FLEXIVOICE: ZERO-SHOT TEXT-TO-SPEECH WITH MULTI-MODALITY CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

The study proposes **FlexiVoice**, a zero-shot text-to-speech synthesis system speaking in any style in one voice. It is achieved by enabling multi-modality control, where speaking styles are controlled by natural language instructions and the voice timbre is provided by a speech reference.

a zero-shot text-to-speech synthesis system with multi-modality control. The multi-modality control enable a TTS system to speak in any style in one voice.

It is a challenging task to have flexible controllability of a TTS system, which can speak in any style

In this work, we study speech synthesis with multi-modality control, where natural-language instructions enable flexible style control and reference speech serves as a timbre prompt. This task faces two primary challenges: (1) the scarcity of high-quality instruction-speech data, and (2) the entanglement among content, timbre, and style, which impedes flexible control. In practice, most training examples bind the style instruction, reference timbre, and textual content under a consistent style, creating a train-inference mismatch when we aim to disentangle these control factors at inference. To address these challenges, we first construct a large-scale and diverse instruction–speech dataset using LLM-based annotation and carefully designed filtering strategies. Building on this foundation, we propose ***FlexiVoice***, an LLM-based TTS model equipped with a novel **Progressive Post-Training (PPT)** paradigm that progressively unlocks accurate and flexible controllability. Specifically, PPT introduces reinforcement learning objectives in three stages: (1) Direct Preference Optimization (DPO) to enhance multimodal alignment, (2) multi-objective GRPO to disentangle style instruction, reference timbre, and textual content for more flexible control, and (3) GRPO with the audio-language model (ALM)-based reward model to strengthen instruction following on complex real-world inputs. Each stage builds on the previous one to strengthen controllability and instruction adherence. Experimental results show that FlexiVoice surpasses open-source baselines, substantially narrows the gap with closed-source systems, and demonstrates strong capability in decoupling control factors. Human evaluations further confirm its naturalness, controllability, and robustness. Audio samples are available at https://flexi-voice.github.io/.

1 INTRODUCTION

Recent advances in text-to-speech (TTS) have been largely driven by the emergence of Large Language Models (LLMs) and refined post-training techniques. A notable breakthrough is zero-shot TTS (Du et al., 2024; Zhou et al., 2025), which enables voice cloning with only a short reference speech, effectively capturing and reproducing a speaker's timbre. Beyond timbre, controlling speaking style has become an important challenge. One direction, exemplified by Vevo (Zhang et al., 2025c), employs two separate speech references to disentangle timbre and style. Another line of work, instruction-based TTS (Vyas et al., 2023; Zhou et al., 2024b; Ji et al., 2024b), leverages natural language instructions to specify the target style. However, existing instruction-driven models often struggle either to faithfully follow the instructions or to maintain stable timbre consistency.

In this work, we aim to achieve **flexible zero-shot speech synthesis with multi-modality control**, where speech generation is jointly guided by natural-language instructions to control speaking

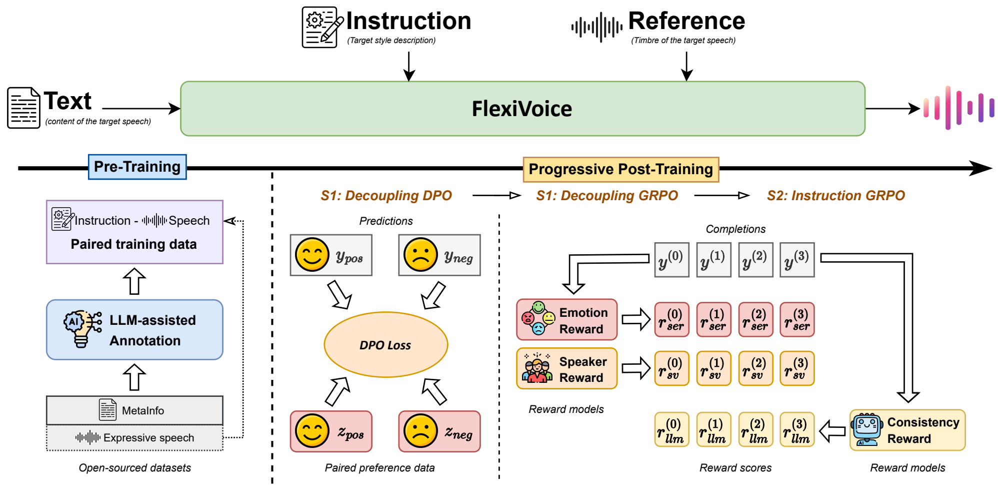

Figure 1: An overview of FlexiVoice. It takes text as input and takes natural language instruction and speech reference as optional controlling instructions to generate a speech signal. It consists of pre-training and Progressive Post-Training (PPT) stages.

style and reference speech to control timbre, enabling controllable and expressive speech generation across diverse scenarios.

Despite its promise, this task still faces two critical challenges. First, the scarcity and limited quality of instruction-speech corpora with rich style annotations limit the ability of models to follow complex and natural instructions resembling real human usage. Second, existing instruction–speech datasets still suffer from an entanglement problem: the instruction and target speech typically share the same style. As a result, models trained under such supervision tend to conflate timbre, style, and content. At inference, when a reference speech is additionally provided, the model naturally follows the style of the reference speech instead of disentangling it from the instruction, leading to a train–inference mismatch. Ideally, training data should decouple these factors, for instance, pairing a cheerful instruction with a neutral reference speech and arbitrary content, so that the model learns to integrate heterogeneous signals in a controlled manner. Without such decoupled supervision, a train-inference mismatch inevitably occurs: models trained on stylistically homogeneous data fail to generalize to the diverse combinations of instructions and references required at inference, undermining both robustness and performance.

In this work, we propose **FlexiVoice**, which is a zero-shot TTS system with multi-modality control. In particular, it can take both natural language instruction and speech reference for controllability. The natural language instruction aims to control speaking styles (e.g. emotion, speaking speed) and the speech reference is to control timbre for speaker identity. We build FlexiVoice on top of a pre-trained large language model (LLM), which equips the model with robust and comprehensive instruction-following ability. To achieve multi-modality controllability, we first construct a large-scale and diverse speech dataset with natural language instructions, named FlexiVoice-Instruct. The FlexiVoice-Instruct dataset is annotated with the help of LLM. We then pre-train FlexiVoice with Emilia (He et al., 2024) and FlexiVoice-Instruct. After that, we propose a **Progressive Post-Training (PPT)** paradigm that progressively unlocks accurate and flexible controllability after pre-training. The PPT paradigm introduces reinforcement learning objectives in three stages: (1) **Direct Preference Optimization (DPO)**, which enhances multi-modality alignment between instruction and reference speech, (2) **multi-objective Group Relative Policy Optimization (GRPO)**, which disentangles the instruction, reference speech, target text to enable flexible and independent control, and (3) **GRPO with the audio-language model (ALM)-based reward model**, which strengthens instruction-following on complex, real-world inputs. Each stage builds upon the previous one, ensuring stable optimization and progressively improving controllability and adherence to instructions.

We evaluate FlexiVoice from multi-modality controllability and instruction following ability aspects, using emotion datasets and the InstructTTSEval (Huang et al., 2025) benchmark. The experimental results demonstrate that FlexiVoice can disentangle speaking style (e.g. emotion) and speaker identity. In comparison with baselines, it achieves large gains in instruction adherence and robustness on the multi-modality control evaluation, and demonstrates strong performance on complex instruction tasks. Human evaluation results further confirm the naturalness and controllability of the generated speech.

