RESTORA: CONTROLLING GENERATIVE FREEDOM FOR FAITHFUL AND UNIVERSAL AUDIO RESTORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Universal audio restoration aims to recover speech, singing, and music from diverse and compounded degradations with a single model. Generative models suit this goal because they synthesize clean content rather than simply filtering the degraded input, but this creative freedom can also produce hallucinations: plausible-sounding output unsupported by the original signal, so controlling how far generation departs from the input is essential. Two standard mechanisms govern this freedom: reward-based post-training refines the model toward perceptual quality targets, and inference-time guidance steers each sample toward the observed condition. As commonly implemented, each has a specific weakness that we address. Uniform reward optimization pushes every audio domain equally hard, so a domain that has saturated continues to receive noisy updates that degrade held-out quality; Headroom-Anchored RL instead tracks each domain on its own and, once reward signals become unreliable, shifts that domain back toward a stable supervised target. Classifier-free guidance (CFG) steers away from a null reference that discards the degraded input, the very evidence that faithful restoration relies on; Condition-Swap Guidance (CSG) replaces that null reference with a permuted copy of the input's own conditioning segments—same content, wrong timing—giving a structurally aware direction that improves perceptual quality and content accuracy together. Combining both mechanisms, RESTORA, a single 500M flow-matching model, raises the mean opinion score of degraded inputs from 2.55 to 4.12 across six test sets and achieves the best spectral-fidelity and aesthetic scores among compared music systems. CSG also transfers to a discrete-token speech enhancement model, showing it is not limited to diffusion guidance. Demo page: `https://restora-demo.netlify.app/`.

1 INTRODUCTION

Audio restoration has long been divided into specialized tasks, with distinct models developed for degradation types such as denoising (Reddy et al., 2020; 2021) and dereverberation (Kinoshita et al., 2013), and for signal domains such as speech (Zhao et al., 2022; Richter et al., 2023; Lemercier et al., 2023; Fu et al., 2026), singing voice (Xu et al., 2023; Jiang et al., 2025), and music (Liu et al., 2026; Kandpal et al., 2022; Chae et al., 2023; Moliner et al., 2024; Melechovsky et al., 2025). Generative modeling can bridge these divisions by changing the objective: instead of learning a deterministic map from a degraded signal to a clean one, a generative restorer learns the distribution of clean audio and resynthesizes a sample conditioned on the observation (Li et al., 2024a; Zhang et al., 2025). Its main advantage is the pre-established knowledge of the prior distribution of the clean data rather than simply filtering the degraded input. Therefore, a single generative model can handle complex, compound degradations and restore content that the original observation no longer carries.

This flexibility also creates a risk. The same capability that allows models to diverge from the input can produce hallucinations: content that sounds plausible but is not supported by the original source (Wang et al., 2026b; Rong et al., 2026a;b;c). The standard training objective lacks an explicit mechanism to ensure faithfulness, teaching the model to prioritize statistical likelihood over factual consistency with the input evidence. In speech, a model may change words and compromise content

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.

integrity; in music, it may alter melodies and violate artistic intent. Faithful restoration therefore requires controlling to avoid introducing hallucinations.

We argue that this loss of control appears at two stages where generative behavior is shaped: during post-training, where the model is refined toward quality targets via reward optimization, and during inference, where guidance steers each sample given the observed condition. **At the post-training stage,** to jointly optimize multiple domains such as speech and music, a single model is often trained with the same reinforcement learning (RL) objective and update strength across domains (e.g., direct preference optimization (DPO) (Zhang et al., 2026a), FlowSE-GRPO (Wang et al., 2026a)). However, we observe that the achievable gains from reward optimization yields uneven and asynchronous outcomes across domains (Gao et al., 2023): speech quality may improve while music fidelity downgrades. The training process lacks a mechanism to recognize whether reward-driven updates can strike a balance across domains. **At inference,** classifier-free guidance (CFG) (Ho & Salimans, 2022) steers conditional generation by extrapolating away from a null condition. Stronger guidance can initially improve faithfulness, but large extrapolation can reduce perceptual quality. At higher weights, both content accuracy and perceptual quality can degrade. This suggests that the negative reference itself, not only the guidance weight, matters.

Our solution to these two problems is to use signals already present within the system. (1) During post-training, for each domain, **Headroom-Anchored RL** tracks whether the gap between top-scoring and mean rollout rewards (Zheng et al., 2025b) is closing. This trend indicates whether RL is still yielding genuine quality gains. When improvements stop, Headroom-Anchored RL shifts the training objective from RL toward the supervised fine-tuning target. (2) At inference, we propose **Condition-Swap Guidance (CSG)**. It replaces CFG's null reference with one formed by permuting conditioning chunks within the sample itself. The permuted condition shares the same content as the original but has wrong timing, so the guidance direction isolates the model's reliance on correct structural alignment. At the evaluated operating points, CSG provides the best joint improvement in content accuracy and perceptual quality among the guidance methods we compare.

Combining these two mechanisms, we present RESTORA, a single 500M flow-matching (Lipman et al., 2023) model that restores speech, singing voice, full-song mixtures, and instrumental music. Headroom-Anchored RL stabilizes multi-domain post-training: over the same 1k–3k training interval, uniform RL degrades music log-spectral distance roughly 2.5$\times$ faster than Headroom-Anchored RL, while Headroom-Anchored RL keeps most of the speech gain. At inference, CSG simultaneously improves naturalness and reduces word error rate on speech, and improves both spectral fidelity and perceptual quality on music. Across six test sets, RESTORA raises the mean opinion score of degraded inputs from 2.55 to 4.12 and attains the best Fréchet audio distance (FAD) and aesthetic scores among the compared music systems. CSG also transfers without retraining to a discrete-token speech enhancement model (Zhang et al., 2025) and to the image restoration model DiffBIR (Lin et al., 2024). To summarize, our contributions are as follows:

- Headroom-Anchored RL, a training-time controller that tracks reward headroom per domain and interpolates between RL and a supervised target accordingly.
- Condition-Swap Guidance, an inference-time guidance method that replaces the null negative reference with a misaligned condition constructed from the input sample.
- The RESTORA system, a single 500M model that restores speech, singing, full-song, and instrumental audio. Transfer experiments further show that CSG generalizes to a discrete-token speech model and to image restoration.

2 RELATED WORK

Universal audio restoration. Audio restoration has moved from task-specific pipelines toward unified models that handle multiple distortions jointly. VoiceFixer (Liu et al., 2022a) was among the first, cascading a ResUNet-based Mel restorer with a GAN vocoder to deal with noise, reverberation, clipping, and bandwidth limitation. UNIVERSE (Serrà et al., 2022) and UNIVERSE++ (Scheibler et al., 2024) extended the scope to codec artifacts and packet loss via score-based diffusion. With the rise of generative modeling, architectures have diversified: MaskSR (Li et al., 2024a) applies masked generative modeling to general speech restoration; AnyEnhance (Zhang et al., 2025) adds in-context-learning prompt guidance and semantic distillation (Liu et al., 2025b; Yu et al., 2025); LLaSE-G1 (Kang et al., 2025) uses a LLaMA backbone for multi-task enhancement; and inspired

by zero-shot text-to-speech (TTS) (Du et al., 2024), SenSE (Li et al., 2025b) autoregressively generates semantic tokens for flow-matching to synthesize the Mel spectrogram. Alongside architecture advances, recent work addresses the distortion–perception tradeoff directly: UniPASE (Rong et al., 2026c) introduces a distilled WavLM phonological prior (Chen et al., 2022; Rong et al., 2026b) to reduce hallucinations in a cascaded restorer–upsampler pipeline, and RE-USE (Fu et al., 2026) adopts a two-stage regression-then-GAN training guided by rate–distortion–perception theory. In the music domain, SonicMaster (Melechovsky et al., 2025) is the first unified generative model for music restoration and mastering, supporting text-conditioned or automatic modes. The URGENT challenge series (Zhang et al., 2024; Saijo et al., 2025; Li et al., 2026) has further catalyzed the community by benchmarking universal speech enhancement across noise, reverb, bandwidth, and clipping conditions. Despite this rapid progress, most systems still train with likelihood-based objectives that can misalign with human perception. When a single model serves multiple domains, this misalignment manifests unevenly—motivating the per-domain reward control we propose.

Reward-driven post-training for audio restoration. Post-training alignment via reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022) and DPO (Rafailov et al., 2023) is now standard for generative language and image models, and has recently entered audio restoration. Proximal policy optimization (PPO) (Schulman et al., 2017) was first applied to a ratio-mask speech enhancer by perturbing the output mask to introduce stochasticity, optimizing toward mean-opinion-score (MOS) rewards (Kumar et al., 2025). DPO was then adopted with a single UTMOS (Saeki et al., 2022) reward (Li et al., 2025a). Subsequent work moved to richer reward signals: multi-metric DPO jointly balances quality, fidelity, and speaker similarity (Zhang et al., 2026a); FlowSE-GRPO (Wang et al., 2026a) applied group relative policy optimization (GRPO) (Shao et al., 2024), which Flow-GRPO (Liu et al., 2025a) adapted to flow-matching models, with multiple rewards to a flow-matching enhancer; and group sequence policy optimization (GSPO) (Zheng et al., 2025a) was brought to speech-enhancement language models (Berdoz et al., 2026). In visual generation, DiffusionNFT (Zheng et al., 2025b) contrasts reward-positive and reward-negative generations on the forward process, providing a likelihood-free alternative to reverse-process RL. All these methods apply **uniform** optimization pressure across domains; when a single model serves speech, singing, and music, each domain saturates at a different rate, and continued optimization past saturation degrades held-out quality (Gao et al., 2023). Headroom-Anchored RL detects this per-domain saturation and gates the learning signal accordingly.

3 RESTORA FOR UNIVERSAL AUDIO RESTORATION

3.1 PRELIMINARIES

Flow matching. Diffusion and flow models (Lipman et al., 2023; Liu et al., 2023) learn to reverse a forward noising process that interpolates between clean data x_0 and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$. Under the rectified-flow schedule $x_t = t\, x_0 + (1{-}t)\, \epsilon$, a velocity network $v_\theta(x_t, t, c)$ is trained to predict the tangent of the interpolation path, i.e. the target velocity $v = x_0 - \epsilon$. The training objective is

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{t,\, x_0,\, \epsilon}\big[\, \|v_\theta(x_t, t, c) - (x_0 - \epsilon)\|^2 \,\big], \tag{1}$$

and samples are generated by solving the ODE $\mathrm{d}x_t/\mathrm{d}t = v_\theta(x_t, t, c)$ from $t{=}0$ (noise) to $t{=}1$ (data).

Classifier-free guidance. To strengthen adherence to the condition c at inference time, CFG (Ho & Salimans, 2022) extrapolates between the conditional and an unconditional ($c{=}\varnothing$) prediction:

$$\hat{v} = v_\theta(x_t, t, c) + \omega\big(v_\theta(x_t, t, c) - v_\theta(x_t, t, \varnothing)\big), \tag{2}$$

where ω controls the guidance strength.

