

StyleSpeaker: Audio-Enhanced Fine-Grained Style Modeling for Speech-Driven 3D Facial Animation

An Yang¹, Chenyu Liu², Pengcheng Xia², Jun Du^{*1}

¹NERC-SLIP, University of Science and Technology of China

²IFLYTEK Research

Abstract

Speech-driven 3D facial animation is challenging due to the diversity in speaking styles and the limited availability of 3D audio-visual data. Speech predominantly dictates the coarse motion trends of the lip region, while specific styles determine the details of lip motion and the overall facial expressions. Prior works lack fine-grained learning in style modeling and do not adequately consider style biases across varying speech conditions, which reduce the accuracy of style modeling and hamper the adaptation capability to unseen speakers. To address this, we propose a novel framework, StyleSpeaker, which explicitly extracts speaking styles based on speaker characteristics while accounting for style biases caused by different speeches. Specifically, we utilize a style encoder to capture speakers' styles from facial motions and enhance them according to motion preferences elicited by varying speech conditions. The enhanced styles are then integrated into the coarse motion features via a style infusion module, which employs a set of style primitives to learn fine-grained style representation. Throughout training, we maintain this set of style primitives to comprehensively model the entire style space. Hence, StyleSpeaker possesses robust style modeling capability for seen speakers and can rapidly adapt to unseen speakers without fine-tuning. Additionally, we design a trend loss and a local contrastive loss to improve the synchronization between synthesized motions and speeches. Extensive qualitative and quantitative experiments on three public datasets demonstrate that our method outperforms existing state-of-the-art approaches.

1. Introduction

Speech-driven 3D facial animation has become an increasingly important research area due to its applications in vir-

^{*}Corresponding author.

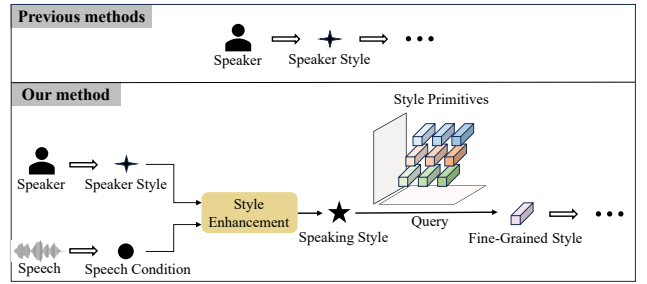


Figure 1. Illustrations of the style modeling process of our proposed StyleSpeaker. Unlike previous methods, StyleSpeaker not only models speaking styles based on both speaker characteristics and specific speech, but also queries fine-grained representations based on style primitives.

tual reality, film or game production, and biomimetic technology. In this task, it is crucial to ensure not only the accuracy of lip motions but also the naturalness of the overall animation and the consistency of the speaking style characteristics. The error in speech-driven 3D facial animation arises from three sources: viseme information provided by the speech, speaking style, and randomness. We focus only on the first two. The viseme information determines the coarse motion trends of the lip region, such as opening and closing, while the speaking style determines detailed lip motions and overall facial expressions. The speaking style can be decomposed into two components: first, the speaker’s inherent style, referred to as “speaker style” in the following sections, which stems from the speaker’s habitual patterns. The speaker style dictates audio-independent motion habits, such as the motion patterns in the upper-face and the symmetry of lip motions during speech. Second, the specific speech condition, which further indicates motion preferences and influences the speaking style to some extent. For example, speech intensity affects the maximum extent of mouth opening and the variance of changes in lip region. To summarize, the speaking style is primarily deter-

mined by the speaker identity but is influenced by the driven speech.

Accurately extracting the speaking style is crucial for synthesizing precise and characteristic-preserving facial animation. Previous works tend to overlook fine-grained style modeling and audio-induced style biases. Early methods [16, 26] can only synthesize facial animation in a single style. Later works [6, 9, 27, 40] embed one-hot encoding for individuals in the training set via an embedding layer and optimize this layer during training to implicitly capture speaker styles. However, the resulting styles become entangled with speech information, and cannot adapt to unseen speakers. Based on the prior methods, Imitator [33] adapts to unseen speakers by fine-tuning some layers of the generalized model, which incurs additional overhead. Recent methods [11, 37] begin to explicitly extract styles from facial motion sequences but ignore the impact of speech conditions and generalization learning in style space, which results in less accurate style extraction and weakens the adaptability to the styles of unseen speakers.

To address these issues, we propose a novel method, StyleSpeaker, which differs from previous methods by incorporating audio conditions into style considerations and using style primitives for fine-grained learning. We illustrate in Figure 1 the improvements in the style modeling process of our method compared to previous methods. Specifically, we employ a speaker style extractor to capture speaker styles. Concurrently, we utilize an audio encoder to extract high-level audio features and learn audio conditions from them via an audio condition extractor. We then enhance the speaker styles with the audio conditions to produce speaking styles, aiming to integrate motion preference information embedded in the audio conditions into speaking styles. Subsequently, we use a style infusion module that employs a set of style primitives to derive the fine-grained style representation for the speaking styles. In practice, we deconstruct the input speaking styles into foundational representation using style primitives, which are simultaneously optimized throughout training. Through iterative deconstruction and learning of encountered styles, these style primitives capture foundational style information to construct a comprehensive style space. For unseen speakers in the training set, we extract their speaking styles, map them within the style space for fine-grained representation, and acquire more precise style information. This approach significantly enhances our model’s adaptability to diverse speaking styles. In addition, we design two constraints: trend loss and local contrastive loss. The trend loss imposes constraints on the higher-order differences in the facial motion sequences. Derived from CLIP [25], the local contrastive loss accounts for the recurrence of many syllables and facial motions within a speech segment. To avoid mismatches, we limit the computation of the contrastive

loss to a local range. These two constraints significantly improve the accuracy of the synthesized animation. We also design a new metric, Fourier Frequency Error, which effectively evaluates style consistency. The main contributions of our work are as follows:

- We are the first to enhance speaking styles by incorporating audio conditions. Our fine-grained learning of the style space endows the model with strong style modeling capabilities, allowing rapid adaptation to unseen speakers without fine-tuning.
- We introduce two novel constraint functions, trend loss and local contrastive loss, to further improve the accuracy and synchronization of the synthesized motions.
- Extensive qualitative and quantitative experiments on three public datasets demonstrate that our method outperforms existing state-of-the-art approaches for both seen and unseen speakers.