The contributions are summarized as follows:

- We propose **FlexiVoice**, a zero-shot TTS system that has multi-modality controllability using natural-language instructions for speaking style and speech reference for timbre.
- We develop a large-scale and diverse speech dataset with natural language instruction, FlexiVoice-Instruct. The annotations are created using a large language model (i.e. Deepseek-V3 (Liu et al., 2024a)). It has a broad coverage of human-like instructions, including expressive scenarios.
- We propose a Progressive Post-Training (PPT) scheme for stable improvement. In particular, DPO is first applied to enable FlexiVoice to have multi-modality controllability. Then, the decoupling GRPO is used to disentangle speaking style and timbre. After that, the instruction GRPO is adopted to improve the instruction following ability of FlexiVoice.

2 Related Work

2.1 Controllable Speech Synthesis

Research on controllable speech synthesis has mainly followed two directions. (i) *Zero-shot TTS* clones timbre from short reference speech (Chen et al., 2024; Wang et al., 2024; Zhou et al., 2025; Zhang et al., 2025c), while partially controlling speaking style. (ii) *Instruction-based TTS* uses natural-language prompts to specify style; for instance, PromptTTS (Guo et al., 2023) and PromptStyle (Liu et al., 2023) enable text-guided style control but within limited spaces, and Parler-TTS (Lyth & King, 2024) scales conditioning with weak labels without supporting timbre control. More recent systems, such as VoxInstruct (Zhou et al., 2024b), AudioBox (Vyas et al., 2023), ControlSpeech (Ji et al., 2024b), and CosyVoice2 (Du et al., 2024), combine instruction with reference speech, but still lack robust disentanglement and broad style diversity. These limitations highlight the need for a unified framework that processes multi-modal inputs, generates speech following instruction-defined style, preserves timbre, and explicitly addresses disentanglement.

2.2 Instruction-Speech Dataset

Instruction-based TTS has driven datasets that couple text descriptions with speech. TextrolSpeech (Ji et al., 2024a) introduced large prompt–speech pairs with five style factors; Audiobox (Vyas et al., 2023) broadened the paradigm to multi-modal audio generation, though its lack of public release limited accessibility. Parler-TTS (Lyth & King, 2024) scaled weak labels for speaker/style/conditions to tens of thousands of hours. SpeechCraft (Jin et al., 2024) and ParaSpeechCaps (Diwan et al., 2025) enriched granularity with automatic captioning and diverse paralinguistic attributes. Nonetheless, most corpora come from homogeneous sources and emphasize templated descriptions, leaving insufficient coverage of natural, diverse instructions. Our dataset targets higher-quality, more natural annotations to strengthen instruction-following.

2.3 Reinforcement Learning in Speech Synthesis

Reinforcement learning has recently been explored to improve controllability in speech synthesis. INTP (Zhang et al., 2025a) applies DPO to challenging zero-shot cases for better intelligibility. Emo-DPO (Gao et al., 2025) extends preference alignment to emotional control. Vevo2 (Zhang et al., 2025b) employs multi-objective post-training to jointly enhance intelligibility and prosody across both speech and singing. These works highlight the promise of RL-based alignment for targeted aspects of TTS. In contrast, our approach adopts a progressive curriculum that leverages reinforcement learning to explicitly address modality disentanglement and complex instruction following, thereby enabling broader controllability in zero-shot multi-modality TTS.

3 OVERVIEW OF FLEXIVOICE

FlexiVoice is built on top of an LLM similar to other recent TTS systems (Du et al., 2024; Zhou et al., 2025; Zhang et al., 2025a). In FlexiVoice, a speech tokenizer converts speech into discrete tokens as inputs and outputs of an LLM. The LLM core processes the input text, natural-language instruction, and reference tokens to generate discrete speech tokens. The generated tokens are transformed into Mel-spectrogram features via flow matching (Lipman et al., 2023), and finally converted into waveform audio with a vocoder. A detailed description is provided in Appendix A.1. FlexiVoice is first pre-trained on a large-scale dataset as **FlexiVoice-Base** and then post-trained using a progressive strategy.

3.1 PRE-TRAINING

FlexiVoice-Base is a pre-trained model of FlexiVoice. It is pre-trained using Emilia (He et al., 2024) and a diverse set of instruction speech dataset as listed in Appendix A.2.

Emilia (He et al., 2024) is a large-scale, multilingual, and diverse speech generation dataset, covering a wide range of speaking styles and scenarios. It is mainly used to facilitate the fundamental speech generation ability. To empower the instruction-guided TTS capacity, we first construct a high-quality and diverse instruction dataset, FlexiVoice-Instruct, covering a wide range of scenarios and human natural-language instructions. The detailed processing procedure is descibed in Section 4. To enrich data diversity in the pre-training phase, we incorporate additional instruction–speech datasets (Diwan et al., 2025) and perform basic processing on specialized resources featuring attributes such as emotion, age, and debate scenarios. Processing details are provided in Appendix A.2. We also include NVSpeech (Liao et al., 2025) to introduce paralinguistic tags and enhance coverage of expressive scenarios. The combined corpus together forms the pre-training data.

Following the traditional pre-training strategy for TTS systems (Du et al., 2024; Zhou et al., 2025), we only train the LLM core and keep other modules frozen, without incorporating the reference speech during pre-training. As shown in Figure 3, the text and instruction are first formatted by the LLM input template. The paired ground-truth speech is preprocessed by a frozen speech tokenizer into discrete tokens, which are used to compute loss with generated tokens during pretraining. To ensure consistency in input format, we apply *Speak the following text* as the default instruction when encountering data from Emilia (He et al., 2024) and NVSpeech (Liao et al., 2025) that lacks an explicit instruction.

3.2 POST-TRAINING

3.2.1 OVERVIEW

After pre-training, the model shows solid zero-shot TTS capability but still struggles with multi-modality inputs and complex instructions. We therefore propose a **Progressive Post-Training (PPT)** scheme inspired by Curriculum Learning (Bengio et al., 2009), which starts from simpler objectives and gradually advances to harder ones for stable optimization and better generalization. PPT has three stages: **S1** aligns multi-modality controllability of instruction and reference speech in controlled emotion-centric tasks with explicit labels. **S2** disentangles the timbre and style in reference speech, and the content and style in target text, within the same scenario as S1, while **S3** extends to complex real-world instructions that are more ambiguous and harder to align. Direct optimization on such data is difficult due to scarce preference annotations and weak reward signals; as shown in Section 5.6, training only on complex instructions brings limited gains. This progressive curriculum ultimately yields a robust multi-modality instruction TTS model, **FlexiVoice**.

3.2.2 S1: MULTI-MODALITY ALIGNMENT

The first two stages target emotional instructions to reduce task complexity and explicitly address the core issue. We restrict instructions to templates like *Use* {*label*} *emotion to read it*, with labels chosen from *Neutral, Happy, Angry, Sad, and Surprised*.

For emotion-related tasks, paired preference data can be directly obtained from Speech Emotion Recognition (SER) datasets so that DPO is a feasible and effective method as the starting point.

We use the Emotional Speech Dataset (ESD) (Zhou et al., 2021), where the same speaker reads identical sentences with different emotions. Following Gao et al. (2025), for each data point, we assign a target emotion label (e.g., Happy) via an instruction template (all templates are listed in Appendix A.5), select a sentence with the target emotion as the preferred sample, the identical sentence with a different emotion (e.g., Angry) as the dis-preferred one, and use a neutral sample from the same speaker as the reference speech (an example pair is provided in Appendix A.5).

DPO directly aligns the model's emotional output with the instruction and reference speech without requiring an explicit reward model (Rafailov et al., 2023). The preference dataset $\mathcal{D}$ consists of tuples (x, y_w, y_l), where x includes instruction, text, and reference; y_w is the "winner" response that matches the instruction, and y_l is the "loser." The DPO loss is defined as:

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

where π_θ is the policy model and π_{ref} is the reference model, both initialized from **FlexiVoice-Base**.

3.2.3 S2: Decoupling of Reference Speech and Target Text

After DPO training, the model follows emotional instructions well under neutral references, but interference remains when references or texts themselves are emotion-laden, conflicted with the target emotion defined in the instruction. To explicitly suppress these effects, we adopt a multi-objective GRPO formulation. We use the same instruction templates, combine neutral and emotional clips from SER datasets as reference speech, and sample texts from NCSSD (Liu et al., 2024b) (details in Appendix A.4).

