FOURIER IS FRONTIER: FREQUENCY-AWARE AUTOENCODING FOR HIGH-FIDELITY MUSIC RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Continuous-latent audio autoencoders form the backbone of latent music generators, and their reconstruction fidelity limits the acoustic detail available to downstream generators. At the compression rates these pipelines require, decoders commonly exhibit three recurring failure modes: high-frequency loss, phase incoherence, and stereo-image collapse. These shortcomings in high-fidelity generation share a structural root: waveform autoencoders, previously regarded as the frontier design, lack an explicit frequency axis, leaving no handle for targeted per-band correction. Among five matched-budget representations in our controlled study, the complex STFT achieves the lowest full-band and high-frequency spectral distances, providing direct access to magnitude and phase at every frequency bin. Building on this finding, we present εar-VAE2, a complex-spectral autoencoder that exploits the explicit frequency structure and introduces cross-channel interaction between the left and right channels. First, Spec-SnakeBeta exploits the additional capacity of the frequency-domain representation by learning a periodic activation for each frequency bin, with frequency-dependent initialization in log-parameter space. This design overall outperforms the other activation variants in our controlled ablation, while using fewer parameters than the fully independent variant. Second, Duplex-Aware Refiner applies band-specific corrections to magnitude and phase. Its band allocation follows duplex theory of sound localization and corrects the decoder's reconstruction error. On the 546-track Song Describer Dataset, εar-VAE2 achieves the best point estimates on five of seven reconstruction metrics and matches the best stereo-coherence score. Adding the Duplex-Aware Refiner module further reduces Mel Distance by 19.4%, and its duplex band allocation uses $\approx 45\%$ fewer residual-output dimensions than the Unconstrained Refiner version. Beyond this efficiency, the banded allocation also lowers spectral distances and targeted spatial-cue error metrics, and it receives higher mean paired ratings from professional mixing and mastering engineers. The downstream generator using εar-VAE2 latents achieves better point estimates on all 12 downstream metrics. Demo page is available at `https://earvae2-2026.github.io/`.

1 INTRODUCTION

Latent-diffusion pipelines are widely used for high-quality music generation (Rombach et al., 2022; Evans et al., 2024). In these systems, a diffusion model operates not on raw audio but on the continuous latent space of a variational autoencoder (VAE), and a decoder maps sampled latents back to a waveform. The decoder therefore limits the acoustic detail available to the downstream system. Indeed, increasing the generator's capacity cannot restore spectral, phase, or stereo information that the autoencoder does not preserve. Under aggressive compression, decoders often lose high-frequency content, degrade phase coherence, and narrow the stereo image. These errors are especially relevant to 48 kHz stereo music, which contains dense harmonics, sharp transients, and spatial cues. Waveform VAEs (Evans et al., 2024; Wang et al., 2025) do not expose a fixed frequency axis for direct per-band parameterization. Many earlier spectral autoencoders instead discarded phase and used a separate vocoder for synthesis. This leads to our central question: *which modeling choices best preserve perceptually salient information at a fixed compression rate?*

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.

To probe this question empirically, we compare five audio representations under matched latent shape and rate, parameter budget, data, and optimization (§4.3). Complex STFT achieves the lowest full-band and high-frequency spectral distances, whereas waveform patches perform better on waveform alignment and low-frequency reconstruction. The comparison therefore identifies wideband spectral detail as the strength of the spectral representation, and this advantage, together with the physical frequency axis it exposes, motivates our spectral-domain design.

Building on this result, we present εar-VAE2, a spectral-domain music autoencoder that preserves the energy distribution of stereo music across both the stereo field and the frequency spectrum, while markedly reducing the metallic high-frequency artifacts common in prior designs. It targets the three failure modes through explicit stereo separation in the complex spectrum, a frequency-aware nonlinearity, and a psychoacoustic refiner. εar-VAE2 compresses 48 kHz stereo music into a 25 Hz sequence of 128-dimensional continuous latents ($1920\times$ temporal downsampling). Its two central contributions are frequency-aware components that the spectral formulation enables.

We make the network's nonlinearity frequency-aware. Spec-SnakeBeta (§3.2) improves frequency-dependent modeling. BigVGAN applies the Snake activation with periodic-residual parameters learned per channel (Lee et al., 2023); the commonly used SnakeBeta variant retains this channel-wise parameterization. Spec-SnakeBeta (§3.2) instead learns one parameter pair per bin, exploiting the exposed frequency axis. This frequency-specific, channel-shared form overall outperforms both a standard version and a fully independent channel–frequency parameterization in the ablation.

The STFT offers a second advantage: it exposes a physical frequency axis while the model compresses along time. Each bin retains a fixed physical interpretation, so a refiner can apply different corrections to different frequency bands. Duplex theory provides a prior for this design: low-frequency localization relies mainly on interaural timing and phase cues, whereas high-frequency localization relies mainly on interaural level cues. Accordingly, the Duplex-Aware Refiner (§3.3) applies phase-only correction below 1.5 kHz, joint magnitude-and-phase correction between 1.5 and 4 kHz, and magnitude-only correction above 4 kHz with fewer output dimensions.

The full system achieves the best point estimates on five of seven reconstruction metrics on the Song Describer Dataset (Manco et al., 2023) and ties for the best stereo coherence. The Duplex-Aware Refiner reduces Mel Distance by 19.4%. In downstream generation, replacing the LeVo 2 VAE with εar-VAE2 yields higher point estimates on all reported automatic metrics under matched evaluation.

Our contributions are:

1. A **frequency-aware continuous audio VAE model** operating on the complex STFT. In a matched-budget study, this representation gives the best spectral distances performance (§4.3).

2. **Spec-SnakeBeta**, a per-bin periodic activation shared across feature channels with frequency-proportional log-space initialization. Across three seeds, it outperforms its fully independent counterpart on all six ablation metrics with $\approx 128\times$ fewer activation parameters (§3.2, §4.6).

3. The **Duplex-Aware Refiner**, a duplex-guided residual module that improves spectral and spatial fidelity with $\approx 45\%$ fewer output residual dimensions than the Unconstrained Refiner (§3.3, §4.5).

2 Related Work

2.1 Discrete neural codecs

The convolutional encoder–decoder with residual vector quantization (RVQ), introduced by SoundStream (Zeghidour et al., 2022) and refined by EnCodec (Défossez et al., 2023) and DAC (Kumar et al., 2023), is the architectural ancestor of most modern audio tokenizers. By quantizing intermediate features into discrete codes at low bitrates, these models produce tokens suited to storage, transmission, and autoregressive generation. The same discreteness, however, fits latent diffusion poorly: quantization commits the representation to a discrete interface, which decreases the reconstruction quality that downstream generation depends on.

2.2 CONTINUOUS WAVEFORM AUTOENCODERS

Continuous-latent autoencoders remove quantization in favor of a smooth latent. RAVE (Caillon & Esling, 2021) learns a KL-regularized continuous VAE latent. More recently, Stable Audio Open (Evans et al., 2024), εar-VAE (Wang et al., 2025), and LeVo 2 (Lei et al., 2026) provide the continuous representations on which modern latent-diffusion music pipelines run. These latents align naturally with diffusion training and match or exceed discrete codecs in downstream generative quality.

These models, however, encode raw waveforms, entangling frequency, phase, and stereo cues sample by sample. Without an explicit frequency axis, the model does not expose direct per-band parameterization by spectral location. This motivates our design for 48 kHz stereo music.

2.3 REVISITING THE SPECTRAL DOMAIN

In a spectral representation, each axis corresponds to a fixed physical frequency, enabling direct per-band operations that waveform models do not natively support. Earlier spectral autoencoders nevertheless operated on magnitude or mel spectrograms, leaving phase reconstruction and waveform synthesis to a separate vocoder (Siuzdak, 2024; Kong et al., 2020; Lee et al., 2023).

Within spectral-domain audio modeling, recent work has focused on conditional vocoding (Siuzdak, 2024) or discrete coding (Ai et al., 2024; Li et al., 2025), leaving spectral representations under-explored for continuous latent autoencoding. Existing representation studies either target GAN generation (Nistal et al., 2020) or change the quantizer together with the representation (Langman et al., 2025), making it difficult to isolate the effect of the representation itself.

Our controlled study addresses this gap by comparing representation choices under matched budgets (§4.3). It identifies the complex STFT as the best fit for full-band and high-frequency spectral preservation, although waveform representations retain advantages on other metrics. Guided by this comparison, we adopt the *complex STFT* for our continuous autoencoder. Its real and imaginary components retain phase information and enable direct iSTFT synthesis without a separate vocoder. The resulting physical frequency axis further supports the frequency-aware operations of §3.

3 METHOD

εar-VAE2 is a spectral-domain VAE-GAN with explicit stereo modeling (Figure 1). A complex-STFT encoder–decoder produces a continuous latent and a coarse reconstruction. The frequency axis is exploited at two complementary levels: Spec-SnakeBeta adapts the internal nonlinear responses across frequency, while the Duplex-Aware Refiner applies frequency-dependent output corrections. The refiner is trained after the encoder–decoder and operates on its frozen reconstruction.

3.1 SPECTRAL ENCODER–DECODER

We transform 48 kHz stereo audio into a complex spectrogram using an STFT with $n_{\text{fft}}{=}960$ ($\Delta f{=}50$ Hz), hop$=480$, and a Hann window. The one-sided STFT of each real-valued channel contains $F_{\text{full}}{=}481$ bins; we exclude the Nyquist bin and operate on the remaining $F{=}480$ retained non-Nyquist frequency bins. The result is $\mathbf{X} \in \mathbb{C}^{2\times F\times T}$, where the leading dimension indexes the left and right stereo signals, f indexes frequency bins, t indexes temporal frames, and T is the number of STFT frames at 100 Hz. Here, $F{=}480$ denotes the input frequency size. Intermediate layers downsample the frequency axis; F_ℓ denotes the number of frequency positions at layer ℓ. C later denotes the network feature-channel count, not the two stereo channels of $\mathbf{X}$. The encoder maps $\mathbf{X}$ to a continuous latent $\mathbf{z} \in \mathbb{R}^{128\times T'}$ at 25 Hz. The decoder reconstructs a coarse complex spectrogram, which is converted to audio by iSTFT. We regularize the latent against a unit Gaussian.

Inspired by SpectroStream (Li et al., 2025), we adopt the 2D-convolutional encoder–decoder as the architecture, but replace the residual-VQ bottleneck with a KL-regularized continuous latent. Delayed channel fusion in the encoder and early channel splitting in the decoder preserve explicit stereo structure. An asymmetric-padding STFT wrapper produces exactly $T{=}L/\text{hop}$ frames while keeping the path differentiable (Appendix D).

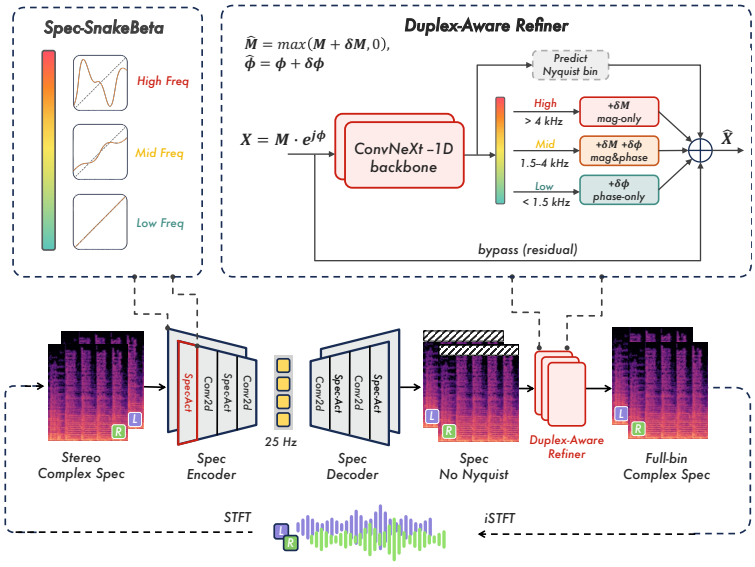

Figure 1: Architecture of εar-VAE2. Stereo audio is transformed via STFT into a complex spectrogram; a Spec Encoder compresses it to a 25 Hz continuous latent and a Spec Decoder reconstructs a coarse complex spectrogram; the highest-frequency (Nyquist) STFT bin is removed to facilitate frequency downsampling. The Duplex-Aware Refiner applies band-specific complex corrections and restores the Nyquist bin to produce the full-bin spectrogram, resynthesized to audio by iSTFT.

