HFSD-V2C: Zero-Shot Visual Voice Cloning Via Hierarchical Face-Styled Diffusion Model

Yaping Liu[1,2], Linqin Wang[1,2], Shengxiang Gao[1,2,*], Zhengtao Yu[1,2], Ling Dong[1,2]
[1]Faculty of Information Engineering and Automation,
Kunming University of Science and Technology, Kunming, China
[2]Yunnan Key Laboratory of Artificial Intelligence, Kunming, China
20232204183@stu.kust.edu.cn, 2424172505@qq.com
gaoshengxiang.yn@foxmail.com, ztyu@hotmail.com, 46761956@qq.com

Abstract

The goal of this work is zero-shot visual voice cloning (ZS-V2C), which aims to generate speech samples with unseen speaker identity and prosody derived from a video clip and an acoustic reference. ZS-V2C presents greater challenges as: 1) unseen speaker modeling; and 2) unseen prosody modeling. Unlike previous works, we propose a novel ZS-V2C framework that incorporates a hierarchical face-styled diffusion model (HFSD-V2C). Specifically, first, we leverage cross-modal biometrics to predict unseen speaker embeddings based on facial features. Then, we jointly model the unseen prosodic features at the text, speech and video levels. Finally, a diffusion model is constructed based on the embeddings of the unseen speaker and prosodic features, enabling the generation of expressive and diverse speech. Extensive experiments on the LRS2 and GRID benchmark dataset demonstrate the superior performance of our proposed method.

Keywords: Visual voice cloning , Zero-shot , Hierarchical face-styled diffusion model

1 Introduction

Visual voice cloning (V2C) (Chen et al., 2022) aims to convert text scripts into speech that matches the target speaker's identity and emotions, and synchronizes with the lip movements in the video, based on reference audio and video. In recent years, advancements in V2C models have notably improved the alignment of speech with video and the simulation of speaker voice characteristics, driving the development of applications such as movie dubbing. However, most existing V2C models rely on data from specific speakers, limiting their generalization ability and hindering their application in open-domain scenarios.

To overcome these limitations, growing interest has emerged in zero-shot visual voice cloning (ZS-V2C), which aims to synthesize speech for previously unseen speakers without requiring their training data. Unlike traditional V2C, ZS-V2C is required to synthesize high-quality speech that accurately preserves previously unseen paralinguistic attributes (such as speaker timbre, emotional expression, and prosodic variation) conditioned on reference audio-visual inputs, which presents two main challenges: 1) unseen speaker modeling: the ability to accurately capture, generate, and reproduce the unique timbral characteristics of unseen speakers; 2) unseen emotion and prosody modeling: the model should be capable of extracting and utilizing previously unseen emotional and prosodic features. During speech generation, it must ensure that the output demonstrates natural prosodic variation. Furthermore, these variations should be consistent with the emotional tone conveyed in the scene or narrative. These challenges pose significant obstacles to existing V2C methods. Specifically, current research related to V2C primarily focuses on audio-video synchronization (Hegde et al., 2022; Lu et al., 2022; Wang and Zhao, 2022) and speaker timbre modeling (Hassid et al., 2022; Hu et al., 2021), with relatively little attention

* Corresponding Author: Shengxiang Gao. gaoshengxiang.yn@foxmail.com
©2025 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

given to deeper levels of fine-grained prosody and emotion modeling. Furthermore, in zero-shot scenarios, existing model architectures show a noticeable decline in fitting the timbre, emotional, and prosodic variations of the reference audio, making it difficult to generate high-quality speech with specific styles.

To address the challenges in the ZS-V2C task, inspired by the recent success of the face-styled diffusion model (Face-TTS) (Lee et al., 2023) applied to speech synthesis, we propose a zero-shot visual voice cloning via hierarchical face-styled diffusion model (HFSD-V2C). Specifically, HFSD-V2C has designed a cross-modal biometric unseen speaker modeling module that generates timbre features matching the target speaker using the speaker's facial images. Then, HFSD-V2C proposed a hierarchical unseen prosody modeling module based on face-styled approach, capturing both global emotional tone and local fine-grained prosody information to control the emotional expression and prosodic variation in the synthesized audio. Additionally, to ensure the naturalness and realism of the synthesized speech, HFSD-V2C introduced a conditional probabilistic diffusion model (Sohl-Dickstein et al., 2015; Ho et al., 2020; Dhariwal and Nichol, 2021) along with three loss mechanisms: speaker timbre binding loss, phoneme duration binding loss, and prosody and emotion binding loss. These improvements significantly enhance the model's generalization ability in zero-shot scenarios, allowing for more personalized and emotionally rich synthesized speech.