2. Related Works

2.1. Speech-Driven 3D Facial Animation

Initially, speech-driven 3D facial animation is synthesized using rule-based methods. The dominance functions [21] are employed to map speech to parameters controlling facial motions. Some methods [7, 22] model facial muscle motions from biological and anatomical perspectives and establish mappings with speech.

With the advent of 4D face datasets, various learning-based methods have emerged [13]. Some methods [16, 26] focus on driving a specific speaker, while more methods are dedicated to driving different speakers. VOCA [6] primarily generates lower face motions. MeshTalk [27] uses a categorical latent space to disentangle audio-correlated and audio-uncorrelated information. FaceFormer [9] first employs the transformer architecture. CodeTalker [40] utilizes a discrete codebook approach to decouple motion space. FaceDiffuser [31], DiffSpeaker [20] and DiffSHEG [3] employ diffusion models. CorrTalk [5] divides faces into two regions based on audio correlation and uses two branches to generate motions separately. TalkingStyle [30] disentangles style codes from motion patterns and proposes a style-conditioned self-attention mechanism. EmoTalk [23] utilizes a paired emotional content dataset to explicit control emotion styles. CSTalk [19] proposes a parameter model for representing faces and learns the motion characteristics of specific emotions. All these methods use one-hot encoding to represent different speaker styles, which fails to generalize to unseen speakers. Subsequently, Imitator [33] fine-tunes the generalized model for the adaptability to unseen speakers. Wu et al. [37] extracts styles from facial motions using the temporal convolutional architecture. Mimic [11] disentangles speaking styles and speech content using two separate latent spaces. DiffPoseTalk [32] extracts styles for

expressions and poses from parameterized facial sequences. Yang et al. [41] propose a probabilistic model to preserve diversity in speech-driven tasks, while Probtalk3D [38] employs a non-deterministic model with emotion control for similar purposes. Recently, some works [8, 39, 42] utilize video or other modality information and pre-trained models to enhance the naturalness of synthesized 3D animation. The aforementioned methods overlook style variations under different speeches for a specific speaker and fail to learn style in the fine-grained manner, limiting the ability to model style and adapt to the new styles of unseen speakers.

2.2. Stylized Talking Head Video Generation

Various methods exist for stylized talking head video generation. Ji et al. [14] disentangle content and emotion information from audio and generate videos guided by the predicted landmarks. Wang et al. [35] and Sinha et al. [29] use emotion labels as style codes but lack fine-grained individualized control. Ji et al. [15] and Liang et al. [18] use audio as content information and employ reference video frames as styles. StyleTalk++ [36] and SyncTalk [24] capture facial expression and head pose information from reference videos and extract them as style codes to control video generation. StyleSync [12] encodes the style information into the W^+ space and generates target frames through StyleGan inversion[1, 2, 28].

3. Method

3.1. Motivation

To explore the distribution of speaking styles, we extract features from facial motion sequences of different speakers in different speeches. Inspired by FDD metric [40], we use upper-face dynamics deviation to reflect the speaking style information in facial motion sequences. Specifically, we calculate the standard deviation for each vertex motion sequence along the temporal dimension and project the results into 2D space for visualization using t-SNE [34], as shown in Figure 2 (a). Additionally, we calculate the discrete Fourier transform and extract the first 20 principal frequency components for each vertex motion sequence in x, y, and z directions, and project the results into 2D space, as shown in Figure 2 (b). Based on the two visualization results, we observe significant style differences among various speakers, while the same speaker exhibits different preferences under distinct speech conditions. This indicates that incorporating the style biases brought by audio can enhance the accuracy of style modeling. Moreover, variations across different speakers and speech conditions result in a multitude of speaking styles. To effectively handle the large variety of styles, we construct a style space using fundamental units to represent all styles, rather than extracting and directly applying styles as done in previous work. We aim for

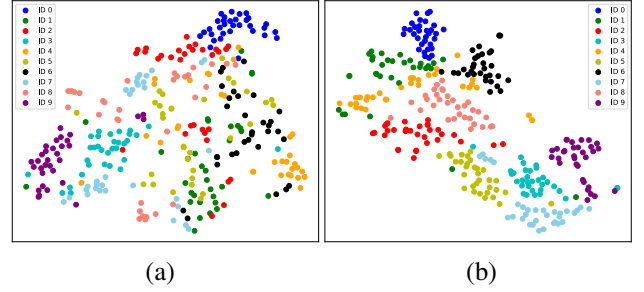


Figure 2. Visualization of motion characteristics under two feature extraction methods.

the fundamental units to learn and utilize all observed styles from the training phase as comprehensively as possible, becoming an expert capable of representing a broad spectrum of styles.

3.2. Overview

We focus on synthesizing 3D facial animation that is highly synchronized with speech and retains the characteristics of the speaking style. To this end, we propose a novel model, StyleSpeaker, which comprises four modules: audio encoder, style encoder, viseme transformer decoder, and style infusion module, as shown in Figure 3. We use four constraint functions for training. In the following, we will provide a detailed introduction to our model architecture, training strategy and objectives, and style adaptation.

Problem Formulation. Let $\mathbf{M}_{1:T} = \{m_1, \dots, m_T\}$ be a sequence of facial motions, where each frame $m_t \in \mathbb{R}^{N \times 3}$ denotes the displacement of N vertices over a neutral-face mesh template $h \in \mathbb{R}^{N \times 3}$ in the xyz three directions at time t . Furthermore, let \mathcal{X} be the speech segment. Our goal is to synthesize the facial motion sequence $\mathbf{M}_{1:T}$ based on the speech segment \mathcal{X} and speaker information. Afterwards, we combine $\mathbf{M}_{1:T}$ with the template h to obtain the target speech-driven 3D facial animation $\mathbf{O}_{1:T} = \{m_1 + h, \dots, m_T + h\}$.

