DiffPoseTalk: Speech-Driven Stylistic 3D Facial Animation and Head Pose Generation via Diffusion Models

ZHIYAO SUN, TIAN LV, SHENG YE, MATTHIEU LIN, and JENNY SHENG, BNRist, Tsinghua University, China
YU-HUI WEN*, Beijing Jiaotong University, China
MINJING YU, Tianjin University, China
YONG-JIN LIU*, BNRist, Tsinghua University, China

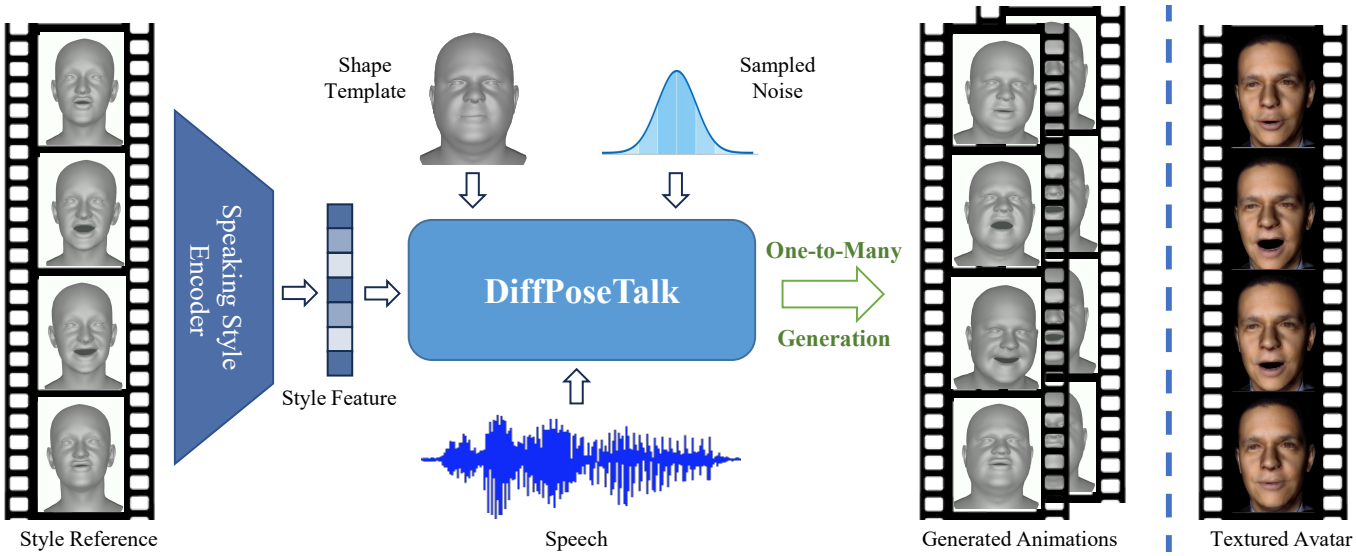

Fig. 1. We present DiffPoseTalk, a novel diffusion-based speech-driven animation system incorporated with a speaking style encoder to extract style features from arbitrary reference videos. Given an input speech and a speaking style, our system generates diverse and stylistic facial animations along with head movements.

The generation of stylistic 3D facial animations driven by speech presents a significant challenge as it requires learning a many-to-many mapping between speech, style, and the corresponding natural facial motion. However, existing methods either employ a deterministic model for speech-to-motion mapping or encode the style using a one-hot encoding scheme. Notably, the one-hot encoding approach fails to capture the complexity of the style and thus limits generalization ability. In this paper, we propose DiffPoseTalk, a generative framework based on the diffusion model combined with a style encoder that extracts style embeddings from short reference videos. During inference, we employ classifier-free guidance to guide the generation process based on the speech and style. In particular, our style includes the generation of head poses, thereby enhancing user perception. Additionally, we address the shortage of scanned 3D talking face data by training our model on reconstructed 3DMM parameters from a high-quality, in-the-wild audio-visual dataset. Extensive experiments and user study demonstrate that our approach outperforms state-of-the-art methods. The code and dataset are at https://diffposetalk.github.io.

*Corresponding authors.

Authors' addresses: Zhiyao Sun; Tian Lv; Sheng Ye; Matthieu Lin; Jenny Sheng, BNRist, Tsinghua University, Beijing, 100084, China, sunzy21@mails.tsinghua.edu.cn, lt22@mails.tsinghua.edu.cn, yec22@mails.tsinghua.edu.cn, yh-lin21@mails.tsinghua.edu.cn, cqq22@mails.tsinghua.edu.cn; Yu-Hui Wen, Beijing Jiaotong University, Beijing, 100044, China, yhwen1@bjtu.edu.cn; Minjing Yu, Tianjin University, Tianjin, 300072, China, minjingyu@tju.edu.cn; Yong-Jin Liu, BNRist, Tsinghua University, Beijing, 100084, China, liuyongjin@tsinghua.edu.cn.

CCS Concepts: • **Computing methodologies** → **Computer graphics**.

Additional Key Words and Phrases: Speech-driven animation, facial animation, diffusion models

1 INTRODUCTION

The domain of speech-driven 3D facial animation has experienced significant growth in both academia and industry, primarily owing to its diverse applications in education, entertainment, and virtual reality. Speech-driven 3D facial animation generates lip-synchronized facial expressions from an arbitrary speech signal. It is a highly challenging research problem due to the cross-modal many-to-many mapping between the speech and the 3D facial animation. However, most existing speech-driven 3D facial animation methods rely on deterministic models [Cudeiro et al. 2019; Fan et al. 2022; Richard et al. 2021; Xing et al. 2023], which often fail to sufficiently capture the complex many-to-many relationships and suffer from the regression-to-mean problem, thereby resulting in over-smoothed face motions. Furthermore, these methods generally employ a one-hot encoding scheme for representing speaking styles during training, thus limiting their adaptability to new speaking styles.

In contrast to deterministic models, diffusion models can fit various forms of distributions, making them better suited to addressing the many-to-many mapping challenge. Recent diffusion models

have shown impressive results in various domains [Yang et al. 2023]. Specifically, the existing diffusion-based audio-driven human motion generation methods have shown appealing results. However, they are not trivially transferable to speech-driven facial animation for three main reasons. First, unlike gestures, which can have a more relaxed temporal correlation with audio (occurring slightly before or after the associated speech), facial movements — particularly lip motions — require much stricter timing. This calls for specifically designed structures to precisely align speech and motion information. Second, lip motions contain richer semantics than gestures or dancing, which necessitate a more robust speech encoder to extract phoneme-level features. Lastly, humans have diverse speaking styles. A strong style encoder should be designed to extract style representation from an arbitrary style clip.

To address the aforementioned limitations and challenges, we introduce DiffPoseTalk, a novel controllable diffusion-based generative model, to generate high-quality, diverse, speech-matched, and stylistic facial motions for speech-driven 3D facial animation (Figure 1). Our method overcomes the inability of existing diffusion models that cannot be directly transferred to speech-driven expression animation. Compared to existing methods, the main improvement of DiffPoseTalk can be characterized as follows. We use an attention-based architecture to align facial motions with speech, and train a diffusion model to predict the facial expression signal itself [Ramesh et al. 2022a; Tevet et al. 2023] instead of predicting the noise; this architecture allows us to facilitate the subsequent addition of geometric constraints to generate more reasonable results. Along with the expression, we also predict the speaker's head pose and design the corresponding loss function to obtain more natural animations. Furthermore, we exploit HuBERT [Hsu et al. 2021] to encode the input speech to improve generalization and robustness. Finally, we develop a style encoder to obtain latent style code from a style video clip, and perform classifier-free guidance [Ho and Salimans 2022] at inference time to achieve example-based style control. To address the scarcity of co-speech 3D facial animation data by motion capture, we collect and build a speech-driven facial animation dataset with varied speaking styles and head poses.

