The Fortieth AAAI Conference on Artificial Intelligence (AAAI-26)

Multi-Metric Preference Alignment for Generative Speech Restoration

Junan Zhang[1], **Xueyao Zhang**[1], **Jing Yang**[2], **Yuancheng Wang**[1], **Fan Fan**[2], **Zhizheng Wu**[1,3,4*]

[1]School of Data Science, Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen
[2]Central Media Technology Institute, Huawei
[3]Shenzhen Loop Area Institute
[4]City University of Macau
junanzhang@link.cuhk.edu.cn, wuzhizheng@cuhk.edu.cn

Abstract

Recent generative models have significantly advanced speech restoration tasks, yet their training objectives often misalign with human perceptual preferences, resulting in suboptimal quality. While post-training alignment has proven effective in other generative domains like text and image generation, its application to generative speech restoration remains largely under-explored. This work investigates the challenges of applying preference-based post-training to this task, focusing on how to define a robust preference signal and curate high-quality data to avoid reward hacking. To address these challenges, we propose a **multi-metric preference alignment** strategy. We construct a new dataset, ***GenSR-Pref***, comprising 80K preference pairs, where each chosen sample is unanimously favored by a complementary suite of metrics covering perceptual quality, signal fidelity, content consistency, and timbre preservation. This principled approach ensures a holistic preference signal. Applying Direct Preference Optimization (DPO) with our dataset, we observe consistent and significant performance gains across three diverse generative paradigms: autoregressive models (AR), masked generative models (MGM), and flow-matching models (FM) on various restoration benchmarks, in both objective and subjective evaluations. Ablation studies confirm the superiority of our multi-metric strategy over single-metric approaches in mitigating reward hacking. Furthermore, we demonstrate that our aligned models can serve as powerful "data annotators", generating high-quality pseudo-labels to serve as a supervision signal for traditional discriminative models in data-scarce scenarios like singing voice restoration.

Demopage — https://gensr-pref.github.io
Extended version — https://arxiv.org/abs/2508.17229

1 Introduction

Speech restoration, which aims to recover high-quality speech from various degradations, is a fundamental task in audio processing (Liu et al. 2022; Zhang et al. 2024b). Recent advances in utilizing generative models for speech restoration, namely *generative speech restoration (GenSR)*, have shown remarkable performance in tasks such as denoising, dereverberation, declipping, super-resolution, etc. (Zhang et al. 2025a; Yang et al. 2024b; Li, Wang, and Liu 2024; Wang et al. 2025c, 2024b; Yao et al. 2025a; Kang et al. 2025; Wang et al. 2024a; Liu et al. 2024b). Unlike traditional discriminative methods that minimize a distance metric to a clean reference, generative speech restoration models are trained to maximize the likelihood of clean speech given a degraded input, thus learning the underlying distribution of clean speech (Lemercier et al. 2023), enabling them to generate high-fidelity audio even from severely degraded inputs.

In parallel, post-training techniques, designed to align generative models with specific downstream objectives or human preferences, have become integral to the advancement of generative modeling in some domains. In natural language processing (Ouyang et al. 2022; Rafailov et al. 2023; Bai et al. 2022), text-to-image/video (Xu et al. 2023; Wallace et al. 2024; Liu et al. 2025), and text-to-speech/audio (Zhang et al. 2025b; Sun et al. 2025; Hussain et al. 2025; Tian et al. 2025; Yao et al. 2025b; Zhang et al. 2024a; Liao et al. 2024) synthesis, preference-based alignment methods have been instrumental in enhancing model quality, safety, and human alignment. However, despite its potential, the application of such post-training alignment to the domain of generative speech restoration remains largely under-explored.

Applying preference-based post-training to GenSR models presents a unique set of challenges. Chief among them are: (1) **Defining a faithful preference signal**: How to construct an accessible, automated proxy that captures the multifaceted nature of human auditory perception (which values clarity, naturalness, and a lack of artifacts)? (2) **Curating high-quality preference data**: Given a preference signal, what is an effective strategy for constructing preference pairs that can robustly guide model optimization? (3) **Mitigating reward hacking**: How to ensure the model achieves holistic, genuine improvement, rather than simply learning to exploit the biases of a specific metric?

To address these challenges, we propose and investigate a **multi-metric preference alignment** strategy. We argue that a robust solution to reward hacking lies in the preference signal itself being multi-dimensional and holistic. To this end, we construct a new preference dataset, which we name ***GenSR-Pref***. A sample is chosen as winner only when it is *unanimously* judged superior by a complementary suite of metrics that assess four distinct aspects of quality: perceptual

*Zhizheng Wu is the corresponding author.

quality, signal-level fidelity, content alignment, and timbre preservation. This strict, unanimous criterion ensures that our model learns from a robust and holistic preference signal, effectively mitigating the risk of reward hacking.

Based on the *GenSR-Pref* dataset, we apply Direct Preference Optimization (DPO) (Rafailov et al. 2023) to align models across three major distinct generative paradigms: the sequential **Autoregressive (AR)** models, the iterative **Masked Generative (MGM)** models (also known as masked or discrete diffusion), and the continuous **Flow-Matching (FM)** models. Our approach yields significant and consistent improvements across all benchmarks even with only 3k pairs, confirmed by objective and subjective tests. Our key findings are threefold. First, our multi-metric signal is crucial for avoiding the **reward hacking** inherent in single-metric optimization. Second, we show that using ground-truth as a fixed winner causes **model collapse**, highlighting the need for learned, relative preferences over absolute targets. Third, models perform best with data from their own architecture, a principle we term **in-paradigm alignment**. We attribute this to unique "alignment directions" for each paradigm, where in-paradigm data provides a more direct optimization path. Finally, we demonstrate that our aligned models can act as powerful "data annotators", generating high-quality pseudo-labels to train discriminative models in data-scarce scenarios like singing voice restoration, thus bridging the gap between generative and discriminative paradigms.

To summarize, our main contributions are:

- We propose a **multi-metric preference alignment** strategy to address the challenges of defining a robust preference signal and mitigating reward hacking in post-training for generative speech restoration. Through extensive experiments, we demonstrate that this strategy yields significant and consistent improvements across three diverse generative paradigms (AR, MGM, and FM).
- We construct and introduce ***GenSR-Pref***, a dataset of 80k preference pairs built upon a strict, **unanimous multi-metric agreement criterion**. This dataset, which captures a holistic aspect of audio quality, will be publicly released to facilitate future research.
- We showcase a practical application of our aligned models, demonstrating their ability to serve as "data annotators" to generate pseudo-labels, thereby empowering the training of traditional discriminative models in data-scarce scenarios.

