

# ReStyle-TTS: Relative and Continuous Style Control for Zero-Shot Speech Synthesis

Haitao Li<sup>1,2</sup>, Chunxiang Jin<sup>3</sup>, Chenglin Li<sup>1,2</sup>, Wenhao Guan<sup>4,2</sup>, Zhengxing Huang<sup>1</sup>, Xie Chen<sup>5,2</sup>

<sup>1</sup>Zhejiang University, <sup>2</sup>Shanghai Innovation Institute, <sup>3</sup>Ant Group,

<sup>4</sup>Xiamen University, <sup>5</sup>Shanghai Jiao Tong University

lihaitao@zju.edu.cn, chenxie95@sjtu.edu.cn

## Abstract

Zero-shot text-to-speech models can clone a speaker’s timbre from a short reference audio, but they also strongly inherit the speaking style present in the reference. As a result, synthesizing speech with a desired style often requires carefully selecting reference audio, which is impractical when only limited or mismatched references are available. While recent controllable TTS methods attempt to address this issue, they typically rely on absolute style targets and discrete textual prompts, and therefore do not support continuous and reference-relative style control. We propose ReStyle-TTS, a framework that enables continuous and reference-relative style control in zero-shot TTS. Our key insight is that effective style control requires first reducing the model’s implicit dependence on reference style before introducing explicit control mechanisms. To this end, we introduce Decoupled Classifier-Free Guidance (DCFG), which independently controls text and reference guidance, reducing reliance on reference style while preserving text fidelity. On top of this, we apply style-specific LoRAs together with Orthogonal LoRA Fusion to enable continuous and disentangled multi-attribute control, and introduce a Timbre Consistency Optimization module to mitigate timbre drift caused by weakened reference guidance. Experiments show that ReStyle-TTS enables user-friendly, continuous, and relative control over pitch, energy, and multiple emotions while maintaining intelligibility and speaker timbre, and performs robustly in challenging mismatched reference–target style scenarios. The project webpage is available at <https://cucl-2.github.io/Restyle-TTS>.

## 1 Introduction

Recent zero-shot text-to-speech (TTS) systems can synthesize speech for unseen speakers from only a short reference audio clip. By conditioning on this reference, these models can preserve the speaker’s

identity (timbre) while following the input text. However, the generated speech is often strongly influenced by the speaking style present in the reference audio, including prosody and emotion, which fundamentally limits controllability in zero-shot TTS. As a result, synthesizing speech with a desired style often requires carefully selecting reference audio that matches the target style, which is time-consuming and sometimes impossible when only limited or mismatched reference audio is available. This issue is especially pronounced when the available reference conveys a different style from the target, such as attempting to generate angry speech when only a happy reference clip is available.

This limitation naturally motivates research on controllable TTS like InstructTTS (Guo et al., 2023; Yang et al., 2024; Liu et al., 2023). While these approaches have demonstrated promising results, most of them assume a fixed or predefined speaker space and therefore do not support true zero-shot speaker generalization from short reference audio. More recent work attempts to bridge the gap between voice cloning and controllability by enabling timbre cloning while allowing style manipulation. IndexTTS2 (Zhou et al., 2025) and Vevo (Zhang et al., 2025) achieves style control through a style prompt audio, but still requires carefully selecting suitable reference samples. ControlSpeech (Ji et al., 2024), EmoVoice (Yang et al., 2025), and CosyVoice (Du et al., 2024b, 2025) instead allow users to specify style through language-based prompts on top of voice cloning, which is more user-friendly. However, text-based style control remains unstable due to the complex and many-to-many relationship between textual descriptions and acoustic realizations. Moreover, these methods typically rely on absolute target styles and do not support continuous and reference-relative style control, where attributes are adjusted incrementally with respect to the reference, which is more intu-

Model	Timbre Source	Style Source	Continuous Control	Control Type
IndexTTS2 / Vevo	Reference Audio	Style Audio	No	Absolute
ControlSpeech / EmoVoice / CosyVoice	Reference Audio	Text Description	No	Absolute
StyleFusion TTS	Reference Audio	Audio or Text	No	Absolute
<b>ReStyle-TTS (Ours)</b>	Reference Audio	Style LoRA	<b>Yes</b>	<b>Relative</b>

Table 1: Comparison of controllable zero-shot TTS methods.

itive and user-friendly. A structured comparison of representative controllable zero-shot TTS methods is summarized in Table 1.

Achieving continuous and reference-relative style control while preserving zero-shot voice cloning capability is challenging due to a fundamental trade-off. If the model remains strongly dependent on the reference audio, the generated speech is tightly constrained by the reference style, leaving little room for flexible control. Conversely, simply weakening the influence of the reference audio often degrades speaker timbre consistency, undermining the core objective of zero-shot TTS. To address this challenge, we propose ReStyle-TTS. We first introduce Decoupled Classifier-Free Guidance (DCFG), which separately controls the guidance strengths from text and from the reference audio, allowing us to reduce the reliance on the reference audio during generation while maintaining text fidelity. Building on this, we apply style-specific LoRAs together with Orthogonal LoRA Fusion to inject explicit and continuously adjustable style factors (e.g., pitch, energy, and emotions) on top of the base model. Orthogonal LoRA Fusion enables the independent and simultaneous control of multiple style attributes. Finally, since reducing reference guidance can introduce timbre drift, we incorporate a Timbre Consistency Optimization module that explicitly reinforces speaker timbre preservation during training.

With these components, ReStyle-TTS enables controllable zero-shot TTS that provides user-friendly, continuous, and relative control over speaking style while preserving speaker timbre. We also evaluate our method on several challenging scenarios, including generating angry speech from happy references, a case that previous approaches have not effectively addressed.

Our contributions are summarized as follows:

- We propose **ReStyle-TTS**, a controllable zero-shot TTS framework that enables user-friendly, continuous, and reference-relative control of speaking style while preserving speaker timbre.

- We demonstrate that ReStyle-TTS can effectively control pitch, energy, and emotions, and it excels in handling scenarios where the reference and target styles are mismatched.

## 2 Related Works

**Zero-shot TTS.** Zero-shot text-to-speech (TTS) aims to synthesize speech for unseen speakers without explicit speaker-specific training and can be broadly categorized into non-autoregressive (NAR), autoregressive (AR), and hybrid architectures. In the NAR domain, Voicebox (Le et al., 2023) formulates TTS as a text-guided speech infilling problem, trained via flow matching (Lipman et al., 2022). E2-TTS (Eskimez et al., 2024) and F5-TTS (Chen et al., 2024a) simplify the alignment process by appending filler tokens to the text sequence, avoiding the need for duration models. In AR-based approaches, models such as AudioLM (Borsos et al., 2023), VALL-E (Wang et al., 2023), and SparkTTS (Wang et al., 2025) model discrete audio semantic and acoustic tokens, leveraging powerful language modeling techniques for speech generation. Hybrid architectures like CosyVoice (Du et al., 2024a,b, 2025), Seed-TTS (Anastassiou et al., 2024), and IndexTTS2 (Zhou et al., 2025) autoregressively model semantic tokens and then employ flow matching to generate mel spectrograms. To reduce the information loss caused by discrete token modeling, continuous token modeling has been explored in DiTAR (Jia et al., 2025) and MELLE (Meng et al., 2024), inspired by continuous representation learning in image generation, such as MAR (Li et al., 2024).

