








# Dubbing Movies via Hierarchical Phoneme Modeling and Acoustic Diffusion Denoising

Liang Li , Gaoxiang Cong , Yuankai Qi , Zheng-Jun Zha , Qi Wu, Quan Z. Sheng , *Member, IEEE*,  
Qingming Huang , *Fellow, IEEE*, and Ming-Hsuan Yang , *Fellow, IEEE*

**Abstract**—Given a piece of text, a video clip, and reference audio, the movie dubbing (also known as Visual Voice Cloning, V2C) task aims to generate speeches that clone reference voice and align well with the video in both emotion and lip movement, which is more challenging than conventional text-to-speech synthesis tasks. To align the generated speech with the inherent lip motion of the given silent video, most existing works utilize each video frame to query textual phonemes. However, such an attention operation usually leads to mumble speech because different phonemes are fused for video frames corresponding to one phoneme (video frames are finer-grained than phonemes). To address this issue, we propose a diffusion-based movie dubbing architecture, which improves pronunciation by Hierarchical Phoneme Modeling (HPM) and generates better mel-spectrogram through Acoustic Diffusion Denoising (ADD). We term our model as HD-Dubber. Specifically, our HPM bridges the visual information and corresponding speech prosody from three aspects: (1) aligning lip movement with the speech duration based on each phoneme unit by contrastive learning; (2) conveying facial expression to phoneme-level energy and pitch; and (3) injecting global emotions captured from video scenes into prosody. On the other hand, ADD exploits a denoising diffusion framework to transform the noise signal into a mel-spectrogram via a parameterized Markov chain conditioned on textual phonemes

and reference audio. ADD has two novel denoisers, the Style-adaptive Residual Denoiser (SRD) and the Phoneme-enhanced U-net Denoiser (PUD), to enhance speaker similarity and improve pronunciation quality. Extensive experimental results on the three benchmark datasets demonstrate the state-of-the-art performance of the proposed method. The source code and trained models will be made available to the public.

**Index Terms**—Visual voice cloning, speech synthesis, hierarchical phoneme modeling, contrastive learning, acoustic diffusion denoising.

## I. INTRODUCTION

**M**OVIE dubbing, also known as visual voice cloning (V2C) [1], aims to convert a paragraph of text to speech with both desired voice specified by reference audio and desired emotion and duration presented in the reference video, as shown in the top panel of Fig. 1. V2C is more challenging than conventional speech synthesis tasks (e.g., text-to-speech or voice cloning) in two aspects: First, it requires synchronization between lip motion and generated speech; Second, it requires proper prosodic variations of the generated speech to convey the speaker's emotion displayed in the video (i.e., movie's plot). V2C promises significant potential in real-world applications, such as personal speech AIGC or movie post-production.

Although significant progress has been made, existing speech synthesis methods cannot handle the challenges in V2C well. For example, text-to-speech synthesis methods [2], [3], [4], [5] construct speeches from given text conditioned on the different speaker embeddings but do not consider audio-visual synchronization. On the other hand, lip-to-speech synthesis schemes [6], [7], [8] predict mel-spectrograms directly from the sequence of lip movements. Since one lip movement may roughly correspond to different words, they suffer a high word error rate in generated speech. As for talking heads generation methods [9], [10], [11], they focus on reconstructing realistic image regions based on audio. However, these methods usually do not re-synthesize a speech to reflect targeted emotion and identity as intended in V2C. Recently, the multi-modal Large Language Models (LLMs) [12], [13], [14] have brought impressive speech synthesis effects. For instance, given a prompt, GPT-4o can synthesize expressive emotional speech. However, such models cannot serve dubbing tasks because the speaking voice is specified during pre-training, which cannot clone from user-specified reference audio.

In our prior work, FL-Dub [15], we present a mel-spectrogram-frame level dubbing architecture to address the

Received 4 July 2024; revised 23 June 2025; accepted 24 July 2025. Date of publication 8 August 2025; date of current version 3 October 2025. This work was supported in part by the National Nature Science Foundation of China under Grant 62322211, and Grant 62236008, in part by “Pioneer” and “Leading Goose” R&D Program of Zhejiang Province under Grant 2024C01023, in part by Key Laboratory of Intelligent Processing Technology for Digital Music (Zhejiang Conservatory of Music), Ministry of Culture and Tourism under Grant 2023DMKLB004. Yuankai Qi, Qi Wu, Quan Z. Sheng, Ming-Hsuan Yang are not supported by aforementioned fundings. Recommended for acceptance by L. Cao. (Corresponding authors: Gaoxiang Cong; Yuankai Qi.)

Liang Li is with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: liang.li@ict.ac.cn).

Gaoxiang Cong is with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: gaoxiang.cong@vipl.ict.ac.cn).

Yuankai Qi and Quan Z. Sheng are with the School of Computing, Macquarie University, Sydney, NSW 2113, Australia (e-mail: yuankai.qi@mq.edu.au; michael.sheng@mq.edu.au).

Zheng-Jun Zha is with the University of Science and Technology of China, Hefei 230052, China (e-mail: zhazj@ustc.edu.cn).

Qi Wu is with the Australian Institute for Machine Learning (AIML), School of Computer Science, The University of Adelaide, Adelaide, SA 5005, Australia (e-mail: qi.wu01g@adelaide.edu.au).

Qingming Huang is with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 101408, China, and also with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: qmhuang@ucas.ac.cn).

Ming-Hsuan Yang is with the EECS, University of California at Merced, Merced, CA 95344 USA (e-mail: mhyang@ucmerced.edu).

Digital Object Identifier 10.1109/TPAMI.2025.3597267

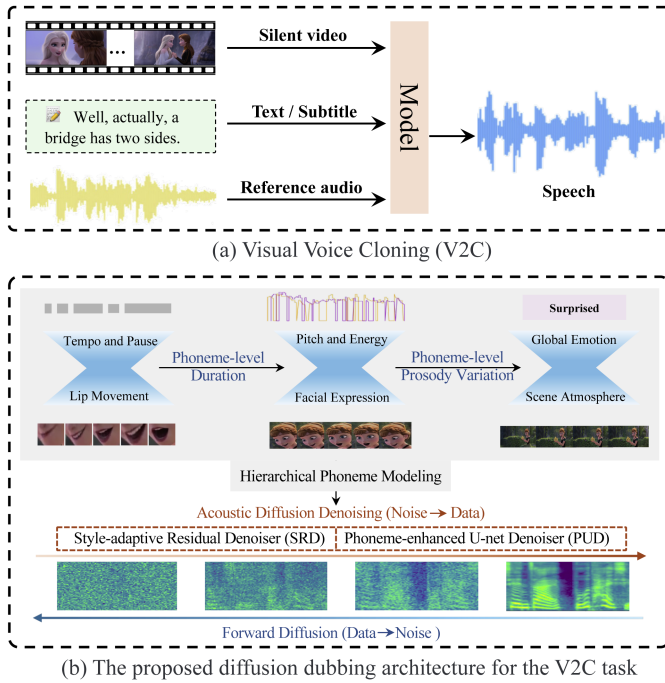


Fig. 1. (a) Illustration of the V2C task. (b) We propose a diffusion-based movie dubbing architecture consisting mainly of Hierarchical Phoneme Modeling (HPM) and Acoustic Diffusion Denoising (ADD).

issues mentioned above. Specifically, FL-Dub uses multi-head attention between the video frame sequence and input text phonemes to achieve temporal alignment, using the video frame sequence as the query. The output of multi-head attention is a video frame sequence containing attended text phoneme information. Then, an upsampling process with a fixed coefficient is used to expand the video frame sequence to the length of mel-spectrogram frame sequence, which serves as intermediate speech representations. To incorporate prosody, FL-Dub predicts pitch and energy on the mel-spectrogram frame level by associating facial expressions with intermediate speech representations. Finally, FL-Dub utilizes cross-modal attention to fuse each mel-spectrogram frame-level feature with the scene representation to absorb the global emotion. Owing to visual-acoustic alignments, the generated speech achieves large improvements over previous methods. Despite progress, it still faces two issues. First, its pronunciation is still not that good. Using video frames to query textual phonemes makes it hard to learn clear pronunciations because different phonemes are fused for video frames. Second, the predicted mel-spectrogram might be blurry and over-smoothing. FL-Dub adopts a transformer decoder for generating mel-spectrograms, which fails to handle the complex and diverse spectrum changes due to its simple objective functions (e.g., L1 or L2).

In this work, we propose a diffusion-based movie dubbing architecture, which improves FL-Dub from two aspects (see Fig. 1(b)). *First*, we propose Hierarchical Phoneme Modeling (HPM), which focuses on phoneme-level modeling while aggregating visual information. It consists of a phoneme duration aligner, a phoneme prosody adaptor, and an affine emotion booster. Specifically, our phoneme duration aligner gets rid of

traditional video-frame level alignment. Instead, it perceives the duration of each phoneme from related lip-phoneme contextual sequences by contrastive learning. Our prosody adaptor learns pitch and energy at the phoneme level, rather than the mel-spectrogram frame level, based on relevant textual phonemes and facial expressions. Our affine emotion booster is designed to introduce global emotion by converting the scene vector into bias and gain without destroying phoneme pronunciation. *Second*, to alleviate spectrum over-smoothing and improve generation quality, we propose Acoustic Diffusion Denoising (ADD). It is a parameterized Markov chain that iteratively converts the noise into mel-spectrograms conditioned on the textual phoneme from the script and style embedding from reference audio. Unlike previous works, our ADD not only improves the quality of mel-spectrograms but also strengthens style and pronunciation during the process of recovering the spectrum from noise. This is achieved by two diffusion denoisers in ADD: Style-adaptive Residual Denoiser (SRD) which learns to enhance the speaker's style similarity by affine transform in each residual block, and Phoneme-enhanced U-net Denoiser (PUD) which focuses on improving pronunciation by duration-based downsampling and phoneme-level attention mechanism. Equipped with these novel designs, the proposed method performs favorably against the state-of-the-art approaches on three widely used benchmark datasets.

The main contributions of this work are:

- We propose a diffusion-based movie dubbing architecture, an enhanced version of FL-Dub, which improves pronunciation and mel-spectrogram quality.
- We design a hierarchical phoneme modeling module, which focuses on phoneme-level modeling while aggregating visual information, including phoneme duration aligner, a phoneme prosody adaptor, and an affine emotion booster.
- We devise an acoustic diffusion denoising module, where the style-adaptive residual denoiser improves the speaker's style similarity in each residual block and phoneme-enhanced U-net denoiser strengthens phoneme level pronunciation details.
- Extensive experimental results demonstrate favorable performance of the proposed method against state-of-the-art models on three benchmark datasets.

## II. RELATED WORK

### A. Text-to-Speech Synthesis

Text-to-speech (TTS) [16], [17] has advanced rapidly and been widely applied in real-life scenarios. The classic TTS approaches [2] address speech synthesis by decoding the mel-spectrogram in parallel. In FastSpeech2s [2], a fully end-to-end model is proposed to eliminate the cascade structure (acoustic model and vocoder) and directly generate waveform from the text. To improve style adaptability, StyleSpeech [18] introduces the transformer-based TTS architecture for multi-speaker scenarios, which utilizes a learnable style encoder [19] and meta-learning. To achieve monotonic alignment between text and speech, Glow-TTS [20] designs a flow-based generative TTS

model, which combines the properties of flows and dynamic programming. Besides, SC-GlowTTS [21] and YourTTS [22] introduce speaker embeddings into the affine coupling layers of flow-based decoder blocks to enhance speaker similarity. Then, StyleTTS 2 [23] takes advantage of pre-trained large speech language models (SLMs) to achieve human-level TTS synthesis by adversarial training. Most recently, NaturalSpeech 3 [24] and MaskCGT [25] achieve state-of-the-art performance by using a factorized diffusion model and masked generative codec transformer, respectively. Despite the impressive progress in TTS, these methods lack visual processing ability and therefore cannot synchronize speech with visual frames required in V2C. In contrast, we propose HD-Dubber with Hierarchical Phoneme Modeling (HPM), which integrates visual information at three granularities (i.e., lip motion, facial expression, and scene emotion) to improve the expressiveness of speech synthesis while ensuring audio-visual synchronization.

### B. Lip-to-Speech Synthesis

Since sound and lip movements usually convey the same speech information [26], [27], several methods generate audio by lip reading [28]. Lip2Wav [29] is a sequence-to-sequence architecture focusing on learning mappings between lip and speech for individual speakers. A few methods [30], [31] focus on obtaining more robust lip movement representations through contrastive learning [32], [33], [34]. ADC-SSL [35] proposes dual-contrastive learning to improve the audio-visual synchronization from local and global embedding with multi-scale temporal convolution networks (MSTCN). To solve the slow inference in autoregressive, FastLTS [7] introduces an end-to-end unconstrained lip-to-speech synthesis system, which adopts a full transformer architecture (both spatial and temporal). Recently, some works have focused on generating speech in challenging wild environments. For example, Kim et al. [36] propose two different types of content supervision (feature-level and output-level) with connectionist temporal classification [37] and pre-train an automatic speech recognition model to correct words. However, these methods struggles to maintain clear pronunciation and accurately clone the specified style from the reference audio when applied to V2C tasks. To address these issues, we propose HD-Dubber with Acoustic Diffusion Denoising (ADD), which includes a Phoneme-enhanced U-net Denoiser (PUD) and a Style-adaptive Residual Denoiser (SRD) to improve pronunciation clarity and enhance speaker similarity.