The contributions of this work are summarized below:

- We propose an innovative hierarchical ZS-V2C framework that jointly models the timbre, prosody, and emotional characteristics of unseen speakers. This allows the model not only to accurately reproduce the unique timbre of the speaker but also to naturally express prosodic variations and emotional information.
- We designed a face-styled diffusion model, which uses the unseen speaker's timbre, prosody, and emotion as conditional inputs for each step of the generation process, thereby improving the naturalness of the synthesized speech.
- Extensive experimental results demonstrate that HFSD-V2C outperforms other V2C models on the LRS2 and GRID datasets.

2 Related work

2.1 Text to speech

The objective of text-to-speech (TTS) is to convert text into natural, fluent, and expressive speech. Recent advancements in TTS models (Huang et al., 2023; Tan et al., 2024; Wang et al., 2017; Ren et al., 2020) have led to substantial improvements in output quality. For example, VITS (Kim et al., 2021) improves the diversity of speech by randomly modeling latent variables and introducing a duration predictor. YourTTS (Casanova et al., 2022) builds upon VITS to achieve zero-shot and multilingual support. Grad-TTS (Popov et al., 2021) introduces diffusion models, utilizing a stepwise denoising process to generate high-quality speech waveforms. Although the naturalness of synthesized speech has steadily improved, TTS models are not designed to handle video signals, which prevents them from synchronizing visual features in videos, and thus limits their application in V2C task.

2.2 Visual voice cloning

Unlike traditional TTS tasks that rely solely on text input, V2C requires processing text, audio and visual information simultaneously. In recent studies, V2C-Net (Chen et al., 2022) extracts text, audio, and video information through a multimodal encoder to generate Mel spectrograms with emotional and vocal features. Neural Dubber (Hu et al., 2021) generates speech synchronized with the video by learning the alignment relationship between the video's lip movements and the text. HPMDubbing (Cong et al., 2023) proposes a hierarchical prosody modeling framework to align visual information from video with the prosodic features of speech.

Although existing research has made significant progress in audio-visual synchronization, speaker timbre modeling, and basic prosodic feature extraction, fine-grained prosody modeling and adaptive

CCL 2025

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1020-1030, Jinan, China, August 11-14, 2025.

(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

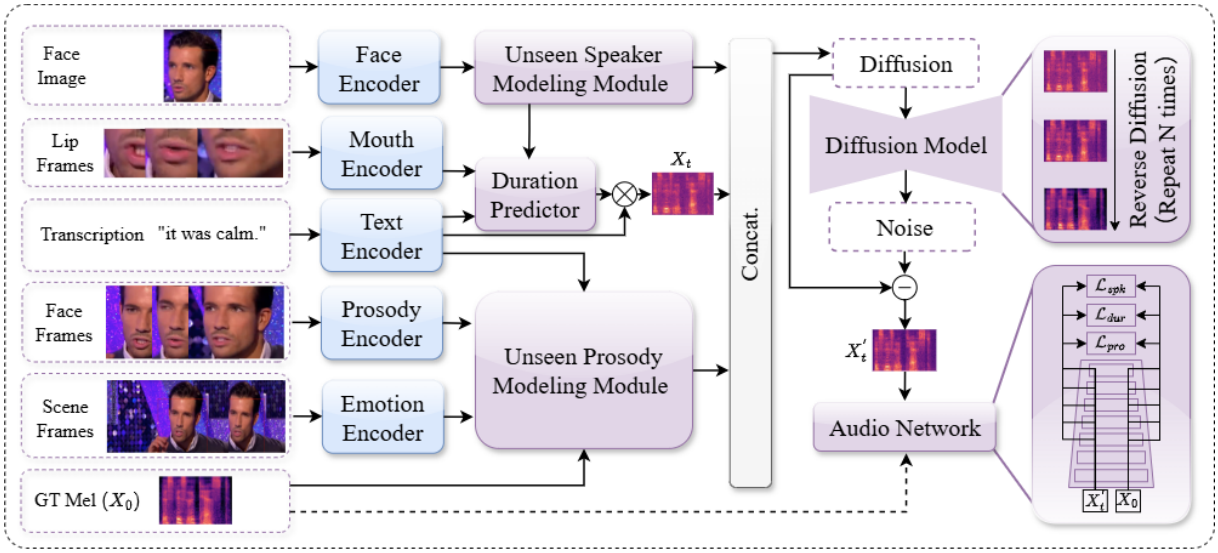

Figure 1: The main architecture of HFSD-V2C.

generation in zero-shot scenarios still present considerable challenges. Unlike these methods, HFSD-V2C combines a hierarchical face-styled diffusion model, focusing on unseen speakers and hierarchical prosody modeling, aiming to improve the adaptability and expressiveness of speech synthesis in zero-shot scenarios.

3 Method

3.1 Overview

We adopted the model based on HPMDubbing (Cong et al., 2023), which is used for the V2C task, as our baseline model and introduced HFSD-V2C on top of it. An overview of the proposed HFSD-V2C framework is depicted in Fig. 1.

