ResoDiff-44k: High-Fidelity Cross-Lingual Speech and Singing Synthesis via Discrete Diffusion

Gyanendra Das **Sai Satyam Jena**

Abstract

While large-scale generative speech models have achieved remarkable semantic coherence, industrial deployment remains constrained by a fidelity ceiling typically capped at lower sampling rates. A fundamental limitation is the reliance on intermediate mel-spectrograms, a low-dimensional bottleneck that discards phase and high-frequency information, causing artifacts in expressive scenarios like singing. In this work, we introduce ResoDiff-44k, a production-grade generative foundation model designed for cinema-quality, 44.1kHz audio synthesis. Departing from standard masked audio modeling and mel-spectrogram inversion, ResoDiff-44k leverages Discrete Diffusion over a pure Descript Audio Codec latent space. We pre-train ResoDiff-44k on a massive 150K -hour multilingual dataset to establish a robust acoustic prior, followed by targeted fine-tuning on a curated regional mixed-language and singing corpus. Our experiments demonstrate that replacing the standard prediction head with a discrete diffusion trajectory significantly reduces misalignment in long sequences. We report a double-blind subjective evaluation showing that ResoDiff-44k achieves a 4.6 Mean Opinion Score in 44.1kHz singing synthesis and a 71% reduction in character error rate on regional mixed-language prompts compared to strong baselines. The proposed pipeline offers a viable path for deploying high-fidelity, culturally adaptive conversational agents.

1 Introduction

Current SOTA models like VALL-E Wang et al. (2023) and Metis Wang et al. (2025) model audio as discrete tokens. However, they introduce an "information bottleneck." Metis, for example, uses a two-stage pipeline that first predicts lossy semantic (SSL) tokens before mapping them to 24kHz acoustic tokens. This approach discards critical paralinguistic details (breath, pitch micro-modulations) and limits the spectral bandwidth, failing to meet the demands of high-fidelity singing or media production.

To address these limitations, we present ResoDiff-44k, a unified framework that synthesizes industrial-grade 44.1kHz audio. It operates directly on the discrete latent space of a high-fidelity audio codec, eliminating intermediate SSL and mel-spectrogram representations. By leveraging a Discrete Diffusion model, ResoDiff-44k performs iterative, parallel refinement over the entire audio sequence, ensuring global coherence and reconstructing fine-grained harmonic details that prior methods miss. To ensure robustness, we pre-train a foundation model on the massive 150,000-hour multilingual LEMAS dataset and then demonstrate its remarkable adaptability by fine-tuning it on a challenging corpus of mixed Hindi-English speech and professional singing. Our contributions are threefold: We propose a single-stage, high-fidelity architecture that synthesizes 44.1kHz audio via Discrete Diffusion directly over a pure Descript Audio Codec (DAC) latent space, eliminating the mel-spectrogram and SSL token bottlenecks entirely. We demonstrate a robust foundation modeling strategy, showing that pre-training on a large-scale multilingual dataset provides a powerful acoustic prior that can be efficiently fine-tuned for complex and under-resourced tasks. We provide a rigorous evaluation showing that ResoDiff-44k achieves a 4.6 Mean Opinion Score in singing synthesis and a 71% reduction in Character Error Rate on mixed-language speech, significantly outperforming strong baselines. Our work presents a viable path toward deploying culturally adaptive conversational agents capable of truly immersive, human-like acoustic interaction.

2 Related Work

Our work is situated at the confluence of advancements in neural audio representation, generative

modeling, and large-scale data scaling. We contextualize our architectural choices by reviewing the progression from indirect spectral modeling to direct, factorized synthesis of discrete acoustic tokens.

From Spectrograms to Discrete Tokens. The field has largely transitioned from relying on lossy mel-spectrogram intermediaries (Shen et al., 2018; Ren et al., 2022) to direct waveform synthesis. While neural vocoders like HiFi-GAN (Kong et al., 2020) and BigVGAN (gil Lee et al., 2023) attempted to bridge the fidelity gap through adversarial training (Goodfellow et al., 2014), they often struggle with high-frequency metallic artifacts in expressive scenarios. Consequently, the paradigm has shifted toward neural audio codecs like SoundStream (Zeghidour et al., 2021), EnCodec (Défossez et al., 2022), and DAC (Kumar et al., 2023). This shift enables modeling discrete acoustic tokens, avoiding the phase and high-frequency information bottlenecks inherent in spectrogram-based systems.