All these models perform well in zero-shot TTS, but they often inherit the speaking style of the reference. This makes synthesizing speech in a desired style time-consuming, as it requires carefully selecting reference audio, which may be infeasible when suitable references are unavailable. Ideally, synthesized speech should allow for flexible style control.

**Controllable Speech Synthesis.** Early controllable TTS models, such as FastSpeech2 (Ren

et al., 2020) and FastPitch (Łańcucki, 2021), primarily controlled prosody by explicitly predicting low-level attributes like pitch, energy, and duration. Later advancements introduced control through discrete textual tags (Kim et al., 2021; Gao et al., 2025) and moved toward prompt-based control, where natural language descriptions specify the desired speaking style. Notable models include InstructTTS (Yang et al., 2024), PromptStyle (Liu et al., 2023), and PromptTTS (Guo et al., 2023). However, these models are either speaker-independent or rely on predefined speaker identities or embeddings for timbre, meaning they cannot perform true zero-shot speaker cloning from a brief reference audio clip.

More recent approaches attempt to bridge the gap between voice cloning and controllability by enabling timbre cloning while allowing style manipulation. SC VALL-E (Kim et al., 2023) achieves style control by adjusting a latent style control vector; however, this vector itself is not interpretable. Vevo (Zhang et al., 2025) and IndexTTS2 (Zhou et al., 2025) achieve style control by providing a separate style prompt audio. ControlSpeech (Ji et al., 2024), EmoVoice (Yang et al., 2025), and CosyVoice (Du et al., 2024b, 2025) instead enable style control through language-based style prompts on top of voice cloning, which is more user-friendly. StyleFusion TTS (Chen et al., 2024b) further supports style control using both textual descriptions and style audio simultaneously.

However, all these methods cannot support continuous or relative control, and they disregard the inherent style of the reference audio. A more user-friendly interaction would instead allow relative adjustments, such as slightly increasing the pitch or making the speech sound a bit angrier.

**Style Control in Image Generation using LoRA.** In the field of image generation, it is common practice to train LoRA models to modify the style of generated images (Gandikota et al., 2024; Frenkel et al., 2024) and to combine multiple LoRA models for controlling a blend of styles (Shah et al., 2024; Zhong et al., 2024; Ouyang et al., 2025; Zheng et al., 2025). However, in the TTS domain, the style of speech is inherently embedded in the reference audio. The model is trained to replicate the style from the reference audio to the generated audio. As a result, the direct application of LoRA-based style control methods from image generation is not suitable for TTS.

## 3 Method

### 3.1 Overview

Style control using LoRA has been widely adopted in image generation, where LoRAs fine-tuned on specific datasets (e.g., anime or Van Gogh paintings) can control image styles during inference. However, this approach doesn’t directly apply to zero-shot TTS systems due to the key difference in the use of reference audio. In image generation, outputs are guided solely by text prompts, while zero-shot TTS systems rely on both text and reference audio during inference. The model learns to replicate not only the timbre but also the style of the reference audio, making precise style control difficult.

To address this issue, we propose ReStyle-TTS, which uses Decoupled Classifier-Free Guidance (DCFG) to reduce the model’s dependency on the reference audio while maintaining faithfulness to the text. After this decoupling step, a series of Style LoRA modules can be trained to control various attributes such as pitch or emotion, and we further introduce an Orthogonal LoRA Fusion mechanism to combine multiple Style LoRAs without mutual interference. However, weakening the reference dependency may reduce timbre consistency. Therefore, we additionally propose a Timbre Consistency Optimization (TCO) module to explicitly preserve the speaker timbre encoded in the reference audio. An overview of ReStyle-TTS is shown in Figure 1.

### 3.2 Decoupled Classifier-Free Guidance

In standard zero-shot TTS systems, classifier-free guidance (CFG) is commonly used to balance conditional and unconditional predictions during generation (Du et al., 2024a,b; Chen et al., 2024a). Let  $f_{a,t}$  denote the predicted audio representation conditioned on both the reference audio  $a$  and the text  $t$ , and let  $f_{\emptyset,\emptyset}$  denote the unconditional prediction. The conventional CFG formulation is:

$$\hat{f} = f_{a,t} + \lambda_{\text{cfg}}(f_{a,t} - f_{\emptyset,\emptyset}), \quad (1)$$

where  $\lambda_{\text{cfg}}$  controls the overall guidance strength. Increasing  $\lambda_{\text{cfg}}$  enhances the influence of the conditional inputs but does not distinguish between text guidance and reference guidance, since both are entangled within  $f_{a,t}$ . As a result, adjusting  $\lambda_{\text{cfg}}$  simultaneously affects text fidelity and style dependency—reducing the weight of the reference also weakens textual alignment.

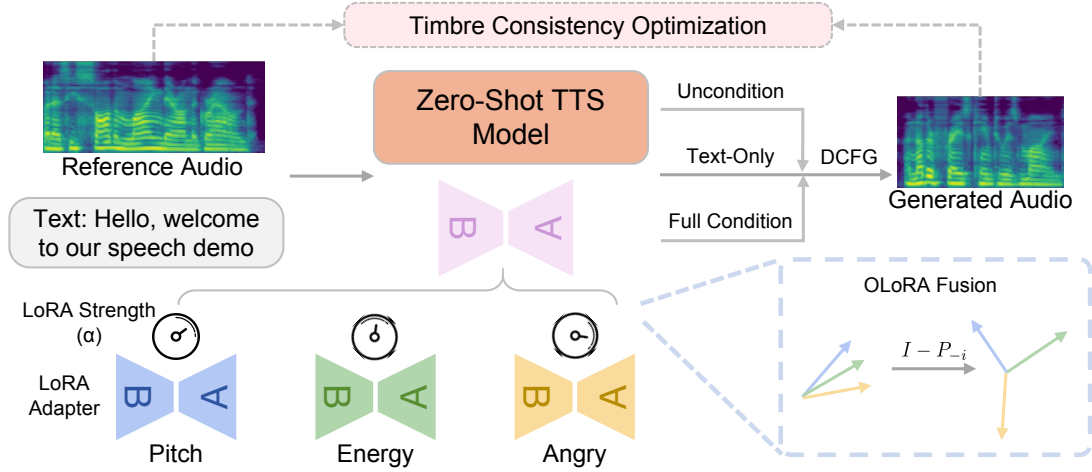


Figure 1: The overall framework of **ReStyle-TTS**. The proposed method consists of three logically coordinated components: (1) **Decoupled Classifier-Free Guidance (DCFG)** reduces the model’s dependency on the reference style while maintaining text fidelity; (2) **Orthogonal LoRA Fusion (OLoRA)** reduces interference among multiple style-specific LoRAs by projecting each LoRA onto the orthogonal complement of the subspace spanned by the others; and (3) **Timbre Consistency Optimization (TCO)** reinforces speaker identity preservation through a similarity-based reward mechanism.

To disentangle these effects, we introduce Decoupled Classifier-Free Guidance. We separately compute intermediate predictions conditioned on (i) text only,  $f_{\emptyset,t}$ , and (ii) both reference and text,  $f_{a,t}$ . DCFG combines them as follows:

$$\hat{f}_{\text{DCFG}} = f_{\emptyset,t} + \lambda_t(f_{\emptyset,t} - f_{\emptyset,\emptyset}) + \lambda_a(f_{a,t} - f_{\emptyset,t}), \quad (2)$$

where  $\lambda_t$  and  $\lambda_a$  are independent guidance strengths for text and reference. Specifically,  $\lambda_t$  controls how strongly the model follows the text, while  $\lambda_a$  determines how much it depends on the reference audio.