3.3. Model Architecture

Audio Encoder. The input audio \mathcal{X} undergoes feature extraction before being fed into the subsequent network. We use a self-supervised pre-training model, WavLM [4], as our audio encoder, which consists of a convolutional feature encoder and a transformer encoder with gated relative position bias. We initialize our audio encoder with the pre-trained WavLM weights and freeze the convolutional feature encoder during training. To match the frame rate of the facial motions, we add a linear interpolation layer after the transformer encoder for resampling the audio features. We obtain the final audio features $\mathbf{A}_{1:T} = \{a_1, \dots, a_T\} \in \mathbb{R}^{d_a \times T}$.

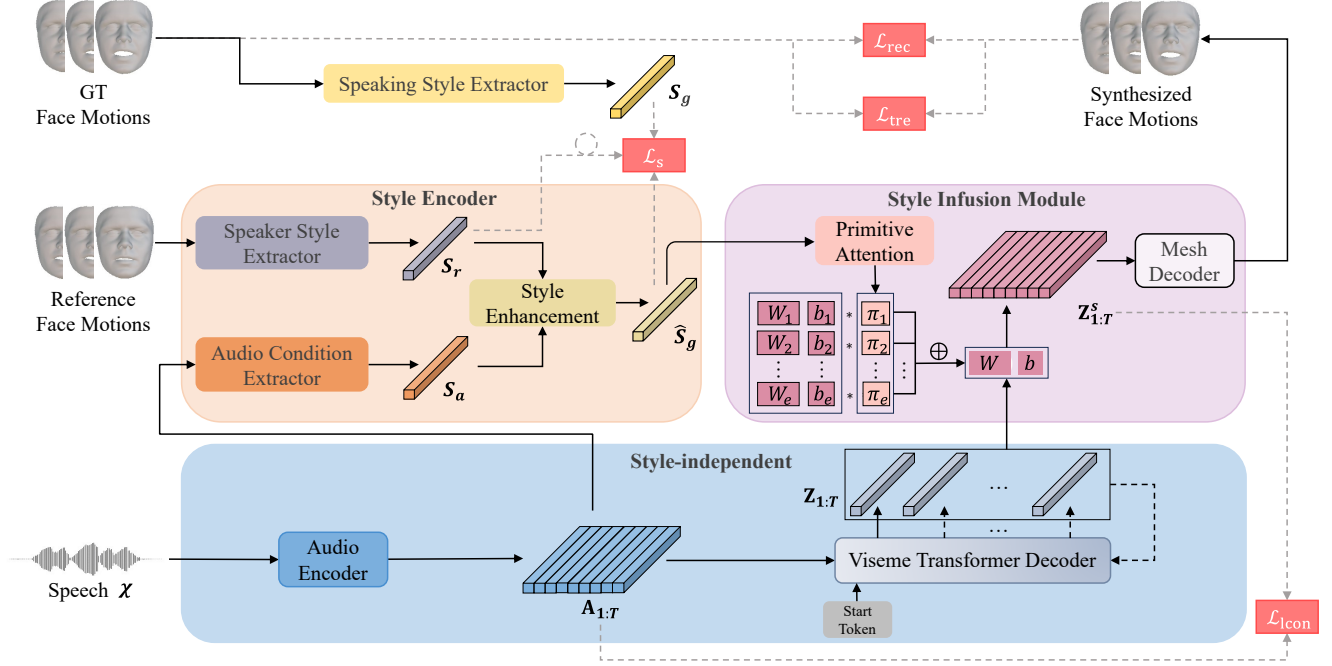


Figure 3. The pipeline of StyleSpeaker. Our framework separately extracts viseme features and style vectors before the final fusion. We use the audio encoder to extract audio features $\mathbf{A}_{1:T}$, which are then fed into the viseme transformer decoder to generate style-independent viseme features $\mathbf{Z}_{1:T}$ in an autoregressive manner. Concurrently, we extract the speaker style S_r from reference face motions and the audio condition vector S_a from $\mathbf{A}_{1:T}$. We then use S_a to enhance S_r , producing the predicted speaking style \hat{S}_g . Finally, the style infusion module integrates \hat{S}_g into $\mathbf{Z}_{1:T}$ using combined style primitives and synthesizes facial motions.

Style Encoder. To guide the synthesis of 3D facial animation, it is essential to capture the speaking style of the target speaker under the target speech, denoted as $S_g \in \mathbb{R}^{d_s}$. We decompose S_g into two components: the speaker style vector $S_r \in \mathbb{R}^{d_s}$, which carries the speaker’s typical speaking habits, and audio condition vector $S_a \in \mathbb{R}^{d_s}$, which carries the condition information of the target speech. We extract S_r from the facial motion sequence using a speaker style extractor, which consists of a temporal convolutional network (TCN) and feature extraction layers. The detailed structure is provided in Appendix A. Given that the input facial motion sequences carry different speech conditions during training, we impose a consistency constraint on S_r for the same speaker, ensuring that the speaker style extractor disregards these variations and captures the shared underlying information. We employ an audio condition extractor, structured similarly to the speaker style extractor, to learn S_a from audio features $\mathbf{A}_{1:T}$. We then enhance S_r with S_a to derive the predicted speaking style $\hat{S}_g \in \mathbb{R}^{d_s}$. The enhancement is as follows:

$$S_{\text{bias}} = W_s \begin{bmatrix} S_r \\ S_a \end{bmatrix} + b_s, \quad (1)$$

where $S_{\text{bias}} \in \mathbb{R}^{d_s}$ represents the style bias. $W_s \in \mathbb{R}^{d_s \times 2d_s}$ and $b_s \in \mathbb{R}^{d_s}$ are trainable parameters. We obtain \hat{S}_g by:

$$\hat{S}_g = S_r + \alpha S_{\text{bias}}, \quad (2)$$

where α is used to control the degree of the style bias. We set $\alpha = 0.1$ in our implementation. Additionally, we employ a speaking style extractor, with the same structure as the speaker style extractor, to learn the actual speaking style S_g under the target speech as a form of supervision.

