Step-Audio-EditX Technical Report

Chao Yan[†*], Boyong Wu[*], Peng Yang, Pengfei Tan, Guoqiang Hu, Yuxin Zhang, Xiangyu(Tony) Zhang, Fei Tian, Xuerui Yang, Xiangyu Zhang, Daxin Jiang, Gang Yu

Abstract

We present Step-Audio-EditX, the first open-source LLM-based audio model excelling at expressive and iterative audio editing—encompassing emotion, speaking style, and paralinguistics—alongside robust zero-shot text-to-speech (TTS) capabilities.Our core innovation lies in leveraging only large-margin synthetic data, which circumvents the need for embedding-based priors or auxiliary modules. This large-margin learning approach enables both iterative control and high expressivity across voices, and represents a fundamental pivot from the conventional focus on representation-level disentanglement. Evaluation results demonstrate that Step-Audio-EditX surpasses both MiniMax-2.6-hd and Doubao-Seed-TTS-2.0 in emotion editing and other fine-grained control tasks. Our code and models are available at `https://github.com/stepfun-ai/Step-Audio-EditX`.
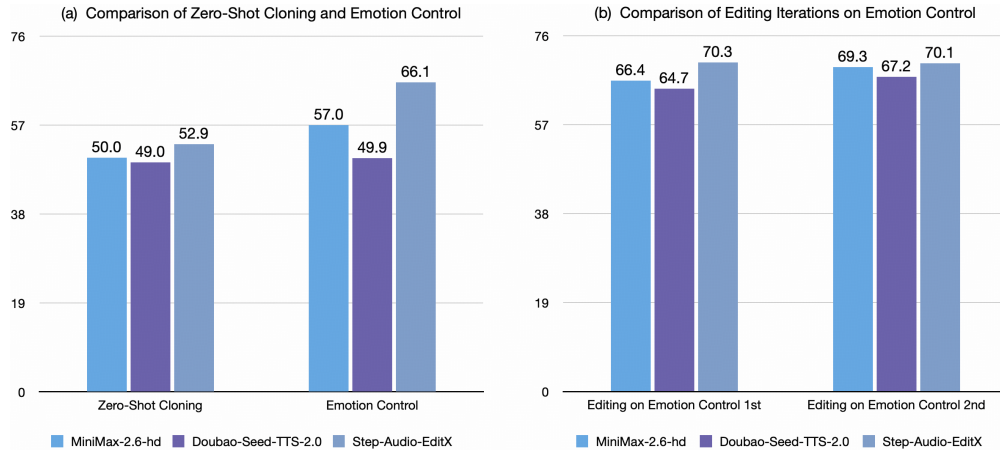
Figure 1: Comparison between Step-Audio-EditX and Closed-Source models.(a) Step-Audio-EditX demonstrates superior performance over Minimax and Doubao in both zero-shot cloning and emotion control.(b) Emotion editing of Step-Audio-EditX significantly improves the emotion-controlled audio outputs of all three models after just one iteration. With further iterations, their overall performance continues to improve.

1 Introduction

In recent years, TTS technology has advanced significantly. A notable development is zero-shot TTS models, which can generate high-quality, natural-sounding speech by mimicking the timbre,

---

* Core Contribution
† Corresponding Author: yanchao@stepfun.com

prosody, and style of a reference speech prompt. Generally, current zero-shot TTS systems fall into three main categories: those that utilize LLMs to model discrete or continuous acoustic tokens [1, 2, 3, 4, 5], those employing diffusion or flow matching models to learn direct text-to-speech mapping [6, 7, 8, 9, 10, 11], and hybrid coarse-to-fine systems, where LLMs first convert text tokens into coarse speech tokens, which are then refined by a diffusion or flow matching model to render fine-grained speech details[12, 13, 14, 15, 16, 17, 18, 19].

Despite considerable progress in zero-shot TTS synthesis, attributes such as emotion, style, accent, and timbre in the synthesized speech are still derived directly from the reference audio. This inherent limitation restricts independent control over these attributes. Although prepending style or emotion instructions to the input text offers a certain degree of controllability and often performs well for in-domain speakers [9, 12, 13, 14], this approach faces challenges in disentangling speech attributes. In particular, the cloned voice often fails to effectively follow the provided style or emotion instructions.

Many previous studies on speech disentanglement have relied on approaches such as adversarial training[20, 21], feature engineering[22, 23], and innovative network architectures[24] to achieve attribute decoupling. In contrast, we propose a simple yet stable data-driven method. Specifically, we design a pipeline for generating high-quality data pairs that preserve identical linguistic content while exhibiting clearly distinguishable variations in one or a few attributes, such as emotion, style, accent, and paralinguistic features. By training models on such data pairs, we achieve effective attribute disentanglement, enabling to edit the attribute of input speech. Moreover, by applying multiple iterative "editing" steps, the intensity of a target attribute can be progressively enhanced or reduced. Beyond emotion, style and paralinguistic editing, we demonstrate that this approach can be extended to other applications, including speed rate adjustment, speech denoising, and silence trimming. In this report, we outline our contributions and findings:

- We present Step-Audio-EditX, the first open-source LLM-based audio model excelling at expressive and iterative audio editing, encompassing emotion, speaking style, and paralinguistics, alongside robust zero-shot TTS capabilities.