When  $\lambda_t = \lambda_{\text{cfg}}$  and  $\lambda_a = 1 + \lambda_{\text{cfg}}$ , DCFG reduces to the standard CFG formulation. By lowering  $\lambda_a$  while keeping  $\lambda_t$  fixed, we explicitly reduce the model’s reliance on the reference style without harming text alignment. This makes subsequent style control feasible, rather than relying entirely on the style present in the reference audio.

### 3.3 Style LoRA and Orthogonal LoRA Fusion

With DCFG reducing the model’s dependency on reference style, the generated audio is no longer bound to the prosody or emotion of the reference. This enables us to introduce controllable style modification using LoRA, inspired by its successful application in image generation (Zhong et al., 2024; Ouyang et al., 2025; Zheng et al., 2025). Similarly, we fine-tune style-specific LoRAs on audio datasets annotated with particular attributes such as

high/low pitch or different emotions. Each LoRA thus captures a single interpretable attribute direction in the model parameter space.

Following the practice in the image domain, the influence of each LoRA can be continuously adjusted by scaling its magnitude, enabling smooth control of style intensity (Gandikota et al., 2024). Moreover, since each LoRA specializes in a single attribute, it is desirable to combine multiple LoRAs to control several attributes simultaneously. However, directly adding LoRA weights often leads to interference between adapters (Shah et al., 2024; Zhong et al., 2024; Ouyang et al., 2025; Zheng et al., 2025), resulting in entangled or unstable styles.

To address this, we propose Orthogonal LoRA Fusion (OLoRA), a training-free mechanism for combining multiple style LoRAs. OLoRA jointly orthogonalizes the parameter subspaces of individual LoRAs and performs weighted fusion to compose multiple style attributes without retraining. Formally, for a linear layer with  $N$  trained LoRAs, let  $\{\Delta W_i\}_{i=1}^N$  denote their low-rank updates, where  $\Delta W_i = B_i A_i$ . We first decorrelate them by projecting each  $\Delta W_i$  onto the orthogonal complement of the subspace spanned by all others. Denote  $v_i = \text{vec}(\Delta W_i) \in \mathbb{R}^D$  and  $V_{-i} = [v_1, \dots, v_{i-1}, v_{i+1}, \dots, v_N]$ , and compute the projection matrix  $P_{-i} = V_{-i}(V_{-i})^+$  using least squares or SVD. The orthogonalized update is then  $\tilde{v}_i = (I - P_{-i})v_i$ , which is reshaped back to  $\tilde{\Delta W}_i$ .

In contrast to sequential projection schemes that are sensitive to the fusion order and may yield inconsistent compositions, OLoRA performs joint orthogonalization by projecting each adapter against the entire subspace spanned by the remaining ones, resulting in an order-independent fusion process. Crucially, since the number of LoRAs is significantly smaller than the parameter dimension ( $N \ll D$ ), the style vectors occupy a sparse subspace within the high-dimensional manifold. This sparsity ensures that orthogonal projection effectively eliminates interference.

The orthogonalized LoRAs are fused through a weighted combination, and the final generation of ReStyle-TTS is expressed using the following unified formulation:

$$\hat{f}_{\text{ReStyle}} = g_{\emptyset,t} + \lambda_t (g_{\emptyset,t} - g_{\emptyset,\emptyset}) + \lambda_a (g_{a,t} - g_{\emptyset,t}), \quad (3)$$

$$g_{a,t} = f_{a,t}^{(\Theta + \Delta W_{\text{fuse}})}, \quad (4)$$

$$\Delta W_{\text{fuse}} = \sum_{i=1}^N \alpha_i \tilde{W}_i. \quad (5)$$

Here,  $\Theta$  denotes the base model parameters,  $\lambda_t$  and  $\lambda_a$  control the text and reference guidance strengths, and each  $\alpha_i$  provides continuous control over its corresponding style attribute.

### 3.4 Timbre Consistency Optimization

While DCFG relaxes the dependency on the reference audio and OLoRA enables flexible style control, these modifications may weaken the preservation of speaker timbre. To explicitly enhance timbre consistency without altering the main training objective, we introduce Timbre Consistency Optimization (TCO), a lightweight reinforcement strategy guided by speaker similarity rewards.

In standard flow-matching training, the model parameters  $\theta$  are optimized by minimizing the mean squared error between the predicted and target flows:  $\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{(x,y)} \|f_{\theta}(x) - y\|_2^2$ . To incorporate timbre feedback, we sample speech generated by the current model and evaluate its speaker similarity against the corresponding reference audio, which serves as a reward signal  $r$ . To reduce reward variance, we maintain an exponential moving average (EMA) baseline  $b_t = \mu b_{t-1} + (1 - \mu)r_t$ , and define the advantage as  $A_t = r_t - b_t$ . To avoid the training instability and computational overhead of policy gradients, we instead adopt an advantage-weighted regression strategy (Peng et al., 2019) that

reweights the flow-matching loss using a smooth, bounded weight  $w_t = 1 + \lambda \tanh(\beta A_t)$ , where  $\lambda$  controls the reward strength and  $\beta$  modulates sensitivity to advantage. The total objective becomes  $\mathcal{L}_{\text{total}} = w_t \cdot \mathcal{L}_{\text{FM}}$ .

This formulation can be viewed as a reward-modulated weighting of the original supervised loss. Samples with higher speaker similarity receive stronger gradient emphasis, while those with lower similarity are naturally down-weighted. Since no gradient is propagated through the generation or reward computation, TCO preserves the stability and efficiency of standard flow-matching training. As a result, TCO effectively reinforces timbre consistency between generated and reference speech.

## 4 Experiments

### 4.1 Experimental Setup

**Dataset.** We trained separate LoRA modules on different subsets of the VccmDataset (Ji et al., 2024). The VccmDataset is composed of LibriTTS (Zen et al., 2019) and several emotion-focused audio datasets (Christophe et al., 2016; Zhou et al., 2021; Dupuis and Pichora-Fuller, 2010). Specifically, we used subsets corresponding to high and low pitch, high and low energy, and multiple emotion categories including *angry*, *disgusted*, *fear*, *happy*, *sad*, *surprised*, and *neutral*, while excluding *contempt* due to insufficient data. For evaluation, we conducted controllable zero-shot speech synthesis experiments on the Seed-TTS test set (Anastassiou et al., 2024) and additionally used the VccmDataset (Ji et al., 2024) test set for the contradictory-style setting, which requires emotional audio to create mismatched reference–target conditions.