2 Related Work

2.1 Generative Speech Restoration

The paradigm for speech restoration has recently shifted towards generative modeling, which can be broadly categorized by the underlying generation mechanism. One major stream leverages **autoregressive language models** to treat restoration as a sequence-to-sequence problem, as seen in SELM (Wang et al. 2024b), TSELM (Tang, Zeng, and Li 2024), GenSE (Yao et al. 2025a), and versatile multi-task systems like SpeechX (Wang et al. 2024a), UniAudio (Yang et al. 2024a), and LLaSE-G1 (Kang et al. 2025). A second family consists of **masked generative models**, or discrete diffusion models, which iteratively refine tokenized audio representations. This approach has been applied to general restoration (e.g., MaskSR (Li, Wang, and Liu 2024; Liu et al. 2024c), AnyEnhance (Zhang et al. 2025a)), speech enhancement (e.g., Genhancer (Yang et al. 2024b)), and target speaker extraction (e.g., Metis (Wang et al. 2025c)). A third direction employs **generative models with continuous dynamics** to learn mappings in a continuous space; this includes methods based on score-SDE (Welker, Richter, and Gerkmann 2022), latent diffusion (Liu et al. 2024b; Wang et al. 2025a), flow-matching (Wang et al. 2025d; Liu et al. 2024a; Ku et al. 2024), and schrödinger bridges (Li et al. 2025a; Jukić et al. 2024) for various restoration tasks. Despite achieving state-of-the-art results, these generative models are typically trained using likelihood-based objectives, which often misalign with human perceptual quality. Our work addresses this gap by introducing a preference-based post-training strategy applicable across these diverse generative paradigms.

2.2 Post-Training for Audio Alignment

Post-training alignment has proven effective across various generative domains, including text (Ouyang et al. 2022; Rafailov et al. 2023; Bai et al. 2022), vision (Xu et al. 2023; Wallace et al. 2024; Liu et al. 2025), speech (Zhang et al. 2025b; Sun et al. 2025; Hussain et al. 2025; Tian et al. 2025; Yao et al. 2025b; Zhang et al. 2024a), music (Cideron et al. 2024; Lei et al. 2025), and audio effects (Majumder et al. 2024; Liao et al. 2024). In the context of speech restoration, early efforts like MetricGAN (Fu et al. 2019) used adversarial training to optimize for metrics like PESQ. More recently, preference-based methods have emerged, such as aligning with a learned NISQA predictor via PPO (Kumar, Perrault, and Williamson 2025) or using DPO with a single metric like UTMOS (Li et al. 2025b).

While single-metric alignment is effective, it risks reward hacking by ignoring that robust restoration is multi-faceted, requiring a balance of quality, fidelity, and timbre (Section 4.3). Grounded in this principle, we propose a **multi-metric alignment strategy** to pursue holistic improvement. We systematically show this approach yields more comprehensive gains across diverse generative paradigms while naturally mitigating reward hacking.

3 Multi-Meric Preference Alignment

Our approach to aligning generative speech restoration models is centered around a **multi-metric preference alignment** strategy, which is illustrated in Figure 1. This strategy involves two key stages: first, the construction of a high-quality preference dataset, and second, the application of Direct Preference Optimization (DPO) to align the model with these preferences. This approach is designed to be model-agnostic, allowing it to be applied to any generative restoration model.

3.1 The *GenSR-Pref* Dataset: Curating a Holistic Preference Signal

A pivotal element of our alignment strategy is the ***GenSR-Pref*** dataset, designed to provide a robust, holistic preference

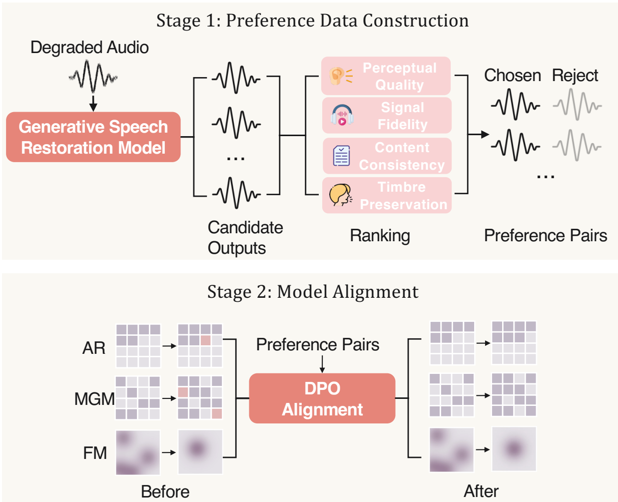


Figure 1: An overview of our multi-metric preference alignment strategy. The process consists of two main stages: (1) constructing the *GenSR-Pref* preference dataset by ranking model outputs based on a unanimous agreement across multiple metrics, and (2) fine-tuning the model with these preferences using Direct Preference Optimization (DPO).

signal while mitigating single-metric reward hacking. To ensure our dataset's breadth and quality, we generate candidate outputs using models from three distinct generative paradigms. We then form preference pairs by ranking outputs exclusively within each paradigm. This approach is chosen to capture the subtle variations in a single model's generations, providing a fine-grained signal tailored to its specific error patterns. This comprehensive strategy yields a dataset for robust alignment and enables detailed in-paradigm versus cross-paradigm analyses.

Model Selection To comprehensively evaluate our alignment framework across different generative paradigms, we adopt one representative model from each of the three major families: autoregressive (AR), masked generative (MGM), and flow-matching (FM) models.

Masked Generative Model (MGM). We use the pre-trained AnyEnhance (Zhang et al. 2025a) model as our MGM model. It predicts acoustic tokens based on the DAC codec (Kumar et al. 2024) from partially masked sequences, effectively capturing non-autoregressive and context-aware dependencies.

Autoregressive Model (AR). We train a new model, AR+Soundstorm, which follows a two-stage pipeline common in text-to-speech (TTS) domain. It first predicts semantic tokens of the Metis codec (Wang et al. 2025c) given noisy inputs, and a second-stage non-autoregressive model similar to Soundstorm (Borsos et al. 2023) then converts these tokens into acoustic tokens for waveform synthesis.

