UNISPEAKER: A UNIFIED SPEECH GENERATION MODEL FOR MULTIMODALITY-DRIVEN VOICE CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advancements in zero-shot speech personalized generation have brought synthetic speech increasingly close to the realism of target speakers' recordings, yet multimodal voice creation remains on the rise. In various scenarios, individuals often seek to control and create voice characteristics through different voice description modalities. To address the limitations in both the versatility and performance of voice control found in previous methods, this paper introduces UniSpeaker, a unifiled multimodality-driven speech generation model that integrates face images, text descriptions, voice attribute descriptions, and reference speech for comprehensive voice control and creation. Specifically, we propose a unified voice aggregator based on KV-Former, applying soft contrastive loss to map diverse voice description modalities into a shared voice space, ensuring that the generated voice aligns more closely with the input descriptions. In addition, multimodal voice control is incorporated within a large-scale speech generation framework, employing self-distillation to enhance voice disentanglement. To evaluate multimodality-driven voice control, we build the first multimodality-based voice control (MVC) benchmark, focusing on voice suitability, voice diversity, and speech quality. UniSpeaker is evaluated across five tasks using the MVC benchmark, and the experimental results demonstrate that UniSpeaker outperforms previous modality-specific models. Speech samples are available at https://UniSpeaker.github.io.

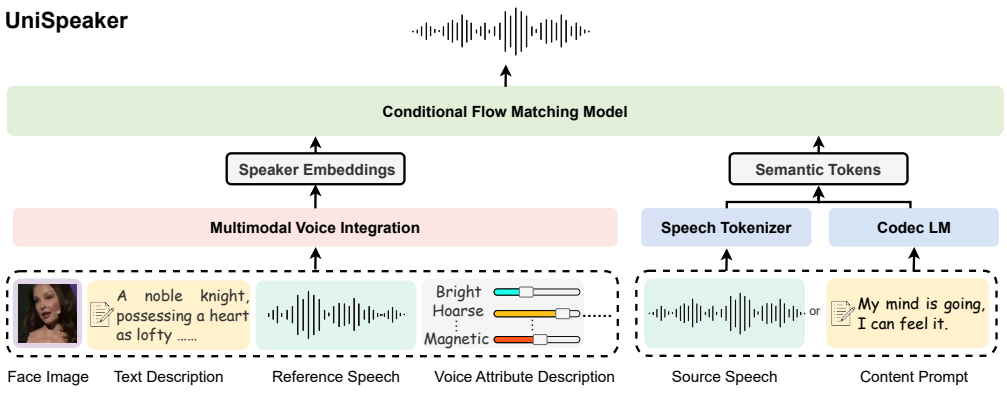


Figure 1: The overview of UniSpeaker, where speaker embeddings and semantic tokens serve as intermediate representations. Speaker embeddings are used to control the voice characteristics of generated speech and can be derived from various voice description modalities. Semantic tokens can be generated from source speech or text, corresponding to the speech-to-speech conversion and text-to-speech synthesis, respectively. Codec LM represents the text-to-token language model.

Table 1: Comparison between UniSpeaker and previous studies on multimodality-driven voice control tasks. TTS stands for text-to-speech synthesis. SSC stands for speech-to-speech conversion, preserving both the content and prosody of source speech.

Model	Voice Description Modality				Speech Generation Approaches	
	Speech	Text	Face	Attribute	TTS	SSC
PromptSpeaker(Zhang et al., 2023)	✓	✓			✓	
Prompttts++ (Shimizu et al., 2024)	✓	✓			✓	
PromptVC (Yao et al., 2024a)	✓	✓				✓
HybridVC (Niu et al., 2024)	✓	✓				✓
Imaginary Voice (Lee et al., 2023)			✓		✓	
Synthe-sees (Park et al., 2024)	✓		✓		✓	
3D-face (Yang et al., 2023)	✓		✓		✓	
FaceVC (Lu et al., 2021)	✓		✓			✓
SP-FaceVC (Weng et al., 2023)			✓			✓
FVMVC (Sheng et al., 2023)	✓		✓			✓
HearFace (Lee et al., 2024a)			✓			✓
VoxEditor (Sheng et al., 2024)	✓			✓		✓
UniSpeaker (Ours)	✓	✓	✓	✓	✓	✓

1 INTRODUCTION

In recent years, the field of speech synthesis has seen remarkable progress, driven by innovations in generative models and the expansion of training data. Some models (Wang et al., 2023a; Du et al., 2024; Ju et al., 2024; Vyas et al., 2023) can clone a voice using only a few seconds of reference speech, achieving a level of naturalness and speaker similarity that closely resembles actual recordings. Despite significant advances in voice cloning, zero-shot speech synthesis still faces limitations in certain scenarios, such as providing voiceovers for artificially created virtual characters, where obtaining ideal reference speech is very difficult or even non-existent (Guo et al., 2023). Compared to reference speech, natural text descriptions offer a more user-friendly approach to express intentions for voice generation (Leng et al., 2024), and facial images, which are easier to obtain, also have a strong correlation with voice characteristics (Goto et al., 2020; Oh et al., 2019). In the absence of reference speech, utilizing other modalities allows for more flexible and convenient control over voice characteristics. Hence, multimodal voice description-based speech generation, which involves generating corresponding voice characteristics from natural text descriptions, face images, or other modalities, presents a promising approach.

Recently, several studies have explored text prompt-based voice control for speech generation. These research (Shimizu et al., 2024; Zhang et al., 2023) have developed internal voice description sets and use BERT (Devlin et al., 2019) networks to extract text embeddings for voice control. Some research (Lu et al., 2021; Sheng et al., 2023) aligns facial recognition representations with speaker embeddings and uses these facial representations for voice control. In addition to the aforementioned absolute voice descriptions, VoxEditor (Sheng et al., 2024) introduces the relative descriptions for voice attributes editing, allowing for more nuanced control over voice characteristics.

Despite notable advancements made in these studies, they still have limitations in two key aspects: (1) **The versatility of voice control**: Current methods often explore different voice description modalities and generation approaches independently. On the one hand, it faces challenges in handling diverse inputs (as shown in Table 1). On the other hand, it fails to combine multimodal voice descriptions for more fine-grained and unique voice generation. (2) **The performance of voice control**: Previous methods were typically trained from scratch on limited paired multimodal data (Zhang et al., 2023; Shimizu et al., 2024), resulting in sparse coverage of the voice space. Effective strategies for fusing multimodality to enrich this space remain unclear. Additionally, these methods often generate speech with voice characteristics that do not align well with the input voice descriptions. While large-scale speech generation models (Wang et al., 2023a; Shen et al., 2024; Du et al., 2024) demonstrate exceptional voice control capabilities, the scalable integration of multimodality to enrich the capabilities of these pre-trained models remains unexplored.

To address these limitations, we present **UniSpeaker**, a speech generation model that aligns multiple modalities into a consistent voice space through a unified framework. As illustrated in Figure 1, speaker embeddings and semantic tokens serve as key representations to generate speech. Speaker embeddings control the voice characteristics and can be extracted from various inputs. Semantic tokens convey the content and prosody of the generated speech, derived from either source speech or content prompt. To effectively integrate multimodal input for voice control, pre-trained modality-specific encoders extract corresponding representations from different modalities, and then a unified multimodal voice aggregator (MVA) aligns these multimodal representations into a consistent voice space. The MVA is built upon the designed KV-Former, a streamlined variant of the Transformer (Vaswani et al., 2017). The KV-Former leverages a set of learnable key-value vectors to build a shared multimodal voice space, where multimodal representations serve as queries. To improve the alignment between voice characteristics and other modalities, soft contrastive learning (SoftCL) is applied to relax the strict one-to-one contrastive constraint and leverage intra-modal discriminative information for guidance. Considering the advantages of supervised semantic-tokens, we use the open-source CosyVoice (Du et al., 2024) as the backbone for UniSpeaker. Prior to integrating multimodal voice descriptions, we employ the simple yet effective self-distillation (Anastassiou et al., 2024) to improve the voice disentanglement of the pre-trained CosyVoice, thereby preserving its general capabilities across different tasks.

Due to the lack of publicly available benchmarks for evaluating multimodality-driven voice control, we established a multimodality-based voice control (MVC) benchmark, encompassing five tasks: face-driven voice conversion (FaceVC), face-driven personalized text-to-speech (FaceTTS), text description-driven voice conversion (TextVC), text description-driven personalized text-to-speech (TextTTS), attribute-driven voice editing (AVE). Following previous works (Sheng et al., 2023; Yao et al., 2024a), MVC benchmark evaluates generated speech using multimodal voice descriptions across three aspects: voice suitability, voice diversity and speech quality. We evaluate UniSpeaker using the MVC benchmark, where it demonstrates superior performance on the aforementioned five tasks compared to previous modality-specific models. Speech samples are available at `https://UniSpeaker.github.io`.