**Implementation.** Instead of training the TTS model from scratch, we fine-tune the well-known F5-TTS (Chen et al., 2024a). During LoRA training, we inject LoRA adapters into all linear layers with a rank of 32 and an alpha value of 64. The AdamW optimizer is used with a learning rate of  $1 \times 10^{-5}$  and a batch size of 30,000 audio frames. Because the amount of audio data varies across subsets, we fixed the total training time to 250 hours rather than keeping the number of epochs constant. In DCFG training, the masked speech input is first dropped with a rate of 0.3, and then the input containing both masked speech and text is dropped with a rate of 0.2. In Timbre Consistency Opti-

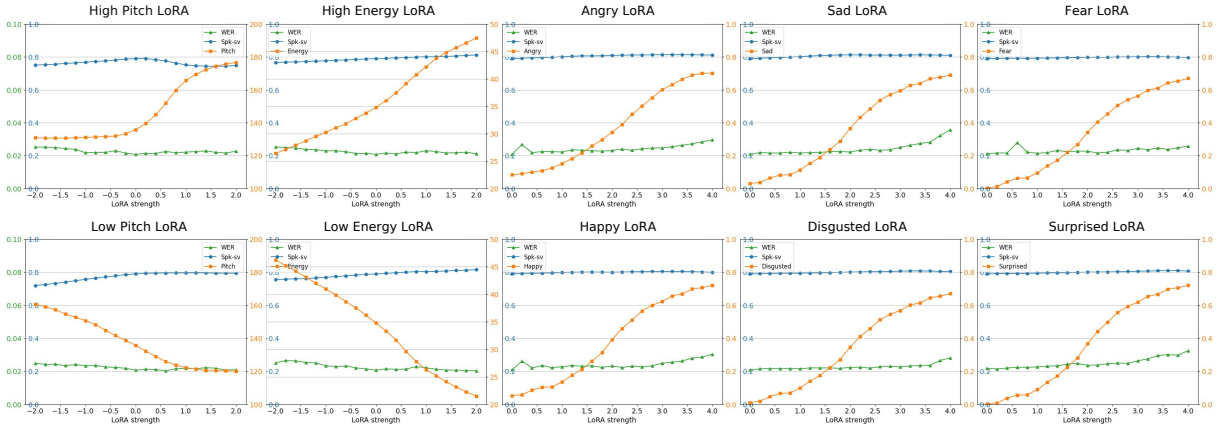


Figure 2: Continuous single-attribute control with style-specific LoRAs.

mization, the reward strength coefficient is set to  $\lambda = 0.2$ , the advantage sensitivity to  $\beta = 5.0$ , and the EMA momentum for the baseline to  $\mu = 0.9$ . For standard CFG, a common choice is  $\lambda_{\text{cfg}} = 2$ . When using our DCFG, the setting  $\lambda_t = 2$  and  $\lambda_a = 3$  is equivalent to this conventional configuration. In order to reduce the model’s dependence on the reference audio, we instead set  $\lambda_a = 0.5$ .

**Evaluation.** Following ControlSpeech (Ji et al., 2024), we report not only the Word Error Rate (WER) and timbre similarity (Spk-sv) between the reference and synthesized speech, but also measure attribute-specific control effectiveness. For subjective evaluations, we conduct MOS-SA (Mean Opinion Score–Style Accuracy) evaluations to measure the accuracy of the synthesized speech’s style. The evaluation details can be found in Appendix B.

## 4.2 Single-Attribute Control

To verify that ReStyle-TTS can continuously control individual attributes without harming intelligibility or speaker identity, we first activate a single style LoRA at a time and sweep its strength over a range of values. Figure 2 summarizes the results for pitch (high/low), energy (high/low), and six emotions (angry, sad, fear, happy, disgusted, surprised). The reported metrics are averaged over the Seed-TTS test set.

For the prosodic LoRAs (pitch and energy), the attribute curves vary smoothly as the LoRA strength changes, while WER and Spk-sv remain almost constant. Notably, negative scaling of a ‘high-attribute’ LoRA naturally produces the opposite effect, effectively enabling bidirectional control even when only one side of the attribute was trained. For emotional LoRAs, we similarly obtain monotonic control over the emotion similar-

ity score as the LoRA strength increases. Unlike text-prompt-based methods, where emotion is typically specified by discrete labels or natural language descriptions and is therefore difficult to adjust continuously, our method yields a smooth intensity knob for each emotion. These results confirm that ReStyle-TTS enables precise and continuous single-attribute manipulation for both low-level prosody and high-level emotion.

## 4.3 Multi-Attribute Composition

To further evaluate whether different Style LoRAs can be jointly applied without introducing noticeable interference, we activate two LoRAs simultaneously and sweep their strengths over a 2D grid. Figure 3 presents representative combinations. The reported metrics are averaged over the Seed-TTS test set. Across all evaluated pairs, the controlled attributes vary smoothly along their respective axes. Modulating the strength of one LoRA primarily influences its target attribute, while the other attribute remains largely stable. Meanwhile, both WER and speaker similarity remain stable over the entire 2D space, suggesting that simultaneous multi-attribute manipulation does not compromise intelligibility or timbre preservation.

To further push the analysis, we activate three Style LoRAs simultaneously and evaluate how the model behaves in the resulting three-dimensional control space. As shown in Figure 4, the surfaces for pitch, energy, and anger each show smooth and monotonic variation along their respective control axes.

## 4.4 Relative Style Control

A key advantage of ReStyle-TTS is its ability to perform relative style control: attributes are ad-

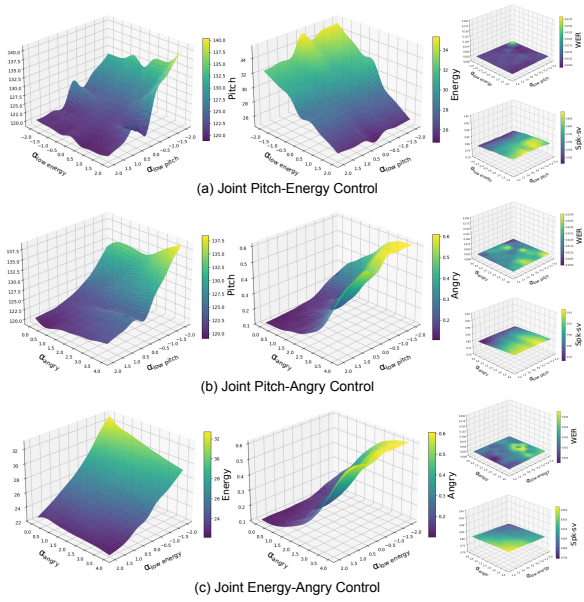


Figure 3: Two-attribute joint control results.

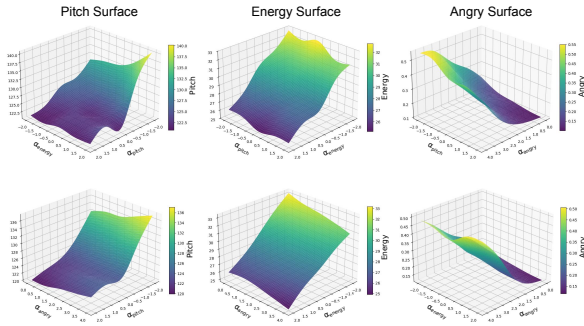


Figure 4: Three-attribute joint control results.

justed with respect to the reference audio rather than pushed toward a fixed absolute target. This interaction paradigm is more predictable and user-friendly. Previously, we reported only averages over the entire Seed-TTS test set. To assess relative control, we examine how the Style LoRA affects the attribute of each individual sample. Figure 5 plots reference energy against generated energy under different LoRA scales. Across all scales, the points form clear linear trends with regression slopes ranging from 0.77 to 1.22 and intercepts consistently near zero. This pattern indicates that the LoRA induces a roughly proportional change. As a result, the relative ordering among reference samples is preserved, in contrast to absolute control, which would drive the slope toward 0 and collapse all samples toward the same target value. Figure 6 shows energy trajectories for eight reference samples. All curves vary smoothly and monotonically while starting from distinct baselines corresponding to each sample’s inherent style, further confirm-