Flow-Matching Model (FM). We train Flow-SR, a model that learns to predict the velocity field of clean mel-spectrograms from noisy inputs via optimal transport interpolation based on the Diffusion Transformer (DiT) architecture (Peebles and Xie 2023). A pre-trained Vocos vocoder (Siuzdak 2023) is then used to synthesize the final waveform from the generated spectrograms. Details on model architectures and training procedures can be found in Appendix.

To ensure a reliable preference signal that captures the multi-faceted nature of audio quality, we evaluate each candidate output across four complementary dimensions:

- **Perceptual Quality (NISQA)** (Mittag et al. 2021): Assesses overall listening quality, including naturalness and the absence of annoying artifacts.
- **Signal-level Fidelity (DNSMOS)** (Reddy, Gopal, and Cutler 2022): A composite metric evaluating signal distortion, background noise, and overall quality for a fine-grained assessment of signal integrity.
- **Content Alignment (SpeechBERTScore)** (Saeki et al. 2024): Measures semantic similarity to the ground-truth transcription to ensure linguistic content is not altered during restoration.
- **Timbre Preservation (Speaker Similarity)**: Uses a pre-trained speaker verification model to compute cosine similarity, ensuring the speaker's identity is preserved.

A preference pair (y_w, y_l) is formed only under a **strict, unanimous agreement criterion**: the winning output y_w must score higher than the losing output y_l on *all* of these metrics simultaneously. This process ensures each preference pair represents a holistic and unambiguous improvement, providing a robust signal for the DPO alignment stage.

In total, *GenSR-Pref* comprises approximately 80K preference pairs, structured for distinct experimental goals. The primary component is a large-scale subset of **69,456** pairs from the AnyEnhance (MGM) model, allowing us to validate our alignment strategy's effectiveness with a substantial amount of data. For fair comparisons and ablation studies—particularly our cross-paradigm analysis—we created smaller, controlled subsets from a shared set of input utterances, yielding **3,354** pairs for AR, **3,428** for FM, and **3,035** for MGM. This dual-component structure enables both an investigation into data scaling effects and rigorous, controlled analyses of alignment dynamics.

3.2 GenSR Model Alignment via Direct Preference Optimization

In the second stage of our approach, we align the generative speech restoration models using the *GenSR-Pref* dataset. We employ **Direct Preference Optimization (DPO)** (Rafailov et al. 2023), a simple yet powerful technique that directly optimizes a policy to satisfy preferences without explicit reward modeling or reinforcement learning. The core idea of DPO is to re-parameterize the reward function in the standard Reinforcement Learning from Human Feedback (RLHF) objective using the optimal policy, which leads to a simple contrastive loss over preferred and dispreferred samples. In this section, we detail how we adapt and apply DPO to the three diverse generative paradigms used in our study: Autoregressive (AR), Masked Generative (MGM), and Flow-Matching (FM) models. Detailed derivations can be found in Appendix.

DPO for Autoregressive Models AR models treat speech restoration as a sequence-to-sequence task. Given a degraded input x, they generate a clean audio sequence $y = (y_1, ..., y_T)$ token by token. They are typically trained by maximizing the log-likelihood of the ground-truth sequence, using a standard cross-entropy loss:

$$\mathcal{L}_{\text{AR}} = -\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{train}}}\left[\sum_{t=1}^{T}\log p_\theta(y_t|y_{<t}, x)\right] \quad (1)$$

DPO Derivation. The standard RLHF objective aims to find a policy p_θ that maximizes the reward from a learned reward model $r_\phi(x, y)$, while constraining its deviation from a reference policy p_{ref} using a KL-divergence term:

$$\max_{p_\theta} \mathbb{E}_{y\sim p_\theta(y|x)}[r_\phi(x, y)] - \beta D_{\text{KL}}\left[p_\theta(y|x) \,\|\, p_{\text{ref}}(y|x)\right] \quad (2)$$

The key insight of DPO is that for this objective, the optimal policy p^*_θ has a closed-form solution: $p^*_\theta(y|x) \propto p_{\text{ref}}(y|x)\exp(\frac{1}{\beta}r_\phi(x, y))$. By rearranging this equation, the unknown reward function $r_\phi(x, y)$ can be expressed in terms of the optimal policy and the reference policy. Substituting this back into the Bradley-Terry loss for the reward model, which models the probability of preferring y_w over y_l, yields the final DPO loss that directly optimizes the policy p_θ:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\Big(\beta\log\frac{p_\theta(y_w|x)}{p_{\text{ref}}(y_w|x)} - \beta\log\frac{p_\theta(y_l|x)}{p_{\text{ref}}(y_l|x)}\Big)\right] \quad (3)$$

where $\mathcal{D}$ is our *GenSR-Pref* preference dataset, and σ is the sigmoid function.

DPO for Masked Generative Models MGM models (Chang et al. 2022; Zhang et al. 2025a; Wang et al. 2025c; Ju et al. 2024; Wang et al. 2025b), also known as discrete diffusion models, learn to restore a clean speech sequence y_0 from a partially masked or corrupted version y_t, where $t \in [0, T]$ is the noise step parameter that controls the amount of corruption, and y_T is the fully masked sequence. They are trained to predict the original tokens in the masked positions, optimizing a masked language modeling objective:

$$\mathcal{L}_{\text{MGM}} = -\mathbb{E}_{(x,y_0),t,m_t}\left[\sum_{i=1}^{T} m_{t,i}\log p_\theta(y_{0,i}|y_t, x)\right] \quad (4)$$

where $m_t \in \{0, 1\}^T$ is a binary mask indicating which tokens in the sequence are masked (1) or unmasked (0). The model learns to predict the original tokens in the masked positions, effectively learning to denoise the corrupted input.

DPO Derivation. Following the INTP framework (Zhang et al. 2025b), we extend DPO to this non-autoregressive setting. The core logic remains the same, but the policy now represents the conditional probability of the full clean sequence y_0 given the corrupted version y_t. The DPO loss for MGM is thus analogous to the AR case, contrasting the log-probabilities of the preferred and dispreferred sequences, conditioned on their respective masked inputs:

$$\mathcal{L}_{\text{DPO-MGM}} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D},\,t}\left[\log\sigma\Big(\beta\log\frac{p_\theta(y_0^w|y_t^w, x)}{p_{\text{ref}}(y_0^w|y_t^w, x)} - \beta\log\frac{p_\theta(y_0^l|y_t^l, x)}{p_{\text{ref}}(y_0^l|y_t^l, x)}\Big)\right] \quad (5)$$

where y_t^w and y_t^l are the masked versions of the winning (y_0^w) and losing (y_0^l) sequences from our preference dataset.