3.2 Spec-SnakeBeta Activation

For periodic signals, the Snake activation embeds a learnable periodic residual in an otherwise identity-like activation, improving the representation of periodic functions beyond piecewise-linear nonlinearities (Ziyin et al., 2020). Building on this, BigVGAN transfers this inductive bias to waveform synthesis, learning one periodicity per feature channel together with anti-aliasing filters (Lee et al., 2023). Channel-wise SnakeBeta, a decoupled variant of this activation, $x + \beta^{-1}\sin^2(\alpha x)$, retains the per-channel parameterization and learns one periodic response per channel. Such channel-wise variation suits waveform features, but in a spectral network it applies the same response across all F frequency bins despite their different physical meanings. To our knowledge, prior work has not parameterized this periodic prior along the spectral-frequency axis.

Spec-SnakeBeta instead learns one parameter pair per frequency position and shares it across feature channels. Here c indexes network feature channels, not the left and right audio channels. At layer ℓ, let $x_{b,c,t,f}$ denote the pre-activation at batch index b, feature channel c, frame t, and frequency position $f \in \{0, \ldots, F_\ell - 1\}$. The frequency coordinate is downsampled with the feature map, and $\mathrm{freq}_{\ell,f}$ denotes the corresponding center frequency. In log-space,

$$Spec - SnakeBeta(x)_{b,c,t,f} = x_{b,c,t,f} + \frac{1}{\exp(\tilde{\beta}_f) + \epsilon_{\text{num}}} \sin^2\big(\exp(\tilde{\alpha}_f) \cdot x_{b,c,t,f}\big), \qquad (1)$$

with $\tilde{\alpha}, \tilde{\beta} \in \mathbb{R}^{F_\ell}$ for each layer and $\epsilon_{\text{num}}{=}10^{-9}$. We optimize $\tilde{\alpha}$ and $\tilde{\beta}$ in log space, with $\alpha = e^{\tilde{\alpha}}$ and $\beta = e^{\tilde{\beta}}$, so that $\alpha, \beta > 0$; ϵ_{num} is a numerical stabilizer in the implemented denominator, distinct from the initialization floor ϵ_{init} below. We initialize $\tilde{\alpha}$ proportionally to normalized frequency and $\tilde{\beta}$ uniformly:

$$\tilde{\alpha}_f = \log\Big(\frac{\mathrm{freq}_{\ell,f}}{\bar{f}_\ell} + \epsilon_{\text{init}}\Big), \qquad \bar{f}_\ell = \frac{1}{F_\ell}\sum_f \mathrm{freq}_{\ell,f}, \qquad \tilde{\beta}_f = 0, \qquad (2)$$

($\epsilon_{\text{init}}{=}10^{-6}$; the Nyquist bin is excluded at the input). In α-space, the initialization is linear in normalized frequency, while $\tilde{\beta}{=}0$ gives $\beta{=}1$ at every frequency. This initialization is identity-like at

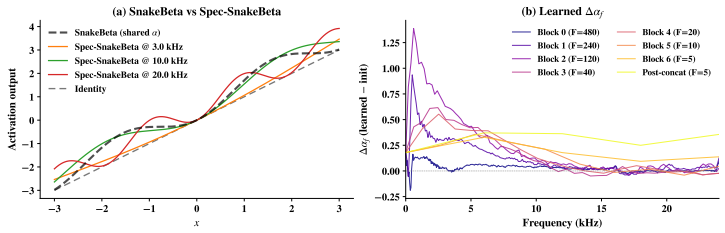

Figure 2: Spec-SnakeBeta analysis. **(a)** Activation curves at representative frequencies (0.5, 3, 10, 20 kHz) vs. channel-wise SnakeBeta (black dashed). **(b)** Learned deviation $\Delta\alpha_f = \alpha_f^{\text{trained}} - \alpha_f^{(0)}$ from frequency-proportional initialization across encoder layers.

low frequencies for bounded pre-activations and increases the feature-space oscillation rate toward high frequencies. Training adapts each frequency response while preserving a common frequency-dependent prior across feature channels. Compared with channel-wise SnakeBeta, this moves the learnable variation from the feature-channel axis to the physical-frequency axis. Because α and β are decoupled, the activation need not remain monotone; the identity skip nevertheless centers the derivative at one, with its range controlled by $\alpha/(\beta + \epsilon_{\text{num}})$. Figure 2(a) illustrates the effect: the frequency-proportional initialization renders the activation identity-like at low frequencies for the plotted input range and progressively more oscillatory toward high frequencies. After training, the learned α_f deviates from initialization primarily in 0–5 kHz (Figure 2(b)).

3.3 Duplex-Aware Refiner

The encoder–decoder already adapts its internal responses across frequency, yet its frozen output retains frequency-structured reconstruction errors. Figure 3 separates these two observations. Panel (a) shows the learned Spec-SnakeBeta parameter $\tilde{\beta}_f$; because it is initialized to zero, the plotted values also give its displacement from initialization. This frequency-dependent adaptation occurs inside the encoder–decoder. In contrast, panels (b–c) measure the error that remains at its output before refinement. Together, these observations motivate a post-hoc refiner for residual output errors.

Our core design insight for the refiner comes from duplex theory, which provides both the cue allocation and its frequency boundaries: human localization relies primarily on interaural timing and phase cues below approximately 1.5 kHz, passes through a transition region from 1.5 to 4 kHz, and relies increasingly on interaural level cues above 4 kHz (Rayleigh, 1907; Macpherson & Middlebrooks, 2002). Interestingly, we find that the frozen decoder's error profile independently follows these theoretically defined regions: magnitude error is largest in the transition band, while phase error rises toward its plateau near the upper boundary. This convergence of theoretical prior and empirical measurement motivates phase-only correction in the Low band, joint magnitude-and-phase correction in the Mid band, and magnitude-only correction in the High band.[1]

As shown in Figure 1, the refiner processes the decoder's complex STFT $X = Me^{j\phi}$ with a lightweight ConvNeXt-1D backbone (Liu et al., 2022). The same module is applied independently to each stereo channel. Its shared output projection allocates residuals by frequency band: the Low band (< 1.5 kHz) receives $\delta\phi$, the Mid band (1.5–4 kHz) receives δM and $\delta\phi$, and the High band (> 4 kHz) receives δM. The projection also predicts the excluded Nyquist bin.

For each stereo channel, the batched input is $X \in \mathbb{C}^{B\times F\times T}$. The backbone maps its real and imaginary components to per-frame residuals over the retained frequency bins. Suppressing the batch, time-frame, and frequency-bin indices, each retained bin is updated as

$$\hat{M} = \max(M + \delta M, \epsilon), \qquad \hat{\phi} = \phi + \delta\phi, \qquad \hat{X} = \hat{M}e^{j\hat{\phi}}. \tag{3}$$

Here, $\epsilon > 0$ ensures numerical stability, $\delta M{=}0$ in the Low band, and $\delta\phi{=}0$ in the High band. The predicted Nyquist bin is appended to these updated bins to form the corrected spectrum. This banded allocation uses fewer output residual dimensions than the Unconstrained Refiner (Appendix F).

[1]These refiner bands (Low/Mid/High, boundaries at 1.5 and 4 kHz) serve a different purpose than the HF (> 8 kHz) / LF (< 2 kHz) metric bands defined in §4.2: the former guides correction, the latter isolates evaluation.

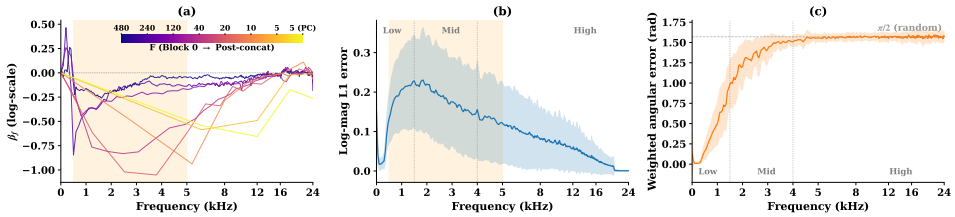

Figure 3: Frequency-resolved diagnosis of the trained encoder–decoder before refinement. **(a)** Learned frequency-indexed $\tilde{\beta}_f$ per encoder layer. **(b)** Per-frequency magnitude error. **(c)** Weighted angular error; the horizontal line marks the $\pi/2$ random-phase baseline. Vertical lines mark the refiner band boundaries, and the shaded region spans 0.5–5 kHz.

All refiner output weights are zero-initialized, so $\hat{X}{=}X$ when training begins and adding the refiner preserves the trained decoder output. The encoder and decoder are frozen; the refiner learns only band-specific corrections.

3.4 Training Objective and Discriminators

3.4.1 Losses

The objective has four base terms. A multi-resolution STFT loss (Yamamoto et al., 2020) combines spectral convergence and log-magnitude L1 at seven FFT scales from 32 to 2048. An Instantaneous Frequency/Group Delay (IF/GD) phase loss follows εar-VAE (Wang et al., 2025) and computes cosine distance on conjugate products along the time and frequency axes. The adversarial (Adv) term combines an LSGAN loss (Mao et al., 2017) with feature matching (FM) (Larsen et al., 2016). The remaining term is KL regularization against a unit Gaussian. Beyond these four terms, we adopt several perceptual techniques: K-weighting (ITU-R, 2015), Mid/Side/Left/Right (MSLR) stereo decomposition (Wang et al., 2025), the adaptive log-magnitude normalization of SAME (Parker et al., 2026), and the mixed-scale spectral loss of SpectroStream (Li et al., 2025). All losses use float64 for numerical stability.

3.4.2 Discriminators

We use three discriminator families along two supervision paths. The waveform-routed path contains multi-scale STFT and constant-Q transform (CQT) discriminators, which operate on the waveform after iSTFT and backpropagate through the synthesis path; we denote them collectively as D_w. The direct spectral path uses a spectral discriminator D_s on the refiner's native complex STFT output before resynthesis. The discriminator architectures are detailed in Appendix G.

3.5 Three-Stage Training Pipeline

Training has three stages: reconstruction pretraining; waveform-adversarial training with STFT and CQT discriminators; and refiner-only training with STFT and IF/GD losses, the retained D_w, and the added D_s. Stage 1 establishes the KL-regularized reconstruction before adversarial gradients are introduced. Stage 2 then optimizes the full encoder–decoder under D_w. Finally, Stage 3 freezes that backbone, drops KL, retains D_w, adds D_s, and trains only the refiner with phase, reconstruction, and adversarial supervision. This separation prevents the refiner from destabilizing the learned latent space. For optimizers, stages 1–2 use the convolution-adapted Muon optimizer (Jordan et al., 2024) with AdamW for remaining parameters; Stage 3 uses AdamW. Details appear in Appendices F and E.

4 Experiments

4.1 Setup

We evaluate reconstruction on 546 tracks in the Song Describer Dataset (Manco et al., 2023). The baselines are four recent open-source audio autoencoders: εar-VAE (Wang et al., 2025), Stable Audio

Open (SA-Open) (Evans et al., 2024), LeVo 2 (Lei et al., 2026), and SAME-L (Parker et al., 2026). Each system uses its native configuration. We report εar-VAE2 both without and with the Duplex-Aware Refiner to measure the refiner's contribution. By contrast, comparisons within εar-VAE2, including the ablations, are controlled. We resample each track to the native sample rate of the model.

The full-scale εar-VAE2 is trained on approximately two million publicly sourced tracks, including a subset of LAION-DISCO-12M (LAION, 2024; Lanzendörfer et al., 2023), plus 10,000 hours of proprietary in-house 48 kHz music. An automated quality pipeline filters the training data (Appendix B).

