REVOC: HIGH-FIDELITY AND EFFICIENT AUDIO GENERATION VIA RESTORATION BRIDGES AND GANS

Anonymous authors
Paper under double-blind review

ABSTRACT

Although reconstructing high-fidelity audio from compressed representations is an underdetermined inverse problem, the corresponding inverse transforms or decoders can recover useful acoustic structure at low computational cost. In this paper, we propose **ReVoC**, a few-step generative framework that combines restoration bridges with adversarial training to refine these approximate reconstructions. Starting from the outputs of Mel inversion or neural codec decoding, ReVoC learns to restore clean audio in the complex STFT domain using an efficient Vocos-style 1D backbone. To improve perceptual quality across inference budgets, we introduce multi-NFE adversarial fine-tuning, where a step-conditioned discriminator guides the joint optimization of complete one-, two-, and four-step inference trajectories. A single generator per task thus supports all three budgets with quality comparable to models fine-tuned separately for each step count. On Mel-conditioned vocoding, ReVoC-Small uses the same eight-block backbone as Vocos and improves one-step PESQ from 3.592 to 4.142. With four-step inference, ReVoC-Base achieves a PESQ of 4.433 and approximately four times the GPU throughput of four-step BridgeVoC. ReVoC also consistently improves reconstruction quality in low-bitrate neural codec restoration.

1 INTRODUCTION

Reconstructing high-fidelity audio from compressed representations is central to speech synthesis and neural audio coding (Kong et al., 2020; Défossez et al., 2023). Mel spectrograms and quantized neural representations discard information, requiring a generator to recover missing acoustic structure and perceptual details. GAN vocoders enable efficient parallel synthesis, while diffusion and flow models offer iterative generation (Lee et al., 2023; Kong et al., 2021; Liu et al., 2025). The practical challenge is to improve reconstruction quality while keeping both the number of function evaluations (NFE) and the cost of each evaluation small.

Useful structure can already be recovered before generative refinement. Approximate Mel inversion provides an initial spectral estimate, and a neural codec decoder produces a coarse waveform. Prior work exploits such reconstructions through spectral prediction, restoration bridges, and generative postfiltering (Lv et al., 2024; Li et al., 2025b;a; Wu et al., 2024; Welker et al., 2025). These results motivate using each representation's own reconstruction as an informative starting point for generation. Combined with efficient spectral processing, as demonstrated by Vocos (Siuzdak, 2024), this offers a promising basis for high-quality restoration under a small inference budget.

An informative starting point alone, however, does not ensure high perceptual quality after only a few inference steps. Adversarial supervision can improve the generated audio, as demonstrated by bridge-based vocoders and adversarially refined flow models (Li et al., 2025a; Lee et al., 2024; Yao et al., 2026). Yet fine-tuning for a fixed step count specializes a generator to one inference budget; Flow2GAN, for example, trains separate models for different step counts. We seek to retain the benefits of adversarial refinement while allowing the inference budget to change without switching generators.

We propose **ReVoC**[1], a few-step audio generation framework combining restoration bridges with joint multi-NFE adversarial fine-tuning. ReVoC restores Mel-inverted spectra or neural-codec-

[1] Audio samples are available at `https://revoc-demo.github.io/`

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.

decoded audio in the complex STFT domain using a Vocos-style 1D ConvNeXt backbone. A discriminator conditioned on NFE guides the joint optimization of complete one-, two-, and four-step trajectories, allowing one generator per task to approach the quality of separately fine-tuned models. Experiments demonstrate a favorable quality–efficiency trade-off: ReVoC-Small retains Vocos's eight-block backbone and improves one-step PESQ from 3.592 to 4.142, while ReVoC-Base reaches 4.433 with four steps and approximately four times the GPU throughput of four-step BridgeVoC under our benchmarking setup. ReVoC also improves both PESQ and UTMOS of EnCodec reconstructions at 3 and 6 kbps, using one restoration model for both bitrates.

2 Related Work

2.1 Adversarial and Spectral-Domain Vocoders

GAN vocoders synthesize waveforms in parallel with adversarial objectives that capture complementary aspects of audio structure. MelGAN uses multi-scale waveform discriminators (Kumar et al., 2019), HiFi-GAN introduces multi-period discrimination (Kong et al., 2020), and UnivNet evaluates spectrograms at multiple resolutions (Jang et al., 2021). BigVGAN improves the generator through periodic activations and anti-aliased representations (Lee et al., 2023). Spectral-domain vocoders reduce the need for learned waveform-rate upsampling: iSTFTNet replaces output-side generator layers with inverse STFT synthesis (Kaneko et al., 2022), while Vocos predicts Fourier coefficients with a frame-level ConvNeXt backbone (Siuzdak, 2024). APNet and APNet2 explicitly predict amplitude and phase spectra (Ai & Ling, 2023; Du et al., 2024). ReVoC builds on efficient spectral generation, using a Vocos-style backbone as the endpoint predictor of a restoration bridge.

2.2 Diffusion, Flow Matching, and Few-Step Generation

DiffWave generates audio by iteratively denoising a Gaussian source (Kong et al., 2021). Subsequent methods improve the source distribution or the generative model: PriorGrad derives an adaptive Gaussian prior from conditioning information (Lee et al., 2022), RFWave applies rectified flow to complex spectrogram subbands (Liu et al., 2025), and PeriodWave incorporates waveform periodicity into flow estimation (Lee et al., 2025). WaveFM combines a Mel-conditioned prior with consistency distillation for single-step synthesis (Luo et al., 2025). These methods demonstrate that both the generation path and its network parameterization influence inference efficiency.

Adversarial refinement directly targets the outputs of shortened inference procedures. PeriodWave-Turbo fine-tunes fixed-step flow generators using reconstruction and adversarial losses (Lee et al., 2024). Flow2GAN combines endpoint-prediction pretraining with adversarial fine-tuning of complete one-, two-, or four-step trajectories, training a separate generator for each step count (Yao et al., 2026). ReVoC jointly optimizes these inference budgets with shared generator parameters and an NFE-conditioned discriminator.

2.3 Reconstruction-Informed Vocoding and Restoration Bridges

Approximate Mel inversion recovers part of the spectral structure needed for vocoding. FreeV uses a pseudoinverse Mel filter to initialize amplitude prediction (Lv et al., 2024). RNDVoC separates a recoverable range-space component from a learned null-space component (Li et al., 2025b), while DegVoC refines a linearly recovered spectral estimate with a learned prior solver (Li et al., 2026a). These methods exploit an explicit reconstruction prior within a direct synthesis framework; such priors do not necessarily define a stochastic generation trajectory.

Restoration bridges connect degraded and clean distributions, as in I^2SB for images (Liu et al., 2023) and A^2SB for audio bandwidth extension and inpainting (Kong et al., 2025). BridgeVoC applies this approach to Mel vocoding by connecting a Mel-derived spectral reconstruction to clean spectra (Lei et al., 2025). Its expanded version combines bridge training with adversarial supervision and adds a separately distilled single-step student trained with distillation, reconstruction, consistency, and adversarial losses (Li et al., 2025a). ReVoC adopts this restoration perspective for both Mel inversion and fixed codec decoder outputs, emphasizing efficient spectral modeling and joint optimization across inference budgets.

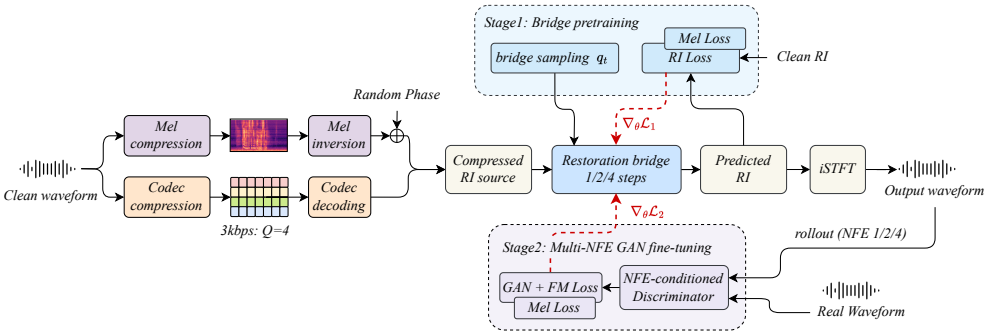

Figure 1: Overview of ReVoC. Mel inversion with random phase or codec decoding provides the source for a spectral restoration bridge. The bridge is pretrained with RI and Mel losses, then jointly fine-tuned for one-, two-, and four-step inference using an NFE-conditioned discriminator.