Viseme Transformer Decoder. Our viseme transformer decoder is a multi-layer transformer decoder with biased causal self-attention to learn the dependencies of the current frame on past frames, and cross-modal attention to align the audio features $\mathbf{A}_{1:T}$ with the viseme features in the motion modality. We obtain the viseme features $\mathbf{Z}_{1:T} = \{z_1, \dots, z_T\} \in \mathbb{R}^{d_m \times T}$, which are independent of style and represent the coarse motions. Formally,

$$\mathbf{Z}_{1:T} = D_v(\mathbf{A}_{1:T}). \quad (3)$$

Style Infusion Module. The style infusion module integrates the style information \hat{S}_g into the viseme features $\mathbf{Z}_{1:T}$ and synthesizes the final facial motion sequence. Most previous methods directly input the extracted style as a

reference into the subsequent network, whereas our approach incorporates fine-grained style learning and maintains a style primitive table during training, enhancing style adaptability. Specifically, the module consists of a primitive attention layer, a style infusion layer, and a mesh decoder. We use the combination of e pairs of base weights (W_i, b_i) to project the style-agnostic $\mathbf{Z}_{1:T}$ to the style-infused motion features $\mathbf{Z}_{1:T}^s$, referring to these pairs of base weights as style primitives. We use these style primitives to decompose speaking styles into more basic representations, and aim to effectively represent diverse styles using the combinations of the style primitives. Specifically, \hat{S}_g first undergoes the primitive attention layer to generate the attention weights for each primitive. These primitives are dynamically aggregated via the attention vector π , resulting in the style infusion layer weights (W, b) corresponding to \hat{S}_g :

$$W = \sum_{i=1}^e \pi_i W_i, b = \sum_{i=1}^e \pi_i b_i, \quad (4)$$

where $0 \leq \pi_i \leq 1$, $\sum_{i=1}^e \pi_i = 1$. π_i is the attention weight of i -th style primitive. $W \in \mathbb{R}^{d_m \times d_m}$, $b \in \mathbb{R}^{d_m}$. The style-infused motion features $\mathbf{Z}_{1:T}^s = \{z_1^s, \dots, z_T^s\} \in \mathbb{R}^{d_m \times T}$ are then generated by:

$$z_i^s = W z_i + b. \quad (5)$$

Finally, we obtain the predicted face motions $\hat{\mathbf{M}}_{1:T} = \{\hat{m}_1, \dots, \hat{m}_T\}$, which are projected from $\mathbf{Z}_{1:T}^s$ via the mesh decoder.

3.4. Training

Training Strategy. For a target speaker and speech, we extract the speaker style S_r from the reference facial motion sequence from other speeches of the same speaker, and learn the speech condition S_a specific to the target speech. These two components are then fused as a style reference \hat{S}_g to synthesize the target motion sequence. We impose consistency constraints on S_r extracted from the same speaker to minimize its distance to the cluster center. Additionally, we extract the speaking style S_g under the target speech condition from the ground-truth facial motion sequence to supervise the predicted speaking style \hat{S}_g , accelerating convergence. This training strategy effectively decouples the speaking style into the speaker style and audio-induced style biases.

Reconstruction Loss. The reconstruction loss \mathcal{L}_{rec} is:

$$\mathcal{L}_{\text{rec}} = \sum_{t=1}^T \|\hat{m}_t - m_t\|_2^2. \quad (6)$$

Style Loss. The style loss is defined as follows:

$$\mathcal{L}_s = \|\hat{S}_g - S_g\|_2^2 + \|S_r - \mu(S_r)\|_2^2, \quad (7)$$

where \hat{S}_g is obtained from Equation (2). Let $\mu(S_r)$ represent the mean of all S_r extracted from the reference facial sequences of the same speaker, with the one-hot encoding of the speaker in the training set as the initial value.

Trend Loss. We impose constraints on the differences at various orders between facial motion sequences. Previous works use only the first-order difference as a constraint, but the facial motions between consecutive frames are minimal and do not effectively reflect the trend information. We believe that using differences at various time intervals can better constrain the facial motion trends. The trend loss is defined as follows:

$$\mathcal{L}_{\text{tre}} = \frac{1}{R} \sum_{r=1}^R \sum_{t=1}^{T-r} \|(\hat{m}_{t+r} - \hat{m}_t) - (m_{t+r} - m_t)\|_2^2, \quad (8)$$

where R denotes the maximum order, we set $R = 5$ in our implementation.

Local Contrastive Loss. Inspired by CLIP [25], we design a local contrastive loss to align sequences with repetitive characteristics. In this work, we use this loss to align audio features $\mathbf{A}_{1:T}$ with motion features $\mathbf{Z}_{1:T}^s$, aiming to improve lip synchronization. The local contrastive loss is defined as follows:

$$\mathcal{L}_{\text{lcon}} = \frac{1}{T} \sum_{t=1}^T \left(\lambda \mathcal{L}_t^{(a \rightarrow m)} + (1 - \lambda) \mathcal{L}_t^{(m \rightarrow a)} \right) + \|W_l\|_1, \quad (9)$$

where $\lambda = 0.5$. $\mathcal{L}_t^{(a \rightarrow m)}$ and $\mathcal{L}_t^{(m \rightarrow a)}$ represent audio-to-motion and motion-to-audio local contrastive loss:

$$\mathcal{L}_t^{(a \rightarrow m)} = -\log \frac{\exp(\langle a_t, W_l z_t^s \rangle / \tau)}{\sum_{i=1}^T \exp(\langle a_t, W_l z_i^s \rangle / \tau) \cdot I_t^k(i)}, \quad (10)$$

$$\mathcal{L}_t^{(m \rightarrow a)} = -\log \frac{\exp(\langle W_l z_t^s, a_t \rangle / \tau)}{\sum_{i=1}^T \exp(\langle W_l z_i^s, a_t \rangle / \tau) \cdot I_t^k(i)}, \quad (11)$$

where $\langle \cdot \rangle$ denotes the cosine similarity. τ is the temperature parameter, which is fixed to 0.1 in our experiments. The matrix $W_l \in \mathbb{R}^{d_a \times d_m}$ is a set of learnable parameters that aligns z_t^s from the motion space to the audio space. Since audio is strongly correlated only with the mouth region, we apply l_1 regularization constraint on W_l to induce sparsity, encouraging the aligned features to focus on local regions. $I_t^k(i)$ is an indicator function that outputs 1 when $|i - t| \leq k$, and 0 otherwise. This setup confines the computation of the contrastive loss within the range of $2k + 1$ frames.