4.2 Metrics

We report Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) (Le Roux et al., 2019), four spectral distances (STFT Distance, $\text{STFT}_{\log}$ Distance, Mel Distance, and $\text{Mel}_{\log}$ Distance), and two stereo metrics: CCPC (Cross-Channel Phase Coherence) (Wang et al., 2025) and Spectral Pan Error (SPE) (Avendano & Jot, 2002), which averages the absolute error in intensity-derived pan position. Higher is better for SI-SDR, CCPC, and subjective paired ratings (↑); lower is better for the distances and SPE (↓). The four spectral distances are the auraloss magnitude L1 distances on linear STFT and mel spectrograms (Steinmetz & Reiss, 2020), with the log variants applying $\log1\text{p}$ before the same L1 distance; tables abbreviate them as STFT, $\text{STFT}_{\log}$, Mel, and $\text{Mel}_{\log}$. In the input-representation study, STFT_{HF} and STFT_{LF} are STFT Distance restricted to bins above 8 kHz (HF) and below 2 kHz (LF). Appendix C gives the definitions and scale settings for the metrics listed above. The refiner ablation (§4.5) additionally reports band-restricted errors, defined in Appendix K.1.

4.3 Controlled Input-Representation Study

We first test the representation choice that motivates the model. The study compares five compact autoencoders: complex STFT, waveform patch, waveform Conv1D, magnitude-only, and a patch Transformer. Each model has approximately 3.0–3.1M parameters, matched within $\pm5\%$, and uses a 128-dimensional latent at 25 Hz. All variants train for 100k steps with the same data, losses, optimizer, and four-scale STFT discriminator. The complex-STFT and waveform-patch variants also share the same Conv2D backbone. This pair directly compares an axis organized by physical frequency with one organized by within-patch sample position. The full five-way study compares practical representation paradigms under the same macro-level budget (Appendix A).

Table 1: Targeted spectral results from the matched-budget representation study on the Song Describer Dataset. STFT_{HF} / STFT_{LF} are STFT Distance above 8 kHz and below 2 kHz. Complex STFT gives the lowest full-band and high-frequency distances, while waveform patch gives the lowest low-frequency distance. **Bold**: best; <u>underline</u>: second best.

Representation	STFT↓	STFT_{HF}↓	STFT_{LF}↓
Waveform Patch (Conv2D)	1.294	1.513	**0.986**
Complex STFT (real/imag)	**1.122**	**1.178**	<u>1.137</u>
Magnitude-only	<u>1.228</u>	<u>1.376</u>	1.213
Waveform (Conv1D)	1.589	1.740	1.358
Patch Transformer	2.222	2.350	1.749

Complex STFT achieves the lowest full-band STFT Distance (1.122) and the lowest high-frequency distance (1.178), the latter being 22% lower than the waveform-patch result (Table 1). The trade-off is equally clear: waveform patch achieves the best values on three of the four metrics (SI-SDR, CCPC, and Mel Distance) and also gives the lowest low-frequency distance. We therefore do not claim that complex STFT is uniformly better. Instead, the study identifies it as the better fit when full-band and high-frequency spectral preservation are the target. Appendix A reports all metrics.

The shared-backbone comparison supports input organization as a source of this spectral advantage: in the complex-STFT model, neighboring rows carry fixed frequency meaning. Figure 4 compares reconstructed spectrograms.

Table 2: Objective reconstruction on the Song Describer Dataset (546 tracks). **Bold**: best point estimate; underline: second-best. The two εar-VAE2 rows report the pipeline without and with the Duplex-Aware Refiner. Systems operate at their native configurations, so this is a competitive comparison rather than a matched-budget comparison.

	Config		**Temporal**	**Spectral**		**Mel**		**Stereo**	
System	SR	Lat. rate	SI-SDR↑	STFT↓	$\text{STFT}_{\log}$↓	Mel↓	$\text{Mel}_{\log}$↓	CCPC↑	SPE↓
εar-VAE	44.1	43.1	12.4	0.880	0.079	0.509	0.095	**0.973**	0.267
SA-Open	44.1	21.5	6.7	1.016	0.089	0.612	0.106	0.933	0.276
LeVo 2	48	25.0	8.1	0.971	0.086	0.599	0.103	0.947	0.273
SAME-L	44.1	10.8	**12.5**	0.986	0.079	0.539	0.096	0.970	0.276
εar-VAE2 (without refiner)	48	25.0	10.9	0.916	0.078	0.572	0.093	0.966	0.268
εar-VAE2 (with refiner)	48	25.0	11.3	**0.870**	**0.075**	**0.461**	**0.089**	**0.973**	**0.264**

4.4 Main Reconstruction Results

As shown in Table 2, the full εar-VAE2 achieves the best point estimates on five of seven metrics (STFT Distance, $\text{STFT}_{\log}$ Distance, Mel Distance, $\text{Mel}_{\log}$ Distance, and SPE) and matches the best CCPC. The refiner accounts for the largest mel-fidelity gain (Mel Distance -19.4%).

4.5 Efficiency and Perceptual Value of Banded Refinement

The refiner's central design choice is to allocate magnitude and phase corrections by frequency band. In Table 3, we compare the default duplex placement (Low-band phase-only correction, Mid-band joint magnitude-and-phase correction, and High-band magnitude-only correction) against the Unconstrained Refiner, which predicts magnitude and phase corrections everywhere.

The banded head predicts 532 rather than 962 residual dimensions per channel, a reduction of $\approx 45\%$. Despite this reduction, it improves Mel Distance and STFT Distance and reduces the band-restricted interaural phase difference (IPD) and interaural level difference (ILD) errors by 2.4% (Low IPD), 8.1% (Low ILD), and 14.7% (High ILD). These are the duplex cues that the band allocation targets; Appendix K.1 defines both metrics. Ten professional mixing and mastering engineers also give the banded design a higher mean rating in the paired test (0.75 vs. 0.66; protocol in Appendix K).

Table 3: Banded versus Unconstrained Refiner. Both variants use the same frozen decoder and training protocol on the Song Describer Dataset. Low/High IPD (interaural phase difference) and ILD (interaural level difference) are mean absolute deviations from the reference, computed over the Low and High refiner bands. IPD errors are in radians and ILD errors in decibels, so magnitudes are comparable only within a column; Appendix K.1 gives both definitions. Rating is the mean paired rating on a 1–10 scale mapped to $[0, 1]$. **Bold**: best.

	Efficiency	**Spectral**		**Duplex-Cue**				**Subjective**
	Out. dims/ch	Mel↓	STFT↓	Low IPD↓	High IPD↓	Low ILD↓	High ILD↓	Rating↑
Banded (ours)	**532**	**0.658**	**1.019**	**0.7963**	1.2907	**0.1690**	**0.9804**	**0.75**
Unconstrained Refiner	962	0.705	1.120	0.8155	**1.2825**	0.1840	1.1495	0.66

4.6 Ablation: Spec-SnakeBeta Activation

We compare seven configurations to distinguish the effects of the periodic prior, the parameter axis, the frequency-proportional initialization, and channel sharing (Table 4). As a reminder, C here indexes network feature channels, not the left and right audio channels of $\mathbf{X}$. The three parameterizations are C (channel-indexed), F (frequency-indexed and channel-shared), and CF (channel–frequency independent). The controlled setting uses a 42.6M-parameter generator (C_0=32 base width, D=64 bottleneck width) trained for 100k steps on the filtered LAION-DISCO-12M subset. All other hyperparameters are shared; F-Uniform is the F parameterization with zero initialization ($\tilde{\alpha}_f$=$\tilde{\beta}_f$=0); Uniform refers to this parameter initialization, not to frequency sampling. F-Log is the F parameterization stored in log space with frequency-proportional initialization. CF-Log uses the F-Log initialization but optimizes every (c, f) pair independently. Thus, F-Uniform

versus F-Log isolates initialization, C versus F-Uniform changes the indexed axis, and F-Log versus CF-Log isolates channel sharing. Appendix J gives the full protocol.

Table 4: Activation ablation. C is channel-indexed (feature channel, not left/right audio), F is frequency-indexed and channel-shared, and CF is channel–frequency independent. **Bold**: best; underline: second best.

	Temporal	*Spectral*		*Mel*		*Stereo*
Variant	SI-SDR↑	STFT↓	$\text{STFT}_{\log}$↓	Mel↓	$\text{Mel}_{\log}$↓	CCPC↑
ELU	2.38	**1.313**	0.106	1.114	0.124	0.886
SiLU	1.78	1.334	0.104	1.126	0.123	0.883
GELU	1.78	1.382	0.107	1.193	0.125	0.881
SnakeBeta-C	3.40	1.379	0.106	1.194	0.124	0.897
Spec-SnakeBeta-F-Uniform	3.46	1.365	0.106	1.206	**0.120**	0.896
Spec-SnakeBeta-CF-Log	3.51	1.380	0.109	1.223	0.128	0.885
Spec-SnakeBeta-F-Log (ours)	**4.40**	1.315	**0.102**	**1.070**	0.125	**0.908**

The proposed F-Log configuration achieves the best three-seed mean on four of six metrics. Frequency-proportional initialization improves five metrics over F-Uniform, including SI-SDR from 3.46 to 4.40 dB. F-Log also outperforms CF-Log on all six metrics with $\approx 128\times$ fewer activation parameters. These results support a shared frequency response as an effective inductive bias, not a parameter-saving compromise (Appendix J).

4.7 Downstream Generation

We keep the downstream renderer fixed and retrain it per VAE latent space. The renderer produces VAE latents from the conditioning signal; the VAE decoder maps them to audio, optionally followed by the Duplex-Aware Refiner. We compare the LeVo 2 and εar-VAE2 latent spaces. LeVo 2 matches εar-VAE2 in sample rate (48 kHz) and latent rate (25 Hz), keeping temporal resolution and output bandwidth fixed. Each system generates 100 English and 100 Chinese songs, scored with the seven-axis SongBench rubric (Wu et al., 2026) and five-axis SongEval (Yao et al., 2025). Under matched conditioning and guidance, replacing the LeVo 2 VAE with εar-VAE2 gives higher point estimates on all 12 metrics. Appending the Duplex-Aware Refiner at decode time further raises every point estimate (Table 5).

Table 5: Downstream generation evaluation (matched conditioning and CFG; 100 English + 100 Chinese songs per system). The same renderer is retrained per latent space; εar-VAE2(-) omits the refiner. **Bold**: best; underline: second best.

System	**SongBench** (↑)							**SongEval** (↑)				
	Mel-ody	Arran-gement	Music-ality	Vo-cal	Instru-ment	Mix-ing	Struc-ture	Coher-ence	Music-ality	Memor-ability	Clar-ity	Natur-alness
LeVo 2 VAE	6.56	6.88	5.71	6.88	6.95	6.59	6.38	4.31	4.17	4.24	4.20	4.10
εar-VAE2(-)	6.88	7.17	6.04	7.30	7.19	7.02	6.75	4.39	4.25	4.33	4.29	4.18
εar-VAE2	**7.02**	**7.24**	**6.15**	**7.39**	**7.23**	**7.11**	**6.81**	**4.42**	**4.27**	**4.38**	**4.44**	**4.25**

5 Discussion and Conclusion

We presented εar-VAE2, a frequency-aware complex-spectral autoencoder for 48 kHz stereo music. The explicit frequency axis enables two components: Spec-SnakeBeta, a periodic nonlinearity parameterized per frequency bin and shared across feature channels, and the Duplex-Aware Refiner, which applies band-specific magnitude and phase corrections guided by duplex theory. On the Song Describer Dataset, the full system achieves the best point estimates on five of seven reconstruction metrics; in downstream generation, replacing the LeVo 2 latent space with εar-VAE2 yields higher scores on all 12 automatic metrics. The controlled representation, activation, and banded-refinement studies establish relative rankings of individual design choices rather than production-scale limits. An exploratory latent probe (Appendix L) additionally reveals that εar-VAE2 exhibits the largest temporal-frequency inversion ratio among tested models.

AI USE STATEMENT

In this work, we used generative AI tools solely for language polishing and copy-editing of the manuscript. All AI-assisted text was reviewed and verified by the authors, who take responsibility for the final content.

ETHICS STATEMENT

εar-VAE2 is a music reconstruction and representation model. While higher-fidelity audio autoencoders can improve accessibility and creative tools, they also lower the barrier to generating convincing synthetic media and may reproduce copyrighted material present in training data. Our training data undergoes an automated quality-and-provenance pipeline (Appendix B); evaluation uses the publicly available Song Describer Dataset, and we release no cloned-voice or identifiable-artist models.