### C. Talking Heads Generation

The talking heads task [38], [39] aims to reconstruct realistic image regions from input audio and align them with audio signals. Some works focus on synthesizing only the lip motion regions, achieved by mapping audio signals directly to lip movements. For example, DiffDub [40] utilizes a person-generic visual dubbing methodology, underpinned by the Denoising Diffusion Probabilistic Model (DDPM), to generate seamlessly blended lower facial regions (i.e., changing lip motion). On the other hand, several methods extend the construction region to encompass a wider range of facial expressions and

head movements driven by the audio input. For instance, the SadTalker [11] separates the generation targets into lips, eye blinks, and head poses. The Neural Emotion Director (NED) [9] emphasizes emotion editing by altering facial expressions while preserving the original mouth motion, enabling the attachment of a specific style to the target actor. Recently, VASA [10] has supported the online generation of  $512 \times 512$  videos at up to 40 FPS, which consists of diffusion-based holistic facial dynamics and head movement generation models. Instead of generating speeches as required by V2C tasks, these methods focus on changing the visual content of the silent video driven by audio. In contrast, we propose a diffusion-based movie dubbing architecture HD-Dubber, which improves pronunciation via hierarchical phoneme modeling (HPM) and generates better mel-spectrograms via acoustic diffusion denoising (ADD).

### D. Visual Voice Cloning

Visual Voice Cloning aims to generate vivid speech for videos/films based on the tone of reference audio [1]. Cong et al. [15] propose a hierarchical prosody dubbing model to enhance audio-visual association by bridging acoustic details with three visual granularities. To handle multi-speaker scenes, Hassidet al. [41] unify identity by normalizing all utterances of each speaker to the unit norm and using an RNN-based autoregressive decoder to generate a mel-spectrogram. In addition, some works focus on automatic video dubbing [8] to generate speech synchronized with a given video without using reference audio. For instance, Neural Dubber [42] adopts image-based speaker embedding to provide gender and age information from face regions. Face-TTS [43] uses biometric information extracted from face images as style to improve identity using a diffusion model. Furthermore, Automatic Voice Over (AVO) [44] uses a learning objective of self-supervised discrete speech unit prediction to provide more direct supervision for the alignment learning. However, these methods have advanced audio-visual aligning and voice cloning through attention-based fusion mechanisms, the generated pronunciation and audio quality are still far from satisfying. Although audio-visual speech recognition [45] and visual forced alignment [46] methods focus on visual feature extraction and word level alignment, they ignore the alignment between lip movement and phoneme-level information by contrastive learning. In this work, we mitigate pronunciation problems and improve acoustic quality by hierarchical phoneme modeling and acoustic diffusion denoising.

### E. Cross-Modal Interaction and Alignment

Visual Voice Cloning (V2C) is closely related to cross-modal interaction and alignment, which requires aligning text modality with visual modality (lip movements). Many methods have been proposed to explore cross-modal alignment. For instance, CMRAN [47] is an effective architecture to leverage both audio and visual information for accurate event localization by a relation-aware module. FiLM [48] proposes feature-wise linear modulation to improve the accuracy of visual reasoning tasks on the CLEVR benchmark. To synthesize sound from videos, REGNET [49] designs a time-dependent module to extract



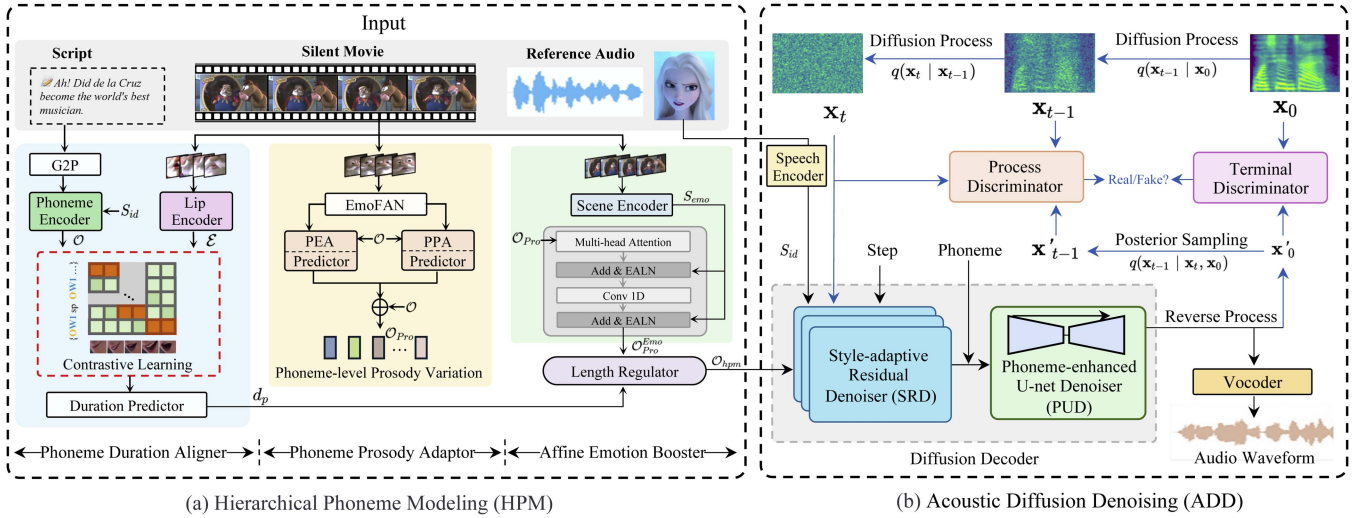


Fig. 2. The proposed HD-Dubber consists of two main components: HPM (Section III-A) and ADD (Section III-B). HPM aggregates visual representation at three granularities (i.e., lip motion, facial expression, surrounding scenes) to generate intermediate speech representations  $\mathcal{O}_{hpm}$ , containing desired duration, prosody, and emotion. ADD consists of a parameter-free diffusion process and a parameterized reverse process implemented by a diffusion decoder. The diffusion decoder contains two carefully designed denoisers (SRD and PUD) to improve style adaptation and pronunciation enhancement. Two discriminators (Section III-C) approximate the noise and mel-spectrogram distribution, respectively. Finally, we use a vocoder to obtain the speech waves from the mel-spectrogram predicted by our diffusion decoder. The black and blue arrows represent the generation and discrimination pipelines, respectively.

visual features and an audio forwarding regularizer to obtain bottlenecked real sound information. Note that “sounds” in these works refer to natural sounds (e.g., dog barking, fireworks, and drums *etc.*), which are different from the human speech of V2C. Besides, Video-LLaMA [50] is a large multimodal model to achieve video content understanding by capturing the temporal changes in visual scenes and integrating audio-visual signals. Recently, RLHMN [51] is proposed for video captioning by a hierarchical modular network and reinforcement learning. It learns multi-level visual representations at four granularities by associating linguistic counterparts: entity, verb, predicate, and sentence. In contrast, we propose a hierarchical phoneme modeling method to align video representations with phoneme-level counterparts directly at three granularities: lip movement, facial expression, and scene atmosphere.

### III. METHOD

As shown in Fig. 2, the proposed diffusion-based movie dubbing architecture (HD-Dubber) consists of two main components: Hierarchical Phoneme Modeling (HPM) and Acoustic Diffusion Denoising (ADD). Given a reference audio  $R_a$ , a raw text sequence  $T_r$ , and a video frame sequence  $V_f$ , the goal of HD-Dubber is to synthesize a piece of time-domain audio  $\hat{Y}$  for dubbing:

$$\hat{Y} = G_{\theta}(R_a, T_r, V_f), \quad (1)$$

where  $G_{\theta}$  denotes the generation pipeline of HD-Dubber (see black arrow in Fig. 2). Specifically, HPM aims to generate intermediate speech representations with duration, prosody variations, and emotional information. Then, the output of HPM is fed into the diffusion decoder, which consists of two denoisers to iteratively recover mel-spectrograms from Gaussian noise conditioned on the textual phoneme sequence and style

embedding from reference audio. There are two discriminators in ADD to improve spectrogram quality by adversarial training with multi-frequency channels and denoising distribution (see blue arrow in Fig. 2). We detail each module in the following sections.

#### A. Hierarchical Phoneme Modeling

HPM contains three modules: 1) Phoneme Duration Aligner, which predicts speech duration via contrastive learning between lip movement and textual phonemes; 2) Phoneme Prosody Adaptor, which predicts phoneme level pitch and energy variations from facial expression; and 3) Affine Emotion Booster, which introduces global emotion embedding to phoneme-level prosody variation from video scenes.

1) *Phoneme Duration Aligner*: Our phoneme duration aligner takes the script and lip motion sequence as input, and then the duration of each phoneme is predicted by lip-phoneme context sequences through contrastive learning.

*Extracting Textual Phoneme and Lip-motion Embedding*:

The open-source grapheme-to-phoneme tool<sup>1</sup> (G2P) is used to obtain the textual phoneme sequence from raw scripts. Then, the phoneme encoder [15], which is composed of stacked Feed-Forward-Transformer (FFT) blocks, is used to extract textual phoneme embeddings:

$$\mathcal{O} = \text{PhoEncoder}(T_r \in \mathbb{R}^P, S_{id}), \quad (2)$$

where  $\mathcal{O} \in \mathbb{R}^{P \times d_m}$  is the textual phoneme embeddings.  $d_m$  and  $P$  represent the hidden dimension and length of the phoneme sequence, respectively.

Compared to the phoneme encoder  $\text{PhoEncoder}(\cdot)$  [1], [15] that obtains phoneme embedding without style information, we

<sup>1</sup><https://github.com/Kyubyong/g2p>



introduce the style affine transform (SAT) to each layer normalization (LN) of PhoEncoder( $\cdot$ ) to improve the speaker identity. The principle of SAT is to perform additional scaling and shifting transformation based on speaker identity vector  $S_{id} \in \mathbb{R}^{1 \times d_m}$  of reference audio during normalizing the phoneme hidden features:

$$\text{SAT}(h, S_{id}) = \gamma(S_{id}) \cdot h_L + \delta(S_{id}), \quad (3)$$

where  $\gamma(S_{id}) = \text{FC}_1(S_{id})$  and  $\delta(S_{id}) = \text{FC}_2(S_{id})$  denote the learnable gain and bias to bring scaling and shifting for style expression by two Fully Connected (FC) layers on  $S_{id}$ , respectively.  $h_L = \frac{h - \mu}{\sigma}$  denotes the normalized feature by traditional layer normalization (LN).  $\mu$  and  $\sigma$  denote the mean and variance of vector  $h$ .

To obtain the lip-motion embedding from the input video frame sequence  $V_f$ , we adopt the same extracting pipeline as [15]:

$$\mathcal{E} = \text{LipEncoder}(M_{roi} \in \mathbb{R}^{T_v \times D_w \times D_h \times D_c}), \quad (4)$$

where  $M_{roi}$  indicates the mouth Region of Interest (ROI) frame sequence cropped by face landmarks from  $V_f$ , following [15].  $D_w$ ,  $D_h$ , and  $D_c$  indicate the number of width, height, and channels of images in the mouth ROI frame sequence.  $T_v$  denotes the total length of mouth ROI frame sequence.  $\mathcal{E} \in \mathbb{R}^{T_v \times d_m}$  denotes the output lip motion embedding from LipEncoder( $\cdot$ ).

*Attention Constraints based on Contrastive Learning:* The phoneme duration aligner aims to align lip motion and textual phonemes. To this end, we first use a multi-head attention to learn the relation between textual phonemes embedding  $\mathcal{O}$  and lip motion embedding  $\mathcal{E}$ :

$$C_{lip} = \text{sim}(\mathcal{E}, \mathcal{O})\mathcal{E}^\top = \text{softmax}\left(\frac{\mathcal{O}^\top \mathcal{E}}{\sqrt{d_m}}\right)\mathcal{E}^\top, \quad (5)$$

where  $C_{lip} \in \mathbb{R}^{P \times d_m}$  denotes the lip-phoneme context sequences.  $\text{sim}(\mathcal{E}, \mathcal{O})$  indicates the weight matrix between textual phonemes embedding  $\mathcal{O}$  and lip motion embedding  $\mathcal{E}$  by multi-head attention. Specifically,  $\mathcal{O}$  and  $\mathcal{E}$  are projected into multiple subspaces by eight attention heads. Each head computes attention independently, and their outputs are concatenated and linearly transformed to produce the final lip-phoneme context representation. The differences from FL-Dub [15] are: (a) We utilize textual phonemes embedding  $\mathcal{O}$  as Query and lip motion embedding  $\mathcal{E}$  as Key and Value to encourage the model to capture related lip movement based on the textual phoneme. This facilitates the preservation of more phoneme pronunciation. (b) We enforce alignment by constraining the attention weight matrix  $\text{sim}(\mathcal{E}, \mathcal{O})$  to meet monotonicity and surjectivity using contrastive learning:

$$\mathcal{L}_{cl} = -\log \frac{\sum \exp((P_{Attn}(\mathcal{E}, \mathcal{O}))/\tau)}{\sum \exp((\text{sim}(\mathcal{E}, \mathcal{O})))}, \quad (6)$$

where

$$P_{Attn}(\mathcal{E}, \mathcal{O}) = \text{sim}(\mathcal{E}, \mathcal{O}) \times M_{lip,pho}^{gt}. \quad (7)$$

In this formulation,  $P_{Attn}(\mathcal{E}, \mathcal{O})$  is a positive attention matrix for contrastive learning, and  $M_{lip,pho}^{gt}$  is a “0-1” matrix with  $P$  rows and  $T_v$  columns and satisfies the monotonicity and surjectivity.

We visualize  $M_{lip,pho}^{gt}$  in Fig. 4(a), where the highlighted part denotes the value “1” of  $M_{lip,pho}^{gt}$ , showing the correct correspondence. The monotonicity means that lip movements progress in a time-ordered sequence, consistently matching phoneme order without jumps or backward movements. The surjectivity means that every phoneme sequence of speech has at least one corresponding lip movement frame. We set the temperature coefficient  $\tau$  as 0.1. As such, contrastive learning loss  $\mathcal{L}_{cl}$  pulls  $\text{sim}(\mathcal{E}, \mathcal{O})$  close to positive pairs and drives away other pairs, thus encouraging surjectivity and monotonicity between phoneme units and their corresponding lip movement sequences.