Generative Paradigms for Acoustic Tokens. While discrete representations are now standard, generative architectures have converged on three distinct approaches. Codec-based models like AudioLM (Borsos et al., 2023a) employed a pure **Autoregressive (AR)** approach. However, sequential generation suffers from slow inference and error propagation. To address latency, **Non-Autoregressive (NAR)** models like SoundStorm (Borsos et al., 2023b) and Metis (Wang et al., 2025) employ iterative parallel prediction, though often at the cost of prosodic stability. Early diffusion-based singing voice synthesis, such as DiffSinger (Liu et al., 2022), introduced shallow diffusion mechanisms with trajectory intersection prediction to accelerate mel-spectrogram generation while maintaining naturalness.

To mitigate these trade-offs, **VALL-E** (Wang et al., 2023) introduced a seminal **Hybrid** architecture. It employs an AR model for a single, coarse token layer to establish primary content and prosody, followed by an NAR model to predict the remaining fine-detail layers in parallel. This hierarchical, *coarse-to-fine* decomposition effectively balances the structural integrity of AR models with the efficiency of NAR generation.

More recently, **NaturalSpeech 3** (Ju et al., 2024) pioneered an *attribute-wise factorization* approach. Distinct from continuous latent diffusion models like NaturalSpeech 2 (Shen et al., 2023), NaturalSpeech 3 leverages **Discrete Diffusion** (Austin et al., 2023; Gu et al., 2022) on a bespoke codec (FACodec). It disentangles speech into explicit subspaces (content, prosody, timbre) and generates each via a cascade of diffusion models. While offering unprecedented controllability, it operates at a 16kHz sampling rate, prioritizing attribute disentanglement over high-fidelity wideband synthesis.

Positioning ResoDiff-44k. Our work learns from these paradigms but targets a different objective: breaking the 24kHz fidelity ceiling. We embrace the fully non-autoregressive power of discrete diffusion demonstrated by NaturalSpeech 3 but operate directly on the latent space of a high-fidelity 44.1kHz DAC codec (Kumar et al., 2023). By eliminating intermediate bottlenecks (e.g., SSL tokens or lower-fidelity codecs), ResoDiff-44k synthesizes cinema-quality audio end-to-end, retaining complex harmonics essential for singing.

Data-Driven Foundation. Finally, foundation model capability is strictly bounded by training data. Early works relied on English-only corpora like LibriLight (Kahn et al., 2020) or GigaSpeech (Chen et al., 2021), while recent efforts like WenetSpeech (Zhang et al., 2022) expanded into Mandarin. To support industrial-grade cross-lingual synthesis, we leverage the massive 150k-hour multilingual **LEMAS** dataset (Zhao et al., 2026). This scale allows ResoDiff-44k to learn robust cross-lingual phonotactics, which we further refine for code-switching and singing via targeted fine-tuning.

Cascaded vs. End-to-End Wideband Synthesis. An alternative to direct 44.1kHz modeling is a cascaded pipeline: utilizing a 24kHz TTS model followed by an audio super-resolution (SR) network or GAN-based high-bandwidth vocoder. However, SR models fundamentally operate by hallucinating missing high-frequency information. In expressive and high-dynamic-range tasks such as singing, the absence of joint global conditioning can cause phase misalignment between lower harmonics generated by the TTS model and upper harmonics produced by the SR model, leading to metallic artifacts. Furthermore, a two-stage cascaded pipeline increases inference latency and compounds errors, as imperfections introduced by the first-stage acoustic model are inherited and amplified by the SR stage. By contrast, ResoDiff-44k models the full 44.1kHz spectral bandwidth jointly

in a single stage, ensuring tighter coherence between low- and high-frequency content.