DiffusionNFT. While policy-gradient methods such as GRPO discretize the reverse sampling process to apply RL, DiffusionNFT (Zheng et al., 2025b) operates directly on the forward process. For each prompt, K samples are generated and scored by a reward function; the raw reward is normalized to an optimality probability $r \in [0, 1]$. A single model v_θ is then optimized via implicit positive and negative branches:

$$\mathcal{L}_{\mathrm{NFT}} = \mathbb{E}\big[\, r\, \|v_\theta^+ - v\|^2 + (1 - r)\, \|v_\theta^- - v\|^2 \,\big], \tag{3}$$

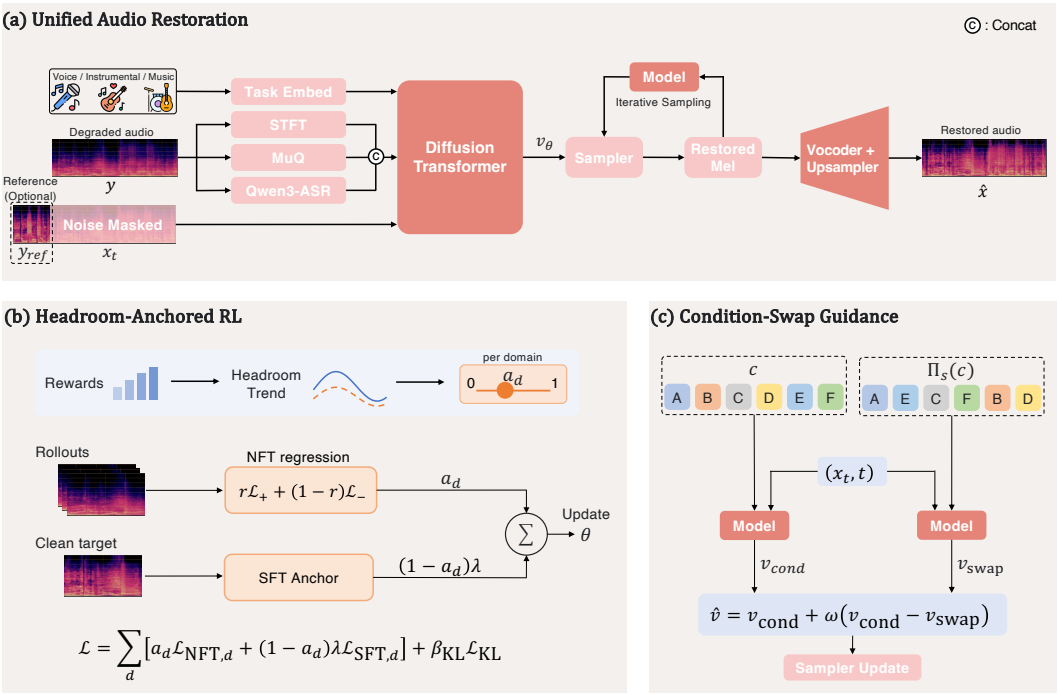

Figure 1: RESTORA overview. (a) A single flow-matching model restores speech, singing, full-song, and instrumental audio, conditioned on three complementary streams extracted from the degraded input: Qwen3-ASR (speech semantics), MuQ (music semantics), and STFT (acoustic features). (b) Headroom-Anchored RL tracks per-domain reward headroom during post-training and gates between RL and a supervised anchor accordingly. (c) Condition-Swap Guidance constructs a negative reference by permuting conditioning chunks within the sample, then steers sampling away from the temporally misaligned prediction.

where $v_\theta^+ = (1-\beta)\, v^{\text{old}} + \beta\, v_\theta$ and $v_\theta^- = (1+\beta)\, v^{\text{old}} - \beta\, v_\theta$. The positive branch drives the policy toward high-reward samples while the negative branch steers it away from low-reward ones, defining an implicit improvement direction without requiring likelihood estimation.

3.2 SYSTEM OVERVIEW

RESTORA is a Diffusion Transformer (Peebles & Xie, 2023) (DiT) that restores speech, singing voice, full-song mixtures, and instrumental music with a single set of weights. The model operates on Mel spectrograms: x_0 is the clean target and the condition c concatenates three streams extracted from the degraded input y—Qwen3-ASR (Shi et al., 2026) (speech semantics), MuQ (Zhu et al., 2025) (music semantics), and power-compressed STFT (acoustic features)—both semantic encoders being stabilized via representation distillation (Rong et al., 2026b;c) (Figure 1a). Optionally, a clean reference clip y_{ref} is concatenated as an in-context prompt to better preserve speaker identity or timbral characteristics (Appendix Tables 16 and 17). The restored Mel $\hat{x}_0$ is converted to 24 kHz audio by a Vocos vocoder (Siuzdak, 2023); for listening, a second Vocos model (Sharma, 2026) extends it to 48 kHz. Training proceeds in three stages: flow-matching pre-training ($\mathcal{L}_{\text{FM}}$), supervised fine-tuning on quality-filtered data, and DiffusionNFT post-training. The latter two stages introduce Headroom-Anchored RL (§3.3) and Condition-Swap Guidance (§3.4), described next. Further implementation details are in Appendix A.1.

3.3 HEADROOM-ANCHORED REINFORCEMENT LEARNING

Applying DiffusionNFT with domain-specific composite rewards uniformly across domains exposes reward over-optimization: the training reward becomes unstable and held-out fidelity degrades, even

Algorithm 1 Headroom-Anchored RL (one iteration)

Require: Policy θ, reference θ_{ref}; per-domain EMA state; τ, q, β_f, β_s, λ, β_{KL}

Roll out K samples per prompt with θ_{old}; decode to waveforms

Score with domain-routed reward; per-key group-normalize; combine and per-domain renormalize $\rightarrow$ implicit target r

for each domain d **do**

  $g_d \leftarrow \overline{R}_d^{(q)} - \text{mean}(R_d)$ ▷ top-q fraction vs. mean gap

  fast $\leftarrow \text{EMA}_{\beta_f}(g_d)$, slow $\leftarrow \text{EMA}_{\beta_s}(g_d)$

  $z \leftarrow (\text{fast} - \text{slow}) \,/\, \text{scale}_d$ ▷ self-normalized slope

  $a_d \leftarrow \text{clip}(\sigma(-z/\tau),\ a_{\min},\ 1)$ ▷ gap closing $\Rightarrow a_d \rightarrow 1$

  if warmup **then**

    $a_d \leftarrow 1$

  end if

end for

$\mathcal{L} \leftarrow \sum_d \big[a_d\, \mathcal{L}_{\text{NFT},d}(r) + (1 - a_d)\, \lambda\, \mathcal{L}_{\text{SFT},d}\big] + \beta_{\text{KL}}\, \text{KL}(\theta \| \theta_{\text{ref}})$

Update θ with $\nabla_\theta \mathcal{L}$

though the RL objective itself does not signal a problem (Figure 2a vs. b). This degradation is domain-dependent—music fidelity deteriorates long before speech quality does—because each domain saturates its reward at a different rate. A natural remedy is to mix RL with supervised fine-tuning so that the data distribution anchors the policy when reward signals become unreliable (Chen et al., 2026; Ma et al., 2026); the remaining question is **when** and **how much** to shift.

We detect this saturation by tracking the **reward headroom**, defined as the gap between the top-q fraction mean and the batch mean of the reward vector R_d within each domain d:

$$g_d = \overline{R}_d^{(q)} - \text{mean}(R_d), \tag{4}$$

where $\overline{R}_d^{(q)}$ averages the highest q-fraction of scores in the batch. A narrowing gap means the policy mean is catching up with the best rollouts—the high-reward behavior is reproducible and RL should continue. A widening gap, by contrast, indicates that top-scoring samples are pulling away from the mean without the mean following: these outliers likely exploit weaknesses in the reward model rather than representing genuine quality gains. We monitor this trend via a fast–slow exponential moving average (EMA) pair:

$$\text{slope}_d = \text{EMA}_{\beta_{\text{fast}}}(g_d) - \text{EMA}_{\beta_{\text{slow}}}(g_d), \tag{5}$$

and convert the self-normalized slope into a per-domain gate:

$$a_d = \text{clip}\left(\sigma\left(\frac{-\,\text{slope}_d}{\text{scale}_d \cdot \tau}\right),\ a_{\min},\ 1\right), \tag{6}$$

where $\text{scale}_d = \text{EMA}_{\beta_{\text{fast}}}(|\text{slope}_d|)$ is the running mean magnitude of the slope, which makes the gate invariant to the reward scale of each domain, and τ controls how sharply the gate responds. The per-domain loss interpolates between DiffusionNFT and a supervised anchor:

$$\mathcal{L} = \textstyle\sum_d \big[\, a_d\, \mathcal{L}_{\text{NFT},d} + (1 - a_d)\, \lambda\, \mathcal{L}_{\text{SFT},d}\,\big] + \beta_{\text{KL}}\, \mathcal{L}_{\text{KL}}, \tag{7}$$

where $\mathcal{L}_{\text{SFT},d}$ is the flow-matching loss on clean targets. When a_d is high, the domain trains via RL; when headroom is exhausted and a_d drops, the SFT term anchors the model to the data distribution, preventing the drift that unchecked RL would cause. Setting $a_d \equiv 1$ recovers Fully-RL; $a_d \equiv 0$ recovers continued SFT. Algorithm 1 summarizes one training iteration.

3.4 Condition-Swap Guidance

In audio restoration, the condition c is the degraded input itself—it carries the content evidence the model must preserve. CFG (Eq. 2) steers generation away from a negative reference (like a null condition $\varnothing$ that discards this evidence entirely). In text-to-image generation, where $\varnothing$ simply means "no text prompt," this is a reasonable unconditional baseline; in restoration, it removes the very signal that defines faithfulness, which means the negative reference needs a more suitable substitute in this case. In practical, the useful guidance range is consequently narrow and modest

Table 1: Speech and singing restoration. RESTORA is a single 500M non-autoregressive model covering four audio domains; all other systems except AnyEnhance are speech-only. **Bold**/underline: best/second-best per column. On speech, speech-only systems (SenSE, RE-USE) lead most perceptual columns; RESTORA remains competitive as the only multi-domain model. On singing, RESTORA leads four of six columns.

Model	SIG↑	BAK↑	OVRL↑	NISQA↑	SBS↑	SIM↑
LibriVox-GSR (speech)						
VoiceFixer	3.365	3.971	3.085	3.770	0.706	0.864
AnyEnhance	3.546	4.142	3.308	4.346	0.822	0.955
LLaSE-G1	3.456	4.034	3.159	3.834	0.711	0.890
PASE	3.359	3.985	3.068	3.656	0.838	0.944
UniPASE	3.553	4.101	3.298	4.259	**0.862**	**0.959**
UniSE	**3.608**	4.091	3.339	3.933	0.747	0.911
SenSE	3.587	**4.175**	**3.375**	**4.719**	0.840	0.957
RE-USE	3.596	4.174	**3.375**	4.485	0.842	0.952
RESTORA	3.594	4.170	3.353	4.590	0.800	0.941
CCMusic-GSR (singing)						
VoiceFixer	2.750	3.094	2.354	2.917	0.636	0.823
AnyEnhance	3.235	3.542	2.790	3.327	**0.811**	**0.914**
LLaSE-G1	2.983	3.606	2.589	3.111	0.627	0.820
PASE	3.047	3.559	2.642	3.451	0.712	0.874
UniPASE	3.184	3.516	2.733	3.304	0.756	0.888
UniSE	3.130	3.431	2.718	3.628	0.726	0.852
SenSE	3.272	3.651	2.871	3.753	0.734	0.855
RE-USE	3.324	**3.840**	2.956	3.793	0.794	0.911
RESTORA	**3.350**	3.816	**2.978**	**4.317**	**0.811**	0.909

strengths help, but stronger extrapolation away from $\varnothing$ degrades both perceptual quality and content accuracy.