2 Relate Work

2.1 Large speech generation models

As speech generation systems (Tan et al., 2022; Kim et al., 2021) have achieved remarkable levels of naturalness and robustness, recent research (Ju et al., 2024; Lee et al., 2024b) has shifted focus towards exploring novel generative models, advanced modeling objectives, and larger-scale datasets to pursue voice diversity. When integrating multimodal voice descriptions, it is crucial to preserve the performance of pre-trained speech generation models in terms of naturalness, robustness, and prosody. Some representative large-scale speech generation (Wang et al., 2023a; Kim et al., 2024; Chen et al., 2024) models typically leverage a neural codec to convert speech waveforms into discrete acoustic token sequences, along with an autoregressive language model to generate discrete tokens from text. However, the discrete acoustic token sequences entangle content, speaker, and prosodic information in this approach, complicating the alignment of multimodal voice characteristics without disrupting the content and prosody of the generated speech. Recently, CosyVoice (Du et al., 2024) has utilized supervised semantic tokens (Radford et al., 2023) as the modeling objectives for a large language model (LLM). Subsequently, a conditional flow matching model (CFM) generates speech based on semantic tokens, speaker embeddings and mel spectrograms prompt. Since the semantic tokens primarily encompass content and prosodic information, the speaker information included is limited. This facilitates further voice disentanglement and the integration of multimodal voice descriptions, making CosyVoice well-suited as the backbone for the UniSpeaker model proposed in this paper.

2.2 Multimodality-driven voice control for speech generation

Modeling diverse voice characteristics has consistently been a critical focus in the field of speech synthesis. Recent works, such as PromptTTS2 (Leng et al., 2024), Audiobox (Vyas et al., 2023), and others (Guan et al., 2024; Yang et al., 2024; Ji et al., 2024), have explored using text prompts to control the style or emotion of generated speech. However, only a few studies have specifically

targeted voice control with text prompt (Shimizu et al., 2024; Zhang et al., 2023). Text prompt-based style control TTS methods typically convert speech attributes like pitch, energy, duration, and emotion into natural style prompts using LLMs. Since these style prompts primarily reflect prosody and capture minimal speaker individuality, achieving the desired voice control remains challenging.

In the field of multimodal voice control, researchers have previously attempted to align different voice description modalities with speaker embeddings using models such as memory networks (Sheng et al., 2023), mixture density networks (Shimizu et al., 2024), and latent diffusion (Yao et al., 2024a), as well as loss functions like MSE loss (Lu et al., 2021), cosine similarity loss (Zhang et al., 2023), and perceptual loss (Weng et al., 2023). However, these alignment methods relied on parallel datasets and were challenging to extend directly to additional modalities. Performance-wise, previous face-based methods (Lee et al., 2023) generally ensured gender accuracy but often produced incongruous voice characteristics, such as generating a youthful voice for an elderly face. Additionally, VoxEditor (Sheng et al., 2024) is limited to performing voice attribute editing on existing source speech, thus offering restricted voice diversity. In response, the proposed UniSpeaker employs a unified voice aggregator to construct a shared voice space that can be easily extended to new modalities, achieving versatile and diverse voice control.

3 METHODS

In this section, we first review the backbone CosyVoice, then introduce how multimodal voice descriptions are integrated into a pre-trained speech generation model, and finally outline our training strategy with SoftCL and self-distillation.

3.1 PRELIMINARIES

Our backbone leverages supervised semantic tokens (Radford et al., 2023; Ye et al., 2024) as modeling objectives, utilizing an LLM for text-to-token generation and a CFM for token-to-speech synthesis. Given a dataset $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}$, where $\mathbf{x}$ is a speech sample and $\mathbf{y}$ is the corresponding text transcription, the sequence input to the LLM is mainly comprised of $\{\mathbf{s}, \mathbf{Y}, \mathbf{C}\}$, where $\mathbf{s}$ represents the speaker embeddings of $\mathbf{x}$, $\mathbf{Y}$ is the text embedding of $\mathbf{y}$ and $\mathbf{C}$ is the semantic tokens of $\mathbf{x}$. The LLM is then trained in an autoregressive manner to minimize the negative log-likelihood of semantic tokens $\mathbf{C}$. The core of CFM is to construct a probability density path from a prior distribution to $p_0(\mathbf{X})$ to the data distribution of the Mel-spectrograms $q(\mathbf{X})$. The probability density path is defined by a time-dependent vector field $\mathbf{v}_t(\mathbf{X})$, which generates the flow ϕ_t through an ordinary differential equation (ODE). The flow matching model is trained using optimal-transport conditional flow matching (OT-CFM) (Tong et al., 2023), which can be written as follows,

$$\mathcal{L}_{\text{OT-CFM}} = \mathbb{E}_{t,p_0(\mathbf{X}_0),q(\mathbf{X}_1)} \left| \omega_t(\phi_t^{OT}(\mathbf{X}_0, \mathbf{X}_1)|\mathbf{X}_1) - \nu_t(\phi_t^{OT}(\mathbf{X}_0, \mathbf{X}_1)|\theta_{CFM}) \right|, \tag{1}$$

where $\phi_t^{OT}(\mathbf{X}_0, \mathbf{X}_1) = (1-t)\mathbf{X}_0 + t\mathbf{X}_1$ and $\omega_t(\phi_t^{OT}(\mathbf{X}_0, \mathbf{X}_1)|\mathbf{X}_1) = \mathbf{X}_1 - \mathbf{X}_0$. The speaker embeddings $\mathbf{s}$, speech tokens $\mathbf{C}$, and masked Mel-spectrogram prompt $\tilde{\mathbf{X}}_1$ are also fed into the neural network to match the vector field with learnable parameters θ_{CFM},

$$\nu_t(\phi_t^{OT}(\mathbf{X}_0, \mathbf{X}_1)|\theta_{CFM}) = \text{NN}\left(\phi_t^{OT}(\mathbf{X}_0, \mathbf{X}_1), t; \mathbf{s}, \mathbf{C}, \tilde{\mathbf{X}}_1\right). \tag{2}$$

Therefore, both the LLM and CFM receive speaker embeddings. Our preliminary experiments revealed the CFM plays a primary role in voice control, while the LLM still exerts some influence (details in Appendix A). To improve voice disentanglement, self-distillation is applied to the pre-trained CFM, permitting multimodal voice descriptions to be integrated exclusively in the CFM.

3.2 MULTIMODAL VOICE DESCRIPTION INTEGRATION

We incorporate multiple modalities into the self-distillation CFM model, allowing various inputs to control the voice characteristics of generated speech. As shown in Figure 2, each modality is first processed by a pre-trained, modality-specific encoder to obtain the corresponding feature. Each kind of feature is then transformed into a latent vector via adaptive average pooling or a multi-layer perceptron (MLP) (Yao et al., 2024b). Those vectors across modalities are mapped into a unified voice space through a shared MVA, producing the corresponding speaker embeddings. These speaker embeddings are then fed into the CFM for speech generation.

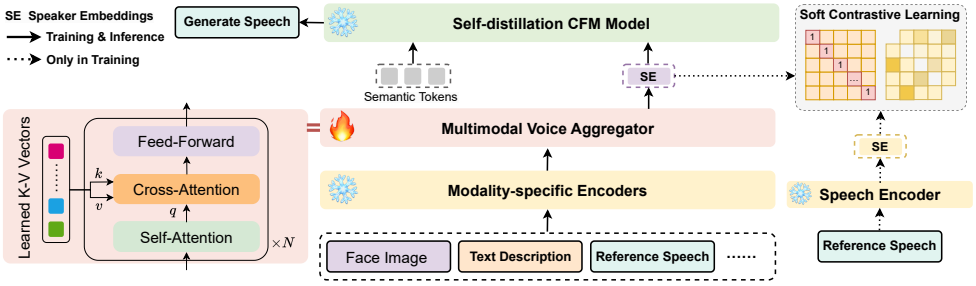


Figure 2: Overview of multimodal voice description integration.

Modality-Specific Encoders UniSpeaker employs three modality-specific encoders to process face images, speech and text. For face images, we use the MTCNN (Zhang et al., 2016) model for face detection, followed by a pre-trained convolutional neural network-based face recognition model (Schroff et al., 2015; Liu et al., 2022) to obtain global representations $\mathbf{s}_f$. For text description, T5 model (Raffel et al., 2020) is utilized to extract variable-length representations $\mathbf{s}_t$. For reference speech, the pre-trained speaker verification network (Wang et al., 2023b) from open-source CosyVoice is leveraged to extract speaker embeddings $\mathbf{s}_r$. More details about voice attribute descriptions are provided in Appendix B.

Multimodal Voice Aggregator Then global representations of different modalities should be aligned with speaker embeddings within the voice space. Previous methods relied on limited datasets that matched only two modalities for alignment, resulting in a sparse distribution in the voice space and weak generalization capabilities. Therefore, effectively utilizing multimodal voice descriptions through joint modeling to share a unified voice space, and thereby enhance the performance of each modality, remains an open question.

