MimicLM: Zero-Shot Voice Imitation through Autoregressive Modeling of Pseudo-Parallel Speech Corpora

Tao Feng[1], Yuxiang Wang[2], Yuancheng Wang[2], Xueyao Zhang[2], Dekun Chen[2], Chaoren Wang[2], Xun Guan[1], Zhizheng Wu[2†]

[1]Tsinghua University, [2]The Chinese University of Hong Kong, Shenzhen

Abstract

Voice imitation aims to transform *source* speech to match a *reference* speaker's timbre and speaking style while preserving linguistic content. A straightforward approach is to train on triplets of *(source, reference, target)*, where *source* and *target* share the same content but *target* matches the *reference*'s voice characteristics, yet such data is extremely scarce. Existing approaches either employ carefully designed disentanglement architectures to bypass this data scarcity or leverage external systems to synthesize pseudo-parallel training data. However, the former requires intricate model design, and the latter faces a quality ceiling when synthetic speech is used as training *targets*. To address these limitations, we propose MimicLM, which takes a novel approach by using synthetic speech as training *sources* while retaining real recordings as *targets*. This design enables the model to learn directly from real speech distributions, breaking the synthetic quality ceiling. Building on this data construction approach, we incorporate interleaved text-audio modeling to guide the generation of content-accurate speech and apply post-training with preference alignment to mitigate the inherent distributional mismatch when training on synthetic data. Experiments demonstrate that MimicLM achieves superior voice imitation quality with a simple yet effective architecture, significantly outperforming existing methods in naturalness while maintaining competitive similarity scores across speaker identity, accent, and emotion dimensions.

1 Introduction

Voice imitation aims to transform *source* speech to match a *reference* speaker's voice characteristics while preserving the linguistic content. Unlike voice conversion that focuses solely on timbre transfer (Sisman et al., 2020), voice imitation additionally captures the complete speaking style including prosodic and expressive patterns (Liu et al., 2020; Zhou et al., 2022; Zhang et al., 2025).

A straightforward approach is to train on triplets (*source*, *reference*, *target*), where *source* and *target* share the same linguistic content but differ in speaker identity, while *reference* provides the target speaker's voice characteristics. However, such triplet data is extremely scarce in real-world speech corpora (Yoshino et al., 2016), and manual collection at scale remains prohibitively expensive.

Existing methods address this parallel data bottleneck through two primary strategies. The first approach bypasses the need for parallel data by explicitly disentangling content, timbre, and prosody through specialized architectural components (Qian et al., 2019, 2020; Ju et al., 2024; Zhang et al., 2025). While effective, these methods require multi-stage training with carefully balanced objectives and complex inference pipelines involving multiple learned modules (Ju et al., 2024; Łajszczak et al., 2024). The second approach leverages external zero-shot text-to-speech (TTS) (Wang et al., 2023a; Anastassiou et al., 2024; Du et al., 2024; Wang et al., 2024) or voice conversion (VC) (Qin et al., 2023) systems to construct pseudo-parallel training pairs by generating synthetic speech with different speaker characteristics, creating (*real source*, *real reference*, *synthetic target*) triplets (Li et al., 2025; Tu et al., 2025). However, since the model learns to reproduce these synthetic targets, its output quality is inherently bounded by the external system's capabilities. Recent advances attempt to preserve real speech as targets to overcome this ceiling. SeedVC (Liu, 2024) employs an external voice conversion system (Qin et al., 2023) to generate timbre-perturbed inputs paired with real *targets*, but initially focuses solely on timbre transfer rather than complete speaking

† Corresponding author.

Resources available at https://fff-ttt.github.io/MimicLM_demo/

Findings of the Association for Computational Linguistics: ACL 2026, pages 20392–20405

style imitation, and requires an external model during training. SynthVC (Guo et al., 2025) trains on synthetic sources with real *targets*, yet is limited to a fixed set of speakers and exhibits lower naturalness than its external VC system (Liu, 2024). These limitations motivate the need for a zero-shot voice imitation approach that learns from real speech distributions without sacrificing naturalness.

To address these limitations, we introduce a novel data construction strategy: inverting the role of synthetic speech from training targets to training sources. Specifically, given a real utterance spoken by speaker A, we use a TTS model (Du et al., 2024) to re-synthesize the same linguistic content in a different voice (speaker B). We then construct training triplets where the synthetic utterance serves as the *source*, another real recording from speaker A serves as the *reference*, and the original real utterance becomes the *target*. The model learns to transform the synthetic source to match the reference speaker's voice while preserving content. This construction is valid because the source and target share identical textual content by design. By learning to generate real human speech rather than synthetic outputs, MimicLM breaks the quality ceiling that constrains methods trained on synthetic *targets*. Moreover, since both *reference* and *target* come from real recordings of the same speaker, the model implicitly learns to capture and transfer voice characteristics without requiring explicit feature extraction or disentanglement. However, this inversion introduces new challenges: training on synthetic inputs while performing inference on real inputs creates a distributional gap that leads to systematic performance degradation in real-world conditions. Additionally, voice imitation itself is inherently more challenging than timbre-only voice conversion: transforming both timbre and prosodic patterns simultaneously alters the temporal structure of speech, making it harder to preserve linguistic content and often resulting in higher word error rates (WER). This elevation is an inherent characteristic of voice imitation, also observed in prior work (Zhang et al., 2025).

To address these challenges, we propose MimicLM, an end-to-end voice imitation model incorporating two key techniques. First, to mitigate content corruption when transforming complete speaking styles, we incorporate interleaved text-audio modeling: text tokens are interleaved with audio tokens in the input sequence, providing explicit content anchors that guide the model to preserve semantic information during voice transformation. Second, to bridge the synthetic-to-real gap (Su et al., 2024), we apply preference alignment during post-training. Specifically, we conduct post-training by feeding the model real *source-reference* pairs, sampling multiple candidate outputs, and ranking them by WER. These ranked candidates form preference pairs that guide the model to generate content-faithful, natural outputs when processing real speech, effectively adapting it from the synthetic-input training regime to real-world conditions.

Experimental results demonstrate that MimicLM achieves strong naturalness while maintaining competitive similarity to state-of-the-art voice imitation systems. Notably, preference alignment yields a significant reduction in WER on real inputs, effectively bridging the distributional gap. Furthermore, scaling analysis reveals consistent improvements as data volume increases, suggesting substantial potential for further gains with larger datasets. Our main contributions are: (1) We propose a role-swapping data construction strategy that uses TTS-generated speech as *source* inputs and real recordings as *targets*, enabling scalable training while learning from high-quality real speech distributions. (2) We introduce interleaved text-audio modeling and preference alignment to address the intelligibility degradation and distributional gap inherent in role-swapped training. (3) Through extensive evaluation, we demonstrate that this unified framework achieves competitive performance across naturalness, intelligibility, and similarity, offering a conceptually simpler alternative to complex disentanglement-based architectures.

2 Related Work

2.1 From Timbre to Voice Imitation

Voice conversion (VC) traditionally transforms only speaker timbre while preserving content and prosody (Sisman et al., 2020), whereas voice imitation (VI) reproduces both timbre and speaking style (Zhang et al., 2025).