REPRODUCIBILITY STATEMENT

The architecture, exact-alignment STFT wrapper, loss configuration, discriminator design, and three-stage training schedule are specified in §3 and Table 8, with additional hyperparameters provided in Appendix F. Metric definitions appear in §4.2 and Appendix C. The controlled studies detail their protocols in Appendices A, J, and K. All controlled studies and the final cross-system evaluation use the public Song Describer Dataset (Manco et al., 2023). The full-scale model is trained on approximately two million publicly obtained training tracks and approximately 10,000 hours of proprietary in-house music (Appendix B). Detailed source composition cannot be released because of data-strategy and licensing restrictions; evaluation data and metric definitions are public and reproducible.

REFERENCES

Yang Ai, Xiao-Hang Jiang, Ye-Xin Lu, Hui-Peng Du, and Zhen-Hua Ling. APCodec: A neural audio codec with parallel amplitude and phase spectrum encoding and decoding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3256–3269, 2024. doi:10.1109/TASLP.2024.3417347.

Carlos Avendano and Jean-Marc Jot. Frequency domain techniques for stereo to multichannel upmix. In *AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, 2002.

Antoine Caillon and Philippe Esling. RAVE: A variational autoencoder for fast and high-quality neural audio synthesis, 2021.

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

Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. *arXiv preprint arXiv:2407.14358*, 2024.

ITU-R. BS.1770-4: Algorithms to measure audio programme loudness and true-peak audio level. Technical report, International Telecommunication Union, 2015.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. `https://kellerjordan.github.io/posts/muon/`, 2024.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis, 2020.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved RVQGAN. In *Advances in Neural Information Processing Systems*, 2023.

LAION. LAION-DISCO-12M. `https://laion.ai/blog/laion-disco-12m/`, 2024.

Ryan Langman, Ante Jukić, Kunal Dhawan, Nithin Rao Koluguri, and Jason Li. Spectral codecs: Improving non-autoregressive speech synthesis with spectrogram-based audio codecs, 2025. URL `https://arxiv.org/abs/2406.05298`.

Luca A. Lanzendörfer, Florian Grötschla, Emil Funke, and Roger Wattenhofer. DISCO-10M: A large-scale music dataset. In *Advances in Neural Information Processing Systems*, 2023.

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *International Conference on Machine Learning*, pp. 1558–1566, 2016.

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

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

Shun Lei, Huaicheng Zhang, Dapeng Wu, Yaoxun Xu, Lishi Zuo, Wei Tan, Hangting Chen, Guangzheng Li, Jianwei Yu, Zhiyong Wu, and Dong Yu. LeVo 2: Stable and melodious song generation via hierarchical representation modeling and progressive post-training, 2026. URL `https://arxiv.org/abs/2606.30642`.

Yunpeng Li, Kehang Han, Brian McWilliams, Zalán Borsos, and Marco Tagliasacchi. SpectroStream: A versatile neural codec for general audio. *arXiv preprint arXiv:2508.05207*, 2025.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022.

Ewan A. Macpherson and John C. Middlebrooks. Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited. *The Journal of the Acoustical Society of America*, 111(5):2219–2236, 2002. doi:10.1121/1.1471898.

Ilaria Manco, Benno Weck, Seungheon Doh, Minz Won, Yixiao Zhang, Dmitry Bogdanov, Yusong Wu, Ke Chen, Philip Tovstogan, Emmanouil Benetos, Elio Quinton, György Fazekas, and Juhan Nam. The song describer dataset: A corpus of audio captions for music-and-language evaluation. In *Machine Learning for Audio Workshop at NeurIPS*, 2023.

Xudong Mao, Qing Li, Haoran Xie, Raymond Y.K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *IEEE International Conference on Computer Vision*, pp. 2794–2802, 2017.

Javier Nistal, Stefan Lattner, and Gaël Richard. Comparing representations for audio synthesis using generative adversarial networks. In *28th European Signal Processing Conference (EUSIPCO)*, pp. 161–165. IEEE, 2020.

Julian D. Parker, Zach Evans, CJ Carr, Zachary Zukowski, Josiah Taylor, Matthew Rice, and Jordi Pons. SAME: A semantically-aligned music autoencoder, 2026. URL https://arxiv.org/abs/2605.18613.

Lord Rayleigh. Xii. on our perception of sound direction. *Philosophical Magazine*, 13(74):214–232, 1907. doi:10.1080/14786440709463595.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Hubert Siuzdak. Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis. In *International Conference on Learning Representations*, 2024.

Christian J. Steinmetz and Joshua D. Reiss. auraloss: Audio focused loss functions in PyTorch. In *Digital Music Research Network One-day Workshop*, 2020.

Kangdi Wang, Zhiyue Wu, Dinghao Zhou, Rui Lin, Junyu Dai, and Tao Jiang. Back to ear: Perceptually driven high fidelity music reconstruction. *arXiv preprint arXiv:2509.14912*, 2025.

Dapeng Wu, Shun Lei, Wei Tan, Guangzheng Li, Yunzhe Wang, Huaicheng Zhang, Lishi Zuo, and Zhiyong Wu. SongBench: A fine-grained multi-aspect benchmark for song quality assessment, 2026. URL https://arxiv.org/abs/2604.25937.

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 *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6199–6203, 2020.

Jixun Yao, Guobin Ma, Huixin Xue, Huakang Chen, Chunbo Hao, Yuepeng Jiang, Haohe Liu, Ruibin Yuan, Jin Xu, Wei Xue, Hao Liu, and Lei Xie. Songeval: A benchmark dataset for song aesthetics evaluation, 2025. URL https://arxiv.org/abs/2505.10793.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. SoundStream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2022.

Liu Ziyin, Tilman Hartwig, and Masahito Ueda. Neural networks fail to learn periodic functions and how to fix it. In *Advances in Neural Information Processing Systems*, 2020.

A Input-Representation Study: Full Protocol and Extended Results

A.1 Protocol

Each autoencoder has approximately 3.0–3.1M parameters, matched within $\pm 5\%$ by counting `sum(p.numel())`. Every model uses the same latent tensor $(B, 128, 25{\cdot}s)$, with 128 dimensions at 25 Hz. All variants train for 100k steps with the same multi-resolution STFT loss, feature-matching loss, optimizer, and 4-scale STFT discriminator. We omit the CQT discriminator to remove it as a confounding factor. The spectral variants also share the same STFT parameters (n_{fft}=960, hop=480). Training uses a public music subset from the collections described in Appendix B. We evaluate reconstruction on the Song Describer Dataset (Manco et al., 2023) and report the mean metric value across files.

We compare five paradigms: **Complex STFT** (real/imaginary, Conv2D), **Waveform Patch** (Conv2D), **Waveform** (Conv1D), **Magnitude-only** (with a Vocos-style iSTFT head), and **Patch Transformer**. The waveform-patch model divides the raw waveform into 480-sample patches, reshapes them into a 2D grid, and uses the same Conv2D backbone as the complex-STFT model. Thus, the vertical axis represents physical frequency in one model and within-patch sample position in the other. The waveform-patch model operates entirely in the time domain (waveform $\rightarrow$ patch grid $\rightarrow$ network $\rightarrow$ grid $\rightarrow$ waveform). It achieves the best values on the waveform-alignment metrics in this compact comparison, while complex STFT achieves the lowest full-band and high-frequency spectral distances.

This controlled study is designed to establish the *relative ranking* of input-representation paradigms. The compact model scale means that absolute reconstruction quality (e.g., negative SI-SDR for spectral paradigms) is far from production-level, and the production system described in the main paper is trained at full scale. The paradigm-to-implementation mapping is given in §4.3.

A.2 Full results (five paradigms)

Table 6 presents the full per-paradigm results. On aggregate metrics, the waveform-patch (Conv2D) paradigm achieves the best SI-SDR, CCPC, and Mel Distance values. Relative to complex STFT, it shows weaker full-band and high-frequency spectral preservation, with STFT and HF distances of 1.294 and 1.513, respectively. The patch transformer trails on every metric at this compact budget, but replacing its Transformer backbone with the same Conv2D yields the waveform-patch paradigm (the aggregate-metric leader), showing that patch tokenization alone does not explain the deficit; backbone inductive bias and optimization may also contribute.

Table 6: Input-representation study on the Song Describer Dataset (matched budget). **Bold**: best; underline: second best. STFT_{HF} / STFT_{LF} are STFT Distance above 8 kHz and below 2 kHz (metric bands, not the refiner Low/Mid/High bands).

	Aggregate				*Freq-Band*	
Representation	SI-SDR↑	CCPC↑	Mel↓	STFT↓	STFT_{HF}↓	STFT_{LF}↓
Waveform Patch (Conv2D)	**+3.98**	**0.865**	**1.184**	1.294	1.513	**0.986**
Complex STFT (real/imag)	−10.49	0.800	1.523	**1.122**	**1.178**	1.137
Magnitude-only	−42.16	0.756	1.568	1.228	1.376	1.213
Waveform (Conv1D)	−33.56	0.732	2.009	1.589	1.740	1.358
Patch Transformer	−52.04	0.697	2.828	2.222	2.350	1.749

A.3 Training stability

Figure 5 shows the encoder gradient ℓ_2-norm and magnitude-spectrogram reconstruction loss recorded during the same 100k-step runs. The complex STFT paradigm maintains the flattest gradient-norm trajectory with the least variance and converges to the lowest magnitude loss under this shared training protocol, complementing its full-band and high-frequency reconstruction results.

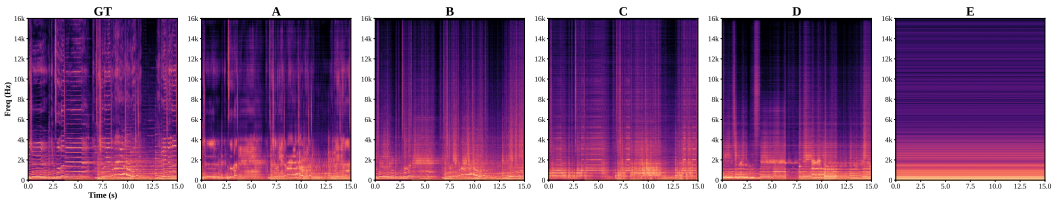

Figure 4: Reconstruction spectrograms from the matched-budget representation study (48 kHz, n_{fft}=2048, hop=128). **(A)** complex STFT; **(B)** waveform patch; **(C)** magnitude-only; **(D)** waveform Conv1D; **(E)** patch Transformer; and **(GT)** ground truth. Complex STFT and waveform patch share the same Conv2D backbone.

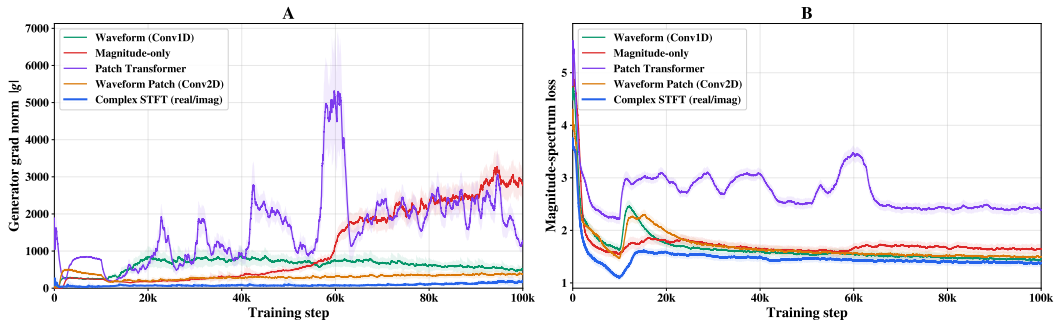

Figure 5: Encoder gradient ℓ_2-norm (left) and magnitude reconstruction loss (right) over training, recorded during the same 100k-step runs under the shared protocol.

B Training Data and Quality Pipeline

The full-scale εar-VAE2 is trained on approximately two million publicly sourced tracks, including LAION-DISCO-12M (LAION, 2024; Lanzendörfer et al., 2023), and approximately 10,000 hours of proprietary in-house 48 kHz stereo music. Detailed source composition is withheld for data-strategy and licensing reasons. Before training, all sources pass through an automated quality and provenance pipeline. Raw collections may contain lossless containers transcoded from lossy sources, duplicated mono tracks labeled as stereo, clipped masters, and upsampled files with empty high-frequency bands. The pipeline (Table 7) checks encoder tags, metadata, bitrate statistics, loudness, true peaks, left–right difference energy, and spectral cutoffs.