DPO for Flow-Matching Models FM models (Lipman et al. 2022) learn a continuous mapping from a simple prior distribution (e.g., Gaussian noise $y_0 \sim \mathcal{N}(0, I)$) to a complex data distribution (e.g., clean speech y_1). They are trained to predict the velocity vector field $v_\theta(y_t, t, x)$ that describes the probabilistic flow of data points along a linear interpolation path from the noise y_0 to the data y_1. The standard training objective minimizes the difference between the predicted velocity and the true velocity:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,y_0,y_1,x}\left[\|v_\theta(y_t, t, x) - (y_1 - y_0)\|_2^2\right] \quad (6)$$

where $y_t = (1-t)y_0 + ty_1$ is a point on the linear trajectory from the noise to the data, and t is a noise step parameter that controls the interpolation between the two. The model learns to predict the velocity field that describes the continuous flow of data points along the linear interpolation path, effectively learning to denoise by predicting the gradient of the data distribution along the linear interpolation.

DPO Derivation. To extend DPO to FM models, we follow recent advances in aligning diffusion-based generators (Wallace et al. 2024; Liu et al. 2025). Instead of modeling the full data likelihood along the entire generative trajectory, we use a tractable surrogate objective that operates on a single noise step. Specifically, we approximate the DPO objective using the model's instantaneous prediction error at a sampled timestep, treating the L2 loss between predicted and true velocities as a proxy for negative log-likelihood. The resulting loss encourages the model to reduce velocity error for preferred samples (y_w) while increasing it for dispreferred ones (y_l), yielding the following expression:

$$\mathcal{L}_{\text{DPO-FM}} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D},t,y_0}\left[\log\sigma\left(-\beta(\Delta_w - \Delta_l)\right)\right] \quad (7)$$

where Δ represents the difference in squared L2 error between the trained model v_θ and the reference model v_{ref} for a given sample (y_w or y_l):

$$\Delta_w = \|v_\theta(y_t^w, t, x) - (y_1^w - y_0)\|_2^2 - \|v_{\text{ref}}(y_t^w, t, x) - (y_1^w - y_0)\|_2^2 \quad (8)$$

and Δ_l is defined analogously for the losing sample y_l. This single-step formulation provides a scalable and effective approximation to DPO in diffusion processes.

4 Experiments

4.1 Experimental Setup

Models and Datasets. We evaluate our multi-metric preference alignment strategy across the three generative paradigms: AnyEnhance for MGM, AR+Soundstorm for AR, and Flow-SR for FM. All alignment experiments are conducted using our newly constructed *GenSR-Pref* dataset. Specifically, we conduct the main experiment (Section 4.2) using the primary preference subset (69k pairs for MGM, approx. 3k for AR/FM), while the ablations (Section 4.3) are conducted using the controlled smaller 3k-pair subsets for all models to ensure fair and direct comparisons.

Evaluation Benchmarks and Metrics. We conduct evaluations on two major kinds of benchmarks: (1) General speech restoration (GSR) that includes denoising, dereveration, declipping and super-resolution at the same time (Voicefixer-GSR, Librivox-GSR, CCMusic-GSR), (2) Speech Enhance-

Dataset	Model	Type	DPO-aligned?	SIG↑	BAK↑	OVRL↑	NISQA↑	Speech-BERTScore↑	Similarity↑
Voicefixer-GSR	Voicefixer	DISC.	-	3.299	3.971	3.003	4.160	0.797	0.882
	MaskSR	MGM	-	3.445	3.971	3.128	-	-	-
	AnyEnhance	MGM	✗	3.406	4.073	3.136	4.308	0.829	0.924
			✓	**3.532**	**4.091**	**3.267**	**4.639**	**0.834**	**0.935**
	AR+Soundstorm	AR	✗	3.550	4.097	3.294	4.556	0.788	0.894
			✓	**3.564**	**4.144**	**3.331**	**4.850**	**0.803**	**0.904**
	Flow-SR	FM	✗	3.398	3.969	3.104	4.010	0.812	0.918
			✓	**3.483**	**4.092**	**3.230**	**4.672**	**0.830**	**0.924**
Librivox-GSR	AnyEnhance	MGM	✗	3.546	4.142	3.308	4.346	0.822	**0.955**
			✓	**3.690**	**4.201**	**3.475**	**4.865**	**0.828**	0.954
	AR+Soundstorm	AR	✗	3.663	4.134	3.419	4.535	0.783	0.922
			✓	**3.693**	**4.193**	**3.478**	**4.923**	**0.793**	**0.924**
	Flow-SR	FM	✗	3.550	4.062	3.281	4.184	0.791	**0.931**
			✓	**3.602**	**4.152**	**3.368**	**4.825**	**0.801**	0.930
CCMusic-GSR	AnyEnhance	MGM	✗	3.243	3.547	2.797	3.345	0.811	**0.915**
			✓	**3.440**	**3.827**	**3.062**	**4.154**	**0.817**	0.909
	AR+Soundstorm	AR	✗	3.378	3.693	2.956	3.948	0.710	**0.854**
			✓	**3.460**	**3.865**	**3.094**	**4.438**	**0.713**	0.853
	Flow-SR	FM	✗	3.298	3.540	2.813	3.897	0.733	**0.883**
			✓	**3.378**	**3.770**	**2.971**	**4.371**	**0.740**	0.881

Table 1: Quantitative comparison on GSR benchmarks across three generative paradigms—MGM (AnyEnhance), AR (AR+Soundstorm), and FM (Flow-SR)—before and after DPO alignment. Results from a discriminative baseline Voicefixer (DISC.) and the MaskSR model are also included. Bold indicates improvements after alignment.

ment (SE) under noisy and reverberant conditions (DNS-No-Reverb and DNS-With-Reverb). For a comprehensive evaluation, we not only include the original unaligned models, but also involve state-of-the-art baselines LLaSE-G1 (Kang et al. 2025), GenSE (Yao et al. 2025a), MaskSR (Li, Wang, and Liu 2024), and FlowSE (Wang et al. 2025d), and a discriminative model Voicefixer (Liu et al. 2022). Performance is measured using a comprehensive suite of objective metrics: DNSMOS (SIG, BAK, OVRL) (Reddy, Gopal, and Cutler 2022) for signal fidelity and quality, NISQA (Mittag et al. 2021) for perceptual quality, SpeechBERTScore (SBERT) (Saeki et al. 2024) for content consistency, and speaker similarity (SIM). For subjective evaluation, we conduct A/B preference tests to assess human perception. Details on the evaluation can be found in Appendix.