Early VI work used sequence-to-sequence models (Zhang et al., 2019; Wang et al., 2023b) or explicit prosody modeling (Choi et al., 2023; Lee et al., 2025). Recently, zero-shot TTS (Wang et al., 2023a; Anastassiou et al., 2024; Wang et al., 2024; Ju et al., 2024; Du et al., 2024) has become dominant but requires external ASR (Radford et al., 2022) for speech-to-speech scenarios, introducing

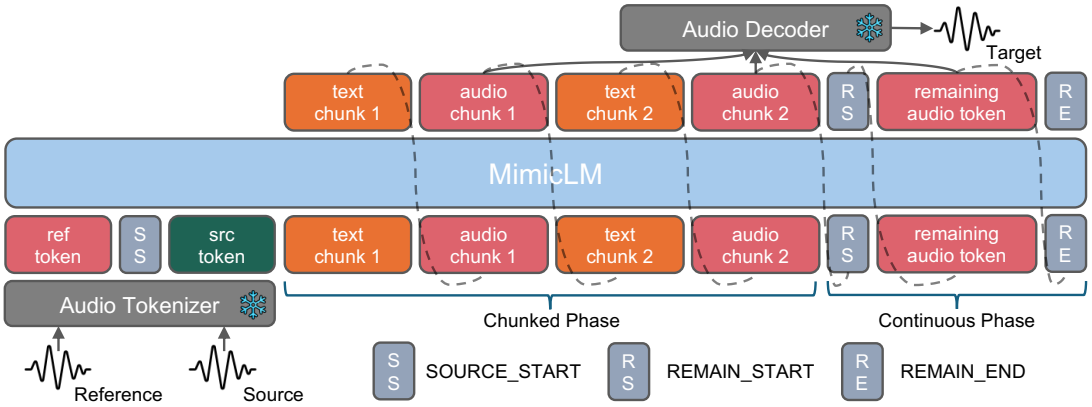

Figure 1: Overview of MimicLM architecture for voice imitation. The model processes *reference* audio (*ref token*) for target timbre/style and *source* audio (*src token*), then generates the conversion *target* in two phases: a chunked phase with interleaved text-audio prediction, followed by a continuous phase for remaining speech tokens. Both the audio tokenizer and decoder are frozen during training (snowflake icons). Special control tokens are annotated at bottom right.

latency and error propagation.

Recent work has revisited VI from the VC perspective through direct speech-to-speech transformation. Vevo (Zhang et al., 2025) achieves state-of-the-art results using VQ-VAE information bottlenecks (Van Den Oord et al., 2017). Similarly, SeedVC (Liu, 2024) introduced a new version with explicit quantization and ASR-supervised feature learning to better support VI tasks. We explore whether simpler end-to-end VC architectures can achieve voice imitation by rethinking training data construction and model design.

2.2 Addressing Parallel Data Scarcity

A fundamental bottleneck in training speech-to-speech conversion models is the scarcity of naturally parallel data. Recent approaches construct pseudo-parallel training pairs through two strategies.

Learning from Synthetic Targets. StarVC (Li et al., 2025) employs a VC system (Qin et al., 2023) to generate speaker variations, while O_O-VC (Tu et al., 2025) uses multi-speaker TTS (Kim et al., 2021) to synthesize paired speech from the same linguistic content. While enabling scalable training, these approaches inherit quality limitations from their external synthesis systems.

Learning from Real Targets. Recent work preserves real speech as training *targets*. SeedVC (Liu, 2024) applies timbre perturbation via external VC (Qin et al., 2023), reducing timbre leakage but requiring auxiliary models. SynthVC (Guo et al., 2025) trains on synthetic sources with real *targets* but supports only fixed speaker sets rather than zero-shot conversion. We instead propose zero-shot voice imitation that learns from real *targets* while capturing both timbre and speaking style, achieving natural output unconstrained by external TTS quality.

2.3 Disentanglement or End-to-End Learning

Current approaches rely on explicit disentanglement through knowledge distillation (Polyak et al., 2021; Ju et al., 2024), information bottlenecks (Qian et al., 2019; Zhang et al., 2025), or acoustic perturbation (Choi et al., 2021). However, complete disentanglement remains difficult and requires multi-stage training with adversarial objectives (Ju et al., 2024; Łajszczak et al., 2024).

The availability of pseudo-parallel data enables end-to-end learning that bypasses explicit disentanglement. Large-scale models (Wang et al., 2023a; Le et al., 2023; Anastassiou et al., 2024) demonstrate that in-context learning can implicitly capture speaker characteristics when sufficient paired data is available.

3 MimicLM

3.1 Overview

As illustrated in Figure 1, our system consists of three components: (1) a frozen audio tokenizer that converts waveforms to discrete tokens, (2) a decoder-only Transformer that transforms source tokens to target tokens conditioned on reference, and (3) a flow-matching-based decoder that reconstructs waveforms. We adopt the tokenizer and

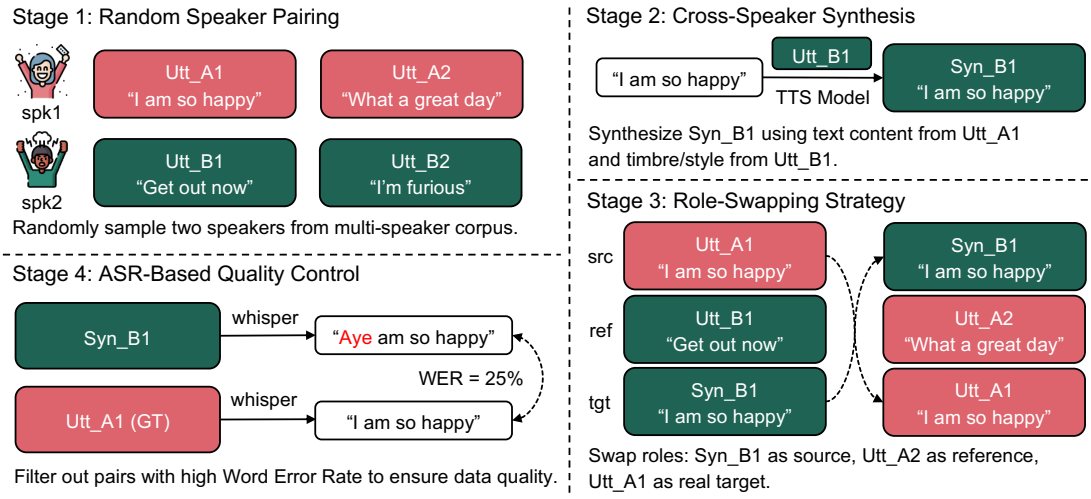

Figure 2: Four-stage pipeline for pseudo-parallel data construction. We randomly sample two speakers with their utterances (Stage 1), synthesize cross-speaker audio using TTS (Stage 2), apply role-swapping to ensure real speech serves as training *target* (Stage 3), and filter low-quality pairs via ASR-based verification (Stage 4).

decoder from CosyVoice 2.0 (Du et al., 2024), which produces 25 tokens per second. Our approach achieves effective voice imitation through three key designs: (1) a role-swapping strategy that constructs pseudo-parallel training pairs with real speech as targets (Section 3.2), (2) interleaved text-audio modeling that enhances content fidelity through explicit textual representation (Section 3.3), and (3) preference alignment that bridges the synthetic-to-real distribution gap during inference (Section 3.4).