Inspired by Q-Former (Li et al., 2023b) and the memory mechanism (Sheng et al., 2023; Lee et al., 2021), we propose the KV-Former architecture as a unified multimodal voice aggregator. This architecture integrates learnable key-value vectors into a simplified Transformer, as shown in Figure 2. The multimodal representations act as queries and perform multi-head cross-attention with the learnable key-value vectors to retrieve the most informative representation in the voice subspace. The formulation of this process is as follows,

$$\mathbf{q} = \mathbf{W}^q \mathbf{s}_m, \mathbf{k} = \mathbf{W}^k \mathbf{f}, \mathbf{v} = \mathbf{W}^v \mathbf{f}, \mathbf{a}_m = \text{Softmax}\left(\frac{\mathbf{q}\mathbf{k}^T}{\sqrt{d}}\right)\mathbf{v}, \tag{3}$$

where $\mathbf{W}$ are the projection matrices in attention, $\mathbf{s}_m \in \{\mathbf{s}_f, \mathbf{s}_r, \mathbf{s}_t\}$ represents various state vectors, $\mathbf{f}$ are learnable key-value vectors, d is the dimension of $\mathbf{f}$, and $\mathbf{a}_m$ is the output of cross attention. In this process, the learnable key-value vectors create an information bottleneck, interacting with the three modalities to build a shared voice space. Additionally, MVA adopts a speech-anchoring mechanism, reference speech is used as input for MVA with a 50% probability. In this way, even without parallel data between all modalities, different modalities achieves potential alignment in the voice space through shared k-v vectors and joint training. The inclusion of additional modality data can facilitate performance of current modality for voice control (evaluated in Section 5.4). Our module also allows for easy expansion to new modalities by adding the a modality-specific encoder.

To integrate multimodal inputs for voice control without losing the general abilities of CFM, we feed the output of MVA to the CFM and adapt the model without changing the CFM weights. The MVA is trained to optimize $\mathcal{L}_{\text{OT-CFM}}$ and Equation (2) is transformed as follows to fit speaker embeddings,

$$\nu_t(\phi_t^{OT}(\mathbf{X}_0, \mathbf{X}_1)|\theta_{MVA}) = \text{NN}\left(\phi_t^{OT}(\mathbf{X}_0, \mathbf{X}_1), t; \mathbf{v}_m, \mathbf{C}\right), \tag{4}$$

where $\mathbf{v}_m \in \{\mathbf{v}_f, \mathbf{v}_r, \mathbf{v}_t\}$ and $\mathbf{v}_f, \mathbf{v}_r, \mathbf{v}_t$ are the outputs of applying MVA to $\mathbf{s}_f, \mathbf{s}_r, \mathbf{s}_t$, respectively. In this manner, CFM can integrate multiple modalities for voice control and keep its ability to generate natural and robust speech.

3.3 TRAINING STRATEGIES

Soft Contrastive Learning Relying solely on OT-CFM to optimize MVA leads to slow convergence, and the generated speech may exhibit voice discordance with the input voice descriptions. Inspired by previous studies (Gao et al., 2024; Wang et al., 2024), we additionally introduce the SoftCL strategy for speech-anchoring multimodal alignment, including both inter-modal and intra-modal alignment, as shown in Figure 2. For inter-modal alignment, we employ InfoNCE (Radford et al., 2021), which pulls the paired multimodal and speaker embeddings closer together while pushing the unpaired ones apart. In addition, to bring cross-modal similarities closer to the distribution within each modality, intra-modal similarities serve as soft labels. Specifically, given a batch of N multimodal-voice speaker embeddings pairs $\{(\mathbf{v}_m^i, \mathbf{s}_r^i)\}_{i=1}^N$, the intra-model self-similarity vector $p_i(\mathbf{s}_r, \mathbf{s}_r) = \{p_{ij}(\mathbf{s}_r, \mathbf{s}_r)\}_{j=1}^N$ can be obtained by:

$$p_{ij}(\mathbf{s}_r, \mathbf{s}_r) = \frac{\exp\left(\text{sim}\left(\mathbf{s}_r^i, \mathbf{s}_r^j\right)/\tau\right)}{\sum_{j=1}^N \exp\left(\text{sim}\left(\mathbf{s}_r^i, \mathbf{s}_r^j\right)/\tau\right)}, \tag{5}$$

where τ is a learnable temperature coefficient, initialized to 0.07, and $\text{sim}()$ denotes the dot product used to calculate similarity. Despite intra-model self-similarity, the confidence of positive samples still outweighs that of negatives, potentially overshadowing negatives in cross-modal relation alignment. To address this, we disentangle the negatives in the distribution to boost the relation alignment. For the self-similarity vector $p_i(\mathbf{s}_r, \mathbf{s}_r) \in \mathbb{R}^{1\times N}$, the neg-disentangled $p_i^*(\mathbf{s}_r, \mathbf{s}_r) \in \mathbb{R}^{1\times N-1}$ distribution is calculated as follows,

$$p_{ij}^* = \frac{\exp\left(p_{ij}\right)}{\sum_{k=1,k\neq i}^N \exp\left(p_{ik}\right)}. \tag{6}$$

We also apply the above negative disentanglement to $p_i(\mathbf{s}_r, \mathbf{v}_m)$, yielding $p_i^*(\mathbf{s}_r, \mathbf{v}_m)$. Then, the intra-modality alignment supervision can be achieved with negative disentanglement as follows,

$$\mathcal{L}_{\text{INTRA}} = \frac{1}{N}\sum_{i=1}^N \text{KL}\left(p_i^*(\mathbf{s}_r, \mathbf{s}_r) \| p_i^*(\mathbf{s}_r, \mathbf{v}_m)\right), \tag{7}$$

where KL represents the Kullback-Leibler Divergence. Generally, UniSpeaker is trained to optimize the following loss function,

$$\mathcal{L} = \mathcal{L}_{\text{OT-CFM}} + \lambda_1 \mathcal{L}_{\text{INTRA}} + \lambda_2 \mathcal{L}_{\text{INTER}}, \tag{8}$$

where $\mathcal{L}_{\text{INTRA}}$ is the InfoNCE loss, λ_1 and λ_2 are hyper-parameters used to balance each loss term.

Self-distillation Before integrating multimodal voice description, self-distillation is applied to fine-tune the CFM to improve voice disentanglement. Specifically, we utilize semantic tokens from the source speech, along with Mel-spectrogram prompt and speaker embeddings from another randomly selected speaker, and feed them into the CFM to perform voice conversion. This converted speech maintains the content and prosody of the source speech while almost entirely removing its speaker information (objective evaluations are shown in Table 5). Then, given the semantic tokens $\bar{\text{C}}$ of source speech and speaker embeddings $\mathbf{s}$ of converted speech as prompt, the CFM is fine-tuned to predict the source speech. Specifically, we removed the masked Mel-spectrogram prompt to improve the voice control by the speaker embeddings, transforming Equation (2) as follows,

$$\nu_t(\phi_t^{OT}(\mathbf{X}_0, \mathbf{X}_1)|\theta_{FM}) = \text{NN}\left(\phi_t^{OT}(\mathbf{X}_0, \mathbf{X}_1), t; \mathbf{s}, \bar{\text{C}}\right). \tag{9}$$

In this way, the voice characteristics of the generated speech is controlled by the speaker embeddings input to the CFM. This allows the integration of multimodal voice description directly into the CFM, simplifying the process without requiring modifications to the LLM.

4 DATASET AND BENCHMARK

Four modality-specific datasets were used to train the UniSpeaker. For the facial modality, we used the LRS3-TED dataset (Afouras et al., 2018), which includes TED Talks from 5,594 speakers, totaling approximately 400 hours of video. For the text description, LibriTTS-P (Kawamura et al.,

2024) was utilized with annotations for both voice characteristics and style, totaling approximately 585 hours of audio data from 2,443 speakers. Additionally, we collected speech-speaker identity description pairs from the internet, totaling about 90 hours. For the voice attribute modality, the VCTK-R (Sheng et al., 2024) dataset was selected, including pairwise comparisons of voice attributes among same-gender speakers, with 40 hours of audio data from 110 speakers.

The MVC Benchmark was proposed to evaluate multimodal voice control in five tasks, including FaceTTS, FaceVC, TextTTS, TextVC, and AVE. For face-related evaluation, we randomly selected 600 face images from the test set of LRS3-TED. In terms of textual descriptions, 600 sentences were randomly picked from the validation set and rewritten by a LLM (GPT-3.5-TURBO), ensuring that the meaning of the sentences remained unchanged. For voice attribute editing, following VoxEditor, 200 sentences were randomly selected from VCTK and edited on all attributes for evaluation. All above samples are unseen during training. More details can be found in the Appendix D.2.

The MVC benchmark evaluates the generated speech from three perspectives: voice suitability, voice diversity, and speech quality. Further details of the following metrics are in Appendix D.3.

1) **Voice suitability** evaluates whether the voice characteristics of the generated speech align with the input multimodal voice description. This includes three specific metrics: Speaker Similarity with Target (SST), Speaker Similarity Consistency (SSC), and MOS-Match. Speaker similarity can be computed by cosine similarity between the speaker embeddings, which are extracted from speech using a speaker verification model[1]. SST is measured by calculating the speaker similarity between the generated speech and reference speech of the target speaker. SSC assesses the consistency of the generated voice with various descriptions for the same speaker by calculating speaker similarity between the speech generated from different face images of the same speaker. MOS-Match is obtained through subjective listening tests for the mean opinion score to quantify how closely the voice characteristics of the generated speech align with the input description.