Rewards are defined as follows: (1) $r_{ser} \in (0, 1)$, the probability score for the instructed emotion given by the emotion recognition result from Emotion2vec-Large (Ma et al., 2024); (2) $r_{sv} \in \{0, 1\}$, the speaker verification result from CAM++ (Wang et al., 2023) to ensure timbre consistency. Following Zhang et al. (2025b), the advantage with multi-objective rewards is:

$$A^i_{emo} = \frac{r^i_{ser} - \text{mean}(r^i_{ser})}{\text{std}(r^i_{ser})} + \frac{r^i_{sv} - \text{mean}(r^i_{sv})}{\text{std}(r^i_{sv})}$$

where i indexes the i-th completion among K candidates for the same input x.

3.2.4 S3: Enhancement on Complex Instruction-following

The second stage enhances instruction following on complex, real-world directives beyond emotion tasks. Since paired preference data are infeasible at this scale, we directly employ GRPO. Unlike the first stage, where rewards are available from SER models, assessing the consistency between speech and open-ended instructions is more difficult. Huang et al. (2025) shows that Gemini-2.5-pro provides reliable judgments aligned with human preferences, but its use in GRPO is cost-prohibitive. We instead adopt the open-sourced Kimi-Audio-7B-Instruct (Ding et al., 2025) as the reward model due to its strong speech comprehension. It is prompted to output a binary yes/no decision on whether the generated speech matches the instruction, which is mapped into a reward $r_{llm} \in \{1, 0\}$.

For this stage, we use only instruction and text as inputs, discarding references, since references may conflict with open-ended constraints (e.g., gender) and destabilize training. Details of the data construction process are given in Appendix A.4. The single-objective advantage in this stage is $A^i_{ins} = \frac{r^i_{llm} - \text{mean}(r^i_{llm})}{\text{std}(r^i_{llm})}$. To mitigate catastrophic forgetting, we mix a small portion of S2-GRPO data during this stage. Thus, the final GRPO training becomes a multi-task, multi-objective optimization: $A^i = \begin{cases} A^i_{emo} & \text{for inputs in S1} \\ A^i_{ins} & \text{for inputs in S2} \end{cases}$

4 FlexiVoice-Instruct Dataset

4.1 Overview

To equip the model with fundamental instruction-following capability in the pre-training phase, it is essential to build a large-scale instruction–speech dataset covering diverse styles and scenarios.

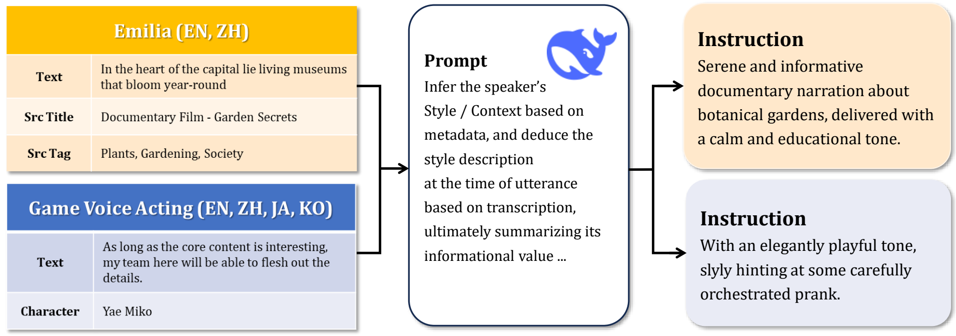

Figure 2: Processing flow and examples for the high-quality and diverse instruction-speech dataset.

We therefore construct a high-quality and diverse dataset totaling 4,316 hours, where speech-related textual metadata is processed with an LLM-based annotator. This approach enables efficient generation of natural, high-level instructions that better reflect real human usage patterns, surpassing traditional bottom-up labeling methods.

4.2 Main Process

We employ two data sources, Emilia and game voice acting, using a unified processing strategy:

Emilia (He et al., 2024) is a large-scale, diverse, in-the-wild speech dataset. Because the audio originates from video platforms and podcasts, some entries include source metadata such as video titles and tag lists. Combined with transcriptions, these cues allow us to infer speaking style and scene context, enabling the generation of open-ended instructions even without direct speech analysis. We employ Deepseek-V3 (Liu et al., 2024a) for instruction annotation. To filter noisy samples (e.g., URLs or conflicting metadata–transcription pairs), the LLM is prompted to evaluate the informational value of metadata for style and scenario inference, ensuring higher data quality.

To enhance expressiveness and stylistic richness, game voice acting data from two popular games is additionally incorporated. A distinctive feature of this data is the strong link between speaking style and character personality. LLMs, trained on large-scale web corpora, can often recognize these characters and capture their iconic styles. We thus use Deepseek-V3 to generate instructions conditioned on both transcription and character name. Concretely, we first supply the full character list and let the LLM identify those it can reliably recognize. For known characters, we prompt the LLM to infer personality and speaking style from the name, then refine the description using the transcription. As with Emilia, we incorporate an informational value check to filter noisy annotations. Prompts are provided in Appendix A.3, while the overall processing flow and examples are shown in Figure 2.

5 Experiments

5.1 Benchmarks

To evaluate the model's ability to disentangle instruction, reference speech, and text, as well as its capacity to follow complex real-world instructions, we employ one self-constructed dataset and one open-source benchmark. For disentanglement, we define two task types with two difficulty levels each, focusing on instruction-based emotional TTS, as summarized in Table 1. For complex instruction following, we adopt the InstructTTSEval (Huang et al., 2025) benchmark.

The first evaluation set is built from MEAD (Wang et al., 2020) (English) and CSEMOTIONS (Tian et al., 2025) (Chinese). In TR tasks, reference speech is randomly selected as either neutral or conflicting emotional clips from the same speaker as the ground-truth. For TO-hard, the text is replaced with sentences carrying emotions different from the target, so no ground-truth audio exists for this case. We retain five emotion categories from both datasets and randomly sample 500 examples each for English and Chinese, ensuring balanced coverage of target emotions.

Table 1: Tasks design and examples for decoupling evaluation.

Task Type	Difficulty	Text	Reference	Example
Single Modality: Text-Only (TO)	Easy	Neutral	-	Instruction: Speak it using **happy** emotion. Text: Today is Monday. Reference Speech: -
	Hard	Emotional	-	Instruction: Speak it using **happy** emotion. Text: I'm so **sad** that it's raining outside. Reference Speech: -
Multi-Modality: Text and Reference (TR)	Easy	Neutral	Neutral	Instruction: Speak it using **happy** emotion. Text: Today is Monday. Reference Speech: [A **neutral** voice]
	Hard	Neutral	Emotional	Instruction: Speak it using **happy** emotion. Text: Today is Monday. Reference Speech: [A **sad** voice]

The second benchmark, InstructTTSEval (Huang et al., 2025), contains 1,000 English and 1,000 Chinese samples. Each sample includes instructions across three distinct task levels, paired with a ground-truth speech recording. The three tasks are: Acoustic-Parameter Specification (APS), Descriptive-Style Directive (DSD), and Role-Play (RP).

5.2 METRICS

For disentanglement, we use Emotion2vec-Large (Ma et al., 2024) for SER and emotion embedding extraction. We report instruction-following accuracy with respect to the target emotion (**ACC-I**, ↑), and in the hard settings also report accuracy against conflicting emotions from text or reference (**ACC-T, ACC-R**, ↓), where lower is better. Cosine similarity between emotion embeddings of ground-truth and generated speech (**E-SIM**, ↑) further reflects adherence. To verify timbre preservation in TR tasks, we use CAM++ (Wang et al., 2023) for speaker verification accuracy (**SV**, ↑). We also evaluate intelligibility and perceptual quality: word error rate (**WER**, ↓) using Whisper-Large-V3 (Radford et al., 2023), Mean Opinion Score of speech quality (**Q-MOS**, 1–5, ↑), and Comparative MOS (**CMOS**, –2 to 2, ↑). Details of the subjective setup are provided in Appendix A.7.