3.2 Cross-Modal Biometric Unseen Speaker Modeling Module

Through facial feature extraction, unique identity information related to the timbre of the target speaker can be inferred (Goto et al., 2020; Wang et al., 2022; Nagrani et al., 2018). Speaker embedding based on facial features has been validated in Neural Dubber (Hu et al., 2021). Further inspired by Face-TTS (Lee et al., 2023), HFSD-V2C proposes a speaker modeling method that combines cross-modal biometrics and design a cross-modal biometric unseen speaker modeling module. This module significantly improves the accuracy of generating speaker embeddings from facial images by leveraging the relationship between facial features and speaker identity. Joint training is applied to the proposed module, the model is enabled to capture and reproduce the unseen speaker's timbre in zero-shot scenarios.

3.2.1 Facial feature extraction

A clear facial image I_i^f is randomly selected from the video frames $V_i^f = \{I_1^f, I_2^f, ..., I_{M_l}^f\}$ and input into the face encoder to extract the speaker's facial features. These features carry rich identity information, providing the model with foundational cues for generating the timbre.

3.2.2 Transforming visual features into timbre embedding

The extracted facial features are processed by a trainable speaker visual network $\mathcal{J}(\cdot)$ and mapped into an embedding vector E_{spk}, which is related to the speaker's vocal characteristics. This embedding vector effectively captures the speaker's timbre features:

$$E_{spk} = \mathcal{J}(I_i^f) \tag{1}$$

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1020–1030, Jinan, China, August 11–14, 2025.

(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

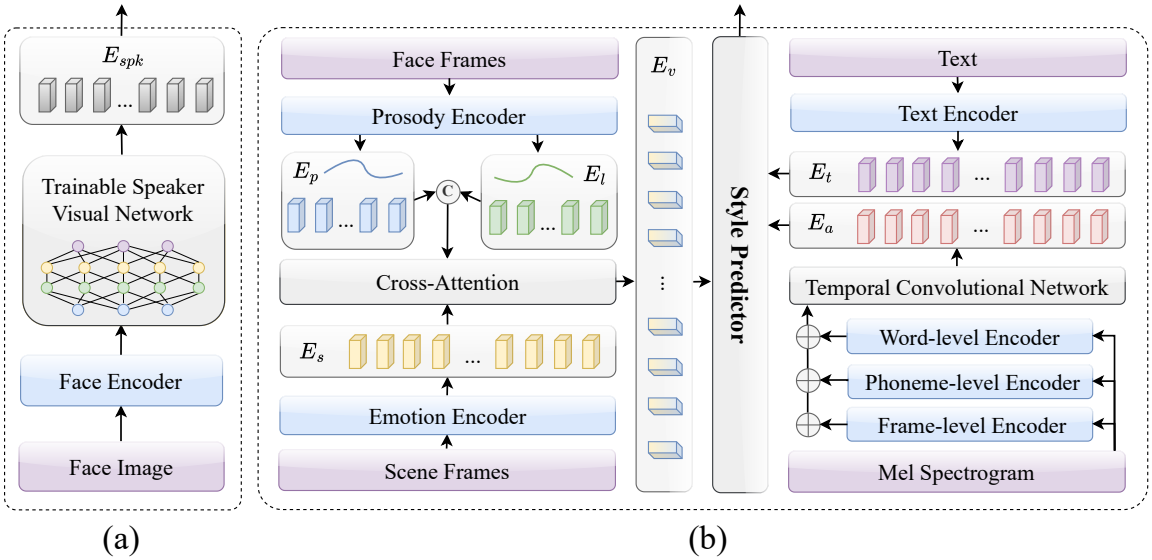

Figure 2: Schematic of module architectures. (a) cross-modal biometric unseen speaker modeling module; (b) hierarchical unseen prosody modeling module based on face-styled.

3.2.3 The application of timbre embedding

The generated speaker embedding vector is then passed into the diffusion model, facilitating the generation of speech that matches the individual in the facial image. This enables the model to establish cross-modal mappings from visual features to acoustic features. Even if the model has never encountered the speaker during training, it can generate highly personalized speech during inference based on facial images.

3.3 Hierarchical Unseen Prosody Modeling Module Based on Face-styled

Inspired by the successful application of attention mechanisms (Vaswani, 2017), hierarchical unseen prosody modeling module based on face-styled combines information from three modalities: audio, video, and text. First, fine-grained prosodic variations are extracted from both audio and video modalities, followed by the generation of a global emotional tone from the text, so as to guarantee that the prosody of the synthesized speech remains consistent with the emotional expression conveyed by the scene or narrative.