CSG replaces $\varnothing$ with a permuted copy $\tilde{c} = \Pi_S(c)$ of the sample's own condition:

$$\hat{v} = v_\theta(x_t, t, c) + \omega\big(v_\theta(x_t, t, c) - v_\theta(x_t, t, \tilde{c})\big), \tag{8}$$

where ω controls guidance strength. To construct $\tilde{c}$, we partition c into temporal chunks and permute dissimilar segments so that each chunk lands in a mismatched position (Figure 1c). The resulting $\tilde{c}$ keeps the content of c but has the wrong timing. The guidance direction $v_\theta(\cdot, c) - v_\theta(\cdot, \tilde{c})$ therefore isolates the model's reliance on correct structural alignment; extrapolating along it pushes the sample toward stronger consistency with the input.

Negative prompting in text-to-image generation (Liu et al., 2022b; Ban et al., 2024) also replaces $\varnothing$ with another condition, but that condition must be defined by hand as a description of what to avoid. CSG needs no such definition: its negative reference is constructed from the sample's own condition, keeping the content and breaking only the alignment. Self-Swap Guidance (SSG) (Zhang et al., 2026b) swaps dissimilar hidden-state tokens inside every transformer block; CSG instead permutes the condition once per input, before it enters the network, and yields larger gains (Table 3). Because the negative reference comes from the input itself, the same rule applies unchanged to discrete tokens and VAE latents (§4.4), for which a negative condition would be hard to specify by hand; the swap recipe and hyperparameters are in Appendices B.2 and A.1.

4 EXPERIMENTS

4.1 SETUP

Data and evaluation. Training data comprises speech and singing voice from the AnyEnhance pipeline (Zhang et al., 2025) augmented with UniPASE-style degradations (Rong et al., 2026c), plus 6,857 hours of music from SingNet (Gu et al., 2025) with a degradation pipeline following SonicMaster (Melechovsky et al., 2025) with environmental noise augmentation. We evaluate on four

Table 2: Music restoration on MUSDB18, whole-utterance inference. RESTORA uses MuQ-only CSG at $\omega = 0.5$. **Bold**/underline: best/second-best per column. Degraded input and clean reference are shown for scale and excluded from ranking.

Model	FAD↓	LSD↓	SSIM↑	PQ↑	CE↑	CU↑
MUSDB18 (full song)						
Degraded input	0.159	9.862	0.679	5.679	5.394	5.424
Clean reference	–	–	–	7.323	6.863	7.155
Mel2Mel + DiffWave	0.514	18.852	0.416	5.794	5.452	5.640
TF-Conformer	0.134	12.182	0.638	6.461	6.193	6.262
SonicMaster	0.181	9.714	0.537	6.121	5.947	5.804
MaskSR	0.104	7.833	0.619	6.631	6.343	6.306
AnyEnhance (MERT)	0.112	8.371	0.624	6.783	6.372	6.505
RESTORA	**0.084**	**6.963**	**0.682**	**7.335**	**6.975**	**7.165**
MUSDB18 (instrumental)						
Degraded input	0.154	9.928	0.671	5.988	5.634	6.068
Clean reference	–	–	–	7.335	6.940	7.490
Mel2Mel + DiffWave	0.422	16.999	0.431	6.238	5.889	6.341
TF-Conformer	0.121	11.517	0.635	6.789	6.414	6.861
SonicMaster	0.145	9.767	0.542	6.541	6.266	6.586
MaskSR	0.069	7.982	0.628	6.972	6.608	7.053
AnyEnhance (MERT)	0.079	8.371	0.624	7.118	6.675	7.230
RESTORA	**0.041**	**6.890**	**0.683**	**7.519**	**7.217**	**7.638**

held-out test sets: LibriVox-GSR (Kearns, 2014) (speech, 500 utterances), CCMusic-GSR (Zhou et al., 2024) (singing, 500 utterances), and MUSDB18 (Rafii et al., 2017) full-song and instrumental (music, 500 each). A listening test with 12 participants covers six sets: real speech, singing, instrumental, and full-song recordings, plus the two simulated MUSDB18 sets (Table 4).

For speech and singing, we report DNSMOS (Reddy et al., 2022) SIG/BAK/OVRL, NISQA (Mittag et al., 2021), SpeechBERTScore (Saeki et al., 2024) (SBS), and speaker similarity (SIM). For music, we report log-spectral distance (LSD), Mel-SSIM, FAD (Kilgour et al., 2019), and Audiobox Aesthetics (Tjandra et al., 2025) PQ/CE/CU. In the guidance ablation (Table 3), we additionally compute WER via Whisper (Radford et al., 2023) to directly measure content-level hallucination. Full definitions are in Appendix A.3.

Models and baselines. RESTORA is a single 500M-parameter DiT trained in three stages: (1) flow-matching pre-training on the full data mixture for 300k steps, (2) supervised fine-tuning (SFT) on quality-filtered data for 50k steps, and (3) DiffusionNFT post-training with Headroom-Anchored RL for 1k steps (see Appendix A.1 for full architecture and training details). The deployed checkpoint (Headroom-1k) uses CSG at inference (ω=1.0 for speech/singing, ω=0.5 MuQ-only for music). We compare against eight speech systems—VoiceFixer (Liu et al., 2022a), AnyEnhance (Zhang et al., 2025), LLaSE-G1 (Kang et al., 2025), PASE (Rong et al., 2026b), UniPASE (Rong et al., 2026c), UniSE (Yan et al., 2025), SenSE (Li et al., 2025b), and RE-USE (Fu et al., 2026)—and five music systems—Mel2Mel+DiffWave (Kandpal et al., 2022), TF-Conformer (Chae et al., 2023), SonicMaster (Melechovsky et al., 2025), MaskSR (Li et al., 2024a), and AnyEnhance (MERT (Li et al., 2024b)). The last two music baselines are retrained on the same SingNet data as RESTORA for a fair comparison. Baseline descriptions are in Appendix A.2.

4.2 MAIN RESULTS

A single RESTORA model is the best system on music, the best or near-best on singing, and competitive with speech-only specialists on speech (Tables 1 and 2). On both MUSDB18 subsets, RESTORA leads all six music metrics; its FAD is 19% lower than that of the next-best system, MaskSR, on full songs (0.084 vs. 0.104) and 41% lower on instrumental tracks (0.041 vs. 0.069). It is also the only system whose Mel-SSIM does not fall below that of the degraded input, and its aesthetic scores exceed those of the clean reference; because its LSD and FAD are also the best, these aesthetic gains

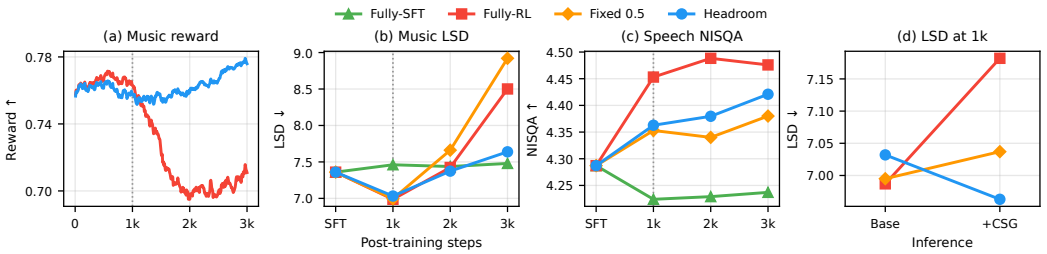

Figure 2: Headroom versus constant-gate schedules, all starting from SFT-50k; Fully-SFT, Fully-RL, and Fixed 0.5 hold the gate at a_d=0, 1, and 0.5. (a) Music training reward (moving average over ±320 steps). (b, c) Held-out music LSD and speech NISQA without guidance. (d) Music LSD at the 1k checkpoint before and after adding CSG. Dotted lines mark the deployed 1k checkpoint.

do not come from trading away fidelity. On singing, where the speech-only baselines operate outside their training domain, RESTORA leads SIG, OVRL, and NISQA (4.317 vs. 3.793 for RE-USE, the next best) and is within 0.03 of the best system on BAK, SBS, and SIM. Speech is the most competitive domain: speech-only specialists hold the best score in every column. RESTORA ranks second or third on all four perceptual metrics, including second on NISQA (4.590 vs. 4.719 for SenSE), but trails the best system, UniPASE, on SBS (0.800 vs. 0.862) and SIM (0.941 vs. 0.959); a 3 s speaker prompt raises SIM to 0.953 (Appendix Table 16).

In the listening test (Table 4), twelve listeners rated the degraded input, AnyEnhance, and RESTORA on a 1–5 scale across six test sets, with the music variant AnyEnhance (MERT) used on the four music sets. RESTORA raises the overall mean opinion score from 2.55 to 4.12 (+1.57) and scores higher than AnyEnhance on every set. Gains are largest on voice (+2.37 averaged over speech and singing), followed by instrumental (+1.25) and full-song music (+1.09).

4.3 Ablation on Headroom training stability

Figure 2 compares Headroom with three constant-gate schedules from the same SFT-50k initialization: continued SFT (a_d≡0), Fully-RL (a_d≡1), and a fixed mixture (a_d≡0.5). Under Fully-RL, the music training reward falls from 0.76 at 1k to 0.70 at 2k, while Headroom's stays between 0.75 and 0.78 (panel a). Held-out fidelity follows the reward: from 1k to 3k, Fully-RL's LSD rises by 1.51 (6.987→8.501; panel b), about 2.5× Headroom's rise (7.032→7.639), and its FAD rises by 0.056 versus 0.011. The fixed mixture does not prevent this drift; its LSD reaches 8.921 at 3k. On speech (panel c), every schedule that uses RL improves NISQA while continued SFT does not, and Headroom keeps about 70% of Fully-RL's gain at 3k.

Without guidance, Fully-RL-1k is slightly ahead of Headroom-1k (NISQA 4.453 vs. 4.363, LSD 6.987 vs. 7.032), but the two respond differently to CSG (panel d): Headroom's music LSD improves (7.032→6.963) while Fully-RL's worsens (6.987→7.182), and Headroom-1k with CSG also has the lowest FAD (0.084 vs. 0.090), at a cost of 0.05 NISQA. We therefore deploy Headroom-1k: it gives the best music fidelity at the deployed operating point and drifts far more slowly if training continues. Appendix B.1 gives the full results, including gate trajectories showing that the controller closes the music gate right after warm-up while keeping speech RL open (by 1k, the running mean of a_d is 0.35 for music and 0.96 for speech).

4.4 Ablation on Condition-Swap Guidance

Guidance comparison. As shown in Table 3, we compare CSG with CFG and five other guidance methods that need no null condition: SSG (Zhang et al., 2026b), PAG (Ahn et al., 2024), SEG (Hong, 2024), SAG (Hong et al., 2023), and autoguidance (Karras et al., 2024), all using Headroom-1k weights; Appendix B.2.1 describes each method and its settings. On speech, CSG has both the best NISQA (4.590) and the best WER (0.2244). SAG raises WER above the unguided baseline (0.2672 vs. 0.2516), indicating that blurred attention maps can harm content fidelity. On music (Appendix Table 8), CSG is best on LSD, PQ, CE, and CU; only SAG has a lower FAD.