For complex instruction following, we follow InstructTTSEval (Huang et al., 2025), which uses Gemini as the judge model to assess consistency, and report macro-average accuracy.

5.3 BASELINES

We compare FlexiVoice against representative open-source instruction TTS systems, including Parler-TTS (Lyth & King, 2024), reproducible variants (Ji et al., 2024b) of PromptStyle (Liu et al., 2023) and PromptTTS (Guo et al., 2023), VoxInstruct (Zhou et al., 2024b), and CosyVoice2 (Du et al., 2024), which jointly processes natural-language instructions and reference speech. In addition, we report results from InstructTTSEval (Huang et al., 2025), which cover both the above open-source systems and closed-source commercial models (gemini-2.5-flash-preview-tts, gemini-2.5-pro-preview-tts, gpt-4o-mini-tts, and Hume), and we further include MiMo-Audio-7B-Instruct, a recent large audio-language model with strong instruction-following capability.

5.4 DECOUPLING ABILITY OF FLEXIVOICE

From Table 2, we observe that FlexiVoice demonstrates strong capabilities in handling multi-modality inputs and disentangling content, timbre, and style across both English and Chinese tasks. The discussion is organized into four perspectives.

Multi-modality Control Comparing the easy sets of these two tasks, under the guidance of the instruction, TO-easy only includes the text condition (where the model uses a random timbre), while TR-easy incorporates both text and reference speech as multi-modality condition inputs (where the model uses the reference's timbre). Most baselines cannot support both simultaneously, and even those that do, together with FlexiVoice-Base, show a large gap, particularly for Chinese. After progressive post-training, FlexiVoice achieves substantial gains, reaching 97.4% ACC-I in English

Table 2: Decoupling ability evaluation result in different tasks and two languages.

Model	Text-Only as input (TO)				Text and Reference Speech as input (TR)						
	Easy		Hard		Easy			Hard			
	ACC-I↑	**E-SIM↑**	**ACC-I↑**	**ACC-T↓**	**ACC-I↑**	**E-SIM↑**	**SV↑**	**ACC-I↑**	**ACC-R↓**	**E-SIM↑**	**SV↑**
EN											
Groud-truth	93.4	1.00	-	-	93.4	1.00	-	93.4	0.6	1.00	-
Parler-TTS	44.6	0.72	12.2	42.0	-	-	-	-	-	-	-
PromptStyple	43.8	0.70	14.0	33.6	-	-	-	-	-	-	-
PromptTTS	57.8	0.79	15.0	41.0	-	-	-	-	-	-	-
CosyVoice2	-	-	-	-	65.6	0.85	**99.8**	61.0	14.4	0.84	**99.8**
VoxInstruct	70.6	0.84	17.8	41.2	58.5	0.81	89.0	49.7	23.9	0.80	90.6
FlexiVoice-Base	72.4	0.83	39.4	30.6	58.8	0.81	99.2	48.8	32.2	0.78	99.4
FlexiVoice	**97.4**	**0.89**	**89.4**	**6.6**	**89.4**	**0.90**	91.0	**78.2**	**10.6**	**0.87**	95.8
ZH											
Groud-truth	61.6	1.00	-	-	61.6	1.00	-	61.6	4.4	1.00	-
CosyVoice2	-	-	-	-	44.4	0.84	**99.8**	47.8	15.3	0.79	**100.0**
VoxInstruct	48.6	**0.76**	29.0	21.2	19.4	0.75	46.8	18.7	23.2	0.73	59.8
FlexiVoice-Base	78.4	**0.76**	66.8	14.2	25.2	0.78	99.6	22.4	38.0	0.74	99.2
FlexiVoice	**99.8**	0.72	**98.4**	**0.8**	**81.8**	**0.85**	98.8	**75.8**	**13.2**	**0.80**	98.4

Table 3: WER scores and subjective evaluation results in decoupling tasks with groud-truth.

Model	Text-Only as input (TO)			Text and Reference Speech as input (TR)					
	Easy			Easy			Hard		
	WER↓	**Q-MOS↑**	**CMOS↑**	**WER↓**	**Q-MOS↑**	**CMOS↑**	**WER↓**	**Q-MOS↑**	**CMOS↑**
EN									
Groud-truth	4.50	3.16 $\pm$0.07	0.00	4.50	3.50 $\pm$0.13	0.00	4.50	4.26 $\pm$0.22	0.00
CosyVoice2	-	-	-	**3.71**	3.50 $\pm$0.36	-0.75 $\pm$0.32	**3.60**	3.68 $\pm$0.43	-0.88 $\pm$0.21
VoxInstruct	7.29	3.02 $\pm$0.34	-0.75 $\pm$0.44	14.44	2.10 $\pm$0.23	-1.50 $\pm$0.19	12.61	2.66 $\pm$0.63	-0.86 $\pm$0.36
FlexiVoice-Base	**5.01**	3.72 $\pm$0.14	-0.12 $\pm$0.39	5.31	**3.90** $\pm$0.26	-1.25 $\pm$0.29	6.55	3.82 $\pm$0.29	-0.56 $\pm$0.31
FlexiVoice	5.99	**4.08** $\pm$0.29	**0.91** $\pm$0.20	5.23	3.62 $\pm$0.25	**0.89** $\pm$0.30	6.99	**4.06** $\pm$0.35	**0.78** $\pm$0.40
ZH									
Groud-truth	4.55	3.78 $\pm$0.14	0.00	4.55	3.38 $\pm$0.14	0.00	4.55	3.92 $\pm$0.07	0.00
CosyVoice2	-	-	-	**0.78**	3.54 $\pm$0.27	-1.14 $\pm$0.24	**0.89**	3.58 $\pm$0.39	0.36 $\pm$0.29
VoxInstruct	**3.37**	3.18 $\pm$0.28	-1.60 $\pm$0.30	10.04	2.62 $\pm$0.28	-1.88 $\pm$0.09	9.40	3.04 $\pm$0.17	-1.62 $\pm$0.19
FlexiVoice-Base	5.01	**4.10** $\pm$0.26	**0.75** $\pm$0.40	3.08	3.74 $\pm$0.28	-1.50 $\pm$0.23	5.63	3.66 $\pm$0.06	-0.88 $\pm$0.45
FlexiVoice	7.59	4.04 $\pm$0.23	0.60 $\pm$0.41	4.34	**3.79** $\pm$0.26	**-0.36** $\pm$0.24	7.02	**3.76** $\pm$0.29	**0.57** $\pm$0.44

and 99.8% in Chinese on TO-easy, surpassing ground-truth in some cases. On TR-easy, it maintains strong accuracy (89.4% EN, 81.8% ZH) while preserving speaker consistency, highlighting robust multi-modality control capacity.

Text Disentanglement For the two difficulty levels of TO, the easy level uses neutral text while the hard level employs text with emotions differing from the instruction. The purpose is to test whether the model can ignore conflicting emotional cues in text. Baselines and FlexiVoice-Base generally fail here, showing low ACC-I and high ACC-T, indicating they are biased toward the text's implied style. In contrast, FlexiVoice substantially improves disentanglement, achieving 89.4% ACC-I with only 6.6% ACC-T in English, and 98.4% vs. 0.8% in Chinese. Notably, the Chinese gap between easy and hard settings is reduced to just 1.4%, showing FlexiVoice's strong ability to ensure style is controlled by instruction alone.

Reference Speech Disentanglement For TR, the easy task uses neutral speech as the timbre reference, while the hard uses emotionally charged reference speech differing from the instruction. Most baselines exhibit large drops in accuracy and high ACC-R, showing disruption by reference's style. FlexiVoice alleviates this issue, achieving 78.2% ACC-I (EN) and 75.8% (ZH) with correspondingly low ACC-R (10.6% EN, 13.2% ZH). This demonstrates that FlexiVoice effectively disentangles reference timbre from style, preserving timbre while adhering to instruction-defined emotion.