Training Objectives. To train our model, we use $\mathcal{L}_{\text{total}}$ as our final loss function, defined as follows:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{rec}} \mathcal{L}_{\text{rec}} + \lambda_s \mathcal{L}_s + \lambda_{\text{tre}} \mathcal{L}_{\text{tre}} + \lambda_{\text{lcon}} \mathcal{L}_{\text{lcon}}. \quad (12)$$

To ensure all loss terms remain at a similar scale, we set the weights as follows: $\lambda_{\text{rec}} = 1.0$, $\lambda_s = 0.001$, $\lambda_{\text{tre}} = 1.0$, and $\lambda_{\text{lcon}} = 0.001$.

3.5. Style Adaptation

Our model can rapidly adapt to the new styles of unseen speakers outside the training set. Given a video without audio of an unseen speaker, we obtain the 3D facial geometry sequence as a style reference sequence through a reconstruction method HRN [17]. We extract S_r from the reference sequence using the trained speaker style extractor without fine-tuning the original model weights.

4. Experiments

4.1. Datasets

We use two widely adopted 4D datasets, BIWI [10] and VOCASET [6], along with the synthetic dataset 3D-MEAD. All three datasets provide paired spoken English audio and 3D facial geometry sequences.

BIWI Dataset. BIWI is a corpus comprising affective speech and corresponding dense dynamic 3D face geometries. The dataset contains 40 sentences, each spoken by 14 subjects (8 females and 6 males) with an average duration of 4.67 seconds. The 3D face geometries are captured at 25 fps, each with 23370 vertices. Following Fan et al. [9], we select sentences with emotional context and partition the data as follows: a training set (BIWI-Train), a validation set (BIWI-Val), and two test sets (BIWI-Test-A and BIWI-Test-B). BIWI-Train, BIWI-Val, and BIWI-Test-A contain 192, 24, and 24 sentences, respectively, from the same 6 subjects. BIWI-Test-B contains 32 sentences from 8 unseen subjects.

VOCASET Dataset. VOCASET contains 480 speech sentences and corresponding 3D facial geometry sequences from 12 subjects, with an average duration of about 4 seconds. The 3D facial geometries are captured at 60 fps, each with 5023 vertices. We follow the data split methodology of VOCA [6] to create a training set (VOCA-Train), a validation set (VOCA-Val), and a test set (VOCA-Test).

3D-MEAD Dataset. 3D-MEAD is synthesized through the 3D facial reconstruction method HRN [17] based on MEAD dataset [35]. The 3D facial geometries are captured at 30 fps, each with 35709 vertices. More details are provided in Appendix B. 3D-MEAD comprises 1760 sequences from 44 speakers, which are divided into a training set (MEAD-Train), a validation set (MEAD-Val) and two test sets (MEAD-Test-A and MEAD-Test-B). MEAD-Train, MEAD-Val, and MEAD-Test-A contain 1152, 144, and 144 sequences, respectively, from the same 36 speakers. MEAD-Test-B contains 136 sequences from 8 unseen speakers.

4.2. Implementation Details

Our framework is implemented by PyTorch. We train our model, StyleSpeaker, on a single NVIDIA A40 GPU

for 50 epochs. We employ Adam optimizer for training, with an initial learning rate set to 0.0001. After 40 epochs, the learning rate is decayed to 25% of the initial rate. We compare our method with FaceFormer [9], CodeTalker [40], FaceDiffuser [31], CorrTalk [5], Imitator [33], and Mimic [11]. More details of baselines and our implementation can be found in Appendix A.

4.3. Quantitative Evaluation

Metric. We quantitatively evaluate the synthesized motions for accuracy and style consistency. Currently, there is no widely accepted metric for style consistency evaluation. Previous studies use FDD [40] for this purpose, but our observations in Figure 2 indicate that motion features extracted using the discrete Fourier transform provide better style distinction between different speakers. Therefore, we propose a new general metric, Fourier Frequency Error (FFE), for evaluating style consistency. Specifically, FFE is calculated by:

$$\text{FFE}(\mathbf{M}_{1:T}, \hat{\mathbf{M}}_{1:T}) = \frac{\sum_{c=1}^{N \times 3} \|\mathcal{F}(\mathbf{M}_{1:T}^c) - \mathcal{F}(\hat{\mathbf{M}}_{1:T}^c)\|_2^2}{N \times 3}, \quad (13)$$

where $\mathbf{M}_{1:T}^c \in \mathbb{R}^T$ denotes the sequence of a vertex motion component in x, y, or z direction. \mathcal{F} represents the discrete Fourier transform, which extracts the first 20 principal frequency components.

We adopt four metrics: (1) Lip Vertex Error (LVE). It calculates the maximal l_2 error in the lip region between each frame of the synthesized motion sequences and the ground truth motion sequences, and averages this over all frames. (2) Face Vertex Error (FVE). It calculates the average l_2 error over the entire face between each frame of the synthesized motion sequences and the ground truth motion sequences, and averages this over all frames. (3) Face Dynamic Time Wrapping (FDTW). It evaluates synchronization by computing temporal sequence similarity using Dynamic Time Warping. (4) Fourier Frequency Error (FFE).

Comparison. We first evaluate the synthesis capability of different models on BIWI-Test-A. According to the results in Table 1, our model StyleSpeaker outperforms other models in terms of lip accuracy, overall facial motion accuracy and synchronization, and style consistency. Moreover, to further compare the comprehensive synthesis capability on

Method	LVE ↓ ($\times 10^{-4}$ mm)	FVE ↓ ($\times 10^{-5}$ mm)	FDTW ↓	FFE ↓ ($\times 10^{-2}$)
FaceFormer [9]	5.3077	8.7978	1.147	1.82
CodeTalker [40]	4.7914	8.2758	1.156	1.63
FaceDiffuser [31]	4.2977	7.6533	1.097	1.46
CorrTalk [5]	4.0858	7.4356	1.090	1.37
Ours	3.8036	6.3635	0.995	1.26

Table 1. Quantitative evaluations on BIWI-Test-A.