*Duration Predictor:* The duration predictor is formulated as:

$$d_p = Total_{Length} \cdot \frac{\text{E}_{\text{Softplus}}(C_{lip})}{\sum_{i=1}^P \text{E}_{\text{Softplus}}(C_{lip}^i)}, \quad (8)$$

where  $C_{lip}^i$  indicates the  $i$ th lip-phoneme context sequences  $C_{lip}$ , and  $i \in [1, P]$ .  $P$  is the total length of the sequence  $C_{lip}$ .  $d_p \in \mathbb{R}^{P \times 1}$  indicates the predicted duration for all phoneme units to ensure the pronunciation boundary.  $\text{E}_{\text{Softplus}}(\cdot)$  denotes the function of duration predictor, which consists of 2-layer 1D convolutional layers, layer normalization, dropout layer, and softplus activate function [52] to predict  $d_p$ . Since the total dubbing time  $Total_{Length}$  is known, we can re-scale duration by dividing the predicted phoneme sum to achieve time consistency.  $d_p$  is optimized using MSE loss  $\mathcal{L}_{dura} = \text{MSE}(d_p, \log(g_d))$ , where  $\log(g_d)$  represents the ground truth duration in the log domain.

2) *Phoneme Prosody Adaptor:* Instead of learning mel-spectrogram frame-level prosody as in FL-Dub [15], we focus on phoneme-level prosody learning. Specifically, we bridge facial arousal with phoneme-level energy and valence with phoneme-level pitch, respectively. The facial arousal embedding  $\mathbf{A}$  and valence embedding  $\mathbf{V}$  are extracted by a pre-trained emotional face-alignment network (EmoFAN) as [15]. Then, we use phoneme level energy attention (PEA) and phoneme level pitch attention (PPA) to compute the relevance between the facial embedding ( $\mathbf{A}$  and  $\mathbf{V}$ ) and textual phoneme embedding  $\mathcal{O}$ , respectively:

$$A^l = \text{softmax}\left(\frac{\mathcal{O}^\top \mathbf{A}}{\sqrt{d_m}}\right)\mathbf{A}^\top, V^l = \text{softmax}\left(\frac{\mathcal{O}^\top \mathbf{V}}{\sqrt{d_m}}\right)\mathbf{V}^\top, \quad (9)$$

where  $A^l \in \mathbb{R}^{P \times d_m}$  and  $V^l \in \mathbb{R}^{P \times d_m}$  are the contextual arousal-phoneme and valence-phoneme, respectively.

Then,  $V^l$  and  $A^l$  are fed into two predictors consisting of several fully connected layers, Conv1D blocks, and layer normalization. The two predictors output the phoneme-level energy embedding  $E_{aro} \in \mathbb{R}^{P \times d_m}$  and pitch embedding  $P_{val} \in \mathbb{R}^{P \times d_m}$ , respectively, optimized by pitch loss  $\mathcal{L}_{pitch} = \text{MSE}(P_{val}, P_{gt})$  and energy loss  $\mathcal{L}_{energy} = \text{MSE}(E_{aro}, E_{gt})$ . The  $P_{gt}$  and  $E_{gt}$  denote the ground truth pitch and energy, respectively. Finally, we fuse the variation information (i.e., pitch and energy) into  $\mathcal{O}$  by elements adding operation:

$$\mathcal{O}_{Pro} = \mathcal{O} \oplus E_{aro} \oplus P_{val}, \quad (10)$$

where  $\mathcal{O}_{Pro} \in \mathbb{R}^{P \times d_m}$  denotes the phoneme level prosody variations by associating the facial expression from the video.

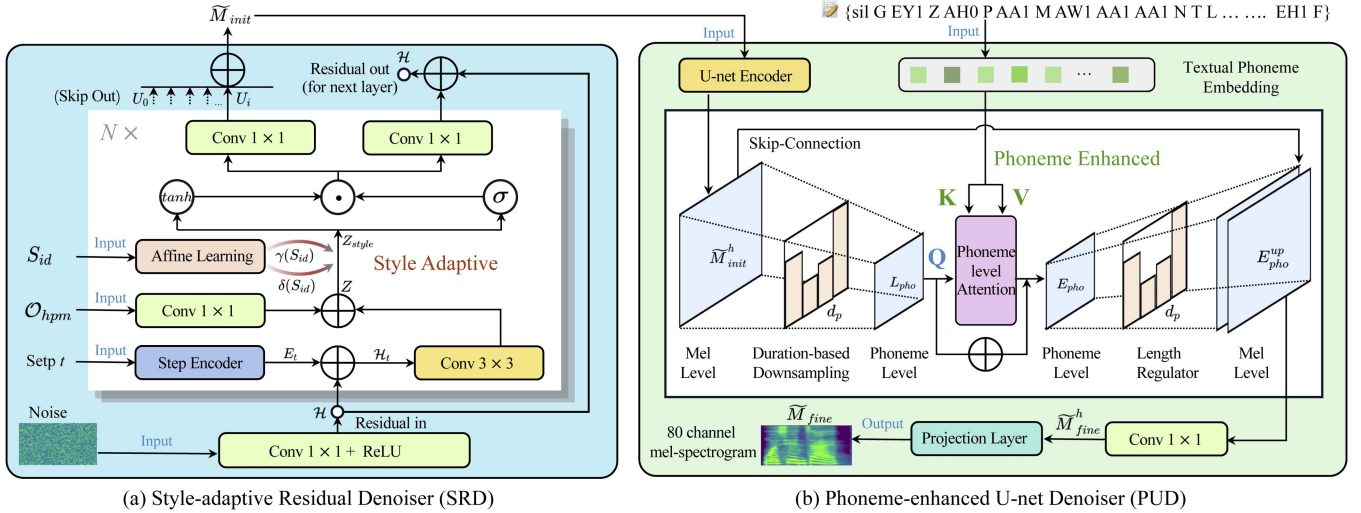


Fig. 3. Architecture of the proposed diffusion decoder in acoustic diffusion denoising (Section III-B). It consists of two main components: SRD (Section III-B1), which learns to strengthen speaker's style similarity when recovering initial mel-spectrogram from noise, and PUD (Section III-B2), which learns to improve the pronunciation by duration-based downsampling and phoneme-level attention mechanism.

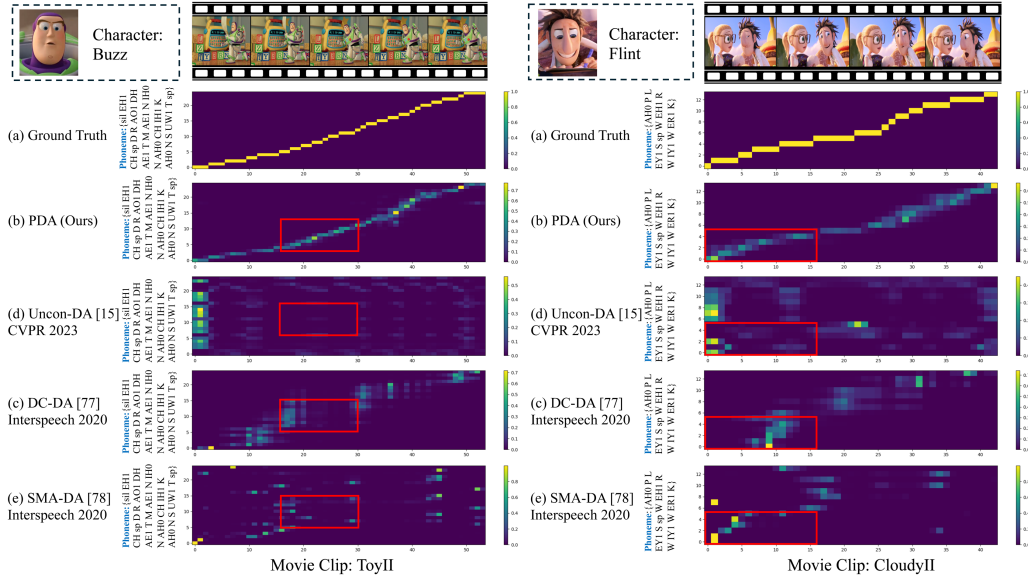


Fig. 4. Visualization of attention matrix of the alignment between lip motion and text phoneme. In each attention matrix, the horizontal axis represents the length of the lip sequence (i.e., video frame), and the vertical axis represents the length of the text phoneme. Generally, the video frame is much longer than the text phoneme length. We use red rectangles to highlight the differences between our and other methods.

3) *Affine Emotion Booster*: We further take into account the effect of movie scenes to inject global emotion into  $\mathcal{O}_{Pro}$ . To this end, we first use the pre-trained I3D model [53] to extract the emotional scene representation  $S_{emo} \in \mathbb{R}^{1 \times d_m}$  from the whole video, following [1], [15]. Then, instead of capturing the global emotion relevance by broadcasting the  $S_{emo}$  as in FL-Dub [15], we utilize the transformer encoder with Emotional Affine Layer Norm (EALN) to fuse the emotional scene representation into  $\mathcal{O}_{Pro}$ :

$$\mathcal{O}_{Pro}^{Emo} = \text{EmoEncoder}(\mathcal{O}_{Pro}, S_{emo}), \quad (11)$$

where the  $\text{EmoEncoder}(\cdot)$  indicates the transformer encoder with EALN, including multi-head attention, EALN, and 1D

convolution layer (see Fig. 2(a)). The EALN aims to replace the original layer norm to introduce learnable gain and bias based on  $S_{emo}$  to the input hidden feature, similar to (3). The  $\mathcal{O}_{Pro}^{Emo} \in \mathbb{R}^{P \times d_m}$  is the output of  $\text{EmoEncoder}(\cdot)$ , which denotes the phoneme level prosody variations with global emotion learned from the whole visual scene.

Finally, we expand phoneme level  $\mathcal{O}_{Pro}^{Emo}$  to mel-spectrogram length representation based on duration  $d_p$ :

$$\mathcal{O}_{hpm} = \text{LR}(\mathcal{O}_{Pro}^{Emo} \oplus \mathcal{O}, d_p), \quad (12)$$

where  $\text{LR}(\cdot)$  is the length regulator [2], used to repeat the corresponding duration of each phoneme.  $\mathcal{O}$  enhances the original textual phoneme representation by elements adding operation.

The whole HPM aims to generate intermediate speech representations  $\mathcal{O}_{hpm}$ . It centers around the phoneme modeling and learns the desired duration  $d_p$  using contrastive learning, the phoneme level prosody variation from facial expression, and global emotion  $\mathcal{O}_{Pro}^{Emo}$  from scene representation.

### B. Acoustic Diffusion Denoising

As shown in Fig. 2(b), the output of HPM ( $\mathcal{O}_{hpm}$ ) is fed into Acoustic Diffusion Denoising (ADD), which is a parameterized Markov chain to generate target mel-spectrograms. ADD consists of a parameter-free  $T$ -step diffusion process and a parameterized  $T$ -step reverse process.

**Diffusion Process:** The diffusion process is a Markov chain with fixed parameters [54], which gradually adds small Gaussian noises into ground truth mel-spectrogram  $\mathbf{x}_0$  until the data structure is destroyed at step  $T$ . Let  $\mathbf{x}_1, \dots, \mathbf{x}_T$  be a sequence of variables with the same dimension where  $t = 0, 1, \dots, T$  is the index for diffusion time steps. Each transition step is predefined with a variance schedule  $\beta = \{\beta_1, \dots, \beta_T\}$ . Each transformation in the diffusion process is performed according to the Markov transition probability  $q(\mathbf{x}_t | \mathbf{x}_{t-1})$  as follows:

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}). \quad (13)$$

The whole diffusion process  $q(\mathbf{x}_{1:T} | \mathbf{x}_0)$  is the Markov process and can be factorized as follows:

$$q(\mathbf{x}_1, \dots, \mathbf{x}_T | \mathbf{x}_0) = \prod_{t=1}^T q(\mathbf{x}_t | \mathbf{x}_{t-1}). \quad (14)$$

After the diffusion process, we obtain the noise mel-spectrogram  $\mathbf{x}_T$ , which is fed into the reverse process.

**Reverse Process:** The reverse process is a Markov chain with learnable parameters from noise  $\mathbf{x}_T$  to data  $\mathbf{x}_0$ . Since the exact reverse transition distribution  $q(\mathbf{x}_{t-1} | \mathbf{x}_t)$  is intractable, it is approximated by diffusion decoder:

$$p_\psi(\mathbf{x}_{t-1} | \mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \mu_\psi(\mathbf{x}_t, t), \sigma_t^2 \mathbf{I}), \quad (15)$$

where  $\mu_\psi(\mathbf{x}_t, t)$  and  $\sigma_t^2$  are the mean and variance for the diffusion decoder and  $\psi$  denotes its parameters. Thus, the whole reverse process can be formulated as:

$$p_\psi(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^T p_\psi(\mathbf{x}_{t-1} | \mathbf{x}_t). \quad (16)$$

To implement the reverse process, we propose a style-adaptive residual denoiser and a phoneme-enhanced U-net denoiser as a diffusion decoder, which learns to recover the mel-spectrogram from the noise of the diffusion process conditioned on textual phoneme from script and style embedding from reference audio.

1) **Style-Adaptive Residual Denoiser (SRD):** As shown in Fig. 3(a), SRD aims to recover the initial mel-spectrogram  $\widetilde{M}_{init}$  from noise data  $\mathbf{x}_t$  conditioned on intermediate speech representations  $\mathcal{O}_{hpm}$  (12), diffusion step  $t$ , and speaker identity vector  $S_{id}$ .