3.2 Pseudo-Parallel Data Construction

Motivation. Parallel voice conversion corpora, recordings of the same content spoken by different speakers, are scarce (Yoshino et al., 2016), and manual collection is expensive and time-consuming. Recent zero-shot TTS models (Wang et al., 2024; Du et al., 2024; Xie et al., 2025; Zhou et al., 2025) offer a scalable alternative through synthetic data generation, but using synthetic speech as training objects introduces a quality ceiling. We address this through a role-swapping strategy that uses real speech as targets while leveraging TTS for content transformation.

Four-Stage Pipeline. As shown in Figure 2, our data construction pipeline comprises four stages:

Stage 1: Random Speaker Pairing. From the filtered Emilia dataset (He et al., 2024) containing over 620K English-speaking speakers (each with at least four utterances), we randomly sample two speakers, denoted as `spk1` and `spk2`, and retrieve three utterances: two adjacent utterances `Utt_A1` and `Utt_A2` from `spk1`, and one utterance `Utt_B1` from `spk2`.

Stage 2: Cross-Speaker Synthesis. We employ CosyVoice 2.0 (Du et al., 2024), a zero-shot TTS model, to synthesize speech `Syn_B1` using the text content of `Utt_A1` (e.g., "I am so happy") and the timbre/style reference from `Utt_B1`. A conventional approach would construct a training triplet (`Utt_A1`, `Utt_B1`, `Syn_B1`), where the model learns to map `Utt_A1` (*source*) to `Syn_B1` (*target*) conditioned on `Utt_B1` (*reference*).

However, this configuration suffers from two fundamental limitations: (1) *Synthetic quality ceiling*: Training the model to generate synthetic speech creates an upper bound on output quality, as the targets themselves are imperfect despite advances in modern TTS systems. (2) *Reference-target mismatch*: `Syn_B1` may not accurately inherit the timbre and style from `Utt_B1` due to inherent voice cloning errors in the TTS model, creating inconsistency between the target and reference during training.

Stage 3: Role-Swapping Strategy. To address these issues, we propose a role-swapping strategy that inverts the conventional data configuration. The key insight is: instead of teaching the model to convert real speech `Utt_A1` into synthetic speech `Syn_B1`, we teach it to convert synthetic speech `Syn_B1` back into real speech `Utt_A1`.

Concretely, we construct the training triplet as (`Syn_B1`, `Utt_A2`, `Utt_A1`), where `Syn_B1` serves as the *source* (synthetic speech providing content), `Utt_A2` as the *reference* (real speech from `spk1` providing target timbre/style), and `Utt_A1` as the

target (real speech from spk1 as ground truth output).

This inversion is valid because Syn_B1 and Utt_A1 share the same textual content by construction (Stage 2), making the task equivalent to voice conversion. This design brings two key advantages: (1) *Real speech targets*: The model learns to generate real human speech directly, removing the quality ceiling imposed by synthetic training objects and potentially exceeding the quality of the TTS system used for data construction. (2) *Better reference alignment*: Since Utt_A1 (*target*) and Utt_A2 (*reference*) are both from spk1, they naturally share the same speaker identity and exhibit similar timbre and style characteristics. This reduces the reference-target mismatch present in the conventional approach, where the synthetic *target* may deviate from the *reference* due to TTS cloning errors.

Stage 4: ASR-Based Quality Control. Prior to ASR filtering, we apply voice activity detection (VAD)[1] to trim leading and trailing non-speech segments from the synthesized utterances. This step addresses a common artifact in TTS-generated speech, where variable-length silence may precede the actual spoken content. We then apply Whisper-large-v3 (Radford et al., 2022) to transcribe both Syn_B1 and Utt_A1, retaining only pairs with WER below 0.1. This filtering removes 33% of the data, yielding 8.5M high-quality training triplets (approximately 18K hours). We chose this threshold to balance data quantity and quality; a stricter threshold of 0.01 would remove 61% of the data.

3.3 Interleaved Text-Audio Modeling

Motivation. Recent work in speech language models has demonstrated that incorporating text prediction as an auxiliary task significantly improves speech intelligibility (Xie and Wu, 2024; Ding et al., 2025; Xiaomi, 2025). For voice imitation tasks, this is particularly important: unlike timbre-only conversion, imitating speaking style while preserving content fidelity presents greater challenges to maintaining intelligibility. We therefore adopt an interleaved text-audio architecture where textual predictions provide semantic guidance during audio synthesis. As shown in Table 4, this design substantially reduces WER compared to audio-only training.

[1] https://huggingface.co/nvidia/frame_vad_multilingual_marblenet_v2.0

Interleaved Sequence Construction. We extend Qwen2's (Yang et al., 2024) vocabulary with 6,561 speech tokens from CosyVoice 2.0's frozen audio tokenizer (Du et al., 2024) and special control tokens for sequence management. As illustrated in Figure 1, the input sequence consists of three components: (1) *reference* tokens providing target timbre/style, (2) *source* tokens prefixed by <|SOURCE_START|>, and (3) interleaved text-audio chunks representing the conversion *target*. This sequence operates in two phases:

Chunked Phase. We alternate between text chunks and audio chunks with sizes $C_{\text{text}} = 5$ and $C_{\text{audio}} = 25$ respectively. Each text chunk is enclosed by <|TEXT_START|> and <|TEXT_END|> tokens (omitted in Figure 1 for clarity), immediately followed by the corresponding audio chunk. This 1:5 ratio is deliberately designed: while natural temporal correspondence exhibits approximately three text tokens per 25 audio tokens (reflecting semantic density differences), we increase text chunk size to five, ensuring text predictions temporally lead audio synthesis. This temporal offset allows the model to leverage richer textual context as guidance for more intelligible audio generation.

Continuous Phase. After the chunked phase, any remaining content is generated continuously: remaining text tokens between <|REMAIN_START|> and <|TEXT_END|>, followed by remaining audio tokens ending with <|REMAIN_END|>. This two-phase design accommodates variable-length inputs while maintaining structured guidance during the critical chunked generation phase.

Dual-Task Learning. The model is trained to simultaneously predict text and audio tokens. For loss computation, control tokens following text chunks (<|TEXT_END|>) contribute to text loss, while those following audio tokens (<|TEXT_START|>, <|REMAIN_START|>, <|REMAIN_END|>) contribute to audio loss. The training objective is:

$$\mathcal{L} = 0.5\mathcal{L}_{\text{text}} + 0.5\mathcal{L}_{\text{audio}}, \tag{1}$$

where both losses are cross-entropy computed at their respective token positions.