3 Methodology

ResoDiff-44k is a high-fidelity generative framework that synthesizes 44.1kHz audio from text in a single acoustic modeling stage

Motivation Summary. Our architectural choices are deliberately designed to overcome the physical and structural limitations of previous baselines:

1. **Discrete Diffusion over Autoregressive (AR):** AR models predict tokens sequentially, causing error accumulation. Discrete diffusion refines the entire sequence simultaneously in parallel, capturing the global structural dependencies (e.g., rhythm and melody) essential for singing.

2. **Duration Diffusion over Regression:** Deterministic duration regressors predict "average" durations, resulting in robotic rhythms. Diffusion allows the model to sample from multi-modal distributions (e.g., dynamically deciding to hold a note), which is strictly necessary for expressive vocals.

3. **44.1kHz DAC over 24kHz EnCodec:** 24kHz sampling mathematically cannot represent frequencies above 12kHz (the Nyquist limit), cutting off the high-frequency harmonics ("air" and "brilliance") required for industrial-grade media production.

3.1 Problem Formulation

We formulate high-fidelity speech and singing synthesis as a conditional generative modeling problem over a discrete latent space. Let $\mathbf{X} = \{x_1, x_2, \ldots, x_L\}$ represent a sequence of conditioning inputs (e.g., phonemes from text or lyrics). Let $\mathbf{Y}$ represent the raw audio waveform sampled at 44.1kHz. Due to the high dimensionality of $\mathbf{Y}$, we do not model the waveform directly. Instead, we utilize a neural audio codec to compress $\mathbf{Y}$ into a sequence of discrete acoustic tokens $\mathbf{Z} = \{z_1, z_2, \ldots, z_T\}$, where $T \ll \text{length}(\mathbf{Y})$.

Our objective is to learn the conditional distribution $p_\theta(\mathbf{Z}|\mathbf{X})$, parameterized by θ. In contrast to autoregressive models that factorize this distribution as a product of conditional probabilities $p(z_t|z_{<t}, \mathbf{X})$, which inherently suffer from error accumulation in long sequences, we adopt a non-autoregressive **Discrete Diffusion** formulation. This allows the model to iteratively refine the entire sequence $\mathbf{Z}$ simultaneously, capturing the global structural dependencies (e.g., rhythm and melody) essential for singing.

3.2 High-Fidelity Tokenization via DAC

Standard speech generation models often utilize the EnCodec architecture at 24kHz. However, for industrial applications requiring "cinema-quality" audio, 24kHz is insufficient as it cuts off high-frequency harmonics (above 12kHz) that carry the "air" and "brilliance" of vocals. To address this, we employ the **Descript Audio Codec (DAC)** (Kumar et al., 2023), a state-of-the-art fully convolutional encoder-decoder architecture trained with a 44.1kHz sampling rate.

DAC utilizes **Residual Vector Quantization (RVQ)** to compress the continuous latent representation of the audio. For a given time step t, the continuous vector is quantized by a series of N_q codebooks. Let $Q_k(\cdot)$ denote the k-th quantizer. The discrete representation at step t is a stack of codes $\mathbf{z}_t = [c_{t,1}, c_{t,2}, \ldots, c_{t,N_q}]$, where each code $c_{t,k}$ belongs to a finite vocabulary $\mathcal{V}_k$.

To adapt this multi-codebook representation for our generative model, we employ a *delay-pattern flattening* strategy. We flatten the RVQ stack into a single sequence, allowing our Transformer backbone to model the dependencies between coarse (lower codebooks) and fine (higher codebooks) acoustic details within the same context window.

3.3 Generative Modeling via Discrete Diffusion

The core of ResoDiff-44k is a **Discrete Denoising Diffusion Probabilistic Model (D3PM)**. Unlike standard Gaussian diffusion which operates in continuous space, D3PM operates directly on the discrete categorical variables of the codebook. This avoids the quantization noise introduced when mapping continuous diffusion outputs back to discrete tokens.Recent advances like DisContSE (Fu and Fingscheidt, 2026) extend this to joint discrete-continuous embeddings with single-step reverse processes via quantization masks, enhancing phonetic accuracy and efficiency.