Table 4: Complex instruction-following evaluation result.

Model	InstructTTSEval-EN				InstructTTSEval-ZH			
	APS	DSD	RP	Avg.	APS	DSD	RP	Avg.
Groud-truth	96.2	89.4	67.2	84.3	90.9	86.7	69.8	82.5
Closed-sourced								
Gemini-flash	92.3	93.8	80.1	88.7	88.2	90.9	77.3	85.4
Gemini-pro	87.6	86.0	67.2	80.3	89.0	90.1	75.5	84.8
GPT-4o-mini-TTS	76.4	74.3	54.8	68.5	54.9	52.3	46.0	51.1
Hume	83.0	75.3	54.3	71.1	-	-	-	-
Open-sourced								
ParlerTTS	60.0	45.9	31.2	45.7	-	-	-	-
PromptStyle	57.4	46.4	30.9	38.2	-	-	-	-
VoxInstruct	54.9	57.0	39.3	50.4	47.5	52.3	42.6	47.5
PromptTTS	64.3	47.2	31.4	47.6	-	-	-	-
MiMo-Aduio-7B-Instruct	80.6	77.6	59.5	72.6	**75.7**	**74.3**	61.5	70.5
FlexiVoice-Base	63.6	75.0	60.6	66.4	56.7	59.1	59.5	58.4
FlexiVoice	**81.2**	**85.2**	**71.4**	**79.3**	71.0	71.8	**69.7**	**70.8**

WER and Subjective Evaluation Table 3 presents intelligibility and perceptual quality. FlexiVoice slightly increases WER compared to FlexiVoice-Base, which may result from ASR models under-performing on expressive speech. Nevertheless, it consistently achieves higher Q-MOS (4.08 EN, 4.04 ZH) and positive CMOS (up to +0.9), indicating better overall expressiveness and instruction adherence, ensuring that the quality of generated speech remains unaffected at the same time. In contrast, baselines often obtain negative CMOS, showing lower preference in human judgments compared with the groud-truth. These results confirm that FlexiVoice not only achieves superior disentanglement but also maintains naturalness and quality close to ground-truth.

5.5 Complex Instruction-following Ability

InstructTTSEval (Huang et al., 2025) defines three levels of complex instruction following. These tasks range from low-level acoustic control to open-ended style generation and character imitation, thereby testing different aspects of instruction adherence. As shown in Table 4, the pre-trained model FlexiVoice-Base already performs competitively, especially on DSD (75.0 EN) and RP (60.6 EN), due to alignment between these tasks and the natural instructions present in our constructed dataset. After applying the Progressive Post-Training strategy, FlexiVoice achieves consistent gains across all task types. On English tasks, it improves by over 12 points on APS (81.2 vs. 63.6) and nearly 11 points on RP (71.4 vs. 60.6), reaching an overall average of 79.3, close to Gemini-pro (80.3). On Chinese tasks, FlexiVoice attains 70.8 average accuracy, surpassing MiMo-Audio-7B-Instruct (70.5) and reducing the gap to Gemini models. Overall, FlexiVoice consistently outperforms all open-source baselines and narrows the gap with closed-source commercial systems, demonstrating robust instruction-following ability in complex, real-world scenarios.

5.6 Effectiveness of Progressive Post-Training

Table 5: Both decoupling and complex instruction-following evaluation results of English subset for models in different training stages. Note that ① is our pre-trained model, and ③→④→⑤ is our progressive post-training (PPT) strategy so that ⑤ is our final model, FlexiVoice.

No.	Model	Decoupling Set (EN)					InstructTTSEVal (EN)			
		TO Easy	TO Hard	TR Easy	TR Hard	Avg.	APS	DSD	RP	Avg.
①	FlexiVoice-Base	72.4	39.4	58.8	48.8	54.9	63.6	75.0	60.6	66.4
②	+ *instruction GRPO*	72.8	59.0	48.6	38.2	54.7	73.3	78.8	64.7	72.3
③	+ *decoupling DPO*	96.0	81.4	83.3	72.2	83.2	63.4	79.3	64.2	69.0
④	+ *decoupling DPO & GRPO*	96.4	88.0	89.4	80.2	88.5	66.8	80.3	67.9	71.7
⑤	+ *decoupling DPO & GRPO* + *instruction GRPO*	97.7	89.4	89.4	78.2	88.7	81.2	85.2	71.4	79.3

Table 5 reports results on the English subset for both disentanglement and InstructTTSEval (Huang et al., 2025). The pre-trained model (①) provides only moderate performance. Adding instruction GRPO alone (②) slightly improves instruction-following, reflecting the difficulty of aligning directly on complex instructions. Decoupling DPO (③) yields large gains on disentanglement (+28.3 avg.), while combining DPO with GRPO (④) further boosts disentanglement (88.5 avg.) and modestly improves instruction-following (71.7 avg.). The full PPT curriculum (⑤) achieves the best balance, with 88.7 average disentanglement accuracy and 79.3 instruction-following accuracy, representing +34 and +13 improvements over the base. These results show that progressively moving from disentanglement to complex instruction following yields a stable and effective optimization path.

6 Conclusion

In this work we introduce FlexiVoice, a zero-shot TTS system for multi-modality control that uses natural-language instructions for style and reference speech for timbre, supported by a large LLM-guided instruction–speech dataset and a Progressive Post-Training (PPT) paradigm. PPT first strengthens disentanglement of instruction, text, and reference via emotion-centric DPO/GRPO, then scales to complex instruction following with an ALM-based reward, yielding stable optimization and broad controllability. Experiments show consistent gains on disentanglement (e.g., large ACC-I improvements with low interference from text/reference) and strong performance on InstructTTSEval, where FlexiVoice surpasses open-source baselines and narrows the gap to closed-source systems, while maintaining naturalness and robustness in human evaluations.

7 Ethics Statement

This work investigates controllable text-to-speech with multi-modality inputs. The instruction–speech dataset was created from publicly available or licensed resources, filtered to remove offensive, biased, or harmful content. No personal or sensitive user data were collected or processed. While advances in speech synthesis can potentially be misused (e.g., generating deceptive or harmful audio), our study focuses on improving controllability, robustness, and transparency for research purposes, and we encourage responsible use of the proposed methods. All authors have read and adhered to the ICLR Code of Ethics , and confirm that this research complies with standards of fairness, privacy, and research integrity.

8 Reproducibility statement

We have taken multiple steps to ensure the reproducibility of our work. Detailed descriptions of the model architecture, training objectives, and evaluation protocols are provided in the main text and appendix. We will release the instruction–speech dataset, model checkpoints, and all training and inference code to facilitate replication and further research. Hyperparameter settings, data processing procedures, and evaluation scripts will also be included in the release materials.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

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

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*, 2024.

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

Anuj Diwan, Zhisheng Zheng, David Harwath, and Eunsol Choi. Scaling rich style-prompted text-to-speech datasets, 2025. URL https://arxiv.org/abs/2503.04713.

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

Xiaoxue Gao, Chen Zhang, Yiming Chen, Huayun Zhang, and Nancy F Chen. Emo-dpo: Controllable emotional speech synthesis through direct preference optimization. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. Prompttts: Controllable text-to-speech with text descriptions. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

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

Kexin Huang, Qian Tu, Liwei Fan, Chenchen Yang, Dong Zhang, Shimin Li, Zhaoye Fei, Qinyuan Cheng, and Xipeng Qiu. Instructtttseval: Benchmarking complex natural-language instruction following in text-to-speech systems. *arXiv preprint arXiv:2506.16381*, 2025.

Yiqiao Huang, Yuancheng Wang, Jiaqi Li, Haotian Guo, Haorui He, Shunsi Zhang, and Zhizheng Wu. Debatts: Zero-shot debating text-to-speech synthesis. *arXiv preprint arXiv:2411.06540*, 2024.