In summary, our contribution is threefold:

- We propose a novel diffusion-based approach to jointly generate diverse and stylistic 3D facial motions with head poses from speech.
- We develop a style encoder to extract personalized speaking styles from reference videos, which can be used to guide the motion generation in a classifier-free manner.
- We build a new audio-visual dataset that encompasses a diverse range of identities, speaking styles, and head poses. This dataset and our code are available for research purposes.

2 RELATED WORK

2.1 Speech-Driven 3D Facial Animation

Existing speech-driven 3D facial animation methods can be roughly divided into procedural and learning-based methods. Procedural approaches generally segment speech into phonemes, which are then mapped to predefined visemes via a set of comprehensive rules. For example, Cohen et al. [2001] use dominance functions to map phonemes to corresponding facial animation parameters, while Edwards et al. [2016] factor speech-related animations into jaw and lip actions, employing a co-articulation model to animate facial rigs. Although these procedural methods offer explicit control over the resulting animations, they often require intricate parameter tuning and lack the ability to capture the diversity of real-world speaking styles.

Meanwhile, learning-based methods have grown rapidly over the recent decade. These approaches typically adopt acoustic features like MFCC or pretrained speech model features [Baevski et al. 2020; Hannun et al. 2014; Hsu et al. 2021] as the speech representation, which is then mapped to 3D morphable model parameters [Peng et al. 2023; Zhang et al. 2023b] or 3D mesh [Cudeiro et al. 2019; Fan et al. 2022; Haque and Yumak 2023; Richard et al. 2021; Xing et al. 2023] through neural networks. However, most current methods are regression models, which are deterministic and tend to generate over-smoothed lip motion. This issue is especially pronounced in large-scale datasets, where these methods are prone to yielding average outcomes, thereby struggling to produce precise and diverse responses to voice data. Typically, these deterministic approaches have been trained on smaller datasets like VOCA [Cudeiro et al. 2019] and BIWI [Fanelli et al. 2013], which unintentionally sidestep the regression-to-mean challenge. However, when these methods are applied to the more extensive dataset used in our study, a significant decline in performance is observed. In contrast, our method proposed in this paper leverages the strong probabilistic modeling capability of diffusion models to generate diverse and stylistic 3D facial animations.

Current learning-based methods typically achieve style control in a label-based or example-based manner. The former class relies on manually predefined style labels. For example, Cudeiro et al. [2019] and Fan et al. [2022] employ one-hot embeddings as the style label for different identities in the training set. However, this limits the model's ability to adapt to new individuals and capture complex fine-grained styles. The latter class (e.g., Imitator [Thambiraja et al. 2023]) is able to generate talking animation in arbitrary styles, even those unseen during training, by imitating examples. Our method falls into this category. Different from Imitator, our method adopts contrastive learning to extract salient style features and does not require optimization or fine-tuning.

2.2 Diffusion Probabilistic Models

Diffusion probabilistic models [Ho et al. 2020; Sohl-Dickstein et al. 2015], which are able to generate high-quality and diverse samples, have achieved remarkable results in various generative tasks [Yang et al. 2023]. They leverage a stochastic diffusion process to gradually add noise to data samples, subsequently employing neural architectures to reverse the process and denoise the samples. A key strength of diffusion models lies in their ability to model various forms of distributions and capture complex many-to-many relationships, making them particularly well-suited for our speech-driven facial animation task.

For conditional generation, classifier-guidance [Dhariwal and Nichol 2021] and classifier-free guidance [Ho and Salimans 2022]

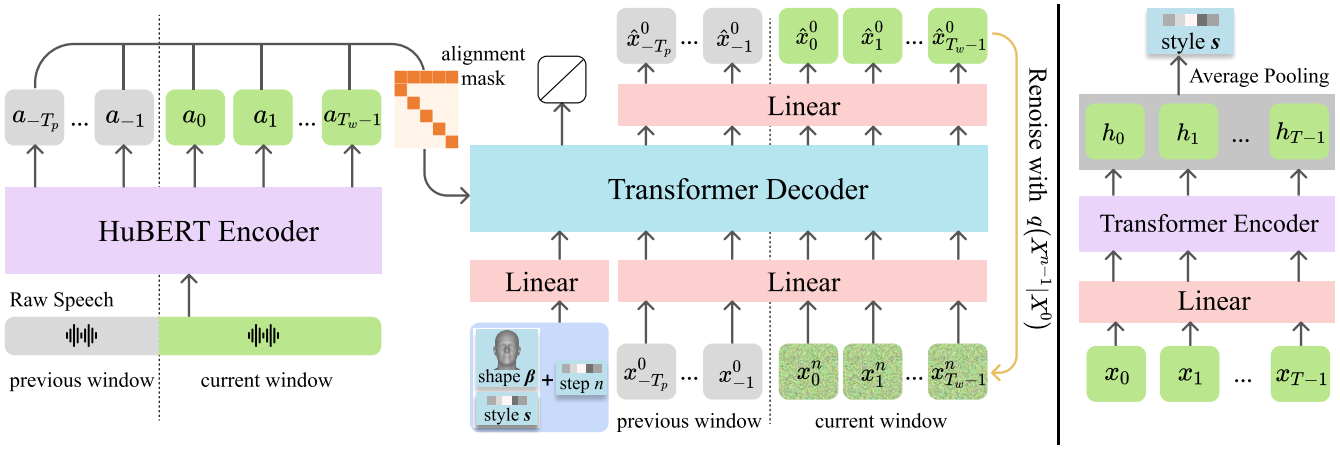

Fig. 2. **(Left) Transformer-based denoising network.** We employ a windowing strategy to generate speech-driven 3D facial animations for inputs of arbitrary length. HuBERT-encoded speech features $\boldsymbol{A}_{-T_p:T_w}$, prior clean motion parameters $\boldsymbol{X}^0_{-T_p:0}$, current noisy motion parameters $\boldsymbol{X}^n_{0:T_w}$, shape parameters $\boldsymbol{\beta}$, style feature $\boldsymbol{s}$, and the diffusion timestep n are fed into the transformer decoder. The decoder then predicts clean motion parameters $\hat{\boldsymbol{X}}^0_{-T_p:T_w}$, which are renoised to $\boldsymbol{X}^{n-1}_{0:T_w}$ for the next denoising iteration. **(Right) The speaking style encoder.** The style feature $\boldsymbol{s}$ can be extracted from a sequence of motion parameters $\boldsymbol{X}_{0:T}$ using a transformer encoder.

are widely employed in tasks such as text-to-image synthesis [Rombach et al. 2022], text-to-motion [Tevet et al. 2023] synthesis, and audio-driven body motion generation [Alexanderson et al. 2023; Zhu et al. 2023]. In particular, diffusion model with classifier-free guidance has achieved impressive results in multi-modal modeling. Recently, diffusion models have also been applied in speech-driven 3D facial animation. FaceDiffuser [Stan et al. 2023] leverages the non-deterministic capabilities of diffusion models to capture the complex many-to-many relationship between audio and face motions. Nevertheless, it lacks the ability to control the head poses of the generated talking faces and does not support novel style conditions. In this paper, we propose a novel diffusion-based model that jointly generates facial motions and head poses while accommodating arbitrary novel style conditions.