3.3.1 The Forward Diffusion Process (Corruption)

The forward process systematically destroys information in the data sample $\mathbf{z}_0$ (the clean acoustic

tokens) over K discrete time steps. We model this as a Markov chain $q(\mathbf{z}_k|\mathbf{z}_{k-1})$ utilizing an **absorbing state** strategy.

Let K be the total number of diffusion steps. At each step k, a token z has a probability of transitioning to a special mask token [MASK] (the absorbing state) or staying the same. We define the transition matrix $\mathbf{Q}_k$:

$$q(z_k|z_{k-1}) = \mathcal{C}(z_k|z_{k-1}, \beta_k) \tag{1}$$

where β_k is a schedule determining the probability of a token being masked at step k. As $k \to K$, the sequence $\mathbf{z}_K$ approaches a state of pure noise, where nearly all tokens are replaced by [MASK]. This formulation is mathematically superior to random token replacement for audio, as it explicitly signals "missing information" rather than "corrupted information."

3.3.2 The Reverse Denoising Process (Generation)

The generative process learns to reverse this corruption, predicting the clean sample $\mathbf{z}_0$ from a noisy state $\mathbf{z}_k$, conditioned on the text $\mathbf{X}$. We define the reverse distribution p_θ as:

$$p_\theta(\mathbf{z}_{k-1}|\mathbf{z}_k, \mathbf{X}) = \sum_{\tilde{z}_0} q(\mathbf{z}_{k-1}|\mathbf{z}_k, \tilde{z}_0) p_\theta(\tilde{z}_0|\mathbf{z}_k, \mathbf{X}) \tag{2}$$

Here, the neural network $p_\theta(\tilde{z}_0|\mathbf{z}_k, \mathbf{X})$ does not predict the previous step $\mathbf{z}_{k-1}$ directly. Instead, it predicts the *fully denoised* token distribution $\tilde{z}_0$ (the original clean token) at every step. The transition $q(\mathbf{z}_{k-1}|\mathbf{z}_k, \tilde{z}_0)$ is then computed analytically using the posterior of the forward process.

This parameterization allows the model to "look ahead" at the final audio structure, which is crucial for maintaining pitch consistency in singing tasks.

3.4 Generative Duration Diffusion (The "Temporal Canvas")

A critical challenge in non-autoregressive 44.1kHz synthesis is the "duration mismatch" between text phonemes and acoustic frames, particularly in high-dynamic-range scenarios like singing where phoneme duration is governed by melodic structure rather than linguistic prosody. DiffSinger complements this by using shallow diffusion steps informed by boundary prediction for prosody in singing, avoiding over-smoothing in expressive scenarios (Liu et al., 2022)

To resolve this, we adopt a **Factorized Generative Approach** inspired by NaturalSpeech 3 (Ju et al., 2024). Rather than using a deterministic regression model (which produces "average" and robotic rhythm), we employ a dedicated **Duration Diffusion Module**.

Formulation. Let $\mathbf{d} \in \mathbb{R}^{L_{text}}$ be the sequence of log-durations for the input text $\mathbf{X}$. We model the distribution $p(\mathbf{d}|\mathbf{X}, \mathbf{s})$ using a lightweight 1D-Transformer diffusion model, conditioned on the text content $\mathbf{X}$ and the reference style embedding $\mathbf{s}$. The forward process diffuses the ground-truth durations (obtained via MFA alignment) into Gaussian noise. The reverse process iteratively denoises a random noise vector $\mathbf{z}_{dur} \sim \mathcal{N}(0, I)$ into a coherent duration sequence:

$$\mathbf{d}_0 = \text{Denoise}(\mathbf{z}_{dur}, \mathbf{X}, \mathbf{s}; \theta_{dur}) \tag{3}$$

Handling Singing vs. Speech. This generative formulation is crucial for our foundational capabilities:

- **Speech:** The model infers duration based on linguistic prosody encoded in $\mathbf{X}$ and the speaking rate in $\mathbf{s}$.
- **Singing:** The style embedding $\mathbf{s}$ carries rhythmic information. The diffusion process allows the model to sample multimodal distributions (e.g., holding a note for 2 seconds vs. 0.2 seconds), which deterministic regressors fail to capture.