First, SRD projects noise data  $\mathbf{x}_t$  to the hidden sequence  $\mathcal{H}$  by  $1 \times 1$  convolution and Relu layer, and then we feed  $\mathcal{H}$  into the residual block. In each residual block, there are three

inputs: 1) Step  $t$  is fed to a step encoder, which consists of sinusoidal position embedding and linear layers to convert the discrete step  $t$  to continuous hidden  $E_t$ . Subsequently,  $\mathcal{H}_t$  is computed by  $E_t \oplus \mathcal{H}$ , where  $\oplus$  denotes the element-wise adding operation. 2) Intermediate speech representations  $\mathcal{O}_{hpm}$  are fed to  $1 \times 1$  convolution to obtain fused hidden sequence  $Z = \text{Conv}_{1 \times 1}(\mathcal{O}_{hpm}) \oplus \text{Conv}_{3 \times 3}(\mathcal{H}_t)$ . 3) Speaker identity vector  $S_{id}$  is introduced to affine transform to provide additional scaling and shifting:

$$Z_{style} = \gamma(S_{id}) \cdot Z_L + \delta(S_{id}), \quad (17)$$

where  $Z_L = (Z - \mu)/\sigma$  denotes the normalized feature.  $\mu$  and  $\sigma$  represent the mean and variance of  $Z$  to perform normalization.  $\gamma(S_{id})$  and bias  $\delta(S_{id})$  are learnable parameters based on speaker identity vector  $S_{id}$  to provide the gain and bias to improve style expression.  $Z_{style} \in \mathbb{R}^{L_m \times C}$  denotes the fused features with style expression, where  $L_m$  is the length of mel-spectrogram and  $C$  is the channel number. Next, the sigmoid function  $\sigma(\cdot)$  and tanh function  $\tanh(\cdot)$  are used to form a gated activation unit [55], [56] to process fused features  $Z_{style}$ . Finally, a residual block is used to split the merged hidden into two branches with  $C$  channels (the residual as the following  $\mathcal{H}$  and another as the current output  $U_i$ ). The initial mel-spectrogram is obtained by  $\widetilde{M}_{init} = \sum_{i=0}^{N-1} U_i$ , where  $N$  indicates the number of the residual blocks.

2) **Phoneme-Enhanced U-Net Denoiser (PUD):** Inspired by the success of U-net architecture in image diffusion denoising [57] and speech enhancement [58], we propose PUD to improve restoration quality further and enhance pronunciation from  $\widetilde{M}_{init}$ .

As shown in Fig. 3(b), the  $\widetilde{M}_{init}$  is first fed into the U-net encoder, which consists of convolutional layers, layer normalization, and Relu activation to obtain hidden initial sequence  $\widetilde{M}_{init}^h$ . Then, a duration-based downsampling is used to extract low-resolution features  $L_{pho} \in \mathbb{R}^{P \times C}$  from  $\widetilde{M}_{init}^h$ , which highlights the pronunciation by average pooling based on duration boundary  $d_p$ . Next, we leverage the phoneme-level attention mechanism to capture the relevance between  $L_{pho}$  and original textual phoneme sequence  $\mathcal{O}$ :

$$E_{pho} = L_{pho} \oplus \text{softmax}\left(\frac{L_{pho}^\top \mathcal{O}}{\sqrt{d_m}}\right) \mathcal{O}^\top, \quad (18)$$

where the  $L_{pho}$  is used as the query, while  $\mathcal{O}$  is used as key and value.  $\oplus$  denotes the element-wise adding.

Then, we use the length regulator  $\text{LR}(\cdot)$  [2] to expand  $E_{pho}$  to mel-spectrogram length:

$$E_{pho}^{up} = \text{LR}(E_{pho}, d_p), \quad (19)$$

where  $E_{pho}^{up}$  indicates the phoneme enhanced feature after up-sampling. Next, the Skip-Connection connecting the “parallel” layers of the Decoder is used to recover more detailed information and improve the accuracy of the final result:

$$\widetilde{M}_{fine}^h = \text{Conv}_{1 \times 1}\left(\text{SkipConnect}\left(\left[\widetilde{M}_{init}^h, E_{pho}^{up}\right]\right)\right), \quad (20)$$

where the initial mel-spectrogram  $\widetilde{M}_{init}^h$  and phoneme enhanced feature  $E_{pho}^{up}$  are concatenated along the channel dimension.



$\widetilde{M}_{fine}^h$  denotes the refined the mel-spectrogram. The role of  $\text{Conv}_{1 \times 1}(\cdot)$  is to reduce the channel dimension to 256. Finally, we use the  $1 \times 1$  convolutional layer to project  $\widetilde{M}_{fine}^h$  to 80-dimensions  $\widetilde{M}_{fine}$  for converting time-domain wave  $\hat{Y}$  by pre-trained vocoder.

The predicted mel-spectrogram  $\widetilde{M}_{fine}$  is optimized by mel-reconstruction loss  $\mathcal{L}_{rec}$  and structural similarity index loss [59], [60]  $\mathcal{L}_{SSIM}$ :

$$\mathcal{L}_{rec} = \left( \mathbb{E} \left[ \left\| \mathbf{x}_0 - \widetilde{M}_{fine} \right\|_1 \right] \right), \quad (21)$$

$$\mathcal{L}_{SSIM} = \left( 1 - \text{SSIM} \left( \mathbf{x}_0 - \widetilde{M}_{fine} \right) \right), \quad (22)$$

where the  $\mathbf{x}_0$  represents the ground-truth mel-spectrogram. The mel-reconstruction loss  $\mathcal{L}_{rec}$  is based on L1 differences. The SSIM loss  $\mathcal{L}_{SSIM}$  is used to measure structural information and texture from mel-spectrograms.

### C. Full-Stage Discriminator

Inspired by the diffusion GAN model [55], [61], we introduce a full-stage discriminator, which contains two sub-discriminators, a process discriminator and a terminal discriminator, to approximate the denoising distribution.

1) *Process Discriminator*: It aims to reduce the number of denoising steps by using predicted noise distribution  $p_\psi(\mathbf{x}_{t-1}|\mathbf{x}_t)$  (15) to approximate the true denoising distribution  $q(\mathbf{x}_t|\mathbf{x}_{t-1})$  (13). Formally, we use an adversarial loss that minimizes a divergence  $D_{adv}$  per denoising step, following [61]:

$$\min_{\theta} \sum_{t \geq 1} \mathbb{E}_{q(\mathbf{x}_t)} [D_{adv}(q(\mathbf{x}_t|\mathbf{x}_{t-1})|p_\psi(\mathbf{x}_{t-1}|\mathbf{x}_t))], \quad (23)$$

where we adopt the Least Squares Generative Adversarial Network (LS-GAN) training formulation [62] to minimize  $D_{adv}$  thanks to its various successful practices in audio generation domain [18]. Specifically, to set up the adversarial training, we denote the process discriminator as  $D_{Pro}(\mathbf{x}_{t-1}, \mathbf{x}_t, t) : \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R} \rightarrow [0, 1]$ . It takes the  $N$ -dimensional  $\mathbf{x}_{t-1}$ ,  $\mathbf{x}_t$ , and time step  $t$  as inputs, and decides whether  $\mathbf{x}_{t-1}$  is a plausible denoised version of  $\mathbf{x}_t$  on  $t$  step. The discriminator is trained by:

$$\begin{aligned} \mathcal{L}_{Pro}^D = & \sum_{t \geq 1} \mathbb{E}_{q(\mathbf{x}_t)q(\mathbf{x}_{t-1}|\mathbf{x}_t)} \left[ (D_{Pro}(\mathbf{x}_t, \mathbf{x}_{t-1}, t) - 1)^2 \right] \\ & + \mathbb{E}_{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} \left[ D_{Pro}(\mathbf{x}_t, Pos(\widetilde{M}_{fine}), t)^2 \right], \end{aligned} \quad (24)$$

where fake samples from  $Pos(\widetilde{M}_{fine})$  are contrasted against real samples from  $\mathbf{x}_t$ . The  $Pos(\cdot)$  represents the posterior sampling distribution  $q(\mathbf{x}'_{t-1}|\mathbf{x}'_0, \mathbf{x}_t)$ , where  $\mathbf{x}'_0$  indicates the predicted 80-dimensions mel-spectrogram  $\widetilde{M}_{fine}$  and  $\mathbf{x}_t$  is the noise distribution at step  $t$  from (13).

2) *Terminal Discriminator*: Although the process discriminator is beneficial for acceleration, the quality of the mel-spectrogram receives less attention in adversarial training. Observing that different frequency bands of the mel-spectrogram contain different details, we propose a terminal discriminator that adopts multi-band discrimination to focus on denoising

quality in the final step output of the reverse process. Specifically, we define the terminal discriminator as  $D_{Ter}(\widetilde{M}_{fine}) : \mathbb{R}^N \rightarrow [0, 1]$ . It takes the  $N$ -dimensional  $\widetilde{M}_{fine}$  as input and evaluates whether it has the same mel-spectrogram details with original mel-spectrogram  $\mathbf{x}_0$  by multi-band discrimination:

$$D_{Ter}(\widetilde{M}_{fine}) = \sum_{j=0} \text{Mul}_{k \times k} \left( \left[ \widetilde{M}_{fine} \right]_{S_w * j}^{S_w * (j+1)} \right), \quad (25)$$

where  $S_w$  indicates the sub-band bandwidth in the channel dimension, and  $j = [0, 1, \dots, \frac{80}{S_w} \in \mathbb{N}^+]$ .  $\text{Mul}_{k \times k}(\cdot)$  denotes the convolution layers with different kernel. The convolution kernel size  $k$  used in the top sub-band is smaller to preserve high-frequency details, and the  $k$  used in the bottom sub-band is larger to capture more contextual information of low-frequency, respectively. The terminal discriminator is trained by:

$$\mathcal{L}_{Ter}^D = \mathbb{E}_{(\mathbf{x}_0)} \left[ (D_{Ter}(\mathbf{x}_0) - 1)^2 + (D_{Ter}(\widetilde{M}_{fine}))^2 \right], \quad (26)$$

it aims to distinguish the real samples from fake ones generated by the final step output of the reverse process. The real sample is the ground truth mel-spectrogram  $\mathbf{x}_0$ , and the fake sample is the predicted mel-spectrogram  $\widetilde{M}_{fine}$ .

Finally, the generation loss of HD-Dubber is defined as the sum of the adversarial loss for each discriminator:

$$\begin{aligned} \mathcal{L}_{adv} = & \mathbb{E} \left[ (D_{Pro}(Pos(G_\theta(R_a, T_r, V_f)), \mathbf{x}_t, t) - 1)^2 \right] \\ & + \mathbb{E} \left[ D_{Ter}(G_\theta(R_a, T_r, V_f) - 1)^2 \right], \end{aligned} \quad (27)$$

where the adversarial loss follows LSGAN [62] to replace the binary cross-entropy terms of the original GAN [63] objective with least squares loss functions. Besides, we adopt the feature matching loss  $\mathcal{L}_{fm}$  to evaluate the difference in intermediate feature maps of each discriminator sub-module between a ground truth sample and a generated sample by summing L1 distances:

$$\mathcal{L}_{fm} = \mathbb{E}_{q(\mathbf{x}_t)} \left[ \sum_{m=1}^N (||D_{Pro}^m(\cdot)||_1, ||D_{Ter}^m(\cdot)||_1) \right], \quad (28)$$

where  $N$  represents the total number of hidden layers in the discriminator. The  $||D_{Pro}^m(\cdot)||_1$  and  $||D_{Ter}^m(\cdot)||_1$  denote the L1 distances of output hidden layer  $m$  of the process discriminator and terminal discriminator, respectively, where  $m \in [1, N]$ .

### D. Training Loss

Overall, we conduct the adversarial training for HD-Dubber by alternating between the updates of the generation part (including HPM and ADD's diffusion decoder, see components with a black arrow in Fig. 2) that minimizes  $L_G$  loss, and the updates of the full-stage discriminator (see components with blue arrow in Fig. 2) that minimizes  $L_D$  loss:

$$\begin{aligned} \mathcal{L}_G = & \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{SSIM} + \lambda_3 \mathcal{L}_{fm} + \lambda_4 \mathcal{L}_{adv} \\ & + \lambda_5 \mathcal{L}_{pitch} + \lambda_6 \mathcal{L}_{energy} + \lambda_7 \mathcal{L}_{cl} + \lambda_8 \mathcal{L}_{dura}, \end{aligned} \quad (29)$$

$$\mathcal{L}_D = \epsilon_1 \mathcal{L}_{Pro}^D + \epsilon_2 \mathcal{L}_{Ter}^D, \quad (30)$$

where the loss weights are empirically set as  $\lambda_1 = 6.5$ ,  $\lambda_2 = 5.5$ ,  $\lambda_3 = 0.2$ ,  $\lambda_4 = 0.45$ ,  $\lambda_5 = 0.1$ ,  $\lambda_6 = 0.1$ ,  $\lambda_7 = 0.01$ ,  $\lambda_8 = 5.0$ ,  $\epsilon_1 = 1.0$ , and  $\epsilon_2 = 1.0$ , respectively.

#### IV. EXPERIMENTS AND RESULTS

##### A. Datasets

We conduct experiments on three dubbing datasets, including Disney cartoon character dubbing, real-person dubbing in a recording studio, and YouTube real-person dubbing. Below, we describe the characteristics of each dataset.

**V2C-Animation:** This dataset [1] is currently the only publicly available dataset for multi-speaker movie dubbing. Specifically, it contains 153 diverse characters extracted from 26 Disney cartoon movies, specified with speaker identity and emotion labels. The dataset has 9,296 video clips, and the audio samples, which contain background music and environment sound, are collected at 48 kHz with 32 bits. The split of training and test sets follows FL-Dub [15].