Shengpeng Ji, Jialong Zuo, Minghui Fang, Ziyue Jiang, Feiyang Chen, Xinyu Duan, Baoxing Huai, and Zhou Zhao. Textrolspeech: A text style control speech corpus with codec language text-to-speech models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10301–10305. IEEE, 2024a.

Shengpeng Ji, Jialong Zuo, Wen Wang, Minghui Fang, Siqi Zheng, Qian Chen, Ziyue Jiang, Hai Huang, Zehan Wang, Xize Cheng, et al. Controlspeech: Towards simultaneous zero-shot speaker cloning and zero-shot language style control with decoupled codec. *arXiv preprint arXiv:2406.01205*, 2024b.

Zeyu Jin, Jia Jia, Qixin Wang, Kehan Li, Shuoyi Zhou, Songtao Zhou, Xiaoyu Qin, and Zhiyong Wu. Speechcraft: A fine-grained expressive speech dataset with natural language description. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 1255–1264, 2024.

Jiaqi Li, Xiaolong Lin, Zhekai Li, Shixi Huang, Yuancheng Wang, Chaoren Wang, Zhenpeng Zhan, and Zhizheng Wu. Dualcodec: A low-frame-rate, semantically-enhanced neural audio codec for speech generation. *arXiv preprint arXiv:2505.13000*, 2025.

Huan Liao, Qinke Ni, Yuancheng Wang, Yiheng Lu, Haoyue Zhan, Pengyuan Xie, Qiang Zhang, and Zhizheng Wu. Nvspeech: An integrated and scalable pipeline for human-like speech modeling with paralinguistic vocalizations. *arXiv preprint arXiv:2508.04195*, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *11th International Conference on Learning Representations, ICLR 2023*, 2023.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

Guanghou Liu, Yongmao Zhang, Yi Lei, Yunlin Chen, Rui Wang, Lei Xie, and Zhifei Li. Prompt-style: Controllable style transfer for text-to-speech with natural language descriptions. In *Proc. Interspeech 2023*, pp. 4888–4892, 2023.

Rui Liu, Yifan Hu, Yi Ren, Xiang Yin, and Haizhou Li. Generative expressive conversational speech synthesis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 4187–4196, 2024b.

Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

R. Lotfian and C. Busso. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483, October-December 2019. doi: 10.1109/TAFFC.2017.2736999.

Dan Lyth and Simon King. Natural language guidance of high-fidelity text-to-speech with synthetic annotations, 2024.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, ShiLiang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 15747–15760, 2024.

Kari Ali Noriy, Xiaosong Yang, and Jian Jun Zhang. Emns/imz/corpus: An emotive single-speaker dataset for narrative storytelling in games, television and graphic novels. *arXiv preprint arXiv:2305.13137*, 2023.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, 2019.

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

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

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

Zhiyuan Tang, Dong Wang, Yanguang Xu, Jianwei Sun, Xiaoning Lei, Shuaijiang Zhao, Cheng Wen, Xingjun Tan, Chuandong Xie, Shuran Zhou, Rui Yan, Chenjia Lv, Yang Han, Wei Zou, and Xiangang Li. Kespeech: An open source speech dataset of mandarin and its eight subdialects. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Fengping Tian, Chenyang Lyu, Xuanfan Ni, Haoqin Sun, Qingjuan Li, Zhiqiang Qian, Haijun Li, Longyue Wang, Zhao Xu, Weihua Luo, et al. Marco-voice technical report. *arXiv preprint arXiv:2508.02038*, 2025.

Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, et al. Audiobox: Unified audio generation with natural language prompts. *arXiv preprint arXiv:2312.15821*, 2023.

Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. Cam++: A fast and efficient network for speaker verification using context-aware masking. In *INTERSPEECH*, 2023.

Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*, 2020.

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

Xueyao Zhang, Yuancheng Wang, Chaoren Wang, Ziniu Li, Zhuo Chen, and Zhizheng Wu. Advancing zero-shot text-to-speech intelligibility across diverse domains via preference alignment. *arXiv preprint arXiv:2505.04113*, 2025a.

Xueyao Zhang, Junan Zhang, Yuancheng Wang, Chaoren Wang, Yuanzhe Chen, Dongya Jia, Zhuo Chen, and Zhizheng Wu. Vevo2: Bridging controllable speech and singing voice generation via unified prosody learning. *arXiv preprint arXiv:2508.16332*, 2025b.

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

Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna. L2-arctic: A non-native english speech corpus. In *Proc. Interspeech*, pp. 2783–2787, 2018. doi: 10.21437/Interspeech.2018-1110. URL `http://dx.doi.org/10.21437/Interspeech.2018-1110`.

Jinming Zhao, Tenggan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou Li. M3ed: Multi-modal multi-scene multi-label emotional dialogue database. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5699–5710, 2022.

Jiaming Zhou, Shiyao Wang, Shiwan Zhao, Jiabei He, Haoqin Sun, Hui Wang, Cheng Liu, Aobo Kong, Yujie Guo, and Yong Qin. Childmandarin: A comprehensive mandarin speech dataset for young children aged 3-5. *arXiv preprint arXiv:2409.18584*, 2024a.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 920–924. IEEE, 2021.

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

Yixuan Zhou, Xiaoyu Qin, Zeyu Jin, Shuoyi Zhou, Shun Lei, Songtao Zhou, Zhiyong Wu, and Jia Jia. Voxinstruct: Expressive human instruction-to-speech generation with unified multilingual codec language modelling. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 554–563, 2024b.

A Appendix

A.1 Model Structure

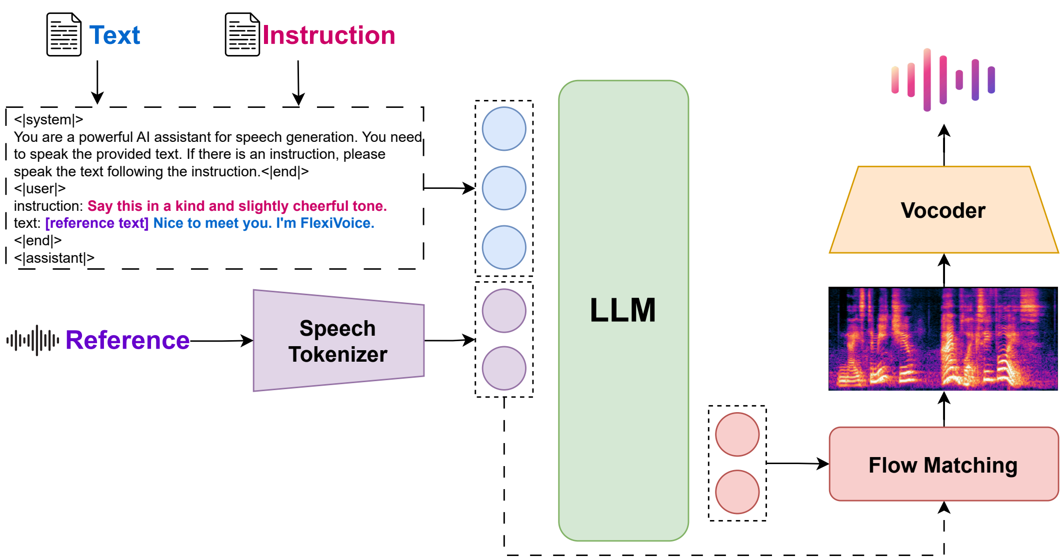

Figure 3: The complete structure of FlexiVoice.