- Our results show that emotion and speaking style can be controlled through post-training with large-margin data alone, eliminating the need for extra presentation modeling or adapter modules.

- We find that post-training with large-margin data enables iterative control and high expressivity across voices, which represents a fundamental pivot from conventional representation-level disentanglement methods.

2 Architecture

2.1 Overview

In our prior work, we introduced an Audio-Edit synthesis model in Step-Audio[25] for nuanced emotional expressions and diverse speaking styles data generation. In this report, we retain the previous model along with the same audio tokenizer. The key modifications include an expanded range of emotions and speaking styles, the addition of zero-shot TTS and paralinguistic editing capabilities, as well as a reduction in model parameters from 130B to 3B. Leveraging large-margin synthetic data, our 3B model demonstrates superior and more stable performance compared to the previous version.

Our system comprises three primary components: (1) a dual-codebook audio tokenizer, which converts reference or input audio into discrete tokens; (2) an audio LLM that generates dual-codebook token sequences; and (3) an audio decoder, which converts the dual-codebook token sequences predicted by the audio LLM back into audio waveforms using a flow matching approach. This integrated architecture enables the Step-Audio-EditX to perform zero-shot TTS and diverse editing tasks within a unified framework. Thus, it can directly capitalize on the rich ecosystem of post-training techniques developed for text LLMs.
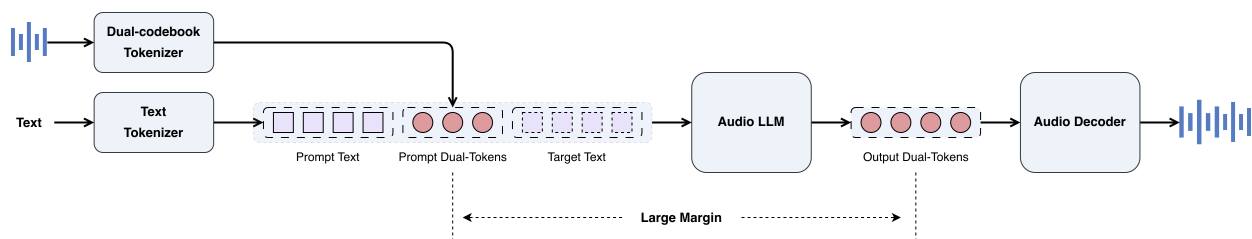

Figure 2: An overview of the architechture of Step-Audio-EditX

2.2 Audio Tokenizer

We investigate the effect of LLMs post-training with large-margin data by retaining the dual-codebook tokenization framework from our previous Step-Audio model, which employs parallel linguistic (16.7 Hz, 1024-codebook) and semantic (25 Hz, 4096-codebook) tokenizers in a 2:3 interleaving ratio. Based on a series of downstream audio tokenizer reconstruction experiments, we observed that the dual-codebook tokenizer retains a considerable amount of emotional, prosodic, and other non-linguistic information, indicating suboptimal disentanglement. This shortcoming makes it particularly suitable for validating the effectiveness of our LLM post-training strategy and the proposed large-margin data-driven methodology.

2.3 Audio LLM

The audio LLM uses the same architecture as our prior Audio-Edit model, differing only in its smaller parameter size of 3B. To capitalize on the powerful language capabilities of pre-trained text LLMs, the 3B model is initialized with a text-based LLM, followed by training on a blended dataset with a 1:1 ratio of text data to audio dual-codebook tokens. The audio LLM processes text tokens along with their corresponding dual-codebook audio tokens in a chat format, subsequently generating dual-codebook tokens as the sole output.

2.4 Audio Decoder

The audio decoder consists of a Flow Matching module and a BigVGANv2[26] vocoder. The flow matching module generates Mel spectrograms, given output audio tokens, reference audio, and speaker embedding as conditions, while the BigVGANv2 vocoder further converts the Mel spectrograms into waveforms. For the flow matching module, we adopt the diffusion transformer (DiT) as the backbone, and train the model on 200,000 hours of high-quality speech. This enhancement significantly improves its Mel spectrogram reconstruction capability, leading to substantial gains in both pronunciation accuracy and timbre similarity.

3 Data

Consistent with prior work on StepAudio's pre-training dataset and methodology, this report focus on the post-training dataset and the corresponding methods.

3.1 SFT Data

We employ SFT to enable the Step-Audio-EditX model for zero-shot TTS and diverse audio editing tasks. The SFT data can be categorized into several parts: zero-shot TTS, emotion editing, speaking style editing and paralinguistic editing. Notably, the large-margin dataset targets editing tasks, particularly on the aspects of emotion and speaking style.

