SFM-Adapter: Style-aware Feature Manipulation Adapter for Speech Style Editing

Yun Chen, Haohe Liu, Qi Chen, Arshdeep Singh, Junqi Zhao, Wenwu Wang, *Fellow, IEEE*, Philip J.B. Jackson, *Member, IEEE*, Mark D. Plumbley, *Fellow, IEEE*

***Abstract*—Speech Style Editing (SSE) aims to modify selected style attributes (e.g., timbre, emotion, pitch) while preserving the linguistic content and all other style attributes that are not given. Many speech applications require flexible control over speech style, making SSE increasingly important. Existing SSE approaches typically follow a style-generation paradigm that synthesizes non-linguistic attributes from style conditions. However, this often results in limited preservation of source attributes and insufficient flexibility when only a subset of style attributes is specified. To overcome these limitations, we adopt a style editing paradigm, in which the target style is achieved by adjusting the source speech instead of producing speech from scratch. Building on this paradigm, we propose a diffusion-based framework with a Style-aware Feature Manipulation Adapter (SFM-Adapter). The SFM-Adapter performs feature-level modulation by integrating user-provided style information with source speech features through multi-layer cross-attention. The resulting modulated features are incorporated into the generation process via mask attention. During inference, a Large Audio-Language Model (LALM)-based length regulation is designed to predict speaking speed and adjust duration. Experiments across multiple speech style editing tasks demonstrate that the SFM-Adapter achieves more natural, accurate, and source-preserving style editing compared with existing methods. Speech samples are provided in https://ychenn1.github.io/SFM-Adapter/.**

***Index Terms*—Speech style editing, Style-aware, Feature manipulation, Diffusion models**

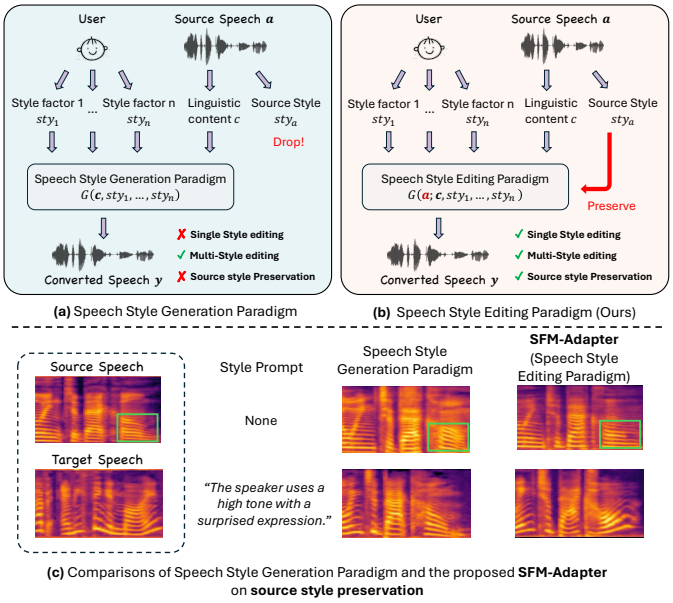


Fig. 1. Motivation. (a) Previous style-generation-based paradigm, (b) style-editing-based paradigm (SFM-Adapter), and (c) evidence that previous methods lack the ability to selectively modify specific aspects of style. As shown in the green box, when the style prompt is absent, the speech style generation paradigm alters the source style, whereas the speech style editing paradigm preserves stylistic characteristics consistent with the source speech.

I. Introduction

Speech style plays an important role in human communication, as it can convey a speaker's emotion, attitude, and even mental state [1]. In speech style modeling, style-related attributes are typically described using two major groups: (1) prosodic or expressive features, such as pitch, duration, and speaking rate, and (2) timbre or voice-quality features that capture speaker-dependent spectral characteristics [1, 2]. While timbre contributes to the identity and texture of the voice, expressive features shape how the message is delivered, making them both essential in speech style modeling. Understanding and modeling speech style holds significant value for a wide range of applications, including personalised speech synthesis [3], voice dubbing [4], and voice cloning [5]. Many of these applications require the ability to control or manipulate specific style attributes of speech, which has led to increasing interest in speech style editing (SSE) [3, 6].

In the past, voice style transfer and voice conversion methods were commonly used as speech style editing approaches, achieving progress in timbre editing [3, 7, 8] and expressiveness editing [9, 10, 11, 12]. However, these methods handle style dimensions such as emotion, timbre, speaking rate, and pitch independently, preventing them from modeling their interactions. For example, these may convert "sadness" to "excitement" without modeling how different style attributes interact. In practice, "sadness" is typically associated with lower pitch, whereas "excitement" often involves rising pitch. Moreover, pitch distributions vary significantly across different timbres (e.g., male vs. female or child vs. adult speakers), even when expressing the same emotion.

Yun Chen is now with the Department of Informatics, King's College London, London, U.K. This work was conducted while she was with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, U.K. (e-mail: yun.2.chen@kcl.ac.uk).

Haohe Liu, Junqi Zhao, Wenwu Wang and Philip J. B. Jackson are with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, U.K. (e-mail: haohe.liu@surrey.ac.uk, junqi.zhao@surrey.ac.uk, w.wang@surrey.ac.uk, p.jackson@surrey.ac.uk).

Qi Chen is with ByteDance Intelligent Creation, Shanghai, China (e-mail: qichen377@gmail.com).

Arshdeep Singh and Mark D. Plumbley are with King's College London, U.K. (e-mail: arshdeep.singh@kcl.ac.uk; mark.plumbley@kcl.ac.uk).

This work was supported by a research scholarship from the China Scholarship Council (CSC), the Engineering and Physical Sciences Research Council [grant numbers EP/T019751/1, EP/Y028805/1], and an Adobe Research Gift.

In recent years, some methods have attempted multi-style editing by disentangling style and linguistic content, where the overall style representation is separated from the content representation to enable controllable recombination [3, 6, 13]. This disentanglement allows users to specify multiple style conditions, such as timbre and emotion, which the model recombines to generate speech with the desired styles, as illustrated in Fig. 1(a). In these approaches, the source speech $\boldsymbol{a}$ is used only to provide the linguistic content $\boldsymbol{c}$, while all other information in $\boldsymbol{a}$ beyond content is discarded. The target stylistic attributes are then generated entirely from the given style conditions. We refer to these works as the **speech style generation paradigm**, which can be formulated as $G(\boldsymbol{c}, sty_1, \ldots, sty_n) \rightarrow \boldsymbol{y}$, where $\boldsymbol{y}$ is the generated speech. In Fig. 1(c), we compare the impact of the expressive condition when the other components remain the same. As highlighted in the green box, in the absence of an explicit expressive condition, the generated speech exhibits lower expressive similarity to the source speech, rather than naturally preserving the original source style. This suggests that, under such a formulation, it is less straightforward to explicitly modify only selected style attributes while preserving other unspecified source-related attributes.

To better support selective style manipulation, more recent approaches [8, 14] incorporate the full source speech $\boldsymbol{a}$ into the model, enabling interaction between the user-specified style conditions and the existing attributes of the source speech. Formally, this can be viewed as a **speech style editing paradigm**, described as $G(\boldsymbol{a}, \boldsymbol{c}, sty_1, \ldots, sty_n) \rightarrow \boldsymbol{y}$, where $\boldsymbol{a}$ provides the complete source speech information and $\boldsymbol{c}$ denotes the linguistic content condition. As illustrated in Fig. 1(b), under this paradigm, the model is designed to modify the style attributes specified by the prompts while preserving the remaining source-related attributes as much as possible. For instance, if the prompt contains only "high pitch", the model changes pitch while attempting to keep emotion and timbre unchanged; if both timbre and emotion are provided, it edits these prompted attributes jointly.

Under the speech style editing paradigm, we aim to provide flexible and diverse style control over the source speech. To support such fine-grained manipulation, natural-language instructions offer a particularly suitable interface, as they can express a wide range of editing intents beyond predefined style labels. However, this setting requires speech data paired with detailed style descriptions, which remain scarce in current public datasets. Among the two mainstream model families, audio Large Language Models (LLMs) and diffusion models, prior studies [15, 16] have shown that diffusion-based methods are generally more suitable under limited-data conditions, whereas audio LLM-based models typically require substantially more training data to achieve robust performance. Therefore, in this work, we adopt a reconstruction-based diffusion framework for speech style editing, where the model is trained to reconstruct the original speech from style conditions derived from the same sample.

Building on these ideas, we propose a Style-aware Feature Manipulation Adapter (**SFM-Adapter**) within a Siamese-like diffusion architecture [17] that can incorporate external style conditions and enable precise style manipulation during conversion. As illustrated in Fig. 2, our framework contains a Source Branch that extracts style-and-content features from the input speech, and a Conversion Branch that generates the edited output. We first employ a Multi-Condition Embedder (MCE) to convert two types of style input (a text prompt and an audio prompt) into style features. The text prompt, described in natural language, conveys comprehensive and detailed expressive features. The audio prompt offers precise timbre attributes, compensating for the limitations of natural language descriptions in capturing timbre nuances. The SFM-Adapter employs a learnable style probe, implemented as a set of learnable query vectors, and progressively extracts stylistic information from the source speech through multi-layer cross-attention. During this process, the stylistic features extracted by the MCE interact with the source speech features, facilitating the manipulation of the desired style while maintaining unchanged style attributes. Finally, the manipulated source features and the provided style features are combined as style conditions and integrated into the Conversion Branch through mask attention to complete the style editing. During inference, we apply Classifier-Free Guidance [18] to enhance the controllability of style editing. Additionally, we introduce a Large Audio-Language Model (LALM) that analyzes expressive features and adaptively determines the appropriate speech rate and relative duration changes to further improve editing performance.

The main contributions of this paper are summarized as follows:

- We identify the challenge of inadequate preservation of source speech attributes in the speech style generation paradigm. To address this limitation, we incorporate the source speech and formulate a speech style editing paradigm that supports multiple style editing tasks, enabling controllable manipulation of speech style.
- We propose a Style-aware Feature Manipulation Adapter that explicitly modulates source speech style features conditioned on user-specified style information, and injects the manipulated styles into the editing process via specially designed masked attention.
- We introduce LALM-based length regulation to adaptively adjust the speed of the generated speech, and employ multi-condition sampling to control the strength of style conditions.
- Experiments show that our framework achieves superior style editing accuracy and produces more natural speech compared to existing methods across multiple SSE tasks.