As illustrated in Section **??**, our model mainly contains two stages: auto-regressive LLM and flow matching. In the first stage, the model receives the inputs of text, instruction, and reference speech. The text and instructions are formatted according to the LLM's input template, with the reference speech transcription concatenated before the text. We use the semantic code extracted from DualCodec (Li et al., 2025) to represent the speech in discrete form within our system. Therefore, the reference speech will first be converted into discrete tokens via DualCodec, and the tokens resulting from processing the formatted text and instructions through the LLM encoder will be concatenated to the front. Together, they will serve as input for the auto-regressive LLM. Here we employ Phi-3.5-mini-instruct (Abdin et al., 2024) due to its suitability for multi-modal tasks. We first expand its vocabulary (equivalent to DualCodec's vocabulary size, 16384), then use the parameters of a text LLM as the initial state for pre-training and post-training, illustrated as the main part in our work.

Mainstream TTS works (Du et al., 2024; Zhou et al., 2025; Zhang et al., 2025a) chose flow matching (Lipman et al., 2023) as the decoder because of its high-quality reconstruction of speech details. For the second stage in this work, we employ a flow matching module trained on Emilia (He et al., 2024) to convert generated code into mel-spectrum, using reference speech code as the condition. Finally, the mel-spectrum is converted into the target audio via a vocoder (using Vocos (Siuzdak, 2023) in this case). The model structure is shown in Figure 3

A.2 DATA PROCESS FOR PRE-TRAINING

Table 6: Instruction-speech datasets used in pre-training stage.

Source	Dur (h)	Description
ParaSpeechCaps (Diwan et al., 2025)	2847	Open-sourced style-prompted speech data
ChildMandarin (Zhou et al., 2024a)	41	Speech clips of child's voice
Debatts (Huang et al., 2024)	67	Speech clips with debating style
Emotion set*	117	Speech clips with different emotions
KeSpeech (Tang et al., 2021)	1541	Speech clips with different dialects, ages, and genders
L2-ARCTIC (Zhao et al., 2018)	27	Speech corpus of non-native English
NVSpeech (Liao et al., 2025)	775	Paralinguistic tagged speech data
FlexiVoice-Instruct	4316	Expressive speech with high-level instructions

* Including CREMA-D (Cao et al., 2014), EMNS (Noriy et al., 2023), ESD (Zhou et al., 2021), IEMOCAP (Busso et al., 2008), M3ED (Zhao et al., 2022), MELD (Poria et al., 2019), MSP-Podcast (Lotfian & Busso, 2019), RAVDESS (Livingstone & Russo, 2018).

For ParaSpeechCaps (Diwan et al., 2025), due to its bottom-up data annotation process, some speech entries possess a detailed acoustic feature dictionary alongside a summarized description. For these data points, we concatenate the feature dictionaries together and randomly sample descriptions. For other data, we utilize their descriptions as instructions.

For ChildMandarin (Zhou et al., 2024a), Detatts (Huang et al., 2024), and the emotion set, since they are all single-label datasets (boy, debate scene, emotion label), we pre-generated several instruction templates using Deepseek-V3 (Liu et al., 2024a) (e.g., "Speak in the voice of {label}," "Imagine you are in a debate scene," "Express the emotion of {label}"). Then, for each speech data point, we randomly selected a template and filled in the corresponding label.

KeSpeech (Tang et al., 2021) and L2-ARCTIC (Zhao et al., 2018) are both multi-label datasets. For example, each speech in KeSpeech has three attributes: age, gender, and accent. For each attribute triplet across the entire dataset, we use Deepseek-V3 (Liu et al., 2024a) to generate three Chinese instructions and three English instructions, which are then randomly applied to the entire dataset.

For NVSpeech, which is a speech corpus with paralinguistic tags, we do not apply any instruction but use their tags to enlarge the vocabulary size of our core LLM module to guarantee expressive paralinguistic generation. All of the instruction-speech paired datasets in the pre-training stage are listed in Table 6.

A.3 Prompts in Instruction Data Construction

For data source of Emilia

Role and Tasks

```
You are a multilingual text analysis expert tasked with generating voice descriptions
required for users' Instruction TTS tasks.  Your core responsibility is:  Generating
voice descriptions based on text and metainfo (including but not limited to scenarios,
styles, emotions, etc.)
```

Detailed Description

```
- The metainfo includes the video title and a list of video tags from which the audio
originates.  If the title is irrelevant (e.g., URLs, gibberish, etc.), ignore the
title.
- The scene and speaking style can be inferred from the tags and title, but in case of
conflict, prioritize the text content.
- Output a concise description of the speech in natural human language instructions.
- Use colloquial, vivid phrasing.  Vary sentence structures to avoid template-like
patterns.
- Determine whether valid or rich speech descriptions can be generated based on the
text and meta information and classify the information value (high/medium/low).
```

Input/Output Specifications

```
Input Structure:
{
    "speech transcript":  "The speech text to analyze",
    "metainfo":  {"title":  [Title of the source video], "tags":  [Tag list of the
source video]}
}
Output Requirement:
{
    "description":  "Natural language description of the speech style",
    "value":  "information value (high/medium/low)"
}
```

Example

```
Example Input:
{
    "speech transcript":  "This match currently shows Ma Chao with four kills, nearly
6,000 in gold.",
    "metainfo":  {"title":  "Honor of Kings KPL Autumn Finals", "tags":
"Game,esports,Honor of Kings,LGD,KPL,DYG"}
}
Example Output:
{
    "description":  "Professional and passionate esports commentary, delivered with
excitement and enthusiasm.",
    "value":  "high"
}
```

Core Constraints

```
- When metainfo conflicts with text, prioritize text content.
- Prohibit any unfounded speculation; prohibit all uncertain detail descriptions.
- The example outputs are for reference only; do not rigidly adhere to their phrasing.
```

For data source of game voice acting

Role and Tasks

You are a multilingual text analysis expert tasked with generating voice descriptions required for users' Instruction TTS tasks. Your core responsibility is: **Generating voice descriptions based on text and metainfo (including but not limited to scenarios, styles, emotions, etc.)**

Detailed Description

- Metainfo refers to a specific character in [GameName], whose dialogue is the speech transcript.
- Infer the character's speaking style based on their name and output a description of the speech.
- When the speaking style conflicts with the text's expressed style, prioritize the text.
- Do NOT include specific character names in the description.
- Use colloquial, vivid phrasing. Vary sentence structures to avoid template-like patterns.
- Determine whether valid or rich speech descriptions can be generated based on the text and meta information and classify the information value (high/medium/low).

Input/Output Specifications

```
Input Structure:
{
    "speech transcript": "Speech transcript (dialogue from a [GameName] character)",
    "metainfo": "Source information (character name in [GameName])"
}
Output Requirement:
{
    "description": "Natural language description of the speech style",
    "value": "information value (high/medium/low)"
}
```

Example

```
Example Input:
{
    "speech transcript": "As long as the content is interesting, these experts can
help spice it up.",
    "metainfo": "Yae Miko"
}
Example Output:
{
    "description": "With a hint of playful elegance in her tone, as if she had long
seen through the other's thoughts yet chose to hint at them subtly.",
    "value": "high"
}
```

Core Constraints

- When metainfo conflicts with text, prioritize text content.
- Prohibit any unfounded speculation; prohibit all uncertain detail descriptions.
- The example outputs are for reference only; do not rigidly adhere to their phrasing.

A.4 Data Construction for GRPO Training

Decoupling Task The training set is constructed from the recording subset of NCSSD (Liu et al., 2024b), comprising about 20,000 Chinese and 10,000 English dialogue speech samples. We selected this dataset primarily because its transcriptions align closely with everyday conversations and implicitly carry emotional tendencies in most cases. We randomly assign emotion labels to each data point, thereby generating two types of data: those where the instruction and transcription emotions align, and those where they conflict. This approach helps guide the model to learn the ability to decouple content and style within text. Simultaneously, we randomly assign neutral and emotional speech from the pre-training data's emotion set (6) to each data point with a 90% and 10% probability, respectively, as reference speech. This enables the model to learn to decouple timbre and style from the references. We utilize half of the Chinese data and all of the English data, forming approximately 20,000 balanced data points as the GRPO training set for the decoupling phase.