B.1 Noise-floor-aware cutoff detection

Container metadata alone cannot distinguish true high-resolution content from an upsampled file. We estimate the noise floor from track *edges* and detect the spectral cutoff from *interior* segments, keeping the two independent. We reserve the first and last 10% of the mono waveform as edges and compute Hann-windowed 8192-point RFFTs over non-overlapping windows.

Define the peak-normalized magnitude in decibels as

$$M_{\text{dB}}(f) = 20 \log_{10} \frac{\bar{M}(f)}{\max_{f'} \bar{M}(f') + \epsilon_M}, \qquad \epsilon_M > 0. \tag{4}$$

The empirical noise floor is

$$\hat{\eta} = \text{median}\left(M_{\text{dB,edge}}[\text{top } 10\% \text{ bins}]\right). \tag{5}$$

The detected cutoff f_c is the highest bin whose mean interior magnitude stays above -80 dB relative to the peak; the local roll-off is

$$\Delta_{\text{dB}} = \max M_{\text{dB,int}}[f_c{-}2\text{kHz} : f_c] - \max M_{\text{dB,int}}[f_c : f_c{+}2\text{kHz}]. \tag{6}$$

We replace the post-cutoff maximum with $\hat{\eta}$ when no reliable content exists above f_c. A file is `hard_cutoff` when $f_c/f_{\text{Nyq}} < 0.85$ and $\Delta_{\text{dB}} \geq 40$ dB (or ≥ 25 dB near $\hat{\eta}$), `suspicious` when $f_c/f_{\text{Nyq}} < 0.90$ with $\Delta_{\text{dB}} \geq 20$ dB, and `natural` otherwise (Figure 6).

Table 7: Metric-level checklist used by the acoustic pipeline. Each row pairs a rule-engine identifier with a brief description and its pass/flag condition.

Check	Metric ID	Evidence	Condition
Meta & encoder	`encoder_tag`	Encoder string in container/stream tags	flag if `lavf/lavc/ffmpeg/lame`
	`mutagen_tags`	Metadata sweep for conversion traces	flag if `transcode/convert`
Bitrate	`has_non_monotonic_pts`	PTS monotonicity	pass if `false`
	`bitrate_utilization`	Realized / nominal bitrate	pass if ≥ 0.6
	`top10_bitrate_mean`	Complex-segment mean bitrate	pass if $\geq 1.2\times$ mean
	`vbr_variation`	Bitrate coefficient of variation	pass if $\sigma/\mu \geq 0.15$
	`quality_tier`	Bitrate-based quality level	HiRes/High/Med/Low/Trash
Loudness	`lufs_i`	Integrated loudness (EBU R128)	$[-25, -5]$ LUFS
	`true_peak`	dBTP level	$[-1, +2]$ dBFS
	`lra`	Loudness range	$[3, 15]$ LU
	`plr`	Peak-to-loudness ratio	≥ 6 LU
Fake stereo	`lr_diff_zero_ratio`	Fraction of zero $\lvert L - R \rvert$ samples	pass if < 0.99
Freq cutoff	`cutoff_ratio`	Cutoff / Nyquist freq ratio	ranking from $0.5 \rightarrow 1.0$
	`drop_db`	Energy gap around f_c	hard ≥ 40 dB; suspicious ≥ 20 dB
	`valid_segments`	Non-silent interior FFT windows	pass if > 0

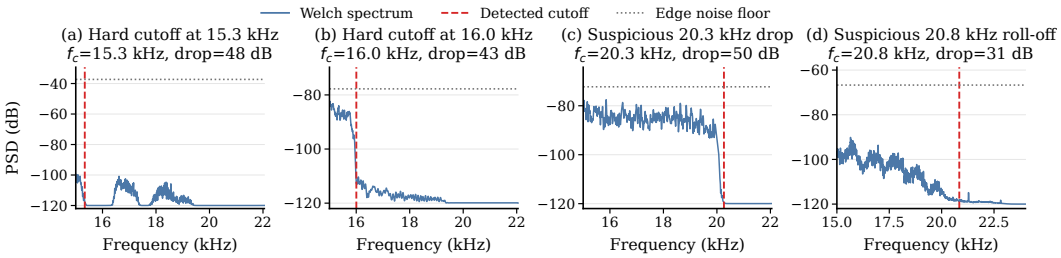

Figure 6: High-frequency cutoff detector examples. Blue: Welch spectrum of a verification segment; red dashed: detected cutoff f_c; gray dotted: independently estimated edge noise floor $\hat{\eta}$.

C Metric Definitions

This appendix gives the metric details omitted from the main text (§4.2). Throughout this appendix, S (with hats for reconstructions) denotes a generic spectrogram in metric formulas, distinct from the native-model STFT $\mathbf{X}$ and the per-channel refiner spectrogram X.

STFT Distance and STFT_{log} Distance. We follow the STFT magnitude distance in `auraloss` (Steinmetz & Reiss, 2020) (`STFTMagnitudeLoss`, L1, mean reduction), not spectral convergence. Spectral convergence is the separate Frobenius-ratio term $\| |S| - |\hat{S}| \|_F / \| |S| \|_F$; it appears in the training multi-resolution STFT loss (§3.4) but is not a reported evaluation metric. Let $|S_k|$ and $|\hat{S}_k|$ denote reference and reconstructed magnitudes at scale k, with $n_{\text{fft}} \in \{64, 128, 256, 512, 1024, 2048\}$, hop $= n_{\text{fft}}/4$, a Hann window, and $N_k = F_k T_k$ time–frequency bins. Averaging over K resolutions,

$$\text{STFT} = \frac{1}{K} \sum_k \frac{1}{N_k} \big\| |S_k| - |\hat{S}_k| \big\|_1. \tag{7}$$

STFT_{log} Distance uses that same L1 magnitude distance after $\log1\text{p}(x) = \log(1 + x)$, matching auraloss log compression with `log_fac`=1 and `log_eps`=1:

$$\text{STFT}_{\text{log}} = \frac{1}{K} \sum_k \frac{1}{N_k} \big\| \log1\text{p}(|S_k|) - \log1\text{p}(|\hat{S}_k|) \big\|_1. \tag{8}$$

Mel Distance and $\text{Mel}_{\log}$ Distance. **Mel Distance** and **$\text{Mel}_{\log}$ Distance** replace $|S_k|$ with a 64-bin mel spectrogram and otherwise use the same two auraloss magnitude L1 distances, abbreviated Mel and $\text{Mel}_{\log}$.

HF and LF metric bands. In the input-representation study (§4.3), STFT_{HF} and STFT_{LF} denote STFT Distance restricted to bins above 8 kHz and below 2 kHz, respectively. These HF/LF metric bands are distinct from the refiner's Low (<1.5 kHz), Mid (1.5–4 kHz), and High (>4 kHz) correction bands (§3.3).

Spectral Pan Error (SPE). The per-bin, per-frame pan position is defined via the intensity panning law (Avendano & Jot, 2002),

$$\text{pan}(f,t) = \frac{|L(f,t)|^2}{|L(f,t)|^2 + |R(f,t)|^2 + \epsilon_{\text{pan}}}, \quad \epsilon_{\text{pan}} = 10^{-8}. \tag{9}$$

At each scale, left and right channel powers are computed from Hann-windowed complex STFTs. SPE is the mean L1 deviation between reconstructed and reference pan profiles over all $N_k = F_k T_k$ time–frequency bins, averaged over K=4 STFT scales ($n_{\text{fft}} \in \{512, 1024, 2048, 4096\}$):

$$\text{SPE} = \frac{1}{K} \sum_k \frac{1}{N_k} \sum_{f,t} \left| \widehat{\text{pan}}_k(f,t) - \text{pan}_k(f,t) \right|, \tag{10}$$

where $\widehat{\text{pan}}_k$ and pan_k denote reconstructed and reference pan profiles, respectively. The additive ϵ_{pan} stabilizes the ratio near zero energy and is distinct from ϵ_{num} and ϵ_{init} in Spec-SnakeBeta; no silence mask is applied before averaging.

D Exact-Alignment STFT Wrapper

Let L be the waveform length, N the FFT size, and H the hop size. For waveforms whose length L is divisible by the hop size H, preserving the $1920\times$ temporal ratio requires exactly $T = L/H$ STFT frames. PyTorch's default centered analysis produces $1 + \lfloor L/H \rfloor = T + 1$ frames. Removing the extra frame to meet the network interface also removes one hop of synthesis support at the boundary, so the waveform must be shortened or padded rather than reconstructed at its original length. We therefore avoid this frame-count–boundary-coverage trade-off.

Disabling implicit padding does not by itself solve the problem. A Hann-windowed `center=False` inverse must satisfy the Nonzero-Overlap-Add (NOLA) condition at every finite-signal boundary: the sum of squared, overlapping synthesis-window values must be nonzero at every reconstructed sample. Because the Hann window vanishes at its endpoint, PyTorch's boundary check rejects the unpadded configuration used here (N=960, H=480), blocking a differentiable STFT–iSTFT path. Likewise, mixing center conventions without an explicit boundary construction changes the temporal origin. We instead define that construction directly in Algorithm 1.

Algorithm 1 Exact-alignment STFT–iSTFT wrapper.

Require: Waveform $x \in \mathbb{R}^{B\times L}$; FFT size N; hop H; window length W; window w
Ensure: Spectrum with $T = L/H$ frames and reconstruction $\hat{x}$ of length L

procedure ANALYSIS(x, N, H, W, w)
  $p_\ell \leftarrow N/2, p_r \leftarrow H/2$
  $x_{\text{pad}} \leftarrow \text{ZEROPAD}(x,\ p_\ell,\ p_r)$
  $X \leftarrow \text{STFT}(x_{\text{pad}}, N, H, W, w,$ `center=False`$)$
  return X
end procedure
procedure SYNTHESIS(X, N, H, W, w, L)
  $p_r \leftarrow H/2$
  $\tilde{x} \leftarrow \text{ISTFT}(X, N, H, W, w,$ `center=True`, `length` $= L + p_r)$
  return $\tilde{x}[:, 0{:}L]$
end procedure

For our $N{=}2H$ setting, the asymmetric analysis padding gives $1 + \lfloor (L + N/2 + H/2 - N)/H \rfloor = L/H$ frames. The left pad moves original sample zero into a nonzero-overlap region and is removed by the centered inverse; requesting $L + H/2$ synthesis samples supplies the explicit tail margin, which is then cropped. The wrapper therefore returns exactly L aligned samples, avoids the rejected NOLA boundary, and remains fully differentiable. Figure 7 shows the sample-level effect.

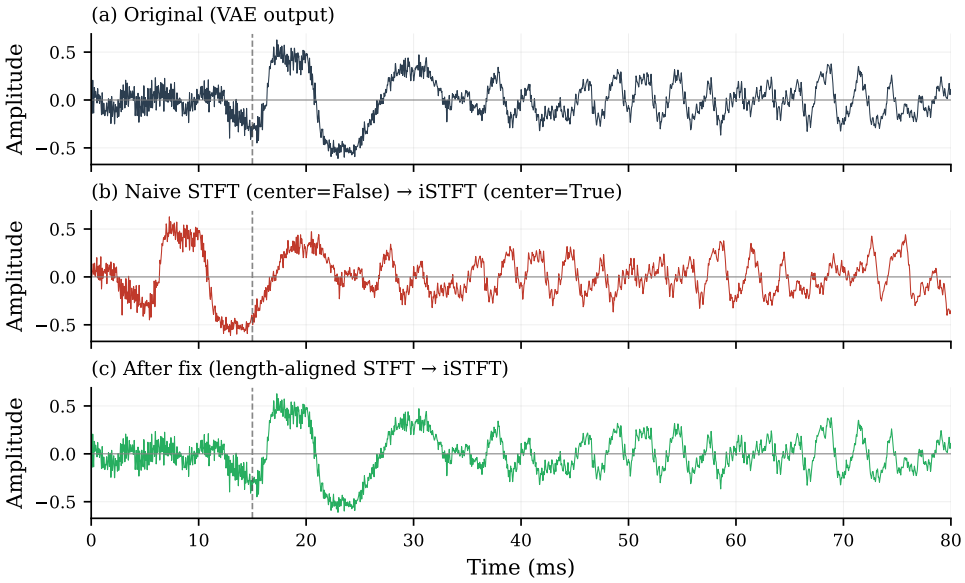

Figure 7: Sample-level comparison over 80 ms at 48 kHz; the dashed line marks the same reference instant in every panel. **(a)** Original waveform. **(b)** A naive mixed-center round-trip (analysis `center=False`, synthesis `center=True`) shifts the waveform left by one hop (≈10 ms). A fully uncentered Hann round-trip is instead rejected by the boundary NOLA check and therefore has no reconstructed trace. **(c)** Our explicitly padded wrapper produces T frames and recovers sample-level alignment.