3.3.1 Video-level prosody extraction

The video layer captures dynamic prosodic features through a prosody encoder and an emotion encoder. Using the emotion face-alignment network (EmoFAN) (Toisoul et al., 2021) to extract the degree of facial pleasure for predicting speech pitch E_p and employing arousal to infer loudness E_l. These two prosodic features are then combined to form the speaker's prosodic information $E_{p,l}$.

$$E_{p,l} = (E_p; E_l) = (\sum_{k=0}^{M_l-1} \xi_{i,k} F^k; \sum_{k=0}^{M_l-1} \psi_{i,k} F^k) \tag{2}$$

$$E_v = Softmax(\frac{E_s E_{p,l}^\top}{\sqrt{D_m}}) E_{p,l} \tag{3}$$

Where i is the index of the video frame, $\xi_{i,k}$ and $\psi_{i,k}$ represent the attention weight of arousal and valence for the k-th phoneme-lip feature F^k corresponding to the i-th frame, and M_l is the desired length of the Mel spectrogram.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1020–1030, Jinan, China, August 11–14, 2025.

(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Subsequently, the emotion encoder analyzes the entire frame, rather than being limited to only the facial region, to further capture the emotional information E_s conveyed by the layout and colors of the scene, and through an attention mechanism, the prosodic and emotional features are integrated to form the video-level prosodic description E_v.

3.3.2 Audio-level prosody extraction

The extraction of prosodic features at the audio level consists of several key stages. First, the Mel spectrogram is segmented at multiple scales, with frame-level segmentation performed using a fixed time window, and fine-grained segmentation at the phoneme and word levels achieved using the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017). Then, a Temporal Convolutional Network (TCN) is employed to capture the dynamic prosodic features. The convolution at each layer is performed using causal convolution (Causal Convolution) to progressively extract higher-order temporal features, which is mathematically expressed as:

$$E_a^{(l)} = \text{ReLU}\left(\text{CausalConv1D}(E_a^{(l-1)}, W^{(l)}, d = 2^{l-1})\right) \tag{4}$$

where $E_a^{(l)}$ denotes the output features of the l-th layer, $W^{(l)}$ represents the convolutional kernel weights, d is the dilation factor, and $l \in \{1, \ldots, L\}$ denotes the layer number.

Subsequently, the output features $E_a^{(L)}$ from the final layer undergo average pooling (AvgPool), followed by a linear transformation (Linear) for feature compression, ultimately resulting in the audio feature vector. To further extract the most representative prosodic features, vector quantization (VQ) is applied to the compressed features, yielding the final prosodic feature representation E_a:

$$E_a = \text{VQ}\left(\text{Linear}(\text{AvgPool}(E_a^{(L)}))\right) \in \mathbb{R}^{N_a \times D_m} \tag{5}$$

where E_a represents the final prosodic features, N_a is the number of features, and D_m is the dimensionality of each feature.

3.3.3 Text-level prosody extraction

The text layer analyzes the input text to predict the overall emotional tone E_t. First, a text encoder encodes the text, extracting the semantic and emotional information. This information helps generate global style features, such as emotional tone and intonation, ensuring that the synthesized speech aligns with the prosodic and emotional expression of the text content.

3.3.4 Style predictor

The style predictor employs a dual-path parallel attention mechanism to achieve multimodal feature fusion. It simultaneously uses E_t as the query vector and performs cross-modal attention computations with E_a and E_v: on the one hand, it enables interaction between the text and audio modalities; on the other hand, it aligns the text and visual modalities. Finally, by summing the two output features, a comprehensive style representation is obtained.

$$E_{style} = \text{CM}_{A\to T}^{\text{mult}} \oplus \text{CM}_{V\to T}^{\text{mult}} = \text{Softmax}\left(\frac{E_t E_a^\top}{\sqrt{D_m}}\right) E_a \oplus \text{Softmax}\left(\frac{E_t E_v^\top}{\sqrt{D_m}}\right) E_v \tag{6}$$

3.4 Diffusion Model

HFSD-V2C adopts a conditional diffusion model to generate speaker-specific mel-spectrograms from multimodal inputs. The model architecture follows a score-based diffusion process, where the model progressively denoises a randomly sampled latent spectrogram under the guidance of multimodal conditions.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1020–1030, Jinan, China, August 11–14, 2025.

(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

3.4.1 Forward Process

Given a ground-truth mel-spectrogram X_0, the forward diffusion process gradually adds Gaussian noise to produce a noisy spectrogram X_t. This process follows the continuous-time formulation:

$$X_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{7}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ represents the accumulated noise schedule at timestep t. This formulation ensures that the latent variable X_t progressively incorporates noise while preserving the underlying structural characteristics of the original speech.