4.2 Main Results: Effectiveness of Multi-Metric Preference Alignment

Table 1 presents the main results, showing that our multi-metric preference alignment strategy is consistently effective across all three paradigms. Notably, the MGM model achieves significant gains when aligned on its large-scale 69k-pair subset (e.g., **+0.519** NISQA on LibriVox-GSR), demonstrating our method's scalability. Impressively, the AR and FM models also show significant improvements (e.g., **+0.388** and **+0.641** NISQA) even with their much smaller, 3k-pair data, highlighting the data efficiency of our approach. These objective gains are confirmed by subjective A/B tests (Figure 2), where human listeners consistently preferred the aligned models, with win rates reaching up to **54.5%**.

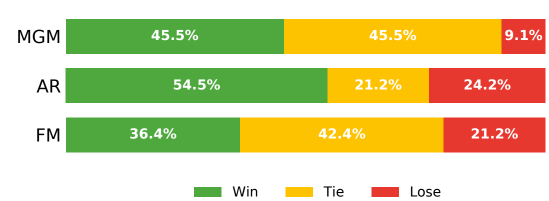


Figure 2: Human AB preference test results for DPO-aligned models on Librivox-GSR testset.

Moreover, this strong performance generalizes beyond GSR tasks. As shown in Table 2, models aligned with our diverse preference data also exhibit seamless performance gains on the downstream DNS speech enhancement benchmark. For example, the aligned AnyEnhance model boosts its OVRL score from 3.204 to 3.438 on the reverberant set, demonstrating the robustness of our alignment strategy. More results and audio visualizations can be found in Appendix.

4.3 Ablation Studies and Analysis

To better understand the factors contributing to our performance gains, we conduct a series of in-depth ablation studies. **Multi-Metric vs. Single-Metric Alignment.** To validate our core hypothesis that a multi-metric signal is crucial for avoid-

Data	Model	Type	Aligned?	SIG	BAK	OVRL
No Reverb	MaskSR	MGM	-	3.586	4.116	3.339
	GenSE	AR	-	3.650	4.180	3.430
	LLaSE-G1	AR	-	3.660	4.170	3.420
	FlowSE	FM	-	3.685	4.201	3.445
	AnyEnhance	MGM	✗	3.640	4.179	3.418
			✓	**3.684**	**4.203**	**3.476**
	AR+Soundstorm	AR	✗	3.648	4.155	3.422
			✓	**3.673**	**4.194**	**3.469**
	Flow-SR	FM	✗	3.581	4.133	3.355
			✓	**3.632**	**4.173**	**3.420**
With Reverb	MaskSR	MGM	-	3.531	4.065	3.253
	GenSE	AR	-	3.490	3.730	3.190
	LLaSE-G1	AR	-	3.590	4.100	3.330
	FlowSE	FM	-	3.601	4.102	3.331
	AnyEnhance	MGM	✗	3.500	4.040	3.204
			✓	**3.670**	**4.178**	**3.438**
	AR+Soundstorm	AR	✗	3.681	4.127	3.431
			✓	**3.709**	**4.189**	**3.496**
	Flow-SR	FM	✗	3.539	4.019	3.255
			✓	**3.629**	**4.163**	**3.399**

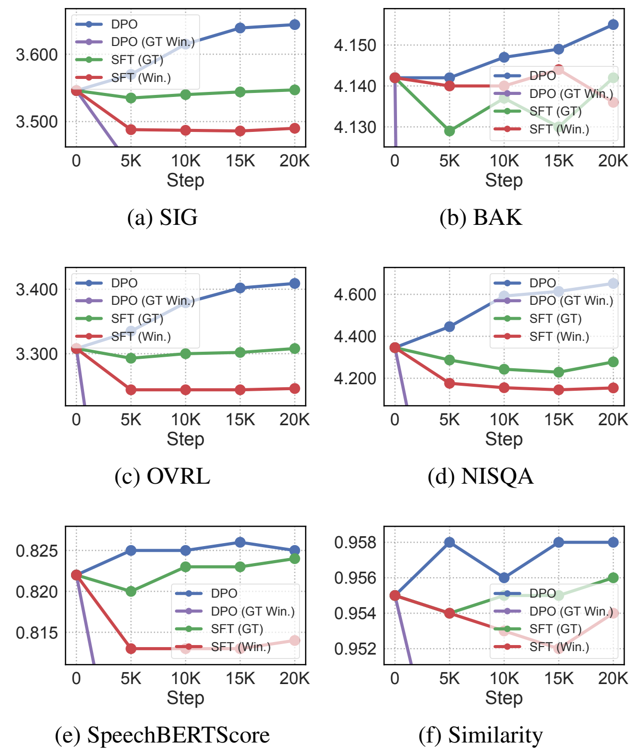
Table 2: Evaluation results on the DNS speech enhancement benchmark. Models aligned on diverse restoration preference pairs (e.g., denoising, declipping) exhibit seamless performance gains when applied to speech enhancement as a downstream task.

ing reward hacking, we compare our Multi-Metric Preference Alignment approach with DPO alignment using preference pairs constructed from single metrics. As shown in Table 3, while optimizing for a single metric improves that specific score, it often leads to stagnation or even degradation in other, un-targeted metrics. For example, the Similarity-aligned model performs worse than the baseline on OVRL and SIG. In contrast, our multi-metric approach achieves robust improvements across all metrics simultaneously. This confirms that our unanimous agreement criterion is effective at mitigating reward hacking and promoting holistic quality improvement. More results can be found in Appendix.

Criterion	SIG	BAK	OVRL	NISQA	SBERT	SIM
-	3.550	4.097	3.294	4.556	0.788	0.894
Multi-Metric	**3.564**	**4.144**	**3.331**	**4.850**	0.803	**0.904**
NISQA	3.531	4.137	3.300	4.810	0.785	0.896
OVRL	3.561	4.117	3.317	4.600	0.792	0.896
SIM	3.537	4.101	3.285	4.577	0.792	0.901
SBERT	3.540	4.109	3.291	4.612	**0.804**	0.901

Table 3: Ablation on preference metric for the AR model on Voicefixer-GSR testset. The Multi-Metric Preference Strategy ("Multi-Metric") consistently outperforms single-metric approaches across all evaluation metrics.