3.4 Preference Alignment

Motivation. While role-swapping ensures high-quality *target* speech via real recordings, the *source* speech remains synthetic during training. This creates a distributional mismatch at inference time. As

shown in Table 3, our SFT model achieves 4.30% WER when evaluated on Syn/Real pairs (matching the training distribution), but performance degrades to 15.80% WER on Real/Real pairs (real-world inference scenario). To bridge this synthetic-to-real gap (Su et al., 2024), we employ Direct Preference Optimization (DPO) (Rafailov et al., 2023) to align the model with real-world acoustic conditions.

Preference Data Construction. We construct two-stage preference data from 150K speaker pairs in Emilia. For each *(source, reference)* pair, we generate $K = 8$ candidate outputs using nucleus sampling and rank them using automatic metrics (Section 4.1). Stage 1 uses the base model to generate candidates and prioritizes speech intelligibility (WER) to align with real source characteristics. After Stage 1 optimization, Stage 2 generates new candidates and focuses on acoustic similarity including voice timbre, accent, and emotional expression. Full construction details are in Appendix B.

Training Objective. Following (Rafailov et al., 2023), the DPO loss is:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_D\Big[\log\sigma\Big(\beta\big(\log\tfrac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \log\tfrac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\big)\Big)\Big], \tag{2}$$

where π_θ is the policy being optimized, π_{ref} is the frozen reference model (our SFT model), and (x, y_w, y_l) denotes the (input, chosen, rejected) triplet. We set $\beta = 0.1$ following the default configuration in (Rafailov et al., 2023).

4 Experiments and Results

4.1 Experimental Settings

Training Data. We construct a pseudo-parallel corpus following Section 3.2 based on Emilia (He et al., 2024), a large-scale multilingual speech dataset with 1.13M English and 0.91M Chinese speakers. We select 620K speakers per language (each with at least four utterances), generating 8.5M English pairs (18K hours) and 0.74M Chinese pairs (1.6K hours). All experiments use English data unless noted; multilingual training results are in Appendix A.

Evaluation Data. We evaluate on two benchmarks: (1) SeedTTS *test-vc-en* (Anastassiou et al., 2024): A zero-shot TTS/VC benchmark from Common Voice (Ardila et al., 2020). We select all utterances longer than 4 seconds. (2) MimicLM-Test: To analyze the synthetic-to-real gap, we construct a diagnostic set with 833 speaker pairs (1,666 utterances) from Emilia. Each pair is evaluated under three conditions based on input types: *Real/Real*, *Syn/Real*, and *Real/Syn*, where the format indicates *Source/Reference* audio types.

Evaluation Metrics. We employ both objective and subjective metrics. For naturalness, we use DNSMOS[2] (Reddy et al., 2021) to evaluate speech quality (SIG), background noise (BAK), and overall quality (OVRL), and UTMOSv2[3] (Baba et al., 2024) for automatic Mean Opinion Score (MOS) prediction. For intelligibility, we transcribe outputs with Whisper-Large-v3 (Radford et al., 2023) and compute Word Error Rate (WER). For similarity, we compute cosine similarity between embeddings of generated and reference audio using WavLM-Large[4] (Chen et al., 2022) for speaker similarity (S-SIM), CommonAccent (Ardila et al., 2020) for accent (A-SIM), and emotion2vec[5] (Ma et al., 2024) for emotion (E-SIM). We also conduct subjective evaluation where human raters score naturalness (N-MOS), speaker similarity (S-MOS), accent similarity (A-MOS), and emotion similarity (E-MOS) on a 1–5 scale. Details are in Appendix C.

Baselines We compare against open-source state-of-the-art zero-shot voice conversion systems in two categories. Timbre-only systems including CosyVoice 2.0 (Du et al., 2024), FACodec (Ju et al., 2024), OpenVoice v2 (Qin et al., 2023), LSCodec (Guo et al., 2024b), and SeedVC (Liu, 2024) transfer speaker identity while preserving source prosody. Full voice imitation systems including SeedVC v2 (Liu, 2024) and Vevo (Zhang et al., 2025) transfer both timbre and style.

Training and Inference Setup We train the model in two stages on NVIDIA A800 GPUs. Stage 1 performs supervised fine-tuning for 4 epochs with effective batch size of 128, learning rate 5×10^{-4}, and warmup ratio 0.03. Stage 2 applies DPO alignment with effective batch size of 32, learning rate 1×10^{-5}, $\beta = 0.1$, and warmup ratio 0.05 for 4 epochs. We perform inference in bfloat16 precision. Text generation uses temperature 0.7, top-p 0.92, and repetition penalty 1.05, while audio generation uses temperature 0.8, top-p

[2] https://github.com/microsoft/DNS-Challenge
[3] https://github.com/sarulab-speech/UTMOSv2
[4] https://github.com/BytedanceSpeech/seed-tts-eval
[5] https://github.com/ddlBoJack/emotion2vec

Table 1: Performance comparison on SeedTTS *test-vc-en*. We compare our method against two groups of baselines: (1) *Timbre-only* systems, and (2) *Full Voice Imitation* systems (our direct baselines: Vevo and SeedVC v2), which transfer both timbre and style. "↑" indicates higher is better, and "↓" indicates lower is better. WER comparison is only made among voice imitation systems.

Model	Naturalness & Speech Quality				Intelligibility	Speaker Similarity		
	UTMOS ↑	OVRL ↑	SIG ↑	BAK ↑	WER (%) ↓	S-SIM ↑	A-SIM ↑	E-SIM ↑
Timbre-only Voice Conversion								
FACodec (Ju et al., 2024)	2.13	3.65	4.22	3.87	4.86	0.372	0.571	0.912
MeanVC (Ma et al., 2025)	2.84	3.65	4.04	4.14	17.82	0.419	0.450	0.909
OpenVoice v2 (Qin et al., 2023)	2.55	**4.15**	**4.47**	4.43	4.15	0.424	0.552	0.918
LSCodec (Guo et al., 2024b)	2.84	3.94	4.33	4.25	8.76	0.445	0.583	0.911
CosyVoice 2.0 (Du et al., 2024)	3.04	3.98	4.31	4.38	4.28	0.539	0.647	0.919
SeedVC (Liu, 2024)	2.79	3.71	4.19	4.03	3.25	0.587	0.684	0.922
Full Voice Imitation (Timbre + Style Transfer)								
SeedVC v2 (Liu, 2024)	2.94	3.65	4.14	4.01	**6.32**	0.553	0.653	0.917
Vevo (Zhang et al., 2025)	2.83	3.77	4.27	4.00	9.10	**0.652**	**0.727**	**0.926**
Ours (SFT)	**3.31**	4.12	4.43	4.42	12.80	0.571	0.692	0.912
Ours (DPO)	3.22	**4.15**	4.45	**4.45**	8.25	0.601	0.699	0.925

Table 2: Subjective evaluation results with Mean Opinion Scores (MOS) on a 1–5 scale. Scores are reported as mean ± 95% confidence interval.

