3DiFACE: Synthesizing and Editing Holistic 3D Facial Animation

Balamurugan Thambiraja[1], Sadegh Aliakbarian[3], Darren Cosker[3], and Justus Thies[1,2]

[1] Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2] Technical University of Darmstadt, Germany
[3] Microsoft Mixed Reality & AI Lab, UK
https://balamuruganthambiraja.github.io/3DiFACE

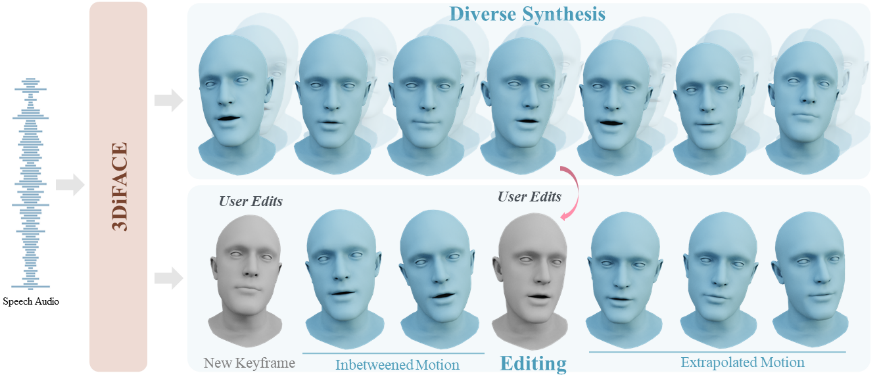


Fig. 1: *3DiFACE* is a novel diffusion-based method for synthesizing and editing holistic 3D facial animation from an audio sequence, wherein one can synthesize a diverse set of facial animations (top), seamlessly edit facial animations between two or multiple user-specified keyframes, and extrapolating motion from past motion (bottom).

Abstract. Creating an animation of a specific person with audio-synced lip motions, realistic head motion and editing via artist-defined keyframes are a set of tasks that challenge existing speech-driven 3D facial animation methods. Especially, editing 3D facial animation is a complex and time-consuming task carried out by highly skilled animators. Also, most existing works overlook the inherent one-to-many relationship between speech and facial motion, where multiple plausible lip and head animations could sync with the audio input. To this end, we present 3DiFACE, a novel method for holistic speech-driven 3D facial animation, which produces diverse plausible lip and head motions for a single audio input, while also allowing editing via keyframing and interpolation. 3DiFACE is a lightweight audio-conditioned diffusion model, which can be fine-tuned to generate personalized 3D facial animation requiring only a short video of the subject. Specifically, we leverage the viseme-level diversity in our training corpus to train a fully-convolutional diffusion model that produces diverse sequences for single audio input. Additionally, we employ a modified guided motion diffusion to enable head-motion synthesis and editing using masking. Through quantitative and qualitative evaluations, we demonstrate that our method is capable of generating and editing diverse holistic 3D facial animations given a single audio input, with control between high fidelity and diversity.

1 Introduction

Synthesizing and editing 3D holistic facial animations is essential for enhancing digital experience in gaming, films, and interactive media. This involves generating realistic lip and head movements that are synchronized with the audio. Early works [9, 15] employed procedural-based rules that map audio features to facial animation parameters. With the evolution of machine learning, data-driven approaches have become prevalent, allowing for audio-conditioned facial animations [10, 18, 39, 42, 58, 62].

Despite these advancements, existing methods only focus on generating facial motion, they do not tackle the issue of editing, such as creating intermediate facial animations between two or multiple, user-specified keyframes (refer to Table 1). Notably, diffusion-based facial motion synthesis methods [1, 6, 48, 49], where editing could be considered as a byproduct of diffusion models, are not demonstrating this. The design choices inherent in these current models, such as the employment of auto-regressive mechanisms or transformer decoders with a look-ahead mask in self-attention and a lack of personalization to a new target identity, are preventing them from effectively addressing the task of editing facial motions. While recent works [1, 6, 48, 49] focus on showcasing diversity in eye-blinks and upper face motion, which has a weak (if any) correlation with the audio, we argue that capturing variations in lip motion in a natural motion sequence is important. In our experiments, we demonstrate how a given audio can be lip-synced in varied but plausible ways, which is a critical part of the animation pipeline and movie dubbing, especially, when edited by an artist.