2.4 Generative Neural Codec Restoration

Generative postfilters refine decoded audio to compensate for information lost during neural compression. ScoreDec combines non-adversarial codec training with a score-based postfilter in the complex spectral domain (Wu et al., 2024). FlowDec uses conditional flow matching for postfiltering, extending this approach to low-bitrate full-band general audio with fewer network evaluations (Welker et al., 2025). Codec resynthesis can also operate before waveform decoding: Liu et al. (2024) compare token prediction and regression, and introduce a Schrödinger bridge from coarse codec embeddings to pre-quantized representations. ReVoC operates on the decoded waveform of an existing, fixed codec. This setting lets it apply the same restoration-bridge formulation used for Mel vocoding to codec artifacts, with a single codec-restoration model covering both evaluated bitrates and all three inference budgets.

3 Preliminaries

Spectral representation. Let $\mathcal{S}$ denote the STFT. We represent complex spectra in power-compressed real–imaginary (RI) coordinates,

$$\mathcal{C}(X) = \mathrm{RI}\big(\kappa |X|^p e^{j\angle X}\big), \qquad \mathcal{V}(z) = \mathcal{S}^{-1}\big(\mathcal{C}^{-1}(z)\big), \tag{1}$$

where RI stacks the real and imaginary components and $\mathcal{V}$ synthesizes a waveform. Thus, a spectral state has shape $2 \times F \times T$, with F frequency bins and T frames.

Paired Gaussian bridges. We use the Gaussian bridge formulation employed in generative restoration (Liu et al., 2023; Li et al., 2025a). Given a clean endpoint z_0 and a reconstruction endpoint z_1, define $B_t = \int_0^t \beta(u)\,du$ and $\bar{B}_t = B_1 - B_t$ for a positive diffusion schedule β. The intermediate-state distribution is

$$q(z_t \mid z_0, z_1) = \mathcal{N}\left(\frac{\bar{B}_t z_0 + B_t z_1}{B_1}, \gamma^2 \frac{B_t \bar{B}_t}{B_1} I\right), \tag{2}$$

where γ controls the noise scale. This closed-form marginal allows intermediate states to be sampled without simulating a forward trajectory. We place clean audio at $t = 0$ and its approximate reconstruction at $t = 1$, so restoration proceeds backward in time.

4 Method

ReVoC first learns a restoration bridge from a representation-specific reconstruction to clean audio, then adversarially fine-tunes its few-step outputs across multiple inference budgets. We apply this formulation to Mel vocoding and neural-codec artifact restoration, with a separate generator for each task.

4.1 Unified Restoration from Approximate Reconstructions

Given a clean waveform x and its compressed representation y, let $\mathcal{R}_k(y, \xi)$ denote an initial complex spectrum for task $k \in \{\text{mel}, \text{codec}\}$, where ξ accounts for initialization randomness when needed. The bridge endpoints are

$$z_0 = \mathcal{C}(\mathcal{S}(x)), \qquad z_1 = \mathcal{C}(\mathcal{R}_k(y, \xi)). \tag{3}$$

The resulting reconstruction distribution serves as the source of the restoration bridge.

Mel vocoding. For a log-Mel spectrogram y, we estimate a nonnegative magnitude spectrum $\hat{a}$ from $v = \exp(y)$ using Mel pseudoinversion followed by two positivity-preserving residual corrections. Since Mel features do not retain phase, we initialize phase using the normalized STFT of a Gaussian waveform ξ. Writing this phase factor as U_ξ, the source spectrum is $\mathcal{R}_{\text{mel}}(y, \xi) = \hat{a} \odot U_\xi$. This retains the recovered spectral magnitude while leaving phase and missing details to the restoration model. The inversion procedure is given in Appendix A.2.

Codec restoration. For codec tokens y_b at bitrate b, we use $\mathcal{R}_{\text{codec}}(y_b) = \mathcal{S}(D_{\text{codec}}(y_b))$, where D_{codec} is the fixed original decoder. One codec-restoration model is trained on decoded audio from both 3 and 6 kbps. In both tasks, conditioning is derived from the reconstruction:

$$c_{\text{mel}} = \text{mag}(z_1), \qquad c_{\text{codec}} = z_1, \tag{4}$$

where $\text{mag}(z) = \sqrt{z_{\text{R}}^2 + z_{\text{I}}^2}$. Mel conditioning therefore uses the recovered magnitude, while codec conditioning retains both RI components.

4.2 Learning the Restoration Bridge

We sample $t \sim \mathcal{U}(0, 1)$ and z_t from Eq. equation 2, and train a time-conditioned predictor to estimate the clean endpoint:

$$\hat{z}_0 = f_\theta(z_t, t; c_k), \qquad \hat{x} = \mathcal{V}(\hat{z}_0). \tag{5}$$

Local spectral encoding. Our predictor uses a Vocos-style ConvNeXt1D backbone (Siuzdak, 2024). Its input concatenates the current RI state and the reconstruction condition, giving three channels for Mel vocoding and four for codec restoration. These channels describe corresponding quantities at the same time–frequency locations. A lightweight patch encoder jointly processes them with a local 2D convolution before frequency flattening and temporal projection. This introduces explicit frequency locality and local interactions between the current state and the reconstruction condition before the 1D backbone. Time-conditioned ConvNeXt blocks preserve the frame rate, and a linear head predicts the clean RI spectrum at a single STFT resolution. Our ablation compares this frontend with direct frequency flattening under an otherwise matched backbone and shows faster convergence. Architectural configurations are listed in Appendix A.3.

Bridge pretraining. We supervise the predicted endpoint in RI coordinates and its synthesized waveform with a multi-resolution log-Mel loss:

$$\mathcal{L}_{\text{pre}} = \mathbb{E}_{x,y,\xi,t,z_t}[\text{MSE}(\hat{z}_0, z_0) + \lambda_{\text{mel}} \mathcal{L}_{\text{mel}}^{\text{pre}}(\hat{x}, x) + \mathcal{L}_{\text{aux}}]\,. \tag{6}$$

The MSE is averaged over all RI elements. The low-weight auxiliary term supervises the spectrum realized after waveform synthesis and encourages agreement with the predicted spectrum; its definition and schedule are given in Appendix A.4.

Few-step restoration. Starting from $z^{(0)} = z_1$, we use a decreasing grid $1 = t_0 > \cdots > t_N = 0$. With $r_i = B_{t_{i+1}}/B_{t_i}$, each reverse update is

$$\begin{aligned} z^{(i+1)} = r_i z^{(i)} + (1 - r_i) f_\theta(z^{(i)}, t_i; c_k) \\ + \gamma \sqrt{B_{t_{i+1}}(1 - r_i)}\, \eta_i, \qquad \eta_i \sim \mathcal{N}(0, I). \end{aligned} \tag{7}$$

The final noise term vanishes because $B_{t_N} = B_0 = 0$. We synthesize the waveform once from $z^{(N)}$, requiring exactly N predictor evaluations. In particular, one-step inference reduces to a single clean-endpoint prediction from z_1.

4.3 JOINT MULTI-NFE ADVERSARIAL FINE-TUNING

To improve perceptual quality at the inference budgets used in deployment, we fine-tune the pretrained predictor through complete restoration trajectories. Each minibatch samples N uniformly from $\mathcal{B} = \{1, 2, 4\}$ and produces $\hat{x}_N = G_\theta^{(N)}(y; \omega) = \mathcal{V}(z^{(N)})$, where ω collects the initialization and sampling noise. All budgets share the same predictor parameters, and generator updates backpropagate through every evaluation in the selected trajectory.