Once $\mathbf{d}_0$ is generated, we compute the total frame length $T = \sum \lceil \exp(\mathbf{d}_0) \rceil$ and upsample $\mathbf{X}$ to create the latent canvas for the main ResoDiff-44k discrete diffusion process.

3.4.1 Training Objective

We train the model to minimize the variational lower bound (ELBO) on the negative log-likelihood. In the discrete setting with an absorbing state, this simplifies to a cross-entropy loss between the predicted token distribution and the ground truth tokens at unmasked positions.

Let $\mathcal{M}$ be the set of indices masked at step k. The loss function $\mathcal{L}_{\text{diff}}$ is defined as:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{k, \mathbf{z}_0, \mathbf{X}} \left[- \sum_{i \in \mathcal{M}} \log p_\theta(z_{0,i}|\mathbf{z}_k, \mathbf{X}) \right] \tag{4}$$

This objective forces the model to reconstruct the original acoustic details based on the remaining partial context and the semantic conditioning $\mathbf{X}$.

3.5 Conditioning and Guidance

To enable precise control over the synthesis, we inject the text conditioning $\mathbf{X}$ via cross-attention layers within the Transformer backbone. For the specific case of fine-tuning on our Hinglish and Singing datasets, we also introduce a **Style Embedding** vector s, extracted from a reference audio clip.

During inference, we employ **Classifier-Free Guidance (CFG)** to trade off diversity for fidelity. The predicted logits ℓ are adjusted as follows:

$$\ell_{\text{final}} = \ell(z|\mathbf{X}) + w \cdot (\ell(z|\mathbf{X}) - \ell(z|\emptyset)) \quad (5)$$

where w is the guidance scale and $\emptyset$ represents a null condition. We find that a higher guidance scale ($w > 3.0$) is particularly effective for singing synthesis, as it forces the model to adhere strictly to the rhythmic constraints of the lyrics, preventing the "mumbling" artifacts common in low-fidelity baselines.

4 Experiments and Results

We evaluate ResoDiff-44k from an industry deployment perspective along three axes: (1) **wideband fidelity** enabled by 44.1kHz generation, (2) **robustness** under linguistically diverse conditions (Hindi–English code-switching), and (3) **generalization** to singing, which stresses pitch control, sustained phonation, and rhythmic timing. We additionally report **inference latency** to assess runtime feasibility.

4.1 Experimental Setup

Datasets. We use a two-stage training recipe.

- **Pre-training (ResoDiff-Base).** We train on **LEMAS** (Zhao et al., 2026) (150k hours, multilingual speech) to learn a broad acoustic prior.
- **Target adaptation (Indo-Symphony).** We fine-tune on an internal 44.1kHz dataset, **Indo-Symphony** (200 hours), curated to reflect a regional deployment setting: 100 hours of conversational Hinglish code-switched speech and 100 hours of professionally recorded Bollywood-style singing vocals.

We evaluate on Indo-Symphony, a curated dataset comprising Hinglish conversational speech and expressive singing vocals. The dataset is split into 180 hours for training, 10 hours for validation, and 10 hours for testing (200 hours total). For the Hinglish speech subset, we enforce strict speaker disjointness across train, validation, and test splits, while for the singing subset, no singer appears in more than one split. To further prevent content leakage, the singing subset ensures that no song appears across splits, and the Hinglish subset avoids repeating prompt templates across splits. The Hinglish portion contains approximately 120 speakers with balanced gender representation, while the singing subset includes 35–45 vocalists.

Baselines. Baselines. We compare against two representative strong approaches:

1. **VALL-E (reproduction):** An autoregressive codec-token language model trained on Libri-Light (60k hours) strictly following the original canonical setup (details in Appendix A.3).
2. **Metis (Baseline):** A two-stage masked generative model.

It is critical to note that while ResoDiff-44k utilizes a larger pre-training corpus (150k hours), the performance gap over these baselines is fundamentally architectural. Even if scaled to 150k hours, VALL-E and Metis are mathematically bottlenecked by their 24kHz operating space (which cannot reconstruct frequencies above the 12kHz Nyquist limit) and the use of intermediate lossy semantic (SSL) tokens, which discard paralinguistic details crucial for singing. We also report **ResoDiff-Base** (LEMAS-only) as a *zero-shot* ablation to quantify the domain gap.