Method	LVE ↓ ($\times 10^{-3}$ mm)	FVE ↓ ($\times 10^{-4}$ mm)	FDTW ↓	FFE ↓ ($\times 10^{-2}$)
Imitator [33]	1.2934	1.2748	1.204	2.29
Mimic [11]	1.2630	1.2194	1.221	2.29
Ours	0.9423	1.0391	1.167	1.87
Imitator	3.9451	4.5677	2.111	7.98
Imitator*	2.2391	2.2515	1.438	3.41
Mimic	2.4400	2.3133	1.440	3.42
Ours	2.0236	2.1251	1.422	3.10

Table 2. Quantitative evaluations on MEAD-Test-A (top 3 rows) and MEAD-Test-B (bottom 4 rows). * denotes the fine-tuned Imitator.

both seen and unseen speakers, we conduct experiments on 3D-MEAD, which encompasses a wider variety of styles. For MEAD-Test-B, we use a single motion sequence from the target speaker as a style reference. According to the results in Table 2, our model outperforms Imitator and Mimic across all metrics on both MEAD-Test-A and MEAD-Test-B, particularly in LVE and FFE. This demonstrates that our model synthesizes more accurate lip motions and possesses stronger style adaptation capability.

4.4. Qualitative Evaluation

Visual Comparison. We visually compare our method with competing methods in Figure 4. Our model synthesizes more accurate lip motions. For example, in the pronunciation of /s/ sound in “expensive”, the synthesized lip motion from our model forms a narrow slit, closely matching the ground truth (GT). The lip closure during /m/ sound in “me” and the opening during /ðə/ sound in “that” are also more closely aligned with the GT. Additionally, our model captures more precise facial details for both seen and unseen speakers, such as preserving facial wrinkles and muscle contractions during speech, as well as the asymmetry of the lips when speaking. This highlights our model’s stronger capabilities in style learning and adaptation. Notably, even without applying specific constraints to the eye region, some of the synthesized motion sequences exhibit blinking actions, suggesting that periodic motions like blinking are effectively captured by our style modeling. See the supplementary video for more visualization examples.

User Study. To evaluate the performance of different methods perceptually, we conduct a user study on BIWI-Test-B, VOCASET-Test, and MEAD-Test-B, focusing on lip synchronization, realism, and style consistency. For BIWI-Test-B and VOCASET-Test, we obtain 30 videos for each method and create 150 A vs. B pairs (30 videos \times 5 competing methods). For MEAD-Test-B, we obtain 32 videos for each method and create 96 A vs. B pairs (with the GT videos for style evaluation). We invite 30 par-

Ours vs. Competitors	BIWI-Test-B		VOCA-Test	
	Lip Sync	Realism	Lip Sync	Realism
Ours vs. FaceFormer [9]	82.22	80.00	82.22	84.44
Ours vs. CodeTalker [40]	76.67	75.56	80.00	72.22
Ours vs. FaceDiffuser [31]	66.67	63.33	68.89	67.78
Ours vs. CorrTalk [5]	65.56	55.56	58.89	63.33
Ours vs. GT	46.67	42.22	43.33	41.11
Ours vs. Competitors	MEAD-Test-B			
	Lip Sync	Realism	Style Consistency	
Ours vs. Imitator* [33]	71.00	83.00	68.00	
Ours vs. Mimic [11]	74.00	78.00	71.00	
Ours vs. GT	42.00	39.00	46.00	

Table 3. User study results on BIWI-Test-B, VOCA-Test, and MEAD-Test-B.

ticipants with good vision and perception to complete the study. Each pair is judged by at least 3 different participants, and 450, 450, 300 entries are collected for BIWI-Test-B, VOCASET-Test, and MEAD-Test-B. The percentage results indicate that our model achieves the best perceptual performance in terms of lip synchronization, realism, and style consistency, as shown in Table 3.

4.5. Ablation Studies

We conduct ablation experiments to assess the effectiveness of different modules and loss constraints, as shown in Table 4.

Constraints. We observe a marked increase in all metrics after removing the trend loss, suggesting that motion trends over different time intervals encapsulate deep motion information and significantly enhance model performance. In contrast, applying the velocity loss from previous work yields only marginal improvements. Similarly, removing the local contrastive loss results in notable metric increases, highlighting its role in aligning audio with synthesized motions. The comparisons with the contrastive loss demonstrate the necessity of our added local constraints.

Audio Enhancement. Audio enhancement further refines the extracted style by capturing motion preference details

Method	LVE ↓ ($\times 10^{-4}$ mm)	FVE ↓ ($\times 10^{-5}$ mm)	FDTW ↓	FFE ↓ ($\times 10^{-2}$)
Full	3.8036	6.3635	0.995	1.26
w/o \mathcal{L}_{tre}	3.9428	6.5712	1.008	1.33
with $\mathcal{L}_{\text{tre}}(R = 1)$	3.9236	6.5653	1.007	1.33
w/o $\mathcal{L}_{\text{icon}}$	3.9092	6.5564	1.006	1.31
with contrastive loss	3.8801	6.5028	1.005	1.30
w/o S_{a} enhancement	3.8943	6.4541	1.004	1.28
w/o style primitives	4.1017	6.7258	1.026	1.33
Full	2.0236	2.1251	1.422	3.10
w/o style primitives	2.1847	2.1973	1.429	3.30

Table 4. Ablation study results on BIWI-Test-A (top 7 rows) and MEAD-Test-B (bottom 2 rows).