**GRID:** This dataset [65] is a popular benchmark for multi-speaker dubbing. The whole dataset has 33 speakers (originally 34, but one is corrupted), each with 1,000 short English samples. All participants are recorded in a noise-free studio with a unified screen background. All audio is sampled at 25 000 Hz. GRID does not have speaker emotion labels or various movie backgrounds compared to V2C-Animation. The train set consists of 29 694 samples, nearly 900 sentences from each speaker. In the test set, there are 100 samples of each speaker.

**Chemistry Lecture (Chem):** This dataset [42] is a single-speaker dataset of a chemistry teacher speaking in the class. It comprises 6,640 short video clips collected from YouTube, with a total video length of approximately nine hours. The Chem dataset is originally used for the unconstrained single-speaker lip-to-speech synthesis [29]. For fluency and complete dubbing, each video clip has sentence-level text and audio based on the start and end timestamps extracted by NeuralDubber [42]. There are 6,132 and 196 dubbing clips for training and testing, respectively.

##### B. Evaluation Metrics

**Objective Metrics:** We adopt the Mel Cepstral Distortion Dynamic Time Warping (MCD-DTW) metric following [1], [66] to measure the difference between the generated dubbing and the ground truth. To further assess the duration consistency between the generated dubbing and the video, we utilize the MCD-DTW-SL metric, which adjusts the weights based on duration consistency [1]. Furthermore, to assess the pronunciation quality of the generated dubbing, we utilize the state-of-the-art automatic speech recognition (ASR) model Whisper [67] from OpenAI for dubbing recognition and computing the word error rate (WER) [68] against the script to evaluate the accuracy of the generated dubbing. To evaluate the timbre consistency between the generated dubbing and the reference audio, we employ the speaker encoder cosine similarity (SECS) following [21], [22] to compute the similarity of speaker identity. In addition, we utilize a speech emotion recognition model [69] to evaluate the

emotion accuracy (EMO-ACC) of the generated dubbing (For the V2C-Animation benchmark only because there is no emotion label in the other datasets).

**Subjective Metrics:** For subjective evaluation, we conduct human evaluations of mean opinion score (MOS) in aspects of naturalness (NMOS) and similarity (SMOS). Both metrics are rated on a 1-to-5 scale and reported with the 95% confidence intervals (CI). Following the settings in [1], [15], participants are asked to assess the dubbing quality of 30 randomly selected audio samples from each test set.

##### C. Implementation Details

**Data Preprocessing:** The video frames are sampled at 25 FPS, and all audios is resampled to 22 050 Hz. Note that for the V2C-Animation dataset, audio compression from multi-channel to mono-channel is required. We use the Montreal Forced Aligner (MFA) [70] to extract the ground truth of phoneme duration. The window length, frame size, and hop length in short-time Fourier transform (STFT) are 1,024, 1,024, and 256, respectively. The max wave value and mel-spectrogram channel are 32 768 and 80, respectively. For energy extraction, we compute the mean L2-norm of the amplitude of each STFT frame within a phoneme duration [2]. The face image is extracted by  $S^3FD$  [71] face detection model, following [15]. The lip region is resized to a  $96 \times 96$  grayscale image, and then a pre-trained temporal convolutional network is used to extract lip embedding with 512-dimension [15], [72], [73].

**Model Architecture:** The phoneme encoder and lip encoder in the hierarchical phoneme modeling module follow Neural Dubber [42], which consists of feed-forward transformers (FFT) blocks to encode phoneme hidden sequences. Note that we use Conv1D  $1 \times 1$  layer to replace the original Position Embedding (PE) in the phoneme encoder and lip encoder to model the relative position information [74]. The one-dimensional convolution's hidden size, number of attention heads, kernel size, and filter size in the FFT block are set as 256, 8, 9, and 1,024, respectively. The pitch predictor and energy predictor have the same network structure and hyper-parameters as in FL-Dub.<sup>2</sup> Diffusion step  $t$  is encoded using the same sinusoidal positional encoding as in [75]. We use VPSDE for the diffusion process of each variable with  $\beta_{\min} = 0.1$  and  $\beta_{\max} = 40.0$ , following [55]. In addition, the diffusion time step is set to 4. The mel-spectrogram feature maps are added with the diffusion step embedding and fed into 20 residual blocks with a hidden dimension of 256. The gated mechanism is used to process the feature maps before the style adaption layer. The phoneme level attention head in PUD is 1, with a drop rate of 0.2. The skip connection is to reconnect  $\widetilde{M}_{init}^h$  and  $E_{pho}^{up}$  along the channel dimension. The process discriminator has the same network structure as joint conditional and unconditional (JCU) network [55], which consists of the 1D convolutional layers with the 64, 128, 512, 128, and 1 channel and 3, 5, 5, 5, and 3 kernel sizes. The sub-band bandwidth  $S_w$  in the terminal

<sup>2</sup>The source code of our prior work is available at <https://github.com/GalaxyCong/HPMDubbing>. All the source code of this paper will be released.

TABLE I  
EXPERIMENTAL SETTINGS FOR V2C

Setting	Explanation	Num.
Setting 1 (Original setting [1])	Ground-truth speaker in test set	2,779
Setting 2 (Reference speaker setting)	Same-speaker from other movie clips	2,626
Setting 3 (Unseen speaker setting)	Unseen speaker for V2C	4,851

discriminator is 20, and the initial convolution kernel sizes are 5, 3, 3, and 2, respectively. The spectral normalization is applied to each sub-discriminator in the terminal discriminator to stabilize training.

*Training Details:* For training, we use the Adam [76] optimizer with  $\beta_1 = 0.5$ ,  $\beta_2 = 0.9$ ,  $\epsilon = 10^{-9}$  to optimize HPM and ADD. The initial learning rate of the Adam optimizer is 0.0001. We use the AdamW [77] optimizer with  $\beta_1 = 0.5$ ,  $\beta_2 = 0.9$ ,  $\epsilon = 10^{-9}$  to optimize the full-stage discriminator. The initial learning rate of the AdamW optimizer is 0.0002. We set the batch size to 32, 64, and 16 on V2C-Animation, GRID, and Chem datasets, respectively. In this work, the vocoder's parameters are frozen, which is used to transform the generated mel-spectrograms into audio samples. Both training and inference are implemented with PyTorch on a GeForce RTX 4090 GPU.

#### D. Performance Comparison

To evaluate the performance of our HD-Dubber, we compare it with different kinds of state-of-the-art speech synthesis methods:

- 1) *FL-Dub (CVPR'23) [15]*: It is a mel-spectrogram frame-level movie dubbing architecture, which models the relevance between three visual representations (i.e., lips, facial expressions, and scenes) with their counterparts.
  - 2) *Face-TTS (ICASSP'23) [43]*: It is a zero-shot speech synthesis model with face-styled diffusion. Face-TTS introduces a score-based diffusion model and cross-modal biometrics to preserve speaker identity between face images and generated speech.
  - 3) *V2C-Net (CVPR'22) [11]*: This movie dubbing model for the V2C task fuses speaker style from reference audio, text information from the raw script, and emotion from video by element-wise addition.
  - 4) *CDFSE (Interspeech'22) [64]*: It is a zero-shot speaker adaptation method for speech synthesis, which learns fine-grained style features by extracting local content embeddings and speaker embeddings from a reference speech.
  - 5) *Meta-StyleSpeech (ICML'21) [18]*: It is a multi-speaker speech synthesis method to clone the style in unseen scenes, which learns speaker style by a mel-style encoder under the meta-learning episode with adversarial training.
  - 6) *Fastspeech2 (ICLR'21) [2]*: It is a transformer-based, non-autoregressive speech synthesis method to improve speech quality by explicitly modeling energy, pitch, and duration as variation information.
- 1) *Experimental Setup:* We evaluate our method in three experimental settings as shown in Table I: 1) Setting 1 is same as in [1], which uses ground-truth audio as reference audio; 2) Setting 2 uses non-ground truth audio of the same speaker as

reference audio, simulating real-world applications; 3) Setting 3 uses the audio of unseen characters (from another dataset) as reference audio simulating real-world applications.

2) *Results on the V2C-Animation: Setting 1 on V2C-Animation:* The results are presented in Table II left. The proposed HD-Dubber achieves the best performance on all metrics. Specifically, in terms of pronunciation quality, our method achieves 28.45% on WER, which significantly surpasses the current SOTA method (FL-Dub) and is much closer to human performance. In terms of MCD-DTW and MCD-DTW-SL, the proposed method achieves relative 7.21% and 25.34% improvements over SOTA baseline CDFSE, respectively. Considering these methods use the same vocoder to generate audio, the above improvement demonstrates the proposed HD-Dubber can achieve better mel-spectrogram reconstruction quality and duration consistency. Additionally, the proposed method achieves 80.03% SECS, which significantly outperforms the previous methods regarding speaker similarity. These experimental results demonstrate the proposed method can better capture and convey the speaker's timbre, which is crucial for movie dubbing.

*Setting 2 on V2C-Animation:* We report the results of setting 2 on the V2C-Animation benchmark in Table II right. Note that setting 2 is more challenging than setting 1, which requires the model to be robust. The proposed method achieves outstanding performance on all metrics. For instance, its SECS and WER are far ahead of the SOTA method FL-Dub, and do not degrade obviously as other methods, showing the effectiveness of the proposed HD-Dubber. In terms of metrics MCD-DTW and MCD-DTW-SL, although all the performance degrades compared to setting 1, our method still achieves the best performance. Last but not least, the human subjective evaluation results (see MOS-N and MOS-S in Table V) also show that the proposed method can generate better speeches according to naturalness and timbre similarity.

3) *Results on the GRID: Setting 1 on GRID:* As shown in Table III left, the proposed model achieves the best performance across all metrics on the real-person dubbing dataset GRID. In terms of MCD-DTW, the proposed method achieves the lowest mel-cepstral distortion, demonstrating that it can generate high-quality mel-spectrograms closer to the ground truth. Regarding speaker similarity (see SECS), the proposed method outperforms other models with an absolute margin of 2.7% over the SOTA baseline Meta-StyleSpeech. Last but not least, our method achieves the lowest WER, producing clearer pronunciation in real-person dubbing.

*Setting 2 on GRID:* As shown in Table III right, the proposed HD-Dubber performs best on almost all metrics under setting 2 on the GRID benchmark. Our method performs slightly worse regarding MCD-DTW than the SOTA TTS method CDFSE. Regarding speaker similarity (see SECS), our method outperforms other models with an absolute improvement of 3.43% over the second-best method CDFSE. In addition, the proposed method improves by 7.2% in terms of WER compared with the second-best method, CDFSE. Furthermore, the lowest MCD-DTW-SL demonstrates that our HD-Dubber achieves better duration sync than other methods. Finally, as shown in Table V, our method



TABLE II  
RESULTS ON THE V2C-ANIMATION BENCHMARK

Methods	Visual	Setting 1 (V2C-Animation benchmark)					Setting 2 (V2C-Animation benchmark)				
		SECS(%) $\uparrow$	WER(%) $\downarrow$	MCD-DTW $\downarrow$	MCD-DTW-SL $\downarrow$	EMO-ACC(%) $\uparrow$	SECS(%) $\uparrow$	WER(%) $\downarrow$	MCD-DTW $\downarrow$	MCD-DTW-SL $\downarrow$	EMO-ACC(%) $\uparrow$
GT	-	100	25.55	0	0	99.96	100	22.55	0	0	99.96
GT Mel+Vocoder	-	96.96	24.4	3.77	3.8	97.09	96.96	24.4	3.77	3.8	97.09
FastSpeech2 [2]	$\times$	24.87	34.48	11.2	14.48	42.21	24.17	35.08	11.2	14.48	42.21
Meta-StyleSpeech [18]	$\times$	54.99	106.73	11.5	15.1	44.12	75.66	76.58	11.56	15.1	41.55
CDFSE [64]	$\times$	48.98	68.81	9.98	12.51	42.75	47.79	58.82	10.68	13.52	39.11
FastSpeech2* [2]	$\checkmark$	25.47	33.53	11.35	14.73	42.39	25.47	34.08	11.35	14.73	42.39
Meta-StyleSpeech* [18]	$\checkmark$	42.53	108	11.62	14.23	42.53	75.67	82.48	11.58	14.73	42.57
CDFSE* [64]	$\checkmark$	48.93	68.05	10.03	12.01	43.97	47.55	58.81	10.76	13.66	39.3
V2C-Net [1]	$\checkmark$	40.61	73.08	14.12	18.49	43.08	34.07	61.61	14.58	18.73	41.01
FL-Dub [15]	$\checkmark$	53.76	164.16	11.12	11.22	46.61	31.42	171.03	11.88		43.97
Face-TTS [43]	$\checkmark$	52.81	201.13	13.44	26.94	44.04	51.98	200.18	13.78	28.03	43.56
HD-Dubber (Ours)	$\checkmark$	80.03	28.45	9.26	9.34	46.63	77.92	28.49	10.28	10.37	42.02

The method with “\*” refers to a variant taking video embedding as an additional input following [1]. For the setting 1, we use the ground truth audio as reference audio, and for the setting 2, we use the non-ground truth audio from the same speaker within the dataset as the reference audio, which is more aligned with practical usage in dubbing. The same setup is applied to the GRID and Chem benchmarks.