2) **Voice diversity** evaluates the model's ability to produce a diverse set of voice characteristics based on the descriptions of different speakers, rather than generating very similar ones. A metric named Speaker Similarity Diversity (SSD) is employed for evaluating voice diversity, which measures the speaker similarity between the speech generated from the descriptions of different speakers.

3) **Speech quality** assesses the robustness and naturalness of the generated speech, using two key metrics: word error rate (WER) and MOS-Nat. We employ an automatic speech recognition model[2] to transcribe the generated speech and compute the WER. MOS-Nat is determined through subjective listening tests for mean opinion scores to evaluate the naturalness of the generated speech.

5 Experiments

5.1 Experiment Settings

We trained the UniSpeaker using 4 NVIDIA TESLA V100 32G GPUs for 30K steps. The models were optimized using the AdamW optimizer with a learning rate of 1e-5 and a 10K warmup steps. The weights λ_1 and λ_2 in Equation (8) were set to 0.05. The model parameters for MVA are detailed in Table 7 of the Appendix. The speech tokenizer and codec LM were the same as those used in CosyVoice. For TTS, the codec LM accepted only text inputs without speaker embeddings.

We compared our UniSpeaker with 11 task-specific expert models in five tasks. We used the official code or pre-trained checkpoints of Imaginary Voice (Lee et al., 2023), FaceVC (Lu et al., 2021), SP-FaceVC (Weng et al., 2023), FVMVC (Sheng et al., 2023), and CosyVoice-Instruct (Du et al., 2024). For the other methods, we reproduced them according to their respective papers and evaluated them on the same dataset. Please refer to Appendix E for more details.

5.2 Evaluation Results

In this section, we conduct experiments comparing the UniSpeaker with the baselines and all objective and subjective evaluation results are reported in Table 2.

[1] https://github.com/modelscope/3D-Speaker

[2] https://huggingface.co/openai/whisper-large-v3

Table 2: Objective and subjective evaluation results of comparison systems. The definitions of all metrics can be found in 4. "-" denotes the results are not available.

Task	Methods	Voice Suitability			Voice Diversity	Speech Quality	
		SST ↑	SSC ↑	MOS-Match ↑	SSD ↓	WER ↓	MOS-Nat ↑
FaceTTS	Imaginary Voice(Lee et al., 2023)	10.08	38.46	2.39 ± 0.09	32.17	8.23	2.45 ± 0.08
	Face-StyleSpeech(Kang et al., 2023)	11.02	37.09	2.78 ± 0.12	30.78	7.09	3.29 ± 0.10
	SYNTHE-SEES(Park et al., 2024)	10.97	38.81	2.92 ± 0.11	31.09	9.14	3.39 ± 0.09
	UniSpeaker(Ours)	**12.48**	**40.75**	$\mathbf{3.18 \pm 0.10}$	**14.09**	**4.01**	$\mathbf{3.82 \pm 0.08}$
FaceVC	FaceVC(Lu et al., 2021)	8.97	50.91	2.21 ± 0.11	30.19	10.90	2.79 ± 0.10
	SP-FaceVC(Weng et al., 2023)	9.52	52.29	2.39 ± 0.09	29.86	14.92	3.04 ± 0.10
	FVMVC(Sheng et al., 2023)	9.49	51.33	2.69 ± 0.07	22.60	11.94	3.31 ± 0.08
	UniSpeaker(Ours)	**11.68**	**55.13**	$\mathbf{3.09 \pm 0.10}$	**15.91**	**4.98**	$\mathbf{3.80 \pm 0.09}$
TextTTS	PromptSpeaker(Zhang et al., 2023)	17.39	-	3.64 ± 0.13	29.84	14.70	3.37 ± 0.10
	Prompttts++(Shimizu et al., 2024)	16.87	-	3.63 ± 0.12	35.42	15.08	3.41 ± 0.11
	CosyVoice-Instruct (Du et al., 2024)	14.51	-	3.71 ± 0.13	34.62	7.03	$\mathbf{3.91 \pm 0.09}$
	UniSpeaker (Ours)	**23.09**	-	$\mathbf{3.85 \pm 0.11}$	**21.10**	**6.46**	3.87 ± 0.13
TextVC	PromptVC(Yao et al., 2024a)	16.59	-	3.47 ± 0.07	36.98	7.08	3.64 ± 0.10
	UniSpeaker(Ours)	**24.45**	-	$\mathbf{3.81 \pm 0.09}$	**24.04**	**6.29**	$\mathbf{3.77 \pm 0.11}$
AVE	VoxEditor(Sheng et al., 2024)	41.48	-	$\mathbf{3.78 \pm 0.09}$	49.92	8.01	3.57 ± 0.10
	UniSpeaker(Ours)	**49.04**	-	$\mathbf{3.79 \pm 0.10}$	**34.92**	**4.09**	$\mathbf{3.92 \pm 0.09}$

In terms of voice suitability, our findings revealed that: 1) Across five tasks, UniSpeaker outperformed previous approaches on all three metrics, except for MOS-Match in the AVE task. While VoxEditor incorporates a complex residual memory network, the performance of our unified and scalable MVA remains competitive in MOS-Match. 2) In terms of face-based voice control, previous methods were generally effective in accurately controlling the gender of the voice characteristics but often exhibited obvious voice inconsistencies in subjective aspects such as age. In contrast, UniSpeaker achieved substantial improvements in both voice-age matching and overall subjective perception. 3) Additionally, we conducted an ABX test, as shown in Figure 5 of the Appendix, the voice characteristics generated by UniSpeaker sometimes can match the face image even more closely than those of the actual speaker. We encourage readers to listen to the samples on the demo page. 4) In text control, CosyVoice-instruct concatenates voice characteristic descriptions with the content prompt in the LLM without utilizing a pre-trained text prompt, resulting in difficulties grasping semantic information effectively and producing ambiguous voice characteristics. In contrast, UniSpeaker achieves excellent semantic-to-voice consistency, where similar semantics generate similar voice characteristics.

In terms of voice diversity, it is clear that UniSpeaker significantly outperforms previous methods across 5 tasks. Furthermore, we visualized the speaker embeddings of the generated speech from both SYNTHE-SEES and UniSpeaker systems using t-SNE (Chan et al., 2019), as shown in Figure 4 (a). The figure reveals that the voice space generated by our method is significantly richer, whereas the voice space of the baseline is relatively sparse. This indicates the voice characteristics generated by the baseline for different faces may being very similar, greatly limiting voice diversity.

In terms of speech quality, by freezing the CFM during training, UniSpeaker preserve the general abilities of our backbone. Consequently, UniSpeaker surpasses previous methods in overall speech quality, only the MOS-Nat slightly lags behind CosyVoice-Instruct. This lag is due to the CFM occasionally learning noise patterns from the dataset. Conversely, CosyVoice-Instruct only integrate multimodal voice descriptions in the LLM, resulting in minimal impact on speech quality.

5.3 Ablation Study

Three ablation studies were conducted in our experiments. 1) To verify the effectiveness of MVA, the output of modality-specific encoders was mapped to the global representation, and it was directly fed into the CFM. 2) To assess the effectiveness of SoftCL, we removed the intra-class and inter-class contrastive losses from the output of MVA. 3) To validate the effectiveness of self-distillation, the performance of UniSpeaker and the open-source CosyVoice model (without self-distillation) was compared on TTS and VC tasks. We report the evaluation results for certain tasks in Table 3, with more evaluation results available in the Appendix F.

Table 3: The ablation study of UniSpeaker, measured by SST, SSD and SSC.

Task	Methods	SST ↑	SSD ↓	SSC ↑
FaceTTS	UniSpeaker	12.48	14.09	40.75
	w.o. MVA	11.40	15.07	40.61
	w.o. SoftCL	11.57	15.94	38.28
FaceVC	UniSpeaker	11.68	15.91	55.13
	w.o. MVA	10.70	19.07	54.61
	w.o. SoftCL	11.08	19.24	51.55
TTS	UniSpeaker	44.30	10.03	33.32
	w.o. self-distillation	38.49	9.80	29.68
VC	UniSpeaker	39.37	10.34	50.64
	w.o. self-distillation	31.07	10.16	43.62

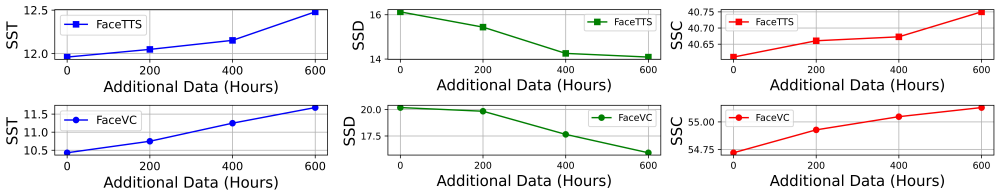


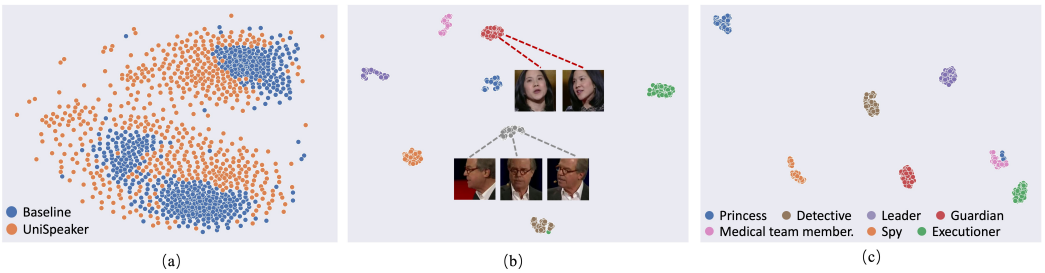
Figure 3: The evaluation results about different multimodal data scales on joint voice modeling. Here, the horizontal axis represents the amount of additional multimodal data, with "0" indicating that only the LRS3 dataset was used.