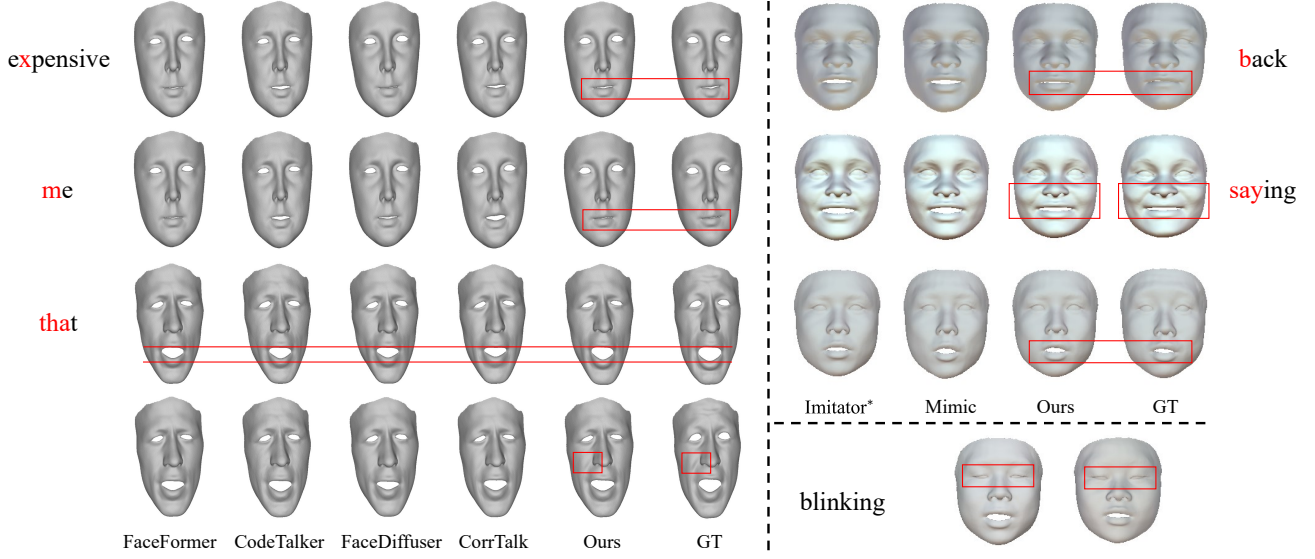


Figure 4. Visual comparisons with competing methods on BIWI-Test-A (left) and MEAD-Test-B (right).

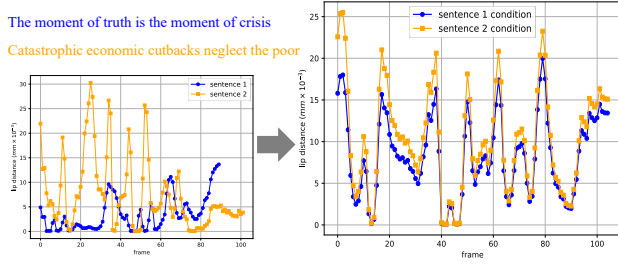


Figure 5. Comparisons of lip distance in synthesized animations under two audio conditions for the same speaker.

specific to the given speech condition based on the speaker style, making style information more accurate and enhancing the motion details in the synthesized facial animation. We select two speech samples from the same speaker in MEAD-Test-A, which correspond to large and slight lip articulations respectively, to enhance the speaker style. Based on these two speaking styles, we synthesize facial animations driven by the same speech and plot the lower-upper lip distances across frames for both styles in Figure 5, from which we observe the subtle impact of the audio conditions on lip motions.

Style Primitives. The style primitives decompose the speaking styles into more fundamental information, which significantly enhances the model’s style learning and adaptation capabilities. Results in Table 4 demonstrate the crucial role of style primitives for both seen and unseen speakers. We present comparative examples of model outputs with and without style primitives in Figure 6. Style primitives enhance large-scale style learning, such as the degree

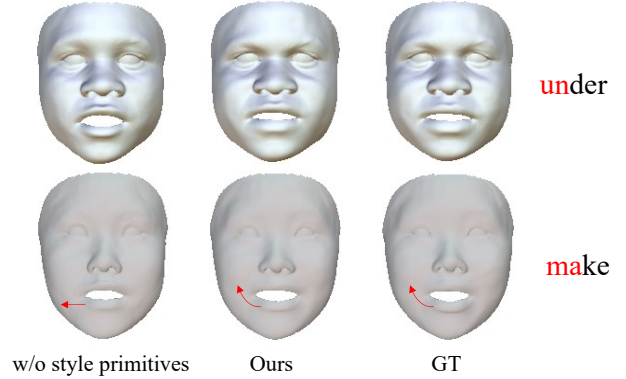


Figure 6. Visualization of the ablation study on the effect of the style primitives.

of mouth opening, while also refining small-scale details, such as the direction of mouth corners, resulting in more expressive and precise facial animation.

5. Conclusion

In this paper, we propose StyleSpeaker, which achieves fine-grained style learning while accounting for audio-induced style biases, endowing our model with strong capabilities in style modeling and adaptation. Additionally, our proposed trend loss and local contrastive loss exhibit effectiveness in improving model performance. Extensive experiments demonstrate that our model outperforms existing state-of-the-art methods in accuracy and style consistency on both seen and unseen speakers. However, the potential

of our model for style modeling may not have been fully realized due to the limited variation in speech conditions within the dataset. In future work, we will focus on enhancing style modeling under varying conditions and emotions.