TABLE III  
RESULTS ON THE GRID BENCHMARK WITH COMPARISONS AGAINST STATE-OF-THE-ART METHODS UNDER THE SETTING 1 AND SETTING 2

Methods	Visual	Setting 1 (GRID benchmark)				Setting 2 (GRID benchmark)			
		SECS(%) $\uparrow$	WER(%) $\downarrow$	MCD-DTW $\downarrow$	MCD-DTW-SL $\downarrow$	SECS(%) $\uparrow$	WER(%) $\downarrow$	MCD-DTW $\downarrow$	MCD-DTW-SL $\downarrow$
GT	-	100	22.41	0	0	100	22.41	0	0
GT Mel+Vocoder	-	97.57	21.41	4.1	4.15	97.57	21.41	4.1	4.15
FastSpeech2 [2]	$\times$	47.41	19.05	7.67	8.43	47.41	19.05	7.67	8.43
Meta-StyleSpeech [18]	$\times$	91.06	24.83	5.87	5.98	74.15	21.42	7.02	7.95
CDFSE [64]	$\times$	86.54	19.13	5.71	5.99	82.25	19.35	6.21	6.76
FastSpeech2* [2]	$\checkmark$	25.47	19.61	11.35	14.73	59.58	19.61	7.24	7.95
Meta-StyleSpeech* [18]	$\checkmark$	90.04	22.62	5.74	5.88	59.58	19.82	7.01	7.82
CDFSE* [64]	$\checkmark$	85.93	20.05	5.75	6.4	81.34	21.05	6.27	7.29
V2C-Net [1]	$\checkmark$	80.98	47.82	6.79	7.23	71.51	49.09	7.29	7.86
FL-Dub [15]	$\checkmark$	85.11	45.51	6.49	6.78	71.99	44.15	6.79	7.09
Face-TTS [43]	$\checkmark$	82.97	44.37	7.44	8.16	34.14	39.05	7.77	8.59
HD-Dubber (Ours)	$\checkmark$	93.76	17.87	5.66	5.73	85.68	17.95	6.37	6.47

Note that the GRID benchmark is a multi-speaker real-person dataset without emotion annotation.

TABLE IV  
RESULTS ON THE CHEM BENCHMARK WITH COMPARISONS AGAINST STATE-OF-THE-ART METHODS UNDER THE SETTING 1 AND SETTING 2

Methods	Visual	Setting 1 (Chem benchmark)				Setting 2 (Chem benchmark)			
		SECS(%) $\uparrow$	WER(%) $\downarrow$	MCD-DTW $\downarrow$	MCD-DTW-SL $\downarrow$	SECS(%) $\uparrow$	WER(%) $\downarrow$	MCD-DTW $\downarrow$	MCD-DTW-SL $\downarrow$
GT	-	100	4.23	0	0	100	4.23	0	0
GT Mel+Vocoder	-	98.45	3.23	2.21	2.22	98.45	3.23	2.21	2.22
FastSpeech2 [2]	$\times$	79.92	15.56	6.04	8.24	73.79	15.56	6.04	8.24
Meta-StyleSpeech [18]	$\times$	66.9	79.86	6.93	11.36	57.31	82.81	7.52	11.86
CDFSE [64]	$\times$	83.64	15.57	5.81	7.97	75.02	17.61	6.08	8.41
FastSpeech2* [2]	$\checkmark$	80.97	16.61	6.07	8.14	75.11	16.62	6.07	8.14
Meta-StyleSpeech* [18]	$\checkmark$	64.16	81.31	7.51	11.83	58.35	89.95	8.16	12.03
CDFSE* [64]	$\checkmark$	84.15	15.49	5.77	7.96	74.61	20.60	6.09	8.14
V2C-Net [1]	$\checkmark$	76.19	39.72	6.25	8.31	67.02	41.34	6.74	9.04
FL-Dub [15]	$\checkmark$	85.09	37.05	6.12	7.25	74.51	38.94	6.68	7.93
Face-TTS [43]	$\checkmark$	62.20	260.1	7.51	12.59	61.54	216.76	7.64	12.79
HD-Dubber (Ours)	$\checkmark$	91.25	14.23	5.53	6.99	86.27	14.80	6.03	7.60

Note that the Chem benchmark is a single-speaker real-person dataset without emotion annotation.

TABLE V  
SUBJECTIVE EVALUATION ON V2C-ANIMATION AND GRID BENCHMARKS

Dataset	V2C-Animation		GRID	
	NMOS $\uparrow$	SMOS $\uparrow$	NMOS $\uparrow$	SMOS $\uparrow$
GT	4.76 $\pm$ 0.09	-	4.69 $\pm$ 0.07	-
GT Mel + Vocoder	4.63 $\pm$ 0.09	4.65 $\pm$ 0.07	4.66 $\pm$ 0.08	4.53 $\pm$ 0.10
FastSpeech2 [2]	3.75 $\pm$ 0.12	2.13 $\pm$ 0.09	3.37 $\pm$ 0.14	3.09 $\pm$ 0.11
Meta-StyleSpeech [18]	3.34 $\pm$ 0.13	3.37 $\pm$ 0.14	3.56 $\pm$ 0.14	3.60 $\pm$ 0.19
CDFSE [64]	3.72 $\pm$ 0.15	3.58 $\pm$ 0.11	3.57 $\pm$ 0.12	3.54 $\pm$ 0.13
FastSpeech2* [2]	3.77 $\pm$ 0.08	2.46 $\pm$ 0.06	3.31 $\pm$ 0.12	3.04 $\pm$ 0.17
Meta-StyleSpeech* [18]	3.31 $\pm$ 0.21	3.35 $\pm$ 0.12	3.50 $\pm$ 0.10	3.58 $\pm$ 0.11
CDFSE* [64]	3.69 $\pm$ 0.09	3.68 $\pm$ 0.14	3.89 $\pm$ 0.14	3.95 $\pm$ 0.11
V2C-Net [1]	2.78 $\pm$ 0.06	3.04 $\pm$ 0.15	3.62 $\pm$ 0.06	3.67 $\pm$ 0.11
FL-Dub [15]	3.06 $\pm$ 0.22	3.19 $\pm$ 0.10	3.77 $\pm$ 0.20	3.74 $\pm$ 0.13
Face-TTS [43]	3.09 $\pm$ 0.06	3.13 $\pm$ 0.12	3.39 $\pm$ 0.21	3.32 $\pm$ 0.17
HD-Dubber (Ours)	3.89 $\pm$ 0.14	3.85 $\pm$ 0.11	4.01 $\pm$ 0.09	4.03 $\pm$ 0.12

achieves the best MOS-N and MOS-S, demonstrating the proposed method can generate speech with high naturalness and identity similarity.

4) *Results on the Chem: Setting 1 on Chem:* As shown in Table IV left, the proposed method achieves the best

performance across all metrics on the Chem benchmark. Unlike V2C-Animation and GRID, samples in Chem are recordings of only one chemistry teacher, which does not present exaggerated prosody variation or diverse speaking styles. Therefore, the pronunciation accuracy of all compared methods is generally better than that of the V2C-Animation and the GRID benchmark. Nevertheless, the proposed model achieves a lower WER than the SOTA method (CDFSE [64]), rendering clearer pronunciation. The lowest MCD-DTW and MCD-DTW-SL demonstrate the ability of our model to generate dubbing closest to ground truth. Furthermore, our model achieves the lowest MCD-DTW-SL, indicating that it can generate speech that is well synchronized with the video.

*Setting 2 on Chem:* We report the setting 2 results on the Chem dataset in Table IV right. Despite setting 2 being much more challenging than setting 1, the proposed method performs best on all metrics. According to the SECS metric, our method significantly improves the performance by 15.78% and 28.72%

TABLE VI  
RESULTS ON ZERO-SHOT TEST (SETTING 3)

Methods	SECS (%) $\uparrow$	WER (%) $\downarrow$	NMOS $\uparrow$	SMOS $\uparrow$
FastSpeech2 [2]	21.11	14.05	$3.02 \pm 0.09$	$2.91 \pm 0.13$
Meta-StyleSpeech [18]	55.81	77.46	$3.22 \pm 0.15$	$3.17 \pm 0.06$
CDFSE [64]	57.23	19.83	$3.35 \pm 0.07$	$3.53 \pm 0.12$
FastSpeech2* [2]	26.79	18.38	$3.03 \pm 0.29$	$2.97 \pm 0.12$
Meta-StyleSpeech* [18]	58.71	89.11	$3.22 \pm 0.10$	$3.31 \pm 0.18$
CDFSE* [64]	61.12	19.25	$3.31 \pm 0.13$	$3.62 \pm 0.09$
V2C-Net [1]	39.43	112.71	$2.83 \pm 0.09$	$3.05 \pm 0.07$
FL-Dub [15]	49.31	106.81	$2.92 \pm 0.09$	$3.11 \pm 0.08$
Face-TTS [43]	33.80	201.98	$3.17 \pm 0.15$	$3.10 \pm 0.05$
HD-Dubber (Ours)	70.59	13.35	$3.84 \pm 0.14$	$3.87 \pm 0.11$

than FL-Dub (CVPR'23) and V2C-Net (CVPR'22), respectively. The proposed method achieves the lowest MCD-DTW, showing minimal acoustic difference even in challenging setting 2. Last but not least, our method achieves the lowest WER, which shows that the proposed HD-Dubber can produce the most accurate pronunciation in real-person dubbing.

### E. Zero-Shot Generalization

For setting 3, we use the speakers from the GRID benchmark as unseen reference audio to evaluate the model's generalizability trained on the V2C-Animation benchmark, following [78]. Since there is no ground-truth audio under this setting, we only report performance on metrics SECS and WER and provide subjective evaluations.

As shown in Table VI, the proposed method achieves the best generation quality on all four metrics. The higher SECS and MOS-S (mean opinion score of similarity) show that our method generates better speeches for unseen speakers. Compared to previous methods (e.g., FL-Dub [15]), the proposed model maintains good pronunciation (see WER). The superior performance of WER and SECS in such challenging scenarios demonstrates the robustness of the proposed HD-Dubber. The proposed HD-Dubber performs favorably in the most challenging zero-shot scenarios.

### F. Ablation Study

1) *Ablation Study of Main Modules*: In this section, we study the effectiveness of the main module: hierarchical phoneme modeling (HPM), diffusion denoiser in Acoustic Diffusion Denoising (ADD), and adding adversarial training. Experiments are carried out on the V2C-Animation benchmark under setting 1. The results are shown in Table VII. It shows all modules of HD-Dubber achieve consistent improvement on all metrics. Comparing Rows 1, 2, and 3 of Table VII, the HPM module contributes the most. Especially in terms of WER, the HPM achieves a performance of 34.11%, which far exceeds the 96.38% of ADD's Diffusion Decoder. This demonstrates that learning on the phoneme level (including duration, prosody, and emotion) improves the consistency of speaker identity while ensuring pronunciation quality. The results on Row 2 and Row 3 of Table VII show that ADD's Diffusion Decoder improves MCD-DTW more significantly than Adversarial Training, which indicates the importance of diffusion-based denoiser for high-quality mel-spectrogram reconstruction. Finally, compared with Row 4 in Table VII, the MCD-DTW in Row 5 shows further improvement,

which is brought by the designed discriminator structure that focuses on the spectral details of multiple frequency bands.

2) *Ablation Study Within HPM*: To verify the effectiveness of each component in HPM, we conduct detailed ablation studies, and the results are shown in Table VIII. Specifically, we conduct the following ablation studies: (1) To evaluate the effectiveness of the style encoder, we remove the style affine transform of each layer of FFT in the phoneme encoder. The results are presented in Row 1 of Table VIII. We find that SECS decreases to 75.93%, indicating that introducing style information in the phoneme encoding stage enhances the similarity of speakers. (2) To verify the effectiveness of the phoneme duration aligner in hierarchical modeling, we remove the contrastive learning and the input of lip motion, forcing the duration prediction to be controlled by the textual phoneme. As shown in Row 2, MCD-DTW-SL significantly worsens, indicating a large audio-visual mismatch. (3) We remove the phoneme prosody adaptor so that the model cannot predict phoneme-level energy and pitch from facial expressions. Row 3 shows a certain degradation in MCD-DTW, demonstrating the effectiveness of incorporating phoneme level prosody variations (i.e., energy and pitch). (4) Finally, we remove the affine emotion booster to verify its impact on emotions. As shown in Row 4, EMO-ACC incurs a significant drop, which indicates that the synthesized audio deviates from the real emotional class without our booster.

3) *Ablation Study Within ADD*: To verify the effectiveness of each component in ADD, we conduct the detailed ablation studies as shown in Table IX: (1) We use DiffGAN-TTS [55] as the baseline, equipped with a basic denoiser conditioned on speaker-ID by broadcasting and includes a process discriminator. The results in Row 1 of Table IX show that ordinary diffusion denoisers are not enough for challenging movie dubbing. (2) We replace the decoder in DiffGAN-TTS [55] with the proposed SRD and keep the remaining modules unchanged. As shown in Row 2, all metrics have been improved, among which SECS has the largest improvement. This result reflects the effectiveness of our Style-adaptive Residual Denoiser (SRD) in diffusion modeling. (3) We add the Phoneme-enhanced U-net Denoiser (PUD) to evaluate its effectiveness. Except for SECS, the model performs better (see row 3), especially regarding WER. Thanks to the unique skip-connection architecture of U-net, the mel-spectrogram quality of SRD output is further enhanced (see MCD-DTW), but SECS has no improvement because PUD does not involve style adaptation.