3.4.2 Conditional Representation via μ

Unlike conventional TTS diffusion models that condition only on text, HFSD-V2C integrates multimodal features as a unified conditional representation. Three types of embedding vectors are concatenated to construct a unified conditional vector μ:

$$\mu = \text{Concat}(E_{\text{spk}}, E_{\text{t}}, E_{\text{style}}) \tag{8}$$

The vector μ serves as the conditioning prior in the diffusion process, providing semantic guidance for speaker identity, linguistic content, rhythm, and expressive style. During denoising, the decoder takes the noisy spectrogram X_t, the timestep t, and the conditional vector μ as input to predict the clean mel-spectrogram X_0':

$$X_0' = \text{Decoder}(X_t, \mu, t) \tag{9}$$

3.5 Training Objective

To guarantee that the synthesized speech faithfully represents the intended speaker characteristics and expressive nuances, HFSD-V2C introduces several loss functions defined in the latent space of a frozen auxiliary audio network. These losses include a speaker timbre binding loss $\mathcal{L}_{\text{spk}}$; a phoneme duration alignment loss $\mathcal{L}_{\text{dur}}$; and a prosody and emotion consistency loss $\mathcal{L}_{\text{pro}}$. The overall training objective is defined as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{diff}} + \lambda_2 \mathcal{L}_{\text{spk}} + \lambda_3 \mathcal{L}_{\text{pro}} + \lambda_4 \mathcal{L}_{\text{dur}} \tag{10}$$

This framework enables fine-grained control over speaker timbre and expressive prosody, and enhances the model's generalization capability to unseen speakers and unseen prosodic styles under zero-shot conditions.

4 Experiment

4.1 Experimental Setup

To validate the effectiveness of HFSD-V2C in terms of timbre similarity and prosody diversity in zero-shot scenarios, we compared its synthesized speech quality with several mainstream visual voice cloning models, as detailed below: 1) V2C-Net(Chen et al., 2022): the first model designed for movie dubbing, capable of fine-grained video understanding and generating corresponding speech; 2) Neural Dubber(Hu et al., 2021): controls speech generation by capturing lip movements from the video to ensure audio-visual synchronization; 3) HPMDubbing (Cong et al., 2023): generates natural prosody that aligns with the movie plot, enhancing the emotional expression of the speech.

4.1.1 Dataset

The LRS2 dataset (Yu et al., 2020) comprises approximately 29 hours of audiovisual material collected from BBC broadcasts, featuring 48,165 video clips from 3,783 unique speakers. Each clip includes synchronized audio and transcriptions, with visible speaker facial regions—particularly the mouth—captured in each sentence (under 100 characters). The dataset is partitioned into training, validation, and test subsets in a 6:1:3 ratio.

CCL 2025

Table 1: Comparison of the proposed model, previous work, and ablation studies on the LRS2 dataset.

Methods	MOS ↑	MCD ↓	LES-D ↓	LES-C ↑	Id.Acc ↑	Emo.Acc ↑
GT	4.72(±0.15)	00.00	5.344	8.013	91.52	83.48
V2C-Net	3.99(±0.04)	12.61	7.784	5.026	36.84	50.41
Neural Dubber	4.14(±0.06)	9.36	6.201	6.861	59.25	58.22
HPMDubbing	4.12(±0.08)	8.66	6.136	6.608	37.75	61.46
HFSD-V2C	**4.29**(±0.08)	**7.25**	**6.012**	**7.003**	**66.87**	**65.63**
w/o US	3.20(±0.03)	12.53	11.835	3.716	30.75	58.33
w/o UP	3.32(±0.05)	10.34	8.514	5.274	57.42	22.08

The GRID corpus (Cooke et al., 2006) is a large-scale multimodal dataset containing 1,000 phonetically structured utterances per speaker, recorded across 34 individuals (18 male, 16 female), resulting in about 17 hours of high-quality audio-visual data and 34,000 recordings. It is split into training, validation, and test sets using a 7:1:2 ratio.

4.1.2 Data Preprocessing

For text data, we first convert the text sequences into phoneme sequences and use a text encoder to transform these sequences into feature representations that contain the necessary semantic information. For audio data, we convert the raw speech waveforms into Mel-spectrograms. As for video data, we sample the videos at a rate of 25 frames per second and use the S^3FD (Zhang et al., 2017) model to detect faces in the video frames. The input images for the mouth encoder are 96 × 96 pixels, focusing only on the speaker's lip region. The input images for the prosody encoder and face encoder are 224 × 224 pixels, covering the entire face. The input images for the emotion encoder are 672 × 448 pixels, capturing the scene information in the video.

4.1.3 Evaluation Metrics

We use the Mean Opinion Score (MOS) as a subjective evaluation metric to assess the perceived quality of the speech and the synchronization between audio and video. Twenty video clips are randomly selected from the test set and rated by 20 evaluators on a five-point scale. MOS was rated on a 5-point scale , where a score of 1 indicates poor quality and a score of 5 represents excellent quality. To avoid subjective bias, the evaluation is conducted under a double-blind setup, where the evaluators are unaware of whether the audio is synthesized or which model it originates from.