We use multi-period and multi-resolution discriminators (Kong et al., 2020; Jang et al., 2021), conditioning each branch on N through a learned projection embedding. Real and generated examples receive the same budget label. Let $\mathcal{L}_{\text{adv}}$ and $\mathcal{L}_{\text{FM}}$ aggregate the adversarial and feature-matching losses across discriminator families. The generator objective is

$$\begin{aligned}\mathcal{L}_G = \mathbb{E}_{x,y,N,\omega}\big[&\mathcal{L}_{\text{adv}}(\hat{x}_N, N) + \lambda_{\text{FM}}\mathcal{L}_{\text{FM}}(\hat{x}_N, x) \\ &+ \lambda_{\text{mel}}^G \mathcal{L}_{\text{mel}}^G(\hat{x}_N, x)\big].\end{aligned} \tag{8}$$

The discriminators use the corresponding hinge objective. This stage replaces the pretraining objective with losses on the final generated audio, optimizing early predictor evaluations through their effect on that output. The generator remains conditioned on bridge time; N selects the rollout length without an additional generator embedding. At deployment, one generator per task supports all three budgets, and the discriminators are discarded.

5 EXPERIMENTS

We evaluate ReVoC on Mel vocoding and low-bitrate codec restoration, examining quality, measured inference efficiency, and support for multiple budgets with one generator. Ablations assess spectral encoding, initialization, and adversarial fine-tuning; out-of-domain experiments test whether the refinement gains extend beyond the training corpus.

5.1 EXPERIMENTAL SETUP

Datasets and evaluation metrics. For Mel vocoding, we train on LibriTTS (Zen et al., 2019) at 24 kHz and compare systems on a fixed 208-utterance subset of `dev-clean` and `dev-other`. Separately trained LJSpeech (Ito & Johnson, 2017) models use 22.05-kHz audio and 80-band Mel conditioning; results on 500 test utterances appear in Appendix B.2. We assess out-of-domain generalization without adaptation on 200 utterances each from EARS (Richter et al., 2024) and VCTK (Yamagishi et al., 2019), resampled to 24 kHz. We report PESQ (Rix et al., 2001; ITU-T, 2005), multi-resolution STFT distance (M-STFT) (Steinmetz & Reiss, 2020), voiced/unvoiced F1 (V/UV F1) and periodicity RMSE using CARGAN (Morrison et al., 2022), UTMOS (Saeki et al., 2022), and ViSQOL (Chinen et al., 2020). These metrics cover spectral reconstruction, voicing and periodicity, and predicted perceptual quality. Periodicity RMSE is reported for the LibriTTS Mel and codec experiments. UTMOS is an automatic quality predictor; human preferences are evaluated separately in Appendix B.3. Data preprocessing and metric implementations are detailed in Appendices A.1 and A.5.

Implementation details. ReVoC-Small and ReVoC-Base use eight width-512 and twelve width-768 ConvNeXt blocks, respectively, with the same patch-encoder and compressed-RI bridge design. The LibriTTS Mel frontend uses 100 bands. Stage 2 jointly fine-tunes complete one-, two-, and four-step trajectories, with one final EMA checkpoint per variant. For LibriTTS, we evaluate released checkpoints of Vocos, BigVGAN-base, BigVGAN, RFWave, BridgeVoC, and Flow2GAN (Siuzdak, 2024; Lee et al., 2023; Liu et al., 2025; Li et al., 2025a; Yao et al., 2026), retaining their native Mel preprocessing. BridgeVoC uses its dedicated single-step and multi-step checkpoints at NFE 1 and 4, respectively; Flow2GAN is evaluated at NFE 1. Codec restoration uses a fixed 24-kHz EnCodec (Défossez et al., 2023) at 3 and 6 kbps. Model and training configurations are summarized in Tables 7 and 8, with the architecture and optimization procedures detailed in Appendices A.3 and A.4.

Table 1: Objective comparison with baseline vocoders on LibriTTS. Baseline scores are recomputed from released checkpoints. Per. denotes periodicity RMSE.

Model	NFE	PESQ↑	M-STFT↓	V/UV F1↑	Per.↓	UTMOS↑	ViSQOL↑
Ground truth	–	–	–	–	–	3.879	–
BigVGAN-base	1	3.838	0.819	0.959	0.108	3.624	4.916
BigVGAN	1	4.266	0.732	0.972	0.077	3.750	4.966
Vocos	1	3.592	0.870	0.952	0.118	3.550	4.884
RFWave	10	4.249	0.961	0.967	0.097	3.383	4.776
BridgeVoC	1	4.319	0.739	0.977	0.066	3.619	4.953
BridgeVoC	4	4.410	0.712	0.978	0.065	3.725	4.960
Flow2GAN	1	4.207	0.792	0.977	0.068	3.627	4.950
ReVoC-Small	1	4.142	0.843	0.971	0.078	3.460	4.906
	2	4.301	0.808	0.975	0.068	3.646	4.921
	4	4.342	0.794	0.977	0.065	3.720	4.929
ReVoC-Base	1	4.357	0.769	0.977	0.064	3.718	4.944
	2	4.409	0.747	0.980	0.060	3.799	4.950
	4	4.433	0.737	0.980	0.057	3.832	4.954

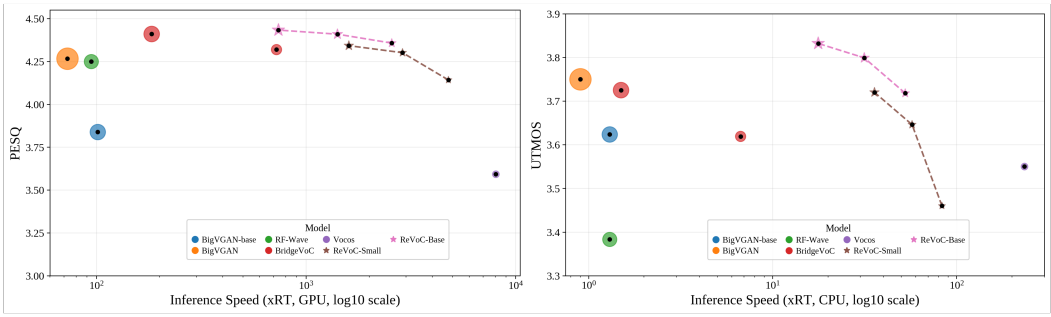

Figure 2: Quality versus inference throughput on LibriTTS: PESQ versus GPU ×RT (left) and UTMOS versus CPU ×RT (right). Measurements use FP32, batch size 16, and nominal one-second inputs. Full results are in Table 9.

5.2 Mel Vocoding: Quality and Efficiency

Few-step refinement. Table 1 shows that ReVoC-Small improves one-step PESQ over Vocos from 3.592 to 4.142 while retaining an eight-block, width-512 backbone. Increasing NFE to four raises its PESQ and UTMOS to 4.342 and 3.720 using the same checkpoint. ReVoC-Base exceeds the one-step BridgeVoC and Flow2GAN configurations in PESQ and UTMOS at NFE 1. At NFE 4, it achieves the highest PESQ (4.433) and UTMOS (3.832), and the lowest periodicity RMSE (0.057), among the reported generators. For both sizes, additional steps also reduce M-STFT and improve or maintain V/UV F1. Most of the PESQ gain occurs between one and two steps, while four steps provide further refinement. This gives each trained model several useful operating points without changing its parameters.

Quality–throughput trade-off. We report measured throughput because equal NFE can entail different computational costs across architectures. Timing includes source initialization and waveform synthesis, following Appendix A.6. At two steps, ReVoC-Base nearly matches four-step BridgeVoC in PESQ (4.409 versus 4.410) with 7.70× its GPU throughput. At four steps, ReVoC-Base improves PESQ and UTMOS over four-step BridgeVoC by 0.023 and 0.107 with 4.03× its GPU throughput. CPU throughput increases from 1.5×RT to 17.7×RT (Figure 2 and Table 9). BridgeVoC retains lower M-STFT and slightly higher ViSQOL at four steps.

Vocos remains the fastest reported baseline. One-step ReVoC-Small reaches 4776.2× real-time GPU throughput with PESQ 4.142, compared with 8036.8× and PESQ 3.592 for Vocos. Together, the two

Table 2: Fixed-NFE versus joint multi-NFE fine-tuning of ReVoC-Base, with matched updates per budget. Fixed-NFE training uses three generators; joint training uses one.