3.1.1 Zero-shot Text to Speech

We employ a high-quality, professionally annotated in-house dataset, primarily in Chinese and English, for zero-shot TTS. Furthermore, a minimal amount of Cantonese and Sichuanese data is employed to elicit dialect capabilities. To ensure diverse and highly expressive styles and emotions in the synthesized speech, as well as robust zero-shot performance, The dataset captures vocal variations within individual speakers as well as across a broad speaker population, comprising approximately 60,000 unique individuals.

3.1.2 Emotion and Speaking Style Editing

Emotion and speaking style present significant challenges for expressive text-to-speech systems, due to the inherent difficulties in both defining their categorical characteristics and collecting high-quality data. We propose a straightforward and efficient large-margin synthetic data approach, which performs zero-shot voice cloning across different emotions and speaking styles for the same speaker, while ensuring a sufficiently large margin between contrastive sample pairs. Only one prompt audio segment per emotion or speaking style is required, eliminating the need for costly data collection. Moreover, this method ingeniously converts complex emotion and style descriptions into a comparative pair-based data construction format. In the following, we introduce the proposed approach:

Voice Actor Recording. Voice actors recorded expressive emotions and speaking styles. For each actor, a single audio clip of approximately 10 seconds was captured for every emotion and style combination.

Zero-shot Cloning. A triplet $\langle text_{\text{prompt}}, audio_{\text{neutral}}, audio_{\text{emotion,style}} \rangle$ is constructed for each emotion and speaking style by selecting corresponding emotional and neutral audio clips from the same speaker as the prompt audio and processing them with the StepTTS voice cloning interface, using a text instruction that describes the target attribute.

Margin Scoring. To evaluate the triplet generated, we developed a scoring model using a small, human-annotated dataset. The model evaluates audio pairs on a 1-10 scale, with higher margin scores corresponding to more desirable outcomes.

Margin Selection. Samples were selected based on a margin score threshold. This threshold was adjusted for different emotions and styles, with a score of 6 serving as the universal lower bound.

Notably, the audio clips in each triplet are generated using the same emotional or stylistic text prompt, which encourages the model to focus solely on the variations in emotion and style itself during the SFT training process.

3.1.3 Paralinguistic Editing

Paralinguistic cues, such as breathing, laughter, and filled pauses (e.g., "uhm"), are crucial for enhancing the naturalness and expressiveness of synthesized speech. We achieved paralinguistic editing capability by using a "semi-synthetic" strategy, which leverages the NVSpeech dataset[27], a highly expressive speech corpus whose rich annotations for numerous paralinguistic types enabled the construction of comparative quadruplets for model training. The quadruplet $\langle text_{\text{without\_tags}}, audio_{\text{without\_tags}}, text_{\text{nv\_source}}, audio_{\text{nv\_source}} \rangle$ construction differs from the triplet by using the NVSpeech original audio and transcript as the target output and the StepTTS voice cloning generated audio as the input, which is synthesized using the original transcript after paralinguistic tag removal.

As paralinguistic editing is an editing task performed in the time domain and exhibits substantial intrinsic margin differences, a margin scoring model is not required for data selection. A small set of quadruplet data is sufficient to effectively elicit the model's paralinguistic editing capabilities.

3.2 Reinforcement Learning Data

To align our model with human preferences, we construct two distinct types of preference datasets using different approaches: one based on human annotation, and the other employing the LLM-as-a-Judge method.

Human Annotation. We first collected real-world prompt audio and corresponding text prompts from users, and generated 20 candidate responses using the SFT model. We then constructed chosen/rejected pairs by having human annotators rate each of the 20 responses on a 5-point scale based on the criteria of correctness, prosody, and naturalness. Only pairs with a score margin greater than 3 were selected.

LLM-as-a-Judge. Model responses were scored on a 1-10 scale for emotion and speaking style editing by a comprehension model. Preference pairs were then generated from these scores, retaining only pairs with a score margin greater than 8 points in the final dataset.

These selected large-margin pairs will be used to train the reward model and PPO.

4 Training

Our post-training process aligns the model's outputs with zero-shot TTS, a variety of editing tasks, and human preferences. This alignment is accomplished through a two-stage approach: SFT followed by proximal policy optimization.

4.1 Supervised Fine-tuning

The SFT stage enhances the model's zero-shot text-to-speech synthesis and editing capabilities through the use of distinct system prompts in a chat format. In the zero-shot TTS task, the prompt waveform is encoded into dual-codebook tokens, which are subsequently detokenized into string format and incorporated into the speaker information within the system prompt. The text to be

synthesized serves as the user prompt in a chat-based format, and the generated dual-code tokens are returned as the system's response. For the editing task, all operations are defined under a unified system prompt. The user prompt includes both the original audio and a descriptive command for the editing operation, and the system response delivers the resulting edited audio tokens. The model is finetuned for one epoch with learning rate from 1×10^{-5} to 1×10^{-6}.