Table 3: Guidance mechanisms compared on LibriVox speech, all using Headroom-1k weights at one strength each; the noisy input has a WER of 0.2109. **Bold**/underline: best/second-best per column.

Method	SIG↑	BAK↑	OVRL↑	NISQA↑	SBS↑	SIM↑	WER↓
Base	3.593	4.162	3.339	4.363	0.789	0.934	0.2516
CFG 1.5	3.569	4.163	3.321	4.460	**0.805**	**0.943**	0.2324
SSG-token 0.5	3.594	4.166	3.344	4.405	0.789	0.934	0.2396
PAG 0.5	**3.604**	4.165	3.350	4.369	0.796	0.935	0.2398
SEG 2.0, $\sigma = 8$	3.602	4.169	3.352	4.457	0.789	0.935	0.2490
SAG 0.75, $\sigma = 1$, $k = 9$	3.601	**4.171**	**3.353**	4.404	0.788	0.934	0.2672
Autoguidance 0.5	3.596	4.169	3.346	4.397	0.789	0.934	0.2495
CSG 1.0	3.594	4.170	**3.353**	**4.590**	0.800	0.941	**0.2244**

Table 4: Subjective MOS evaluation (12 listeners, 1–5 scale). RESTORA raises the overall mean MOS from 2.55 to 4.12 (+1.57) and exceeds the AnyEnhance baseline on all six test sets.

Category	Test set	Degraded	AnyEnhance	RESTORA
Voice	Singing (10)	2.325	4.075	**4.537**
	Speech (10)	1.975	4.275	**4.500**
Instrumental	MUSDB18-Instr. (10)	2.500	3.300	**4.220**
	Real music (10)	2.974	3.171	**3.750**
Full song	MUSDB18-Full (10)	2.525	2.775	**4.081**
	Real full-song (10)	3.013	2.974	**3.640**
Overall mean		**2.552**	3.428	**4.121**

Further ablations. The condition-stream ablation (Appendix B.2.3) confirms the domain recipe: speech benefits from swapping all three streams; music must preserve the STFT anchor and swap only MuQ. Appendix B.2.4 further ablates the swap ratio, pairing policy, and guidance window.

Transfer to other models. On AnyEnhance (Zhang et al., 2025), a discrete masked-token speech enhancement model, CSG reduces WER from 0.207 to 0.188 and raises SBS from 0.821 to 0.829 (Appendix Table 13); on DiffBIR (Lin et al., 2024), an image restoration model, it improves SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), DISTS (Ding et al., 2022), and NIQE (Mittal et al., 2013) on three benchmarks (Appendix Table 14).

5 CONCLUSION & DISCUSSION

We presented RESTORA, a 500M flow-matching model that restores speech, singing voice, full-song mixtures, and instrumental music with a single set of weights. **Headroom-Anchored RL** gates each domain between RL and a supervised anchor according to its reward headroom, slowing music fidelity degradation by roughly $2.5\times$ compared to uniform RL. **Condition-Swap Guidance** replaces the null reference with a temporally permuted copy of the sample's own conditioning, improving both quality and content accuracy without additional training. Together, they raise the mean opinion score of degraded inputs from 2.55 to 4.12 across six test sets.

Limitations. Headroom slows but does not stop music fidelity drift when post-training continues (Figure 2b). Content fidelity on speech remains the main gap: WER stays above that of the noisy input (Table 3), and speech-only systems built for low hallucination, such as UniPASE, preserve content better (Table 1). The music CSG recipe requires per-domain stream choices, and transfer is shown on only one speech model and one image model.

Future Work. Text or visual prompts, such as transcripts, lyrics, or lip videos, could supply the content evidence that degraded audio lacks; rewards that track fidelity more reliably would let RL continue after the current music rewards saturate; and CSG could be applied to other tasks whose conditions are aligned with the output in time or space, such as video restoration.

AI USE STATEMENT

We used large language models & agents in this work. They helped polish the writing and drafted parts of the manuscript, which the authors revised. They assisted retrieval and discovery, including searches for related work; the authors checked every citation in the paper. They also assisted research ideation and execution, including writing plotting and experiment-automation scripts; the authors designed the study, ran the experiments, and verified the code and results. We did not use generative AI to create synthetic datasets or to prove mathematical claims. The authors take full responsibility for the content of this paper.

ETHICS STATEMENT

Generative restoration can change content: our analysis shows that the restored speech can contain words that differ from the original. Restored audio should therefore not be treated as a faithful record of the original recording, for example in forensic or archival use where exact content matters. The listening test was conducted with 12 volunteer listeners who gave informed consent, received no compensation, and rated anonymized audio clips; no personal information was collected.

REPRODUCIBILITY STATEMENT

Appendix A.1 specifies the architecture, training data and degradation pipelines, training schedule, rewards, and all hyperparameters; Algorithm 1 gives one iteration of the Headroom controller; Appendix A.3 describes the test sets and metrics.

REFERENCES

Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention guidance. In *European Conference on Computer Vision*, 2024.

Kaito Baba, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari. The T05 system for the VoiceMOS Challenge 2024: Transfer learning from deep image classifier to naturalness MOS prediction of high-quality synthetic speech. In *IEEE Spoken Language Technology Workshop (SLT)*, pp. 818–824, 2024.

Evelina Bakhturina, Vitaly Lavrukhin, Boris Ginsburg, and Yang Zhang. Hi-Fi multi-speaker English TTS dataset. In *Interspeech*, 2021.

Yuanhao Ban, Ruochen Wang, Tianyi Zhou, Minhao Cheng, Boqing Gong, and Cho-Jui Hsieh. Understanding the impact of negative prompts: When and how do they take effect? In *European Conference on Computer Vision (ECCV)*, pp. 190–206, 2024.

Frédéric Berdoz, Luca A Lanzendörfer, Antonis Asonitis, and Roger Wattenhofer. Post-training speech enhancement language models with perceptual rewards. In *International Conference on Machine Learning (ICML)*, 2026.

Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3086–3095, 2019.

Ruizhe Cao, Sherif Abdulatif, and Bin Yang. CMGAN: Conformer-based metric GAN for speech enhancement. In *Interspeech*, pp. 936–940, 2022.

Yunkee Chae, Junghyun Koo, Sungho Lee, and Kyogu Lee. Exploiting time-frequency conformers for music audio enhancement. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 2362–2370, 2023.

Rong Chao, Wen-Huang Cheng, Moreno La Quatra, Sabato Marco Siniscalchi, Chao-Han Huck Yang, Szu-Wei Fu, and Yu Tsao. An investigation of incorporating Mamba for speech enhancement. In *IEEE Spoken Language Technology Workshop (SLT)*, 2024.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

Yifu Chen, Shengpeng Ji, Qian Chen, Tianle Liang, Yangzhuo Li, Ziqing Wang, Wen Wang, Jingyu Lu, Haoxiao Wang, Xueyi Pu, Fan Zhuo, and Zhou Zhao. WavAlign: Enhancing intelligence and expressiveness in spoken dialogue models via adaptive hybrid post-training. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2022.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. CosyVoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*, 2024.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 12606–12633, 2024.

Szu-Wei Fu, Rong Chao, Xuesong Yang, Sung-Feng Huang, Ryandhimas E Zezario, Rauf Nasretdinov, Ante Jukić, Yu Tsao, and Yu-Chiang Frank Wang. Rethinking training targets, architectures and data quality for universal speech enhancement. *arXiv preprint arXiv:2603.02641*, 2026.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 2023.

Yicheng Gu, Chaoren Wang, Junan Zhang, Xueyao Zhang, Zihao Fang, Haorui He, and Zhizheng Wu. SingNet: Towards a large-scale, diverse, and in-the-wild singing voice dataset. *arXiv preprint arXiv:2505.09325*, 2025.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, et al. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *IEEE Spoken Language Technology Workshop (SLT)*, 2024.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Susung Hong. Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of attention. In *Advances in Neural Information Processing Systems*, 2024.

Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7462–7471, 2023.

Shaohan Jiang, Junan Zhang, Yunjia Zhang, Jing Yang, Fan Fan, and Zhizheng Wu. Singverse: A diverse, real-world benchmark for singing voice enhancement. *arXiv preprint arXiv:2509.20969*, 2025.

Nikhil Kandpal, Oriol Nieto, and Zeyu Jin. Music enhancement via image translation and vocoding. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3124–3128. IEEE, 2022.

Boyi Kang, Xinfa Zhu, Zihan Zhang, Zhen Ye, Mingshuai Liu, Ziqian Wang, Yike Zhu, Guobin Ma, Jun Chen, Longshuai Xiao, et al. LLaSE-G1: Incentivizing generalization capability for LLaMA-based speech enhancement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 13292–13305, 2025.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. In *Advances in Neural Information Processing Systems*, 2024.

Jodi Kearns. Librivox: Free public domain audiobooks. *Reference Reviews*, 2014.

Kevin Kilgour, Maria Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet Audio Distance: A reference-free metric for evaluating music enhancement algorithms. In *Interspeech*, 2019.

Keisuke Kinoshita, Marc Delcroix, Takuya Yoshioka, Tomohiro Nakatani, Emanuel Habets, Reinhold Haeb-Umbach, Volker Leutnant, Armin Sehr, Walter Kellermann, Roland Maas, et al. The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013.

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.

Anurag Kumar, Andrew Perrault, and Donald S Williamson. Using RLHF to align speech enhancement approaches to mean-opinion quality scores. In *ICASSP*, pp. 1–5, 2025.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved rvqgan. *Advances in Neural Information Processing Systems*, 2024.

Jean-Marie Lemercier, Julius Richter, Simon Welker, and Timo Gerkmann. Storm: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

Chenda Li, Wei Wang, Marvin Sach, Wangyou Zhang, Kohei Saijo, Samuele Cornell, Yihui Fu, Zhaoheng Ni, Tim Fingscheidt, Shinji Watanabe, et al. Icassp 2026 urgent speech enhancement challenge. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 21919–21921. IEEE, 2026.

Haoyang Li, Nana Hou, Yuchen Hu, Jixun Yao, Sabato Marco Siniscalchi, and Eng Siong Chng. Aligning generative speech enhancement with human preferences via direct preference optimization. *arXiv preprint arXiv:2507.09929*, 2025a.

Xingchen Li, Hanke Xie, Ziqian Wang, Zihan Zhang, Longshuai Xiao, and Lei Xie. SenSE: Semantic-aware high-fidelity universal speech enhancement. *arXiv preprint arXiv:2509.24708*, 2025b.

Xu Li, Qirui Wang, and Xiaoyu Liu. MaskSR: Masked language model for full-band speech restoration. In *Interspeech*, pp. 2275–2279, 2024a.

Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, et al. MERT: Acoustic music understanding model with large-scale self-supervised training. In *International Conference on Learning Representations*, 2024b.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. DiffBIR: Toward blind image restoration with generative diffusion prior. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Fei Liu, Yang Ai, and Zhen-Hua Ling. Neural music enhancement with dual time-frequency spectral representations for prediction and discrimination. *arXiv preprint arXiv:2609.03357*, 2026.

Haohe Liu, Xubo Liu, Qiuqiang Kong, Qiao Tian, Yan Zhao, DeLiang Wang, Chuanzeng Huang, and Yuxuan Wang. VoiceFixer: A unified framework for high-fidelity speech restoration. In *Interspeech*, pp. 4232–4236, 2022a.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training flow matching models via online RL. In *Advances in Neural Information Processing Systems*, 2025a.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision (ECCV)*, 2022b.