Training	NFE	PESQ↑	M-STFT↓	V/UV F1↑	Per.↓	UTMOS↑	ViSQOL↑
Fixed	1	4.357	0.768	0.978	0.063	3.709	4.946
Joint	1	4.357	0.769	0.977	0.064	3.718	4.944
Fixed	2	4.411	0.742	0.980	0.059	3.807	4.953
Joint	2	4.409	0.747	0.980	0.060	3.799	4.950
Fixed	4	4.442	0.730	0.981	0.056	3.832	4.959
Joint	4	4.433	0.737	0.980	0.057	3.832	4.954

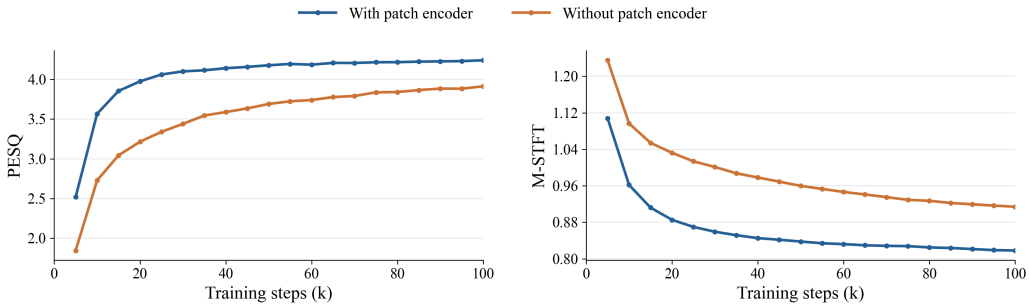

Figure 3: Effect of the patch encoder during the first 100k Stage 1 updates of ReVoC-Small. Left: PESQ; right: M-STFT. Both models use four-step EMA inference on the same 100 validation utterances.

panels of Figure 2 show how model size and inference budget expose different quality–throughput trade-offs on CPU and GPU.

Supplementary evaluations. Separately trained LJSpeech models exhibit the same improvement with additional steps across all five reported metrics (Appendix B.2, Table 11). In the pairwise listening study, four-step ReVoC-Base receives 71.25%, 56.25%, and 53.75% of votes against Vocos, BigVGAN-base, and four-step BridgeVoC, respectively. The protocol and complete win/tie/loss proportions are given in Appendix B.3 and Figure 6.

5.3 Joint Training across Inference Budgets

We compare joint fine-tuning with three ReVoC-Base models specialized for $N \in \{1, 2, 4\}$ from the same Stage 1 checkpoint. Each fixed-NFE run uses 50k minibatches, while the joint run uses 150k with balanced exposure. Both provide 24k generator and 26k discriminator updates per budget, including warm-up (Appendix A.4). Table 2 shows that joint training closely matches the specialized models, with maximum PESQ and UTMOS differences of 0.009. One 52.947M-parameter generator supports all budgets, compared with 158.841M parameters for three separately deployed generators, allowing NFE to change without switching checkpoints.

5.4 Ablation Studies

Patch encoder. We compare the ReVoC-Small patch encoder with direct frequency flattening under the same backbone and training settings. Parameter counts are nearly identical: 19.309M with patches and 19.316M without. At 100k Stage 1 updates, patch encoding improves validation PESQ from 3.912 to 4.239 and reduces M-STFT from 0.914 to 0.818 (Figure 3), indicating faster early convergence. This supports processing local spectral structure before the temporal backbone under a comparable parameter budget. Appendix A.3 gives the frontend configurations. The comparison precedes the scheduled consistency losses described in Appendix A.4.

Table 3: Effect of Mel magnitude initialization with the same ReVoC-Base Stage 1 checkpoint at NFE 4. Phase initialization is fixed, and no retraining is performed.

Initialization	PESQ↑	UTMOS↑	M-STFT↓
Pseudoinverse + absolute value	4.142	3.729	0.742
Pseudoinverse + ReLU, 0 corrections	4.365	3.773	0.731
Pseudoinverse + ReLU, 1 correction	4.399	3.776	0.729
Pseudoinverse + ReLU, 2 corrections	4.403	3.783	0.729

Mel initialization. We vary the magnitude initializer while keeping the ReVoC-Base Stage 1 checkpoint, four-step sampler, and waveform-derived phase fixed (Table 3). Replacing the absolute value of the pseudoinverse with ReLU raises PESQ from 4.142 to 4.365. Positive residual corrections further improve PESQ and UTMOS, reaching 4.403 and 3.783 after two corrections. One correction achieves most of the PESQ gain, reaching 4.399. These corrections are applied once during source construction and leave neural NFE unchanged. The inversion procedure and its error/runtime measurements appear in Appendices A.2 and B.1 (Table 10).

Adversarial fine-tuning. Table 4 compares ReVoC-Base before and after joint Stage 2 fine-tuning from the same Stage 1 checkpoint. Both checkpoints are evaluated directly at NFE 1, 2, and 4. Joint fine-tuning improves PESQ and UTMOS at every budget, with the largest gains at one step: PESQ rises from 4.115 to 4.357, and UTMOS from 3.624 to 3.718. At four steps, the scores increase from 4.403 to 4.433 and from 3.783 to 3.832, respectively. The larger improvement at one step is consistent with directly supervising the final output of each short trajectory. Together with Table 2, this shows that joint fine-tuning improves few-step quality while retaining most of the benefit of budget-specific training.

Table 4: ReVoC-Base before and after joint adversarial fine-tuning, evaluated on the same LibriTTS subset. Stage 2 starts from the Stage 1 checkpoint shown here.

		Stage 1		Joint Stage 2	
Model	NFE	PESQ↑	UTMOS↑	PESQ↑	UTMOS↑
ReVoC-Base	1	4.115	3.624	4.357	3.718
	2	4.375	3.753	4.409	3.799
	4	4.403	3.783	4.433	3.832

5.5 Out-of-Domain Generalization

We apply LibriTTS-trained checkpoints to EARS and VCTK with unchanged weights and inference procedures. Using the fixed, resampled utterances in Appendix A.1, these experiments test whether refinement remains effective on new corpora. Additional steps improve PESQ and UTMOS for both ReVoC variants on both datasets (Table 5). On EARS, four-step ReVoC-Base achieves the highest PESQ (4.405) and UTMOS (3.242), the lowest M-STFT (0.666), and the highest V/UV F1 (0.977) among the reported configurations. On VCTK, it obtains the highest generated-audio UTMOS (4.003), while its PESQ is close to four-step BridgeVoC (4.372 versus 4.379).

5.6 Low-Bitrate Codec Restoration

To evaluate the restoration formulation beyond Mel conditioning, we train a separate ReVoC-Small model on decoded EnCodec audio at 3 and 6 kbps, corresponding to four and eight codebooks. The codec remains fixed, and the decoded RI spectrum supplies both the source and the condition. This reuses the spectral backbone and joint multi-NFE training procedure while changing the source construction; preprocessing is described in Appendix A.1. One-step ReVoC improves all reported metrics over the original decoder and raises PESQ over token-conditioned Vocos by 0.129 and 0.124 at 3 and 6 kbps, respectively (Table 6). Four steps further increase UTMOS to 3.648 and 3.829. Most of the improvement over the original decoder is already obtained in one step, with smaller gains from additional steps than in Mel vocoding.

Table 5: Out-of-domain Mel vocoding on 200 EARS and 200 VCTK utterances, without adaptation. Bold indicates the best generated-audio result in each column, including ties.

Model	NFE	EARS					VCTK				
		PESQ↑	M-STFT↓	V/UV F1↑	UTMOS↑	ViSQOL↑	PESQ↑	M-STFT↓	V/UV F1↑	UTMOS↑	ViSQOL↑
Ground truth	–	–	–	–	3.295	–	–	–	–	4.031	–
BigVGAN-base	1	3.752	0.814	0.957	2.949	4.863	3.797	0.835	0.945	3.908	4.901
BigVGAN	1	4.231	0.718	0.972	3.147	**4.939**	4.213	0.747	0.953	3.963	**4.960**
Vocos	1	3.567	0.844	0.953	2.857	4.844	3.479	0.883	0.942	3.875	4.872
BridgeVoC	1	4.177	0.752	0.963	2.958	4.848	4.342	0.712	0.955	3.918	4.951
BridgeVoC	4	4.260	0.737	0.963	3.119	4.860	**4.379**	**0.700**	**0.958**	3.967	4.952
ReVoC-Small	1	4.068	0.783	0.965	2.834	4.877	4.176	0.821	0.951	3.862	4.900
	2	4.229	0.745	0.969	3.018	4.890	4.260	0.798	0.955	3.918	4.909
	4	4.278	0.725	0.971	3.107	4.898	4.301	0.786	0.956	3.937	4.915
ReVoC-Base	1	4.264	0.714	0.973	3.073	4.919	4.315	0.754	0.957	3.970	4.938
	2	4.378	0.680	0.975	3.207	4.929	4.350	0.741	**0.958**	3.994	4.943
	4	**4.405**	**0.666**	**0.977**	**3.242**	4.935	4.372	0.734	**0.958**	**4.003**	4.947

Table 6: Codec restoration on LibriTTS at 3 and 6 kbps. One ReVoC-Small checkpoint supports both bitrates and all three NFEs. Per. denotes periodicity RMSE.