3 METHOD

An overview of our proposed method is illustrated in Figure 2. We adopt a well-established, pretrained encoder to extract speech features, while using 3DMM as the face representation (Section 3.1). A transformer-based denoising network is used for the reverse diffusion process (Section 3.2), where we guide the conditional generation in a classifier-free manner (Section 3.3).

3.1 Problem Formulation

Our method takes a speech feature[1] $\boldsymbol{A}_{0:T}$, the 3DMM shape parameter $\boldsymbol{\beta}$ of a template face, and a speaking style vector $\boldsymbol{s}$ as input, and generates a 3DMM-based 3D facial animation represented by a sequence of 3DMM expression and pose parameters $\boldsymbol{X}_{0:T}$. The style vector $\boldsymbol{s}$ can be extracted from a short reference video using our speaking style encoder (Section 3.3.1).

[1]We use Python style indexing and slicing in this paper, *i. e.*, "$0 : T$" includes "$0, 1, \ldots, T-1$".

Speech Representation. Extensive facial animation studies have shown that self-supervised pretrained speech model features like Wav2Vec2 [Baevski et al. 2020] and HuBERT [Hsu et al. 2021] outperform traditional ones such as MFCC. Based on these findings, we utilize HuBERT as our chosen speech encoder for facial animation generation, as it has been proven to be superior to Wav2Vec2 in this regard [Haque and Yumak 2023]. HuBERT consists of a temporal convolutional audio feature extractor and a multi-layer transformer encoder. To align the audio features with the facial motions' sampling rate, we introduce a resampling layer after the temporal convolutions. Thus, for a given raw audio clip that matches a facial motion sequence of length T, HuBERT generates a speech representation $\boldsymbol{A}_{0:T}$ that also spans T time steps.

3D Face Representation. We use a 3D morphable model FLAME [Li et al. 2017] with $N = 5,023$ vertices and $K = 4$ joints, whose geometry can be represented using parameters $\{\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\theta}\}$, where $\boldsymbol{\beta} \in \mathbb{R}^{100}$ is the shape parameter, $\boldsymbol{\psi} \in \mathbb{R}^{50}$ is the expression parameter, and $\boldsymbol{\theta} \in \mathbb{R}^{3K+3}$ is the head pose parameter. Given a set of FLAME parameters, the 3D face mesh can be obtained with $M(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}) = W\left(T_P(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}), \mathbf{J}(\boldsymbol{\beta}), \boldsymbol{\theta}, \mathcal{W}\right)$, where T_P outputs vertices by combining blend shapes, the standard skinning function $W(\mathbf{T}, \mathbf{J}, \boldsymbol{\theta}, \mathcal{W})$ rotates the vertices of $\mathbf{T}$ around joints $\mathbf{J}$, and $\mathcal{W}$ performs linear smoothing. Specifically, we use the shape parameter $\boldsymbol{\beta}$ to serve as the neutral template face for the speaker. For facial animation, we predict the expression parameters $\boldsymbol{\psi}$ as well as the jaw and global rotation components within the pose parameters $\boldsymbol{\theta}$. To simplify notation, we compose $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$ as the *motion parameter* $\boldsymbol{x}$ and rewrite the mesh construction function as $M(\boldsymbol{\beta}, \boldsymbol{x})$. The reasons for choosing 3DMM parameters over mesh vertices as the face representation are discussed in Section 5.

To reconstruct accurate 3DMM parameters from the audio-visual dataset, we make comprehensive use of several state-of-the-art 3D face reconstruction and pose estimation works, similar to Daněček

et al. [2023]. We adopt MICA [Zielonka et al. 2022] for identity shape prediction, SPECTRE [Filntisis et al. 2022] for accurate expression reconstruction of lip movements and jaw pose prediction, and 6DRepNet [Hempel et al. 2022] for head pose prediction. Then, we follow EmoTalk [Peng et al. 2023] to apply a Savitzky-Golay filter to the predicted expressions and poses, which markedly improves motion smoothness.

3.2 Facial Animation Diffusion Model

We propose to use a diffusion model to generate speech-driven facial animation. The diffusion model involves two processes. The forward process is a Markov chain $q\left(\boldsymbol{X}^n|\boldsymbol{X}^{n-1}\right)$ for $n \in \{1,\ldots,N\}$ that progressively adds Gaussian noise to an initial data sample $\boldsymbol{X}^0$ according to a variance schedule. The original sample is gradually substituted by noises, eventually reaching a standard normal distribution $q\left(\boldsymbol{X}^N|\boldsymbol{X}^0\right)$. The reverse process, on the contrary, leverages the distribution $q\left(\boldsymbol{X}^{n-1}|\boldsymbol{X}^n\right)$ to recreate the sample from noise. This distribution depends on the entire dataset and hence is intractable. Therefore, a denoising network is used to approximate this distribution. In practice, the denoising network is trained to predict the noise [Ho et al. 2020] or the clean sample $\boldsymbol{X}^0$ [Ramesh et al. 2022b]. We opt for the latter, as it enables us to incorporate geometric losses that offer more precise constraints on facial motions. The effectiveness of this scheme has been validated by prior works on human body motion generation [Tevet et al. 2023] and our experiments.

Transformer-Based Denoising Network. Our transformer-based denoising network, as illustrated in Figure 2, consists of two components: a pretrained HuBERT encoder for extracting speech features $\boldsymbol{A}$, and a transformer decoder for sampling predicted motions $\hat{\boldsymbol{X}}^0$ from noisy observations $\boldsymbol{X}^n$ $(n = N, N-1,\ldots,1)$ in an iterative manner. A notable design is an alignment mask between the encoder and the decoder [Fan et al. 2022], which ensures proper alignment of the speech and motion modalities. Specifically, the motion feature at position t only attends to the speech feature $\boldsymbol{a}_t$. The initial token, which is composed of diffusion timestep n and other conditions, attends to all speech features. We allow the transformer part of the HuBERT speech encoder to be trainable, which enables HuBERT to better capture motion information directly from speech. To accommodate sequences of arbitrary lengths, we implement a windowing strategy for the inputs.