We have the following observations: 1) MVA proved beneficial for voice control with a shared multimodal voice space. It utilizes multimodal data for joint modeling through shared k-v vectors, resulting in a uniform distribution of the voice space. This promotes alignment between different modalities and enhances the model's performance in both voice diversity and voice suitability. 2) Removing SoftCL resulted in a decline across various metrics, specifically creating a significant mismatch between the generated voice and the input voice descriptions. 3) Eliminating self-distillation also had notable effects. Experimental results indicated that self-distillation significantly enhanced voice control, particularly in terms of SST. However, due to the limited data used for self-distillation, there was a slight reduction in voice diversity.

5.4 Discussions

We investigated the impact of different multimodal data scales on the shared voice space. For face-driven voice control, we trained UniSpeaker using various datasets: solely LRS3, and additional datasets of varying sizes. The results, presented in Figure 3, show that increasing the amount of multimodal data improves the performance of FaceVC and FaceTTS, highlighting the benefits of multimodal joint modeling. Furthermore, the influence of additional multimodal data on SSC is less pronounced for SST and SSD, as SSC primarily relies on intra-modal relationships.

We randomly selected 8 unseen speakers and sampled 100 different face images from each for FaceTTS. The t-SNE visualization of speaker embeddings extracted from generated speech is presented in Figure 4 (b). We observed that for each speaker, the voice remained consistent across various facial images with different angles and backgrounds. This indicates that UniSpeaker demonstrates strong robustness to noisy information in facial images. Similarly, we used the LLM to rewrite the identity descriptions 60 times, ensuring consistent semantics with varied phrasing. Figure 4 (c) shows the visualization of the speech generated by TextTTS using these identity descriptions. The results indicate that for identity descriptions with the same semantics, the generated voices are consistent. Additionally, due to the presence of a multimodal shared space, UniSpeaker can accept multiple modalities simultaneously, allowing for more flexible and nuanced voice control. For details, please refer to the Appendix G and demo page.



(a) (b) (c)

Figure 4: The visual analysis of UniSpeaker is presented here. Figure (a) uses t-SNE to visualize the voice space distributions of Baseline and UniSpeaker. In Figure (b), points of the same color represent the speech generated from different facial images of the same speaker. In Figure (c), points of the same color represent the speech generated from different identity descriptions of the same speaker, with the annotations serving as abbreviations of these text descriptions.

6 CONCLUSION

In this paper, we propose the UniSpeaker, a speech generation model that leverages multimodal voice description for voice control. Through a unified voice aggregator and designed training strategies, UniSpeaker outperforms previous modality-specific models across five tasks, generating voices that better match the input voice descriptions. In the future, we will explore how to more effectively utilize multiple voice descriptions of different modalities for one speaker simultaneously and apply our method on other more modalities for voice control.

REFERENCES

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. LRS3-TED: a large-scale dataset for visual speech recognition. *CoRR*, abs/1809.00496, 2018. URL http://arxiv.org/abs/1809.00496.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, Mingqing Gong, Peisong Huang, Qingqing Huang, Zhiying Huang, Yuanyuan Huo, Dongya Jia, Chumin Li, Feiya Li, Hui Li, Jiaxin Li, Xiaoyang Li, Xingxing Li, Lin Liu, Shouda Liu, Sichao Liu, Xudong Liu, Yuchen Liu, Zhengxi Liu, Lu Lu, Junjie Pan, Xin Wang, Yuping Wang, Yuxuan Wang, Zhen Wei, Jian Wu, Chao Yao, Yifeng Yang, Yuanhao Yi, Junteng Zhang, Qidi Zhang, Shuo Zhang, Wenjie Zhang, Yang Zhang, Zilin Zhao, Dejian Zhong, and Xiaobin Zhuang. Seed-tts: A family of high-quality versatile speech generation models. *CoRR*, abs/2406.02430, 2024. doi: 10.48550/ARXIV.2406.02430. URL https://doi.org/10.48550/arXiv.2406.02430.

David M. Chan, Roshan Rao, Forrest Huang, and John F. Canny. GPU accelerated t-distributed stochastic neighbor embedding. *J. Parallel Distributed Comput.*, 131:1–13, 2019. doi: 10.1016/J.JPDC.2019.04.008. URL https://doi.org/10.1016/j.jpdc.2019.04.008.

Sanyuan Chen, Shujie Liu, Long Zhou, Yanqing Liu, Xu Tan, Jinyu Li, Sheng Zhao, Yao Qian, and Furu Wei. VALL-E 2: Neural codec language models are human parity zero-shot text to speech synthesizers. *CoRR*, abs/2406.05370, 2024. doi: 10.48550/ARXIV.2406.05370. URL https://doi.org/10.48550/arXiv.2406.05370.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1423. URL https://doi.org/10.18653/v1/n19-1423.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, Zhifu Gao, and Zhijie Yan. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *CoRR*, abs/2407.05407, 2024. doi: 10.48550/ARXIV.2407.05407. URL https://doi.org/10.48550/arXiv.2407.05407.

Yuting Gao, Jinfeng Liu, Zihan Xu, Tong Wu, Enwei Zhang, Ke Li, Jie Yang, Wei Liu, and Xing Sun. Softclip: Softer cross-modal alignment makes CLIP stronger. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pp. 1860–1868. AAAI Press, 2024. doi: 10.1609/AAAI.V38I3.27955. URL https://doi.org/10.1609/aaai.v38i3.27955.

Shunsuke Goto, Kotaro Onishi, Yuki Saito, Kentaro Tachibana, and Koichiro Mori. Face2speech: Towards multi-speaker text-to-speech synthesis using an embedding vector predicted from a face image. In Helen Meng, Bo Xu, and Thomas Fang Zheng (eds.), *21st Annual Conference of the International Speech Communication Association, Interspeech 2020, Virtual Event, Shanghai, China, October 25-29, 2020*, pp. 1321–1325. ISCA, 2020. doi: 10.21437/INTERSPEECH.2020-2136. URL https://doi.org/10.21437/Interspeech.2020-2136.

Wenhao Guan, Yishuang Li, Tao Li, Hukai Huang, Feng Wang, Jiayan Lin, Lingyan Huang, Lin Li, and Qingyang Hong. MM-TTS: multi-modal prompt based style transfer for expressive text-to-speech synthesis. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pp. 18117–18125. AAAI Press, 2024. doi: 10.1609/AAAI.V38I16.29769. URL https://doi.org/10.1609/aaai.v38i16.29769.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. Prompttts: Controllable text-to-speech with text descriptions. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pp. 1–5. IEEE, 2023. doi: 10.1109/ICASSP49357.2023.10096285. URL https://doi.org/10.1109/ICASSP49357.2023.10096285.

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 *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pp. 10301–10305. IEEE, 2024. doi: 10.1109/ICASSP48485.2024.10445879. URL https://doi.org/10.1109/ICASSP48485.2024.10445879.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Eric Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiangyang Li, Wei Ye, Shikun Zhang, Jiang Bian, Lei He, Jinyu Li, and Sheng Zhao. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=dVhrnjZJad.

Minki Kang, Wooseok Han, and Eunho Yang. Face-stylespeech: Improved face-to-voice latent mapping for natural zero-shot speech synthesis from a face image. *CoRR*, abs/2311.05844, 2023. doi: 10.48550/ARXIV.2311.05844. URL https://doi.org/10.48550/arXiv.2311.05844.

Masaya Kawamura, Ryuichi Yamamoto, Yuma Shirahata, Takuya Hasumi, and Kentaro Tachibana. Libritts-p: A corpus with speaking style and speaker identity prompts for text-to-speech and style captioning. *CoRR*, abs/2406.07969, 2024. doi: 10.48550/ARXIV.2406.07969. URL https://doi.org/10.48550/arXiv.2406.07969.

Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5530–5540. PMLR, 2021. URL http://proceedings.mlr.press/v139/kim21f.html.

Jaehyeon Kim, Keon Lee, Seungjun Chung, and Jaewoong Cho. Clam-tts: Improving neural codec language model for zero-shot text-to-speech. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=ofzeypWosV.