References

- [1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4432–4441, 2019. 3
- [2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8296–8305, 2020. 3
- [3] Junming Chen, Yunfei Liu, Jianan Wang, Ailing Zeng, Yu Li, and Qifeng Chen. Diffshg: A diffusion-based approach for real-time speech-driven holistic 3d expression and gesture generation. In *CVPR*, 2024. 2
- [4] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022. 3
- [5] Zhaojie Chu, Kailing Guo, Xiaofen Xing, Yilin Lan, Bolun Cai, and Xiangmin Xu. Corrtalk: Correlation between hierarchical speech and facial activity variances for 3d animation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 2, 6, 7
- [6] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10101–10111, 2019. 2, 6
- [7] Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. Jali: an animator-centric viseme model for expressive lip synchronization. *ACM Transactions on graphics (TOG)*, 35(4):1–11, 2016. 2
- [8] Han EunGi, Oh Hyun-Bin, Kim Sung-Bin, Corentin Nivelet Etcheberry, Suekyeong Nam, Janghoon Joo, and Tae-Hyun Oh. Enhancing speech-driven 3d facial animation with audio-visual guidance from lip reading expert. *arXiv preprint arXiv:2407.01034*, 2024. 3
- [9] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18770–18780, 2022. 2, 6, 7
- [10] Gabriele Fanelli, Juergen Gall, Harald Romsdorfer, Thibaut Weise, and Luc Van Gool. A 3-d audio-visual corpus of affective communication. *IEEE Transactions on Multimedia*, 12(6):591–598, 2010. 6
- [11] Hui Fu, Zeqing Wang, Ke Gong, Keze Wang, Tianshui Chen, Haojie Li, Haifeng Zeng, and Wenxiong Kang. Mimic: Speaking style disentanglement for speech-driven 3d facial animation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1770–1777, 2024. 2, 6, 7
- [12] Jiazhi Guan, Zhanwang Zhang, Hang Zhou, Tianshu HU, Kaisiyuan Wang, Dongliang He, Haocheng Feng, Jingtuo Liu, Errui Ding, Ziwei Liu, and Jingdong Wang. Stylesync: High-fidelity generalized and personalized lip sync in style-based generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3
- [13] Jewoong Hwang and Kyoungju Park. Audio-driven facial animation: A survey. In *2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*, pages 614–617. IEEE, 2022. 2
- [14] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14080–14089, 2021. 3
- [15] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 3
- [16] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (ToG)*, 36(4):1–12, 2017. 2
- [17] Biwen Lei, Jianqiang Ren, Mengyang Feng, Miaomiao Cui, and Xuansong Xie. A hierarchical representation network for accurate and detailed face reconstruction from in-the-wild images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 394–403, 2023. 6
- [18] Borong Liang, Yan Pan, Zhizhi Guo, Hang Zhou, Zhibin Hong, Xiaoguang Han, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. Expressive talking head generation with granular audio-visual control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3387–3396, 2022. 3
- [19] Xiangyu Liang, Wenlin Zhuang, Tianyong Wang, Guangxing Geng, Guangyue Geng, Haifeng Xia, and Siyu Xia. Cstalk: Correlation supervised speech-driven 3d emotional facial animation generation. *arXiv preprint arXiv:2404.18604*, 2024. 2
- [20] Zhiyuan Ma, Xiangyu Zhu, Guojun Qi, Chen Qian, Zhaoxiang Zhang, and Zhen Lei. Diffspeaker: Speech-driven 3d facial animation with diffusion transformer. *arXiv preprint arXiv:2402.05712*, 2024. 2
- [21] DW Massaro, MM Cohen, M Tabain, J Beskow, and R Clark. Animated speech: research progress and applications. *Audiovisual Speech Processing*, pages 309–345, 2012. 2
- [22] Wesley Matthyses and Werner Verhelst. Audiovisual speech synthesis: An overview of the state-of-the-art. *Speech Communication*, 66:182–217, 2015. 2
- [23] Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20687–20697, 2023. 2

- [24] Ziqiao Peng, Wentao Hu, Yue Shi, Xiangyu Zhu, Xiaomei Zhang, Hao Zhao, Jun He, Hongyan Liu, and Zhaoxin Fan. Synctalk: The devil is in the synchronization for talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 666–676, 2024. [3](#)
- [25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. [2](#), [5](#)
- [26] Alexander Richard, Colin Lea, Shugao Ma, Jorgen Gall, Fernando De la Torre, and Yaser Sheikh. Audio-and gaze-driven facial animation of codec avatars. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 41–50, 2021. [2](#)
- [27] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1173–1182, 2021. [2](#)
- [28] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on graphics (TOG)*, 42(1):1–13, 2022. [3](#)
- [29] Sanjana Sinha, Sandika Biswas, Ravindra Yadav, and Brojeshwar Bhowmick. Emotion-controllable generalized talking face generation. *arXiv preprint arXiv:2205.01155*, 2022. [3](#)
- [30] Wenfeng Song, Xuan Wang, Shi Zheng, Shuai Li, Aimin Hao, and Xia Hou. Talkingstyle: Personalized speech-driven 3d facial animation with style preservation. *IEEE Transactions on Visualization and Computer Graphics*, 2024. [2](#)
- [31] Stefan Stan, Kazi Injamamul Haque, and Zerrin Yumak. Facediffuser: Speech-driven 3d facial animation synthesis using diffusion. In *Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games*, pages 1–11, 2023. [2](#), [6](#), [7](#)
- [32] Zhiyao Sun, Tian Lv, Sheng Ye, Matthieu Lin, Jenny Sheng, Yu-Hui Wen, Mingjing Yu, and Yong-jin Liu. Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models. *ACM Transactions on Graphics (TOG)*, 43(4):1–9, 2024. [2](#)
- [33] Balamurugan Thambiraja, Ikhsanul Habibie, Sadegh Aliakbarian, Darren Cosker, Christian Theobalt, and Justus Thies. Imitator: Personalized speech-driven 3d facial animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20621–20631, 2023. [2](#), [6](#), [7](#)
- [34] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. [3](#)
- [35] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*, pages 700–717. Springer, 2020. [3](#), [6](#)
- [36] Suzhen Wang, Yifeng Ma, Yu Ding, Zhipeng Hu, Changjie Fan, Tangjie Lv, Zhidong Deng, and Xin Yu. Styletalk++: A unified framework for controlling the speaking styles of talking heads. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. [3](#)
- [37] Haozhe Wu, Songtao Zhou, Jia Jia, Junliang Xing, Qi Wen, and Xiang Wen. Speech-driven 3d face animation with composite and regional facial movements. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6822–6830, 2023. [2](#)
- [38] Sichun Wu, Kazi Injamamul Haque, and Zerrin Yumak. Probtalk3d: Non-deterministic emotion controllable speech-driven 3d facial animation synthesis using vq-vae. In *The 17th ACM SIGGRAPH Conference on Motion, Interaction, and Games (MIG '24)*, November 21–23, 2024, Arlington, VA, USA, New York, NY, USA, 2024. ACM. [3](#)
- [39] Sijing Wu, Yunhao Li, Yichao Yan, Huiyu Duan, Ziwei Liu, and Guangtao Zhai. Mmhead: Towards fine-grained multi-modal 3d facial animation. *arXiv preprint arXiv:2410.07757*, 2024. [3](#)
- [40] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12780–12790, 2023. [2](#), [3](#), [6](#), [7](#)
- [41] Karren D Yang, Anurag Ranjan, Jen-Hao Rick Chang, Raviteja Vemulapalli, and Oncel Tuzel. Probabilistic speech-driven 3d facial motion synthesis: New benchmarks methods and applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27294–27303, 2024. [3](#)
- [42] Qingcheng Zhao, Pengyu Long, Qixuan Zhang, Dafei Qin, Han Liang, Longwen Zhang, Yingliang Zhang, Jingyi Yu, and Lan Xu. Media2face: Co-speech facial animation generation with multi-modality guidance. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–13, 2024. [3](#)