II. Related works

A. Speech Style Editing

Timbre editing. Timbre editing aims to convert the speech of a source speaker to resemble that of a target speaker while preserving linguistic content. A key challenge in this task is to disentangle speaker-specific characteristics from speech without compromising speaker-agnostic representations [7]. Phonetic posteriorgrams (PPGs), which represent frame-level

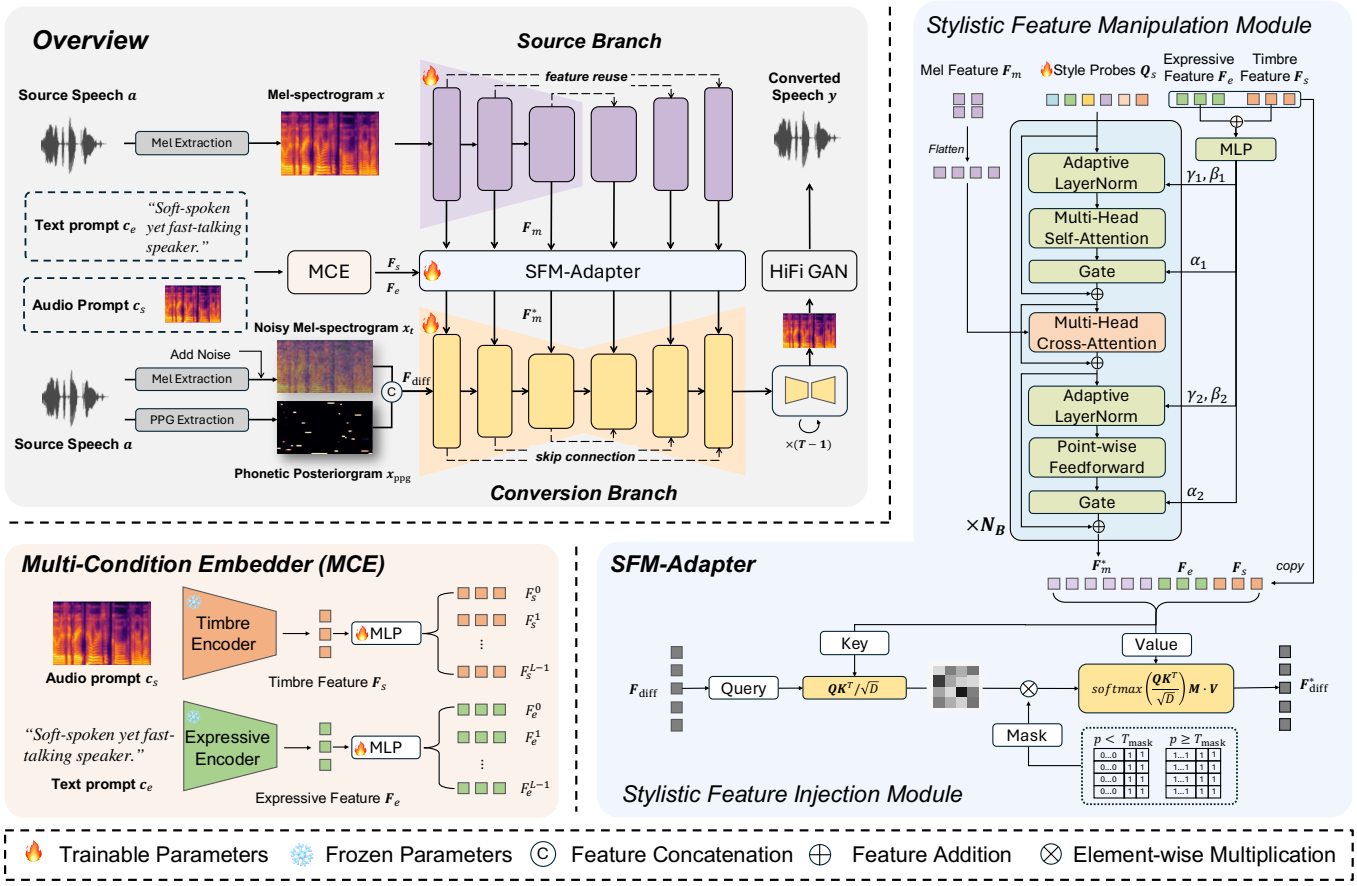


Fig. 2. **Overview of the proposed framework.** The framework is built on a Siamese architecture comprising two branches: the Source Branch for extracting source speech features and a Conversion Branch for reconstructing the edited speech with the guidance of style signals. These two branches are connected through the Multi-Condition Embedder and the SFM-Adapter.

posterior probabilities over phonetic units extracted by an automatic speech recognition model, have been widely adopted for speaker-independent conversion [19, 20]. More recently, self-supervised learning (SSL) methods have been used to enhance content disentanglement in timbre editing models [21, 22, 23], with some leveraging HuBERT [24] and others utilizing WavLM [25].

Expressive style editing. Expressive style editing is a widely studied task that aims to modify the emotional expression of speech while preserving its linguistic content. Some studies use emotion labels [9, 10, 26] to control the expressiveness of generated speech, while others [10, 27] employ averaged emotion representations derived from emotion classifiers by computing the mean feature vector for a specific emotion category. Additionally, some approaches [8, 11, 12] utilize reference speech as an emotion source, allowing the generated speech to align with the expressive characteristics of the reference.

Multi-style editing. To generate expressive speech, recent studies have explored methods that jointly edit multiple speech attributes, such as timbre, emotion, and speaking speed. Some works adopt a sequence-to-sequence formulation [28, 29] or introduce additional modeling for prosody features [30, 31] to achieve multi-style editing. However, these methods rely on a reference speech to provide the style, which limits their flexibility. For example, if no reference speech with the desired style is available, conversion to that style becomes impossible. Since text can describe speech style from multiple perspectives, and is not constrained by the limitations of using reference speech alone, text prompts have attracted growing attention in recent style editing research. Kuan et al. [32] introduced TextguidedVC, which employs natural language instructions as guidance for the style editing process. PromptVC [33] and HybridVC [6] propose speech-to-speech style editing models that use text prompts to generate style representations for controlling the generated speech. However, these approaches often fail to preserve source speech style attributes that are not explicitly controlled by the text prompts.

B. Denoising Diffusion Probability Models (DDPMs)

Diffusion models [34] have achieved the state-of-the-art sample quality in various tasks, e.g., image generation [35, 36], video generation [37, 38, 39] and audio generation [40, 41, 42]. They model the data distribution by iteratively denoising samples from Gaussian noise, gradually transforming noise into realistic data through a learned reverse diffusion process [43]. Diffusion models have recently been introduced into voice conversion (VC) to model complex speech distributions through iterative denoising processes [43]. Diff-HierVC [30]

employs a hierarchical diffusion framework to generate F0 trajectories for expressive zero-shot voice style transfer, achieving improved pronunciation and natural intonation, but its controllability is limited to pitch-related attributes. PromptVC [33] extends diffusion-based style editing by modeling global style representations conditioned on natural language prompts, enabling flexible control over multiple style attributes beyond pitch. In contrast, PromptEVC [44] integrates diffusion models at the emotion embedding level, focusing on fine-grained expressive style editing with explicit control over prosodic patterns and intensity. ClapFM-EVC [45] combines language-audio contrastive pretraining with a conditional flow-matching generative model to enable high-fidelity emotional voice conversion driven by natural language prompts or reference speech.

III. Method

A. Problem Definition

Given an input speech sample $\boldsymbol{a}$ and target styles, the goal is to edit its style to match the target style conditions while preserving all other aspects of $\boldsymbol{a}$. We use two types of prompts to guide the editing process. For timbre information, we use a reference speech sample $\boldsymbol{c}_s$ to provide the target speaking timbre characteristics. We also use natural language text descriptions $\boldsymbol{c}_e$ to provide expressive style attributes such as emotion, speaking rate, and pitch. By using our model G, the generated speech $\boldsymbol{y}$ can be obtained as:

$$\boldsymbol{y} = G(\boldsymbol{a}, \boldsymbol{c}_e, \boldsymbol{c}_s). \tag{1}$$

The values $\boldsymbol{c}_e$ and $\boldsymbol{c}_s$ can be set to a zero vector $\emptyset$ if needed, indicating that the corresponding condition is not provided, enabling the model to edit only the specified styles while preserving the other aspects of the speech.

B. Overview

As shown in Fig. 2, the proposed framework is built on a Siamese architecture comprising two branches: the Source Branch $\mathcal{S}(\cdot)$ and the Conversion Branch $\mathcal{C}(\cdot)$. The Conversion Branch adopts a U-Net-based DDPM to generate the edited speech, while the Source Branch, which has the same encoder architecture as the U-Net backbone but with separate model weights, is responsible for extracting multi-scale features from the source speech. These two branches are connected through the Multi-Condition Embedder (MCE) and a Style-aware Feature Manipulation Adapter (SFM-Adapter) (See Sec. III-D).

During training, the model follows a reconstruction-based DDPM paradigm without paired edited speech supervision. For each speech sample $\boldsymbol{a}$, the audio prompt $\boldsymbol{c}_s$ is randomly cropped from $\boldsymbol{a}$ to provide timbre guidance, and the text prompt $\boldsymbol{c}_e$ describes the expressive style of the speech sample. Conditioned on the speech sample and these prompts, the Conversion Branch is trained to predict the Gaussian noise $\epsilon \sim N(0, I)$ added to the mel-spectrogram of $\boldsymbol{a}$, thereby learning to reconstruct the original mel-spectrogram from its noisy version. During inference, any reference audio prompt and text description can be used to specify the target timbre and expressive style. The audio prompt can be any speech from the target speaker, regardless of its transcript or expressive style. Specifically, the editing process consists of three stages. **First**, the source speech $\boldsymbol{a}$ is processed to extract a mel-spectrogram $\boldsymbol{x}$ and PPG features $\boldsymbol{x}_{\text{ppg}}$, using conventional signal processing methods and a neural PPG model [46], respectively. In the Source Branch, the mel-spectrogram is input to extract multi-scale source speech features $\boldsymbol{F}_m$. In the Conversion Branch, we add Gaussian noises ϵ to the extracted mel-spectrogram, constructing the noisy mel-spectrogram for training diffusion models. Since PPG features primarily encode linguistic content information [46], they can be treated as content conditions for speech generation. Therefore, we concatenate the PPG features with the noisy mel-spectrogram along the channel dimension and feed them into the Conversion Branch to enhance linguistic accuracy. Simultaneously, both text and audio prompts are processed through the Multi-Condition Embedder, generating stylistic features. **Then** $\boldsymbol{F}_m$ is interacted with the extracted timbre features $\boldsymbol{F}_s$ and expressive features $\boldsymbol{F}_e$ to obtain $\boldsymbol{F}_m^*$. Subsequently, the noisy features $\boldsymbol{F}_{\text{diff}}$ integrate the stylistic features from $\boldsymbol{F}_m^*$, $\boldsymbol{F}_s$, and $\boldsymbol{F}_e$ to the conversion diffusion process via the mask-attention. **Finally**, the Conversion Branch outputs the corresponding denoised mel-spectrogram, which is then decoded to a waveform using a HiFi-GAN [47], a high-fidelity neural vocoder widely adopted in previous speech synthesis and voice conversion works due to its high reconstruction quality [7, 12].

C. Multi-Condition Embedder

Given the text prompt $\boldsymbol{c}_e$ and audio prompt $\boldsymbol{c}_s$, a Multi-Condition Embedder is employed to extract stylistic features. Specifically, a pretrained timbre encoder $\mathcal{E}_s$ is utilized to extract the timbre features $\boldsymbol{F}_s = \mathcal{E}_s(\boldsymbol{c}_s) \in \mathbb{R}^{N_s \times D_s}$, where N_s denotes the number of timbre features, and D_s represents the dimension of the timbre features. Similarly, a pretrained text encoder $\mathcal{E}_e$ is used to extract the expressive features $\boldsymbol{F}_e = \mathcal{E}_e(\boldsymbol{c}_e) \in \mathbb{R}^{N_e \times D_e}$, where N_e and D_e denote the number and dimension of the expressive features, respectively.