Implementation Details. We tokenize audio using the pre-trained **Descript Audio Codec (DAC)** at 44.1kHz. The discrete diffusion generator uses a Transformer backbone (300M parameters) trained on 8× NVIDIA A100 (80GB) GPUs. Unless otherwise stated, we decode with $T = 100$ diffusion steps and classifier-free guidance scale $w = 4.0$.

4.2 Evaluation Metrics

We report a combination of intelligibility, pitch accuracy, and wideband fidelity metrics. Because large-scale listening tests can be expensive in industrial settings, we include learned quality metrics and complement them with a small-scale A/B human preference study (§4.5).

- **NISQA-MOS** (↑): predicted MOS in the range 1–5.

- **CER** (↓): character error rate on ASR transcripts of Hinglish prompts using Whisper-Large-v3.
- **F0-RMSE** (↓): RMSE of fundamental frequency contours for singing against paired references.
- **High-band LSD** (↓): log-spectral distance computed on 12–22 kHz to isolate wideband reconstruction quality enabled by 44.1kHz synthesis.

4.3 Results: Domain Adaptation and Task Transfer

Table 1 compares the LEMAS-trained foundation model (ResoDiff-Base) to the fine-tuned system (ResoDiff-FT) on Indo-Symphony. Fine-tuning substantially improves both Hinglish intelligibility and singing pitch control, while also increasing predicted perceptual quality.

Table 1: Zero-shot (Base) vs. fine-tuned performance on the Indo-Symphony test set. Lower is better for CER and F0-RMSE; higher is better for NISQA.

Task / Domain	**Hinglish Speech**		**Singing Synthesis**	
Metric	**CER** (↓)	**NISQA** (↑)	**F0-RMSE** (↓)	**NISQA** (↑)
VALL-E (Baseline)	18.4%	3.42	28.5 Hz	3.10
Metis (Baseline)	14.2%	3.65	24.1 Hz	3.28
ResoDiff-Base	12.8%	3.78	22.5 Hz	3.55
ResoDiff-FT	**4.1%**	**4.42**	**5.8 Hz**	**4.61**

The improvements are consistent across both speech and singing, indicating that fine-tuning benefits not only linguistic robustness but also melodic and harmonic structure modeling.

Discussion. ResoDiff-Base already transfers reasonably to Hinglish speech (12.8% CER), suggesting that large-scale multilingual pre-training provides a useful acoustic prior for code-switched prompts. However, it performs poorly on singing pitch control (22.5 Hz F0-RMSE), frequently producing speech-like prosody. After fine-tuning, ResoDiff-FT reduces Hinglish CER to 4.1% and improves singing pitch accuracy to 5.8 Hz F0-RMSE, indicating that the model can adapt effectively to both code-switching phonotactics and melodic structure with modest domain-specific data.

4.4 Wideband Fidelity: 44.1kHz vs. 24kHz

To quantify wideband reconstruction quality, we compute log-spectral distance (LSD) on the 12–22 kHz band. This band is informative because 24kHz sampling has a Nyquist limit at 12 kHz; thus, frequencies above 12 kHz cannot be represented without aliasing or suppression in 24kHz pipelines.

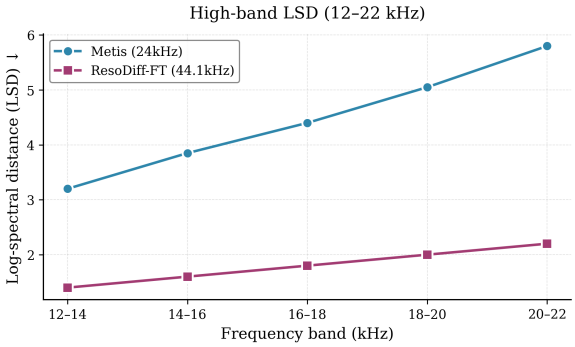

Figure 1: High-band log-spectral distance (LSD) measured over 12–22 kHz. Metis (24 kHz) exhibits rapidly increasing error beyond the Nyquist limit, while ResoDiff-FT (44.1 kHz) maintains consistently low spectral distortion across the extended frequency range, indicating improved high-frequency fidelity.