Model	N-MOS	S-MOS	A-MOS	E-MOS
SeedVC v2	3.14 ± 0.11	3.03 ± 0.12	3.82 ± 0.12	3.61 ± 0.16
Vevo	3.85 ± 0.14	4.32 ± 0.13	$\mathbf{4.64 \pm 0.09}$	$\mathbf{4.23 \pm 0.09}$
Ours (DPO)	$\mathbf{4.71 \pm 0.08}$	$\mathbf{4.62 \pm 0.10}$	4.53 ± 0.11	3.94 ± 0.13

Table 3: Synthetic-to-real gap analysis on MimicLM-Test. Column headers indicate *source/reference* input types, where *source* provides linguistic content and *reference* provides target voice. All values are WER (%).

Model	Real/Real	Syn/Real	Real/Syn
Vevo	17.99	13.90	20.44
Ours (SFT)	15.80	4.30	18.48
Ours (DPO)	**13.81**	**3.63**	**15.58**

0.9, and repetition penalty 1.2. More details are in Appendix D.

4.2 Main Results

Comparison with Baselines. Table 1 presents our method's performance on SeedTTS *test-vc-en* against state-of-the-art baselines. Our SFT model is trained on 8.5M pseudo-parallel pairs (18K hours) constructed from real speech via role-swapping, while our DPO model further applies preference alignment as post-training. Overall, our method achieves competitive performance across naturalness, intelligibility, and similarity metrics. (1) *Naturalness and Speech Quality.* Our method demonstrates strong naturalness, with both SFT and DPO versions achieving competitive scores across quality metrics. After DPO alignment, these naturalness metrics remain stable, indicating that preference optimization maintains speech quality while improving other performance dimensions. (2) *Intelligibility.* Voice imitation systems generally exhibit higher WER than timbre-only approaches, as transforming both timbre and speaking style increases content preservation complexity. Among full imitation systems, our SFT model's WER falls between SeedVC v2 and Vevo. The DPO process substantially reduces WER, demonstrating that post-training alignment on real speech inputs effectively bridges the synthetic-to-real gap encountered during inference. (3) *Similarity Performance.* Our method achieves competitive similarity performance across all dimensions. For speaker identity similarity (S-SIM), our DPO model outperforms timbre-only VC systems and achieves results comparable to other full imitation methods, surpassing SeedVC v2 and approaching Vevo's performance. For accent similarity (A-SIM), our DPO model demonstrates advantages over timbre-only baselines, with performance positioned between SeedVC v2 and Vevo among full imitation systems. For emotion similarity (E-SIM), our DPO model closely matches the best-performing Vevo on this test set. The consistent improvements from SFT to DPO across all similarity metrics demonstrate that preference alignment effectively enhances the model's ability to capture comprehensive voice characteristics. (4) *Subjective Evaluation.* We conduct human evaluation to complement the objective metrics. As shown in Table 2, our DPO model

Table 4: Ablation study on key components using *base*-scale data. RS = Role-Swapping, IT = Interleaved Text modeling. The baseline uses standard triplets (`Utt_A1`, `Utt_B1`, `Syn_B1`) without IT. RS uses role-swapped triplets (`Syn_B1`, `Utt_A2`, `Utt_A1`). All configurations are evaluated on SeedTTS *test-vc-en*. Training hyperparameters follow the SFT stage setup described in Appendix D.

Configuration	OVRL ↑	SIG ↑	BAK ↑	WER (%) ↓	S-SIM ↑	A-SIM ↑	E-SIM ↑
w/o RS, w/o IT	3.99	4.39	4.25	18.25	0.547	0.678	0.903
w/ RS, w/o IT	4.05	4.41	4.33	20.69	0.555	0.684	0.910
w/o RS, w/ IT	4.03	4.41	4.31	15.34	0.547	0.681	0.896
w/ RS, w/ IT (SFT)	4.11	4.43	4.41	18.64	0.560	**0.691**	**0.913**
SFT + DPO	**4.12**	**4.44**	**4.42**	**14.73**	**0.573**	0.688	0.905

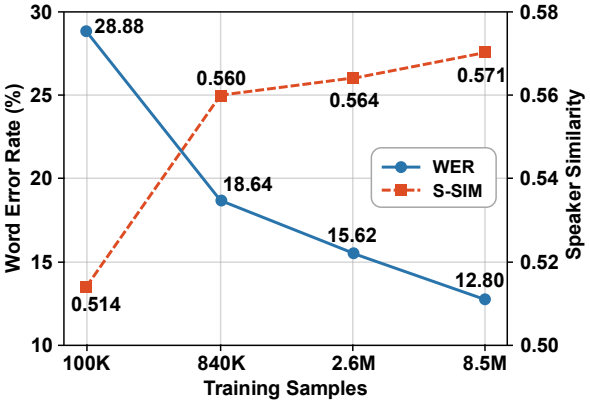

Figure 3: Impact of training data scale on WER and speaker similarity (S-SIM) evaluated on SeedTTS *test-vc-en*.

achieves the highest scores in N-MOS and S-MOS. These subjective results validate the effectiveness of our approach.

Addressing Synthetic-to-Real Gap. Table 3 quantifies the distributional shift between training and inference conditions on our diagnostic set MimicLM-Test. As discussed in Section 3.4, our SFT model performs well on Syn/Real pairs (matching the training distribution) but exhibits substantially higher WER on Real/Real pairs (real-world inference scenario), confirming the synthetic-to-real gap. DPO effectively bridges this gap, achieving the lowest WER across all input configurations and surpassing Vevo under Real/Real conditions despite Vevo being trained on real data. Notably, DPO maintains strong performance on Syn/Real pairs while successfully adapting to Real/Real pairs, demonstrating that preference optimization generalizes across different input distributions without sacrificing quality on the training domain.

Data Scaling. To investigate scaling behavior, we train models on datasets of varying sizes: *tiny* (100K samples), *base* (840K), *medium* (2.6M), and *large* (8.5M), keeping architecture and hyperparameters fixed. Figure 3 shows that performance consistently improves with scale. The transition from *tiny* to *base* demonstrates strong improvements in both WER and speaker similarity, indicating substantial benefits in the small-data regime. As data scale increases further, gains follow typical scaling law patterns with diminishing returns, though neither metric saturates at the largest scale, suggesting potential for further improvement with more data. Notably, WER exhibits steeper improvements than S-SIM across scales, suggesting that content preservation and voice characteristic modeling may have different scaling dynamics.

4.3 Ablation Studies

We validate the effectiveness of role-swapping (RS), interleaved text modeling (IT), and DPO through systematic ablations on *base*-scale data (840K samples). Table 4 presents results on SeedTTS *test-vc-en*. All ablated models are trained from scratch under identical training configurations, ensuring a fair comparison of each component's individual contribution. (1) *Role-swapping* (rows 1 vs. 2, rows 3 vs. 4) consistently improves naturalness and similarity, confirming that real target speech enhances voice characteristic transfer. (2) *Interleaved text modeling* (rows 1 vs. 3, rows 2 vs. 4) substantially reduces WER by providing explicit linguistic anchors that prevent content collapse. Their combination in the SFT model achieves the best naturalness with strong similarity and reasonable WER, validating synergistic effects. (3) *Preference alignment* further reduces WER and improves speaker similarity while maintaining naturalness. Note that this DPO uses *base*-scale data optimized for WER and S-SIM, whereas our main model uses *large*-scale data with multi-stage optimization (Section 3.4).