Xiaoyu Liu, Xu Li, Joan Serrà, and Santiago Pascual. Joint semantic knowledge distillation and masked acoustic modeling for full-band speech restoration with improved intelligibility. In *ICASSP*, 2025b.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *Proceedings of the Eleventh International Conference on Learning Representations*, 2023.

Lu Ma, Hao Liang, Meiyi Qiang, Lexiang Tang, Xiaochen Ma, Zhen Hao Wong, Junbo Niu, Chengyu Shen, Runming He, Yanhao Li, Bin Cui, and Wentao Zhang. Learning what reinforcement learning can't: Interleaved online fine-tuning for hardest questions. In *Proceedings of the Fourteenth International Conference on Learning Representations*, 2026.

Jan Melechovsky, Ambuj Mehrish, Abhinaba Roy, and Dorien Herremans. SonicMaster: Towards controllable all-in-one music restoration and mastering. *arXiv preprint arXiv:2508.03448*, 2025.

Josh Meyer, David Ifeoluwa Adelani, Edresson Casanova, Alp Öktem, Daniel Whitenack, Julian Weber, Salomon Kabongo, Elizabeth Salesky, Iroro Orife, Colin Leong, Perez Ogayo, et al. BibleTTS: a large, high-fidelity, multilingual, and uniquely African speech corpus. In *Interspeech*, 2022.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. In *Proceedings of the Conference of the International Speech Communication Association*, 2021. doi: 10.21437/INTERSPEECH.2021-299.

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013.

Eloi Moliner, Maija Turunen, Filip Elvander, and Vesa Välimäki. A diffusion-based generative equalizer for music restoration. *arXiv preprint arXiv:2403.18636*, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 2022.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4172–4182. IEEE, 2023.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning (ICML)*, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. The MUSDB18 corpus for music separation. Zenodo, 2017.

Chandan KA Reddy, Vishak Gopal, Ross Cutler, Ebrahim Beyrami, Roger Cheng, Harishchandra Dubey, Sergiy Matusevych, Robert Aichner, Ashkan Aazami, Sebastian Braun, et al. The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results. In *Proceedings of the Conference of the International Speech Communication Association*, 2020.

Chandan KA Reddy, Harishchandra Dubey, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner, and Sriram Srinivasan. Icassp 2021 deep noise suppression challenge. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021.

Chandan KA Reddy, Vishak Gopal, and Ross Cutler. Dnsmos p.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022.

Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay, and Timo Gerkmann. Speech enhancement and dereverberation with diffusion-based generative models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

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*, pp. 4873–4877, 2024.

Xiaobin Rong, Jun Gao, Zheng Wang, Mansur Yesilbursa, Kamil Wojcicki, and Jing Lu. Stupase: Towards low-hallucination studio-quality generative speech enhancement. *arXiv preprint arXiv:2603.09234*, 2026a.

Xiaobin Rong, Qinwen Hu, Mansur Yesilbursa, Kamil Wojcicki, and Jing Lu. PASE: Leveraging the phonological prior of WavLM for low-hallucination generative speech enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 32826–32834, 2026b.

Xiaobin Rong, Zheng Wang, Yushi Wang, Jun Gao, and Jing Lu. UniPASE: A generative model for universal speech enhancement with high fidelity and low hallucinations. *IEEE Transactions on Audio, Speech and Language Processing*, 34:3901–3915, 2026c.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. UTMOS: UTokyo-SaruLab system for VoiceMOS challenge 2022. In *Interspeech*, 2022.

Takaaki Saeki, Soumi Maiti, Shinnosuke Takamichi, Shinji Watanabe, and Hiroshi Saruwatari. SpeechBERTScore: Reference-aware automatic evaluation of speech generation leveraging NLP evaluation metrics. In *Interspeech*, 2024.

Kohei Saijo, Wangyou Zhang, Samuele Cornell, Robin Scheibler, Chenda Li, Zhaoheng Ni, Anurag Kumar, Marvin Sach, Yihui Fu, Wei Wang, et al. Interspeech 2025 URGENT speech enhancement challenge. In *Interspeech*, pp. 858–862, 2025.

Robin Scheibler, Yusuke Fujita, Yuma Shirahata, and Tatsuya Komatsu. Universal score-based speech enhancement with high content preservation. In *Proc. Interspeech*, 2024.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Joan Serrà, Santiago Pascual, Jordi Pons, R. Oguz Araz, and Davide Scaini. Universal speech enhancement with score-based diffusion. *arXiv preprint arXiv:2206.03065*, 2022.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Yatharth Sharma. Fast and flexible audio bandwidth extension via vocos. *arXiv preprint arXiv:2603.07285*, 2026.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, et al. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*, 2026.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. Aishell-3: A multi-speaker mandarin tts corpus and the baselines. *arXiv preprint arXiv:2010.11567*, 2020.

Hubert Siuzdak. Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis. *arXiv preprint arXiv:2306.00814*, 2023.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. Meta Audiobox Aesthetics: Unified automatic quality assessment for speech, music, and sound. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2025.

Haoxu Wang, Biao Tian, Yiheng Jiang, Zexu Pan, Shengkui Zhao, Bin Ma, Daren Chen, and Xiangang Li. Flowse-grpo: Training flow matching speech enhancement via online reinforcement learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 16182–16186. IEEE, 2026a.

Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. CAM++: A fast and efficient network for speaker verification using context-aware masking. In *Proceedings of the Conference of the International Speech Communication Association*, pp. 5301–5305, 2023.

Zheng Wang, Xiaobin Rong, Hang Su, Tianyi Tan, Junnan Wu, Lichun Fan, Zhenbo Luo, Jian Luan, and Jing Lu. Reducing linguistic hallucination in lm-based speech enhancement via noise-invariant acoustic-semantic distillation. *arXiv preprint arXiv:2605.08608*, 2026b.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *European Conference on Computer Vision*, pp. 101–117, 2020.

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

Weiming Xu, Zhouxuan Chen, Zhili Tan, Shubo Lv, Runduo Han, Wenjiang Zhou, Weifeng Zhao, and Lei Xie. Mbtfnet: multi-band temporal-frequency neural network for singing voice enhancement. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8. IEEE, 2023.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit, 2019.

Haoyin Yan, Chengwei Liu, Shaofei Xue, Xiaotao Liang, and Zheng Xue. Unise: A unified framework for decoder-only autoregressive lm-based speech enhancement. *arXiv preprint arXiv:2510.20441*, 2025.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *The Thirteenth International Conference on Learning Representations*, 2025.

Marcely Zanon Boito, Vivek Iyer, Nikolaos Lagos, Laurent Besacier, and Ioan Calapodescu. mHuBERT-147: A compact multilingual HuBERT model. In *Proceedings of the Conference of the International Speech Communication Association*, pp. 3939–3943, 2024.

Junan Zhang, Jing Yang, Zihao Fang, Yuancheng Wang, Zehua Zhang, Zhuo Wang, Fan Fan, and Zhizheng Wu. Anyenhance: A unified generative model with prompt-guidance and self-critic for voice enhancement. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

Junan Zhang, Xueyao Zhang, Jing Yang, Yuancheng Wang, Fan Fan, and Zhizheng Wu. Multi-metric preference alignment for generative speech restoration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 34737–34745, 2026a.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Wangyou Zhang, Robin Scheibler, Kohei Saijo, Samuele Cornell, Chenda Li, Zhaoheng Ni, Jan Pirklbauer, Marvin Sach, Shinji Watanabe, Tim Fingscheidt, and Yanmin Qian. Urgent challenge: Universality, robustness, and generalizability for speech enhancement. In *Proceedings of the Conference of the International Speech Communication Association*, 2024. doi: 10.21437/Interspeech.2024-1239.

Weijia Zhang, Yuehao Liu, Shanyan Guan, Wu Ran, Yanhao Ge, Wei Li, and Chao Ma. Guiding a diffusion model by swapping its tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14263–14272, 2026b.

Shengkui Zhao, Bin Ma, Karn N Watcharasupat, and Woon-Seng Gan. Frcrn: Boosting feature representation using frequency recurrence for monaural speech enhancement. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025a.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. DiffusionNFT: Online diffusion reinforcement with forward process. *arXiv preprint arXiv:2509.16117*, 2025b.

Monan Zhou, Shenyang Xu, Zhaorui Liu, Zhaowen Wang, Feng Yu, Wei Li, and Baoqiang Han. Ccmusic: An open and diverse database for chinese and general music information retrieval research, 2024.

Haina Zhu, Yizhi Zhou, Hangting Chen, Jianwei Yu, Ziyang Ma, Rongzhi Gu, Yi Luo, Wei Tan, and Xie Chen. Muq: Self-supervised music representation learning with mel residual vector quantization. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

A Implementation and Experiment Details

A.1 Implementation details

Representations. All audio is processed at 48 kHz on a shared 50 Hz frame grid. The target x_0 is a 128-bin log-Mel spectrogram of the clean signal at 24 kHz (window and FFT size 1920, hop 480, 0–12 kHz), normalized by a global mean and variance. The degraded input provides three frame-aligned condition streams: a power-law compressed STFT magnitude (FFT size 2048, hop 960, exponent 0.3; 1025 bins), layer-18 features of the Qwen3-ASR-1.7B audio encoder (Shi et al., 2026) (1024 dimensions at 12.5 Hz, computed from 16 kHz audio), and layer-11 features of MuQ-large (Zhu et al., 2025) (1024 dimensions at 25 Hz, computed from 24 kHz audio); both semantic streams are linearly interpolated to 50 Hz. A frozen Vocos vocoder (Siuzdak, 2023) converts the restored Mel spectrogram to 24 kHz audio, on which all objective metrics are computed; for the listening test, a Vocos-based bandwidth extension model (Sharma, 2026) further upsamples it to 48 kHz.

Denoised semantic encoders. The Qwen3-ASR and MuQ encoders are replaced by students trained with denoising representation distillation (Rong et al., 2026b;c). A frozen teacher encodes the clean waveform, and a student initialized from the same weights is fully fine-tuned on the paired degraded waveform, produced by the same degradation pipeline as the restoration data, to minimize

$$\mathcal{L}_{\mathrm{DRD}} = \frac{1}{N}\sum_{n=1}^{N}\Big[\tfrac{1}{D}\big\|h_n^{\mathrm{s}} - h_n^{\mathrm{t}}\big\|_2^2 + 1 - \cos\big(h_n^{\mathrm{s}}, h_n^{\mathrm{t}}\big)\Big], \tag{9}$$

where h_n^{s} and h_n^{t} are the D-dimensional features of frame n at the tapped layer, computed by the student from the degraded waveform and by the teacher from the clean one. Each student is trained for 100k steps on four 32 GB GPUs with Adam (β_1=0.9, β_2=0.95, no weight decay), a peak learning rate of 5×10^{-5} after 1k warm-up steps, gradient clipping at 1.0, and an average batch size of 16 (Qwen) or 8 (MuQ). Both students are frozen afterwards.

Backbone. The backbone stacks 25 LLaMA-style transformer blocks (hidden size 1024, 16 heads, feed-forward width 4096) with bidirectional attention. Each sample forms a sequence of a task token (voice, full-song, or instrumental restoration), an optional clean prompt, and the target frames; at each target frame, x_t and the three condition streams are projected, concatenated, and fused by a linear layer. The timestep enters every block through adaptive RMSNorm. A small convolutional prenet denoises the STFT stream before fusion and is trained with an auxiliary MSE loss toward the clean STFT. The model has 470.5M parameters (referred to as 500M). The 200M model used in the model-scale and EMA ablations has 12 blocks (236M parameters) and is trained with a squared velocity loss, larger batches, and no condition dropout.