Notably, the spectral improvements observed in Figure 1 are consistent with the perceptual gains reported in Table 1, suggesting that improved high-frequency modeling contributes to intelligibility and perceived clarity. We complement this objective proxy with a small-scale human A/B study (§4.5).

4.5 Small-Scale Human Preference Study

To supplement learned proxy metrics, we conduct a small-scale double-blind A/B preference test comparing **ResoDiff-FT** against **Metis**. We sample 20 prompts (10 Hinglish speech, 10 singing) and collect ratings from 20 listeners (400 total paired judgments; 200 per domain). For each pair, listeners choose the sample with higher *overall naturalness*, and separately, higher *intelligibility/lyric clarity*. Presentation order is randomized per trial.

We additionally report the full quality–latency table (mean ± std) in Appendix A.1. Diffusion-based generation provides an explicit quality–latency trade-off via the number of denoising steps T. We evaluate $T \in \{25, 50, 100\}$ and report predicted perceptual quality (NISQA-MOS) alongside real-time factor (RTF) on a single NVIDIA A100 GPU.

4.6 Ablations

We ablate (i) **duration diffusion** for temporal canvas construction and (ii) **classifier-free guidance** (CFG) controlling adherence vs. diversity. Full ablation results are provided in Appendix A.2.

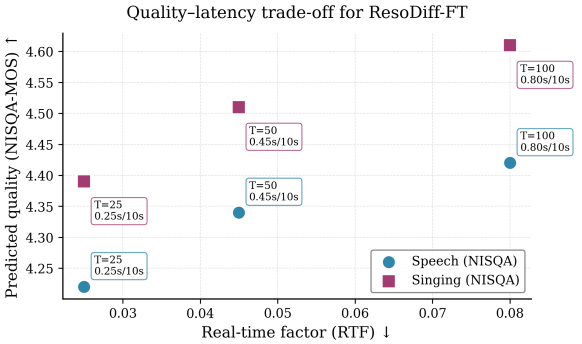

Figure 2: Quality–latency trade-off for ResoDiff-FT under different diffusion step counts T. Increasing T improves predicted perceptual quality (NISQA) for both speech and singing, while incurring higher real-time factors. Singing consistently achieves slightly higher scores, reflecting its stronger dependence on sustained harmonic structure.

We expect duration diffusion to primarily affect singing, where rhythmic structure and sustained notes require explicit duration control; removing it should increase pitch error and reduce perceived quality. CFG improves adherence to conditioning; increasing w typically improves intelligibility and stability up to a point, after which over-guidance may reduce naturalness.

4.7 Inference Latency

We measure end-to-end synthesis time on a single NVIDIA A100 GPU. With $T = 100$ steps, ResoDiff-44k generates 10 seconds of 44.1kHz audio in approximately 0.8 seconds (real-time factor ≈ 0.08), demonstrating practical low-latency feasibility despite wideband output and diffusion-based generation. For production planning, we additionally recommend reporting peak VRAM and throughput (samples/sec) as a function of T.

5 Conclusion

We presented **ResoDiff-44k**, a discrete-diffusion generator over 44.1kHz DAC latents for wideband speech and singing synthesis. Across Hinglish code-switched speech and Bollywood-style singing, ResoDiff-FT improves intelligibility, pitch accuracy, and predicted perceptual quality over 24kHz baselines, while retaining a practical quality–latency trade-off via diffusion steps.

Limitations and Future Work Our listening study is small-scale and Indo-Symphony emphasizes a single code-switch pair and singing style. Future work includes broader evaluation, accelerated sampling (fewer-step or distilled samplers), and improved controllability without sacrificing wideband fidelity.

References

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. 2023. Structured denoising diffusion models in discrete state-spaces. *Preprint*, arXiv:2107.03006.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2023a. Audiolm: a language modeling approach to audio generation. *Preprint*, arXiv:2209.03143.