E Convolution-Adapted Muon Optimizer

The Muon optimizer (Jordan et al., 2024) orthogonalizes each momentum update with a quintic Newton–Schulz iteration. It was designed for the 2D weight matrices used in transformers. To apply it to convolutional spectral networks, we flatten every $\geq$2D kernel into an output-channel-by-fan-in matrix, orthogonalize the update, and reshape it to the original kernel dimensions.

The Muon optimizer updates all $\geq$2D convolutional and linear weights except output projections and discriminator post-convolutions. These output layers remain on the internal AdamW branch, together with bias and normalization parameters. We retain the standard $0.2\sqrt{\max(A, B)}$ shape-dependent learning-rate scaling. Stages 1–2 use the Muon optimizer to reduce gradient spikes during adversarial training and to improve reconstructed high-frequency detail. Stage 3 uses AdamW for the refiner.

F Architecture, Hyperparameters, and Training Configuration

The production εar-VAE2 comprises a Spec Encoder with base width $C_0{=}256$ ($\approx$549M parameters), a Spec Decoder with $C_0{=}128$ ($\approx$134.8M), and a banded Duplex-Aware Refiner ($\approx$8.2M), totaling $\approx$692M parameters. The STFT front end employs $n_{\text{fft}}{=}960$ ($\Delta f{=}50$ Hz at 48 kHz), hop$=$480, and the exact-alignment wrapper of Appendix D. The latent is 128-dimensional at 25 Hz. The refiner is a 12-layer, 256-dimensional ConvNeXt-1D backbone reading a 960-dimensional per-frame real–imaginary spectral vector and producing 532 per-frame residual dimensions per channel (1064 across the stereo pair) with the band assignment described in §3.3. Training uses 1.28 s segments; the three-stage schedule and loss weights are specified in Table 8. For the generator in Stages 1–2, we

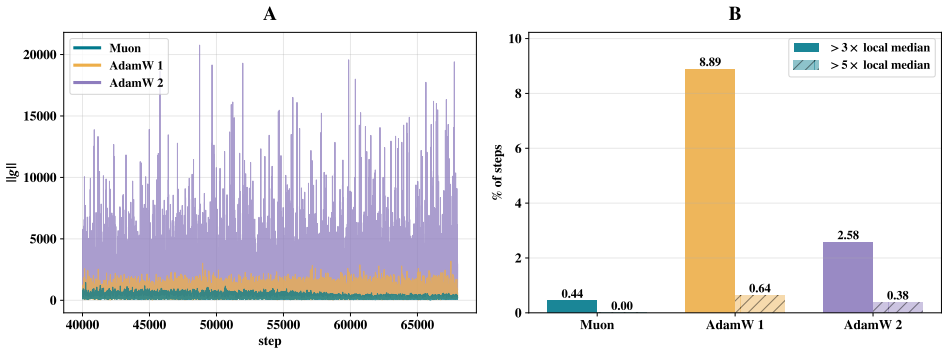

Figure 8: Generator gradient-norm dynamics for Muon and AdamW, measured over a common training window (steps 40k–68k) in which all three runs are stable. **(A)** Per-step generator gradient norm $\|g\|$ for Muon and two AdamW runs (AdamW 1, AdamW 2), shown on a shared linear scale; Muon and AdamW 2 use the same global batch size and AdamW 1 uses a larger batch. **(B)** Local spike rate: the percentage of steps at which $\|g\|$ exceeds $3\times$ (solid) or $5\times$ (hatched) its trailing 500-step median.

employ the Muon optimizer (momentum orthogonalized by a quintic Newton–Schulz iteration) with implementation adaptations for the stereo spectral setting. All Spec-SnakeBeta layers use the F-Log configuration: layer ℓ learns one α_f, β_f pair for each of its F_ℓ frequency positions and shares it across feature channels. We store $\tilde{\alpha}$ and $\tilde{\beta}$ and recover $\alpha = e^{\tilde{\alpha}}$, $\beta = e^{\tilde{\beta}}$.

Table 8 summarizes the three-stage schedule. The multi-resolution STFT loss uses seven FFT sizes from 32 to 2048. The adversarial stack contains an eight-scale STFT discriminator with per-channel $[\mathrm{Re}(X), \mathrm{Im}(X), |X|]$ inputs. Five scales use $n_{\text{fft}} \in \{128, 256, 512, 1024, 2048\}$ with proportionally scaled hops; three scales use n_{fft}=4096 with hops 2048, 1024, and 256. The stack also includes three CQT sub-discriminators at 24, 36, and 48 bins per octave over nine octaves. Stage 3 additionally applies four-scale adaptive frequency pooling to the refiner's native complex STFT output. Training runs on one node with 8 NVIDIA A100 GPUs; optimizer state is reset between stages.

Table 8: εar-VAE2 training configuration across three stages. Stage 1 is reconstruction-only; Stage 2 adds waveform-domain adversarial training; Stage 3 freezes the encoder–decoder and trains the refiner with D_w and D_s. Adv/FM denotes adversarial (Adv) / feature-matching (FM) loss weights.

Category	Param	Stage 1	Stage 2	Stage 3
Optimizer	Generator optim.	Muon	Muon	AdamW
	Generator LR	10^{-4}	1.5×10^{-4}	1.5×10^{-4}
	Disc. optim.	—	Muon	AdamW
	Disc. LR	—	3×10^{-5}	$10^{-6}(D_w)$ / $10^{-5}(D_s)$
Schedule	Training duration	1 epoch	5 epochs	3 epochs
	Batch size / GPU	4	1	1
	Segment length	1.28 s	1.28 s	1.28 s
	Disc. warmup	—	0	0(D_w) / 15k(D_s)
	G:D interleave	—	1:1	4:1
Disc.	STFT disc. scales	—	8	8
	CQT disc.	×	✓	✓
	Spectral disc.	×	×	✓
Loss weights	λ_{STFT}	1.0	1.0	0.5
	$\lambda_{\text{IF/GD}}$	0	0	0.1
	λ_{KL}	10^{-6}	10^{-6}	0
	$\lambda_{\text{time-GAN}}$ (Adv/FM)	—	1.0/100	0.175/20
	$\lambda_{\text{spec-GAN}}$ (Adv/FM)	—	—	0.5/35

Table 9: Discriminability probe at the Stage 3 boundary (n=48 segments): mean real/fake logits and real–fake gap. $D_w^{(2)}$/$D_w^{(3)}$ are D_w at the end of Stage 2 and after Stage 3 continuation; D_s is the trained spectral-domain discriminator; $D_s \circ P$ scores inputs pre-projected onto the consistent subspace (gauge component removed); D_s^{rand} is a randomly initialized spectral-domain discriminator.

Critic	logit(real)	logit(fake)	gap	Note
$D_w^{(2)}$	0.619	0.409	0.210	Stage 2 checkpoint
$D_w^{(3)}$	0.551	0.366	0.186	after Stage 3 continuation
D_s	0.485	0.034	**0.451**	trained spectral
$D_s \circ P$	0.485	0.449	0.036	gauge removed
D_s^{rand}	−0.003	−0.002	0.001	random init

G Discriminator Architectures

The **multi-scale STFT discriminator** addresses the inability of a single STFT resolution to capture both short transients and fine harmonic structure. Five sub-discriminators use dyadic FFT sizes $n_{\text{fft}} \in \{128, 256, 512, 1024, 2048\}$ with hop=$n_{\text{fft}}/2$, maintaining 50% overlap while spanning multiple time–frequency scales. Three additional sub-discriminators fix n_{fft}=4096 and use hops $\{2048, 1024, 256\}$. At 48 kHz, this FFT size gives $\Delta f \approx 11.7$ Hz; varying the hop provides three temporal sampling densities at the same frequency resolution. For the per-channel waveform w, each sub-discriminator receives $[\mathrm{Re}(S_r(w)), \mathrm{Im}(S_r(w)), |S_r(w)|]$, where S_r denotes its STFT configuration.

The **CQT discriminator** complements these linear-frequency views with a log-frequency representation aligned with musical pitch. It contains three sub-discriminators with $\{24, 36, 48\}$ bins per octave over nine octaves.

The **spectral discriminator** D_s, introduced only in Stage 3, evaluates the refiner's native complex STFT output using adaptive frequency pooling at four scales. Waveform-routed objectives can constrain only changes observable after iSTFT, whereas D_s directly regularizes native spectral directions that are not identifiable through the waveform path. At the Stage 3 boundary, the identity-initialized refiner reproduces the frozen Stage 2 decoder output, on which D_w has already been trained. After its warmup, D_s therefore provides a newly learned, complementary direct-spectral signal for optimizing the residual correction. Appendix H analyzes these two motivations through synthesis-null-space and Stage 3 boundary probes.

H Why a Spectral-Domain Discriminator in Stage 3?

This section is a *diagnostic exploration* of the Stage 3 design, not a core claim. All probes are training-free (forward and backward passes only; no weights are updated) and use our trained generator and discriminators; probe inputs are 4 s excerpts from our internal evaluation set. The structural quantities of §H.2 depend only on the STFT configuration, not on model capacity.

H.1 Starting point: supervision at the Stage 3 boundary

The first two rows of Table 9 probe D_w at the ends of Stages 2 and 3, respectively. At the Stage 2 checkpoint, its mean real and fake logits are 0.62 and 0.41, respectively, leaving a gap of 0.21. This gap does not establish equilibrium, but D_w has already been trained on these outputs. Stage 3 then attaches the refiner, whose zero-initialized residuals make its initial output equal to the Stage 2 decoder output (§3.3). We therefore test whether direct spectral supervision provides output-space gradients that are unavailable through the iSTFT path.

H.2 Making the conjecture precise: the synthesis-null subspace

With n_{fft}=960 and hop 480, the one-sided complex spectrogram has approximately twice as many real output coordinates as the waveform representation under this STFT configuration. The exact finite-signal ratio depends on boundary handling and on the real constraints at the DC and Nyquist

Table 10: Gradient anatomy at the Stage 3 boundary (n=24): fraction of each generator-side gradient's energy inside the synthesis-null subspace. $g_w^{(2)}, g_w^{(3)}$ are waveform-path gradients (through the iSTFT) from Stage 2 and Stage 3 D_w; g_s is D_s's direct gradient.

Gradient	null-space energy fraction
$g_w^{(2)}$ (Stage 2 D_w)	0.0002
$g_w^{(3)}$ (Stage 3 D_w)	0.0002
g_s (D_s, direct)	**0.9529**

bins, but the synthesis-null subspace occupies approximately half of the real output dimension. Define the round-trip operator $P = \mathrm{STFT} \circ \mathrm{iSTFT}$, which is a near-orthogonal projector onto the consistent-spectrogram subspace. Any differentiable objective that reaches a spectrogram S only through $w = \mathrm{iSTFT}(S)$ has zero directional derivative along $\ker(\mathrm{iSTFT})$, regardless of the transform applied by the critic. Perturbations in this kernel leave the resynthesized waveform unchanged. They are redundant directions of the representation, not lost acoustic information. Thus, waveform-routed supervision cannot constrain these directions.

H.3 What the spectral discriminator supervises

Three probes characterize what D_s responds to at the Stage 3 boundary.

The null subspace is populated. The Stage 2 output lies measurably outside the consistent subspace: its inconsistency $\|S - PS\|/\|S\|$ is 0.234 ± 0.035 (n=48), i.e., about 5.5% of the squared spectral norm lies in the null component that waveform-routed objectives cannot differentiate.

The discriminator's gradient concentrates in the null subspace. Table 10 dissects the generator-side gradients (adversarial plus feature-matching terms, weighted as in training) with respect to S. Gradients routed through the iSTFT retain only 0.02% of their energy in the null subspace, which is the predicted zero up to boundary effects. In contrast, 95.3% of the spectral discriminator's gradient energy lies inside it. The two paths are nearly orthogonal: $\cos(g_w^{(3)}, g_s) = -0.002 \pm 0.003$ over the full space and -0.012 ± 0.013 within the consistent subspace (n=24).