Jaejun Lee, Yoori Oh, Injune Hwang, and Kyogu Lee. Hear your face: Face-based voice conversion with f0 estimation. *arXiv preprint arXiv:2408.09802*, 2024a. URL https://arxiv.org/pdf/2408.09802.

Jiyoung Lee, Joon Son Chung, and Soo-Whan Chung. Imaginary voice: Face-styled diffusion model for text-to-speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pp. 1–5. IEEE, 2023. doi: 10.1109/ICASSP49357.2023.10094745. URL https://doi.org/10.1109/ICASSP49357.2023.10094745.

Keon Lee, Dong Won Kim, Jaehyeon Kim, and Jaewoong Cho. Ditto-tts: Efficient and scalable zero-shot text-to-speech with diffusion transformer. *CoRR*, abs/2406.11427, 2024b. doi: 10.48550/ARXIV.2406.11427. URL https://doi.org/10.48550/arXiv.2406.11427.

Sangmin Lee, Hak Gu Kim, Dae Hwi Choi, Hyung-Il Kim, and Yong Man Ro. Video prediction recalling long-term motion context via memory alignment learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 3054–3063. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00307. URL https://openaccess.thecvf.com/content/CVPR2021/html/Lee_Video_Prediction_Recalling_Long-Term_Motion_Context_via_Memory_Alignment_Learning_CVPR_2021_paper.html.

Yichong Leng, Zhifang Guo, Kai Shen, Zeqian Ju, Xu Tan, Eric Liu, Yufei Liu, Dongchao Yang, Leying Zhang, Kaitao Song, Lei He, Xiangyang Li, Sheng Zhao, Tao Qin, and Jiang Bian. Prompttts 2: Describing and generating voices with text prompt. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=NsCXDyv2Bn.

Jingyi Li, Weiping Tu, and Li Xiao. Freevc: Towards high-quality text-free one-shot voice conversion. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pp. 1–5. IEEE, 2023a. doi: 10.1109/ICASSP49357.2023.10095191. URL https://doi.org/10.1109/ICASSP49357.2023.10095191.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023b. URL https://proceedings.mlr.press/v202/li23q.html.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 11966–11976. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01167. URL https://doi.org/10.1109/CVPR52688.2022.01167.

Hsiao-Han Lu, Shao-En Weng, Ya-Fan Yen, Hong-Han Shuai, and Wen-Huang Cheng. Face-based voice conversion: Learning the voice behind a face. In Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo César, Florian Metze, and Balakrishnan Prabhakaran (eds.), *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pp. 496–505. ACM,

2021. doi: 10.1145/3474085.3475198. URL https://doi.org/10.1145/3474085.3475198.

Xinlei Niu, Jing Zhang, and Charles Patrick Martin. Hybridvc: Efficient voice style conversion with text and audio prompts. *CoRR*, abs/2404.15637, 2024. doi: 10.48550/ARXIV.2404.15637. URL https://doi.org/10.48550/arXiv.2404.15637.

Tae-Hyun Oh, Tali Dekel, Changil Kim, Inbar Mosseri, William T. Freeman, Michael Rubinstein, and Wojciech Matusik. Speech2face: Learning the face behind a voice. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 7539–7548. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00772. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Oh_Speech2Face_Learning_the_Face_Behind_a_Voice_CVPR_2019_paper.html.

Jae Hyun Park, Joon-Gyu Maeng, Taejun Bak, and Young-Sun Joo. SYNTHE-SEES: face based text-to-speech for virtual speaker. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pp. 10321–10325. IEEE, 2024. doi: 10.1109/ICASSP48485.2024.10448433. URL https://doi.org/10.1109/ICASSP48485.2024.10448433.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 815–823. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7298682. URL https://doi.org/10.1109/CVPR.2015.7298682.

Kai Shen, Zeqian Ju, Xu Tan, Eric Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=Rc7dAwVL3v.

Zhengyan Sheng, Yang Ai, Yan-Nian Chen, and Zhen-Hua Ling. Face-driven zero-shot voice conversion with memory-based face-voice alignment. In Abdulmotaleb El-Saddik, Tao Mei, Rita Cucchiara, Marco Bertini, Diana Patricia Tobon Vallejo, Pradeep K. Atrey, and M. Shamim Hossain (eds.), *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pp. 8443–8452. ACM, 2023. doi: 10.1145/3581783.3613825. URL https://doi.org/10.1145/3581783.3613825.

Zhengyan Sheng, Yang Ai, Li-Juan Liu, Jia Pan, and Zhen-Hua Ling. Voice attribute editing with text prompt. *CoRR*, abs/2404.08857, 2024. doi: 10.48550/ARXIV.2404.08857. URL https://doi.org/10.48550/arXiv.2404.08857.

Reo Shimizu, Ryuichi Yamamoto, Masaya Kawamura, Yuma Shirahata, Hironori Doi, Tatsuya Komatsu, and Kentaro Tachibana. Prompttts++: Controlling speaker identity in prompt-based text-to-speech using natural language descriptions. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pp. 12672–12676. IEEE, 2024. doi: 10.1109/ICASSP48485.2024.10448173. URL https://doi.org/10.1109/ICASSP48485.2024.10448173.

Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, Frank K. Soong, Tao Qin, Sheng Zhao, and Tie-Yan Liu. Naturalspeech: End-to-end text to speech synthesis with human-level quality. *CoRR*, abs/2205.04421, 2022. doi: 10.48550/ARXIV.2205.04421. URL https://doi.org/10.48550/arXiv.2205.04421.

Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. Conditional flow matching: Simulation-free dynamic optimal transport. *CoRR*, abs/2302.00482, 2023. doi: 10.48550/ARXIV.2302.00482. URL https://doi.org/10.48550/arXiv.2302.00482.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, Jeff Wang, Ivan Cruz, Bapi Akula, Akinniyi Akinyemi, Brian Ellis, Rashel Moritz, Yael Yungster, Alice Rakotoarison, Liang Tan, Chris Summers, Carleigh Wood, Joshua Lane, Mary Williamson, and Wei-Ning Hsu. Audiobox: Unified audio generation with natural language prompts. *CoRR*, abs/2312.15821, 2023. doi: 10.48550/ARXIV.2312.15821. URL https://doi.org/10.48550/arXiv.2312.15821.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural codec language models are zero-shot text to speech synthesizers. *CoRR*, abs/2301.02111, 2023a. doi: 10.48550/ARXIV.2301.02111. URL https://doi.org/10.48550/arXiv.2301.02111.

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 Naomi Harte, Julie Carson-Berndsen, and Gareth Jones (eds.), *24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023*, pp. 5301–5305. ISCA, 2023b. doi: 10.21437/INTERSPEECH.2023-1513. URL https://doi.org/10.21437/Interspeech.2023-1513.

Qian Wang, Jia-Chen Gu, and Zhen-Hua Ling. Multiscale matching driven by cross-modal similarity consistency for audio-text retrieval. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pp. 11581–11585. IEEE, 2024. doi: 10.1109/ICASSP48485.2024.10446302. URL https://doi.org/10.1109/ICASSP48485.2024.10446302.

Shao-En Weng, Hong-Han Shuai, and Wen-Huang Cheng. Zero-shot face-based voice conversion: Bottleneck-free speech disentanglement in the real-world scenario. In Brian Williams, Yiling Chen, and Jennifer Neville (eds.), *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 13718–13726. AAAI Press, 2023. doi: 10.1609/AAAI.V37I11.26607. URL https://doi.org/10.1609/aaai.v37i11.26607.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Chao Weng, and Helen Meng. Instructtts: Modelling expressive TTS in discrete latent space with natural language style prompt. *IEEE ACM*

Trans. Audio Speech Lang. Process., 32:2913–2925, 2024. doi: 10.1109/TASLP.2024.3402088. URL https://doi.org/10.1109/TASLP.2024.3402088.

Zhihan Yang, Zhiyong Wu, Ying Shan, and Jia Jia. What does your face sound like? 3d face shape towards voice. In Brian Williams, Yiling Chen, and Jennifer Neville (eds.), *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 13905–13913. AAAI Press, 2023. doi: 10.1609/AAAI.V37I11.26628. URL https://doi.org/10.1609/aaai.v37i11.26628.

Jixun Yao, Yuguang Yang, Yi Lei, Ziqian Ning, Yanni Hu, Yu Pan, Jingjing Yin, Hongbin Zhou, Heng Lu, and Lei Xie. Promptvc: Flexible stylistic voice conversion in latent space driven by natural language prompts. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pp. 10571–10575. IEEE, 2024a. doi: 10.1109/ICASSP48485.2024.10445804. URL https://doi.org/10.1109/ICASSP48485.2024.10445804.

Linli Yao, Lei Li, Shuhuai Ren, Lean Wang, Yuanxin Liu, Xu Sun, and Lu Hou. Deco: Decoupling token compression from semantic abstraction in multimodal large language models. *CoRR*, abs/2405.20985, 2024b. doi: 10.48550/ARXIV.2405.20985. URL https://doi.org/10.48550/arXiv.2405.20985.