Model	NFE	PESQ↑	M-STFT↓	V/UV F1↑	Per.↓	UTMOS↑	ViSQOL↑
3 kbps							
EnCodec	–	2.073	1.303	0.918	0.171	2.290	4.091
Vocos	1	2.575	1.177	0.940	0.137	3.459	4.219
ReVoC-Small	1	2.704	1.135	0.943	0.132	3.597	4.227
	2	2.710	1.135	0.944	0.131	3.622	4.228
	4	2.721	1.132	0.944	0.132	3.648	4.232
6 kbps							
EnCodec	–	2.711	1.183	0.950	0.118	2.933	4.352
Vocos	1	3.118	1.088	0.957	0.109	3.612	4.444
ReVoC-Small	1	3.242	1.052	0.960	0.102	3.814	4.423
	2	3.238	1.053	0.959	0.103	3.816	4.421
	4	3.244	1.052	0.960	0.102	3.829	4.424

6 Conclusion

We introduced ReVoC, a few-step audio generation framework that restores approximate reconstructions in the complex STFT domain. An efficient spectral predictor combines local patch encoding with a Vocos-style backbone, while joint multi-NFE adversarial fine-tuning enables one generator per task to support one-, two-, and four-step inference. The experiments show that patch encoding improves early convergence, positive Mel corrections improve initialization, and joint fine-tuning retains most of the quality of separately specialized generators. ReVoC offers a favorable quality–throughput trade-off in Mel vocoding, preserves refinement gains on unseen speech corpora, and improves low-bitrate codec reconstruction with one model covering both bitrates. These results support efficient spectral refinement from informative reconstructions for audio generation with adjustable inference budgets.

Limitations. ReVoC's Mel-vocoding pipeline constructs the initial spectrum through explicit Mel inversion and positivity-preserving residual corrections. These operations improve reconstruction quality but introduce additional frontend computation (Appendix B.1). This overhead does not decrease with the number of neural refinement steps, limiting further acceleration at low NFE. More efficient correction algorithms or lightweight initialization modules could help retain the quality benefits of initialization at a lower computational cost.

REFERENCES

Yang Ai and Zhen-Hua Ling. APNet: An all-frame-level neural vocoder incorporating direct prediction of amplitude and phase spectra. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023. URL https://arxiv.org/abs/2305.07952.

Michael Chinen, Felicia S. C. Lim, Jan Skoglund, Nikita Gureev, Feargus O'Gorman, and Andrew Hines. ViSQOL v3: An open source production ready objective speech and audio metric. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, 2020. doi: 10.1109/QoMEX48832.2020.9123150. URL https://arxiv.org/abs/2004.09584.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.13438.

Hui-Peng Du, Ye-Xin Lu, Yang Ai, and Zhen-Hua Ling. APNet2: High-quality and high-efficiency neural vocoder with direct prediction of amplitude and phase spectra. In Jia Jia, Zhenhua Ling, Xie Chen, Ya Li, and Zixing Zhang (eds.), *Man-Machine Speech Communication*, volume 2006 of *Communications in Computer and Information Science*, pp. 66–80. Springer, 2024. doi: 10.1007/978-981-97-0601-3_6. URL https://link.springer.com/chapter/10.1007/978-981-97-0601-3_6.

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

ITU-T. P.862.2: Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs. Technical report, International Telecommunication Union, 2005. URL https://www.itu.int/ITU-T/recommendations/rec.aspx?lang=en&rec=8670.

Won Jang, Dan Lim, Jaesam Yoon, Bongwan Kim, and Juntae Kim. UnivNet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation. In *Proceedings of Interspeech*, pp. 2207–2211, 2021. doi: 10.21437/Interspeech.2021-1016. URL https://www.isca-archive.org/interspeech_2021/jang21_interspeech.html.

Takuhiro Kaneko, Kou Tanaka, Hirokazu Kameoka, and Shogo Seki. iSTFTNet: Fast and lightweight mel-spectrogram vocoder incorporating inverse short-time fourier transform. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6207–6211, 2022. doi: 10.1109/ICASSP43922.2022.9746713. URL https://arxiv.org/abs/2203.02395.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17022–17033, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c5d736809766d46260d816d8dbc9eb44-Abstract.html.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. DiffWave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2009.09761.

Zhifeng Kong, Kevin J. Shih, Weili Nie, Arash Vahdat, Sang-gil Lee, Joao Felipe Santos, Ante Jukić, Rafael Valle, and Bryan Catanzaro. A2SB: Audio-to-audio schrödinger bridges. *arXiv preprint arXiv:2501.11311*, 2025. URL https://arxiv.org/abs/2501.11311.

Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C. Courville. MelGAN: Generative adversarial networks for conditional waveform synthesis. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/6804c9bca0a615bdb9374d00a9fcba59-Abstract.html.

Sang-gil Lee, Heeseung Kim, Chaehun Shin, Xu Tan, Chang Liu, Qi Meng, Tao Qin, Wei Chen, Sungroh Yoon, and Tie-Yan Liu. PriorGrad: Improving conditional denoising diffusion models with data-dependent adaptive prior. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=_BNiN4IjC5.

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. URL https://arxiv.org/abs/2206.04658.

Sang-Hoon Lee, Ha-Yeong Choi, and Seong-Whan Lee. Accelerating high-fidelity waveform generation via adversarial flow matching optimization. *arXiv preprint arXiv:2408.08019*, 2024. URL https://arxiv.org/abs/2408.08019.

Sang-Hoon Lee, Ha-Yeong Choi, and Seong-Whan Lee. PeriodWave: Multi-period flow matching for high-fidelity waveform generation. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/4cddc8fc57039f8fe44e23aba1e4df40-Abstract-Conference.html.

Tong Lei, Zhiyu Zhang, Rilin Chen, Meng Yu, Jing Lu, Chengshi Zheng, Dong Yu, and Andong Li. BridgeVoC: Neural vocoder with schrödinger bridge. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 8122–8130. International Joint Conferences on Artificial Intelligence Organization, 2025. doi: 10.24963/ijcai.2025/903. URL https://www.ijcai.org/proceedings/2025/903.

Andong Li, Tong Lei, Rilin Chen, Kai Li, Meng Yu, Xiaodong Li, Dong Yu, and Chengshi Zheng. BridgeVoC: Revitalizing neural vocoder from a restoration perspective. *arXiv preprint arXiv:2511.07116*, 2025a. URL https://arxiv.org/abs/2511.07116.

Andong Li, Tong Lei, Zhihang Sun, Rilin Chen, Erwei Yin, Xiaodong Li, and Chengshi Zheng. Learning neural vocoder from range-null space decomposition. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 8131–8140. International Joint Conferences on Artificial Intelligence Organization, 2025b. doi: 10.24963/ijcai.2025/904. URL https://www.ijcai.org/proceedings/2025/904.

Andong Li, Tong Lei, Lingling Dai, Kai Li, Rilin Chen, Meng Yu, Xiaodong Li, Dong Yu, and Chengshi Zheng. DegVoC: Revisiting neural vocoder from a degradation perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31510–31518, 2026a. doi: 10.1609/aaai.v40i37.40416. URL https://ojs.aaai.org/index.php/AAAI/article/view/40416.

Andong Li, Tong Lei, Zhihang Sun, Rilin Chen, Xiaodong Li, Dong Yu, and Chengshi Zheng. Scalable neural vocoder from range-null space decomposition. *arXiv preprint arXiv:2603.08574*, 2026b. URL https://arxiv.org/abs/2603.08574.