To provide multi-level stylistic guidance for style manipulation, we employ several projection layers to transform the input style features $\boldsymbol{F}_e$ and $\boldsymbol{F}_s$ into a set of features $\{\boldsymbol{F}_e^l\}_{l=0}^{L-1}$ and $\{\boldsymbol{F}_s^l\}_{l=0}^{L-1}$, where L is the number of layers in Conversion Branch.

D. SFM-Adapter

We propose the SFM-Adapter to manipulate the source speech features $\boldsymbol{F}_m$ using the stylistic features $\boldsymbol{F}_s$ and $\boldsymbol{F}_e$. It is also designed to preserve useful source information while mitigating potential source style leakage. The resulting manipulated source features are then injected into the Conversion Branch. As illustrated in Fig. 2, the SFM-Adapter comprises two modules: the Stylistic Feature Manipulation Module and the Stylistic Feature Injection Module. These modules are integrated into every layer of the Conversion Branch. For simplicity, we omit the layer index in the notation in this section.

Stylistic Feature Manipulation Module. Inspired by Querying Transformer (Q-Former) [48], we randomly initialize

learnable style queries $\boldsymbol{Q}_0 \in \mathbb{R}^{N_q \times D}$ to query out features from the source branch and input style conditions for manipulating stylistic features. N_q is the number of learnable query tokens, and D represents the hidden feature dimension. As shown in Fig. 2, the Stylistic Feature Manipulation Module consists of N_B transformer blocks. Each block contains one multi-head self-attention layer, one multi-head cross-attention layer, one point-wise feedforward layer, and two adaptive layernorm (AdaLN) [49] layers. This design enables controlled feature interaction by selectively extracting style-related information from the source branch through cross-attention, while utilizing AdaLN-based modulation to align the extracted source features with the target style features. Instead of directly reusing the full source representation, the learnable queries encourage the model to focus on source information that is most relevant for style manipulation, which helps reduce potential source style leakage.

The computational process of the stylistic feature manipulation module is as follows:

$$\begin{aligned}
&(\alpha_1^n,\ \alpha_2^n,\ \beta_1^n,\ \beta_2^n,\ \gamma_1^n,\ \gamma_2^n) = \mathrm{MLP}_n(\boldsymbol{F}_e, \boldsymbol{F}_s),\\
&\boldsymbol{Q}_n'' = \boldsymbol{Q}_{n-1} + \alpha_1^n \times \mathrm{MSA}_n\left(\mathrm{AdaLN}\left(\boldsymbol{Q}_{n-1},\ \gamma_1^n,\ \beta_1^n\right)\right),\\
&\boldsymbol{Q}_n' = \boldsymbol{Q}_n'' + \mathrm{MCA}_n(\boldsymbol{Q}_n'', \boldsymbol{F}_m),\\
&\boldsymbol{Q}_n = \boldsymbol{Q}_n' + \alpha_2^n \times \mathrm{FFN}_n\left(\mathrm{AdaLN}(\boldsymbol{Q}_n',\ \gamma_2^n,\ \beta_2^n)\right),
\end{aligned} \tag{2}$$

where $n \in [1, 2, \cdots, N_B]$ indicates the n-th block, $\boldsymbol{F}_m$ are the source speech features extracted from the Source Branch, $\mathrm{MSA}_n(\cdot)$ is a multi-head self-attention layer, $\mathrm{MCA}_n(\cdot)$ is a multi-head cross-attention layer, $\mathrm{MLP}_n(\cdot)$ is a multilayer perceptron, $\mathrm{AdaLN}(\cdot)$ is adaptive layer normalization, and $\mathrm{FFN}(\cdot)$ is a feed-forward network. After N_B iterations, we can obtain $\boldsymbol{Q}_{N_B}$. We denote $\boldsymbol{Q}_{N_B}$ as $\boldsymbol{F}_m^*$, since it represents the fused representation of the source style information and the target style information.

Stylistic Feature Injection Module. To integrate the target style into the denoising Conversion Branch, we employ a mask-aware cross-attention mechanism. Given the noise features $\boldsymbol{F}_{\text{diff}} \in \mathbb{R}^{N \times D}$ as a query, which are obtained by concatenating the noised mel-spectrogram at a diffusion timestep t with the corresponding PPGs, the attention module aggregates stylistic information, guiding the model towards the desired style patterns during the denoising process.

Specifically, we concatenate expressive features $\boldsymbol{F}_e \in \mathbb{R}^{N_e \times D}$, the timbre features $\boldsymbol{F}_s \in \mathbb{R}^{N_t \times D}$, and the manipulated source features $\boldsymbol{F}_m^*$ to strengthen the influence of the target style: $\boldsymbol{F}_{\text{cond}} = \mathrm{concat}(\boldsymbol{F}_m^*, \boldsymbol{F}_e, \boldsymbol{F}_s)$. By explicitly strengthening the target style features during feature integration, the model is encouraged to rely more on the desired target attributes rather than implicitly preserving the original source style. Then, the output $\boldsymbol{F}_{\text{diff}}^*$ of the cross-attention is formulated as follows:

$$\boldsymbol{F}_{\text{diff}}^* = \mathrm{softmax}\left(\frac{\boldsymbol{F}_{\text{diff}}(\boldsymbol{F}_{\text{cond}})^T}{\sqrt{D}}\right)\boldsymbol{M} \cdot \boldsymbol{F}_{\text{cond}}, \tag{3}$$

where $\boldsymbol{M}$ is an attention mask applied to the attention weights to selectively suppress contributions from features derived from the Source Branch during training, defined as follows:

$$\boldsymbol{M} = \begin{cases} \begin{bmatrix}\mathbf{1}_{L\times N_q} & \mathbf{1}_{L\times(N_e+N_s)}\end{bmatrix}, & \text{if } \ p \geq T_{\text{mask}} \\ \begin{bmatrix}\mathbf{0}_{L\times N_q} & \mathbf{1}_{L\times(N_e+N_s)}\end{bmatrix}, & \text{if } \ p < T_{\text{mask}} \end{cases}, \tag{4}$$

where $\mathbf{1}$ and $\mathbf{0}$ denote the mask with all ones and zeros elements, respectively. The value $p \sim \mathcal{U}(0,1)$ is a random number uniformly sampled for each sample during the training process, and T_{mask} is a threshold used to control the masking ratio. When p is lower than T_{mask}, the query cannot attend the feature of $\boldsymbol{F}_m^*$. This random feature discarding helps prevent the model from over-relying on source-derived features and reduces the risk of blindly copying undesired source style cues during training.

E. Training Objectives and Classifier-Free Guidance

During training, we optimize the conditional denoising model $\epsilon_\theta(\mathrm{concat}(\boldsymbol{x}_t, \boldsymbol{x}_{\text{ppg}}), t, \boldsymbol{x}, \boldsymbol{c}_e, \boldsymbol{c}_s)$ to predict the added noise using a standard MSE loss:

$$\mathcal{L}_{\text{mse}} = \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{\epsilon}}[||\boldsymbol{\epsilon} - \epsilon_\theta(\mathrm{concat}(\boldsymbol{x}_t, \boldsymbol{x}_{\text{ppg}}), t, \boldsymbol{x}, \boldsymbol{c}_e, \boldsymbol{c}_s)||^2], \tag{5}$$

where $t \sim [1, T]$ is the denoising timestep that is sampled from the uniform distribution, $\boldsymbol{x}_t$ is the noisy mel features at timestep t, and $\boldsymbol{\epsilon}$ is the ground truth noise. The symbols $\boldsymbol{x}_{\text{ppg}}$, $\boldsymbol{x}$, $\boldsymbol{c}_e$ and $\boldsymbol{c}_s$ denote PPG feature, source mel-spectrogram, text condition and audio condition, respectively.

To improve DDPM with fewer sampling steps, we employ an additional loss term $\mathcal{L}_{\text{vlb}}$, corresponding to the variational lower bound, to learn the model variance Σ_θ, following [50].

$$\begin{aligned}
\mathcal{L}_{\text{vlb}} = \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{\epsilon}}\Big[D_{\text{KL}}\Big(&q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) \parallel \\
&p_\theta(\boldsymbol{x}_{t-1} \mid \mathrm{concat}(\boldsymbol{x}_t, \boldsymbol{x}_{\text{ppg}}), t, \boldsymbol{x}, \boldsymbol{c}_e, \boldsymbol{c}_s)\Big)\Big],
\end{aligned} \tag{6}$$

where $D_{\text{KL}}(\cdot\|\cdot)$ denotes the Kullback–Leibler (KL) divergence, which measures the discrepancy between the true posterior distribution $q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0)$ and the model-predicted distribution $p_\theta(\boldsymbol{x}_{t-1} \mid \cdot)$.

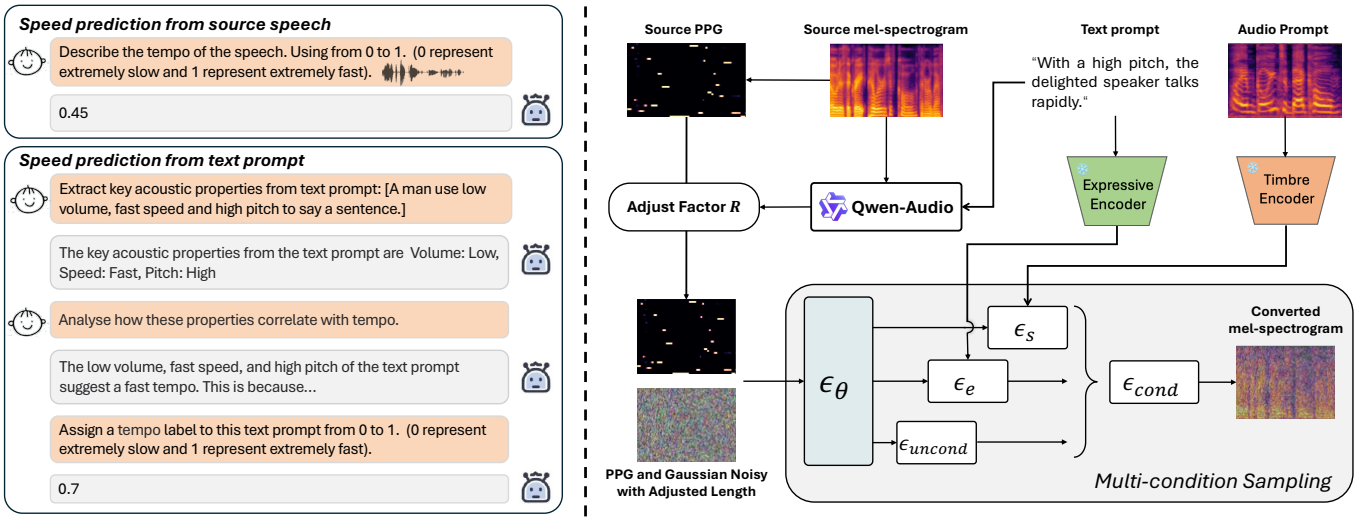
The overall training objective of the proposed framework is as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{mse}} + \mathcal{L}_{\text{vlb}}. \tag{7}$$

To enable controllable generation, we further adopt Classifier-Free Guidance (CFG) [18, 41, 42, 50]. CFG is a guidance strategy for diffusion models that improves conditional generation by jointly training conditional and unconditional denoising models, without relying on an external classifier. During training, the conditional inputs $\boldsymbol{c}_e$ and $\boldsymbol{c}_s$ are randomly dropped with a fixed probability (e.g., 10%) in Eq. 5, allowing the model to learn both conditional and unconditional denoising behaviors.