5 Conclusion

We presented MimicLM, a voice imitation system that learns from real speech distributions by using synthetic speech as training sources rather than targets. This role-swapping data construction strategy breaks the quality ceiling imposed by synthesis systems while ensuring natural alignment between references and targets. Combined with interleaved text-audio modeling for content preservation and preference alignment for handling real speech inputs, MimicLM achieves competitive performance compared to state-of-the-art systems while offering a simpler alternative to complex disentanglement-based architectures. Our work demonstrates that inverting the role of synthetic data enables effective voice imitation without sacrificing naturalness or requiring elaborate architectural designs.

Limitations

Our approach has several limitations worth noting. First, while avoiding synthetic *targets*, our method still depends on TTS quality for generating training sources. Poor TTS outputs require filtering (33% removal rate), potentially limiting data efficiency and introducing biases from the external synthesis system. Second, despite DPO alignment, our model exhibits higher WER than timbre-only systems when processing real inputs, indicating that the synthetic-to-real gap and the complexity of joint timbre-prosody transformation remain challenging. Third, our two-stage pipeline and large-scale data construction (8.5M pairs with TTS synthesis and preference optimization) demand substantial computational resources, which may limit accessibility for researchers with limited budgets. Finally, our evaluation focuses primarily on English with limited multilingual analysis. Real-world diversity in accents, speaking styles, and acoustic conditions may present challenges not fully covered by current benchmarks, warranting further investigation into edge cases and challenging scenarios.

References

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, and 1 others. 2024. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference*, pages 4218–4222.

Kaito Baba, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari. 2024. The t05 system for the voicemos challenge 2024: Transfer learning from deep image classifier to naturalness mos prediction of high-quality synthetic speech. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 818–824. IEEE.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. 2025. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6255–6271.

Ha-Yeong Choi, Sang-Hoon Lee, and Seong-Whan Lee. 2023. Diff-hiervc: Diffusion-based hierarchical voice conversion with robust pitch generation and masked prior for zero-shot speaker adaptation. *arXiv preprint arXiv:2311.04693*.

Hyeong-Seok Choi, Juheon Lee, Wansoo Kim, Jie Lee, Hoon Heo, and Kyogu Lee. 2021. Neural analysis and synthesis: Reconstructing speech from self-supervised representations. *Advances in Neural Information Processing Systems*, 34:16251–16265.

Tri Dao. 2024. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*.

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

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, and 1 others. 2025. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, and 1 others. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

Hao-Han Guo, Yao Hu, Kun Liu, Fei-Yu Shen, Xu Tang, Yi-Chen Wu, Feng-Long Xie, Kun Xie, and Kai-Tuo Xu. 2024a. Fireredtts: A foundation text-to-speech framework for industry-level generative speech applications. *arXiv preprint arXiv:2409.03283*.

Yiwei Guo, Zhihan Li, Chenpeng Du, Hankun Wang, Xie Chen, and Kai Yu. 2024b. Lscodec: Low-bitrate and speaker-decoupled discrete speech codec. *arXiv preprint arXiv:2410.15764*.

Zhao Guo, Ziqian Ning, Guobin Ma, and Lei Xie. 2025. Synthvc: Leveraging synthetic data for end-to-end low latency streaming voice conversion. *arXiv preprint arXiv:2510.09245*.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, and 1 others. 2024. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 885–890. IEEE.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, and 1 others. 2024. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR.

Mateusz Łajszczak, Guillermo Cámbara, Yang Li, Fatih Beyhan, Arent Van Korlaar, Fan Yang, Arnaud Joly, Álvaro Martín-Cortinas, Ammar Abbas, Adam Michalski, and 1 others. 2024. Base tts: Lessons from building a billion-parameter text-to-speech model on 100k hours of data. *arXiv preprint arXiv:2402.08093*.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and 1 others. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. *Advances in neural information processing systems*, 36:14005–14034.

Sang-Hoon Lee, Ha-Yeong Choi, Seung-Bin Kim, and Seong-Whan Lee. 2025. Hierspeech++: Bridging the gap between semantic and acoustic representation of speech by hierarchical variational inference for zero-shot speech synthesis. *IEEE Transactions on Neural Networks and Learning Systems*.

Fengjin Li, Jie Wang, Yadong Niu, Yongqing Wang, Meng Meng, Jian Luan, and Zhiyong Wu. 2025. Starvc: A unified auto-regressive framework for joint text and speech generation in voice conversion. *arXiv preprint arXiv:2506.02414*.

Songting Liu. 2024. Zero-shot voice conversion with diffusion transformers. *arXiv preprint arXiv:2411.09943*.

Songxiang Liu, Disong Wang, Yuewen Cao, Lifa Sun, Xixin Wu, Shiyin Kang, Zhiyong Wu, Xunying Liu, Dan Su, Dong Yu, and 1 others. 2020. End-to-end accent conversion without using native utterances. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6289–6293. IEEE.

Guobin Ma, Jixun Yao, Ziqian Ning, Yuepeng Jiang, Lingxin Xiong, Lei Xie, and Pengcheng Zhu. 2025. Meanvc: Lightweight and streaming zero-shot voice conversion via mean flows. *arXiv preprint arXiv:2510.08392*.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2024. emotion2vec: Self-supervised pre-training for speech emotion representation. *Proc. ACL 2024 Findings*.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. Speech resynthesis from discrete disentangled self-supervised representations. *arXiv preprint arXiv:2104.00355*.

Kaizhi Qian, Yang Zhang, Shiyu Chang, Mark Hasegawa-Johnson, and David Cox. 2020. Unsupervised speech decomposition via triple information bottleneck. In *International Conference on Machine Learning*, pages 7836–7846. PMLR.

Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. 2019. Autovc: Zero-shot voice style transfer with only autoencoder loss. In *International Conference on Machine Learning*, pages 5210–5219. PMLR.

Zengyi Qin, Wenliang Zhao, Xumin Yu, and Xin Sun. 2023. Openvoice: Versatile instant voice cloning. *arXiv preprint arXiv:2312.01479*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

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

Chandan KA Reddy, Vishak Gopal, and Ross Cutler. 2021. Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors.

In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6493–6497. IEEE.

Berrak Sisman, Junichi Yamagishi, Simon King, and Haizhou Li. 2020. An overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:132–157.

Hsuan Su, Hua Farn, Fan-Yun Sun, Shang-Tse Chen, and Hung-yi Lee. 2024. Task arithmetic can mitigate synthetic-to-real gap in automatic speech recognition. *arXiv preprint arXiv:2406.02925*.

Huu Tuong Tu, Huan Vu, Nguyen Tien Cuong, Ngo Dien Hy, and Nguyen Thi Thu Trang. 2025. O_o-vc: Synthetic data-driven one-to-one alignment for any-to-any voice conversion. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16197–16208.

Aaron Van Den Oord, Oriol Vinyals, and 1 others. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, and 1 others. 2023a. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, and 1 others. 2025a. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *arXiv preprint arXiv:2503.01710*.