Selectivity controls. We tested the alternative reading in which D_s mainly provides a finer content view. The controls disfavor this as the dominant explanation. Pre-projecting its inputs onto the consistent subspace collapses the trained discriminator's real–fake gap from 0.45 to 0.04 (Table 9), while a randomly initialized spectral discriminator shows no separation at all (gap 0.001): the margin is learned, and it is learned from the gauge component. Sensitivity probes point the same way. We define z as the absolute mean logit change divided by the baseline logit standard deviation. Pure null-space perturbations that leave the waveform unchanged (relative change $\leq 2.4\times10^{-5}$) are invisible to D_w (z=0.00) yet move D_s by z=5.1 at a 1% spectral norm ratio and z=87 at 30%. Four waveform-calibrated content-artifact families (additive noise, phase jitter, spectral ripple, phase walk; relative waveform change 0.1%–10%) produce $z \leq 0.16$ for both D_w and D_s. A per-position mean $|\Delta\text{logit}|$ analysis gives the same ordering, excluding sign cancellation.

H.4 Trajectory during Stage 3 training

Figure 9 tracks the inconsistency of the refiner output across Stage 3 checkpoints, scored by the fixed final spectral discriminator. At the measured checkpoints, the null-space component decreases from 0.224 to 0.044, a fivefold reduction. The largest declines occur in the first checkpoint intervals after the discriminator's warmup ends. The discriminator's fake logit on the refiner output, 0.03 at the boundary (Table 9), rises to 0.46 by the end of training, against a real baseline of 0.485: the gauge-driven separability that marked the Stage 3 boundary disappears as training proceeds. Over the same window the spectral evaluation metrics (Mel Distance and STFT Distance) improve sharply. We report this as a temporal coincidence, not an attribution. The direct spectral reconstruction loss is active throughout Stage 3 and also shapes the null directions. Isolating the discriminator's causal contribution therefore requires a controlled training arm.

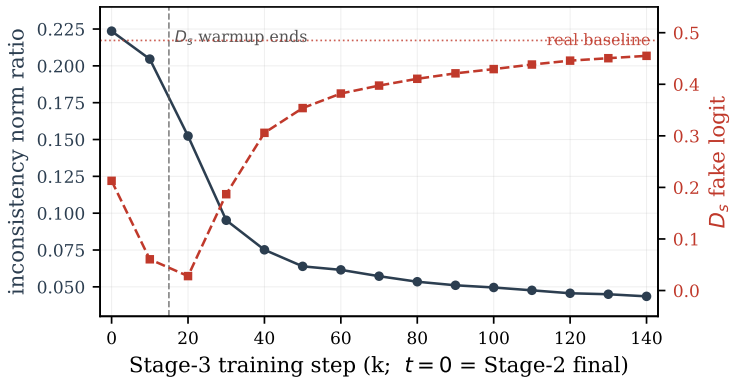

Figure 9: Stage 3 consistency trajectory (n=18 segments), scored by the fixed final spectral discriminator D_s. **Left axis**: inconsistency $\|S-PS\|/\|S\|$ of the refiner output (t=0 is the Stage 2 final generator). **Right axis**: D_s's fake logit; the dotted line marks its real baseline (0.485). The steepest descent follows the end of D_s's warmup. The coincident spectral-metric improvement is correlational, not an attribution (§H.4).

H.5 Summary and limitations

At the Stage 3 boundary, D_w retains a real/fake logit gap of 0.21 on the outputs that the identity-initialized refiner reproduces. The redundant spectral parameterization has an approximately half-dimensional subspace that waveform-routed objectives cannot observe, and the Stage 2 outputs contain energy in this subspace. In our probes, 95.3% of the spectral discriminator's output-gradient energy lies in these directions, nearly orthogonal to the waveform path. The measured null component then falls fivefold during training. These results suggest that D_s acts mainly as a learned regularizer of waveform-unidentifiable spectral inconsistency (gauge fixing), rather than as a source of complementary audible-content discrimination.

Several limitations apply. The probes use a small evaluation set with 18–48 segments per condition. Sensitivity is summarized only through mean logits and per-position mean $|\Delta\text{logit}|$. The evidence is diagnostic: it identifies the signal that D_s uses in these probes but does not establish the size of its causal contribution. That question requires the controlled training arms discussed in §H.4.

I Spec-SnakeBeta: Formulation and Initialization Analysis

This appendix separates exact algebraic properties from architectural design choices. Snake adds a periodic residual in the feature-value coordinate x; SnakeBeta separately parameterizes the residual period and range; and Spec-SnakeBeta indexes these parameters by spectral position while sharing them across feature channels. The first two steps are algebraic, whereas the frequency indexing and its initialization are architectural priors whose utility is evaluated empirically. Throughout, x denotes a scalar pre-activation. We optimize $\tilde{\alpha}$ and $\tilde{\beta}$ in log space, with $\alpha = e^{\tilde{\alpha}}$ and $\beta = e^{\tilde{\beta}}$, and use ϵ_{num}=10^{-9} in the denominator and ϵ_{init}=10^{-6} in the initialization.

A finite ReLU network on $\mathbb{R}$ defines a piecewise-linear (PWL) map with finitely many pieces. Any such map that is globally periodic must be constant; hence a finite ReLU network cannot exactly represent a nonconstant periodic function on the entire real line (Ziyin et al., 2020). On compact intervals, however, ReLU networks can approximate periodic functions, so this observation motivates an explicit periodic basis rather than establishing its necessity. Importantly, it concerns periodicity in the feature-value coordinate x, not harmonic spacing along the spectral-frequency coordinate f. We therefore use Snake's periodic residual as an audio inductive bias, and use the explicit STFT frequency axis for a separate purpose: allowing this bias to vary with spectral position.

Snake adds a periodic residual to an identity skip,

$$\text{snake}_\alpha(x) = x + \tfrac{1}{\alpha}\sin^2(\alpha x) = x + \tfrac{1-\cos 2\alpha x}{2\alpha}, \tag{11}$$

where $\alpha > 0$. Its residual has period π/α and range $[0, 1/\alpha]$, while the complete activation has derivative $1 + \sin(2\alpha x) \in [0, 2]$ and is therefore monotone. Moreover, for fixed x,

$$\tfrac{1}{\alpha}\sin^2(\alpha x) = \alpha x^2 + O(\alpha^3 x^4), \qquad \alpha \to 0, \tag{12}$$

so the activation converges pointwise to the identity. As the period π/α and the range $1/\alpha$ show, Snake couples the residual period and its maximum range through the same parameter α.

SnakeBeta introduces $\beta > 0$ so that the residual period π/α and maximum range $1/(\beta + \epsilon_{\text{num}})$ are separately parameterized. This separation removes Snake's unconditional monotonicity guarantee. Differentiating the implemented activation gives

$$\tfrac{d}{dx}\Big[x + \tfrac{1}{\beta+\epsilon_{\text{num}}}\sin^2(\alpha x)\Big] = 1 + \tfrac{\alpha}{\beta+\epsilon_{\text{num}}}\sin(2\alpha x) \;\in\; \Big[1 - \tfrac{\alpha}{\beta+\epsilon_{\text{num}}},\; 1 + \tfrac{\alpha}{\beta+\epsilon_{\text{num}}}\Big]. \tag{13}$$

The derivative is non-negative for all x if and only if $\alpha \le \beta + \epsilon_{\text{num}}$. This condition is tight because the lower bound is reached when $\sin(2\alpha x) = -1$. We permit non-monotonicity because hidden features need not preserve scalar order. The identity skip centers the derivative interval at one; it does not impose a parameter-uniform derivative bound.

For clarity, we suppress the layer index below and write $F = F_\ell$ and $\text{freq}_f = \text{freq}_{\ell,f}$ for a given layer. Conventional channel-wise SnakeBeta applies one curve (α_c, β_c) to all F spectral positions. Spec-SnakeBeta instead learns $\tilde{\alpha}, \tilde{\beta} \in \mathbb{R}^F$ (Eq. 1), yielding one curve per physical frequency and applying it to every feature channel. We use C for channel-indexed parameters, F for frequency-indexed parameters shared across channels, and CF for independent channel–frequency parameters. F-Log, F-Uniform, and CF-Log denote the corresponding initialized configurations. C contains $2C$ activation parameters per layer and is not the left/right audio channel. F contains $2F$ parameters per layer. CF learns a separate curve for every (c, f) pair and therefore contains $2CF$ parameters per layer. CF-Log begins from the same frequency-proportional values as F-Log, replicated across channels, but optimizes the copies independently. The default Spec-SnakeBeta is the channel-shared F-Log configuration. Its log-space parameterization guarantees positivity and gives

$$\frac{\partial \mathcal{L}}{\partial \tilde{\alpha}_f} = \alpha_f \frac{\partial \mathcal{L}}{\partial \alpha_f}, \qquad \frac{\partial \mathcal{L}}{\partial \tilde{\beta}_f} = \beta_f \frac{\partial \mathcal{L}}{\partial \beta_f}. \tag{14}$$

These identities describe the reparameterization and do not imply a convergence advantage.

We initialize α_f in proportion to normalized physical frequency using Eq. 2, so that

$$\alpha_f^{(0)} = \frac{\text{freq}_f}{\bar{f}} + \epsilon_{\text{init}}. \tag{15}$$

For equally spaced retained bins $\text{freq}_f = f\Delta$, $f = 0, \ldots, F-1$, the mean frequency is $\bar{f} = (F-1)\Delta/2$, yielding

$$\alpha_f^{(0)} = \frac{2f}{F-1} + \epsilon_{\text{init}}. \tag{16}$$

The effective initialization is thus linear in normalized frequency, although the positive parameter is stored in log-space.

At initialization, $\tilde{\beta}_f = 0$ and hence $\beta_f = 1$. To isolate the effect of the α_f schedule, consider an analysis slice in which the marginal pre-activation distribution is $x_{c,f} \sim \mathcal{N}(0, \sigma^2)$ with the same $\sigma > 0$ across frequency bins. Define the residual correction $\Delta_f(x) = \sin^2(\alpha_f x)/(1 + \epsilon_{\text{num}})$. Using $\mathbb{E}[\cos(tx)] = e^{-t^2\sigma^2/2}$ with $t = 2\alpha_f$ gives

$$\mathbb{E}[\Delta_f(x)] = \frac{1 - e^{-2\alpha_f^2\sigma^2}}{2(1 + \epsilon_{\text{num}})}, \tag{17}$$

which is strictly increasing in α_f. Under the stated common-distribution assumption, the expected correction therefore increases with spectral position. This is a conditional distributional statement, not a pointwise one: if feature means or variances differ across bins, the ordering need not hold.

At DC, $\alpha_0^{(0)} = \epsilon_{\text{init}}$ and, for fixed x,

$$\Delta_0(x) = \frac{\epsilon_{\text{init}}^2 x^2}{1 + \epsilon_{\text{num}}} + O(\epsilon_{\text{init}}^4 x^4), \tag{18}$$

so the activation is effectively identity-like at initialization for bounded pre-activations. In contrast, $\sup_x \Delta_f(x) = 1/(1+\epsilon_{\text{num}})$ is identical for every f. Thus, the uniform β initialization introduces no frequency-dependent maximum residual range: α_f controls the feature-space oscillation rate, while the frequency ordering of the expected correction in Eq. 17 depends on the stated input-distribution assumption.

In summary, the F parameterization assigns a scalar nonlinearity to each physical frequency and shares it across feature channels; F-Log applies frequency-proportional initialization in α-space while storing and optimizing the positive parameters in log space. Its initialization orders feature-space oscillation rates by normalized spectral position while remaining identity-like near DC. In the controlled ablation (§4.6), F-Log outperforms channel-wise C on five of six metrics and fully independent CF on all six metrics.