Formally, the inputs to the denoising network are processed as follows. For a given speech feature sequence of length T, we partition it into windows of length T_w (padding is added to the audio if it is not long enough). To ensure seamless transitions between consecutive windows, we include the last T_p frames of speech features $\boldsymbol{A}_{-T_p:0}$ and motion parameters $\boldsymbol{X}^0_{-T_p:0}$ from the preceding window as conditional inputs. Note that for the first window, the speech features and motion parameters are replaced with learnable start features $\boldsymbol{A}_{\text{start}}$ and $\boldsymbol{X}_{\text{start}}$. Within each window at diffusion step n, the network receives both previous and current speech features $\boldsymbol{A}_{-T_p:T_w}$, the previous motion parameters $\boldsymbol{X}^0_{-T_p:0}$, and the current noisy motion parameters $\boldsymbol{X}^n_{0:T_w}$ sampled from $q\left(\boldsymbol{X}^n_{0:T_w}|\boldsymbol{X}^0_{0:T_w}\right)$. The denoising network then outputs the clean sample as:

$$\hat{\boldsymbol{X}}^0_{-T_p:T_w} = D\left(\boldsymbol{X}^n_{0:T_w}, \boldsymbol{X}^0_{-T_p:0}, \boldsymbol{A}_{-T_p:T_w}, n\right). \tag{1}$$

Losses. We use the simple loss [Ho et al. 2020] for the predicted sample:

$$\mathcal{L}_{\text{simple}} = \left\|\hat{\boldsymbol{X}}^0_{-T_p:T_w} - \boldsymbol{X}^0_{-T_p:T_w}\right\|^2. \tag{2}$$

To better constrain the generated face motion, we convert the FLAME parameters into zero-head-posed 3D mesh sequences. Formally, $\boldsymbol{M}_{-T_p:T_w} = M_0\left(\boldsymbol{\beta}, \boldsymbol{X}^0_{-T_p:T_w}\right)$ and $\hat{\boldsymbol{M}}_{-T_p:T_w} = M_0\left(\boldsymbol{\beta}, \hat{\boldsymbol{X}}^0_{-T_p:T_w}\right)$. We then apply the following geometric losses in 3D space: the vertex loss $\mathcal{L}_{\text{vert}}$ [Cudeiro et al. 2019] for the positions of the mesh vertices, the velocity loss $\mathcal{L}_{\text{vel}}$ [Cudeiro et al. 2019] for better temporal consistency, and a smooth loss $\mathcal{L}_{\text{smooth}}$ to penalize large acceleration of the predicted vertices:

$$\mathcal{L}_{\text{vert}} = \left\|\boldsymbol{M}_{-T_p:T_w} - \hat{\boldsymbol{M}}_{-T_p:T_w}\right\|^2, \tag{3}$$

$$\mathcal{L}_{\text{vel}} = \left\|\left(\boldsymbol{M}_{-T_p+1:T_w} - \boldsymbol{M}_{-T_p:T_w-1}\right) - \left(\hat{\boldsymbol{M}}_{-T_p+1:T_w} - \hat{\boldsymbol{M}}_{-T_p:T_w-1}\right)\right\|^2, \tag{4}$$

$$\mathcal{L}_{\text{smooth}} = \left\|\hat{\boldsymbol{M}}_{-T_p+2:T_w} - 2\hat{\boldsymbol{M}}_{-T_p+1:T_w-1} + \hat{\boldsymbol{M}}_{-T_p:T_w-2}\right\|^2. \tag{5}$$

We apply geometric losses $\mathcal{L}_{\text{head}}$ to head motions in a similar way. Please refer to the Appendix for more details.

In summary, our overall loss is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{simple}} + \lambda_{\text{vert}}\mathcal{L}_{\text{vert}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}} + \mathcal{L}_{\text{head}}. \tag{6}$$

3.3 Style-Controllable Diffusion Model

The facial animation diffusion model generates facial motions conditioned on input speech. In addition to the speech, we use speaking style and template face shape as control conditions. The shape parameters $\boldsymbol{\beta}$ and speaking style feature $\boldsymbol{s}$ are shared across all windows. The denoising network then outputs the clean sample as:

$$\hat{\boldsymbol{X}}^0_{-T_p:T_w} = D\left(\boldsymbol{X}^n_{0:T_w}, \boldsymbol{X}^0_{-T_p:0}, \boldsymbol{A}_{-T_p:T_w}, \boldsymbol{s}, \boldsymbol{\beta}, n\right). \tag{7}$$

3.3.1 Speaking Style Encoder. We introduce a novel speaking style encoder designed to capture the unique speaking style of a given speaker from a brief video clip. Speaking style is a multifaceted attribute that manifests in various aspects such as the size of the mouth opening [Cudeiro et al. 2019], facial expression dynamics — especially in the upper face [Xing et al. 2023] — and head movement patterns [Yi et al. 2023; Zhang et al. 2023b]. Given the complexity and difficulty in quantitatively describing speaking styles, we opt for an implicit learning approach through contrastive learning. We operate under the assumption that the short-term speaking styles of the same person at two proximate times should be similar.

Architecture. The speaking style encoder (Figure 2 right) utilizes a transformer encoder to extract style features from a sequence of motion parameters $\boldsymbol{X}_{0:T}$. The encoder features $\{\boldsymbol{h}_i\}$ are aggregated by average pooling into the style embedding $\boldsymbol{s}$. Formally, this is described as:

$$\boldsymbol{s} = SE\left(\boldsymbol{X}_{0:T}\right). \tag{8}$$

Contrastive Learning. We use the NT-Xent loss [Chen et al. 2020] for contrastive learning. Each training minibatch consists of N_s samples of speech features and motion parameters of length $2T$. We

split the sample length in half to get N_s pairs of positive examples. Given a positive pair, the other $2(N_s - 1)$ examples are treated as negative examples. We use cosine similarity as the similarity function. The loss function for a positive pair of examples (i, j) is defined as:

$$\mathcal{L}_{i,j} = -\log \frac{\exp\left(\text{cos\_sim}(\boldsymbol{s}_i, \boldsymbol{s}_j)/\tau\right)}{\sum_{k=1}^{2N_s} \mathbf{1}_{k\neq i} \exp\left(\text{cos\_sim}(\boldsymbol{s}_i, \boldsymbol{s}_k)/\tau\right)}, \tag{9}$$

where $\mathbf{1}_{k\neq i}$ is an indicator function and τ represents a temperature parameter. The overall loss is computed across all positive pairs for both (i, j) and (j, i).

3.3.2 Training Strategy. In our window-based generation approach, our network faces two different scenarios: (a) generating the initial window, where the previous window conditions are learnable start features, and (b) generating subsequent windows, where the conditions are speech features and motions parameters from the preceding window. The network also requires a shape parameter $\boldsymbol{\beta}$ and a speaking style feature $\boldsymbol{s}$ in both scenarios. Therefore, we propose a novel training strategy to meet this demand. Specifically, each training sample includes a frame of shape parameter $\boldsymbol{\beta}$, a speech clip (which will be encoded into speech features $\boldsymbol{A}_{0:2T_w}$ by the HuBERT encoder), and a corresponding motion parameter sequence $\boldsymbol{X}^0_{0:2T_w}$. We partition the sample into two windows and employ the speaking style encoder to derive style features for each, resulting in $(\boldsymbol{s}_a, \boldsymbol{s}_b)$. The tuple $(\boldsymbol{A}_{0:T_w}, \boldsymbol{X}^0_{0:T_w}, \boldsymbol{s}_b)$ is used to train the first window, while $(\boldsymbol{A}_{T_w:2T_w}, \boldsymbol{X}^0_{T_w:2T_w}, \boldsymbol{s}_a)$ with the previous window conditions $(\boldsymbol{A}_{T_w-T_p:T_w}, \boldsymbol{X}^0_{T_w-T_p:T_w})$ is used to train the second window. Taking into account that the actual length of speech during generation may not fully occupy the window, we introduce a random truncation of samples during training. This approach ensures that the model is robust to variations in speech length.