4.2 Reinforcement Learning

Reinforcement learning has further amplified the model's stability in zero-shot TTS, as well as its capability and expressiveness in following editing instructions. These enhancements are particularly noticeable when there is a substantial divergence between the emotional and stylistic characteristics of the source prompt waveform and the target editing output, such as generating sad speech from a happy prompt or converting loud speech into a whisper. This reinforcement learning approach offers a novel perspective to address these challenges by shifting the focus from achieving ideal speech representation disentanglement to improving both the construction of large-margin pairs and the efficacy of the reward model evaluation.

Reward Model Training. The reward model is initialized from a 3B SFT model and is trained using a combination of human-annotated and LLM-as-a-judge-generated large-margin data, optimized with the Bradley-Terry loss. The model is a token-level reward model trained directly on large-margin dual-codebook token pairs. This approach obviates the need to convert tokens back into waveform using an audio decoder during reward computation. The model is finetuned for one epoch and the learning rate is adjusted using a cosine decay strategy, initialized at 2×10^{-5} with a lower bound set at 1×10^{-5}.

PPO Training. Following the acquisition of the reward model, we employ the PPO algorithm for further training, using the same prompt seeds as in the reward model training, except for the selection of only the most challenging prompts for the SFT model. In the PPO training stage, the critic model is warmed up for 80 steps ahead of the actor. The optimizer uses an initial learning rate of 1×10^{-6}, which follows a cosine decay schedule with a lower bound of 2×10^{-7}. A PPO clip threshold of $\epsilon = 0.2$ and a KL divergence penalty with a coefficient of $\beta = 0.05$ are applied.

5 Evaluation

The accurate and comprehensive evaluation of a model's performance in synthesizing emotional, stylistic, and paralinguistic speech represents a substantial challenge. To address this, we first introduce the construction of a comprehensive and reproducible benchmark in Section 5.1. We then employ this benchmark in Section 5.2 to demonstrate the advantages of our Step-Audio-EditX model.

5.1 Evaluation Benchmark

We introduce Step-Audio-Edit-Test, a benchmark that leverages LLM-as-a-judge model to evaluate model performance on emotion, speaking style, and paralinguistics. All evaluation audio is generated via zero-shot voice cloning and subsequently scored using the Gemini-2.5-Pro[1] model.

---

[1]https://blog.google/products/gemini/gemini-2-5-pro-latest-preview/

Speaker Selection. The speaker set for zero-shot cloning consisted of eight speakers (2 male and 2 female per language, for both Chinese and English). The Chinese speakers were sourced from the Wenet-Speech4TTS[28] corpus, whereas the English speakers were sourced from the open-source GLOBE-V2[29] and Libri-Light[30] datasets, respectively.

Emotion. The emotional test set covers five categories: happiness, anger, sadness, fear, and surprise. Each category includes 50 Chinese and 50 English prompts, with the textual content of each prompt designed to be consistent with its corresponding target emotion.

Speaking Style. The test set includes seven speaking styles: childlike, elderly, exaggerated, recitative, passionate, coquettish, and whisper. Each style contains 50 Chinese and 50 English prompts, with content matched to its target style.

Paralinguistic. The paralinguistic test set includes ten paralinguistic labels per speaker: breathing, laughter, surprise-oh, confirmation-en, uhm, surprise-ah, surprise-wa, sigh, question-ei, and dissatisfaction-hnn. Each label contains 50 relevant LLM-generated samples in Chinese and 50 in English.

Emotion and Speaking Style Evaluation. To evaluate emotion and speaking style, predefined category sets (5 emotions and 7 styles) are provided to the Gemini-2.5-Pro model in the prompts, instructing it to classify the audio. The final accuracy for each category is calculated as the average across all speakers.

Paralinguistic Style Evaluation. To evaluate the performance of paralinguistic editing, a specialized evaluation prompt has been designed for the Gemini-2.5-Pro model, employing a rigorous 1–3 scoring scale (3 = perfect, 2 = flawed, 1 = failed). The prompt directs the model to actively examine specific assessment points in the audio—such as whether annotations like [laughter] or [sigh] have been accurately inserted. Particular emphasis is placed on the most common failure mode, "omission," where the audio may remain fluent but lacks required paralinguistic elements specified in the instructions. Finally, model performance in the paralinguistic editing task is assessed by calculating the overall average score generated by Gemini-2.5-Pro model.

5.2 Evaluation Results

This section details our model's performance on the Step-Audio-Edit-Test, benchmark and demonstrates its superior editing accuracy and scalability when used to edit audio generated by various closed-source TTS systems.

5.2.1 Emotion and Speaking Style Editing Results

This evaluation employs an iterative approach to audio editing for emotion and speaking style. The process begins with a zero-shot clone as the initial audio $iteration_0$, which then undergoes N rounds of iterative editing. The output of the N-th round is denoted as $iteration_N$. In this specific setup, N is configured as 3. For most use cases, two editing iterations are sufficient to meet the desired criteria.