J Activation Ablation: Protocol

The activation ablation (§4.6) uses a shared mid-scale encoder–decoder with base width C_0=32, bottleneck width D=64, and approximately 42.6M generator parameters. Each model trains for 100k steps with the multi-resolution STFT loss, LSGAN loss, feature-matching (FM), and KL regularization. Training uses 48 kHz stereo segments of 1.28 s with a batch size of 4. The data are drawn from the LAION-DISCO-12M subset (LAION, 2024; Lanzendörfer et al., 2023) after the filtering described in Appendix B. We train each activation independently with seeds 42, 233, and 666. Table 4 reports the mean across the three runs without a confidence interval. All configurations use the same data, architecture, optimization, and evaluation set. F-Uniform and F-Log share one parameter pair across all feature channels at each frequency bin; they differ only in initialization (F-Uniform is zero-initialized; Uniform does not refer to frequency sampling). CF-Log instead learns an independent parameter pair per channel–frequency combination while retaining the F-Log initialization. Across the full ablation network, F-Log uses 2,730 activation parameters and CF-Log uses 350,720. Evaluation uses the same 100 Song Describer Dataset files for every seed and configuration. For CF-Log, per-channel coherence is 0.892 and CCPC is 0.885; Table 4 reports the latter.

K Banded-Refiner Efficiency and Cue-Aligned Evaluation

The Banded and Unconstrained Refiner variants are trained only at the refiner stage from the same frozen decoder under the same optimization protocol (not the production three-epoch refiner). They differ in output residual dimensions, so the comparison evaluates the efficiency of the complete structured design rather than isolating band placement at matched capacity. As reported in the main text (§4.5), the banded design uses fewer residual dimensions while improving magnitude distances and the duplex-aligned probes most directly targeted by its constraints. Relative to the Unconstrained Refiner, it reduces Low IPD, Low ILD, and High ILD errors by 2.4%, 8.1%, and 14.7%, respectively, while High IPD remains within 0.7%. We complement these measurements with a controlled paired listening test. Ten professional mixing and mastering engineers scored both versions of each excerpt on a 1–10 scale. Scores are mapped to $[0, 1]$ as $(s-1)/9$, and the default placement receives the higher mean paired rating (0.75 vs. 0.66).

K.1 Band-Restricted Duplex-Cue Metrics

The duplex-cue columns of Table 3 measure how faithfully a reconstruction preserves the two interaural cues that the band allocation targets. Both metrics are computed from the complex STFTs of the reference and the reconstruction, truncated to a common signal length and frame count, and both are restricted to a bin range $[f_0, f_1)$: the Low band runs from DC to the bin of the 1.5 kHz refiner edge, and the High band runs from the bin of the 4 kHz edge through the Nyquist bin. Following Eq. 9, $L(f,t)$ and $R(f,t)$ denote the left and right complex STFT coefficients over T frames, with hats marking the reconstruction. Neither metric applies a silence mask.

Interaural phase difference (IPD) error. The interaural phase difference at each time–frequency bin is $\Delta\phi(f,t) = \phi_L(f,t) - \phi_R(f,t)$, where $\phi_L = \arg L$ and $\phi_R = \arg R$. Phase differences are

circular, so the deviation from the reference is reduced modulo 2π by taking the principal argument of the corresponding unit complex number, which maps it into $(-\pi, \pi]$:

$$\mathrm{IPD}_{[f_0,f_1)} = \frac{1}{(f_1 - f_0)\,T} \sum_{f=f_0}^{f_1-1} \sum_{t} \Big| \arg\exp\big(j\,[\,\widehat{\Delta\phi}(f,t) - \Delta\phi(f,t)\,]\big) \Big|. \tag{19}$$

Values are in radians and are therefore bounded above by π. Every time–frequency bin in the band contributes equally, independent of its energy.

Interaural level difference (ILD) error. The ILD is a per-frame quantity: band energies are aggregated across the bins of the band before the level ratio is taken in decibels,

$$\mathrm{ILD}_{[f_0,f_1)}(t) = 10\log_{10} \frac{\sum_{f=f_0}^{f_1-1} |L(f,t)|^2 + \epsilon_{\text{ild}}}{\sum_{f=f_0}^{f_1-1} |R(f,t)|^2 + \epsilon_{\text{ild}}}, \qquad \epsilon_{\text{ild}} = 10^{-8}, \tag{20}$$

and the reported error is the mean absolute deviation from the reference across frames,

$$\text{ILD-Err}_{[f_0,f_1)} = \frac{1}{T} \sum_{t} \big|\, \widehat{\mathrm{ILD}}_{[f_0,f_1)}(t) - \mathrm{ILD}_{[f_0,f_1)}(t) \,\big|. \tag{21}$$

Values are in decibels. The additive ϵ_{ild} stabilizes the ratio in near-silent frames and is distinct from ϵ_{pan} in Eq. 9. Unlike Eq. 19, the level cue is energy-aggregated within each band and frame, but frames are still averaged with equal weight. For both metrics, per-track values are averaged over the evaluation tracks.

K.2 Subjective Test Protocol

Participants and equipment. Ten professional mixing and mastering engineers (all with $\geq$5 years of commercial credits) participated. Each listener used calibrated studio headphones in a treated listening environment with ambient noise $\leq$30 dB SPL. Playback was routed through a dedicated digital-to-analog (D/A) converter at 48 kHz/24-bit with the monitoring level calibrated to 70 dB SPL (K-weighted).

Stimuli and randomization. From the Song Describer Dataset evaluation set we sampled 20 excerpts (10 s each), stratified across genre (pop, rock, electronic, jazz, classical) and dynamic characteristics. Each engineer evaluated all excerpts for both the duplex-placement and Unconstrained Refiner conditions as matched A/B pairs. For each excerpt, the assignment of the two conditions to A and B and the order of presentation were randomized independently for each engineer. Condition labels remained hidden throughout the blind evaluation. A 2-second silent gap separated the two members of each pair.

Rating scale. Engineers scored both members of each pair on a 1–10 integer scale according to the rubric in Table 11. The scale assesses four perceptual dimensions: spectral balance (frequency), dynamics, spatial imaging, and harmonic/timbral quality. Scores are mapped to $[0, 1]$ as $(s-1)/9$ for reporting.

Score aggregation. Each excerpt was rated by all ten engineers under both conditions. The reported score per condition is the arithmetic mean of the normalized paired ratings over all engineers and excerpts.

L A Latent Temporal-Frequency Probe (Exploratory)

As an exploratory investigation of what an audio autoencoder latent encodes, we probe how each model distributes information along the *temporal* axis of its latent sequence. Audio content varies at different rates. Percussion, onsets, and rhythmic "groove" change rapidly, whereas sustained harmony and timbre change slowly. A latent may therefore encode fast and slow temporal changes in different components. This section constitutes a *diagnostic exploration*, not a core claim.

Table 11: Subjective quality rubric. Engineers assign a single holistic score on a 1–10 scale; anchor descriptions guide consistent interpretation.

Score	Anchor Description
8–10	Frequency balance suits the genre; mid/side energy ratio appropriate; no masking between instruments; dynamics well controlled with natural transients; spatial staging coherent with adequate depth; harmonic coloration (if any) enhances rather than muddies.
5–7	Minor spectral gaps or slight imbalance in one band; compression artifacts occasionally audible (pumping, over-limited transients); spatial image somewhat flat but without obvious panning errors; timbral quality acceptable with minor coloration.
2–4	Noticeable frequency-band dropout or harshness (e.g., sibilance, muddy low-mids); dynamics feel unnatural (over-compressed or under-controlled); spatial image collapsed or inconsistent; timbral distortion clearly audible.
1	Severe artifacts across multiple dimensions; reconstruction quality unacceptable for professional evaluation.

Probe. For each model, we encode a clip to its latent $\mathbf{z} \in \mathbb{R}^{D\times T}$, with model-specific latent dimension D, and take the real FFT along time. We then split the temporal-frequency bins at the midpoint. Zeroing the upper half gives a slowly varying latent $\mathbf{z}_{\text{low}}$; zeroing the lower half gives a rapidly varying latent $\mathbf{z}_{\text{high}}$. An inverse FFT returns each partial latent to the time domain. We decode $\mathbf{z}_{\text{low}}$ and $\mathbf{z}_{\text{high}}$ *separately* and measure the spectral centroid of each decoded signal. We summarize with an *inversion ratio* $\rho = \text{centroid}(\mathbf{z}_{\text{low}} \to \text{audio})/\text{centroid}(\mathbf{z}_{\text{high}} \to \text{audio})$; $\rho > 1$ ("inverted") means the rapidly varying latent decodes to lower-centroid audio and the slowly varying latent decodes to higher-centroid audio.

Table 12: Latent temporal-frequency probe (spectral centroid of the separately decoded low-/high-temporal-frequency latent halves; averaged over our internal evaluation set). $\rho > 1$ ("inverted") means the rapidly varying latent decodes to audio with a lower spectral centroid. εar-VAE2 and εar-VAE are inverted; the other three models are not.

Model	$\text{centroid}_{\text{low}}$ (Hz)	$\text{centroid}_{\text{high}}$ (Hz)	ρ	inverted
εar-VAE2 (complex STFT)	323	137	**2.36**	✓
εar-VAE (waveform)	477	249	1.82	✓
LeVo 2	392	736	0.53	×
SAME-L	391	487	0.79	×
SA-Open	363	393	0.91	×

Observation and interpretation. εar-VAE2 (ρ=2.36) and εar-VAE (ρ=1.82) are inverted. LeVo 2, SAME-L, and Stable Audio Open are not ($\rho < 1$), and εar-VAE2 has the largest inversion ratio. Both εar-VAE2 and εar-VAE use the perceptual training losses introduced by εar-VAE, but this comparison does not isolate their effect. Figure 10 visualizes the measured separation.

This result is correlational and does not establish a causal mechanism or a quality advantage. Spectral centroid measures brightness, not rhythm or onset structure. The models differ in architecture, latent configuration, sample rate, objective, and compression, and partial latents are outside the decoder's training distribution. A causal study requires matched architectures, loss ablations, multiple checkpoints, and a larger track set.

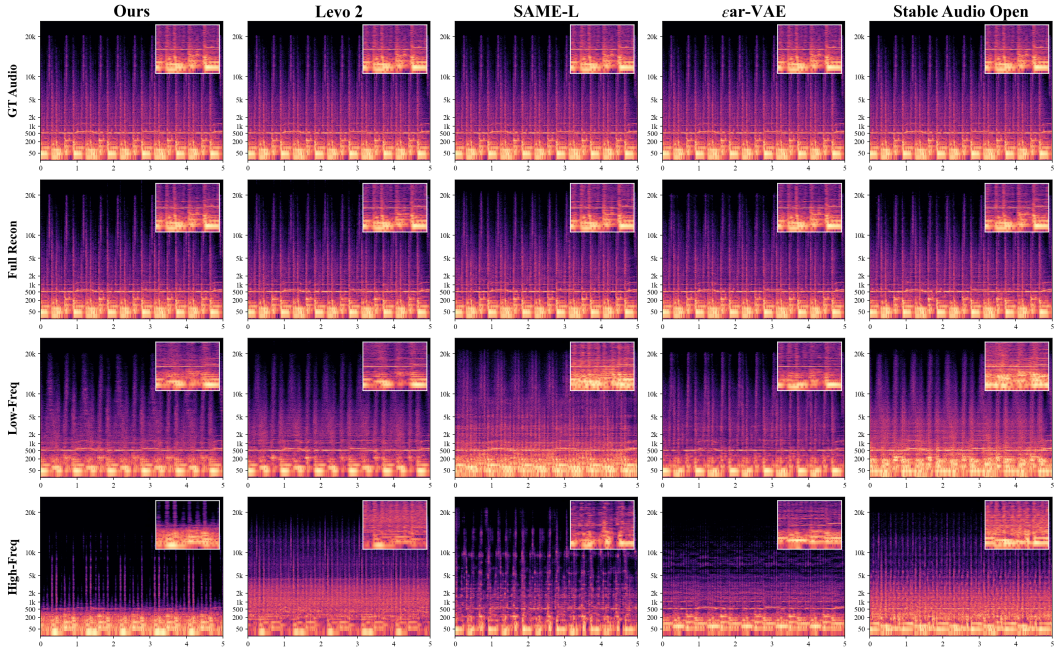

Figure 10: Latent temporal-frequency probe for one track (Track 1004034; temporal split at 0.5). **Rows** (top to bottom): ground-truth audio; full reconstruction; low-temporal-frequency latent half decoded; high-temporal-frequency latent half decoded. **Columns** (left to right): εar-VAE2, LeVo 2, SAME-L, εar-VAE, and Stable Audio Open. For εar-VAE2, the rapidly varying (high) latent half decodes to audio with a lower spectral centroid, matching Table 12.