3.3.3 Sampling with Incremental Classifier-Free Guidance. During generation, we sample the result $\boldsymbol{X}^0$ conditioned on $(\boldsymbol{A}, \boldsymbol{s}, \boldsymbol{\beta})$ in an iterative manner. Specifically, we estimate the clean sample as $\hat{\boldsymbol{X}}^0 = D\left(\boldsymbol{X}^n, \boldsymbol{A}, \boldsymbol{s}, \boldsymbol{\beta}, n\right)$ and subsequently reintroduce noise to obtain $\boldsymbol{X}^{n-1}$. This process is repeated for $n = N, N-1, \ldots, 1$, in accordance with the process in Tevet et al. [2023].

Furthermore, we find it beneficial to apply classifier-free guidance [Ho and Salimans 2022] with the incremental scheme [Brooks et al. 2023], which has been successfully applied to image generation from multiple conditions, where

$$\begin{aligned}\hat{\boldsymbol{X}}^0 = D\left(\boldsymbol{X}^n, \emptyset, \emptyset, \boldsymbol{\beta}, n\right) &+ w_a\left[D\left(\boldsymbol{X}^n, \boldsymbol{A}, \emptyset, \boldsymbol{\beta}, n\right) - D\left(\boldsymbol{X}^n, \emptyset, \emptyset, \boldsymbol{\beta}, n\right)\right] \\ &+ w_s\left[D\left(\boldsymbol{X}^n, \boldsymbol{A}, \boldsymbol{s}, \boldsymbol{\beta}, n\right) - D\left(\boldsymbol{X}^n, \boldsymbol{A}, \emptyset, \boldsymbol{\beta}, n\right)\right].\end{aligned} \tag{10}$$

The w_a and w_s are the guidance scales for audio and style, respectively. During training, we randomly set the style condition to $\emptyset$ with 0.45 probability, and set both the audio and style conditions to $\emptyset$ with 0.1 probability.

4 EXPERIMENTS

4.1 Datasets

We introduce a new dataset— Talking Face with Head Poses (TFHP) — which contains 1,052 videos of 588 subjects, totaling 26.5 hours. In TFHP, 348 videos are collected from the downloading script provided by High-Definition Talking Face (HDTF) dataset [Zhang et al. 2021]. Compared with HDTF, our TFHP dataset is more diversified in content, featuring video clips from lectures, online courses, interviews, and news programs, thereby capturing a wider array of speaking styles and head movements. Moreover, all videos are converted to 25 fps. In total, approximately 2,385,000 frames of FLAME parameters are reconstructed from the videos with our carefully designed data processing pipeline as previously mentioned. We split the combined dataset by speakers, resulting in 460 for training, 64 for validation, and 64 for testing.

4.2 Experiment Setup

Implementation Details. We use a four-layer transformer encoder with four attention heads for the speaking style encoder, with feature dimension $d_s = 128$, sequence length $T = 100$ (4 seconds), and temperature $\tau = 0.1$. We train the encoder with the Adam optimizer [Kingma and Ba 2015] for 26k iterations, with a batch size of 32 and a learning rate of 1e−4.

We use an eight-layer transformer decoder with eight attention heads for the denoising network, with the feature dimension $d = 512$, the window length $T_w = 100$, and $T_p = 10$. We adopt a cosine noise schedule with diffusion $N = 500$ steps. We train the denoising network with the Adam optimizer for 90k iterations, using 5k warmup steps, batch size 16, and learning rate 1e−4. We set $\lambda_{\text{vert}} = 2\text{e}6$, $\lambda_{\text{vel}} = 1\text{e}7$, and $\lambda_{\text{smooth}} = 1\text{e}5$ to balance the magnitude of the losses. The overall training on an Nvidia 3090 GPU takes about 12 hours.

Baselines. We compare our approach with state-of-the-art 3D facial animation methods: FaceFormer [Fan et al. 2022], CodeTalker [Xing et al. 2023], and FaceDiffuser [Stan et al. 2023], which are trained with 3D mesh data generated from 3DMM parameters of our TFHP dataset. Recognizing the scarcity of speech-driven 3D animation methods that account for head poses, we also compare with two 2D talking face generation methods, Yi et al. [2023] and SadTalker [Zhang et al. 2023a], which incorporate head movements and utilize a 3DMM as an intermediate face representation. To compare with these two types of methods, we train two versions of our model: "Ours" (with head pose prediction) and "Ours (no HP)" (without head pose prediction).

4.3 Quantitative Evaluation

Following previous studies, we employ two established metrics — lip vertex error (LVE) [Richard et al. 2021] and upper face dynamics deviation (FDD) [Xing et al. 2023] — for the quantitative evaluation of generated facial expressions. LVE measures lip synchronization by calculating the maximum L2 error across all lip vertices for each frame. FDD evaluates the upper face motions, which are closely related to speaking styles, by comparing the standard deviation of each upper face vertex's motion over time between the prediction and the ground truth. To assess head motion, we use beat alignment

Table 1. Quantitative evaluation of the comparative methods, our proposed method, and ablation study variants. We run the evaluation 10 times and report the average score with a 95% confidence interval when applicable. We also report the diversity scores of expression and head pose generation. Note that the vertex-related metrics are not comparable with SadTalker due to its different face topology.

	Methods	LVE (mm) ↓	FDD ($\times 10^{-5}$m) ↓	MOD (mm) ↓	BA ↑	Div (exp) ($\times 10^{-4}$) ↑	Div (HP) ↑
w/o HP	FaceFormer	$9.90^{\pm 0.030}$	$16.95^{\pm 0.016}$	$2.63^{\pm 0.015}$	N/A	0	N/A
	CodeTalker	$12.71^{\pm 0.057}$	$12.44^{\pm 0.064}$	$2.87^{\pm 0.034}$	N/A	0	N/A
	FaceDiffuser	$12.12^{\pm 0.038}$	$15.48^{\pm 0.048}$	$3.50^{\pm 0.052}$	N/A	5.93×10^{-5}	N/A
	Ours (no HP)	$\mathbf{8.81^{\pm 0.008}}$	$10.13^{\pm 0.038}$	$1.72^{\pm 0.009}$	N/A	**2.83**	N/A
w/ HP	Yi et al.	9.99	21.50	2.42	0.26	0	0
	SadTalker	—	—	—	$0.24^{\pm 0.001}$	0	0.808
	Ours	$8.94^{\pm 0.013}$	$9.60^{\pm 0.027}$	$1.62^{\pm 0.009}$	$\mathbf{0.29^{\pm 0.005}}$	**2.19**	**1.16**
Ablations	Ours w/o $\mathcal{L}_{geo}$	$11.29^{\pm 0.012}$	$15.11^{\pm 0.043}$	$2.14^{\pm 0.013}$	$0.28^{\pm 0.009}$	—	—
	Ours w/o AM	$12.81^{\pm 0.011}$	$12.58^{\pm 0.049}$	$2.18^{\pm 0.007}$	$0.24^{\pm 0.006}$	—	—
	Ours w/o CFG	$9.58^{\pm 0.014}$	$\mathbf{9.59^{\pm 0.032}}$	$\mathbf{1.56^{\pm 0.011}}$	$\mathbf{0.29^{\pm 0.010}}$	—	—
	Ours w/o SSE	$11.33^{\pm 0.013}$	$12.97^{\pm 0.052}$	$2.03^{\pm 0.017}$	$0.28^{\pm 0.005}$	—	—

(BA) [Li et al. 2022; Zhang et al. 2023a], albeit with a minor modification: we compute the synchronization of detected head movement beats between the predicted and the actual outcomes. For examining the diversity of facial expressions and head poses generated from *identical* input, we follow Ren et al. [2023] to compute a diversity score. Since the size of the mouth opening can also indicate speaking style [Cudeiro et al. 2019], we introduce a new metric called *mouth opening difference* (MOD), which measures the average difference in the size of the mouth opening between the prediction and ground truth.