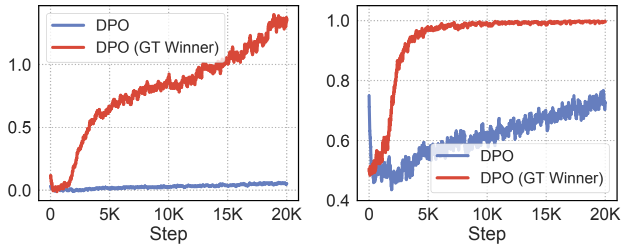
DPO vs. Supervised Fine-Tuning (SFT). To highlight the benefits of preference-based optimization, we compare DPO against two SFT baselines: fine-tuning on ground-truth audio (SFT-GT) and on the "winner" samples from our preference data (SFT-Winner). As shown in Figure 3, DPO consistently outperforms both SFT variants, which show only marginal gains before stagnating. This suggests that simply exposing the model to high-quality examples is insufficient for effective alignment. Furthermore, a naive DPO strategy using ground-truth as the fixed "winner" leads to model collapse, as detailed in Figure 4. This behavior is driven by the core DPO objective, which seeks to maximize the log-probability ratio, $\beta \log \frac{p_\theta(y_0^w|y_t^w,x)}{p_{\text{ref}}(y_0^w|y_t^w,x)} - \beta \log \frac{p_\theta(y_0^l|y_t^l,x)}{p_{\text{ref}}(y_0^l|y_t^l,x)}$. When y_w is a fixed ground-truth target, the model learns a pathological shortcut—drastically suppressing the probability of all outputs other than the GT to maximize this ratio. This highlights the necessity of learning from relative, nuanced preferences.



(a) SIG (b) BAK (c) OVRL (d) NISQA (e) SpeechBERTScore (f) Similarity

Figure 3: Ablation study on training objectives for the MGM model. DPO demonstrates consistent improvements across training steps, while SFT tends to stagnate or degrade. Notably, a naive DPO variant that treats ground-truth outputs as winners ("GT Winner") results in model collapse.

The Role of Data Source in Alignment. Finally, we explore how the source of preference data affects alignment performance. Table 4 shows the results of aligning each model with preference pairs generated from itself (in-paradigm) versus from other paradigms (cross-paradigm). This distinction is analogous to the concepts of on-policy and off-policy learning in reinforcement learning. We observe a clear trend: all three models achieve their optimal performance when aligned with their own in-paradigm preference data. To quantitatively investigate this phenomenon, we formally define



(a) Reward Margin ($\times 10^5$) (b) Reward Accuracy

Figure 4: Detailed training curves between normal DPO and DPO (GT Winner) training (smoothing factor = 0.99). Using ground truth as the unconditional winner leads to inflated reward margins and saturated reward accuracy, indicating model collapse.

Model (Source)	SIG	BAK	OVRL	NISQA	SBERT	SIM
AR (-)	3.663	4.134	3.419	4.535	0.783	0.922
AR (AR)	3.672	**4.171**	**3.449**	**4.867**	**0.791**	**0.925**
AR (MGM)	**3.675**	4.148	3.435	4.604	0.786	0.919
AR (FM)	3.668	4.166	3.440	4.783	0.788	0.918
FM (-)	3.550	4.062	3.281	4.184	0.791	**0.931**
FM (AR)	3.575	4.103	3.324	4.473	0.794	0.932
FM (MGM)	**3.587**	4.048	3.305	4.269	0.789	0.930
FM (FM)	3.580	**4.126**	**3.337**	**4.645**	**0.797**	0.929
MGM (-)	3.546	4.142	3.308	4.346	0.822	0.955
MGM (AR)	3.589	4.147	3.346	4.321	0.809	0.951
MGM (MGM)	**3.644**	**4.155**	**3.409**	**4.652**	**0.825**	**0.958**
MGM (FM)	3.478	4.154	3.232	4.468	0.802	0.946

Table 4: Ablation on the source of preference data, evaluated on the Librivox-GSR testset. For each base model, performance is compared against its baseline (-) when aligned with data from different model paradigms (AR, MGM, FM).

	AR	FM	MGM
AR	**0.0075**	0.0072	0.0018
FM	0.0072	**0.0102**	0.0005
MGM	0.0018	0.0005	**0.0045**

Table 5: Average pairwise cosine similarity of preference vectors ($v_{\text{pref}} = f(y_w) - f(y_l)$) within and across generative paradigms. Higher values indicate greater alignment in the 'alignment direction'. The strong diagonal highlights that in-paradigm preference data is more internally consistent.

the principle of **in-paradigm alignment**. We hypothesize that each architecture possesses a unique "preference direction"—an optimal path for improvement. To quantify this, we define a "preference vector" as the feature-space difference between winner and loser samples ($v_{\text{pref}} = f(y_w) - f(y_l)$) using pre-trained w2v-bert-2 model[1] and compute their average pairwise cosine similarity. The results (Table 5) provide strong quantitative evidence for our hypothesis: the average **in-paradigm similarity (diagonal values) is consistently higher than the cross-paradigm similarity** (e.g., AR-AR similarity of 0.0075 vs. AR-MGM of 0.0018). This indicates that data from a model's own paradigm provides a more consistent and targeted optimization signal.

Furthermore, these similarities correlate with alignment performance (Table 4): paradigms with more aligned preference vectors (e.g., AR and FM, similarity 0.0072) show better cross-paradigm transfer performance than less aligned ones (e.g., FM and MGM, similarity 0.0005). This analysis provides strong backing for our principle, suggesting that alignment is most effective when the preference data's direction matches the target model's intrinsic generative manifold.

[1]https://huggingface.co/facebook/w2v-bert-2.0

4.4 Application: Empowering Discriminative Models via Pseudo-Labeling

Model	SIG	BAK	OVRL	NISQA
Voicefixer (before)	2.657	3.080	2.295	2.919
Voicefixer (after)	**3.096**	**3.745**	**2.756**	**3.312**

Table 6: Evaluation on real-world singing-voice recordings before and after DPO model's annotation

We demonstrate our aligned model's utility as a "data annotator" in a data-scarce scenario by fine-tuning the Voicefixer restoration model on singing recordings. Instead of ground-truth data, we used pseudo-labels generated by our DPO-enhanced AnyEnhance model as supervision. This resulted in dramatic performance improvements across all metrics (Table 6). Our findings show a powerful generative model can effectively create supervision signals to train smaller, discriminative models when paired data is prohibitively difficult to obtain. This bridges the gap between the two modeling paradigms and opens up new avenues for practical speech restoration. More detail on this application can be found in the Appendix.