Lingxuan Ye, Changfeng Gao, Gaofeng Cheng, Liuping Luo, and Qingwei Zhao. ASQ: an ultra-low bit rate asr-oriented speech quantization method. *IEEE Signal Process. Lett.*, 31:221–225, 2024. doi: 10.1109/LSP.2023.3347148. URL https://doi.org/10.1109/LSP.2023.3347148.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.*, 23(10):1499–1503, 2016. doi: 10.1109/LSP.2016.2603342. URL https://doi.org/10.1109/LSP.2016.2603342.

Yongmao Zhang, Guanghou Liu, Yi Lei, Yunlin Chen, Hao Yin, Lei Xie, and Zhifei Li. Promptspeaker: Speaker generation based on text descriptions. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2023, Taipei, Taiwan, December 16-20, 2023*, pp. 1–7. IEEE, 2023. doi: 10.1109/ASRU57964.2023.10389772. URL https://doi.org/10.1109/ASRU57964.2023.10389772.

Table 4: Ablation experiments to explore the impact of the LLM and CFM on voice characteristics under different conditions. The ✓ indicates the input is a regular speaker embeddings, while the ✗ denotes random noise input.

Condition	LLM	CFM	SSIM
With Mel-spectrogram Prompt	✓	✓	62.76
	✗	✓	57.03
	✓	✗	29.39
	✗	✗	24.04
Without Mel-spectrogram Prompt	✓	✓	44.07
	✗	✓	34.51
	✓	✗	8.44
	✗	✗	4.42

Table 5: Performance of different models on the voice conversion task, where * indicates the absence of Mel-spectrogram prompt. Note that these results are not comparable to those in Table 3 due to different test samples

model	SSIM
Groud Truth	69.67
CosyVoice (Du et al., 2024)	72.63
CosyVoice*	43.59
FreeVC (Li et al., 2023a)	36.31
FACodec (Ju et al., 2024)	52.73

A Analysis about CosyVoice

A.1 Impact of the LLM and CFM modules on voice characteristics

In the zero-shot speech synthesis task, the speaker embeddings input to either the LLM or Flow were replaced with random tensors of the same size. For evaluation, 500 sentences from the LRS3 dataset were selected, and the speaker similarity between the generated speech and the source speech was computed, as shown in Table 4. The results indicate that, compared to CFM, LLM has a significantly smaller impact on voice characteristics due to the limited voice characteristics contained in semantic tokens. Additionally, the balance between semantic and voice characteristics within semantic tokens across different scenarios is worth further exploration.

Additionally, by comparing the performance under both conditions in Table 4, we found that the Mel-spectrogram prompt carries more voice information than the speaker embeddings. In fact, the Mel-spectrograms offers a more detailed representation of voice characteristics, while the speaker embeddings provides a coarser one. For multimodal voice alignment tasks, multimodal voice descriptions are inherently incomplete and imprecise (Leng et al., 2024; Sheng et al., 2024), with a one-to-many mapping to voice characteristics. Thus, a coarse speaker embeddings is sufficient to serve as an anchor for multimodal alignment.

A.2 Performance of CosyVoice on zero-shot Voice Conversion

Before self-distillation, we evaluated the zero-shot voice conversion performance of CosyVoice. We extracted semantic tokens from the source speech and speaker embeddings, along with the Mel-spectrogram prompt from the reference speech, as inputs for the CFM. The generated speech retained the content and prosody of the source while altering the speaker's identity. We randomly selected 500 sentences from the LibriTTS test set to evaluate the performance of the CosyVoice, FreeVC[3] (Li et al., 2023a), and FAcodec[4] (Ju et al., 2024) models, with the experimental results presented in Table 5. During inference, when both the Mel-spectrogram prompt and speaker embeddings were provided, CFM-generated speech surpasses real speech in objective metrics. This

[3] https://github.com/OlaWod/FreeVC

[4] https://github.com/Plachtaa/FAcodec

is attributed to the fact that voice characteristics can exhibit local variations driven by content, rhythm, and emotion. This suggests that the audio produced by CFM is independent of the speaker information in the source semantic tokens, achieving exceptional disentanglement of voice characteristics. This makes it well-suited for self-distillation.

Without the Mel-spectrogram prompt, performance was inferior to FAcodec, which can be attributed to the inconsistency between training and inference, as the model was trained with both the Mel-spectrogram prompt and speaker embeddings as input. After self-distillation, the performance relying solely on speaker embeddings showed a marked improvement, as indicated in Table 3.

A.3 Preliminary experiment on face-based voice description integration

In our preliminary experiments, we directly integrated face embeddings into the CFM of the official CosyVoice[5]. Specifically, we utilized a pre-trained face encoder to extract the global face embeddings, replacing the speaker embeddings in the CFM as input. We evaluated the trained model on the zero-shot voice conversion task and found that the resulting SSIM score was around 4.8, indicating that the generated speech retained the identity information of the source speaker and did not achieve speaker conversion. This suggests that due to the cross-modal gap, CFM tends to extract voice information from the semantic tokens while neglecting the speaker information contained in the face. Therefore, to enable CFM to effectively utilize multimodal voice description integration, further voice disentanglement are necessary.

B Details of Voice Attribute Descriptions

For voice attribute description input, the model receives an input tuple consisting of two speech segments (A and B) and a text description t. The text description states that A exhibits a certain attribute more prominently than B. For example, t refers to "sounds more magnetic" meaning that voice characteritics of sample A is more magnetic than that of B. Following VoxEditor (Sheng et al., 2024), we first concatenate the speaker embeddings $\mathbf{s}_r^A, \mathbf{s}_r^B$ of two given speech samples and the text representation $\mathbf{s}_t$. Through MLP and Gaussian sampling, we predict the density difference $\alpha \in [0, 1]$ in attribute x between the two speeches samples, and then obtain the target speaker embeddings $\mathbf{s}_r$ via linear interpolation: $\mathbf{s}_r = (1-\alpha) \cdot \mathbf{s}_r^B + \alpha \cdot \mathbf{s}_t$. During inference, we can control the density of the target voice attribute by adjusting α within a range of 0 to 1.

C Details about InfoNCE

Specifically, given a batch of N multimodal-voice speaker embeddings pairs $\{(\mathbf{v}_m^i, \mathbf{s}_r^i)\}_{i=1}^N$, the multimodal-voice similarity vector $p_i(\mathbf{v}_m, \mathbf{s}_r) = \{p_{ij}(\mathbf{v}_m, \mathbf{s}_r)\}_{j=1}^N$ and voice-to-multimodal similarity vector $p_i(\mathbf{s}_r, \mathbf{v}_m) = \{p_{ij}(\mathbf{s}_r, \mathbf{v}_m)\}_{j=1}^N$ can be calculated as follows,

$$p_{ij}(\mathbf{v}_m, \mathbf{s}_r) = \frac{\exp\left(\text{sim}\left(\mathbf{v}_m^i, \mathbf{s}_r^j\right)/\tau\right)}{\sum_{j=1}^N \exp\left(\text{sim}\left(\mathbf{v}_m^i, \mathbf{s}_r^j\right)/\tau\right)}, p_{ij}(\mathbf{s}_r, \mathbf{v}_m) = \frac{\exp\left(\text{sim}\left(\mathbf{s}_r^i, \mathbf{v}_n^j\right)/\tau\right)}{\sum_{j=1}^N \exp\left(\text{sim}\left(\mathbf{s}_r^i, \mathbf{v}_n^j\right)/\tau\right)} \tag{10}$$

where τ is a learnable temperature coefficient, initialized to 0.07, and $\text{sim}()$ denotes the dot product used to calculate similarity. The inter-modal alignment loss is computed using cross-entropy as follows,

$$\mathcal{L}_{\text{INTER}} = \frac{1}{2N}\sum_{i=1}^{N}\mathcal{L}_{CE}\left(\mathbf{y}_i, p_i(\mathbf{v}_m, \mathbf{s}_r)\right) + \frac{1}{2N}\sum_{i=1}^{N}\mathcal{L}_{CE}\left(\mathbf{y}_i, p_i(\mathbf{s}_r, \mathbf{v}_m)\right) \tag{11}$$

where $\mathcal{L}_{CE}$ denotes the cross-entropy operation and $\mathbf{y}_i$ the one-hot label of i_{th} pair.

[5] https://github.com/FunAudioLLM/CosyVoice

Table 6: An Example of using LLM to generate synonymous sentences.

Diglogue	
LLM prompts:	Rewrite the following sentence, keeping the meaning unchanged, with a variety of sentence structures and styles. Please replace key words with synonyms: Princess X is honored as a priestess of the winter sea god, portrayed as a woman imbued with deep nostalgia and melancholy, while also being a contemporary fashion designer who cherishes traditional craftsmanship.
Response:	Princess X is revered as a high priestess of the deity of the winter sea, depicted as a figure filled with profound wistfulness and sorrow, yet she is also a modern fashion designer who values artisanal traditions.