### G. Evaluation of Duration Aligner

To evaluate the effectiveness of the proposed contrastive learning-based duration aligner, we compare it with several counterparts in the literature: (1) The unconstrained Duration Aligner (Uncon-DA) [15]: no constraints are added between the aligned lips and the textual phonemes in the multi-head attention mechanism. (2) The Diagonal Constraint-based Duration Aligner (DC-DA) [79] is used to solve the text-to-speech alignments in multi-speaker scenarios, which constrains the attention weight matrix of text and lips through diagonal loss with a fixed bandwidth. (3) The Stepwise Monotonic Attention Duration Aligner (SMA-DA) [80], which meets the criteria

TABLE VII  
ABLATION STUDY OF THE MAIN MODULES ON THE V2C-ANIMATION BENCHMARK UNDER THE SETTING 1

Methods		Settings			V2C-Animation benchmark				
#		HPM	ADD's Diffusion Decoder	Adversarial Training	SECS (%) ↑	WER (%) ↓	EMO-ACC (%) ↑	MCD-DTW ↓	MCD-DTW-SL ↓
FL-Dub [15] (CVPR 2023)					53.76	164.16	46.61	11.12	11.22
1			✓		72.25	96.38	46.62	10.63	10.82
2				✓	56.13	159.26	46.61	11.05	11.18
3		✓			75.86	34.11	46.53	9.39	9.47
4		✓	✓		79.65	29.16	46.52	9.28	9.37
5		✓	✓	✓	<b>80.03</b>	<b>28.45</b>	<b>46.63</b>	<b>9.26</b>	<b>9.34</b>

TABLE VIII  
ABLATION STUDY OF THE HIERARCHICAL PHONEME MODELING (HPM) ON THE V2C-ANIMATION BENCHMARK UNDER THE SETTING 1

		HPM Settings				V2C-Animation benchmark				
#	Methods	Identity	Lip motion	Facial Expression	Scene Emotion	SECS (%) ↑	WER (%) ↓	EMO-ACC (%) ↑	MCD-DTW ↓	MCD-DTW-SL ↓
1	w/o Style-PE		✓	✓	✓	75.93	29.55	45.83	9.37	9.46
2	w/o Duration Aligner	✓		✓	✓	79.89	33.34	45.26	9.32	12.16
3	w/o Prosody Adaptor	✓	✓		✓	79.74	28.57	45.51	9.56	9.64
4	w/o Scene Atmosphere Booster	✓	✓	✓		77.22	29.40	42.18	9.31	9.40
5	Full-model	✓	✓	✓	✓	<b>80.03</b>	<b>28.45</b>	<b>46.63</b>	<b>9.26</b>	<b>9.34</b>

TABLE IX  
ABLATION STUDY OF THE ACOUSTIC DIFFUSION DENOISING (ADD) ON THE V2C-ANIMATION BENCHMARK UNDER THE SETTING 1

		ADD Settings			V2C-Animation benchmark			
#	Methods	Process	Style Identity	Textual Phoneme	SECS (%) ↑	WER (%) ↓	MCD-DTW ↓	MCD-DTW-SL ↓
1	DiffGAN-TTS (2 stages) [55]	✓			49.56	31.91	9.87	12.59
2	SRD (Ours)	✓	✓		79.85	32.34	9.39	9.48
3	SRD + PUD (Ours)	✓	✓	✓	79.88	28.47	9.31	9.39

TABLE X  
ABLATION STUDY OF THE DURATION ALIGNER ON THE V2C-ANIMATION BENCHMARK UNDER THE SETTING 1

Methods	SECS (%) ↑	WER (%) ↓	MCD-DTW ↓	MCD-DTW-SL ↓
Uncon-DA [15]	78.70	34.57	9.38	9.49
DC-DA [79]	79.15	33.31	9.40	9.57
SMA-DA [80]	79.03	34.32	9.94	10.04
PDA (Ours)	<b>80.03</b>	<b>28.45</b>	<b>9.26</b>	<b>9.34</b>

of locality and monotonicity. It helps reduce misalignment and improve the robustness of the speech synthesis system. (4) This paper proposes Phoneme Duration Aligner (PDA), which constrains the relevance between lip motion and text sequences by contrastive learning.

The quantitative results are shown in Table X. It shows that the proposed PDA achieves the best MCD-DTW-SL, demonstrating better audio-visual synchronization when capturing related lip motion spans than others. In addition, our method achieves lower WER than other methods in Table X, which shows that the proposed PDA can provide clear pronunciation. In contrast, although the MCD-DTW-SL of the frame-level alignment in Uncon-DA is relatively small, its WER is high. We also visualize the attention matrix obtained by our method and other speech alignment methods in Fig. 4. We highlight the regions where significant differences are observed (see red rectangle). Our method achieves attention closest to the ground truth: the curve shows monotonicity and trend consistency. In contrast, all other methods produce obvious disconnections. That is, the other approaches fail to capture the correspondence well.

TABLE XI  
ABLATION STUDY ON DIFFERENT HYPERPARAMETERS OF  $\mathcal{L}_{CL}$  WEIGHTS

Loss Weight	WER (%) ↓	MCD-DWT ↓	MCD-DWT-SL ↓	SECS (%) ↑
0.001	28.74	9.30	9.44	79.83
0.005	<b>26.41</b>	9.28	9.36	79.89
0.01	28.45	<b>9.26</b>	<b>9.34</b>	<b>80.03</b>
0.05	29.92	9.29	9.38	79.45
0.1	30.44	9.39	9.47	79.06

TABLE XII  
ABLATION STUDY ON DIFFERENT HYPERPARAMETERS OF  $\mathcal{L}_{SSIM}$  WEIGHTS

Loss Weight	WER (%) ↓	MCD-DWT ↓	MCD-DWT-SL ↓	SECS (%) ↑
1.0	28.74	9.40	9.51	78.62
1.5	28.47	9.39	9.48	79.23
2.0	<b>26.56</b>	9.38	9.46	79.37
2.5	27.54	9.37	9.45	79.57
3.0	26.74	9.37	9.44	79.40
3.5	29.07	9.35	9.45	79.70
4.0	28.41	9.34	9.43	79.15
4.5	27.48	9.30	9.41	79.49
5.0	27.64	9.29	9.38	79.80
5.5	28.45	<b>9.26</b>	<b>9.34</b>	<b>80.03</b>
6.0	29.54	9.27	9.38	80.01
6.5	29.42	9.31	9.40	79.96
7.0	28.14	9.33	9.42	79.88

#### H. Evaluation of Different Loss Weights

To ensure the stability of adversarial training, we empirically determine the GAN loss weights according to previous works [15], [55], [81]. Furthermore, we conduct a detailed ablation study across a wide range of values for the contrastive learning loss  $\mathcal{L}_{CL}$  and the structural similarity index loss  $\mathcal{L}_{SSIM}$ . As shown in Tables XI and XII, the results demonstrate that a loss weight of 0.01 for  $\mathcal{L}_{CL}$  and 5.5 for  $\mathcal{L}_{SSIM}$  achieve



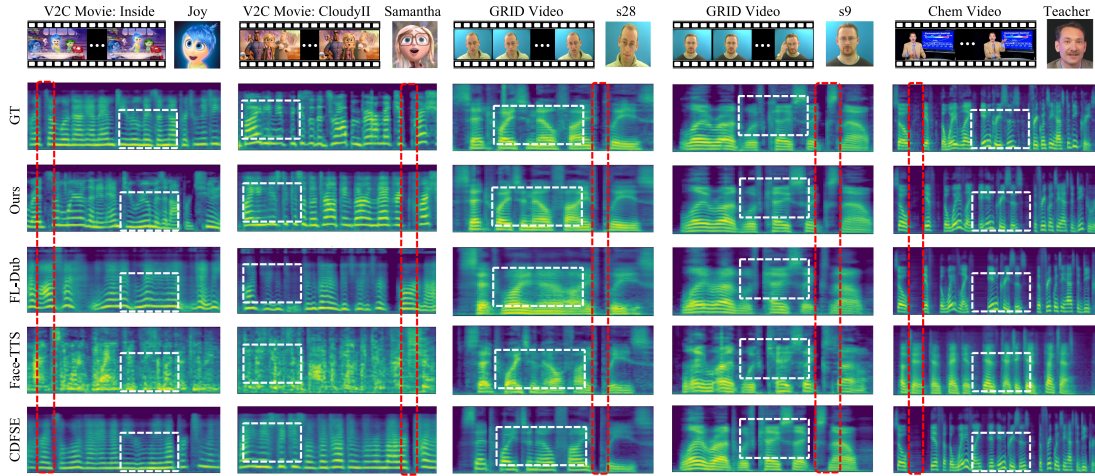


Fig. 5. The visualization of the mel-spectrograms of ground truth (GT) and synthesized audios obtained by different models. The red and white bounding boxes highlight regions where different models exhibit significant differences in duration pausing and mel-spectrograms details.

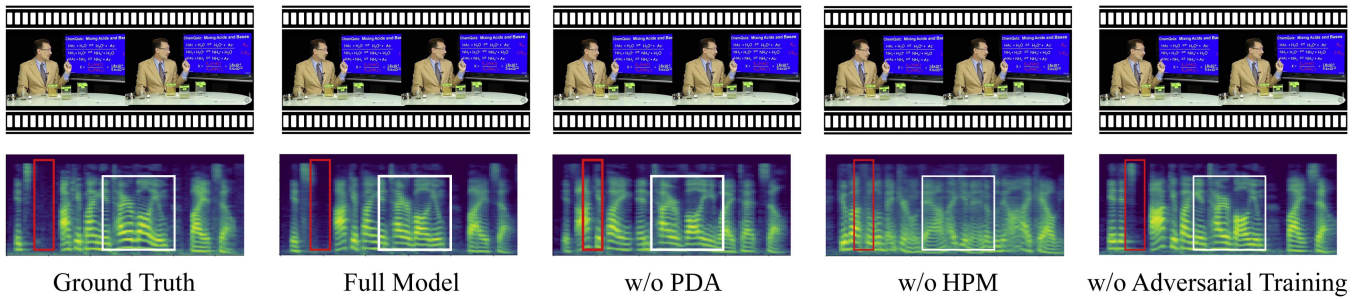


Fig. 6. Additional qualitative examples for core components of HD-Dubber. The white and red rectangles highlight key regions that have significant differences in reconstruction details and duration pause, respectively.

the best trade-off across multiple metrics. For the SSIM loss, we observe that gradually increasing its weight significantly improves the MCD-DWT metric, as it encourages the predicted mel-spectrogram to better match the structural and textural details of the ground truth. However, this trend does not hold indefinitely. When the weight exceeds 6.0, the performance begins to degrade noticeably in our experiments.

### I. Qualitative Analysis

We visualize the mel-spectrogram of reference audio, ground-truth audio, and the audio generated by the proposed method and the other two state-of-the-art methods in Fig. 5. The red and white rectangles highlight regions where different models exhibit significant differences in duration consistency and mel-spectrogram details compared to the ground truth. It shows that our model outperforms others in maintaining duration pauses. Besides, comparing white rectangle regions, we observe that the proposed method achieves distinct horizontal lines and clearer details in the mel-spectrogram, which is more similar to the ground truth than other methods.

Besides, we provide qualitative results to analyze the impact of each component of HD-Dubber on mel-spectrograms generation in Fig. 6. It shows that when the Duration Aligner is

removed, the model can synthesize clear mel-spectrograms, but it cannot ensure a reasonable duration (see the red bounding box part). This shows the key role of the Phoneme Duration Aligner (PDA) in maintaining audio-visual synchronization. In addition, when HPM is removed, the entire mel-spectrogram becomes very blurry, indicating that the speech intelligibility is poor.

### J. Failure Case and Future Work

Although the proposed HD-Dubber achieves SOTA performance on all three publicly available dubbing benchmarks, failure cases may still occur. Compared to the ground truth (Fig. 7(a)), the generated result (Fig. 7(b)) lacks a clear high-pitched reconstruction in the corresponding high frequency region of the mel-spectrogram (see white bounding boxes). Overall, singing-style dubbing differs from traditional speech-style dubbing. It often involves dramatic turns, such as high-pitched transitions. In addition, the ground truth of singing clips is usually accompanied by background music, which further increases the difficulty for the model to learn singing reconstruction. Future work can focus on improving robustness to such challenges by incorporating music-voice separation and expressive modeling.

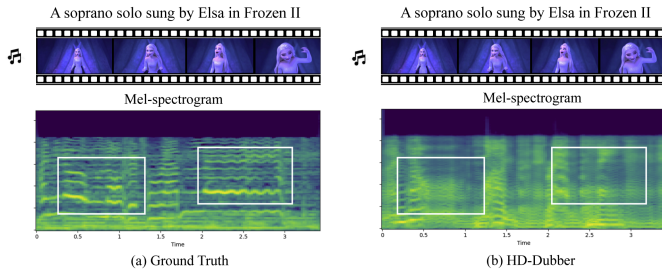


Fig. 7. Visualization of a representative failure case. This clip is taken from the movie “Frozen II” in the V2C-Animation benchmark. This clip shows the character Elsa performing a soprano solo accompanied by background music. The white bounding boxes highlight regions with large discrepancies in high-frequency reconstruction.

## V. CONCLUSION

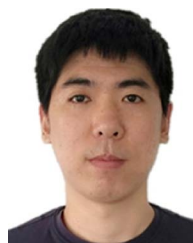
In this work, we propose a diffusion-based movie dubbing architecture, which consists of two main components: Hierarchical Phoneme Modeling (HPM) and Acoustic Diffusion Denoising (ADD). HPM learns to directly align video representations with phoneme-level counterparts at three granularities: lip movement (phoneme level duration), facial expression (pitch and energy), and scene atmosphere (global emotion). The ADD module with a parameterized Markov chain iteratively converts noise into the mel-spectrogram conditioned on textual phoneme from script and style embedding from reference audio. Extensive experiments on three widely adopted benchmarks show the favorable performance of the proposed method.