5 Conclusion

In this work, we introduced a **multi-metric preference alignment** strategy to align generative speech restoration models with human perception while mitigating the risk of reward hacking. To this end, we constructed a new dataset, ***GenSR-Pref***, using a strict unanimous agreement criterion across four complementary metrics to ensure a holistic and robust preference signal. Applying DPO with our dataset yields significant improvements across diverse generative paradigms (AR, MGM, and FM), outperforming single-metric preference and demonstrating the principle of in-paradigm alignment. Furthermore, we demonstrated a novel and practical application where our aligned models serve as powerful "data annotators", generating high-quality pseudo-labels to effectively train discriminative models in data-scarce scenarios. Future work includes exploring more advanced alignment algorithms, investigating preference data properties, and leveraging our annotation capabilities for large-scale data curation in other domains like TTS.

Acknowledgments

This work is funded by the NSFC under Grant 62376237, Shenzhen Science and Technology Program ZDSYS20230626091302006, 2023 Shenzhen stability Science Program, Internal Project Fund from Shenzhen Research Institute of Big Data, Grant No. T00120230002, and Program for Guangdong Introducing Innovative and Enterpreneurial Teams, Grant No. 2023ZT10X044.

References

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Borsos, Z.; Sharifi, M.; Vincent, D.; Kharitonov, E.; Zeghidour, N.; and Tagliasacchi, M. 2023. Soundstorm: Efficient parallel audio generation. *arXiv preprint arXiv:2305.09636*.

Chang, H.; Zhang, H.; Jiang, L.; Liu, C.; and Freeman, W. T. 2022. Maskgit: Masked generative image transformer. In *CVPR*.

Cideron, G.; Girgin, S.; Verzetti, M.; Vincent, D.; Kastelic, M.; Borsos, Z.; Mcwilliams, B.; Ungureanu, V.; Bachem, O.; Pietquin, O.; et al. 2024. MusicRL: Aligning Music Generation to Human Preferences. In *ICML*, 8968–8984. PMLR.

Fu, S.-W.; Liao, C.-F.; Tsao, Y.; and Lin, S.-D. 2019. Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement. In *ICML*, 2031–2041.

Hussain, S.; Neekhara, P.; Yang, X.; Casanova, E.; Ghosh, S.; Desta, M. T.; Fejgin, R.; Valle, R.; and Li, J. 2025. Koel-TTS: Enhancing LLM based Speech Generation with Preference Alignment and Classifier Free Guidance. *arXiv preprint arXiv:2502.05236*.

Ju, Z.; Wang, Y.; Shen, K.; Tan, X.; Xin, D.; Yang, D.; Liu, Y.; Leng, Y.; Song, K.; Tang, S.; et al. 2024. NaturalSpeech 3: zero-shot speech synthesis with factorized codec and diffusion models. In *ICML*.

Jukić, A.; Korostik, R.; Balam, J.; and Ginsburg, B. 2024. Schrödinger bridge for Generative Speech Enhancement. *arXiv preprint arXiv:2407.16074*.

Kang, B.; Zhu, X.; Zhang, Z.; Ye, Z.; Liu, M.; Wang, Z.; Zhu, Y.; Ma, G.; Chen, J.; Xiao, L.; et al. 2025. LLaSE-G1: Incentivizing generalization capability for llama-based speech enhancement. *arXiv preprint arXiv:2503.00493*.

Ku, P.-J.; Liu, A. H.; Korostik, R.; Huang, S.-F.; Fu, S.-W.; and Jukić, A. 2024. Generative Speech Foundation Model Pretraining for High-Quality Speech Extraction and Restoration. *arXiv preprint arXiv:2409.16117*.

Kumar, A.; Perrault, A.; and Williamson, D. S. 2025. Using RLHF to align speech enhancement approaches to mean-opinion quality scores. In *ICASSP*, 1–5.

Kumar, R.; Seetharaman, P.; Luebs, A.; Kumar, I.; and Kumar, K. 2024. High-fidelity audio compression with improved rvqgan. *NeurIPS*.

Lei, S.; Xu, Y.; Lin, Z.; Zhang, H.; Tan, W.; Chen, H.; Yu, J.; Zhang, Y.; Yang, C.; Zhu, H.; et al. 2025. LeVo: High-Quality Song Generation with Multi-Preference Alignment. *arXiv preprint arXiv:2506.07520*.

Lemercier, J.-M.; Richter, J.; Welker, S.; and Gerkmann, T. 2023. Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration. In *ICASSP*, 1–5.

Li, C.; Chen, Z.; Bao, F.; and Zhu, J. 2025a. Bridge-sr: Schrödinger bridge for efficient sr. In *ICASSP*, 1–5.

Li, H.; Hou, N.; Hu, Y.; Yao, J.; Siniscalchi, S. M.; and Chng, E. S. 2025b. Aligning Generative Speech Enhancement with Human Preferences via Direct Preference Optimization. *arXiv preprint arXiv:2507.09929*.

Li, X.; Wang, Q.; and Liu, X. 2024. MaskSR: Masked Language Model for Full-band Speech Restoration. *arXiv preprint arXiv:2406.02092*.

Liao, H.; Han, H.; Yang, K.; Du, T.; Yang, R.; Xu, Z.; Xu, Q.; Liu, J.; Lu, J.; and Li, X. 2024. Baton: Aligning text-to-audio model with human preference feedback. *arXiv preprint arXiv:2402.00744*.

Lipman, Y.; Chen, R. T.; Ben-Hamu, H.; Nickel, M.; and Le, M. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.

Liu, A. H.; Le, M.; Vyas, A.; Shi, B.; Tjandra, A.; and Hsu, W.-N. 2024a. Generative Pre-training for Speech with Flow Matching. In *ICLR*.

Liu, H.; Chen, K.; Tian, Q.; Wang, W.; and Plumbley, M. D. 2024b. AudioSR: Versatile audio super-resolution at scale. In *ICASSP*.

Liu, H.; Liu, X.; Kong, Q.; Tian, Q.; Zhao, Y.; Wang, D.; Huang, C.; and Wang, Y. 2022. Voicefixer: A unified framework for high-fidelity speech restoration. *arXiv preprint arXiv:2204.05841*.