Data. Pre-training uses 1,021 h of fullband speech from HiFi-TTS (Bakhturina et al., 2021), BibleTTS (Meyer et al., 2022), HQ-TTS, VCTK (Yamagishi et al., 2019), AISHELL-3 (Shi et al., 2020), and EARS (Richter et al., 2024); a six-language subset of Emilia (He et al., 2024); singing from the studio singing collection of AnyEnhance (Zhang et al., 2025) (607 h) and from SingNet (Gu et al., 2025) vocals (2,135 h); and 6,857 h of SingNet music whose vocal and accompaniment stems are obtained by source separation, from which we build instrumental and full-song samples. Speech and singing are degraded with the AnyEnhance simulation pipeline (Zhang et al., 2025) (noise, reverberation, clipping, bandwidth limitation, and equalization), extended with packet loss and wind noise as in UniPASE (Rong et al., 2026c); music degradations follow SonicMaster (Melechovsky et al., 2025) (dynamics, equalization, reverberation, bandwidth limitation, noise, and clipping). SFT keeps the top-ranked clips of each corpus: 500 h of speech ranked by UTMOSv2 (Baba et al., 2024), 750 h of singing ranked by DNSMOS OVRL, and 2,000 h of music ranked by Audiobox Aesthetics (Tjandra et al., 2025) production quality. Post-training uses speech and music prompts in equal proportion.

Training objective. Pre-training and SFT minimize the L1 form of Eq. 1 plus the prenet loss, with timesteps biased toward high noise levels. To learn the null condition used by CFG, 20% of the samples zero each condition stream independently with probability 0.9. With probability 0.2, the clean first part of a clip serves as an in-context prompt for the rest.

Training stages. Pre-training (300k steps) and SFT (50k steps) run on four 80 GB GPUs with Adam (β_1=0.9, β_2=0.95, weight decay 0.1) and about 96 s of audio per GPU per batch; pre-training uses an inverse-square-root schedule with a peak learning rate of 10^{-4} after 32k warm-up steps, and SFT a constant learning rate of 10^{-5} after 1k warm-up steps. Post-training runs DiffusionNFT with Headroom from SFT-50k on a single 80 GB GPU for up to 3k steps (deployed at 1k), with a learning rate of 2×10^{-6} after 50 warm-up steps and cosine decay.

Post-training and Headroom. Each iteration generates K=16 rollouts for each of 8 prompts, balanced between speech and music, with a 10-step SDE Euler sampler (noise scale 0.2, applied for $t < 0.5$), and uses them for 8 optimizer steps. The NFT loss uses β=0.1, the KL term penalizes the velocity difference to the frozen SFT-50k model (β_{KL}=10^{-3}), and the SFT anchor is the flow-matching loss on clean targets. Headroom uses q=0.10, β_f=0.90, β_s=0.99, τ=0.2, $a_{\min}$=0, and λ=0.5; the normalized slope is clipped to $[-4, 4]$, and the gate is fixed to a_d=1 for the first 20 iterations.

Rewards. Speech rewards combine DNSMOS OVRL (weight 0.2), CAM++ (Wang et al., 2023) speaker similarity (0.4), and SpeechBERTScore on layer-9 features of mHuBERT-147 (Saeki et al., 2024; Zanon Boito et al., 2024) (0.4), the last two computed against the clean reference. Music rewards combine Mel-SSIM (Wang et al., 2004) (0.55), CLAP (Wu et al., 2023) similarity to the clean reference (0.40), and the mean Audiobox Aesthetics score (0.05), computed after RMS-normalizing the audio. Each term is standardized within its rollout group, and the combined advantage is clipped at 1.5.

Inference and CSG. Sampling uses a 20-step Euler solver; objective test clips (at most 15 s) are restored in a single pass, and longer listening-test recordings in 10 s chunks with 0.5 s crossfades. CSG permutes the condition in blocks of four frames (80 ms), swapping the most dissimilar block pairs, measured by the cosine similarity of Qwen features (MuQ for music), until half of the blocks have moved. It swaps all three streams for speech and singing ($\omega = 1.0$) and only MuQ for music ($\omega = 0.5$), and runs the negative branch only for $t \in [0.1, 0.7]$, which costs 32 forward evaluations per 20-step solve versus 40 for CFG.

Bandwidth extension for listening. The bandwidth extension model adopts the Vocos-based architecture (Sharma, 2026) (14.2M parameters) but takes the Mel spectrogram of the restoration model as input and outputs 48 kHz audio. It is trained as a GAN on band-limited inputs, first on 10.6k h of speech, singing, and music and then on 173 h of fullband recordings.

A.2 Baseline details

Speech. VoiceFixer (Liu et al., 2022a)[1] is a unified speech restoration framework; AnyEnhance (Zhang et al., 2025) is a prompt-guided masked generative model for speech and singing; LLaSE-G1 (Kang et al., 2025)[2] is a LLaMA-based enhancement model; PASE (Rong et al., 2026b)[3] and UniPASE (Rong et al., 2026c)[4] are generative models designed for low hallucination, building on the phonological prior of WavLM and denoising representation distillation (DRD); UniSE (Yan et al., 2025)[5] is a decoder-only autoregressive language model; SenSE (Li et al., 2025b)[6] generates semantic tokens that condition a flow-matching Mel generator; and RE-USE (Fu et al., 2026)[7] is a universal speech enhancement model based on SEMamba (Chao et al., 2024). All speech baselines except AnyEnhance are speech-only. **Music.** Mel2Mel+DiffWave (Kandpal et al., 2022)[8] is a cascaded system that first enhances the Mel spectrogram and then synthesizes the waveform with a DiffWave vocoder (Kong et al., 2021); TF-Conformer (Chae et al., 2023)[9] is a music enhancement model based on the CMGAN architecture (Cao et al., 2022); SonicMaster (Melechovsky et al., 2025)[10] is a rectified-flow model built on MM-DiT blocks (Esser et al., 2024) that supports both blind and text-guided restoration; and MaskSR (Li et al., 2024a) and AnyEnhance aligned with MERT (Li et al., 2024b) are masked generative models that predict DAC audio tokens (Kumar et al., 2024). We train MaskSR and AnyEnhance (MERT) on our music data, building on the open-source AnyEnhance implementation (Zhang et al., 2025); all other baselines are run on our test sets with their official checkpoints and inference code. For AnyEnhance, we use the full model of the original paper, provided by its authors, because the public repository releases only a smaller challenge-baseline version. The released RE-USE checkpoint differs from the one reported in its paper.

[1] https://github.com/haoheliu/voicefixer
[2] https://github.com/ASLP-lab/LLaSE-G1
[3] https://github.com/cisco-open/pase
[4] https://github.com/Xiaobin-Rong/unipase
[5] https://github.com/alibaba/unified-audio
[6] https://github.com/ASLP-lab/SenSE
[7] https://huggingface.co/nvidia/RE-USE
[8] https://github.com/nkandpa2/music_enhancement
[9] https://github.com/yoongi43/music_audio_enhancement_conformer
[10] https://github.com/AMAAI-Lab/SonicMaster

A.3 Evaluation details

Test sets. We use four simulated test sets for automatic evaluation. **LibriVox-GSR** (speech, 500 utterances) and **CCMusic-GSR** (singing, 500 utterances) are general speech/singing restoration benchmarks built by applying compound degradations—additive noise, reverberation, clipping, and bandwidth limitation—to clean sources from LibriVox (Kearns, 2014) and CCMusic acapella (Zhou et al., 2024), following the simulation pipeline of AnyEnhance (Zhang et al., 2025). **MUSDB18** full-song and instrumental (500 clips each) use music from the MUSDB18 test split with the SonicMaster degradation pipeline (Melechovsky et al., 2025). For the subjective listening test (Table 4), 12 participants rate six sets on a 1–5 MOS scale: four sets of real recordings (speech, singing, instrumental music, and full songs) and two simulated MUSDB18 sets (instrumental and full song).

Automatic metrics. Metrics fall into two families. **Perceptual quality** (reference-free): DNS-MOS (Reddy et al., 2022) SIG/BAK/OVRL (1–5, evaluated at 16 kHz); NISQA (Mittag et al., 2021) (1–5, evaluated at 48 kHz); FAD (Fréchet Audio Distance, lower is better); Audiobox Aesthetics (Tjandra et al., 2025) PQ/CE/CU. **Content fidelity** (reference-based): WER via faster-whisper large-v3, computed over the whole test set (total errors divided by total reference words); Speech-BERTScore (SBS) (Saeki et al., 2024) using HuBERT-base; speaker similarity (SIM) via WavLM-based cosine similarity; log-spectral distance (LSD, lower is better); Mel-SSIM. For music, both families should be read together: PQ rising while LSD or FAD worsens typically indicates stronger "beautification" rather than more faithful restoration.

B Extended Experiment Results

B.1 Headroom-Anchored RL

This section collects the Headroom results summarized in §4.3: the controller's state during training, the response of each schedule to CSG, a speech-only RL baseline, and an ablation of the controller's EMA time constants.

B.1.1 Controller dynamics

Figure 3 traces the controller during the Headroom run of Figure 2. After the 20-iteration warm-up, the music gap widens and the music gate closes: over steps 160–1k it averages 0.23 and is below 0.5 in 78% of iterations, so music is trained mostly through the SFT anchor. The speech gap keeps narrowing and the speech gate stays open (mean 0.95). By the deployed 1k checkpoint, the running mean of a_d is 0.35 for music and 0.96 for speech. After 1k the pattern reverses: the speech gate falls to 0.71 on average over 1k–2k, while the music gap starts to narrow and the music gate reopens, averaging 0.50 over 1k–2k and 0.86 over 2k–3k. Headroom's music LSD keeps rising over this interval (Figure 2b), so the controller slows the drift relative to Fully-RL but does not stop it; this is why we deploy the 1k checkpoint.

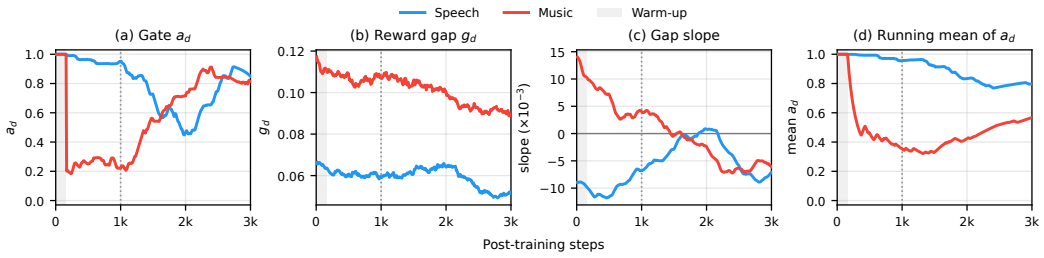

Figure 3: Headroom controller during the run of Figure 2. (a) Gate a_d; (b) reward gap g_d (Eq. 4); (c) gap slope (Eq. 5), which closes the gate when positive; (d) running mean of a_d, the average weight on the RL term so far. Curves in (a)–(c) are moving averages over $\pm$320 steps; shading marks the 20-iteration warm-up, during which a_d=1. Dotted lines mark the deployed 1k checkpoint.