Yuancheng Wang, Dekun Chen, Xueyao Zhang, Junan Zhang, Jiaqi Li, and Zhizheng Wu. 2025b. Tadicodec: Text-aware diffusion speech tokenizer for speech language modeling. *arXiv preprint arXiv:2508.16790*.

Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. 2024. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. *arXiv preprint arXiv:2409.00750*.

Zhichao Wang, Yuanzhe Chen, Lei Xie, Qiao Tian, and Yuping Wang. 2023b. Lm-vc: Zero-shot voice conversion via speech generation based on language models. *IEEE Signal Processing Letters*, 30:1157–1161.

LLM-Core-Team Xiaomi. 2025. Mimo-audio: Audio language models are few-shot learners.

Kun Xie, Feiyu Shen, Junjie Li, Fenglong Xie, Xu Tang, and Yao Hu. 2025. Fireredtts-2: Towards long conversational speech generation for podcast and chatbot. *arXiv preprint arXiv:2509.02020*.

Zhifei Xie and Changqiao Wu. 2024. Mini-omni: Language models can hear, talk while thinking in streaming. *arXiv preprint arXiv:2408.16725*.

Detai Xin, Xu Tan, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2024. Bigcodec: Pushing the limits of low-bitrate neural speech codec. *arXiv preprint arXiv:2409.05377*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Zhen Ye, Xinfa Zhu, Chi-Min Chan, Xinsheng Wang, Xu Tan, Jiahe Lei, Yi Peng, Haohe Liu, Yizhu Jin, Zheqi Dai, and 1 others. 2025. Llasa: Scaling train-time and inference-time compute for llama-based speech synthesis. *arXiv preprint arXiv:2502.04128*.

Koichiro Yoshino, Naoki Hirayama, Shinsuke Mori, Fumihiko Takahashi, Katsutoshi Itoyama, and Hiroshi G Okuno. 2016. Parallel speech corpora of japanese dialects. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4652–4657.

Jing-Xuan Zhang, Zhen-Hua Ling, and Li-Rong Dai. 2019. Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:540–552.

Xueyao Zhang, Xiaohui Zhang, Kainan Peng, Zhenyu Tang, Vimal Manohar, Yingru Liu, Jeff Hwang, Dangna Li, Yuhao Wang, Julian Chan, and 1 others. 2025. Vevo: Controllable zero-shot voice imitation with self-supervised disentanglement. *arXiv preprint arXiv:2502.07243*.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. 2022. Emotional voice conversion: Theory, databases and esd. *Speech Communication*, 137:1–18.

Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. 2025. Indextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. *arXiv preprint arXiv:2506.21619*.

A Multilingual Training Results

To demonstrate language generalizability, we train models on Chinese data constructed following the same Role-Swapping procedure (Section 3.2). We generate 740K Chinese pairs (1.6K hours) from Emilia and evaluate on SeedTTS *test-zh* (Anastassiou et al., 2024), reporting Word Character Error Rate (CER) for Chinese. Following the English evaluation setup, we filter the test set to retain only

Table 5: Performance on Chinese (SeedTTS *test-zh*) and English (SeedTTS *test-vc-en*) test sets for different training data configurations.

Training Data	English		Chinese	
	WER (%) ↓	S-SIM ↑	CER (%) ↓	S-SIM ↑
English-only (0.84M)	18.64	0.560	-	-
English-only (2.6M)	15.62	0.564	-	-
Chinese-only (0.74M)	-	-	13.16	0.630
Chinese (0.74M) + English (0.84M)	19.49	0.558	12.84	0.632
Chinese (0.74M) + English (2.6M)	15.21	0.566	13.68	0.631

examples with reference utterances longer than 4 seconds.

Table 5 presents results for monolingual and multilingual training configurations. The Chinese-only model demonstrates our approach works effectively for this language. For multilingual training, outcomes vary with data balance: on Chinese evaluation, adding balanced English data improves performance, but disproportionately large English data causes degradation; on English evaluation, Chinese data slightly hurts performance with small English datasets but provides gains with larger ones.

B DPO Preference Pair Construction

We conducted two rounds of DPO training, each utilizing 150,000 input samples from the training set. The preference pair construction process consists of two stages:

Stage 1: Candidate Generation. For each input sample, we employ a multi-process inference framework to generate 8 candidate responses (both text and audio). The generation process uses nucleus sampling with configurable temperature and top-p parameters for both text ($T_{\text{text}} = 0.7$, $p_{\text{text}} = 0.92$) and audio modalities ($T_{\text{audio}} = 0.8$, $p_{\text{audio}} = 0.9$). The framework supports distributed inference across 8 GPUs with checkpoint-based resumption capabilities.

Stage 2: Pareto-Optimal Pair Selection. We employ a multi-objective optimization strategy based on Pareto dominance to construct preference pairs. For each sample, all candidate pairs are evaluated using three metrics: (1) *Audio WER* (lower is better), measuring speech recognition accuracy; (2) *SIM* (higher is better), measuring overall similarity to the reference; and (3) *eSIM* (higher is better), capturing fine-grained similarity.

A candidate c_1 is considered to dominate c_2 (forming a chosen-rejected pair) if and only if:

- c_1 is no worse than c_2 on all metrics (Pareto condition);
- c_1 is strictly better than c_2 on at least one metric;
- The improvement on each metric exceeds a minimum threshold δ_{min};
- Both c_1 (chosen) and c_2 (rejected) satisfy quality constraints.

Filtering Criteria. We denote the minimum improvement threshold as δ_{min}, the maximum allowed value for chosen samples as v^c_{max}, and the maximum allowed value for rejected samples as v^r_{max}. The specific criteria are:

Round 1: Audio WER: $\delta_{\text{min}} = 0.05$, $v^c_{\text{max}} = 0.30$, $v^r_{\text{max}} = 0.60$; SIM: $\delta_{\text{min}} = 0.01$, no v^c_{max} or v^r_{max} constraints; eSIM: $\delta_{\text{min}} = 0.01$, no v^c_{max} or v^r_{max} constraints.

Round 2: Audio WER: $\delta_{\text{min}} = 0.00$, $v^c_{\text{max}} = 0.30$, $v^r_{\text{max}} = 0.60$; SIM: $\delta_{\text{min}} = 0.02$, no constraints; eSIM: $\delta_{\text{min}} = 0.02$, no constraints.

Round 1 enforces stricter minimum improvements ($\delta_{\text{min}} = 0.05$ for WER) to ensure high-quality pairs, while Round 2 relaxes the WER threshold to 0.00 but increases similarity requirements to 0.02, encouraging more diverse training signals. The quality constraints ensure that chosen samples maintain low WER (≤ 0.30) while preventing overly poor rejected samples (WER ≤ 0.60).

C Subjective Evaluation Protocol

We conduct subjective evaluation on a randomly selected subset of 20 audio pairs from the SeedTTS *test-vc-en* dataset. Each sample is evaluated by 10 trained listeners using a 5-point Likert scale (1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent).