F. Adaptive Sampling

1) LALM-based Length Regulation: The speed and style of speech are correlated, as style can influence the speed of speech [1]. Given that LALMs exhibit strong comprehension of both text and audio [51], we propose LALM-based Length



(a) LALM-based Length Regulation

(b) Inference stage of the proposed SFM-Adapter

Fig. 3. Illustration of LALM-based Length Regulation and the inference process of SFM-Adapter. In the inference stage of SFM-Adapter, we adopt an adaptive sampling strategy, including LALM-based Length Regulation and Multi-condition Sampling.

Regulation, which leverages LALMs' understanding of style to predict the target speaking speed and adjust the duration of the speech accordingly during style editing. In our work, we use Qwen-Audio [51] as it demonstrates strong capabilities in speech understanding.

To modulate the length based on the source speech, we first extract its speed. As shown in Fig. 3 (a), we design a prompt to obtain the speed of the source speech, resulting in a speed $r_{\text{src}} = \text{LALM}(\boldsymbol{x})$, where $\text{LALM}(\cdot)$ denotes the large audio-language model. To extract the target speed from text prompts, we employ a chain-of-thought (CoT) approach [52]. As illustrated in Fig. 3 (a), the extraction task is decomposed into multiple subtasks, allowing us to progressively obtain the target speed, represented as $r_{\text{trg}} = \text{LALM}(\boldsymbol{c}_e)$. We adopt a logarithmic speed adjustment approach, which better aligns with human sensitivity to speed changes compared to linear adjustment, particularly in transitions between fast and slow speeds [53]. Thus, the speed adjustment factor is defined as $R = \left(\frac{r_{\text{trg}}}{r_{\text{src}}}\right)^{\eta}$, where $\eta \in (0, 1)$ is a smoothing coefficient used to reduce the sensitivity to extreme rate differences. In our experiments, we set $\eta = \frac{1}{\ln 10}$.

By constraining extreme speed changes, the formulation is designed to produce speed transformations that are more natural and better aligned with human perception during speed editing. As illustrated in Fig. 3, the factor R is applied to PPG features. We perform speed interpolation by scaling the length of the PPG features according to R. Given that L_{src} represents the original length of the PPG features and L_{trg} denotes the desired length, the interpolated length is computed as: $L_{\text{trg}} = \frac{L_{\text{src}}}{R}$. The randomly initialized noise $\boldsymbol{x}_T$ is also manipulated in length using R to align with the PPG features, resulting in adjusted Gaussian noise $\boldsymbol{x}'_T$. The adjusted noise is then concatenated with the PPG features along the channel dimension and used as the input for network inference, *i.e.,* $\text{concat}(\boldsymbol{x}'_T, \boldsymbol{x}_{\text{ppg}})$. In this setting, if r_{trg} is greater than r_{src}, the resulting L_{trg} will be shorter than L_{src}, reflecting a faster speech speed. This scaling adjusts the temporal length of the PPG features to match the target speed, enabling the generated speech to follow the speed specified by the text prompt $\boldsymbol{c}_e$.

2) Multi-condition Sampling: Once the model learns style-based speech editing, inference is performed by first sampling a Gaussian noise $\boldsymbol{x}_T$ and then denoising $\boldsymbol{x}_T$ in an iterative manner using the DDPM. To flexibly control the influence of conditions $\boldsymbol{F}_s$ and $\boldsymbol{F}_e$, we adopt classifier-free guidance [18] in our multi-conditional sampling process. The procedure is illustrated in Fig. 3 (b). The denoising procedure is defined as follows:

$$\boldsymbol{\epsilon}_{\text{cond}} = \boldsymbol{\epsilon}_{\text{uncond}} + w_e \boldsymbol{\epsilon}_e + w_s \boldsymbol{\epsilon}_s \ , \tag{8}$$

where $\boldsymbol{\epsilon}_{\text{uncond}} = \epsilon_\theta(\text{concat}(\boldsymbol{x}'_t, \boldsymbol{x}_{\text{ppg}}), t, \boldsymbol{x}, \emptyset, \emptyset)$ denotes the model's unconditioned prediction, and $\boldsymbol{x}'_t$ is the noisy mel-spectrogram at noising step t. Here, both the text and audio prompt conditions are set to zero. The text-guided prediction and the speech-guided prediction are represented by $\boldsymbol{\epsilon}_e = \epsilon_\theta(\text{concat}(\boldsymbol{x}'_t, \boldsymbol{x}_{\text{ppg}}), t, \boldsymbol{x}, \boldsymbol{c}_e, \emptyset) - \boldsymbol{\epsilon}_{\text{uncond}}$ and $\boldsymbol{\epsilon}_s = \epsilon_\theta(\text{concat}(\boldsymbol{x}'_t, \boldsymbol{x}_{\text{ppg}}), t, \boldsymbol{x}, \emptyset, \boldsymbol{c}_s) - \boldsymbol{\epsilon}_{\text{uncond}}$, respectively. The guidance scales corresponding to the text prompt and audio prompt are denoted as w_e and w_s, respectively. A higher w_e enhances style expressiveness, while a larger w_s strengthens speaker timbre preservation.

IV. Experiments

A. Datasets

We conduct our experiments on PromptSpeech [2], Emotional Speech Dataset (ESD) [54], and VCTK [55]. (1) **PromptSpeech** contains over 26,000 real speech samples from LibriTTS [56], each paired with a corresponding text style prompt. The text style prompt in PromptSpeech covers four style factors: gender, pitch, speaking speed, and loudness. (2) **ESD** includes speech samples from 10 native English speakers and 10 native Chinese speakers, covering five emotion categories: neutral, happy, angry, sad, and surprised. ESD

contains 350 distinct utterances in total, where 300 are for training, 20 for validation, and 30 for testing. TextrolSpeech [13] extends ESD by providing text style prompts for its speech samples, incorporating not only the four style factors from PromptSpeech but also an emotion factor. (3) **VCTK** dataset comprises speech samples from 107 speakers. During training, we select 19,400 speech samples from PromptSpeech and 16,500 from ESD, yielding a total of 35,900 training samples. For multi-style editing evaluation, we construct 1,000 speech pairs from ESD. Each pair consists of a source and target speech with the same linguistic content. The expressive style editing evaluation set follows a setup similar to the multi-style setting, with the only difference being that the source and target speech originate from the same speaker. Both evaluation settings use speech samples whose linguistic content is unseen during training. For timbre editing evaluation, we randomly constructed 1,000 speech pairs from VCTK, each comprising a source and a target speech with the same linguistic content but different timbre.

B. Implementation Details

In our experiments, all speech clips are sampled at 16 kHz. The mel-spectrograms are computed from the raw waveforms using a frame size of 1024, a hop size of 256, and 80 mel frequency channels. We employ the HiFi-GAN vocoder [47] to reconstruct waveforms from the generated mel-spectrograms. To extract text prompt features, we utilize a pretrained AnglE model [57], producing 1024-dimensional embeddings. For speaker prompt features, we use a pretrained ECAPA-TDNN model [58], trained on the VoxCeleb2 dataset [59], to extract a 192-dimensional global timbre embedding representing the target speaker.

We train our network on two NVIDIA A100 GPUs for 30,000 iterations with a total batch size of 64, and the AdamW optimizer [60] is used with a learning rate of 0.0001. For a trade-off between computational efficiency and editing quality, we set the number of Q_s to 8 and use $N_B = 2$ blocks in the stylistic feature manipulation module. The masking threshold T_{mask} is fixed at 0.3. During inference, mel-spectrograms are generated using $T = 50$ denoising steps, with guidance scales set to $w_e = 2$ for the text prompt and $w_s = 2$ for the audio prompt. The LALM-based length regulation adopts deterministic decoding. The exact prompts are provided in our demo page.

C. Compared Methods

To comprehensively evaluate our method under different editing scenarios, we conduct three groups of experiments: (1) multi-style editing, (2) expressive style editing, and (3) timbre editing. Since these tasks differ in their objectives and assumptions, we select task-specific baseline methods for fair comparison.

In particular, Vevo [8] is a unified framework that supports multiple editing functionalities, including three distinct settings: joint timbre–expressive style editing, expressive style editing only, and timbre editing only. Following the original paper, we denote these variants as Vevo-Voice, Vevo-Style, and Vevo-Timbre, respectively, and use them in the corresponding experimental settings.

For multi-style editing, we compare our method with Vevo-Voice, which jointly performs timbre and expressive style editing. For expressive style editing, we compare against StyleVC [61], AINN [12], and Vevo-Style, all of which are designed for expressive style editing and trained on the ESD dataset. For timbre editing, we include StyleVC [61], DDDM-VC [62], Diff-HierVC [30], Vevo-Timbre, and FreeVC [23], as these methods specifically focus on speaker timbre editing. All baselines are trained and evaluated on the same datasets as our method to ensure fair comparison.

D. Evaluation Metrics

We perform both objective evaluation and human subjective evaluation. The objective metrics include Word Error Rate (**WER**), Speaker Embedding Cosine Similarity (**SECS**), Pitch Pearson Correlation (**Corr**) [3], Emotion Classification Accuracy (**ACC**) and Unified Speech Mean Opinion Score Predictor (**UTMOS**) [3]. Specifically, WER is calculated based on Whisper-large-v3 [63] to evaluate the generated speech's intelligibility. SECS measures the speaker similarity between the generated speech and the audio prompt, which is calculated using the cosine similarity between speaker embeddings extracted from the ECAPA-TDNN model [58]. To evaluate the quality of the generated speech, we use UTMOS to predict the Mean Opinion Score (MOS). We additionally adopt the Corr and ACC metrics to measure the performance of the style editing. Corr is employed to measure the correlation between the pitch contours of generated speech and target speech. It evaluates how well the pitch of the generated speech aligns with that of the target speech. Additionally, we use a pretrained emotion classifier ACRNN [64] to evaluate the accuracy of the generated speech (ACC). For subjective evaluation, we employ the naturalness Mean Opinion Score (**nMOS**) to assess the naturalness of the generated samples. We further use two similarity MOS metrics: **sMOS-e**, which evaluates the expressive similarity between the generated speech and the text style prompt, and **sMOS-s**, which evaluates the timbre similarity between the generated speech and the reference audio prompt. To evaluate cross-attribute preservation, we define additional metrics for non-target attribute preservation. Specifically, in expressive style editing, $\mathrm{SECS}_{\mathrm{src}}$ is used to measure whether the source speaker identity is preserved after modifying expressive style attributes. In timbre editing, $\mathrm{Corr}_{\mathrm{src}}$ and $\mathrm{ACC}_{\mathrm{src}}$ are used to measure whether the original expressive style is preserved after editing speaker timbre.