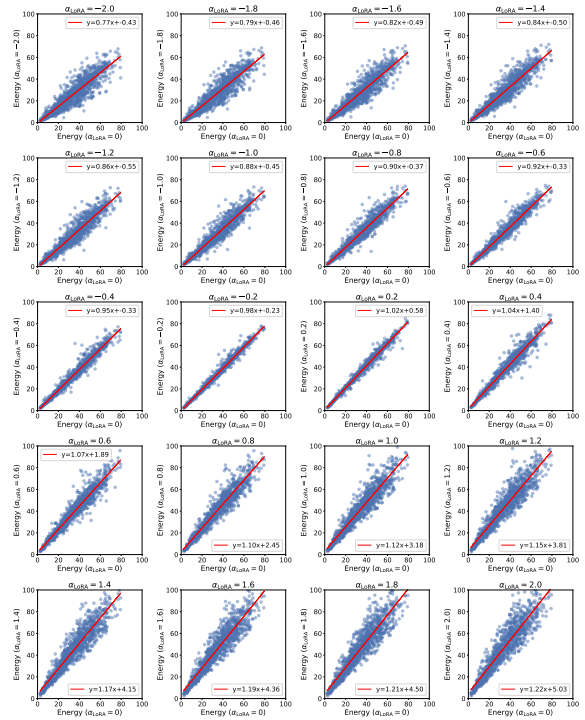


Figure 5: Reference energy vs. generated energy under different LoRA strengths.

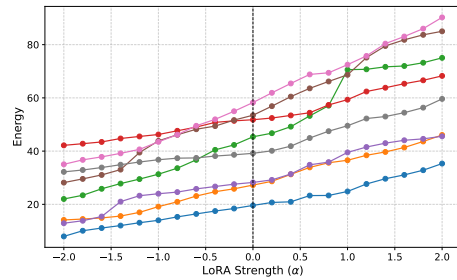


Figure 6: Energy trajectories of five randomly selected samples as the High Energy LoRA scale varies.

ing our relative control.

#### 4.5 Contradictory-Style Generation

We further evaluate ReStyle-TTS in a contradictory-style setting on the VccmDataset test set, where the reference audio and the target style intentionally do not match. Concretely, the reference provides the speaker’s timbre but carries an emotion or prosodic pattern that is different from the desired target. A more detailed explanation is provided in Appendix D.

Table 2 reports the results for emotion transfer under such mismatched conditions. Compared with text-prompt-based controllable TTS, ReStyle-TTS more reliably follows the target emotion instead of the emotion implied by the reference audio, in-

Ref \ Target	Angry	Disgusted	Fear	Happy	Sad	Surprised	Neutral
Angry	–	48.4/63.4/78.6/82.4	61.2/71.2/80.3/88.7	78.4/84.2/86.2/92.1	65.8/72.1/79.4/86.8	64.7/74.5/80.7/90.3	58.4/74.2/78.9/84.6
Disgusted	61.9/70.2/80.2/87.1	–	57.8/70.6/80.4/89.9	83.5/86.5/89.3/96.8	73.5/78.3/80.6/85.6	59.8/75.2/81.3/92.7	61.9/73.9/80.5/88.2
Fear	51.4/60.9/79.2/85.7	65.7/75.9/78.5/83.5	–	82.9/85.7/90.4/100.0	80.0/84.3/86.7/92.9	55.7/73.4/81.2/95.1	54.3/68.2/78.6/81.7
Happy	65.2/73.5/88.5/100.0	55.8/65.2/79.3/86.2	52.1/70.2/80.6/90.8	–	81.1/82.0/81.4/83.9	62.6/70.5/80.2/87.6	52.3/65.4/81.6/91.4
Sad	55.4/65.2/78.8/81.6	66.8/75.5/80.3/88.0	67.2/75.2/80.5/85.3	81.1/83.2/86.6/97.4	–	60.5/72.5/79.7/83.8	57.1/72.3/80.4/90.9
Surprised	72.0/78.5/83.6/100.0	57.0/65.4/80.2/90.4	58.9/70.0/79.5/82.7	72.0/79.2/79.8/84.0	62.0/78.9/82.3/92.0	–	62.0/73.6/80.7/88.5
Neutral	53.3/72.9/80.6/94.4	57.8/70.1/79.2/85.6	64.2/80.2/83.7/91.2	70.7/78.5/80.3/88.9	66.7/75.4/78.4/83.3	67.8/72.9/80.5/86.9	–

Table 2: Emotion transfer matrix for contradictory-style generation. Each off-diagonal cell reports the ACC (%) in the format **CosyVoice / EmoVoice / IndexTTS2 / ReStyle-TTS**.

Ref → Target	CosyVoice	EmoVoice	ReStyle-TTS
<b>Pitch</b>			
Low → High	74.9	72.4	<b>90.2</b>
High → Low	76.9	73.1	<b>92.8</b>
<b>Energy</b>			
Low → High	87.5	76.1	<b>92.4</b>
High → Low	88.6	75.9	<b>93.0</b>

Table 3: Contradictory-style generation results for pitch and energy.

dicating that weakening the reference-style dependence via DCFG and then applying Style-LoRAs is effective for overriding the original style. We also examine contradictory-style control over pitch and energy in Table 3. The results show that our method can consistently move pitch and energy in the desired opposite direction. We also provide a subjective evaluation of MOS-SA in Appendix D. These experiments confirm that ReStyle-TTS can handle challenging contradictory-style generation scenarios for both emotion and prosody.

#### 4.6 Ablation Studies

We conducted ablation studies on DCFG and TCO in Table 4 and provide additional ablation studies on Orthogonal LoRA Fusion and the hyperparameter selection of  $\lambda_a$  in DCFG in Appendix E. The reported metrics in Table 4 are averaged over the 10 attributes shown in Figure 2, with LoRA strengths set to 2.0 for prosody control and 4.0 for emotion control. Regarding the control intensity metric Attr  $\Delta$  (rel.), we calculate the relative percentage change for prosody attributes and the absolute change in logits for emotion attributes, ensuring that the magnitudes remain comparable across different attribute types.

With standard CFG, text and reference guidance are coupled, preventing independent control. A high CFG weight (e.g.,  $\lambda_{\text{cfg}} = 2$ , equivalent to  $\lambda_t = 2$  and  $\lambda_a = 3$ ) enforces strong text fidelity and speaker similarity but severely limits style controllability. Conversely, a low CFG weight

Setting	Attr $\Delta$ (rel.) $\uparrow$	WER(%) $\downarrow$	Spk-sv $\uparrow$
default ( $\lambda_t = 2, \lambda_a = 0.5$ )	51.2%	2.31	0.79
w/o DCFG ( $\lambda_{\text{cfg}} = 2$ )	2.1%	1.83	0.90
w/o DCFG ( $\lambda_{\text{cfg}} = 0.5$ )	7.6%	2.67	0.85
w/o TCO	51.0%	2.32	0.71

Table 4: Ablation study on DCFG and TCO.