Iterative Editing Results. As shown in Table 1, there was a significant boost in both emotion and speaking style accuracy after the initial edit of the $Iter_0$ audio. Furthermore, with successive iterations of editing, the accuracy for both emotion and speaking style was further enhanced.

Table 1: Performance of Step-Audio-EditX on Emotion and Speaking Style Editing.

Language	Emotion ↑				Style ↑			
	$Iter_0$	$Iter_1$	$Iter_2$	$Iter_3$	$Iter_0$	$Iter_1$	$Iter_2$	$Iter_3$
Chinese	57.0	71.7	74.5	**77.7**	41.6	62.1	65.8	**69.2**
English	49.9	60.5	61.5	**63.7**	50.3	62.4	**64.3**	63.1
Average	53.5	66.1	68.0	**70.7**	46.0	62.3	65.1	**66.2**
Chinese (Prompt-Fixed)	57.0	72.8	73.6	**74.0**	41.6	61.8	62.4	**62.6**
English (Prompt-Fixed)	49.9	60.4	61.3	**61.3**	50.3	**65.1**	63.1	64.0
Average	53.5	66.6	67.5	**67.7**	46.0	**63.5**	62.8	63.3

Prompt Audio Ablation. Since the performance improvement in later iterations (starting from $Iter_2$) were attributed to both the dual-code and the prompt audio. To isolate the effect of the prompt audio, we conducted an ablation study in which the prompt audio was held constant across all iterations. As presented in the Prompt-Fixed section of Table 1, the accuracy for both emotion and speaking style continues to improve with an increasing number of editing iterations. This clearly demonstrates the effectiveness of our large-margin method.

Table 2: Generalization of Emotion and Speaking Style Editing on Closed-Source Models.

Language	Model	Emotion ↑				Style ↑			
		$Iter_0$	$Iter_1$	$Iter_2$	$Iter_3$	$Iter_0$	$Iter_1$	$Iter_2$	$Iter_3$
Chinese	MiniMax-2.6-hd	71.6	78.4	77.6	**82.4**	37.9	58.6	64.0	**66.9**
	Doubao-Seed-TTS-2.0	67.4	77.8	80.4	**81.8**	39.6	60.7	**68.0**	67.1
	GPT-4o-mini-TTS	62.6	75.6	79.2	**81.4**	47.2	66.0	67.8	**71.8**
	ElevenLabs-v2	60.4	70.8	**79.0**	76.8	43.1	61.8	69.4	**70.9**
English	MiniMax-2.6-hd	55.0	64.8	64.6	**66.0**	50.9	62.1	63.1	**64.3**
	Doubao-Seed-TTS-2.0	53.8	66.0	**69.0**	65.2	49.0	60.4	61.3	**63.6**
	GPT-4o-mini-TTS	56.8	61.6	**63.2**	62.4	53.6	**63.3**	62.9	62.4
	ElevenLabs-v2	51.0	61.6	64.0	**64.0**	49.0	63.0	**63.7**	62.0
Average	MiniMax-2.6-hd	63.3	71.6	71.1	**74.2**	44.4	60.4	63.6	**65.6**
	Doubao-Seed-TTS-2.0	60.6	71.9	**74.7**	73.5	44.3	60.6	64.7	**65.4**
	GPT-4o-mini-TTS	59.7	68.6	71.2	**71.9**	50.4	64.7	65.4	**67.1**
	ElevenLabs-v2	55.7	66.2	**71.5**	70.4	46.1	62.4	**66.6**	66.5

Generalization on Closed-Source Models. The emotion and speaking style generalization of the Step-Audio-EditX model was evaluated on several leading closed-source TTS systems, including GPT-4o-mini-TTS[1], Eleven_Multilingual_v2[2], Doubao-Seed-TTS-2.0[3], and MiniMax-speech-2.6-hd[4]. For each TTS system, one male and one female built-in voice were selected for direct speech synthesis of the source text. Subsequently, three iterations of editing were applied to the resultant audio outputs. As presented in Table 2, the built-in voices of these closed-source systems possess considerable in-context capabilities, allowing them to partially convey the emotions in the text. After a single editing round with Step-Audio-EditX, the emotion and style accuracy across all voice models exhibited significant improvement. Further enhancement was observed over the next two iterations, robustly demonstrating our model's strong generalization.