Human Evaluation. We conduct a human subjective evaluation to assess the quality of generated speech across three tasks: multi-style editing, expressive style editing, and zero-shot timbre editing. The evaluation set used for the listening test consists of 60 groups, randomly sampled from the full evaluation set, with 20 groups for each task. Naturalness (nMOS), emotional similarity (sMOS-e), and timbre similarity (sMOS-s) scores are assigned by 20 participants who are proficient in English.

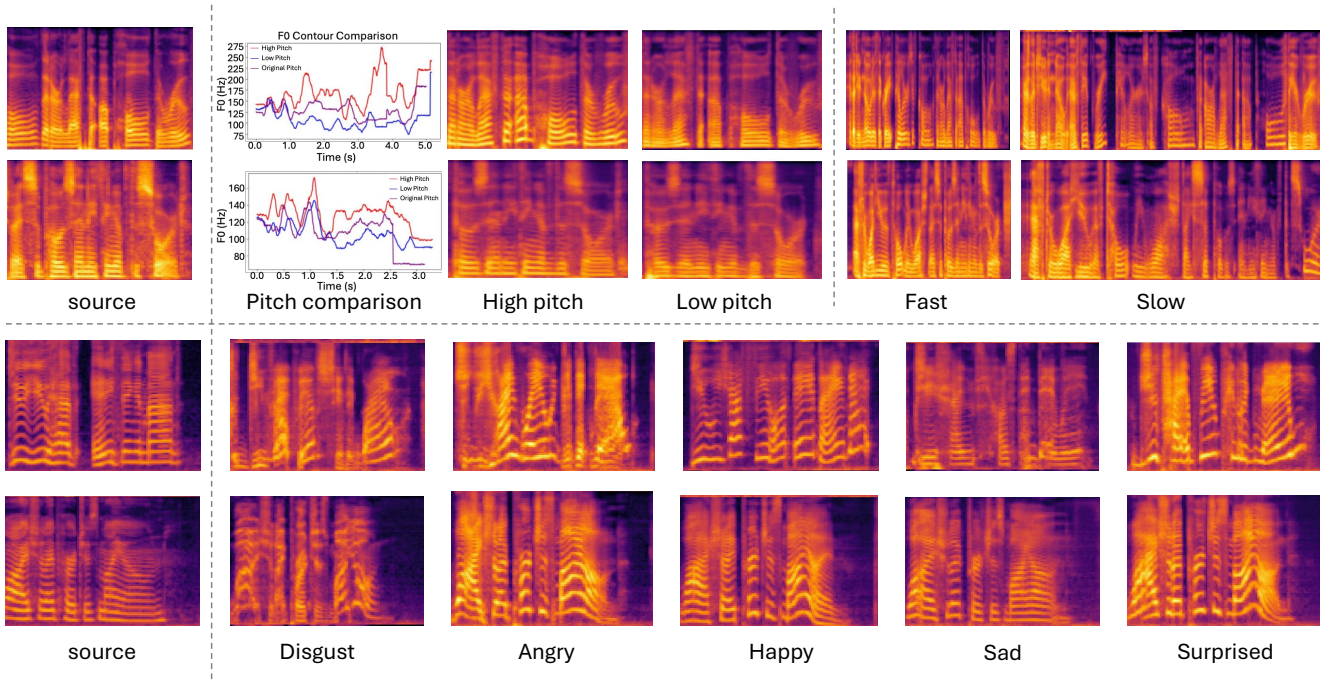


Fig. 4. Visualization of speech style editing. The upper section illustrates pitch and speed editing. In the pitch comparison, the red line represents high pitch, the purple line corresponds to the pitch of the source speech, and the blue line represents low pitch. The lower section presents expressive style editing across five emotion categories.

TABLE I
THE SUBJECTIVE AND OBJECTIVE EVALUATION RESULTS FOR OUR METHODS AND THE BASELINE SYSTEMS IN **MULTI-STYLE EDITING**. ALL SUBJECTIVE METRICS ARE COMPUTED WITH 95 % CONFIDENCE INTERVALS AND "GT" REFERS TO GROUND TRUTH SAMPLES.

Methods	WER ↓	SECS ↑	UTMOS ↑	Corr ↑	ACC ↑	nMOS ↑	sMOS-s ↑	sMOS-e ↑
GT	1.77	-	3.82	-	0.98	4.01 ± 0.12	-	-
Vevo-Voice [8]	**4.91**	0.61	3.60	0.17	0.53	3.74 ± 0.23	3.64 ± 0.26	3.38 ± 0.23
SFM-Adapter (ours)	6.17	**0.64**	**3.73**	**0.38**	**0.70**	**3.78 ± 0.26**	**3.70 ± 0.29**	**3.70 ± 0.30**

E. Performance

We conduct performance comparisons on both multi- and single-style editing settings to demonstrate the editing ability of our framework. Fig. 4 presents visualizations of the generated speech under single-style editing settings. For pitch editing, speech generated with the "high pitch" text prompt exhibits clearly elevated pitch contours compared to that generated with the "low pitch" prompt. Similarly, for expressive style editing, the mel-spectrograms show distinct acoustic differences: the "Sad" style is characterized by flatter pitch contours and lower energy, whereas the "Surprised" style displays rising pitch and higher energy. These observations demonstrate that the proposed model can manipulate the intended acoustic features in response to different style prompts.

Performance on multi-style editing. We conduct multi-style editing experiments to evaluate the effectiveness of our model in handling diverse style editing tasks. We use the caption of the target speech as the text prompt to provide expressive style, and use the target speech as the audio prompt to provide timbre style. As shown in Table I, [1] compared with the state-of-the-art Vevo-Voice [8], SFM-Adapter achieves a higher SECS score of 0.64, along with a Corr of 0.38 and an ACC of 0.70, showing an improvement over Vevo-Voice. SFM-Adapter achieves an sMOS-e of 3.70 and an sMOS-s of 3.70, compared to 3.64 and 3.38 achieved by Vevo-Voice, respectively. This indicates that the generated speech from SFM-Adapter is perceived as more similar to the target in both expressive style and timbre style. Meanwhile, the naturalness of the generated speech is well preserved, with an nMOS of 3.78. Compared to Vevo-Voice, these results highlight the effectiveness of our proposed method. This result also indicates that diffusion-based editing remains competitive and practical under limited paired-data conditions, even when compared with large-scale LLM-based approaches.

Performance on expressive style editing. As shown in Table II, our proposed SFM-Adapter outperforms the previous methods across all three style-related metrics. In terms of style, the previous best method, Vevo-style, achieves an ACC score of 0.50, while our proposed SFM-Adapter method surpasses it with a higher ACC score of 0.73. SFM-Adapter also attains the best Corr score of 0.35, improving over Vevo-Style by 0.11. In addition, SFM-Adapter achieves the highest $\mathrm{SECS}_{\mathrm{src}}$, indicating that it can effectively preserve the source speaker

[1] TextguidedVC [32], HybridVC [6] and PromptVC [33], did not release implementation code and training datasets, so we cannot use them for direct comparison.

TABLE II
THE SUBJECTIVE AND OBJECTIVE EVALUATION RESULTS FOR OUR METHODS AND THE BASELINE SYSTEMS IN **EXPRESSIVE STYLE EDITING**. "CONDITION" REFERS TO THE TYPE OF USER-PROVIDED INPUT USED DURING EDITING. ALL SUBJECTIVE METRICS ARE COMPUTED WITH 95 % CONFIDENCE INTERVALS AND "GT" REFERS TO GROUND TRUTH SAMPLES.

Methods	Condition	WER ↓	Corr ↑	ACC ↑	SECS_{src} ↑	UTMOS ↑	nMOS ↑	sMOS-e ↑
GT	-	1.77	-	0.99	-	3.97	4.12 ± 0.13	-
StyleVC [61]	audio	10.80	0.10	0.23	0.64	3.69	2.97 ± 0.38	2.51 ± 0.41
AINN [12]	audio	13.67	0.21	0.50	0.62	2.37	2.92 ± 0.29	2.41 ± 0.36
Vevo-Style [8]	audio	**3.57**	0.24	0.50	0.70	3.60	3.49 ± 0.27	3.37 ± 0.33
SFM-Adapter (ours)	text	5.30	**0.35**	**0.73**	**0.72**	**3.74**	$\mathbf{3.95 \pm 0.28}$	$\mathbf{4.09 \pm 0.27}$

TABLE III
THE SUBJECTIVE AND OBJECTIVE EVALUATION RESULTS FOR SFM-ADAPTER AND THE BASELINE SYSTEMS IN ZERO-SHOT **TIMBRE EDITING**. ALL SUBJECTIVE METRICS ARE COMPUTED WITH 95 % CONFIDENCE INTERVALS AND "GT" REFERS TO GROUND TRUTH SAMPLES. $*$ INDICATES FREEVC IS NOT IN A ZERO-SHOT SETTING.

Methods	WER ↓	SECS ↑	Corr_{src} ↑	ACC_{src} ↑	UTMOS ↑	nMOS ↑	sMOS-s ↑
GT	3.21	-	-	-	4.01	4.03 ± 0.06	-
StyleVC [61]	25.66	0.41	0.34	0.56	3.38	2.90 ± 0.38	3.02 ± 0.37
DDDM-VC [62]	9.35	0.39	0.39	0.74	3.46	3.04 ± 0.33	3.02 ± 0.32
Diff-HierVC [30]	8.40	0.40	0.41	0.78	3.67	3.02 ± 0.28	2.99 ± 0.35
Vevo-Timbre [8]	8.25	0.61	0.43	0.79	3.85	$\mathbf{3.81 \pm 0.34}$	3.67 ± 0.30
FreeVC* [23]	8.04	0.47	0.37	0.76	3.84	3.43 ± 0.23	3.28 ± 0.34
SFM-Adapter (ours)	**7.95**	**0.66**	**0.46**	**0.80**	**3.88**	3.72 ± 0.27	$\mathbf{3.69 \pm 0.32}$

identity while modifying the target style attribute. In terms of speech quality, SFM-Adapter achieves the highest UTMOS of 3.74 and an nMOS of 3.95, compared to 3.97 and 4.12 for real speech, respectively. These results suggest that our method generates speech with a relatively high degree of naturalness, while still falling slightly short of real human speech. The subjective evaluation results on emotion similarity are mostly consistent with the objective metrics, which achieve an sMOS-e of 4.09, confirming the effectiveness of SFM-Adapter.

Performance on timbre editing. We evaluate our timbre editing performance against several methods in a zero-shot setting, where both the source and target speakers are unseen during training. As shown in Table III, SFM-Adapter achieves a SECS score of 0.66 for speaker similarity, while most baseline methods fall below 0.5, reflecting a noticeable gap in performance. A similar trend is observed in the sMOS-s scores, further indicating that our method effectively enhances resemblance to the target speaker. Additionally, our method achieves a competitive WER compared to baseline methods, suggesting that it maintains strong intelligibility in timbre editing. In terms of naturalness, SFM-Adapter surpasses the baseline methods, reaching a UTMOS score of 3.88, merely 0.13 below the ground truth, indicating that our method can generate speech with reasonable naturalness. We further observe that SFM-Adapter achieves the highest Corr_{src} and a competitive ACC_{src}, indicating that it better preserves the original expressive style of the source speech during timbre editing than compared methods.