## REFERENCES

- [1] Q. Chen, M. Tan, Y. Qi, J. Zhou, Y. Li, and Q. Wu, “V2C: Visual voice cloning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 21210–21219.
- [2] Y. Ren et al., “FastSpeech 2: Fast and high-quality end-to-end text to speech,” in *Proc. Int. Conf. Learn. Representations*, 2021.
- [3] J. Shen et al., “Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2018, pp. 4779–4783.
- [4] Y. Wang et al., “Tacotron: Towards end-to-end speech synthesis,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 4006–4010.
- [5] Y. Ren et al., “FastSpeech: Fast, robust and controllable text to speech,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 3165–3174.
- [6] S. B. Hegde, K. R. Prajwal, R. Mukhopadhyay, V. P. Nambodiri, and C. V. Jawahar, “Lip-to-speech synthesis for arbitrary speakers in the wild,” in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 6250–6258.
- [7] Y. Wang and Z. Zhao, “FastLTS: Non-autoregressive end-to-end unconstrained lip-to-speech synthesis,” in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 5678–5687.
- [8] J. Lu, B. Sisman, R. Liu, M. Zhang, and H. Li, “VisualTTS: TTS with accurate lip-speech synchronization for automatic voice over,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2022, pp. 8032–8036.
- [9] F. P. Papantoniou, P. P. Filntisis, P. Maragos, and A. Roussos, “Neural emotion director: Speech-preserving semantic control of facial expressions in “in-the-wild” videos,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18759–18768.
- [10] S. Xu et al., “VASA-1: Lifelike audio-driven talking faces generated in real time,” 2024, *arXiv:2404.10667*.
- [11] W. Zhang et al., “SadTalker: Learning realistic 3D motion coefficients for stylized audio-driven single image talking face animation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 8652–8661.
- [12] S. Hu et al., “WavLLM: Towards robust and adaptive speech large language model,” 2024, *arXiv:2404.00656*.
- [13] Y. Shu et al., “LLaSM: Large language and speech model,” 2023, *arXiv:2308.15930*.
- [14] X. Wang et al., “Large-scale multi-modal pre-trained models: A comprehensive survey,” *Mach. Intell. Res.*, vol. 20, no. 4, pp. 447–482, 2023.
- [15] G. Cong et al., “Learning to dub movies via hierarchical prosody models,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 14687–14697.
- [16] X. Tan et al., “NaturalSpeech: End-to-end text-to-speech synthesis with human-level quality,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 6, pp. 4234–4245, Jun. 2024.
- [17] C. Wang et al., “Neural codec language models are zero-shot text to speech synthesizers,” 2023, *arXiv:2301.02111*.
- [18] D. Min, D. B. Lee, E. Yang, and S. J. Hwang, “Meta-StyleSpeech: Multi-speaker adaptive text-to-speech generation,” in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 7748–7759.
- [19] S. Ö. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, “Neural voice cloning with a few samples,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 10040–10050.
- [20] J. Kim, S. Kim, J. Kong, and S. Yoon, “Glow-TTS: A generative flow for text-to-speech via monotonic alignment search,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 8067–8077.
- [21] E. Casanova et al., “SC-GlowTTS: An efficient zero-shot multi-speaker text-to-speech model,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 3645–3649.
- [22] E. Casanova, J. Weber, C. D. Shulby, A. C. Júnior, E. Gölge, and M. A. Ponti, “YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone,” in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 2709–2720.
- [23] Y. A. Li, C. Han, V. S. Raghavan, G. Mischler, and N. Mesgarani, “StyleTTS 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2023, pp. 19594–19621.
- [24] Z. Ju et al., “NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models,” 2024, *arXiv:2403.03100*.
- [25] Y. Wang et al., “MaskGCT: Zero-shot text-to-speech with masked generative codec transformer,” in *Proc. Int. Conf. Learn. Representations*, 2025.
- [26] P. Belin, S. Fecteau, and C. Bedard, “Thinking the voice: Neural correlates of voice perception,” *Trends Cogn. Sci.*, vol. 8, no. 3, pp. 129–135, 2004.
- [27] H. McGurk and J. MacDonald, “Hearing lips and seeing voices,” *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.
- [28] T. Afouras, J. S. Chung, A. W. Senior, O. Vinyals, and A. Zisserman, “Deep audio-visual speech recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8717–8727, Dec. 2022.
- [29] K. R. Prajwal, R. Mukhopadhyay, V. P. Nambodiri, and C. V. Jawahar, “Learning individual speaking styles for accurate lip to speech synthesis,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13793–13802.
- [30] H. Zhao et al., “Audio-visual contrastive pre-train for face forgery detection,” *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 21, 2024, Art. no. 45.
- [31] Y. Zhang, W. Lin, and J. Xu, “Joint audio-visual attention with contrastive learning for more general deepfake detection,” *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 20, no. 5, pp. 137:1–137:23, 2024.
- [32] Y. Wang et al., “A low rank promoting prior for unsupervised contrastive learning,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 2667–2681, Mar. 2023.
- [33] C. Yang, Z. An, H. Zhou, F. Zhuang, Y. Xu, and Q. Zhang, “Online knowledge distillation via mutual contrastive learning for visual recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 10212–10227, Aug. 2023.
- [34] X. Wang and G. Qi, “Contrastive learning with stronger augmentations,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5549–5560, May 2023.
- [35] C. Sheng, M. Pietikäinen, Q. Tian, and L. Liu, “Cross-modal self-supervised learning for lip reading: When contrastive learning meets adversarial training,” in *Proc. ACM Int. Conf. Multimedia*, 2021, pp. 2456–2464.
- [36] M. Kim, J. Hong, and Y. M. Ro, “Lip-to-speech synthesis in the wild with multi-task learning,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023, pp. 1–5.
- [37] A. Graves, S. Fernández, F. J. Gomez, and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” in *Proc. Int. Conf. Mach. Learn.*, 2006, pp. 369–376.
- [38] S. Wang et al., “StyleTalk: A unified framework for controlling the speaking styles of talking heads,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 6, pp. 4331–4347, Jun. 2024.



- [39] F. Hong, L. Shen, and D. Xu, "DaGAN: Depth-aware generative adversarial network for talking head video generation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 5, pp. 2997–3012, May 2024.
- [40] T. Liu, C. Du, S. Fan, F. Chen, and K. Yu, "DiffDub: Person-generic visual dubbing using inpainting renderer with diffusion auto-encoder," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2024, pp. 3630–3634.
- [41] M. Hassid, M. T. Ramanovich, B. Shillingford, M. Wang, Y. Jia, and T. Remez, "More than words: In-the-wild visually-driven prosody for text-to-speech," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10577–10587.
- [42] C. Hu, Q. Tian, T. Li, Y. Wang, Y. Wang, and H. Zhao, "Neural dubber: Dubbing for videos according to scripts," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 16582–16595.
- [43] J. Lee, J. S. Chung, and S.-W. Chung, "Imaginary Voice: Face-styled diffusion model for text-to-speech," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023, pp. 1–5.
- [44] J. Lu, B. Sisman, M. Zhang, and H. Li, "High-quality automatic voice over with accurate alignment: Supervision through self-supervised discrete speech units," 2023, *arXiv:2306.17005*.
- [45] P. Ma, S. Petridis, and M. Pantic, "End-to-end audio-visual speech recognition with conformers," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 7613–7617.
- [46] M. Kim, C. W. Kim, and Y. M. Ro, "Deep visual forced alignment: Learning to align transcription with talking face video," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 8273–8281.
- [47] H. Xu, R. Zeng, Q. Wu, M. Tan, and C. Gan, "Cross-modal relation-aware networks for audio-visual event localization," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 3893–3901.
- [48] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. C. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 3942–3951.
- [49] P. Chen, Y. Zhang, M. Tan, H. Xiao, D. Huang, and C. Gan, "Generating visually aligned sound from videos," *IEEE Trans. Image Process.*, vol. 29, pp. 8292–8302, 2020.
- [50] H. Zhang, X. Li, and L. Bing, "Video-LLaMA: An instruction-tuned audio-visual language model for video understanding," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2023, pp. 543–553.
- [51] G. Li et al., "Learning hierarchical modular networks for video captioning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 2, pp. 1049–1064, Feb. 2024.
- [52] W. Song et al., "Dian: Duration informed auto-regressive network for voice cloning," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 8598–8602.
- [53] J. Carreira and A. Zisserman, "Quo Vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6299–6308.
- [54] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 6840–6851.
- [55] S. Liu, D. Su, and D. Yu, "DiffGAN-TTS: High-fidelity and efficient text-to-speech with denoising diffusion GANs," 2022, *arXiv:2201.11972*.
- [56] J. Liu, C. Li, Y. Ren, F. Chen, and Z. Zhao, "DiffSinger: Singing voice synthesis via shallow diffusion mechanism," in *Proc. AAAI Conf. Artif. Intell.*, AAAI Press, 2022, pp. 11020–11028.
- [57] F. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10850–10869, Sep. 2023.
- [58] H. Choi, J. Kim, J. Huh, A. Kim, J. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex U-net," in *Proc. Int. Conf. Learn. Representations*, 2019.
- [59] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
- [60] R. Huang, Z. Zhao, H. Liu, J. Liu, C. Cui, and Y. Ren, "ProDiff: Progressive fast diffusion model for high-quality text-to-speech," in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 2595–2605.
- [61] Z. Xiao, K. Kreis, and A. Vahdat, "Tackling the generative learning trilemma with denoising diffusion GANs," in *Proc. Int. Conf. Learn. Representations*, 2022.
- [62] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2813–2821.
- [63] I. J. Goodfellow et al., "Generative adversarial nets," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
- [64] Y. Zhou et al., "Content-dependent fine-grained speaker embedding for zero-shot speaker adaptation in text-to-speech synthesis," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2022, pp. 2573–2577.
- [65] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoustical Soc. Amer.*, vol. 120, no. 5, pp. 2421–2424, 2006.
- [66] M. Müller, "Dynamic time warping," in *Information Retrieval for Music and Motion*, Berlin, Germany: Springer, 2007, pp. 69–84.
- [67] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 28492–28518.
- [68] A. C. Morris, V. Maier, and P. D. Green, "From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2004, pp. 2765–2768.
- [69] J. Ye, X. Wen, Y. Wei, Y. Xu, K. Liu, and H. Shan, "Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023, pp. 1–5.
- [70] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using Kaldi," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 498–502.
- [71] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S3FD: Single shot scale-invariant face detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 192–201.
- [72] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 6319–6323.
- [73] P. Ma, B. Martinez, S. Petridis, and M. Pantic, "Towards practical lipreading with distilled and efficient models," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 7608–7612.
- [74] N. Li, Y. Liu, Y. Wu, S. Liu, S. Zhao, and M. Liu, "RobuTrans: A robust transformer-based text-to-speech model," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 8228–8235.
- [75] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
- [76] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.
- [77] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.
- [78] G. Cong et al., "StyleDubber: Towards multi-scale style learning for movie dubbing," 2024, *arXiv:2402.12636*.
- [79] M. Chen et al., "MultiSpeech: Multi-speaker text to speech with transformer," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 4024–4028.
- [80] X. Liang, Z. Wu, R. Li, Y. Liu, S. Zhao, and H. Meng, "Enhancing monotonicity for robust autoregressive transformer TTS," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 3181–3185.
- [81] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 17022–17033.



**Liang Li** received the BS degree from Xi'an Jiaotong University, in 2008, and the PhD degree from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China, in 2013. From 2013 to 2015, he held a post-doc position with the Department of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing, China. Currently, he is serving as the full professor with ICT, CAS. He has served on several committees of international journals and conferences.





**Gaoxiang Cong** received the MS degree from Shandong University, China, in 2024. He is currently working toward the PhD degree with the Institute of Computing Technology, Chinese Academy of Sciences. He is also with the Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences. His research interests include visual voice cloning, diffusion modeling, and speech enhancement.



**Quan Z. Sheng** (Member, IEEE) received the PhD degree in computer science from the University of New South Wales and did his post-doc as a research scientist with CSIRO ICT Centre. He is a full professor and head of the School of Computing, Macquarie University. Before moving to Macquarie, he spent ten years with the School of Computer Science, The University of Adelaide, serving in several senior leadership roles, including acting head and deputy head of the School of Computer Science.



**Yuankai Qi** received the MS and PhD degrees from the Harbin Institute of Technology, China, in 2013 and 2018, respectively. His research interests include vision-language navigation, object counting, and video captioning. He serves as area chair of top-tier conferences such as the CVPR, NeurIPS, ICCV, and AAAI, etc. He is also a regular reviewer for *IEEE Transactions on Pattern Analysis and Machine Intelligence* and *IEEE Transactions on Image Processing*, etc. He was awarded the best paper of ACM Multimedia 2024, and the world's top 2% scientist.



**Qingming Huang** (Fellow, IEEE) received the bachelor's degree in computer science and the PhD degree in computer engineering from the Harbin Institute of Technology, China, in 1988 and 1994, respectively. He is currently a professor with the University of Chinese Academy of Sciences and an adjunct research professor with the Institute of Computing Technology, Chinese Academy of Sciences. He has authored or coauthored more than 400 papers in prestigious international journals and conferences.



**Zheng-Jun Zha** received the BE and PhD degrees from the University of Science and Technology of China (USTC), Hefei, China, in 2004 and 2009, respectively. He is a full professor with the School of Information Science and Technology, USTC, and the vice director of the National Engineering Laboratory for Brain-Inspired Intelligence Technology and Application. He was a researcher with the Hefei Institutes of Physical Science, CAS, from 2013 to 2015 and a senior research fellow with the School of Computing, National University of Singapore (NUS), from 2011 to 2013.



**Qi Wu** received the MSc and PhD degrees in computer science from the University of Bath, United Kingdom, 2011 and 2015, respectively. He is a senior lecturer (assistant professor) with the University of Adelaide, and he is an associate investigator with the Australia Centre for Robotic Vision (ACRV). His educational background is primarily in computer science and mathematics. He works on vision and language problems, including image captioning, question answering, and dialog.



**Ming-Hsuan Yang** (Fellow, IEEE) is a professor in electrical engineering and computer science with the University of California, Merced, Merced, California. He serves as a program co-chair of the IEEE International Conference on Computer Vision (ICCV), in 2019 and general co-chair of ACCV 2016. He received the Longuet-Higgins Prize in 2023, NSF CAREER award, in 2012, and Google Faculty Award, in 2009. He is a fellow of the ACM and AAAI.