We present the quantitative results and the diversity scores in Table 1. Our method outperforms all others across all metrics, achieving the best lip synchronization and head pose beat alignment. Additionally, the FDD, MOD, and BA metrics suggest that our method most effectively captures speaking styles. As for diversity, the other methods employ deterministic approaches for motion generation, with the exception of SadTalker (which samples head poses from a VAE) and FaceDiffuser. However, we find FaceDiffuser produces nearly identical results for the same input. Consequently, these methods are unable to produce varied expressions and head poses from identical inputs, falling short in capturing this many-to-many mapping.

4.4 Qualitative Evaluation

We show the comparison of our method with other comparative methods in Figure 3 and 4. Our method excels in handling challenging cases, such as articulating bilabial consonants and rounded vowels. Moreover, the generated results have the closest speaking style to the ground truth in aspects like upper facial expressions and mouth opening size. Notably, our approach can also spontaneously produce random movements like eye blinks, which are implicitly learned from the data distribution. Our generated head motions also align well with the stress and rhythm in speech, in a way similar to the ground truth. More results can be found in the supplementary demo video.

Table 2. User study results.

Method	Lip Sync ↑	Style Sim ↑	Natural ↑
FaceFormer	2.56	2.60	2.36
CodeTalker	2.88	3.00	2.90
FaceDiffuser	2.71	2.51	2.35
Ours (no HP)	**4.23**	**4.07**	**4.43**
Yi et al.	1.94	2.02	1.99
SadTalker	3.25	2.91	2.96
Ours	**4.52**	**4.25**	**4.43**

4.5 User Study

To conduct a more comprehensive assessment of our approach, we designed a user study with the following experiment setup. The methods to be compared are categorized into two groups based on whether head motion are generated. The group without head motion includes FaceDiffuser, FaceFormer, CodeTalker, and Ours (no HP) and consists of 12 sets of questions. The group with head motion involves Yi et al., SadTalker, and our approach, comprising 8 sets of questions. In each set, participants are shown the ground truth animation as well as animations generated by each method. Participants are then asked to rate on a scale of 1-5 the lip synchronization, similarity to the speaking style of the ground truth, and the naturalness of facial expressions and head movements. For more information about the settings, please refer to the Appendix. Twenty-six participants took part in the study, and the results are presented in Table 2. The results demonstrate that our method significantly outperforms existing works in terms of lip synchronization, similarity to the ground truth speaking style, and the naturalness of facial expressions and head movements.

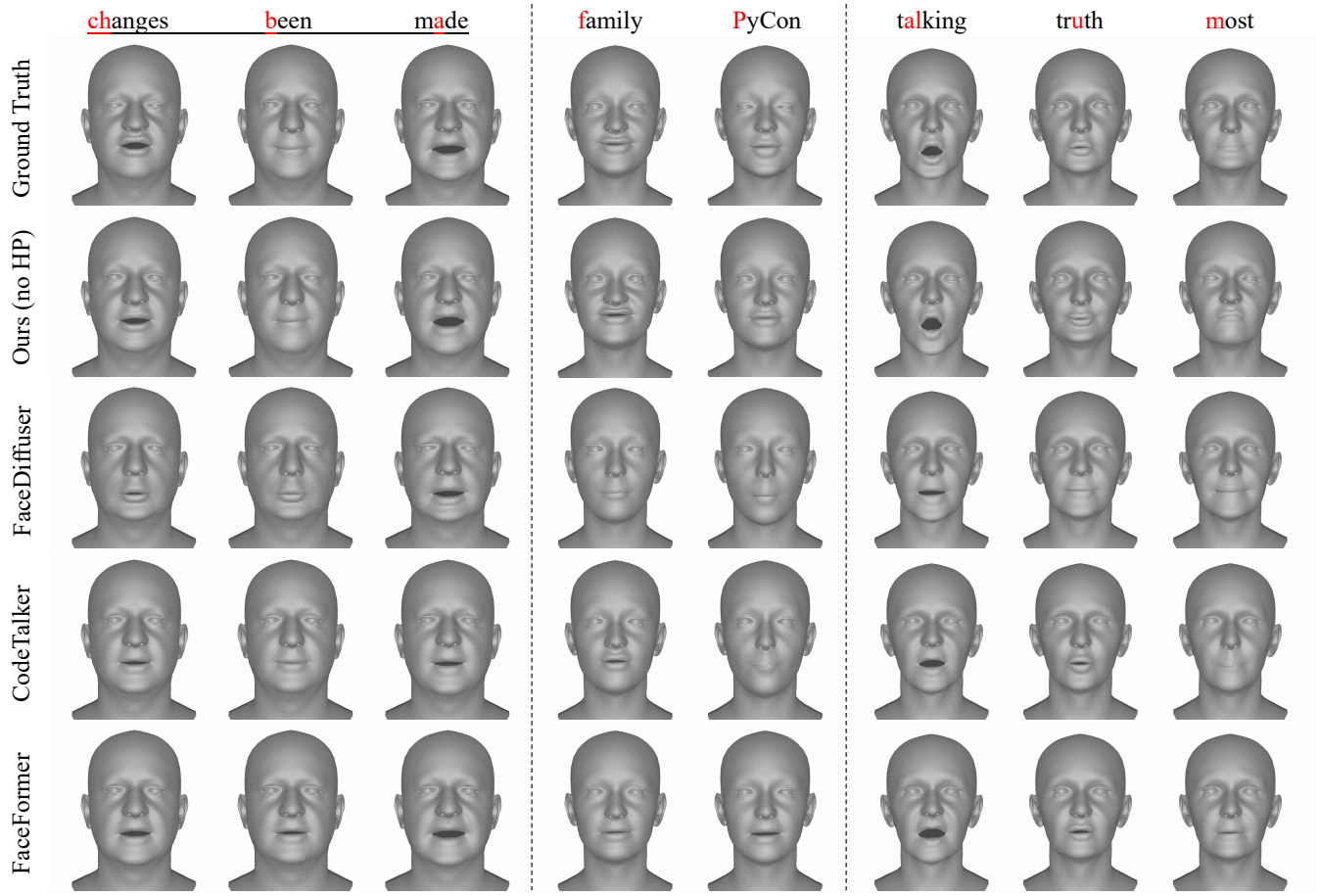

Fig. 3. Qualitative comparison with the state of the arts (w/o head pose prediction). Results of different identities are split by dashed lines.

4.6 Ablation Study

The results are summarized in Table 1, showing that the removal of the speaking style encoder (Ours w/o SSE) leads to a decline in performance across all metrics, as our method is no longer capable of generating animations tailored to the speaking style of individuals. Removing all geometric losses (Ours w/o $\mathcal{L}_{geo}$) results in our method being unable to produce precise facial animations. Removing the alignment mask (Ours w/o AM) causes a serious out-of-sync problem. However, we observe that excluding classifier-free guidance (Ours w/o CFG) yields a mixed impact on the metrics. This is probably due to the fact that CFG is a technique in diffusion models that aims to reduce the diversity of the generated samples while enhancing the quality of each individual sample. Thus, CFG improves LVE for lip synchronization, while slightly diminishing performance on FDD and MOD evaluations, which we speculate are metrics closely related to speaking styles.