Alexander H. Liu, Qirui Wang, Yuan Gong, and James Glass. A closer look at neural codec resynthesis: Bridging the gap between codec and waveform generation. In *NeurIPS Workshop on Audio Imagination*, 2024. URL https://arxiv.org/abs/2410.22448.

Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos Theodorou, Weili Nie, and Anima Anandkumar. I^2SB: Image-to-image schrödinger bridge. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22042–22062. PMLR, 2023. URL https://proceedings.mlr.press/v202/liu23ai.html.

Peng Liu, Dongyang Dai, and Zhiyong Wu. RFWave: Multi-band rectified flow for audio waveform reconstruction. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/99088dffd5eab0babebcda4bc58bbcea-Abstract-Conference.html.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Liu_A_ConvNet_for_the_2020s_CVPR_2022_paper.html.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Tianze Luo, Xingchen Miao, and Wenbo Duan. WaveFM: A high-fidelity and efficient vocoder based on flow matching. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2187–2198. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.110. URL https://aclanthology.org/2025.naacl-long.110/.

Yuanjun Lv, Hai Li, Ying Yan, Junhui Liu, Danming Xie, and Lei Xie. FreeV: Free lunch for vocoders through pseudo inversed mel filter. In *Proceedings of Interspeech*, pp. 3869–3873, 2024. doi: 10.21437/Interspeech.2024-2407. URL https://www.isca-archive.org/interspeech_2024/lv24_interspeech.html.

Takeru Miyato and Masanori Koyama. cGANs with projection discriminator. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1802.05637.

Max Morrison, Rithesh Kumar, Kundan Kumar, Prem Seetharaman, Aaron Courville, and Yoshua Bengio. Chunked autoregressive GAN for conditional waveform synthesis. In *International Conference on Learning Representations*, 2022. URL https://www.maxrmorrison.com/pdfs/morrison2022chunked.pdf.

Tan Dat Nguyen, Ji-Hoon Kim, Youngjoon Jang, Jaehun Kim, and Joon Son Chung. FreGrad: Lightweight and fast frequency-aware diffusion vocoder. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10736–10740, 2024. doi: 10.1109/ICASSP48485.2024.10447251. URL https://arxiv.org/abs/2401.10032.

Julius Richter, Yi-Chiao Wu, Steven Krenn, Simon Welker, Bunlong Lay, Shinji Watanabe, Alexander Richard, and Timo Gerkmann. EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation. In *Interspeech 2024*, pp. 4873–4877, 2024. doi: 10.21437/Interspeech.2024-153. URL https://www.isca-archive.org/interspeech_2024/richter24_interspeech.html.

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, 2001. doi: 10.1109/ICASSP.2001.941023.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. UTMOS: UTokyo-SaruLab system for VoiceMOS challenge 2022. In *Interspeech 2022*, pp. 4521–4525, 2022. doi: 10.21437/Interspeech.2022-439. URL https://www.isca-archive.org/interspeech_2022/saeki22c_interspeech.html.

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. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/6db0903efdfe9b1bbafb015c10990b78-Abstract-Conference.html.

Christian J. Steinmetz and Joshua D. Reiss. auraloss: Audio focused loss functions in PyTorch. In *Digital Music Research Network One-day Workshop (DMRN+15)*, 2020. URL https://github.com/csteinmetz1/auraloss.

Simon Welker, Matthew Le, Ricky T. Q. Chen, Wei-Ning Hsu, Timo Gerkmann, Alexander Richard, and Yi-Chiao Wu. FlowDec: A flow-based full-band general audio codec with high perceptual quality. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2503.01485.

Yi-Chiao Wu, Dejan Marković, Steven Krenn, Israel D. Gebru, and Alexander Richard. ScoreDec: A phase-preserving high-fidelity audio codec with a generalized score-based diffusion post-filter. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 361–365, 2024. doi: 10.1109/ICASSP48485.2024.10448371. URL https://arxiv.org/abs/2401.12160.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92), 2019. URL https://doi.org/10.7488/ds/2645.

Zengwei Yao, Wei Kang, Han Zhu, Liyong Guo, Lingxuan Ye, Fangjun Kuang, Weiji Zhuang, Zhaoqing Li, Zhifeng Han, Long Lin, and Daniel Povey. Flow2GAN: Hybrid flow matching and GAN with multi-resolution network for few-step high-fidelity audio generation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/01f819fa28564f5ba4039658b4e43c69-Abstract-Conference.html.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. LibriTTS: A corpus derived from LibriSpeech for text-to-speech. In *Interspeech 2019*, pp. 1526–1530, 2019. doi: 10.21437/Interspeech.2019-2441. URL https://www.isca-archive.org/interspeech_2019/zen19_interspeech.html.

A Implementation Details

A.1 Data and Preprocessing

Mel vocoding. The main LibriTTS (Zen et al., 2019) models use 24-kHz audio and 100-band Mel features with a 1024-point Hann-window STFT and hop size 256. Separately trained LJSpeech (Ito & Johnson, 2017) models use 22.05-kHz audio and 80-band Mel features with $f_{\max} = 8$ kHz. Evaluation uses fixed lists of 208 LibriTTS and 500 LJSpeech utterances. The out-of-domain sets contain 200 EARS (Richter et al., 2024) utterances of at most 30 s and 200 VCTK (Yamagishi et al., 2019) utterances shorter than 10 s, sampled from the original 48-kHz recordings and resampled to 24 kHz. All systems recomputed in this work use the same reference utterances within each evaluation set. LibriTTS training uses peak-normalized crops of 2.72 s in Stage 1 and 1.5 s in Stage 2; the gain is restored before full-utterance evaluation. Baselines retain their released preprocessing.

Codec restoration. The fixed 24-kHz EnCodec (Défossez et al., 2023) uses four and eight codebooks at 3 and 6 kbps, respectively. Clean and decoded segments share crop indices and a normalization gain derived from the decoded audio. The decoded RI spectrum supplies both the bridge source and the reconstruction condition; the codec encoder, quantizer, and decoder remain fixed.

A.2 Source Construction

For a Mel filterbank $M \in \mathbb{R}_+^{K\times F}$ and linear Mel magnitude $v = \exp(y)$, we precompute $M^\dagger$ and initialize

$$a^{(0)} = \max(M^\dagger v, \epsilon). \tag{9}$$

Two positivity-preserving corrections reduce the Mel residual introduced by clipping:

$$\begin{aligned} \big[M\,\mathrm{diag}(a^{(r)})M^\top\big]\delta^{(r)} &= v - Ma^{(r)}, \\ a^{(r+1)} &= a^{(r)} \odot \exp\big[\mathrm{clip}(M^\top\delta^{(r)}, -4, 4)\big], \qquad r = 0, 1. \end{aligned} \tag{10}$$

The triangular filterbank gives a tridiagonal system on its active support. We use a numerical floor of $\epsilon = 10^{-5}$ and initialize phase from the STFT of a real Gaussian waveform ξ:

$$U_\xi = \frac{\mathcal{S}(\xi)}{\max(|\mathcal{S}(\xi)|, \epsilon)}, \qquad \mathcal{R}_{\mathrm{mel}}(y, \xi) = a^{(2)} \odot U_\xi. \tag{11}$$

DC and Nyquist imaginary components are set to zero. Spectral compression uses $(\kappa, p) = (0.33, 0.5)$, and the Mel condition is the compressed source magnitude. Codec sources are obtained directly from the STFT of decoded waveforms, as in Eq. equation 4.

A.3 Spectral Predictor Architecture

The predictor uses a Vocos-style ConvNeXt backbone (Siuzdak, 2024; Liu et al., 2022). Its patch encoder processes the first 512 input frequency bins with a $3 \to 24$ convolution, kernel $(16, 3)$, stride $(8, 1)$, and padding $(4, 1)$, followed by PReLU. Flattening 64 patches gives 1536 features per frame, which a kernel-7 temporal convolution projects to the backbone width. The output head predicts all 513 RI frequency bins. The direct-flattening reference, labeled Vocos(Flow) in Figure 4, retains all 513 input bins and uses the same temporal projection and backbone. For codec restoration, the Small model receives four channels: the current RI state and decoded-source RI.

A.4 Training Details

Table 7 gives the model configurations, and Table 8 summarizes the optimization settings for the main LibriTTS experiments. Both stages use AdamW (Loshchilov & Hutter, 2019) and EMA. Stage 2 alternates generator and discriminator updates after discriminator warm-up.