The evaluation covers four dimensions: Naturalness (N-MOS), Speaker Similarity (S-MOS), Accent Similarity (A-MOS), and Emotion Similarity (E-MOS). For N-MOS, listeners assess the overall naturalness and audio quality of the generated speech, disregarding any differences in speaker characteristics or content accuracy. For S-MOS, listeners evaluate how closely the generated speech matches the target speaker's timbre and voice characteristics in the reference prompt, while disregarding differences in content or audio quality. For A-MOS, listeners assess how well the generated speech preserves the accent and pronunciation patterns of the target speaker. For E-MOS, listeners evaluate how closely the generated speech matches the emotional tone and expressiveness of the target speaker. For each dimension, listeners are instructed to focus solely on that specific aspect while ignoring other characteristics. The final MOS score

for each model is computed as the average of all listener ratings, with confidence intervals calculated based on the standard error.

D Training Details

Stage 1: Supervised Fine-Tuning We train the MimicLM model based on the Qwen2.5-0.5B (Yang et al., 2024) architecture on 8 NVIDIA A800 GPUs for 4 epochs. The training configuration is as follows: per-device batch size of 4 with gradient accumulation steps of 4, resulting in an effective batch size of 128. We employ the AdamW optimizer with a learning rate of 5×10^{-4}, $\beta_2 = 0.999$, weight decay of 0.01, warmup ratio of 0.03, and cosine learning rate scheduling. The model uses Flash Attention 2 (Dao, 2024) and gradient checkpointing for training efficiency. The maximum sequence length is set to 2560 tokens, and the text and audio loss weights are both set to 0.5. Text chunks and audio chunks are configured with sizes of 5 and 25 tokens, respectively.

Stage 2: DPO Alignment We perform Direct Preference Optimization (DPO) training on 4 GPUs using the checkpoint from Stage 1. The training runs for 4 epochs with per-device batch size of 4 and gradient accumulation steps of 2, yielding an effective batch size of 32. We reduce the learning rate to 1×10^{-5} with the DPO β parameter set to 0.1, warmup ratio of 0.05, weight decay of 0.01, and maximum gradient norm clipping of 1.0. Both training stages utilize bfloat16 mixed precision training.

E Broader Impact and Potential Risks

While MimicLM advances voice imitation technology for legitimate applications such as personalized voice assistants, audiobook narration, and accessibility tools, we acknowledge potential misuse risks that warrant careful consideration.

The primary concern is unauthorized voice cloning for impersonation or fraud. Our model can generate speech that imitates a target speaker's voice using only a 3-second prompt, which could potentially be exploited for social engineering attacks, spreading misinformation through fake audio, or creating non-consensual voice replicas. Although our system is designed for research purposes and legitimate use cases, malicious actors could adapt similar techniques for harmful purposes.

To mitigate these risks, we recommend several safeguards for practical deployment. First, implement speaker verification and consent mechanisms before allowing voice cloning of any individual. Second, incorporate watermarking or fingerprinting techniques to trace generated audio back to its source. Third, develop and deploy detection systems capable of identifying synthetic speech generated by voice imitation models. Fourth, establish clear usage policies and legal frameworks that criminalize unauthorized voice cloning and impersonation.

We emphasize that voice imitation technology itself is neutral—its impact depends on how it is deployed and regulated. The research community, policymakers, and technology developers must work collaboratively to ensure these technologies are used responsibly. We release our work to advance scientific understanding while encouraging ongoing dialogue about ethical deployment practices and appropriate regulatory frameworks.

Our model's current limitations (higher WER on real inputs, dependency on high-quality prompts) provide some natural barriers against trivial misuse, but these technical limitations should not be relied upon as primary safeguards. As the technology matures, proactive measures become increasingly important to prevent harm while preserving beneficial applications.

F Use of Large Language Models

During the preparation of this manuscript, a Large Language Model (LLM) was utilized as a writing aid to improve the overall linguistic quality and clarity. This assistance was confined to copy-editing tasks, such as correcting grammatical and spelling errors, rephrasing sentences for enhanced flow and readability, and ensuring conciseness. All scientific contributions, including the research ideas, experimental design, analysis, and conclusions presented herein, are entirely the original work of the human authors.

G TTS Model Selection

To select the TTS system used for cross-speaker synthesis in Stage 2, we evaluated several state-of-the-art open-source zero-shot TTS models on the SeedTTS *test-en* and *test-zh* benchmarks (Anastassiou et al., 2024), following the evaluation protocol in (Wang et al., 2025b). Table 6 reports WER and speaker similarity (SIM) for each system.

Table 6: Comparison of zero-shot TTS systems on SeedTTS *test-en* and *test-zh*. Results are from (Wang et al., 2025b). SIM denotes cosine speaker similarity.

System	Frame Rate	English		Chinese	
		WER ↓	SIM ↑	WER ↓	SIM ↑
FireRedTTS (Guo et al., 2024a)	25	8.53	0.46	1.27	0.65
SparkTTS (Wang et al., 2025a)	50	2.50	0.57	1.78	0.66
Llasa (Ye et al., 2025) (16 kHz)	50	3.94	0.58	8.02	0.64
F5-TTS (Chen et al., 2025)	93.75	3.02	0.63	3.87	0.71
CosyVoice 2 (Du et al., 2024)	25	2.89	0.66	1.29	0.76

CosyVoice 2 achieves the best bilingual speaker similarity while maintaining competitive WER on both test sets, and operates at a low frame rate of 25 Hz—reducing synthesis cost relative to higher-rate alternatives. These properties make it the most suitable choice for large-scale pseudo-parallel data construction in our pipeline.

H Audio Tokenizer Selection

We selected the audio tokenizer by evaluating multiple codec systems on the SeedTTS *test-en* benchmark using a reconstruction protocol, following (Wang et al., 2025b). Table 7 reports WER, speaker similarity (SIM), and UTMOS for each tokenizer.

Table 7: Comparison of audio tokenizers evaluated via speech reconstruction on SeedTTS *test-en*. Results are from (Wang et al., 2025b). "Semantic Distill" indicates whether the tokenizer incorporates semantic distillation.

System	Rate (Hz)	Codebooks	Semantic	WER ↓	SIM ↑	UTMOS ↑
EnCodec (Défossez et al., 2022)	75	2	×	5.36	0.48	1.54
Mimi (Défossez et al., 2024)	12.5	6	✓	4.51	0.52	3.09
BigCodec (Xin et al., 2024)	80	1	×	3.25	0.61	3.59
X-Codec 2 (Ye et al., 2025)	50	1	✓	2.63	0.62	3.68
Vevo Tokenizer (Zhang et al., 2025)	50	1	✓	3.04	0.53	3.50
CosyVoice 2 (Du et al., 2024)	25	1	✓	4.10	0.68	3.65

CosyVoice 2's semantic tokenizer achieves the highest speaker similarity (0.68) among all compared tokenizers, with competitive WER (4.10) and UTMOS (3.65). Its low frame rate of 25 Hz with a single codebook is particularly advantageous for autoregressive language modeling: the reduced sequence length lowers both training and inference cost, while a single codebook eliminates the architectural complexity of multi-codebook generation strategies (e.g., delayed patterns or residual prediction). The tokenizer is decoupled from our main contributions and can be replaced as better alternatives emerge.