We use Lip Sync Error Distance (LSE-D) and Lip Sync Error Confidence (LSE-C) to evaluate the synchronization between the audio and video. Additionally, we use Mel-Cepstral Distortion (MCD) (Kubichek, 1993) to measure the spectral similarity between the generated speech and the ground truth speech.

4.1.4 Implementation Details

The training of the HFSD-V2C was conducted on an NVIDIA GTX 3090Ti GPU. We used the Adam (Kinga et al., 2015) optimizer, setting the learning rate to 0.00005 and the batch size to 16, with the model reaching convergence after 300k steps. In this work, the model's encoder consists of 4 FFT blocks, with the feature dimension set to 256. In the duration predictor, we utilized 8 attention heads to align the lip movements with the phoneme sequence. A pre-trained ResNet50 (Cao et al., 2018) was used to extract visual features. Additionally, we employed a 2D convolutional layer with a kernel size of 7 × 7, along with the EmoFAN network, composed of three convolutional blocks with a kernel size of 3 × 3 and an average pooling stride of 2 × 2, to capture facial expressions. Finally, we used the pre-trained HiFiGAN (Kong et al., 2020) to convert the generated Mel-spectrograms into speech samples.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1020–1030, Jinan, China, August 11–14, 2025.

(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Table 2: Comparison of the proposed model, previous work, and ablation studies on the GRID dataset.

Methods	MOS ↑	MCD ↓	LES-D ↓	LES-C ↑	Id.Acc ↑	Emo.Acc ↑
GT	4.71(±0.13)	00.00	5.484	8.665	90.78	81.45
V2C-Net	3.87(±0.05)	11.16	7.753	5.368	38.44	51.97
HPMDubbing	4.17(±0.08)	8.43	6.231	6.781	39.25	63.66
HFSD-V2C	**4.31**(±0.07)	**7.33**	**6.092**	**7.122**	**68.99**	**66.32**
w/o US	3.46(±0.03)	12.74	11.479	3.754	31.25	59.64
w/o UP	3.35(±0.07)	10.48	8.325	5.468	56.68	27.64

4.2 Results and Discussion

4.2.1 LRS2 benchmark dataset results

As shown in Table 1, HFSD-V2C achieves the best performance across all metrics. Specifically, in unseen speaker modeling, HFSD-V2C achieved an identity accuracy (Id.Acc) of 66.87. In unseen prosody modeling, the emotion accuracy (Emo.Acc) reached 65.63. The LSE-D and LSE-C scores are 6.012 and 7.003, respectively. Additionally, HFSD-V2C showed significant improvement in MCD. Subjective evaluation results further demonstrate that the proposed method can generate high-quality speech that closely resembles the reference audio.

4.2.2 GRID benchmark dataset results

As shown in Table 2, HFSD-V2C achieves the best performance across six evaluation metrics. Specifically, the LSE-D and LSE-C are 6.092 and 7.122, respectively. Additionally, in terms of Identity Accuracy (Id.Acc.) and Emotion Accuracy (Emo.Acc.), HFSD-V2C reaches 68.99 and 66.32, respectively.

4.2.3 Speaker Embedding Visualization

To analyze the representations of unseen speakers, we applied t-SNE (Van der Maaten and Hinton, 2008) to project their audio embeddings into a 2D space, as illustrated in Fig. 3. The resulting embeddings revealed clear speaker-specific clusters, with evident boundaries separating male and female speakers.

4.2.4 Speech Diversity

While Neural Dubber and HPMDubbing generate speech with a determined prosodic distribution (such as pitch and rhythm), HFSD-V2C introduces a sampling process during the denoising steps to accommodate variations in the generated speech. By running the HFSD-V2C model 10 times for a particular speaker, we computed a set of F0 contours for that speaker. Speech samples of ten speakers were generated from the L2S2 dataset, with their F0 contours visualized in Fig. 4. The results demonstrate that HFSD-V2C produces distinct prosody patterns, capturing each speaker's accent characteristics in different ways. This highlights that the diffusion model significantly enhances the diversity of prosody in synthesized speech, increasing its naturalness and making it sound more like human speech.

4.2.5 Ablation Study

To validate the effectiveness of the unseen speaker modeling and unseen prosody modeling modules, we conducted ablation studies by removing these modules and retraining the model. The results in Table 1 and Table 2 show that removing the unseen speaker modeling module led to the most significant drop in identity accuracy, while removing the unseen prosody modeling module caused the largest decrease in emotion accuracy. These findings indicate that each proposed module contributes significantly to the overall performance of the model, with different focuses for each.