Bridge pretraining. The diffusion schedule and its integral are

$$\beta(t) = 0.01 + (20 - 0.01)t, \qquad B_t = 0.01t + \tfrac{1}{2}(20 - 0.01)t^2, \tag{12}$$

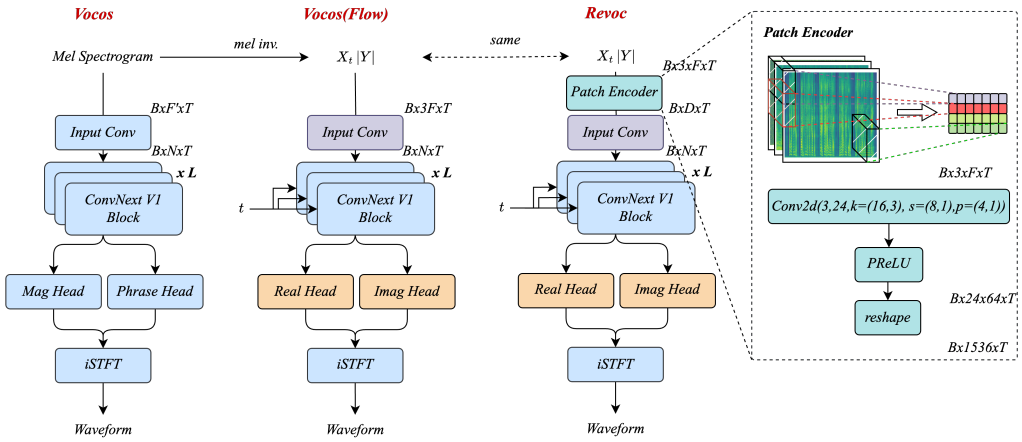

Figure 4: Spectral generator architectures. Vocos predicts magnitude and phase; the two time-conditioned predictors output RI spectra. Vocos(Flow) denotes the direct-flattening reference, and ReVoC adds a local patch encoder.

Table 7: Mel-conditioned predictor configurations. Patch kernels and strides are given in frequency–time order. Vocos(Flow) is the direct-flattening reference; its Small variant is used in the patch-encoder ablation. Parameter counts exclude discriminators.

Configuration	ReVoC-Small	ReVoC-Base	Vocos(Flow), Small / Base
Input channels	$[\mathrm{Re}\, z_t, \mathrm{Im}\, z_t, \lvert z_1\rvert]$	same	same
STFT bins / hop	513 / 256	513 / 256	513 / 256
Patch convolution	$3 \to 24$	$3 \to 24$	none
Patch kernel / stride	$(16,3)/(8,1)$	$(16,3)/(8,1)$	none
Frequency patches	64	64	none
Features before projection	1536	1536	1539
Input projection kernel	7	7	7
Backbone width	512	768	512 / 768
Backbone blocks	8	12	8 / 12
Depthwise kernel	7	7	7
MLP expansion	3	3	3
Time embedding width	128	128	128
Output RI channels	1026	1026	1026
Parameters (M)	19.309	52.947	19.316 / 52.960

with noise scale $\gamma = 1$. Inference uses uniformly spaced predictor times for NFE 1, 2, and 4, followed by a final update to $t = 0$. The multi-resolution log-Mel loss has weight 0.1 and averages seven resolution-wise losses. With $\mathcal{P}(z) = \mathcal{C}(\mathcal{S}(\mathcal{V}(z)))$, the auxiliary term is

$$\mathcal{L}_{\mathrm{aux}} = \rho(u)\big[0.1\ \mathrm{MSE}(\mathcal{P}(\hat{z}_0), z_0) + 0.05\ \mathrm{MSE}(\hat{z}_0, \mathcal{P}(\hat{z}_0))\big]. \tag{13}$$

The weight $\rho(u)$ is zero for the first 100k updates and increases linearly to one over the next 20k. Thus, the patch-encoder comparison in Figure 3 covers training before these auxiliaries take effect.

Joint adversarial fine-tuning. The multi-period discriminator uses periods $(2, 3, 5, 7, 11)$; the multi-resolution discriminator uses FFT sizes $(2048, 1024, 512)$ with quarter-window hops and five frequency bands. Both receive NFE labels through projection conditioning (Miyato & Koyama, 2018). For each branch D, the adversarial losses are

$$\begin{aligned} \mathcal{L}_D &= \mathbb{E}[\max(0, 1 - D(x, N)) + \max(0, 1 + D(\hat{x}_N, N))], \\ \mathcal{L}_{G,\mathrm{branch}} &= \mathbb{E}[\max(0, 1 - D(\hat{x}_N, N))]. \end{aligned} \tag{14}$$

Feature matching uses L_1 distances between real and generated discriminator features. The multi-period and multi-resolution families have weights 1 and 0.1, respectively, for both adversarial

Table 8: Training configurations for the main LibriTTS experiments. G and D denote the generator and discriminators. Stage 2 counts refer to training minibatches.

Setting	Stage 1	Joint Stage 2
Sample rate	24 kHz	24 kHz
Crop length	65,280 samples (2.72 s)	36,000 samples (1.5 s)
Global batch size	64	16
Optimizer	AdamW	AdamW (G and D)
Initial learning rate	5×10^{-4}	G: 2×10^{-5}; D: 2×10^{-4}
Adam coefficients	$(0.8, 0.99)$	$(0.8, 0.99)$
Weight decay	0.01	0
Gradient norm clipping	disabled	2
EMA decay	0.999	0.999
Precision	FP32	FP32
Training minibatches	500k	150k
Discriminator warm-up	–	6k minibatches
G / D updates	–	72k / 78k total
Fine-tuning budgets	–	balanced $N \in \{1, 2, 4\}$

and feature-matching terms. In Eq. equation 8, $(\lambda_{\mathrm{FM}}, \lambda^{G}_{\mathrm{mel}}) = (1, 45)$. The Mel loss sums seven resolutions with FFT sizes $(32, 64, 128, 256, 512, 1024, 2048)$, quarter-window hops, and $(5, 10, 20, 40, 80, 160, 320)$ Mel bands. Generator updates backpropagate through the complete selected trajectory.

Joint fine-tuning uses 150k minibatches, including 6k discriminator warm-up minibatches. Balanced exposure to NFE 1, 2, and 4 gives each budget 24k generator and 26k discriminator updates. Each fixed-NFE run uses 50k minibatches, including 2k warm-up, and therefore matches these per-budget update counts. The comparison in Table 2 starts from the same Stage 1 checkpoint. Each final joint EMA checkpoint is evaluated at all three NFEs; the codec checkpoint also supports both bitrates.

A.5 Metric Implementations and Baselines

PESQ (Rix et al., 2001; ITU-T, 2005) uses wideband mode at 16 kHz. M-STFT uses the default multi-resolution STFT loss in `auraloss` (Steinmetz & Reiss, 2020). V/UV F1 and periodicity RMSE use the CARGAN evaluation pipeline (Morrison et al., 2022) at 22.05 kHz. UTMOS uses the UTMOS22-strong predictor (Saeki et al., 2022), and ViSQOL (Chinen et al., 2020) uses the Python speech-mode implementation supplied with Flow2GAN (Yao et al., 2026) at 16 kHz. Prediction and reference waveforms are length matched before scoring. PESQ, M-STFT, UTMOS, and ViSQOL are averaged over utterances; V/UV F1 and periodicity RMSE aggregate across frames.

The LibriTTS, out-of-domain, and codec comparisons use released baseline checkpoints with their native frontends and synthesis procedures. BridgeVoC uses its dedicated one-step checkpoint and released four-step configuration; Flow2GAN uses its released one-step checkpoint. The LJSpeech baseline values are taken from the reports cited in Table 11.

A.6 Inference Efficiency

Throughput is the nominal audio duration of a batch divided by its mean inference time, reported as a multiple of real time (×RT). Timing includes source initialization, the complete neural rollout, and waveform synthesis, and excludes Mel extraction, file I/O, and host–device transfer. Parameter counts include only the generator. GMACs count convolutional and linear operations over the rollout; measured throughput also includes spectral transforms and other synthesis operations. These are batched-throughput measurements.