Compared with Vevo, SFM-Adapter shows a different trade-off between intelligibility and style controllability. Although SFM-Adapter yields slightly higher WER than Vevo in Tables I and II, this gap is likely related to the increased difficulty of language-guided style editing, where the model must simultaneously preserve linguistic accuracy and follow flexible style descriptions. This interpretation is further supported by the timbre editing results in Table III: without natural-language style control, SFM-Adapter achieves lower WER than Vevo-Timbre (7.95 vs. 8.25). Moreover, Vevo is an LLM-based method trained on approximately 60,000 hours of speech, whereas SFM-Adapter uses only 60 hours of training data. Despite this large difference in training scale, SFM-Adapter still achieves comparable WER while obtaining stronger style-related performance.

TABLE IV
ABLATION STUDY ON MODEL DESIGN. THE BASELINE IS THE MODEL WITHOUT USING SOURCE BRANCH FOR SOURCE MEL-SPECTROGRAM, AND ONLY USING CROSS-ATTENTION MECHANISM FOR FEATURE FUSION.

Model	WER ↓	SECS ↑	UTMOS ↑	Corr ↑	ACC ↑
baseline	8.32	0.58	3.72	0.33	0.55
SFM-Adapter (Full)	6.17	**0.64**	3.73	**0.38**	**0.70**
w/o Modulation	7.63	0.61	3.72	0.35	0.68
w/o Mask	**5.12**	0.42	3.73	0.26	0.53
w/o Cat style Fea.	7.68	0.47	**3.74**	0.31	0.62
w/o CFG	6.35	0.61	3.72	0.34	0.64
w/o Length Regulation	6.26	0.63	3.71	0.32	0.66

F. Ablation Studies and Analysis

Effect of model design. We conduct an ablation study under the multi-style editing setting to prove the effectiveness of the proposed designs. The baseline method directly injects style features using naive cross-attention. As shown in Table IV, we observe a clear performance drop across all metrics for the baseline method, suggesting that relying solely on the cross-attention mechanism may be insufficient for effective

TABLE V
ABLATION STUDY OF MASKING RATIO ACROSS DIFFERENT TASKS. SECS/CORR/ACC METRICS ARE TASK-SPECIFIC.

Masking Ratio	Multi-style editing				Expressive style editing				Timbre style editing				Normalized Average
	WER ↓	SECS ↑	Corr ↑	ACC ↑	WER ↓	SECS_{src} ↑	Corr ↑	ACC ↑	WER ↓	SECS ↑	Corr_{src} ↑	ACC_{src} ↑	
0%	5.12	0.42	0.26	0.53	4.23	0.80	0.24	0.43	5.81	0.49	0.57	0.87	0.50
15%	5.69	0.51	0.32	0.64	4.78	0.77	0.30	0.63	7.09	0.55	0.52	0.81	0.57
30%	6.17	0.64	0.38	0.70	5.30	0.72	0.35	0.73	7.95	0.66	0.46	0.80	0.65
45%	6.19	0.67	0.39	0.71	5.51	0.64	0.36	0.74	8.32	0.65	0.40	0.70	0.57
60%	6.23	0.68	0.39	0.72	5.62	0.48	0.35	0.74	8.39	0.68	0.36	0.64	0.49

TABLE VI
ROBUSTNESS ANALYSIS UNDER EXTREME SOURCE-TARGET STYLE CONFLICTS "F" INDICATES FEMALE AND "M" INDICATES MALE.

Source → Target	SECS ↑	ACC ↑	Corr ↑
happy, high-pitch F → sad, low-pitch M	0.61	0.66	0.35
sad, low-pitch M → surprise, high-pitch F	0.65	0.69	0.37
angry, high-pitch M → sad, low-pitch F	0.63	0.68	0.34

TABLE VII
SOURCE SPEECH ATTRIBUTES PRESERVATION IN EXPRESSIVE STYLE EDITING.

Methods	ACC ↑	Corr ↑	SECS_{src} ↑
baseline	0.71	0.32	0.18
SFM-Adapter	0.73	0.35	0.72

TABLE VIII
SOURCE SPEECH ATTRIBUTES PRESERVATION IN TIMBRE EDITING.

Methods	SECS ↑	Corr_{src} ↑	ACC_{src} ↑
baseline	0.63	0.12	0.21
SFM-Adapter	0.66	0.46	0.80

style editing. In the design of Stylistic Feature Manipulation Module, we compare the results "**w/o Modulation**" by removing the AdaLN, it can be observed that the style expression in the generated speech is weakened. This is evident from the reductions in SECS, Corr, and ACC compared to SFM-Adapter, highlighting that modulation plays an important role in improving style editing. Thus, combining the manipulated features with style features is the most effective approach to guide the network in style editing, allowing the generated speech to closely reflect the intended styles. In the design of the Stylistic Feature Injection Module, we ablate the effect of mask attention and the integration of style features. Compared to SFM-Adapter, removing the input style feature ("**w/o Cat style Fea.**") results in a decrease in style-related metrics, with SECS decreasing by 0.17, ACC by 0.08, and Corr by 0.07. This demonstrates that the provided style features can further reinforce the performance of style editing. Additionally, when the mask in the SFM-Adapter block is removed ("**w/o Mask**"), the WER improves, decreasing from 6.17 to 5.12. However, SECS, ACC, and Corr show noticeable declines, suggesting that without the mask, the network relies more heavily on the input mel-spectrogram during training and pays less attention to the style information provided by the input conditions. This highlights the effectiveness of the mask in guiding the network to better utilize the provided style conditions for editing. We also evaluate the impact of classifier-free guidance (CFG) and Length Regulation. Removing CFG slightly reduces SECS, Corr, and ACC, indicating it strengthens conditional guidance during sampling. Removing Length Regulation mainly degrades expressive style metrics (Corr and ACC), confirming its role in realizing the target expressive style.

Effect on mask ratio. In our SFM-Adapter, the masking ratio controls the proportion of source features dropped before being incorporated into the conversion branch. To examine its impact, we conduct an ablation study with masking ratios of 0%, 15%, 30%, 45%, and 60%. To provide an overall comparison across different masking ratios, we compute an overall score by first normalizing each metric independently across all masking-ratio settings and then averaging the normalized values over all metrics. The results in Table V show a clear trade-off controlled by the masking ratio. Increasing the masking ratio generally improves style editing performance, as reflected by higher SECS/Corr/ACC scores, but also leads to higher WER and weaker preservation of some source-related attributes. In particular, for expressive style editing, excessive masking causes a notable drop in SECS_{src}, indicating reduced preservation of source speaker information. Overall, 30% masking gives the best normalized average score, suggesting the best balance between target style control and source information preservation.

Robustness to source-style leakage. To evaluate robustness under source-target style conflicts, we construct three challenging editing settings with explicitly conflicting source and target attributes, including emotion, pitch, and gender/timbre. For each setting, we randomly sample 25 source-target pairs, where each pair consists of a source speech sample and a target style specification with strongly mismatched attributes. Table VI shows that the performance of SFM-Adapter on these extreme source-target conflict cases remains close to its average performance across all multi-style editing cases shown in Table I. Averaged over the three conflict cases, the model achieves a SECS of 0.63, an ACC of 0.68, and a Corr of 0.35, compared with 0.64, 0.70, and 0.38 in Table I, respectively. These results indicate that SFM-Adapter remains robust even under highly conflicting editing conditions.

G. Disentanglement Analysis

To further validate the disentanglement capability of the proposed method, we conduct both quantitative analysis and feature-level visualization.

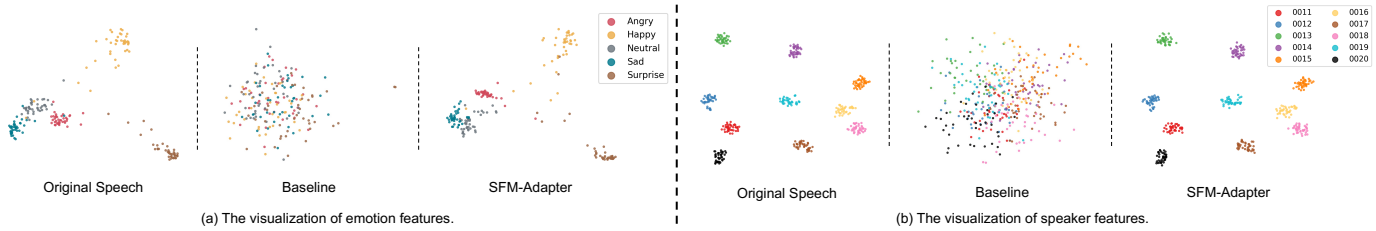


Fig. 5. Visualization of emotion and speaker features under pitch editing. Sub-figure (a) shows the clustering of emotion features, and sub-figure (b) shows the clustering of speaker features for the original speech, the baseline model, and SFM-Adapter. Compared with the baseline, SFM-Adapter preserves clearer cluster structures that remain closer to those of the original speech.

TABLE IX
COMPARISON OF MODEL SIZE AND INFERENCE TIME AMONG DIFFERENT METHODS.

Methods	Parameters (M)	Inference time (s)
Diff-HierVC [30]	18.28	1.05
FreeVC [23]	40.80	0.31
DDDM-VC [62]	79.41	0.73
StyleVC [61]	347.88	0.38
AINN [12]	39.83	0.35
Vevo [8]	819.18	2.56
SFM-Adapter (Ours)	226.46 (Adapter: 57.69)	1.54

Quantitative analysis. For each editing task, we modify only the specified attribute, while leaving the remaining attributes unspecified. We evaluate (1) editing success by comparing the edited attribute with the *target style*, and (2) attribute preservation by comparing the non-edited attributes with the *source style*. Expressive style editing is evaluated by target-style similarity and source-speaker preservation, while timbre editing is evaluated by target-speaker similarity and source-style preservation. As shown in Table VII and Table VIII, our method consistently achieves better cross-attribute preservation compared to baseline method. Specifically, during expressive style editing, our model maintains higher speaker similarity, while during timbre editing, it better preserves expressive attributes such as emotion. These results demonstrate that the proposed method can selectively modify the intended attribute while keeping unrelated attributes unchanged, whereas baseline method tend to entangle multiple attributes during editing.

Feature visualization. We further visualize the distributions of emotion and speaker features during pitch editing. As shown in Fig. 5, the features produced by baseline methods become more dispersed and exhibit mixed cluster structures after editing, indicating that non-target attributes are also affected. In contrast, our method preserves clearer cluster boundaries and maintains distributions closer to the original speech. This indicates that the proposed method can effectively modify pitch while preserving emotion and speaker identity, providing further evidence of disentangled representation.

H. Efficiency Analysis

We measure model size by the total number of parameters and evaluate computational complexity in terms of inference time. As shown in Table IX, The proposed SFM-Adapter has 226.46 M parameters in total, of which 57.69 M come from the adapter module. This is substantially smaller than large-scale models like Vevo (819.18 M) and also smaller than StyleVC (347.88 M). In terms of inference efficiency, SFM-Adapter requires only 1.54s per speech, outperforming Vevo, which takes 2.56s per speech. SFM-Adapter consistently achieves superior style editing performance and maintains comparable speech intelligibility to baseline methods. This demonstrates that the added architectural complexity is justified, as it leads to better control and flexibility in multi-conditional style editing without sacrificing efficiency.