5 DISCUSSIONS

Choice of Face Representation. Unlike closely related works [Cudeiro et al. 2019; Fan et al. 2022; Xing et al. 2023] that use 3D mesh vertices, our approach leverages a widely used 3DMM — specifically the FLAME model [Li et al. 2017] — for face representation. There are several reasons for this choice. First, given the computational intensity of diffusion models, the lower-dimensional 3DMM parameters offer a substantial advantage in terms of computational speed when compared to predicting mesh vertices. Second, data collection and coverage present challenges. Capturing real-world 3D mesh data requires professional motion capture systems and considerable investments of time and effort, thereby constraining the scale and diversity of data that can be collected. For example, VOCASET [Cudeiro et al. 2019] only provides less than 30 minutes of data from just 12 subjects. Conversely, 2D audio-visual data are much simpler to collect, and numerous off-the-shelf methods exist for reconstructing 3DMM parameters from these data, offering considerably wider coverage of identities, phonemes, and styles than scanned mesh data. Additionally, the reduced dimensionality, along with the blendshapes as a prior, simplifies the learning process and improves generalization [Peng et al. 2023]. Lastly, using 3DMM parameters facilitates integration with downstream applications such

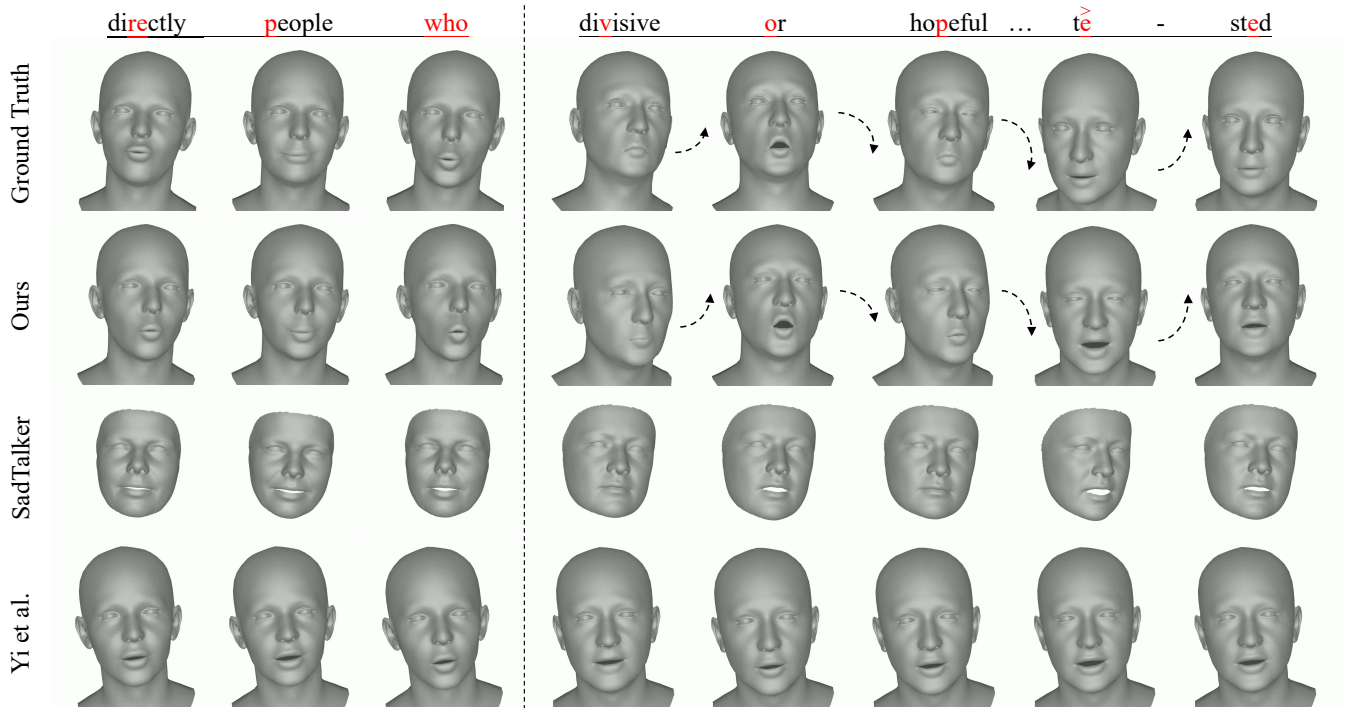

Fig. 4. Qualitative comparison with the state of the arts (w/ head pose prediction). Results of different identities are split by dashed lines. The ">" indicates stress in speech.

as driving blendshape-based avatars [Qian et al. 2023] or facial expression editing [Geng et al. 2019; Sun et al. 2023]. We demonstrate such rendered avatar animations in the Appendix and demo video.

Limitations. Although our method is able to generate high-quality stylistic 3D talking face animation with vivid head poses, there are still some limitations within our framework that could be addressed in follow-up works. Firstly, the computational cost of inference is relatively high due to the sequential nature of the denoising process. To mitigate this, future research can explore more advanced denoisers such as DPM-solver++ [Lu et al. 2022]. Secondly, our method may produce vague or overly smoothed lip motions when encountering very high noise. Possible solutions include incorporating noise suppression during preprocessing or augmenting the training data with noisy audio. Additionally, like existing SOTAs, our approach focuses on animating the face shape while ignoring the inner mouth (including teeth and tongue). Exploring the representation and animation of the inner mouth can lead to more realistic results. Lastly, a promising direction for future research would be to collect real-world 3D talking data that encompasses a broader range of identities and styles, which would further enhance the effectiveness of our approach and contribute to the research community.

Ethical Considerations. Since our approach is able to generate realistic talking head sequences, there are risks of misusing, such as deepfake generation and deliberate manipulation. Therefore, we firmly require that all talking face sequences generated by our method be marked or noted as synthetic data. Moreover, we will make our code publicly available to the deep fake detection community to further ensure that our proposed method can be applied positively. We also hope to raise awareness of the risks and support regulations preventing technology abuse involving synthetic videos.

6 CONCLUSION

Speech-driven expression animation has a wide range of applications in daily life and has received extensive attention from the research community. It involves a challenging many-to-many mapping across modalities between speech and expression animation. In this paper, we present DiffPoseTalk, a novel diffusion-based method for generating diverse and stylistic 3D facial animations and head poses from speech. We leverage the capability of the diffusion model to effectively replicate the distribution of diverse forms, thereby addressing the many-to-many mapping challenge. Additionally, we resolve the limitations associated with current diffusion models that hinder its direct application to speech-driven expression animation. Leveraging the power of the diffusion model and the speaking style encoder, our approach excels in capturing the many-to-many relationships between speech, style, and motion.

REFERENCES

Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. 2023. Listen, Denoise, Action! Audio-Driven Motion Synthesis with Diffusion Models. *ACM Trans. Graph.* 42, 4 (2023), 44:1–44:20.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems* 33 (2020), 12449–12460.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023. InstructPix2Pix: Learning to Follow Image Editing Instructions. In *CVPR*. IEEE, Vancouver, BC, Canada, 18392–18402.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, Virtual Event, 1597–1607.