(e.g.,  $\lambda_{\text{cfg}} = -0.5$ , equivalent to  $\lambda_t = -0.5$  and  $\lambda_a = 0.5$ ) leads to severe distortion and unusable WER exceeding 1.0, which is omitted in the Table. An intermediate setting ( $\lambda_{\text{cfg}} = 0.5$ , i.e.,  $\lambda_t = 0.5$  and  $\lambda_a = 1.5$ ) maintains intelligibility but still relies too heavily on the reference to enable effective style control. Overall, under CFG, improving controllability inevitably degrades text fidelity, and there exists no suitable value that can simultaneously achieve both controllability and text faithfulness. This motivates DCFG, which decouples and independently calibrates text and reference guidance. Furthermore, removing the Timbre Consistency Optimization module leads to a marked decline in speaker similarity, demonstrating its critical role in preserving timbre when reference guidance is reduced.

## 5 Conclusion

In this paper, we propose ReStyle-TTS designed to enable continuous and relative style control in zero-shot speech synthesis. To achieve this, we first introduced Decoupled Classifier-Free Guidance (DCFG) to relax reference audio dependency while maintaining text fidelity. To achieve flexible manipulation, we leveraged Style-LoRAs with Orthogonal LoRA Fusion, allowing for the precise, simultaneous adjustment of multiple attributes. Furthermore, Timbre Consistency Optimization (TCO) was incorporated to ensure robust identity preservation. Experiments demonstrate that ReStyle-TTS effectively supports user-friendly style control and excels in challenging contradictory-style generation scenarios, offering a practical solution for expressive and controllable speech synthesis.

## Limitations

Although ReStyle-TTS successfully enables user-friendly relative and continuous style control, a primary limitation lies in its scalability to new attributes. Specifically, introducing control for a new style dimension requires collecting a corresponding dataset and performing additional LoRA fine-tuning.

## Ethics and Potential Risks

The advancement of high-fidelity, zero-shot text-to-speech systems with flexible style control, such as ReStyle-TTS, brings significant ethical considerations. While our framework enhances user interaction and creative content generation, the ability to clone a speaker's identity from a short audio clip and manipulate their emotional expression poses potential risks for misuse, such as unauthorized voice cloning, deepfake generation, and the spread of misinformation. To mitigate these risks, it is crucial to ensure that such models are deployed responsibly. We strongly advocate for the integration of robust audio watermarking, the continuous development of synthetic speech detection models, and the establishment of strict protocols requiring explicit consent from voice contributors prior to synthesis.

## References

- Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, and 1 others. 2024. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*.
- Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and 1 others. 2023. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533.
- Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.
- Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2024a. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*.
- Zhiyong Chen, Xinnuo Li, Zhiqi Ai, and Shugong Xu. 2024b. Stylefusion tts: Multimodal style-control and enhanced feature fusion for zero-shot text-to-speech synthesis. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 263–277. Springer.
- Veaux Christophe, Yarnagishi Junichi, and M Kirsten. 2016. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. *The Centre for Speech Technology Research (CSTR)*, pages 807–814.
- Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, and 1 others. 2024a. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*.
- Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, and 1 others. 2025. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*.
- Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, and 1 others. 2024b. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.
- Kate Dupuis and M Kathleen Pichora-Fuller. 2010. Toronto emotional speech set (tess)-younger talker\_happy.
- Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, and 1 others. 2024. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 682–689. IEEE.
- Yarden Frenkel, Yael Vinker, Ariel Shamir, and Daniel Cohen-Or. 2024. Implicit style-content separation using b-lora. In *European Conference on Computer Vision*, pages 181–198. Springer.
- Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. 2024. Concept sliders: Lora adaptors for precise control in diffusion models. In *European Conference on Computer Vision*, pages 172–188. Springer.
- Xiaoxue Gao, Chen Zhang, Yiming Chen, Huayun Zhang, and Nancy F Chen. 2025. Emo-dpo: Controllable emotional speech synthesis through direct preference optimization. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. 2023. Prompttts: Controllable text-to-speech with text descriptions. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

- Shengpeng Ji, Jialong Zuo, Wen Wang, Minghui Fang, Siqi Zheng, Qian Chen, Ziyue Jiang, Hai Huang, Zehan Wang, Xize Cheng, and 1 others. 2024. Controlspeech: Towards simultaneous zero-shot speaker cloning and zero-shot language style control with decoupled codec. *arXiv preprint arXiv:2406.01205*.
- Dongya Jia, Zhuo Chen, Jiawei Chen, Chenpeng Du, Jian Wu, Jian Cong, Xiaobin Zhuang, Chumin Li, Zhen Wei, Yuping Wang, and 1 others. 2025. Ditar: Diffusion transformer autoregressive modeling for speech generation. *arXiv preprint arXiv:2502.03930*.
- Daegyom Kim, Seongho Hong, and Yong-Hoon Choi. 2023. Sc vall-e: Style-controllable zero-shot text to speech synthesizer. *arXiv preprint arXiv:2307.10550*.
- Minchan Kim, Sung Jun Cheon, Byoung Jin Choi, Jong Jin Kim, and Nam Soo Kim. 2021. Expressive text-to-speech using style tag. *arXiv preprint arXiv:2104.00436*.
- Adrian Łańcucki. 2021. Fastpitch: Parallel text-to-speech with pitch prediction. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6588–6592. IEEE.
- Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and 1 others. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. *Advances in neural information processing systems*, 36:14005–14034.
- Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. 2024. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37:56424–56445.
- Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.
- Guanghou Liu, Yongmao Zhang, Yi Lei, Yunlin Chen, Rui Wang, Zhifei Li, and Lei Xie. 2023. Promptstyle: Controllable style transfer for text-to-speech with natural language descriptions. *arXiv preprint arXiv:2305.19522*.
- Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2023. emotion2vec: Self-supervised pre-training for speech emotion representation. *arXiv preprint arXiv:2312.15185*.
- Lingwei Meng, Long Zhou, Shujie Liu, Sanyuan Chen, Bing Han, Shujie Hu, Yanqing Liu, Jinyu Li, Sheng Zhao, Xixin Wu, and 1 others. 2024. Autoregressive speech synthesis without vector quantization. *arXiv preprint arXiv:2407.08551*.
- Ziheng Ouyang, Zhen Li, and Qibin Hou. 2025. K-lora: Unlocking training-free fusion of any subject and style loras. *arXiv preprint arXiv:2502.18461*.
- Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. 2019. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.
- Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.
- Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. 2024. Ziplora: Any subject in any style by effectively merging loras. In *European Conference on Computer Vision*, pages 422–438. Springer.
- Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, and 1 others. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.
- Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, and 1 others. 2025. Sparktts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *arXiv preprint arXiv:2503.01710*.
- Dongchao Yang, Songxiang Liu, Rongjie Huang, Chao Weng, and Helen Meng. 2024. Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2913–2925.
- Guanrou Yang, Chen Yang, Qian Chen, Ziyang Ma, Wenxi Chen, Wen Wang, Tianrui Wang, Yifan Yang, Zhikang Niu, Wenrui Liu, and 1 others. 2025. Emovoice: Llm-based emotional text-to-speech model with freestyle text prompting. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 10748–10757.
- Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*.
- Xueyao Zhang, Xiaohui Zhang, Kainan Peng, Zhenyu Tang, Vimal Manohar, Yingru Liu, Jeff Hwang, Dangna Li, Yuhao Wang, Julian Chan, and 1 others. 2025. Vevo: Controllable zero-shot voice imitation with self-supervised disentanglement. *arXiv preprint arXiv:2502.07243*.
- Peng Zheng, Ye Wang, Rui Ma, and Zuxuan Wu. 2025. Freelora: Enabling training-free lora fusion for autoregressive multi-subject personalization. *arXiv preprint arXiv:2507.01792*.