---

[1] https://platform.openai.com/docs/guides/text-to-speech

[2] https://elevenlabs.io/docs/api-reference/text-to-speech/convert

[3] https://www.volcengine.com/docs/6561/1871062

[4] https://platform.minimaxi.com/docs/api-reference/speech-t2a-http

Emotion Control on Closed-Source Models. Due to the limited availability of closed-source systems with emotion and speaking style control, this section presents a comparative evaluation of Doubao-Seed-TTS-2.0 and MiniMax-speech-2.6-hd, selected for their capability in both zero-shot cloning and emotion control. To meet the minimum audio length constraints of the closed-source models and to ensure a fair evaluation, the prompt audios for all the speakers in the Step-Audio-Edit-Test were extended in duration. These extended audios were employed for zero-shot cloning followed by two emotion editing iterations. Additionally, the cloned voices were used to generate emotional speech via each closed-source model's native emotion control.The outputs from this native emotion control subsequently underwent one round of editing with our model. It can be observed from Table 3 that:

- Our Step-Audio-EditX demonstrates better emotional accuracy in its zero-shot cloning capability compared to the other two models.

- The emotional accuracy of all audio samples was significantly improved after just one editing iteration.

- One emotional editing iteration applied to the zero-shot cloned audio outperformed the results generated by the native emotion control functions of the closed-source models.

Table 3: Performance Comparison Between Step-Audio-EditX and Closed-Source Models on Emotion Editing.

Language	Model	Emotion ↑			
		$Iter_0$	$Iter_1$	$Iter_2$	$Iter_3$
Chinese	Step-Audio-EditX	57.4	71.7	**76.1**	75.6
	MiniMax-2.6-hd (Clone)	49.4	70.9	74.2	**75.9**
	MiniMax-2.6-hd (Emotion Control)	-	59.9	73.2	**76.5**
	Doubao-Seed-TTS-2.0 (Clone)	50.8	68.8	75.0	**77.2**
	Doubao-Seed-TTS-2.0 (Emotion Control)	-	51.8	71.0	**72.9**
English	Step-Audio-EditX	48.3	60.4	64.5	**64.6**
	MiniMax-2.6-hd (Clone)	50.6	60.0	61.8	**63.2**
	MiniMax-2.6-hd (Emotion Control)	-	54.0	59.5	**62.0**
	Doubao-Seed-TTS-2.0 (Clone)	47.1	58.3	60.9	**62.1**
	Doubao-Seed-TTS-2.0 (Emotion Control)	-	47.9	58.4	**61.4**
Average	Step-Audio-EditX	52.9	66.1	**70.3**	70.1
	MiniMax-2.6-hd (Clone)	50.0	65.5	68.8	**69.6**
	MiniMax-2.6-hd (Emotion Control)	-	57.0	66.4	**69.3**
	Doubao-Seed-TTS-2.0 (Clone)	49.0	63.6	68.0	**69.7**
	Doubao-Seed-TTS-2.0 (Emotion Control)	-	49.9	64.7	**67.2**

5.2.2 Paralinguistic Results

Paralinguistic editing can be considered a time-domain operation. We evaluated the effect of a single editing iteration using Step-Audio-EditX and assessed its generalization across other closed-source models.

Paralinguistic Editing Results. As shown in Table 4, a significant performance gain is obtained by adding paralinguistic tags in a single editing iteration.

Generalization on Closed-Source Models. The generalization evaluation was conducted identically to the prior one. For each closed-source model, we employed one female and one male built-in voice to synthesize speech from texts with paralinguistic labels removed. The resultant audio

Table 4: Performance of Step-Audio-EditX on Paralinguistic Editing.
(Evaluated by LLM-Judge on a 1-3 scale)

Language	Paralinguistic ↑	
	$Iter_0$	$Iter_1$
Chinese	1.80	2.89
English	2.02	2.88
Average	1.91	2.89

then underwent a single editing iteration. Additionally, for comparison, extra audio samples were synthesized by substituting paralinguistic tags with onomatopoeic words (e.g., "[Laughter]" → "haha"). After one iteration of paralinguistic editing with Step-Audio-EditX, the performance of paralinguistic reproduction is comparable to that achieved by the built-in voices of closed-source models when synthesizing native paralinguistic content directly.

Table 5: Generalization of Paralinguistic Editing on Close-Source Models.
(Evaluated by LLM-Judge on a 1-3 scale)

Language	Model	Paralinguistic ↑		
		$Iter_0$	Substitution	$Iter_1$
Chinese	MiniMax-speech-2.6-hd	1.73	2.80	2.88
	Doubao-Seed-TTS-2.0	1.67	2.81	2.88
	GPT-4o-mini-TTS	1.71	2.88	2.86
	ElevenLabs-v2	1.70	2.71	2.87
English	MiniMax-speech-2.6-hd	1.72	2.87	2.86
	Doubao-Seed-TTS-2.0	1.72	2.75	2.90
	GPT-4o-mini-TTS	1.90	2.90	2.85
	ElevenLabs-v2	1.93	2.87	2.83
Average	MiniMax-speech-2.6-hd	1.73	2.84	2.87
	Doubao-Seed-TTS-2.0	1.70	2.78	2.89
	GPT-4o-mini-TTS	1.81	2.89	2.86
	ElevenLabs-v2	1.82	2.79	2.85

The evaluation results across emotion, speaking style, and paralinguistic editing tasks confirm that our simple yet powerful approach—large-margin learning with reinforcement learning enhancement—delivers high accuracy and strong generalization. This methodology demonstrates considerable promise for both advancing research and enabling practical applications.