B.1.2 ADAPTIVE GATE VS. FIXED MIXTURE

Table 5 compares the adaptive Headroom gate against a fixed $a_d \equiv 0.5$ SFT–RL mixture at the deployment setting; the unguided trajectories are in Figure 2(b, c). Both start from identical initialization, data, reward, and budget. At 1k the two are comparable, with the fixed mixture slightly ahead on speech (NISQA 4.611 vs. 4.590, WER 0.2200 vs. 0.2244); by 3k its music LSD reaches 8.819, while Headroom's is 7.698.

Table 5: Adaptive Headroom vs. a fixed $a_d \equiv 0.5$ SFT–RL mixture with CSG ($\omega = 1.0$ for speech, $\omega = 0.5$ for music). Same initialization, data, reward, and training budget; only the mixing rule differs. **Bold**: better of the two schedules at each checkpoint.

Training	NISQA↑	WER↓	LSD↓	FAD↓	PQ↑
Headroom-1k	4.590	0.2244	**6.963**	**0.084**	7.335
Fixed-0.5-1k	**4.611**	**0.2200**	7.037	0.088	**7.436**
Headroom-2k	**4.608**	0.2246	**7.262**	0.101	**7.297**
Fixed-0.5-2k	4.603	**0.2217**	7.587	**0.097**	7.207
Headroom-3k	4.619	**0.2205**	**7.698**	**0.097**	**7.319**
Fixed-0.5-3k	**4.635**	0.2274	8.819	0.118	7.251

B.1.3 INTERACTION WITH CSG

Table 6 crosses two training schedules with two inference modes; its last three rows compare the gain from CSG under each schedule. On music, CSG improves Headroom's LSD by 0.069 but worsens Fully-RL's by 0.195, a difference of 0.264; the gains in NISQA, SSIM, PQ, and CE are also larger under Headroom. The same trend holds against the SFT-50k starting point: CSG raises NISQA by 0.289 there, by 0.227 on Headroom-1k, and by 0.189 on Fully-RL-1k. CSG thus has less to add the more RL updates a model has received, and Headroom, which gates RL and keeps the SFT anchor, leaves more of this room for guidance.

Table 6: Training × inference 2×2 factorial. Both arms start from SFT-50k with the same 1k-step budget; speech uses CSG $\omega = 1.0$, music uses $\omega = 0.5$. **Bold**: best of the four runs per column. The last three rows give the gain from adding CSG under each schedule, signed so that positive means improvement, and the difference between the two gains.

Training	Inference	NISQA↑	WER↓	LSD↓	SSIM↑	FAD↓	PQ↑	CE↑
Fully-RL-1k	base	4.453	0.2385	6.987	0.6899	0.098	7.356	6.995
Fully-RL-1k	+CSG	**4.642**	**0.2244**	7.182	**0.6904**	0.090	**7.480**	**7.063**
Headroom-1k	base	4.363	0.2516	7.032	0.6794	0.090	7.172	6.873
Headroom-1k	+CSG	4.590	**0.2244**	**6.963**	0.6821	**0.084**	7.335	6.975
CSG gain, Headroom		+0.227	+0.0272	+0.069	+0.0027	+0.006	+0.163	+0.102
CSG gain, Fully-RL		+0.189	+0.0141	−0.195	+0.0005	+0.008	+0.124	+0.068
Difference		+0.038	+0.0131	+0.264	+0.0022	−0.002	+0.039	+0.034

B.1.4 SPEECH-ONLY RL

An alternative to per-domain gating is to optimize speech alone. Starting from the same SFT-50k checkpoint, we disable music rollouts and the SFT anchor and train with Fully-RL on speech for 10k steps, using a six-term speech reward (DNSMOS SIG/BAK/OVRL, NISQA, SBS, and SIM with weights 0.05/0.05/0.10/0.20/0.35/0.25); we select the 7k checkpoint, where OVRL plateaus. With CSG, this model reaches LibriVox OVRL 3.415 and NISQA 4.896, versus 3.353 and 4.590 for Headroom-1k. Its music fidelity, however, degrades because the shared weights drift without any music signal: unguided on MUSDB18, LSD rises from 7.032 to 7.641 and SSIM falls from 0.679 to 0.668 relative to Headroom-1k, while PQ (7.317 vs. 7.172) and FAD (0.072 vs. 0.090) improve—the pattern of aesthetic gains without fidelity described in Appendix A.3. A model that serves all domains with one set of weights therefore cannot be tuned for speech alone without giving up music fidelity.

B.1.5 EMA TIME CONSTANTS

Table 7 varies the two EMA decay rates of the controller. This ablation uses the 200M model with one training seed per setting; all other settings, including τ=0.2, q=0.10, and $a_{\min}$=0, match the 500M runs. A faster fast EMA (β_f=0.80) protects music later in training (MUSDB18 LSD at 2k: 8.075 vs. 8.722) but overreacts to short-term fluctuations in the speech gap: the speech gate drops to 0.67–0.71 and LibriVox NISQA is lower at every checkpoint (4.498 vs. 4.580 at 3k). A slower slow EMA (β_s=0.995) keeps the speech gate near 1 throughout and otherwise stays close to the default; with a single seed, these differences cannot be separated from noise. The default is therefore the most balanced shared setting, and the trade-off seen with β_f=0.80 suggests that per-domain time constants could help. We did not sweep τ, q, or $a_{\min}$.

Table 7: Headroom EMA time-constant ablation on the 200M model (one training seed per setting; unguided). The default uses β_f=0.90 and β_s=0.99. Mean gate: average a_d over the preceding 1k steps. **Bold**: best within each checkpoint.

		LibriVox-GSR		MUSDB18			Mean gate	
Step	**Setting**	**OVRL↑**	**NISQA↑**	**LSD↓**	**SSIM↑**	**FAD↓**	**Speech**	**Music**
1k	Default	**3.350**	4.532	7.150	**0.678**	0.099	0.969	0.529
	β_f=0.80	3.339	4.490	7.309	0.674	0.107	0.708	0.543
	β_s=0.995	3.345	**4.539**	**6.981**	**0.678**	**0.096**	0.968	0.494
2k	Default	**3.377**	**4.581**	8.722	0.651	**0.107**	0.851	0.840
	β_f=0.80	3.344	4.471	**8.075**	**0.663**	0.118	0.675	0.694
	β_s=0.995	3.364	4.576	8.610	0.657	0.108	0.965	0.856
3k	Default	**3.392**	4.580	8.017	0.663	0.113	0.851	0.349
	β_f=0.80	3.361	4.498	**7.746**	**0.667**	**0.106**	0.670	0.486
	β_s=0.995	3.387	**4.639**	7.954	0.665	0.110	0.980	0.418

B.2 CONDITION-SWAP GUIDANCE

This section collects the CSG results summarized in §4.4: details of the guidance comparison, the guidance-strength sweep, and ablations of the swapped streams and the swap settings. Transfer to other models follows in Appendix B.3.

B.2.1 GUIDANCE COMPARISON DETAILS

All methods in Table 3 use the same Headroom-1k model and 20-step solver, and every perturbation is restricted to the target frames, leaving the task and prompt tokens intact. CFG zeroes the same condition streams that CSG permutes; its strength 1.5 is the conventional scale $1 + \omega$, i.e., $\omega = 0.5$ in Eq. 2. PAG perturbs target-to-target attention; SEG smooths the query on the 1D packed sequence; SAG recomputes $\mathrm{softmax}(QK^\top)$ at an intermediate layer with 1D temporal blur; and autoguidance uses the PT-300k checkpoint as the weaker model. PAG, SEG, and SAG act on the middle block (layer 12 of 25). SSG swaps hidden-state tokens along the sequence; we pair tokens by mutual selection, because exact greedy pairing would have to be recomputed in every block at every solver step (500 times per file for our 25-block model with 20 steps), whereas CSG computes its pairing once per input. Starting from the hyperparameters recommended in each original paper, we tried about three settings per method and report the best-performing one in Table 3; SAG uses its published reference configuration (scale 0.75, kernel 9, $\sigma = 1$) directly. The NISQA of CSG (4.590) is higher than that of all five condition-free methods at every setting we tried; the strongest is SEG at $s = 2.0$, $\sigma = 8$ (4.457). At its best-NISQA strength ($s = 2.0$), autoguidance raises WER to 0.2751, above the unguided 0.2516, so its perceptual gain comes at the cost of content accuracy. Table 8 repeats the comparison on MUSDB18 full songs.

Table 8: Guidance mechanisms compared on MUSDB18 full songs. Same setup as Table 3. **Bold**/underline: best/second-best per column. CSG attains the best LSD–aesthetics trade-off. SSG-token wins SSIM; SAG wins FAD under the 24 kHz corpus-level setup used here but its native-setup FAD is 0.557 vs. 0.247 for CSG.

Method	FAD↓	LSD↓	SSIM↑	PQ↑	CE↑	CU↑
Base	0.090	7.032	0.679	7.172	6.873	7.019
CFG 1.5	0.086	7.101	0.677	7.284	6.933	7.112
SSG-token 0.5	0.087	7.031	**0.689**	7.292	6.910	7.126
PAG 0.5	0.094	6.964	0.683	7.098	6.785	6.957
SEG 0.5, $\sigma = 8$	0.086	7.052	0.678	7.272	6.946	7.118
SAG 0.75, $\sigma = 1$, $k = 9$	**0.062**	7.085	0.678	7.172	6.851	7.009
Autoguidance 0.5	0.084	7.092	0.676	7.267	6.933	7.097
CSG 0.5	0.084	**6.963**	0.682	**7.335**	**6.975**	**7.165**

B.2.2 Guidance-strength sweep

Table 9 and Figure 4 show the full ω sweep. On speech, NISQA gains flatten beyond ω=1.0, while WER rebounds at $\omega \geq 1.5$ for the Headroom and SFT checkpoints. On music, PQ rises monotonically with ω, whereas LSD improves only at ω=0.5 (for SFT-50k and Headroom-1k) and worsens steadily beyond it; ω=0.5 is therefore the only strength that improves both.

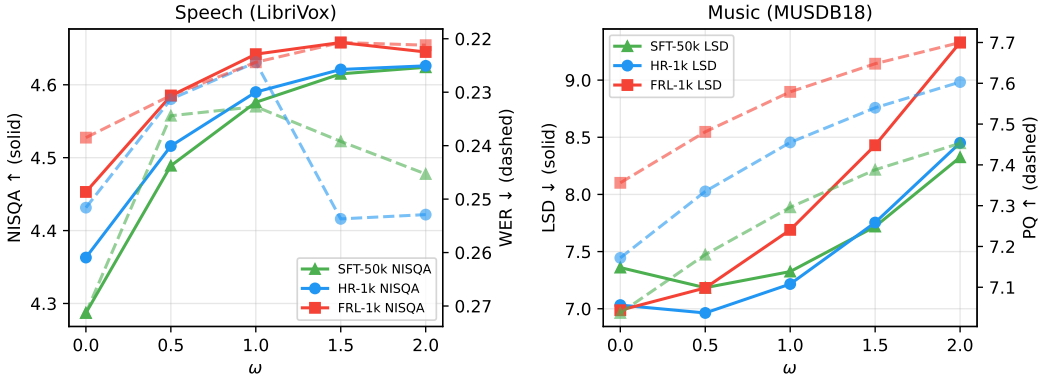

Figure 4: Guidance strength (ω) sweep. Left: speech NISQA (solid) and WER (dashed, inverted). Right: music LSD (solid) and PQ (dashed).