Ming Zhong, Yelong Shen, Shuohang Wang, Yadong Lu, Yizhu Jiao, Siru Ouyang, Donghan Yu, Jiawei Han, and Weizhu Chen. 2024. Multi-lora composition for image generation. *arXiv preprint arXiv:2402.16843*.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. 2021. 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)*, pages 920–924. IEEE.

Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. 2025. In-dextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. *arXiv preprint arXiv:2506.21619*.

## A Equivalence Between DCFG and Standard CFG

We show that standard classifier-free guidance (CFG) is a special case of our Decoupled CFG (DCFG).

$$\begin{aligned}\hat{f}_{\text{CFG}} &= f_{a,t} + \lambda_{\text{cfg}}(f_{a,t} - f_{\emptyset,\emptyset}) \\ &= (1 + \lambda_{\text{cfg}}) f_{a,t} - \lambda_{\text{cfg}} f_{\emptyset,\emptyset}.\end{aligned}\quad (1)$$

$$\begin{aligned}\hat{f}_{\text{DCFG}} &= f_{\emptyset,t} + \lambda_t(f_{\emptyset,t} - f_{\emptyset,\emptyset}) + \lambda_a(f_{a,t} - f_{\emptyset,t}) \\ &= (1 + \lambda_t - \lambda_a) f_{\emptyset,t} + \lambda_a f_{a,t} - \lambda_t f_{\emptyset,\emptyset}.\end{aligned}\quad (2)$$

To recover the CFG form, the coefficients of the three terms must match between (1) and (2):

$$\begin{cases} 1 + \lambda_t - \lambda_a = 0, \\ \lambda_a = 1 + \lambda_{\text{cfg}}, \\ \lambda_t = \lambda_{\text{cfg}}. \end{cases}$$

Solving yields:

$$\lambda_a = 1 + \lambda_{\text{cfg}}, \quad \lambda_t = \lambda_{\text{cfg}}.$$

Substituting these values into (2) gives exactly the standard CFG expression.

## B Evaluation Details

For objective evaluations, following Control-Speech (Ji et al., 2024), we report not only the Word Error Rate (WER) and timbre similarity (Spk-sv) between the reference and synthesized speech, but also measure attribute-specific control effectiveness. For WER, we employ Whisper-large-v3 (Radford et al., 2023) for transcription. To evaluate timbre similarity (Spk-sv) between the original prompt

and the synthesized speech, we utilize the base-plus-sv version of WavLM (Chen et al., 2022). For volume, we compute the  $\ell_2$  norm of the amplitude of each short-time Fourier transform (STFT) frame. Pitch values are estimated using the Parselmouth toolkit, which extracts the fundamental frequency ( $f_0$ ) and computes the geometric mean across all voiced regions. To evaluate emotion, we employ the official Emotion2Vec model (Ma et al., 2023) to compute speech emotion logits and classification accuracy. For subjective evaluations, we conduct MOS-SA (Mean Opinion Score – Style Accuracy) evaluations to measure the accuracy of the synthesized speech’s style via crowdsourcing. We randomly select 30 samples from the test set for subjective evaluation, and each audio sample is listened to by at least 10 testers. Testers are asked to rate the style accuracy on a 5-point scale ranging from 1 to 5.

## C Relative Style Control

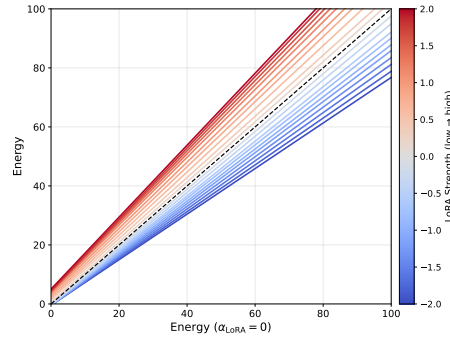


Figure 7: Linear regression analysis of energy control across different LoRA scales.

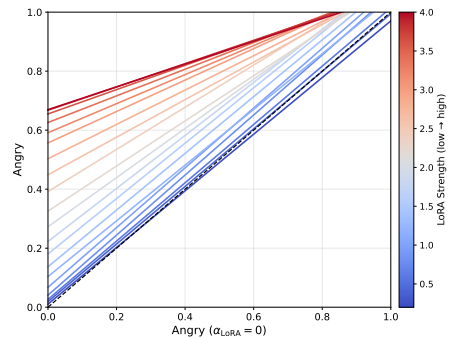


Figure 8: Linear regression analysis of angry control across different LoRA scales.

As illustrated in Figure 7, we visualize the relationship between the baseline style and the modified style. The x-axis represents the energy of the

Ref \ Target	Angry	Disgusted	Fear	Happy	Sad	Surprised	Neutral
Angry	–	3.42/3.61/3.95/ <b>4.18</b>	3.58/3.79/4.05/ <b>4.31</b>	3.91/4.12/4.32/ <b>4.52</b>	3.49/3.72/4.00/ <b>4.23</b>	3.63/3.84/4.15/ <b>4.41</b>	3.38/3.68/3.95/ <b>4.09</b>
Disgusted	3.61/3.82/4.05/ <b>4.29</b>	–	3.47/3.76/4.10/ <b>4.36</b>	4.02/4.21/4.40/ <b>4.61</b>	3.71/3.93/4.10/ <b>4.24</b>	3.46/3.81/4.15/ <b>4.47</b>	3.59/3.74/4.05/ <b>4.28</b>
Fear	3.33/3.54/3.95/ <b>4.17</b>	3.68/3.91/4.00/ <b>4.08</b>	–	3.98/4.23/4.45/ <b>4.68</b>	3.87/4.06/4.25/ <b>4.39</b>	3.41/3.69/4.10/ <b>4.46</b>	3.29/3.57/3.95/ <b>4.02</b>
Happy	3.74/3.92/4.35/ <b>4.69</b>	3.36/3.58/3.95/ <b>4.21</b>	3.31/3.69/4.05/ <b>4.43</b>	–	4.01/4.12/4.20/ <b>4.29</b>	3.62/3.83/4.10/ <b>4.34</b>	3.35/3.62/4.10/ <b>4.42</b>
Sad	3.45/3.64/3.95/ <b>4.13</b>	3.72/3.91/4.10/ <b>4.31</b>	3.63/3.82/4.05/ <b>4.25</b>	4.03/4.14/4.35/ <b>4.58</b>	–	3.52/3.79/4.00/ <b>4.16</b>	3.44/3.71/4.10/ <b>4.37</b>
Surprised	3.88/4.09/4.40/ <b>4.71</b>	3.41/3.63/4.00/ <b>4.27</b>	3.49/3.71/3.95/ <b>4.12</b>	3.91/4.03/4.20/ <b>4.35</b>	3.58/3.97/4.25/ <b>4.49</b>	–	3.61/3.83/4.10/ <b>4.32</b>
Neutral	3.39/3.78/4.20/ <b>4.46</b>	3.51/3.69/4.00/ <b>4.24</b>	3.67/3.99/4.20/ <b>4.41</b>	3.79/4.01/4.20/ <b>4.43</b>	3.55/3.77/4.00/ <b>4.14</b>	3.66/3.81/4.10/ <b>4.34</b>	–

Table 5: Emotion transfer matrix for contradictory-style generation. Each off-diagonal cell reports the MOS-SA (5-point scale) in the format **CosyVoice / EmoVoice / IndexTTS2 / ReStyle-TTS**.

generated speech when the LoRA scale is set to 0 (serving as the reference baseline), while the y-axis displays the energy values obtained under varying LoRA strengths.