Michael M. Cohen, Rashid Clark, and Dominic W. Massaro. 2001. Animated speech: research progress and applications. In *AVSP*. ISCA, Aalborg, Denmark, 200.

Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J. Black. 2019. Capture, Learning, and Synthesis of 3D Speaking Styles. In *CVPR*. Computer Vision Foundation / IEEE, Long Beach, CA, USA, 10101–10111.

Radek Daněček, Kiran Chhatre, Shashank Tripathi, Yandong Wen, Michael Black, and Timo Bolkart. 2023. Emotional Speech-Driven Animation with Content-Emotion Disentanglement. In *SIGGRAPH Asia 2023 Conference Papers* (, Sydney, NSW, Australia,) *(SA '23)*. Association for Computing Machinery, New York, NY, USA, Article 41, 13 pages. https://doi.org/10.1145/3610548.3618183

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

Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. 2016. JALI: an animator-centric viseme model for expressive lip synchronization. *ACM Trans. Graph.* 35, 4 (2016), 127:1–127:11.

Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. 2022. FaceFormer: Speech-Driven 3D Facial Animation with Transformers. In *CVPR*. IEEE, New Orleans, LA, USA, 18749–18758.

Gabriele Fanelli, Matthias Dantone, Juergen Gall, Andrea Fossati, and Luc Van Gool. 2013. Random Forests for Real Time 3D Face Analysis. *Int. J. Comput. Vision* 101, 3 (February 2013), 437–458.

Panagiotis Paraskevas Filntisis, George Retsinas, Foivos Paraperas Papantoniou, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. 2022. Visual Speech-Aware Perceptual 3D Facial Expression Reconstruction from Videos. *CoRR* abs/2207.11094 (2022), 1–17.

Zhenglin Geng, Chen Cao, and Sergey Tulyakov. 2019. 3D Guided Fine-Grained Face Manipulation. In *CVPR*. Computer Vision Foundation / IEEE, Long Beach, CA, USA, 9821–9830.

Awni Y. Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. 2014. Deep Speech: Scaling up end-to-end speech recognition. *CoRR* abs/1412.5567 (2014), 1–12.

Kazi Injamamul Haque and Zerrin Yumak. 2023. FaceXHuBERT: Text-less Speech-driven E(X)pressive 3D Facial Animation Synthesis Using Self-Supervised Speech Representation Learning. In *ICMI*. ACM, Paris, France, 282–291.

Thorsten Hempel, Ahmed A. Abdelrahman, and Ayoub Al-Hamadi. 2022. 6d Rotation Representation For Unconstrained Head Pose Estimation. In *ICIP*. IEEE, Bordeaux, France, 2496–2500.

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

Jonathan Ho and Tim Salimans. 2022. Classifier-Free Diffusion Guidance. *CoRR* abs/2207.12598 (2022), 1–14.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE ACM Trans. Audio Speech Lang. Process.* 29 (2021), 3451–3460.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR (Poster)*. ICLR, San Diego, CA., 1–15. http://arxiv.org/abs/1412.6980

Siyao Li, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. 2022. Bailando: 3D Dance Generation by Actor-Critic GPT with Choreographic Memory. In *CVPR*. IEEE, New Orleans, Louisiana, USA, 11040–11049.

Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.* 36, 6 (2017), 194:1–194:17.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022. DPM-Solver++: Fast Solver for Guided Sampling of Diffusion Probabilistic Models. *CoRR* abs/2211.01095 (2022), 1–17.

Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Hongyan Liu, Jun He, and Zhaoxin Fan. 2023. EmoTalk: Speech-Driven Emotional Disentanglement for 3D Face Animation. In *Proceedings of the IEEE/CVF international conference on computer vision*. IEEE, Vancouver, 20687–20697.

Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. 2023. GaussianAvatars: Photorealistic Head Avatars with Rigged 3D Gaussians. *CoRR* abs/2312.02069 (2023), 13 pages.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022a. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022), 3.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022b. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022), 3.

Zhiyuan Ren, Zhihong Pan, Xin Zhou, and Le Kang. 2023. Diffusion motion: Generate text-guided 3d human motion by diffusion model. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Rhodes, Greece, 1–5.

Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. 2021. MeshTalk: 3D Face Animation from Speech using Cross-Modality Disentanglement. In *ICCV*. IEEE, Montreal, QC, Canada, 1153–1162.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*. IEEE, New Orleans, LA, USA, 10674–10685.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *ICML (JMLR Workshop and Conference Proceedings, Vol. 37)*. JMLR.org, Lille, France, 2256–2265.

Stefan Stan, Kazi Injamamul Haque, and Zerrin Yumak. 2023. FaceDiffuser: Speech-Driven 3D Facial Animation Synthesis Using Diffusion. In *MIG*. ACM, Rennes, France, 13:1–13:11.

Zhiyao Sun, Yu-Hui Wen, Tian Lv, Yanan Sun, Ziyang Zhang, Yaoyuan Wang, and Yong-Jin Liu. 2023. Continuously Controllable Facial Expression Editing in Talking Face Videos. *IEEE Transactions on Affective Computing* (2023), 1–14. https://doi.org/10.1109/TAFFC.2023.3334511

Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. 2023. Human Motion Diffusion Model. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=SJ1kSyO2jwu, Kigali, Rwanda, 1–12.

Balamurugan Thambiraja, Ikhsanul Habibie, Sadegh Aliakbarian, Darren Cosker, Christian Theobalt, and Justus Thies. 2023. Imitator: Personalized speech-driven 3d facial animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, Vancouver, 20621–20631.

Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. 2023. CodeTalker: Speech-Driven 3D Facial Animation with Discrete Motion Prior. In *CVPR*. IEEE, Vancouver, 12780–12790.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. 2023. Diffusion models: A comprehensive survey of methods and applications. *Comput. Surveys* 56, 4 (2023), 1–39.

Ran Yi, Zipeng Ye, Zhiyao Sun, Juyong Zhang, Guoxin Zhang, Pengfei Wan, Hujun Bao, and Yong-Jin Liu. 2023. Predicting Personalized Head Movement From Short Video and Speech Signal. *IEEE Transactions on Multimedia* 25, 1 (2023), 6315–6328. https://doi.org/10.1109/TMM.2022.3207606

Chenxu Zhang, Saifeng Ni, Zhipeng Fan, Hongbo Li, Ming Zeng, Madhukar Budagavi, and Xiaohu Guo. 2023b. 3D Talking Face With Personalized Pose Dynamics. *IEEE Trans. Vis. Comput. Graph.* 29, 2 (2023), 1438–1449.

Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. 2023a. SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation. In *CVPR*. IEEE, Vancouver, BC, Canada, 8652–8661.

Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. 2021. Flow-Guided One-Shot Talking Face Generation With a High-Resolution Audio-Visual Dataset. In *CVPR*. Computer Vision Foundation / IEEE, Virtual Event, 3661–3670.

Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu. 2023. Taming Diffusion Models for Audio-Driven Co-Speech Gesture Generation. In *CVPR*. IEEE, Taming Diffusion Models for Audio-Driven Co-Speech Gesture Generation, 10544–10553.

Wojciech Zielonka, Timo Bolkart, and Justus Thies. 2022. Towards Metrical Reconstruction of Human Faces. In *ECCV (13) (Lecture Notes in Computer Science, Vol. 13673)*. Springer, Tel Aviv, Israel, 250–269.