D Details of Datasets and Benchmark

D.1 Training Datasets

For the LRS3-TED video dataset, 100 facial images per speaker were randomly selected from the videos, and a facial attribute detection model FairFace[6] was used to further clean the data. Specifically, the speaker's age and gender were estimated based on the 100 images, calculating the mean and variance. If the variance was too large, indicating poor video quality for that speaker, all samples from that speaker were discarded. Anomalies in these 100 images, often blurry pictures or images of a different speaker, were also filtered out. During training, a random image from the given speaker's image set was selected as input. FFmpeg[7]was used to extract 16kHz audio from the video. Additionally, the LRS3 dataset is also utilized for self-distillation of the CFM. For libritts-p, following prompttts2 (Leng et al., 2024), we converted the these word-level annotations about voice characteristics into natural descriptive language using a language model.

D.2 Evaluation Datasets

To evaluate the effectiveness of the text descriptions, we used a language model to rewrite sentences from the validation set while maintaining their original meaning, as shown in Table 6. This approach allows us to assess the model's generalization ability while providing targeted audio for comparison. Additionally, we prompted a large language model to randomly generate 100 character descriptions and voice characteristics descriptions, which can be considered out-of-domain. To further validate out-of-domain face image, we selected an Asian face dataset[8] for testing, given that the LRS3 dataset was collected from TED. The generated speech are available on the demo website.

D.3 Evaluation Metrics

For SST, when performing FaceTTS, FaceVC, TextTTS, and TextVC tasks, a multimodal voice description is provided along with a corresponding target speech. This allows us to directly calculate the speaker similarity between the generated speech and the target speech. However, for the AVE task, as there are no real voice characteristics, we calculate the speaker similarity with the source speech. The AVE task aims to edit specific voice attributes while preserving other characteristics as much as possible, so SST is used to assess whether the edited speech retains the original voice characteristics. Therefore, we need to combine SST and MOS-Match to comprehensively evaluate the performance of AVE.

For SSD, we matched generated speech with voice descriptions for different speakers to calculate speaker similarity, and then averaged the results across the evaluation dataset. A smaller average indicates greater voice diversity within the dataset. Specifically, for the AVE task, the diversity of the generated speech is assessed by applying the same voice attribute editing with the same weights to different speech inputs.

[6] https://github.com/dchen236/FairFace

[7] https://ffmpeg.org/

[8] https://github.com/X-zhangyang/Asian-Face-Image-Dataset-AFD-dataset

For SSC, pairwise matching of different images of the same person was performed to calculate their speaker similarity. These values were then averaged across the entire evaluation dataset. A higher average indicates greater voice similarity between different photos of the same individual, suggesting that the model is robust to background noise and other variations in the images.

For MOS-Match and MOS-Nat, subjective evaluation were conducted on Amazon Mechanical Turk[9]. Twenty sentences were randomly selected, and 20 listeners were asked to score each generated utterance on a scale from 1 (completely mismatched or completely unnatural) to 5 (completely matched or completely natural) for both metrics.

E Comparative Methods

FaceTTS baselines:

- Imaginary Voice (Lee et al., 2023) is based on a score-based diffusion model, specifically Grad-TTS. Imaginary Voice used perceptual loss applied to the Mel-spectrograms to further align facial features and language.
- Face-StyleSpeech (Kang et al., 2023) proposes the disentangling of prosody and timbre, using facial features to control timbre and reference audio to control prosody. It also employs a contrastive learning to align facial and speaker embeddings.
- SYNTHE-SEES (Park et al., 2024) utilizes three types of losses—contrastive learning, speaker classification, and perceptual loss—to align face and speaker embeddings.

FaceVC baselines:

- FaceVC (Lu et al., 2021) employed a three stage training strategy, including face-voice reparameterization and facial-to-audio transformation, to align the face and voice characteristics.
- SP-FaceVC (Weng et al., 2023) first employed a bottleneck-free strategy for speech disentanglement. Then, multi-Scale discriminator and feature matching loss was proposed to improve performance.
- FVMVC Sheng et al. (2023) used FaceNet to extract general face embeddings and employ the memory net to align the face embeddings and speaker embeddings.

TextTTS baselines:

- PromptSpeaker (Zhang et al., 2023) annotated an internal dataset of speaker descriptions on LibriTTS-R. Building on this dataset, PromptSpeaker employed a pre-trained BERT network in conjunction with a Glow model to achieve alignment with speaker embeddings.
- Prompttts++ (Shimizu et al., 2024) integrated a BERT network with a Gaussian mixture model to predict speaker embeddings based on text descriptions, utilizing cosine loss for alignment.
- CosyVoice-Instruct (Du et al., 2024) concatenated the speaker's description before the text content in the LLm module of CosyVoice during training.

TextVC baseline:

- PromptVC (Yao et al., 2024a) utilized HuBERT and k-means clustering to represent semantic intermediate representations, and employed a diffusion model to predict style representations based on text input. Here, we replaced the dataset with ours to predict speaker embeddings using the diffusion model.

AVE baseline:

- VoxEditor (Sheng et al., 2024) first annotated a dataset describing timbre characteristics and utilized a residual memory network to accomplish the voice attribute editing.

[9] https://www.mturk.com/

Table 7: The detailed model configurations of MVA.

Configuration	Value
Layer	8
Attention Dim	768
Attention Heads	16
Linear Dim	2048
Dropout	0.1
KV Size	128

Table 8: The results of ablation studies on TextTTS and TextVC tasks, measured by SST, SSD.

Task	Methods	SST ↑	SSD ↓
TextTTS	UniSpeaker	23.09	21.10
	w.o. MVA	21.07	21.18
	w.o. SoftCL	22.57	34.51
TextVC	UniSpeaker	24.45	24.04
	w.o. MVA	21.50	24.26
	w.o. SoftCL	22.06	35.07

F Further Ablation Studies

We present the ablation experiment results for the TextTTS and TextVC tasks in Figure 8. This indicates that MVA and SoftCL are also beneficial for text-based timbre control. Additionally, we conducted ablation experiments on the size of learnable key-value vectors and the number of MVA layers, and found that within a certain range, the performance of voice control is not significantly affected, yet no clear patterns could be derived.

G Further Discussion

A unified voice space is constructed through a unified voice compressor. To validate the benefits of this shared space, voice interpolation on the speaker embeddings from different modalities is performed, allowing for manually adjusting the interpolation weights α. As shown in Figure 6, we achieve voice control by interpolating the speaker embeddings obtained from face and textual descriptions. We observe that the voice characteristics vary as α changes, speech samples are available in the demo page.

By mapping multiple modalities to a unified voice space, we can leverage these different modalities to more comprehensively describe voice characteristics. Both face images and textual descriptions maintain a one-to-many relationship with the voice characteristics themselves. This means that given a face image or a textual description, the model can generate multiple matching voice characteristics. When both the target speaker's face and the textual voice description are input simultaneously, the generated voice characteristics that align with both modalities will better meet user expectations. Furthermore, we can editing specific voice attributes for more refined optimization. In the future, we will explore how to more finely utilize multiple modalities for voice control concurrently.

Figure 5: Average preference scores (%) of ABX tests about voice suitability in comparison, where participants were asked to select which of two speech samples—one generated based on the reference speaker's face image and one from the reference speaker's recording—better matched the speaker's appearance. "N/P" stands for "no preference". "Ground Truth" represents the real recording of the reference speaker.

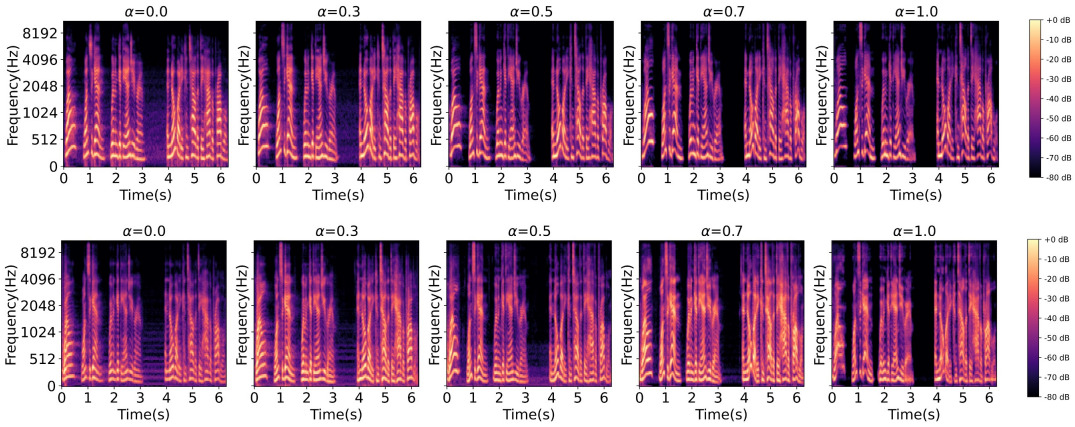


Figure 6: The voice characteristics controlled by both face and textual descriptions varies as α changes. When $\alpha = 0$, the voice characteristics are fully controlled by the face; when $\alpha = 1$, the voice characteristics are fully controlled by the textual description. We can observe the changes in voice characteristics and manually adjust α to achieve the desired voice characteristics.