It can be observed that as the LoRA strength increases, the slope of the fitted regression line steepens, rising from 0.77 to 1.22. This monotonic increase in slope demonstrates that ReStyle-TTS achieves true relative control by scaling the inherent attributes of the reference audio rather than overwriting them with fixed absolute values. In addition to relative prosody control, we also present the results for angry in Figure 8 as an example of relative emotion control.

## D Contradictory-Style Generation

In this section, we provide a detailed explanation of the experimental setup for Contradictory-Style Generation. Our approach diverges from the original usage of the VccmDataset in ControlSpeech. Each sample in the VccmDataset consists of an audio clip, its corresponding transcription, and a style prompt. However, the original ControlSpeech evaluation did not address scenarios where the style of the reference audio conflicts with the target style. For instance, attempting to synthesize angry speech when only a happy reference clip is available. To evaluate this capability, we utilize the audio samples from the VccmDataset as reference audio. For each reference, we attempt to synthesize speech targeting every emotion category that differs from the reference’s inherent emotion. The synthesis accuracy is then quantitatively evaluated using the Emotion2Vec model. Regarding the specific control configurations: for ReStyle-TTS, we simply apply the Style-LoRA corresponding to the target emotion. For the natural language-controlled baselines, CosyVoice and EmoVoice, we provide the specific style instruction: ‘I’m saying this with great {emotion}.’ We also provide a subjective evaluation of MOS-SA in Table 5. The results are consistent with the objective evaluation, and our

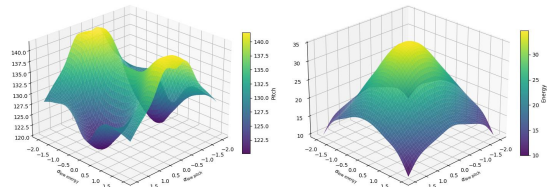


Figure 9: Ablation Study of Orthogonal LoRA Fusion.

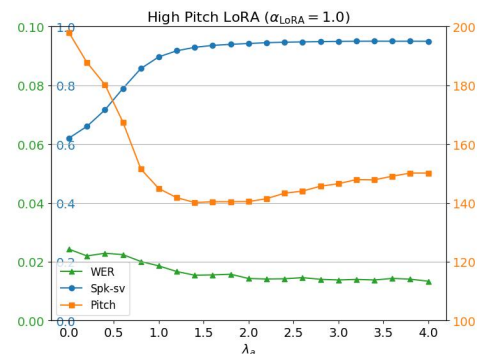


Figure 10: Ablation Study of  $\lambda_a$ .

ReStyle-TTS achieves the best performance across all contradictory-style generation scenarios.

## E Ablation Studies

For Orthogonal LoRA Fusion, we also conducted ablation studies. As shown in Figure 9, the control of pitch and energy becomes fully entangled, making it impossible to adjust them independently.

We further performed ablation experiments on the parameter  $\lambda_a$ . Since DCFG decouples the model’s reliance on the text and the reference audio, we fixed  $\lambda_t = 2$ , a commonly used setting to maintain strong text dependency, and varied  $\lambda_a$ , which governs the trade-off between timbre similarity and controllability. The results are shown in Figure 10. It can be observed that  $\lambda_a$  governs the trade-off between timbre similarity and controllability. We ultimately chose  $\lambda_a = 0.5$ , sacrificing some timbre similarity to achieve better controllability, while using Timbre Consistency Optimization to compensate for the lost timbre similarity.

Table 6: Single-attribute control results using CosyVoice backbone.

<b>Strength</b>	<b>-2.0</b>	<b>-1.5</b>	<b>-1.0</b>	<b>-0.5</b>	<b>0.0</b>	<b>+0.5</b>	<b>+1.0</b>	<b>+1.5</b>	<b>+2.0</b>
Pitch	165.1	160.8	155.4	148.9	142.5	136.2	130.8	125.6	122.7
Energy	49.8	47.2	43.5	39.1	35.8	31.4	28.2	25.1	23.6
<b>Strength</b>	<b>0.0</b>	<b>0.5</b>	<b>1.0</b>	<b>1.5</b>	<b>2.0</b>	<b>2.5</b>	<b>3.0</b>	<b>3.5</b>	<b>4.0</b>
Angry	0.10	0.16	0.28	0.42	0.56	0.68	0.74	0.77	0.79
Happy	0.06	0.12	0.24	0.38	0.52	0.64	0.71	0.75	0.76

## F Generalization to Different Model Backbones

We further verify the generalizability of our proposed framework by applying it to a different model backbone, CosyVoice (Du et al., 2024b). For rapid validation, we evaluated Low Pitch, Low Energy, Angry, and Happy attributes. Notably, we rely solely on LoRA strength to modulate the style of CosyVoice, bypassing its native text-based instructions. The results demonstrate that ReStyle-TTS generalizes effectively to different backbones for both single-attribute and multi-attribute control.

As shown in Table 6, sweeping the LoRA strength on the CosyVoice backbone yields smooth and monotonic transitions for both prosodic and emotional attributes, consistent with the observations on F5-TTS. Furthermore, we applied Orthogonal LoRA Fusion (OLoRA) to CosyVoice for joint Energy-Pitch control. As illustrated in Table 7, adjusting the strength of one attribute has minimal impact on the other (e.g., varying Pitch strength across a 30Hz range results in less than 1.0 unit of Energy fluctuation). These results confirm that our decoupling and fusion mechanism is backbone-agnostic and maintains high disentanglement across different model architectures.

Table 7: Joint Energy-Pitch control on CosyVoice using OLoRA.

<b>Energy \ Pitch</b>	<b>-2.0</b>	<b>-1.0</b>	<b>0.0</b>	<b>+1.0</b>	<b>+2.0</b>
<i>Measured Pitch</i>					
<b>-2.0</b>	147.0	142.5	135.5	128.5	121.5
<b>-1.0</b>	147.8	142.7	135.7	128.7	121.7
<b>0.0</b>	148.5	143.0	136.0	129.0	122.0
<b>+1.0</b>	149.3	143.3	136.3	129.3	122.3
<b>+2.0</b>	150.0	143.5	136.5	129.5	122.5
<i>Measured Energy</i>					
<b>-2.0</b>	42.5	42.2	42.0	41.8	41.5
<b>-1.0</b>	38.5	38.2	38.0	37.8	37.5
<b>0.0</b>	34.5	34.2	34.0	33.8	33.5
<b>+1.0</b>	30.5	30.2	30.0	29.8	29.5
<b>+2.0</b>	26.5	26.2	26.0	25.8	25.5