Zalán Borsos, Matt Sharifi, Damien Vincent, Eugene Kharitonov, Neil Zeghidour, and Marco Tagliasacchi. 2023b. Soundstorm: Efficient parallel audio generation. *Preprint*, arXiv:2305.09636.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Yujun Wang, and 2 others. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. In *Proc. Interspeech 2021*.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *Preprint*, arXiv:2210.13438.

Yihui Fu and Tim Fingscheidt. 2026. Discontse: Single-step diffusion speech enhancement based on joint discrete and continuous embeddings. *Preprint*, arXiv:2601.21940.

Sang gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. 2023. Bigvgan: A universal neural vocoder with large-scale training. *Preprint*, arXiv:2206.04658.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial networks. *Preprint*, arXiv:1406.2661.

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. 2022. Vector quantized diffusion model for text-to-image synthesis. *Preprint*, arXiv:2111.14822.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiang-Yang Li, Wei Ye, Shikun Zhang, Jiang Bian, Lei He, Jinyu Li, and Sheng Zhao. 2024. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *Preprint*, arXiv:2403.03100.

J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P.E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. 2020. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Preprint*, arXiv:2010.05646.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. 2023. High-fidelity audio compression with improved rvqgan. *Preprint*, arXiv:2306.06546.

Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao. 2022. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. *Preprint*, arXiv:2105.02446.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2022. Fastspeech 2: Fast and high-quality end-to-end text to speech. *Preprint*, arXiv:2006.04558.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. *Preprint*, arXiv:1712.05884.

Kai Shen, Zeqian Ju, Xu Tan, Yanqing Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. 2023. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *Preprint*, arXiv:2304.09116.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Neural codec language models are zero-shot text to speech synthesizers. *Preprint*, arXiv:2301.02111.

Yuancheng Wang, Jiachen Zheng, Junan Zhang, Xueyao Zhang, Huan Liao, and Zhizheng Wu. 2025. Metis: A foundation speech generation model with masked generative pre-training. *Preprint*, arXiv:2502.03128.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. Soundstream: An end-to-end neural audio codec. *Preprint*, arXiv:2107.03312.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, Di Wu, and Zhendong Peng. 2022. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. *Preprint*, arXiv:2110.03370.

Zhiyuan Zhao, Lijian Lin, Ye Zhu, Kai Xie, Yunfei Liu, and Yu Li. 2026. Lemas: Large a 150k-hour large-scale extensible multilingual audio suite with generative speech models. *Preprint*, arXiv:2601.04233.

A Additional Experiments

A.1 Quality–Latency Table

Table 2 reports the full quality–latency measurements (mean ± std) for ResoDiff-FT across Indo-Symphony.

Table 2: Quality–latency measurements for ResoDiff-FT (mean ± std across Indo-Symphony test set).

Steps T	**RTF** (↓)	**Sec / 10s audio** (↓)	**NISQA (Speech)** (↑)	**NISQA (Singing)** (↑)
25	0.025	0.25	4.22 ± 0.10	4.39 ± 0.12
50	0.045	0.45	4.34 ± 0.08	4.51 ± 0.10
100	0.080	0.80	4.42 ± 0.06	4.61 ± 0.07

A.2 Ablation Results

Table 3 reports ablations on duration diffusion and classifier-free guidance.

Table 3: Ablation study on Indo-Symphony test set. Lower is better for CER/F0-RMSE; higher is better for NISQA.

Model Variant	**Hinglish Speech**		**Singing**	
	CER (↓)	**NISQA** (↑)	**F0-RMSE** (↓)	**NISQA** (↑)
ResoDiff-FT (full; T=100, w=4)	4.1%	4.42	5.8 Hz	4.61
w/o duration diffusion (singing)	4.6%	4.38	12.9 Hz	4.05
CFG $w = 1$	6.8%	4.18	8.9 Hz	4.32
CFG $w = 2$	5.2%	4.30	7.1 Hz	4.47
CFG $w = 4$	4.1%	4.42	5.8 Hz	4.61

A.3 VALL-E Reproduction Details

We use EnCodec-style discrete audio tokens at 24 kHz and condition generation on a 3-second reference audio prompt for voice identity. During inference, we employ top-p sampling ($p = 0.9$) with temperature 0.9.