6 Extensions

This large-margin learning method can be straightforwardly extended to various downstream applications. By enforcing a sufficiently large margin between paired data samples, the model can rapidly acquire target editing capabilities through SFT. Reinforcement learning can then be seamlessly integrated to further enhance performance on challenging cases. This section details two practical extensions: (1) speed editing for speech rate control, and (2) denoising and silence trimming.

6.1 Speed Editing

Speed editing addresses the need for adjustable speech rates across different speakers and scenarios. This is achieved by constructing $\langle text, audio_{\text{source}}, audio_{\text{faster,slower}} \rangle$ triplet, where the speed-modified versions for a given speaker are generated through controlled speed perturbation using the SoX-toolkit[31]. Since speech rate variations directly lead to substantial disparities in token sequence lengths, even SFT alone is sufficient for effective speed editing.

6.2 Denoising and Silence Trimming

Background noise and silence segments in prompt audio can substantially influence the performance of zero-shot voice cloning. The model tends to interpret these acoustic features as part of the speaker's characteristics, subsequently reproducing them in synthesized audio. While such imitation is desirable in some use cases, it is undesirable in others. To address this, we integrated denoising and silence trimming using a generative approach, which enables targeted editing of both the prompt and the synthesized audio.

Denoising. The triplet $\langle text, audio_{\text{augment}}, audio_{\text{source}} \rangle$ is constructed for denoising, with audio$_{\text{source}}$ serving as the ground-truth reference and audio$_{\text{augment}}$ being generated through additive noise and reverberation simulation.

Silence Trimming. The triplet is defined as $\langle text, audio_{\text{source}}, audio_{\text{trimming}} \rangle$, where audio$_{\text{source}}$ corresponds to the source audio containing silent segments, and audio$_{\text{trimming}}$ refers to the processed version generated by extracting and concatenating speech segments according to the timestamps produced by Silero-VAD[32].

7 Conclusion

In this work, we present Step-Audio-EditX, an LLM-based audio model trained on large-margin data and enhanced through reinforcement learning. The model enables zero-shot TTS, iterative editing of emotion and speaking style, and paralinguistic editing. We have identified that the capabilities of LLMs and the use of large-margin data, which have often been overlooked in previous studies, allow the model to overcome the limitations of audio representations. Furthermore, the proposed framework can be easily extended to a variety of tasks, including dialect editing, accent editing, vocal editing, and imitation. Finally, it should be noted that our audio editing process is not strictly conventional "editing" in the traditional sense. Instead, it functions as a form of conditional regeneration or transfer. For tasks that require partial modifications while preserving the rest of the content, our approach provides a straightforward yet effective mask-based editing method by reconstructing paired data to ensure only specific portions of the edited tokens differ from the original sequence.

References

[1] Sanyuan Chen et al. "Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers". In: *IEEE Transactions on Audio, Speech and Language Processing* 33 (2025), pp. 705–718. DOI: 10.1109/TASLPRO.2025.3530270.

[2] Sanyuan Chen et al. "VALL-E 2: Neural Codec Language Models are Human Parity Zero-Shot Text to Speech Synthesizers". In: *CoRR* abs/2406.05370 (2024).

[3] Lingwei Meng et al. "Autoregressive Speech Synthesis without Vector Quantization". In: *CoRR* abs/2407.08551 (2024).

[4] Yuancheng Wang et al. "MaskGCT: Zero-Shot Text-to-Speech with Masked Generative Codec Transformer". In: *CoRR* abs/2409.00750 (2024).

[5] Xinsheng Wang et al. "Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens". In: *arXiv preprint arXiv:2503.01710* (2025).

[6] Matthew Le et al. "Voicebox: Text-guided multilingual universal speech generation at scale". In: *Advances in neural information processing systems* 36 (2024).

[7] Yiwei Guo et al. "VoiceFlow: Efficient Text-to-Speech with Rectified Flow Matching". In: *CoRR* abs/2309.05027 (2023). DOI: 10.48550/ARXIV.2309.05027. arXiv: 2309.05027. URL: https://doi.org/10.48550/arXiv.2309.05027.

[8] Shivam Mehta et al. "Matcha-TTS: A Fast TTS Architecture with Conditional Flow Matching". In: *ICASSP*. IEEE, 2024, pp. 11341–11345.

[9] Philip Anastassiou et al. "Seed-TTS: A Family of High-Quality Versatile Speech Generation Models". In: *CoRR* abs/2406.02430 (2024).

[10] Yushen Chen et al. "F5-TTS: A Fairytaler that Fakes Fluent and Faithful Speech with Flow Matching". In: *CoRR* abs/2410.06885 (2024).