CCL 2025

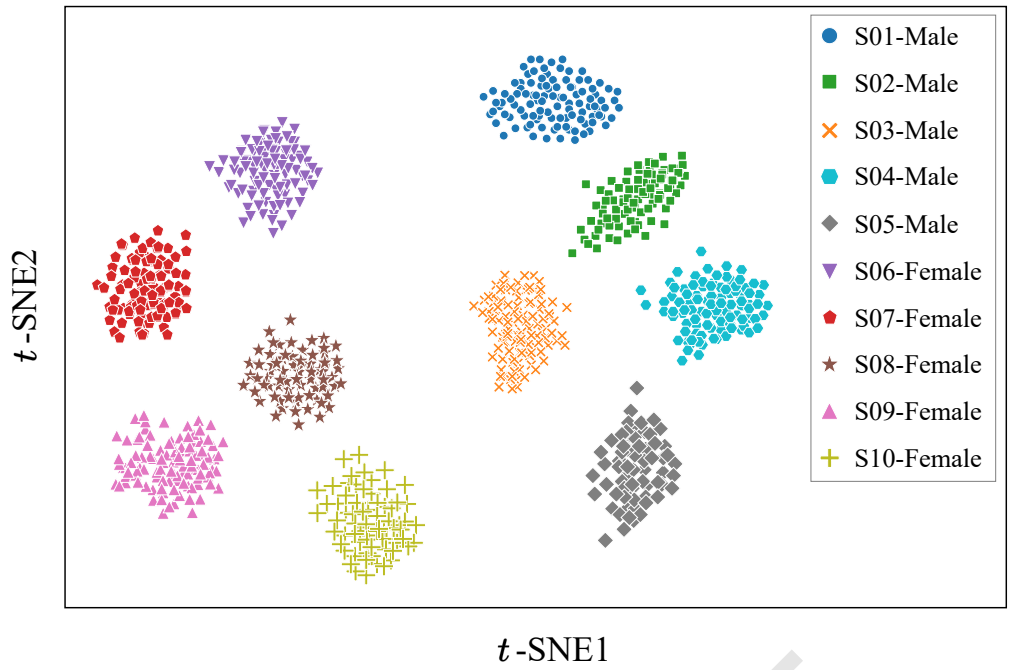

Figure 3: t-SNE visualization of utterance-level speaker vectors for the 10 unseen speakers.

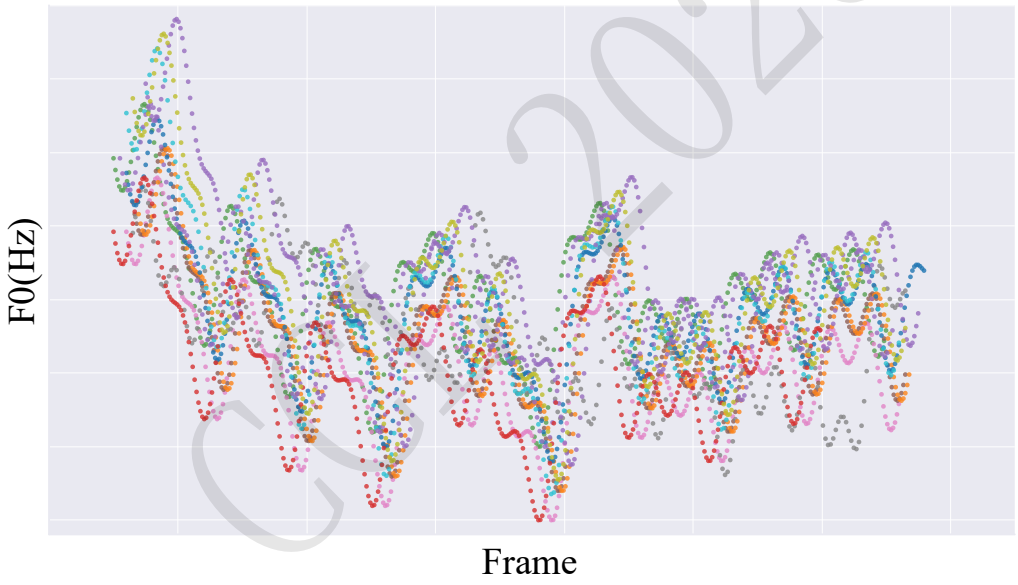

Figure 4: F0 contours and pitch tracks of speech generated by HFSD-V2C for ten speakers.

5 Conclusion

In this work, we propose HFSD-V2C, a hierarchical face-styled diffusion model for zero-shot visual voice cloning. We design a cross-modal biometric unseen speaker modeling module and a hierarchical unseen prosody modeling module based on face-styled to generate speech samples with unseen speaker identity and prosody features by integrating multi-modal information such as text, audio, and video. Additionally, we construct a diffusion model based on unseen speaker embeddings and prosodic features, which enables the generation of expressive and diverse speech. Extensive experiments on the LRS2 and GRID benchmark dataset demonstrate the superior performance of the proposed model in terms of generated speech quality.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1020–1030, Jinan, China, August 11–14, 2025.