I. Evaluation of LALM-Based Length Regulation

We conducted an ablation study to assess the impact of the LALM-based length regulation module on style editing performance. When removing this component, we observed a degradation in ACC. Without length regulation, the ACC has decreased from 0.70 to 0.66, demonstrating that speed alignment plays an important role in perceived expressiveness. Regarding robustness to phrasing variations, we evaluated the LALM's speed estimation results using text prompts, with some prompts containing explicit speed cues and others lacking them. For example, the text prompt "*A stunned female voice, fast and high-pitched.*" yields a predicted speed of 0.9, while "*Her voice was soft and her sad speech unfolded gradually.*" results in a speed of 0.2. This demonstrates that LALM can distinguish between prompts that imply speaking rate and those that do not. In addition, the overall accuracy of text-based speed prediction is 0.82, further confirming the reliability of LALM in interpreting speed from natural language descriptions

V. CONCLUSION

In this paper, we have presented SFM-Adapter, a novel framework that supports multiple style editing tasks, en- abling controllable manipulation of speech style. Our framework consists of a Source Branch for extracting source speech features and a Conversion Branch for generating speech. We introduce two modules to extract rich style descriptions and integrate stylistic features into the conversion process, respectively. During inference, Adaptive Sampling enhances editing controllability, while a Large Audio-Language Model (LALM) based Length Regulation refines speech rate and duration. Experimental results show that SFM-Adapter achieves more

natural and accurate style editing compared to existing methods, while also enabling flexible control across multi-styles. In future work, we plan to explore finer-grained modeling of prosodic attributes, such as intonation patterns, rhythm, and local speaking rate variations, to further enhance style controllability.

REFERENCES

[1] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, no. 1-2, pp. 227–256, 2003.

[2] Z. Guo, Y. Leng, Y. Wu, S. Zhao, and X. Tan, "PromptTTS: Controllable text-to-speech with text descriptions," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[3] J. Yao, Y. Yang, Y. Pan, Z. Ning, J. Ye, H. Zhou, and L. Xie, "StableVC: Style controllable zero-shot voice conversion with conditional flow matching," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[4] C. Hu, Q. Tian, T. Li, Y. Wang, Y. Wang, and H. Zhao, "Neural dubber: Dubbing for videos according to scripts," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 16 582–16 595.

[5] Q. Chen, M. Tan, Y. Qi, J. Zhou, Y. Li, and Q. Wu, "V2C: Visual voice cloning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 242–21 251.

[6] X. Niu, J. Zhang, and C. P. Martin, "HybridVC: Efficient voice style conversion with text and audio prompts," in *Interspeech*, 2024, pp. 4368–4372.

[7] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "AutoVC: Zero-shot voice style transfer with only autoencoder loss," in *International Conference on Machine Learning*, 2019, pp. 5210–5219.

[8] X. Zhang, X. Zhang, K. Peng, Z. Tang, V. Manohar, Y. Liu, J. Hwang, D. Li, Y. Wang, J. Chan *et al.*, "Vevo: Controllable zero-shot voice imitation with self-supervised disentanglement," in *International Conference on Learning Representations*, 2025.

[9] K. Zhou, B. Sisman, and H. Li, "Transforming spectrum and prosody for emotional voice conversion with non-parallel training data," in *The Speaker and Language Recognition Workshop*, 2020.

[10] K. Zhou, B. Sisman, and H. Li, "Limited data emotional voice conversion leveraging text-to-speech: Two-stage sequence-to-sequence training," in *Interspeech*, 2021, pp. 811–815.

[11] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, "Emotion intensity and its control for emotional voice conversion," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 31–48, 2022.

[12] Y. Chen, L. Yang, Q. Chen, J.-H. Lai, and X. Xie, "Attention-based interactive disentangling network for instance-level emotional voice conversion," in *Interspeech*, 2023, pp. 2068–2072.

[13] S. Ji, J. Zuo, M. Fang, Z. Jiang, F. Chen, X. Duan, B. Huai, and Z. 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*, 2024, pp. 10 301–10 305.

[14] R. Huang, R. Hu, Y. Wang, Z. Wang, X. Cheng, Z. Jiang, Z. Ye, D. Yang, L. Liu, P. Gao *et al.*, "InstructSpeech: Following speech editing instructions via large language models," in *International Conference on Machine Learning*, 2024.

[15] M. Prabhudesai, M. Wu, A. Zadeh, K. Fragkiadaki, and D. Pathak, "Diffusion beats autoregressive in data-constrained settings," in *NeurIPS 2025 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2025.

[16] J. Tian, S.-g. Lee, Z. Kong, S. Ghosh, A. Goel, C.-H. H. Yang, W. Dai, Z. Liu, H. Ye, S. Watanabe *et al.*, "UALM: Unified audio language model for understanding, generation and reasoning," *arXiv preprint arXiv:2510.12000*, 2025.

[17] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.

[18] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[19] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *IEEE International Conference on Multimedia and Expo*, 2016, pp. 1–6.

[20] S. Liu, Y. Cao, D. Wang, X. Wu, X. Liu, and H. Meng, "Any-to-many voice conversion with location-relative sequence-to-sequence modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1717–1728, 2021.

[21] W.-C. Huang, S.-W. Yang, T. Hayashi, and T. Toda, "A comparative study of self-supervised speech representation based voice conversion," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1308–1318, 2022.

[22] X. Zhang, Z. Fang, Y. Gu, H. Chen, L. Zou, J. Zhang, L. Xue, and Z. Wu, "Leveraging diverse semantic-based audio pretrained models for singing voice conversion," in *IEEE Spoken Language Technology Workshop*, 2024, pp. 758–765.

[23] J. Li, W. Tu, and L. Xiao, "FreeVC: Towards high-quality text-free one-shot voice conversion," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[24] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[25] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[26] G. Rizos, A. Baird, M. Elliott, and B. Schuller, "StarGAN for emotional speech conversion: Validated by data augmentation of end-to-end emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 3502–3506.

[27] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 920–924.

[28] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, "Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 540–552, 2019.

[29] Z. Wang, Y. Chen, L. Xie, Q. Tian, and Y. Wang, "LM-VC: Zero-shot voice conversion via speech generation based on language models," *IEEE Signal Processing Letters*, vol. 30, pp. 1157–1161, 2023.

[30] H.-Y. Choi, S.-H. Lee, and S.-W. Lee, "Diff-HierVC: Diffusion-based hierarchical voice conversion with robust pitch generation and masked prior for zero-shot speaker adaptation," in *Interspeech*, 2023, pp. 2283–2287.

[31] S.-H. Lee, H.-Y. Choi, S.-B. Kim, and S.-W. Lee, "HierSpeech++: Bridging the gap between semantic and acoustic representation of speech by hierarchical variational inference for zero-shot speech synthesis," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 10, pp. 18 422–18 436, 2025.

[32] C.-Y. Kuan, C.-A. Li, T.-Y. Hsu, T.-Y. Lin, H.-L. Chung, K.-W. Chang, S.-Y. Chang, and H.-y. Lee, "Towards general-purpose

text-instruction-guided voice conversion," in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2023, pp. 1–8.
[33] J. Yao, Y. Yang, Y. Lei, Z. Ning, Y. Hu, Y. Pan, J. Yin, H. Zhou, H. Lu, and L. Xie, "PromptVC: Flexible stylistic voice conversion in latent space driven by natural language prompts," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024, pp. 10 571–10 575.
[34] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2021.
[35] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models," *arXiv preprint arXiv:2308.06721*, 2023.
[36] A. Hertz, A. Voynov, S. Fruchter, and D. Cohen-Or, "Style aligned image generation via shared attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4775–4785.
[37] A. Gupta, L. Yu, K. Sohn, X. Gu, M. Hahn, F.-F. Li, I. Essa, L. Jiang, and J. Lezama, "Photorealistic video generation with diffusion models," in *European Conference on Computer Vision*, 2024, pp. 393–411.
[38] R. Henschel, L. Khachatryan, H. Poghosyan, D. Hayrapetyan, V. Tadevosyan, Z. Wang, S. Navasardyan, and H. Shi, "StreamingT2V: Consistent, dynamic, and extendable long video generation from text," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 2568–2577.
[39] B. Lu, Z. Chen, J. Xiao, and J.-Y. Zhu, "Input-aware sparse attention for real-time co-speech video generation," in *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, 2025.
[40] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "AudioLDM: Text-to-audio generation with latent diffusion models," in *International Conference on Machine Learning*, 2023.
[41] H. Liu, Y. Yuan, X. Liu, X. Mei, Q. Kong, Q. Tian, Y. Wang, W. Wang, Y. Wang, and M. D. Plumbley, "AudioLDM 2: Learning holistic audio generation with self-supervised pretraining," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2871–2883, 2024.
[42] J. Xue, Y. Deng, Y. Gao, and Y. Li, "Auffusion: Leveraging the power of diffusion and large language models for text-to-audio generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.
[43] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 6840–6851.
[44] T. Qi, S. Wang, C. Lu, T. Song, H. Yang, Z. Wu, and W. Zheng, "PromptEVC: Controllable emotional voice conversion with natural language prompts," in *Interspeech*, 2025, pp. 4588–4592.
[45] Y. Pan, Y. Hu, Y. Yang, J. Yao, J. Ye, H. Zhou, L. Ma, and J. Zhao, "ClapFM-EVC: High-fidelity and flexible emotional voice conversion with dual control from natural language and speech," in *Interspeech*, 2025, pp. 4583–4587.
[46] C. Churchwell, M. Morrison, and B. Pardo, "High-fidelity neural phonetic posteriorgrams," in *IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops*, 2024, pp. 823–827.
[47] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 17 022–17 033.
[48] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proceedings of the 40th International Conference on Machine Learning*, 2023.
[49] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4195–4205.
[50] A. K. Bhunia, S. Khan, H. Cholakkal, R. M. Anwer, J. Laaksonen, M. Shah, and F. S. Khan, "Person image synthesis via denoising diffusion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5968–5976.
[51] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, "Qwen-Audio: Advancing universal audio understanding via unified large-scale audio-language models," *arXiv preprint arXiv:2311.07919*, 2023.
[52] J. Bai, H. Liu, M. Wang, D. Shi, W. Wang, M. D. Plumbley, W.-S. Gan, and J. Chen, "AudioSetCaps: Enriched audio captioning dataset generation using large audio language models," in *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, 2024.
[53] J. London, *Hearing in Time: Psychological Aspects of Musical Meter*. Oxford University Press, 2012.
[54] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and ESD," *Speech Communication*, vol. 137, pp. 1–18, 2022.
[55] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," *University of Edinburgh, The Centre for Speech Technology Research (CSTR)*, 2017.
[56] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," in *Interspeech*, 2019, pp. 1526–1530.
[57] X. Li and J. Li, "AoE: Angle-optimized embeddings for semantic textual similarity," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, Aug. 2024, pp. 1825–1839.
[58] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN-based speaker verification," in *Interspeech*, 2020, pp. 3830–3834.
[59] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Interspeech*, 2018.
[60] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.
[61] I.-S. Hwang, S.-H. Lee, and S.-W. Lee, "StyleVC: Non-parallel voice conversion with adversarial style generalization," in *International Conference on Pattern Recognition*, 2022, pp. 23–30.
[62] H.-Y. Choi, S.-H. Lee, and S.-W. Lee, "DDDM-VC: Decoupled denoising diffusion models with disentangled representation and prior mixup for verified robust voice conversion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, 2024, pp. 17 862–17 870.
[63] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*, 2023, pp. 28 492–28 518.
[64] M. Chen, X. He, J. Yang, and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.