Table 9: Mel-vocoding efficiency on a single RTX 4090 GPU and eight pinned threads of an Intel Xeon Gold 6346 CPU. Measurements use FP32, batch size 16, and nominal one-second inputs. GMACs count neural operations per item over the full rollout.

Model	NFE	Params (M)	GMACs	CPU ×RT↑	GPU ×RT↑
BigVGAN-base	1	14.02	32.95	1.3	101.4
BigVGAN	1	112.42	89.98	0.9	72.6
Vocos	1	13.53	1.27	233.5	8036.8
RFWave	10	18.14	25.27	1.3	94.3
BridgeVoC	1	7.89	8.51	6.7	722.8
BridgeVoC	4	7.89	34.04	1.5	183.4
Flow2GAN	1	78.95	10.81	26.0	1113.2
ReVoC-Small	1	19.31	1.78	83.3	4776.2
	2	19.31	3.55	57.3	2884.8
	4	19.31	7.10	35.8	1600.0
ReVoC-Base	1	52.95	4.88	52.6	2563.9
	2	52.95	9.75	31.5	1411.6
	4	52.95	19.50	17.7	738.3

Table 10: Mel-inversion error and latency on 100 two-second clips. Latency uses an RTX 4090 or one CPU thread and excludes phase construction and neural inference. The pseudoinverse is precomputed.

Magnitude construction	Mel error (%)↓	GPU (ms)↓	CPU (ms)↓
Pseudoinverse + absolute value	3.343	0.045	0.260
+ ReLU	1.671	0.035	0.198
+ ReLU + 1 correction	0.504	0.534	3.973
+ ReLU + 2 corrections	0.061	0.999	7.636
+ ReLU + 2 corrections, cached structure	0.061	0.434	6.903

B Additional Results

B.1 Mel Initialization Cost

Table 10 complements the downstream quality comparison in Table 3. Two positive corrections reduce the relative Mel reconstruction error from 1.671% to 0.061% on 100 two-second clips.

Caching the filterbank structure reduces GPU initialization time from 0.999 to 0.434 ms while retaining the same Mel residual. These timings isolate Mel inversion and use a single CPU thread; the system measurements in Table 9 include the full synthesis pipeline.

B.2 LJSpeech Evaluation

We additionally train ReVoC on LJSpeech with 80-band Mel conditioning. The comparison includes BigVGAN, RNDVoC, FreGrad, PeriodWave, APNet, APNet2, FreeV, and Vocos (Lee et al., 2023; Li et al., 2025b; Nguyen et al., 2024; Lee et al., 2025; Ai & Ling, 2023; Du et al., 2024; Lv et al., 2024; Siuzdak, 2024). Baseline scores are taken from the BridgeVoC and RNDVoC reports (Li et al., 2025a; 2026b). Within both ReVoC variants, increasing NFE consistently improves all five metrics. Four-step ReVoC-Base reaches PESQ 4.367 and UTMOS 4.367 with the same checkpoint used at NFE 1 and 2.

B.3 Pairwise Listening Preferences

We compare four-step ReVoC-Base with Vocos, BigVGAN-base, and four-step BridgeVoC using the interface in Figure 5. Listeners hear samples A and B and choose one of five responses: strongly or slightly prefer A, no difference, or slightly or strongly prefer B. We pool the two preference

Table 11: Objective Mel-vocoding results on LJSpeech. Baseline scores are taken from prior reports (Li et al., 2025a; 2026b); ReVoC is evaluated on 500 test utterances. Each ReVoC variant shares one checkpoint across NFEs. Bold indicates the best reported value in each column.

Model	NFE	PESQ↑	M-STFT↓	V/UV F1↑	UTMOS↑	ViSQOL↑
BigVGAN-base	1	3.603	0.978	0.956	4.208	4.822
BigVGAN	1	4.107	0.900	0.972	4.314	**4.958**
RNDVoC	1	3.987	0.910	0.971	4.161	4.837
FreGrad	50	3.775	1.344	0.943	3.934	4.574
PeriodWave	16	4.235	1.098	0.967	**4.374**	4.722
APNet	1	3.390	1.266	0.945	3.170	4.695
APNet2	1	3.492	0.982	0.959	3.984	4.752
FreeV	1	3.593	1.008	0.960	4.010	4.743
Vocos	1	3.522	1.012	0.956	3.968	4.774
ReVoC-Small	1	4.100	0.955	0.975	3.890	4.916
	2	4.257	0.929	0.977	4.234	4.923
	4	4.294	0.917	0.978	4.322	4.928
ReVoC-Base	1	4.290	0.897	0.978	4.197	4.949
	2	4.351	0.881	0.980	4.353	4.953
	4	**4.367**	**0.876**	**0.981**	4.367	4.955

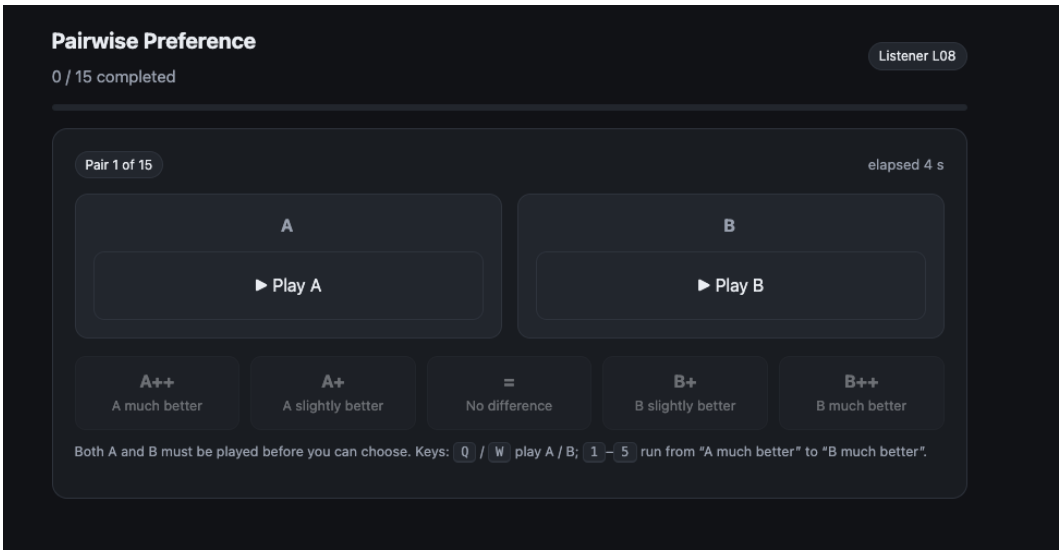

Figure 5: Web interface for the pairwise listening test, with five response choices from strongly preferring A to strongly preferring B.

strengths in each direction into wins and losses and retain no-difference responses as ties. Eight listeners contribute 80 votes per model comparison, for 240 votes in total.

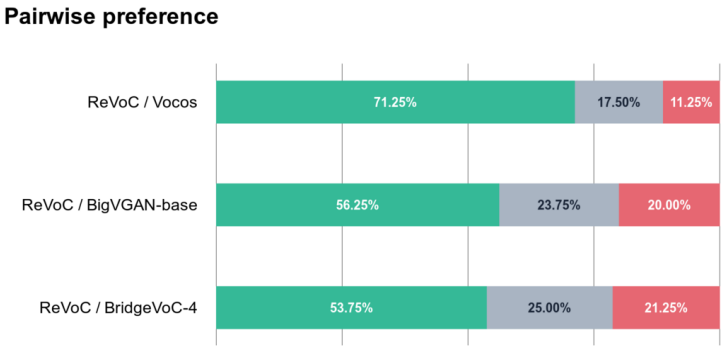

Figure 6: Pairwise preferences for four-step ReVoC-Base against baseline vocoders, with 80 votes per comparison. Green, gray, and red indicate preference for ReVoC, no difference, and preference for the baseline, respectively.

ReVoC receives 71.25%, 56.25%, and 53.75% of preference votes against Vocos, BigVGAN-base, and four-step BridgeVoC, respectively (Figure 6). The corresponding baseline preference rates are 11.25%, 20.00%, and 21.25%, with the remaining votes recorded as ties.

C Use of Large Language Models

We used a large language model to assist with manuscript editing, evaluation-code inspection, and organizing experimental results. Reported metrics were computed using the stated evaluation procedures. The authors remain responsible for the manuscript and its conclusions.