(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (62376111, U24A20334, U23A20388, U21B2027 and 62366027), Science and Technology Planning Projects of Yunnan Province (202502AD080014, 202401BC070021, 202303AP140008 and 202302AD080003).

References

Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. 2018. Vggface2: A dataset for recognising faces across pose and age. In *Proceedings of FG Conference*, pages 67–74.

Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. 2022. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *Proceedings of ICML Conference*, pages 2709–2720.

Qi Chen, Mingkui Tan, Yuankai Qi, Jiaqiu Zhou, Yuanqing Li, and Qi Wu. 2022. V2c: Visual voice cloning. In *Proceedings of CVPR Conference*, pages 21242–21251.

Gaoxiang Cong, Liang Li, Yuankai Qi, Zheng-Jun Zha, Qi Wu, Wenyu Wang, Bin Jiang, Ming-Hsuan Yang, and Qingming Huang. 2023. Learning to dub movies via hierarchical prosody models. In *Proceedings of CVPR Conference*, pages 14687–14697.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794.

Shunsuke Goto, Kotaro Onishi, Yuki Saito, Kentaro Tachibana, and Koichiro Mori. 2020. Face2speech: Towards multi-speaker text-to-speech synthesis using an embedding vector predicted from a face image. In *Proceedings of INTERSPEECH Conference*, pages 1321–1325.

Michael Hassid, Michelle Tadmor Ramanovich, Brendan Shillingford, Miaosen Wang, Ye Jia, and Tal Remez. 2022. More than words: In-the-wild visually-driven prosody for text-to-speech. In *Proceedings of CVPR Conference*, pages 10587–10597.

Sindhu B Hegde, KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 2022. Lip-to-speech synthesis for arbitrary speakers in the wild. In *Proceedings of ACM MM Conference*, pages 6250–6258.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Chenxu Hu, Qiao Tian, Tingle Li, Wang Yuping, Yuxuan Wang, and Hang Zhao. 2021. Neural dubber: Dubbing for videos according to scripts. *Advances in neural information processing systems*, 34:16582–16595.

Rongjie Huang, Yi Ren, Ziyue Jiang, Chenye Cui, Jinglin Liu, and Zhou Zhao. 2023. Fastdiff 2: Revisiting and incorporating gans and diffusion models in high-fidelity speech synthesis. In *Proceedings of ACL Conference*, pages 6994–7009.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proceedings of ICML Conference*, pages 5530–5540.

D Kinga, Jimmy Ba Adam, et al. 2015. A method for stochastic optimization. In *Proceedings of ICLR Conference*, volume 5, page 6.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033.

Robert Kubichek. 1993. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of PACRIM Conference*, pages 125–128.

Jiyoung Lee, Joon Son Chung, and Soo-Whan Chung. 2023. Imaginary voice: Face-styled diffusion model for text-to-speech. In *Proceedings of ICASSP Conference*, pages 1–5.

Junchen Lu, Berrak Sisman, Rui Liu, Mingyang Zhang, and Haizhou Li. 2022. Visualtts: Tts with accurate lip-speech synchronization for automatic voice over. In *Proceedings of ICASSP Conference*, pages 8032–8036.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502.

Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. 2018. Learnable pins: Cross-modal embeddings for person identity. In *Proceedings of ECCV Conference*, pages 71–88.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. 2021. Grad-tts: A diffusion probabilistic model for text-to-speech. In *Proceedings of ICML Conference*, pages 8599–8608.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of ICML Conference*, pages 2256–2265.

Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, et al. 2024. Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4234–4245.

Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. 2021. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, 3(1):42–50.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Yongqi Wang and Zhou Zhao. 2022. Fastlts: Non-autoregressive end-to-end unconstrained lip-to-speech synthesis. In *Proceedings of ACM MM Conference*, pages 5678–5687.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

Jianrong Wang, Zixuan Wang, Xiaosheng Hu, Xuewei Li, Qiang Fang, and Li Liu. 2022. Residual-guided personalized speech synthesis based on face image. In *Proceedings of INTERSPEECH Conference*, pages 4743–4747.

Jianwei Yu, Shi-Xiong Zhang, Jian Wu, Shahram Ghorbani, Bo Wu, Shiyin Kang, Shansong Liu, Xunying Liu, Helen Meng, and Dong Yu. 2020. Audio-visual recognition of overlapped speech for the lrs2 dataset. In *Proceedings of ICASSP Conference*, pages 6984–6988.

Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. 2017. S3fd: Single shot scale-invariant face detector. In *Proceedings of ICCV Conference*, pages 192–201.

CCL 2025