Yun Chen (Student Member, IEEE) received the master's degree in computer science from Sun Yat-sen University, Guangzhou, China, in 2024. She is currently pursuing the Ph.D. degree with the Department of Informatics, King's College London, London, U.K. She was previously with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, U.K., in 2025. Her research interests include controllable speech generation and multi-modal generation.

Haohe Liu (Graduate Student Member, IEEE) received the B.Eng. degree from Northwestern Polytechnical University, Xi'an, China, in 2020, and the Ph.D. degree from the Centre for Vision Speech and Signal Processing, University of Surrey, Guildford, U.K., in 2025. His first-author work has been published in leading journals and conferences such as IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE/ACM Transactions on Audio, Speech, and Language Processing, IEEE Journal of Selected Topics in Signal Processing, ICML, AAAI, ICASSP, and INTERSPEECH. Notable projects include AudioLDM, VoiceFixer, AudioSR, and NaturalSpeech. His research interests include audio quality enhancement, audio generation, source separation, and audio recognition. He is best known for developing AudioLDM for text-to-audio generation, which has attracted wide attention in the open-source community.

Qi Chen is currently a Researcher in ByteDance. He received the M.S. degree from Wuhan University, Wuhan, China, in 2020, and the PhD degree from the Sun Yat-sen University, Guangzhou, China, in 2024. His work has been published in leading journals and conferences such as IEEE Conference on Computer Vision and Pattern Recognition, IEEE Conference on Computer Vision, International Journal of Computer Vision, ICASSP, and INTERSPEECH. His research interests include computer vision, and multi-modal generation.

Arshdeep Singh is currently a Research Fellow in Generative Audio AI at Informatics department, King's College London (KCL), UK, affiliated with Engineering and Physical Sciences Research Council (EPSRC)-funded AI Hub in Generative Models. Previously, he was a Research Fellow in Machine Learning for Sound under the EPSRC-funded "AI for Sound" project at the Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK. He received his B.Tech degree in Electronics and Communication Engineering from Punjabi University, Patiala, in 2013, and his M.E. degree (Gold Medalist) from Panjab University, Chandigarh. He received his Ph.D. in 2022 from IIT Mandi, India. His research interests include sustainability and efficiency of AI models, machine listening, generative AI, audio privacy, and acoustic scene/event classification.

Junqi Zhao is currently working toward the Ph.D. degree at the Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, U.K. He has coauthored several papers in journals and conferences, including IEEE/ACM Transactions on Audio, Speech, and Language Processing, ICASSP, and Interspeech. His research interests include multimodal audio generation and audio understanding.

Wenwu Wang (M'02-SM'11-F'26) was born in Anhui, China. He received the B.Sc., M.E., and the Ph.D. degrees, all in the field of automation, from Harbin Engineering University, China, in 1997, 2000, and 2002, respectively. He then worked with King's College London, Cardiff University, Tao Group Ltd. (now Antix Labs Ltd.), and Creative Labs, before joining University of Surrey, U.K., in May 2007, where he is currently a Professor in Signal Processing and Machine Learning, and an Associate Head in External Engagement, School of Computer Science and Electronic Engineering, University of Surrey, UK. He is also a Core AI Fellow at the Surrey Institute for People Centred Artificial Intelligence. His current research interests include signal processing, machine learning and perception, artificial intelligence, machine audition (listening), human-AI collaboration, and statistical anomaly detection. He has (co)-authored over 400 papers in these areas. His works have been recognized with various awards, including the Meta Distinguished Faculty Award (2026), Audio Engineering Society Best Technical Paper Award (2025), IEEE Signal Processing Society Young Author Best Paper Award (2022), DCASE Judge's Award (2020, 2023, and 2024), DCASE Reproducible System Award (2019 and 2020), and LVA/ICA Best Student Paper Award (2018). He has been elected to IEEE Fellow for contributions to audio classification, generation and source separation, since 2026. He has been a keynote or plenary speaker at about 30 international conferences and workshops.

Philip Jackson (MA, Cambridge University; PhD, University of Southampton) is professor of Machine Audition at CVSSP, and principal fellow of People-Centred Artificial Intelligence, University of Surrey, UK; expert in acoustical systems, spatial reverb, flexible media, audio-visual AI (4507 citations, h-index 32: Scholar); top-rated research impact through collaborations with Bang Olufsen and BBC; ethics lead for CDT on AI for digital media inclusion; co-director Leverhulme doctoral network on AI-Enabled Digital Accessibility (ADA). Current projects include 'AI4ME', 'GRACE' and 'AURIC' partnerships, 'CoSTAR' National Lab, and 'AURORA3' audio-acoustic-AI infrastructure.

Mark D. Plumbley (S'88-M'90-SM'12-F'15) received the B.A.(Hons.) degree in electrical sciences and the Ph.D. degree in neural networks from University of Cambridge, Cambridge, U.K., in 1984 and 1991, respectively. He is a Professor of Signal Processing and Head of Department of Informatics department at King's College London, UK. His current research concerns AI, machine learning and signal processing for analysis, recognition and generation of sound. He led the first international data challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) and recently held an Engineering and Physical Sciences Research Council (EPSRC) Fellowship on "AI for Sound". He currently co-leads the EPSRC-funded Noise Network Plus, and is part of the EPSRC AI Hub in Generative Models. He is a Member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing, and a Fellow of the IET and IEEE.

T-ASL-12513-2025, Supplementary Material

Yun Chen, Haohe Liu, Qi Chen, Arshdeep Singh, Junqi Zhao, Wenwu Wang, *Fellow, IEEE*,
Philip J.B. Jackson, *Member, IEEE*, Mark D. Plumbley, *Fellow, IEEE*

A. Additional Speaker Similarity Evaluation

In addition to ECAPA-TDNN, we also use WavLM-TDNN [S1], following prior work [S2, S3], to extract speaker features for speaker similarity evaluation, thereby providing a more reliable assessment. Specifically, we extract embeddings of both the converted speech and the reference audio prompts using a frozen WavLM-TDNN model, and compute their cosine similarity to obtain the speaker similarity score. The results are shown as follows:

TABLE S.I
SPEAKER SIMILARITY COMPARISON USING WAVLM-TDNN.

Task	Method	SECS
Multi-style Editing	Vevo-Voice	0.66
	SFM-Adapter	0.67
Timbre Editing	StyleVC	0.45
	DDDM-VC	0.48
	Diff-HierVC	0.53
	Vevo-Timbre	0.67
	FreeVC	0.52
	SFM-Adapter	0.69

From the Table S.I, we can observe that SFM-Adapter achieves the best performance across both multi-style editing and timbre editing tasks. These results confirm that the improvement of SFM-Adapter is not tied to a specific speaker verification model and demonstrates the robustness of our reported speaker similarity evaluation.

B. Zero-Shot Evaluation on MEAD Dataset

To further assess the generalization of our method, we conduct additional experiments on the MEAD dataset [S4], which is entirely excluded from training. In this strict zero-shot setting, both the speakers and the linguistic content are unseen during training. We construct 500 source-target speech pairs and perform multi-style editing following the same protocol as in the main experiments.

TABLE S.II
MULTI-STYLE EDITING RESULTS ON THE MEAD DATASET.

Method	WER ↓	SECS ↑	UTMOS ↑	Corr ↑	ACC ↑
GT	-	0.61	2.54	0.35	0.90
Vevo-Voice	8.34	0.54	3.60	0.29	0.71
SFM-Adapter	8.52	0.58	3.57	0.30	0.76

From the results in Table S.II, we observe that the proposed SFM-Adapter achieves consistently better performance than the baseline in terms of speaker similarity (SECS) and style-related metrics (Corr and ACC), while maintaining comparable intelligibility (WER). The relatively low UTMOS values on MEAD are likely due to the lower recording quality of the dataset.

These results demonstrate that the proposed method generalizes well to unseen speakers and content, and is not dependent on training data overlap, alleviating concerns about potential data leakage.

C. Implementation Details of LALM-based Length Regulation

To enable the reproducibility of the proposed LALM-based length regulation module, we provide the exact inference setting and prompting details used in our experiments.

Deterministic decoding. For all experiments, the LALM is decoded with `do_sample=False`. Therefore, the inference process is fully deterministic, without any randomness introduced by temperature, top-k, or other stochastic sampling strategies.

Prompt for audio input. When the input is an audio clip, we prompt the LALM to evaluate the perceived speech tempo based on a step-by-step reasoning process. The exact prompt is given below:

```
You are a perceptive assistant trained to
evaluate the speech tempo of an audio clip.
Your goal is to reason step by step like a
human listener and assign a tempo score
between 0 (extremely slow) and 1 (extremely
fast). The higher the score, the faster the
perceived speech tempo. Do not rely solely on
raw speed metrics--consider how the speech
feels holistically, including rhythm, clarity,
 pausing, and overall delivery style. The
score must be a single number from 0 to 1,
rounded to one decimal place.

Step 1: Pausing Pattern
- Are there long silences or frequent pauses
between words or phrases?
- Or does the speaker talk continuously with
minimal interruption?

Step 2: Articulation Clarity
- Are the words clearly enunciated and easy to
 understand?
- Or are they rushed, slurred, or overly
compressed?

Step 3: Information Density & Rhythm
- Does the speaker convey a large amount of
information in a short time?
- Does the rhythm feel calm and measured, or
fast and pressured?
```

Prompt for text input. When the input is a text description, the LALM is prompted to infer the likely speech tempo implied by the text. The exact prompt is as follows:

```
You are a perceptive assistant trained to
evaluate the likely speech tempo implied by a
text prompt.
Your task is to:
1. Extract key acoustic properties implied by
the text (such as speech rate, pause
expectation, clarity, emotional tone, energy
level, etc.).
2. Analyze how these properties would
influence the pacing of the corresponding
speech.
3. Assign a tempo score between 0 and 1 based
on your analysis:
- 0 represents extremely slow delivery (very
slow, calm, deliberate speech).
- 1 represents extremely fast delivery (very
rapid, energetic, compressed speech).
- 0.5 represents a moderate pace (between slow
 and fast).
- Values in between indicate intermediate
tempo levels.
```

REFERENCES

[S1] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[S2] A. Avdeeva and A. Gusev, "Improvement of speaker similarity for zero-shot any-to-any voice conversion of whispered and regular speech," in *Interspeech*, 2024.

[S3] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar *et al.*, "Voicebox: Text-guided multilingual universal speech generation at scale," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 14 005–14 034.

[S4] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy, "MEAD: A large-scale audio-visual dataset for emotional talking-face generation," in *European Conference on Computer Vision*, 2020, pp. 700–717.