Table 9: Full ω sweep on LibriVox (speech) and MUSDB18 (music). Three training checkpoints $\times$ five guidance strengths. Speech CSG uses all-stream swap; music CSG uses MuQ-only. $\omega = 0$ is the unguided baseline.

	NISQA↑ (speech)			**WER↓ (speech)**		
ω	SFT-50k	HR-1k	FRL-1k	SFT-50k	HR-1k	FRL-1k
0	4.287	4.363	4.453	0.2713	0.2516	0.2385
0.5	4.489	4.516	4.585	0.2344	0.2313	0.2306
1.0	4.576	4.590	4.642	**0.2327**	**0.2244**	0.2244
1.5	4.615	4.621	**4.658**	0.2392	0.2537	**0.2207**
2.0	**4.624**	**4.626**	4.645	0.2453	0.2529	0.2212
	LSD↓ (music)			**PQ↑ (music)**		
ω	SFT-50k	HR-1k	FRL-1k	SFT-50k	HR-1k	FRL-1k
0	7.360	7.032	**6.987**	7.037	7.172	7.356
0.5	**7.184**	**6.963**	7.182	7.180	7.335	7.480
1.0	7.325	7.216	7.690	7.296	7.455	7.579
1.5	7.719	7.754	8.430	7.388	7.540	7.648
2.0	8.325	8.451	9.329	**7.453**	**7.603**	**7.700**

B.2.3 CONDITION-STREAM ABLATION

Tables 10 and 11 ablate which conditioning streams to swap. On speech, swapping all three streams gives the best joint NISQA–WER point. On music, swapping MuQ alone or STFT alone gives similar LSD (7.216 and 7.214), swapping both raises it sharply (8.394), and only swapping MuQ also raises the aesthetic scores; we therefore swap MuQ alone.

Table 10: Condition-stream ablation on LibriVox speech ($\omega = 1.0$). **Bold**: best per column.

Swapped conditions	**SIG↑**	**BAK↑**	**OVRL↑**	**NISQA↑**	**SBS↑**	**SIM↑**	**WER↓**
None	3.593	4.162	3.339	4.363	0.789	0.934	0.2516
Qwen	3.593	4.165	3.341	4.392	0.794	0.936	0.2411
MuQ	3.596	4.164	3.343	4.408	0.791	0.935	0.2396
STFT	3.589	4.158	3.334	4.432	0.794	0.938	0.2307
Qwen + MuQ	**3.601**	**4.171**	**3.355**	4.466	0.795	0.938	0.2407
MuQ + STFT	3.588	4.164	3.342	4.529	0.798	0.939	0.2404
Qwen + MuQ + STFT	3.594	4.170	3.353	**4.590**	**0.800**	**0.941**	**0.2244**

Table 11: Condition-stream ablation on MUSDB18 polyphonic music. **Bold**: best per column. All rows fix $\omega = 1.0$ to isolate stream choice; the deployed music setting is MuQ-only at $\omega = 0.5$.

Swapped conditions	**FAD↓**	**LSD↓**	**SSIM↑**	**PQ↑**	**CE↑**	**CU↑**
None	0.090	**7.032**	0.679	7.172	6.873	7.019
MuQ	**0.087**	7.216	0.678	7.455	7.039	7.261
STFT	0.097	7.214	**0.690**	7.141	6.811	6.937
MuQ + STFT	0.129	8.394	0.675	7.578	7.068	7.315
Qwen + MuQ + STFT	0.143	8.670	0.668	**7.596**	**7.080**	**7.327**

B.2.4 SWAP SETTINGS

Table 12 varies one CSG setting at a time from the deployed configuration. Larger swap ratios help: swapping every chunk (ratio 1.0) gives the best WER and NISQA, slightly better than the ratio of 0.5 used for all reported results. The guidance window must start early: guiding only in the second half ($[0.5, 1.0]$) is worse than no guidance on WER, the deployed window $[0.1, 0.7]$ gives the best WER, and $[0.0, 0.5]$ gives the best NISQA. Similar pairing is clearly worse than dissimilar pairing. Random pairing comes close in this run (WER 0.2307 vs. 0.2244), but its WER varies across sampling seeds (0.2307, 0.2531, and 0.2313), whereas dissimilar pairing is deterministic.

Table 12: Swap-setting ablation on LibriVox (500M, HR-1k, $\omega = 1.0$). Each row changes one setting from the deployed configuration (policy=dissimilar, ratio=0.5, window=$[0.1, 0.7]$). **Bold**: best in each group.

Axis	Setting	WER↓	ΔNISQA	NISQA↑
—	base (no guidance)	0.2516	—	4.363
—	deployed	0.2244	+0.227	4.590
policy	random	0.2307	+0.226	4.588
policy	similar	0.2583	+0.192	4.555
ratio	0.25	0.2410	+0.153	4.516
ratio	0.75	0.2389	+0.256	4.618
ratio	**1.0**	**0.2215**	**+0.258**	**4.621**
window	[0.0, 0.5]	**0.2374**	**+0.316**	**4.679**
window	[0.0, 1.0]	0.2435	+0.302	4.665
window	[0.5, 1.0]	0.2583	+0.030	4.393

B.3 TRANSFERABILITY

We apply CSG, without retraining, to two other restoration models: a discrete-token speech enhancement model and a latent-diffusion image restoration model.

B.3.1 DISCRETE-TOKEN SPEECH ENHANCEMENT

Table 13 shows the full results of applying CSG to AnyEnhance (Zhang et al., 2025), a discrete masked-token speech enhancement model. CSG and CFG both improve content metrics (WER, SBS) while slightly reducing background quality (BAK, NISQA). Combining the two (ω=0.5 each) yields the best OVRL and SIM.

Table 13: CSG on a discrete masked-token audio model (LibriVox, Whisper-large WER). **Bold**: best per column. WER uses a different ASR pipeline from Table 1 and is not cross-comparable.

Configuration	WER↓	SIG↑	BAK↑	OVRL↑	NISQA↑	SBS↑	SIM↑
Baseline	0.207	3.547	**4.142**	3.309	**4.345**	0.821	0.957
+ CSG (swap, $\omega = 1.0$)	0.188	**3.561**	4.113	3.310	4.224	**0.829**	0.960
+ CFG (drop, $\omega = 1.0$)	**0.187**	3.553	4.109	3.301	4.171	**0.829**	0.960
+ CSG +CFG (0.5+0.5)	0.197	3.560	4.120	**3.313**	4.180	**0.829**	**0.961**

B.3.2 IMAGE RESTORATION

To test whether CSG generalizes beyond audio, we apply it to DiffBIR (Lin et al., 2024), a blind image restoration model based on a ControlNet-conditioned latent diffusion architecture. We permute the VAE-encoded condition along spatial blocks and apply guidance within diffusion steps $[0.1, 0.7]$, mirroring the temporal-chunk strategy used for audio. Table 14 shows results on three super-resolution benchmarks, RealSR (Cai et al., 2019), DRealSR (Wei et al., 2020), and DIV2K (Agustsson & Timofte, 2017), measured by PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), DISTS (Ding et al., 2022), and NIQE (Mittal et al., 2013). CSG consistently improves SSIM, LPIPS, DISTS, and NIQE across all three datasets; only DRealSR-Val shows a small PSNR drop (23.410→23.220). The same pattern observed in audio—steering away from a structurally misaligned condition improves perceptual fidelity—transfers directly to the image domain without any architecture-specific modification.

Table 14: CSG transfer to DiffBIR (image restoration). **Bold**: better of Baseline / CSG per metric.

Method	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	NIQE↓
RealSR-Val					
DiffBIR	21.489	0.5840	0.3681	0.2431	6.317
+ CSG	**21.490**	**0.6104**	**0.3398**	**0.2324**	**6.155**
DRealSR-Val					
DiffBIR	**23.410**	0.5601	0.4761	0.2781	7.863
+ CSG	23.220	**0.5832**	**0.4455**	**0.2682**	**7.594**
DIV2K-Val					
DiffBIR	20.822	0.4960	0.3411	0.2136	5.680
+ CSG	**21.015**	**0.5197**	**0.3275**	**0.2083**	**5.351**

B.4 Model scale: 200M vs. 500M

Table 15 compares 200M and 500M at matched training steps without guidance. The larger model consistently improves music fidelity (LSD, SSIM) while matching speech perceptual quality.

Table 15: 200M vs. 500M at matched training steps (unguided). 500M improves music fidelity (LSD, SSIM) at every checkpoint while matching speech perceptual quality; 200M's higher NISQA comes at the cost of music LSD instability.

Scale	Arm	NISQA↑	SBS↑	LSD↓	SSIM↑
200M	Headroom-1k	4.532	0.790	7.150	0.678
500M	Headroom-1k	4.363	0.789	**7.032**	**0.679**
200M	Headroom-2k	**4.581**	0.788	8.722	0.651
500M	Headroom-2k	4.380	**0.791**	**7.374**	**0.687**
200M	Headroom-3k	**4.580**	0.788	8.017	0.663
500M	Headroom-3k	4.421	**0.794**	**7.639**	**0.694**
200M	Fully-RL-1k	4.496	0.792	9.104	0.663
500M	Fully-RL-1k	4.453	**0.798**	**6.987**	**0.690**
200M	Fully-RL-2k	**4.556**	0.795	8.899	0.686
500M	Fully-RL-2k	4.488	**0.802**	**7.424**	**0.692**
200M	Fully-RL-3k	**4.647**	0.798	8.895	0.670
500M	Fully-RL-3k	4.476	**0.805**	**8.501**	**0.678**

B.5 Prompt-conditioned enhancement

Restora can take an optional clean reference clip as an in-context prompt, placed before the target frames. We test it with the deployed Headroom-1k checkpoint and CSG settings. On LibriVox (Table 16), a 3 s clip of the same speaker raises NISQA by a further 0.155 over CSG alone (4.589 to 4.744) and mainly improves speaker identity, with SIM rising from 0.941 to 0.953. On MUSDB18 full songs (Table 17), the prompt is the first 3 s of the clean version of the same clip, so this setting is an upper bound; it lowers FAD from 0.198 to 0.066 and raises PQ, CE, and CU by 0.33, 0.78, and 0.39, at a small LSD cost (6.958 to 7.147). The FAD and aesthetics values in Table 17 come from a separate evaluation run, so we compare them only within the table.

Table 16: Speaker-prompted result on LibriVox-GSR. **Bold**: best per column.

Configuration	SIG↑	BAK↑	OVRL↑	NISQA↑	SBS↑	SIM↑
Base (no guidance)	3.593	4.162	3.338	4.364	0.789	0.934
+ CSG $\omega = 1.0$	3.594	**4.169**	3.352	4.589	0.800	0.941
+ CSG $\omega = 1.0$ + prompt	**3.646**	**4.169**	**3.415**	**4.744**	**0.808**	**0.953**

Table 17: Music-prompted result on MUSDB18 full songs. **Bold**: best per column.

Configuration	FAD↓	LSD↓	SSIM↑	PQ↑	CE↑	CU↑
Base (no guidance)	0.205	7.032	0.679	7.031	6.115	6.775
+ CSG $\omega = 0.5$	0.198	**6.958**	0.682	7.198	6.293	6.911
+ CSG $\omega = 0.5$ + prompt	**0.066**	7.147	**0.690**	**7.528**	**7.074**	**7.303**