Liu, J.; Liu, G.; Liang, J.; Yuan, Z.; Liu, X.; Zheng, M.; Wu, X.; Wang, Q.; Qin, W.; Xia, M.; et al. 2025. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*.

Liu, X.; Li, X.; Serrà, J.; and Pascual, S. 2024c. Joint Semantic Knowledge Distillation and Masked Acoustic Modeling for Full-band Speech Restoration with Improved Intelligibility. *arXiv preprint arXiv:2409.09357*.

Majumder, N.; Hung, C.-Y.; Ghosal, D.; Hsu, W.-N.; Mihalcea, R.; and Poria, S. 2024. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. In *ACM Multimedia*, 564–572.

Mittag, G.; Naderi, B.; Chehadi, A.; and Möller, S. 2021. NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets. In *INTERSPEECH*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *NeurIPS*, 35: 27730–27744.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *ICCV*, 4195–4205.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 36: 53728–53741.

Reddy, C. K.; Gopal, V.; and Cutler, R. 2022. DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP*.

Saeki, T.; Maiti, S.; Takamichi, S.; Watanabe, S.; and Saruwatari, H. 2024. SpeechBERTScore: Reference-Aware Automatic Evaluation of Speech Generation Leveraging NLP Evaluation Metrics. *arXiv preprint arXiv:2401.16812*.

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

Sun, X.; Xiao, R.; Mo, J.; Wu, B.; Yu, Q.; and Wang, B. 2025. F5R-TTS: Improving flow-matching based text-to-speech with group relative policy optimization. *arXiv preprint arXiv:2504.02407*.

Tang, B.; Zeng, B.; and Li, M. 2024. TSELM: Target Speaker Extraction using Discrete Tokens and Language Models. *arXiv preprint arXiv:2409.07841*.

Tian, J.; Zhang, C.; Shi, J.; Zhang, H.; Yu, J.; Watanabe, S.; and Yu, D. 2025. Preference alignment improves language model-based tts. In *ICASSP*, 1–5. IEEE.

Wallace, B.; Dang, M.; Rafailov, R.; Zhou, L.; Lou, A.; Purushwalkam, S.; Ermon, S.; Xiong, C.; Joty, S.; and Naik, N. 2024. Diffusion model alignment using direct preference optimization. In *CVPR*, 8228–8238.

Wang, H.; Hai, J.; Yang, D.; Chen, C.; Li, K.; Peng, J.; Thebaud, T.; Velazquez, L. M.; Villalba, J.; and Dehak, N. 2025a. SoloSpeech: Enhancing Intelligibility and Quality in Target Speech Extraction through a Cascaded Generative Pipeline. *arXiv preprint arXiv:2505.19314*.

Wang, X.; Thakker, M.; Chen, Z.; Kanda, N.; Eskimez, S. E.; Chen, S.; Tang, M.; Liu, S.; Li, J.; and Yoshioka, T. 2024a. Speechx: Neural codec language model as a versatile speech transformer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Wang, Y.; Zhan, H.; Liu, L.; Zeng, R.; Guo, H.; Zheng, J.; Zhang, Q.; Zhang, X.; Zhang, S.; and Wu, Z. 2025b. MaskGCT: Zero-Shot Text-to-Speech with Masked Generative Codec Transformer. In *ICLR*.

Wang, Y.; Zheng, J.; Zhang, J.; Zhang, X.; Liao, H.; and Wu, Z. 2025c. Metis: A Foundation Speech Generation Model with Masked Generative Pre-training. *arXiv preprint arXiv:2502.03128*.

Wang, Z.; Liu, Z.; Zhu, X.; Zhu, Y.; Liu, M.; Chen, J.; Xiao, L.; Weng, C.; and Xie, L. 2025d. FlowSE: Efficient and High-Quality Speech Enhancement via Flow Matching. *arXiv preprint arXiv:2505.19476*.

Wang, Z.; Zhu, X.; Zhang, Z.; Lv, Y.; Jiang, N.; Zhao, G.; and Xie, L. 2024b. SELM: Speech enhancement using discrete tokens and language models. In *ICASSP*.

Welker, S.; Richter, J.; and Gerkmann, T. 2022. Speech enhancement with score-based generative models in the complex STFT domain. *arXiv preprint arXiv:2203.17004*.

Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; and Dong, Y. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. *NeurIPS*, 36: 15903–15935.

Yang, D.; Tian, J.; Tan, X.; et al. 2024a. UniAudio: Towards Universal Audio Generation with Large Language Models. In *ICML*.

Yang, H.; Su, J.; Kim, M.; and Jin, Z. 2024b. Genhancer: High-Fidelity Speech Enhancement via Generative Modeling on Discrete Codec Tokens. In *INTERSPEECH*.

Yao, J.; Liu, H.; Chen, C.; Hu, Y.; Chng, E.; and Xie, L. 2025a. GenSE: Generative Speech Enhancement via Language Models using Hierarchical Modeling. In *ICLR*.

Yao, J.; Yang, Y.; Pan, Y.; Feng, Y.; Ning, Z.; Ye, J.; Zhou, H.; and Xie, L. 2025b. Fine-grained Preference Optimization Improves Zero-shot Text-to-Speech. *arXiv preprint arXiv:2502.02950*.

Zhang, D.; Li, Z.; Li, S.; Zhang, X.; Wang, P.; Zhou, Y.; and Qiu, X. 2024a. SpeechAlign: Aligning Speech Generation to Human Preferences. In *NeurIPS*.

Zhang, J.; Yang, J.; Fang, Z.; Wang, Y.; Zhang, Z.; Wang, Z.; Fan, F.; and Wu, Z. 2025a. AnyEnhance: A Unified Generative Model With Prompt-Guidance and Self-Critic for Voice Enhancement. *IEEE Transactions on Audio, Speech and Language Processing*.

Zhang, W.; Scheibler, R.; Saijo, K.; Cornell, S.; Li, C.; Ni, Z.; Pirklbauer, J.; Sach, M.; Watanabe, S.; Fingscheidt, T.; and Qian, Y. 2024b. URGENT Challenge: Universality, Robustness, and Generalizability For Speech Enhancement. In *INTERSPEECH*.

Zhang, X.; Wang, Y.; Wang, C.; Li, Z.; Chen, Z.; and Wu, Z. 2025b. Advancing zero-shot text-to-speech intelligibility across diverse domains via preference alignment. In *ACL*, 12251–12270.