Complex Instruction To enhance the model's performance in complex instruction-following tasks, we constructed a rich and diverse GRPO training dataset. Specifically, for each language (Chinese and English), we first randomly sampled 1,000 existing instruction-text inputs from the pre-training data to ensure optimization stability. Next, we prompted Deepseek-V3 (Liu et al., 2024a) to randomly generate 6,000 instructions across three distinct configurations: (1) Generate detailed and comprehensive acoustic feature descriptions in dictionary format, (2) Generate natural language descriptions incorporating 3/4/5/6 explicit acoustic features, (3) Freely generate instructions for arbitrary scenarios consistent with human usage patterns. Collectively, we generated 14,000 instruction-text paired inputs as GRPO training data for the second stage.

A.5 DETAILS IN EMOTIONAL CONTROL DPO

Instruction Templates

- Say the sentence with the emotion of {label}.
- Say this with a {label} tone.
- Speak this sentence in a {label} manner.
- Deliver this text with {label} emotion.
- Use a {label} voice.
- Express this sentence with {label} feeling.
- Read it with {label} inflection.
- Recite this with {label} sentiment.
- Voice this in a {label} style.
- {label} (Use only one emotion label)

- 用{label}的情感说出这句话。
- 以{label}的情绪表达这句话。
- 带着{label}的情感说出这个文本。
- 使用{label}的情感。
- 以{label}的心情表达这句话。
- 用{label}的情感色彩说出这句文本。
- 带着{label}的感情说出这句话。
- 用{label}的情感基调传达此句。
- 以{label}的情绪状态说出这句话。
- {label}

One Paired Sample (emo=angry, speaker_id=0013)

{
”prompt_text”: ”Chapter eleven on the doorstep.”,
”prompt_wav_path”: ”Emotion_Speech_Dataset/**0013**/<u>Neutral</u>/0013_000267.wav”,
”target_text”: ”What are you waiting for? man.”,
”instruction”: ”Say this with a <u>angry</u> tone.”,
”chosen_wav_path”: ”Emotion_Speech_Dataset/**0013**/<u>Angry</u>/0013_000697.wav”
”rejected_wav_path”: ”Emotion_Speech_Dataset/**0013**/<u>Sad</u>/0013_001397.wav”
}

A.6 REWARD SELECTION IN DECOUPLING GRPO

In decoupling GRPO, we first use a SER signal as one of the rewards to control the correct emotion synthesis. With the reference speech input, we also need to guarantee that the generated result is from the same speaker with the reference. There are two choices: speaker verification signal (0 or 1) and speaker similarity signal (values between 0 and 1), using CAM++ (Wang et al., 2023) and WavLM-large-finetuned (Chen et al., 2022) respectively. Given the DPO optimized model, we train the multi-objective GRPO using SER reward with two separate speaker-related rewards for four epochs, then test them on two difficulty levels of TR in the decoupling evaluation set, as shown in Figure 4.

Results show that models using speaker verification as the reward signal achieve significant improvements, while those using speaker similarity yield largely unchanged or even reduced performance. This occurs because models computing speaker embeddings (WavLM in this case) do not rely solely on speech timbre but incorporate additional acoustic features beyond timbre, such as pitch and emotion. When humans speak with varying emotions, acoustic attributes like pitch, speech rate, and volume inevitably differ. Consequently, expressive emotion delivery yields high SER reward scores but low speaker similarity signals. This phenomenon can thus be interpreted as a conflict between

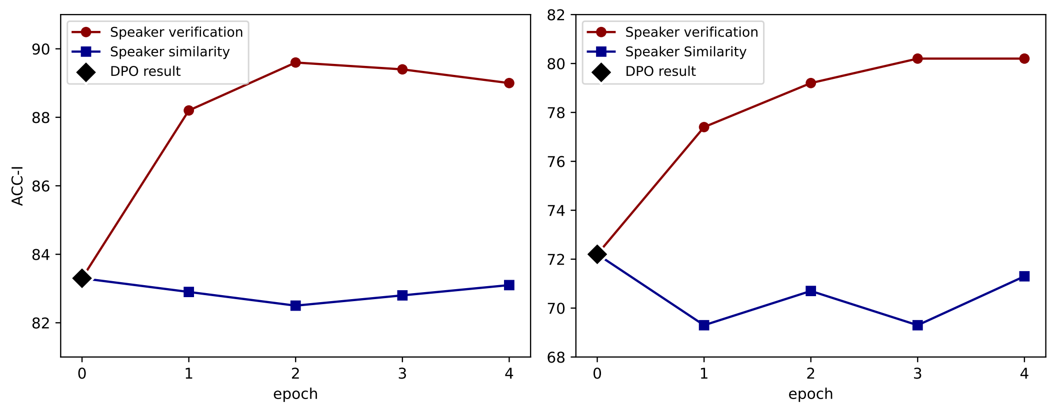

Figure 4: Comparison of results for two reward signals in decoupling GRPO, on the decoupling evaluation set's TR-easy (left) and TR-hard (right) tasks.

SER reward and speaker similarity reward, hindering the normal optimization process of the multi-objective GRPO task.

Speaker verification signals, as a more loose form of speaker similarity, can accommodate models that generate more expressive emotional speech while ensuring speaker identity. Therefore, we select speaker verification as one of the final reward signals in the S1 GRPO stage.

A.7 Subjective Evaluation Configuration

Subjective evaluations are carried out by a group of compensated participants, all of whom had strong speech and audio expertise. We selected two baseline models with relatively comprehensive task support, our pre-trained model, and FlexiVoice for subjective evaluation. For each task in each language, we randomly select five paired samples from each model and the ground-truth. To ensure reliability, every audio sample was independently rated by at least three different individuals for both subjective evaluation.

Q-MOS Participants are asked to evaluate the overall quality of each generated speech sample on a 5-point scale, considering aspects such as clarity, naturalness, and absence of distortion/artifacts, ignoring the style instruction and timbre reference. The meaning of each score is defined as follows:

- **5** - Speech is highly natural, clear, and pleasant to listen to. No noticeable artifacts or distortions. Comparable to professionally recorded human speech.
- **4** – Speech is generally natural and intelligible, with only minor imperfections or occasional artifacts that do not interfere with understanding or listening comfort.
- **3** – Speech is intelligible but has moderate issues such as slight distortion, unnatural prosody, or mild background noise. Quality is acceptable but clearly below high-standard human recordings.
- **2** – Speech is somewhat difficult to understand due to significant artifacts, distortions, or unnatural delivery. Quality issues noticeably affect the listening experience.
- **1** – Speech is largely unintelligible or highly unnatural, with severe artifacts or distortions that make evaluation difficult.

CMOS In this task, participants are asked to compare two audio samples (one ground-truth and one generated by a model) under the condition of a given target emotion. The primary focus is on the richness and naturalness of emotional expression, which reflects the instruction-following and decoupling ability of the models.

The additional evaluation rules are: (1) When no reference speech is provided, ignore timbre consistency. (2) When reference speech is provided, timbre similarity is a secondary criterion (acceptable as long as both samples sound like the same speaker). (3) Slight mispronunciations or noise should be disregarded; the main comparison is the accuracy and expressiveness of emotion. And the scoring scale is defined as follows: (Note that the order of paired audio demonstrations is random.)

- $+$**2** - Audio B is much better than Audio A in conveying the target emotion.
- $+$**1** – Audio B is slightly better than Audio A in emotional expression.
- **0** – Both samples are comparable in terms of emotional richness and naturalness.
- $-$**1** – Audio A is slightly worse than Audio B in emotional expression.
- $-$**2** – Audio A is much worse than Audio B in emotional expression.

A.8 LLM Usage

Large Language Models (LLMs) were used solely as an assistive tool for grammar correction and language polishing of the manuscript. They did not contribute to research ideation, methodology, experiments, analyses, or the generation of scientific content. All conceptual and technical contributions are entirely those of the authors.