[11] Xiaohui Sun et al. "F5R-TTS: Improving Flow Matching based Text-to-Speech with Group Relative Policy Optimization". In: *arXiv preprint arXiv:2504.02407* (2025).

[12] Zhihao Du et al. "CosyVoice: A Scalable Multilingual Zero-shot Text-to-speech Synthesizer based on Supervised Semantic Tokens". In: *CoRR* abs/2407.05407 (2024).

[13] Haohan Guo et al. "FireRedTTS: A Foundation Text-To-Speech Framework for Industry-Level Generative Speech Applications". In: *CoRR* abs/2409.03283 (2024).

[14] Zhihao Du et al. "Cosyvoice 2: Scalable streaming speech synthesis with large language models". In: *arXiv preprint arXiv:2412.10117* (2024).

[15] Kun Xie et al. *FireRedTTS-2: Towards Long Conversational Speech Generation for Podcast and Chatbot*. 2025. arXiv: 2509.02020 [cs.SD]. URL: https://arxiv.org/abs/2509.02020.

[16] Xiangyu Zhang et al. "Speaking in Wavelet Domain: A Simple and Efficient Approach to Speed up Speech Diffusion Model". In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2024, pp. 159–171.

[17] Zhihao Du et al. *CosyVoice 3: Towards In-the-wild Speech Generation via Scaling-up and Post-training*. 2025. arXiv: 2505.17589 [cs.SD]. URL: https://arxiv.org/abs/2505.17589.

[18] Dongya Jia et al. "DiTAR: Diffusion Transformer Autoregressive Modeling for Speech Generation". In: *arXiv preprint arXiv:2502.03930* (2025).

[19] Wei Deng et al. "IndexTTS: An Industrial-Level Controllable and Efficient Zero-Shot Text-To-Speech System". In: *arXiv preprint arXiv:2502.05512* (2025).

[20] Zeqian Ju et al. "NaturalSpeech 3: Zero-Shot Speech Synthesis with Factorized Codec and Diffusion Models". In: *ICML*. OpenReview.net, 2024.

[21] Tao Li et al. *Cross-speaker emotion disentangling and transfer for end-to-end speech synthesis*. 2022. arXiv: 2109.06733 [cs.SD]. URL: https://arxiv.org/abs/2109.06733.

[22] Ha-Yeong Choi, Sang-Hoon Lee, and Seong-Whan Lee. *DDDM-VC: Decoupled Denoising Diffusion Models with Disentangled Representation and Prior Mixup for Verified Robust Voice Conversion*. 2023. arXiv: 2305.15816 [eess.AS]. URL: https://arxiv.org/abs/2305.15816.

[23] Philip Anastassiou et al. "VoiceShop: A Unified Speech-to-Speech Framework for Identity-Preserving Zero-Shot Voice Editing". In: *ArXiv* abs/2404.06674 (2024). URL: https://api.semanticscholar.org/CorpusID:269032883.

[24] Dongya Jia et al. "Zero-Shot Accent Conversion using Pseudo Siamese Disentanglement Network". In: *Interspeech*. 2022. URL: https://api.semanticscholar.org/CorpusID:254564036.

[25] Ailin Huang et al. *Step-Audio: Unified Understanding and Generation in Intelligent Speech Interaction*. 2025. arXiv: 2502.11946 [cs.CL]. URL: https://arxiv.org/abs/2502.11946.

[26] Sang-gil Lee et al. *BigVGAN: A Universal Neural Vocoder with Large-Scale Training*. 2023. arXiv: 2206.04658 [cs.SD]. URL: https://arxiv.org/abs/2206.04658.

[27] Huan Liao et al. "NVSpeech: An Integrated and Scalable Pipeline for Human-Like Speech Modeling with Paralinguistic Vocalizations". In: *arXiv preprint arXiv:2508.04195* (2025).

[28] Linhan Ma et al. *WenetSpeech4TTS: A 12,800-hour Mandarin TTS Corpus for Large Speech Generation Model Benchmark*. 2024. arXiv: 2406.05763 [eess.AS]. URL: https://arxiv.org/abs/2406.05763.

[29] Wenbin Wang, Yang Song, and Sanjay Jha. *GLOBE: A High-quality English Corpus with Global Accents for Zero-shot Speaker Adaptive Text-to-Speech*. 2024. arXiv: 2406.14875 [cs.SD]. URL: https://arxiv.org/abs/2406.14875.

[30] J. Kahn et al. "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)*. IEEE, May 2020, pp. 7669–7673. DOI: 10.1109/icassp40776.2020.9052942. URL: http://dx.doi.org/10.1109/ICASSP40776.2020.9052942.

[31] cbagwell mansr robs rrt uklauer. *SoX-Sound eXchange*. https://github.com/chirlu/sox. 2024.

[32] Silero Team. *Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier*. https://github.com/snakers4/silero-vad. 2024.