In this work, we propose a diffusion-based architecture for speech-driven holistic 3D facial animation synthesis and editing to address this gap. In doing so, we face three main challenges: (i) Facial movements are highly person-specific. For facial motion editing, if the speaking style of the edited region doesn't match the style to the target sequence, the sudden shift in speaking style between the edited and unedited motion (the keyframes) leads to unrealistic animations (refer to Figure 6). (ii) Standard diffusion based inbetweening on head motion data struggles to reproduce the imputation signal in the unedited areas resulting in jittery and unrealistic transitions, which is similar to boundary artifact in the image domain as discussed in [7]. Additionally, in case of editing using keyframes, the model often completely ignores the sparse keyframe signal, as also observed in [31] (refer to Figure 7). (iii) Diffusion models are known to require large training sets [43], yet the size of existing high-quality speech-to-3D-animation datasets is limited. Additionally, for personalization of speaking-style the proposed model should be capable of fine-tuning on short video ($1min$) of the target subject. Recent works such as EMOTE [13] and DiffPoseTalk [49] employ head trackers to annotate large-scale-video datasets with pseudo ground truth data and train their models on the resulting dataset. While effectively solving data scarcity, the synthesis fidelity is limited by the quality of the trackers and it is inferior to models that were trained on smaller datasets with higher quality [10], as reported by EMOTE [13].

To address the aforementioned challenges (i) and (iii), we propose a 1D convolutional U-net architecture that can be trained on smaller VOCAset and fine-tuned using short $1min$ reference video of the target subjects. Especially, the fully convolutional nature of our method allows to sub-divide the input sequence into viseme-level motion segments (e.g., 30 frames) during training and generalize to sequences of arbitrary length at inference time. We empirically found this combination of dividing the input sequence into smaller segments and employing a fully convolutional architecture as a key factor for successfully training the diffusion model in both the unconditional synthesis and the style-personalization. Intuitively, we leverage the viseme-level motion diversity present in the dataset to train/fine-tune our method and generate diverse sequence-level samples for a single audio input. In addition, we utilize the classifier-free-guidance [27] in our approach to offer an extra control that can blend between fidelity (lip-sync) and diversity based on the use-case (refer to Figure 8). To address the challenge of editing head motion, we draw inspiration from Guided Motion Diffusion [31] and employ a modified Guided motion diffusion approach. Specifically, we replace portions of the noisy sequence with ground truth data and enforce the model to precisely replicate the samples within the ground truth region. This enables the model to faithfully reproduce the sample in the unedited region during diffusion sampling allowing for smoother and natural head-motion editing.

Through quantitative, qualitative, and perceptual evaluation, we demonstrate the superiority of our method in producing diverse personalized facial animation with natural head motions, enabling the synthesis and editing holistic 3D facial animation. We demonstrate the importance of our architecture design choices, data-efficiency and robustness in detailed ablation studies.

In summary, our contributions are twofold:

- We propose a speech-driven diffusion model for synthesizing diverse, realistic and temporally coherent holistic 3D facial animations.
- To the best of our knowledge, this work constitutes the first attempt to demonstrate pioneering results on two relatively unexplored and challenging research problems: (1.) 3D facial animation editing, such as seamless motion interpolation, keyframing and (2.) unconditional facial animation synthesis.

2 Related Work

Numerous studies have explored speech-driven generation, primarily focusing on synthesizing 2D talking head videos. However, for applications in 3D content creation like games, movies, and immersive telepresence, speech-driven 3D facial animation has raised significant attention in the research community.

Approaches for **talking head video generation** can be broadly categorized into two groups: direct generation of RGB videos from speech and the use of a

	VOCA [10]	Faceformer [18]	CodeTalker [58]	Imitator [53]	FaceDiffuser [48]	EMOTE [13]	TalkSHOW [60]	SadTalker [64]	DiffPoseTalk [49]*	FaceTalk [1]*	DiffusionTalker [6]*	**Ours**
Personalization	✗	✗	✗	✓	✗	✗	✗	✗	✓	✗	✓	✓
Head-motion	✗	✗	✗	✗	✗	✗	✓	✓	✓	✗	✗	✓
Diversity	✗	✗	✗	✗	✓	✗	✓	✓	✓	✓	✓	✓
Editing	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓

Table 1: Summary of the state-of-the-art 3D facial animation synthesis methods. From the table, we can see that only our proposed method is capable of expression personalization, head-motion generation, synthesizing diverse samples and editing 3D animation. * - represents concurrent works.

3D Morphable Model (3DMM) for guided rendering. Suwajanakorn et al. [50] propose a recurrent method to to predict person-specific 2D lip landmarks to guide 2D image generation. Chung et al. [8] propose a real-time method mapping audio input directly to RGB video output. Temporal generative adversarial networks (GANs) have also been employed [55, 66] to address this problem. In another line of work, Zhou et al. [66] disentangles content from style and speaker identity, enabling diverse speech-driven generation. In the second category, an intermediate 3DMM [4, 16] guides 2D neural rendering of talking heads from audio [47, 54, 61, 64, 65], with a focus on facial expressions. Extending these, Wang et al. [57] add the head movements of the speaker to the synthesis. Several works [24, 59] leverage dynamic neural radiance fields [21] to learn personalized audio-driven talking head models.

In addition to 2D talking head generation, a number of studies explored **speech-driven 3D facial animation**. Traditional procedural techniques [14, 15, 17, 29] animate pre-defined facial rigs through procedural rules. With the advent of deep learning, however, these methods have been extended by learning-based approaches [5,10,18,30,42,51,53,54], directly learning viseme patterns from data. Procedural techniques used hierarchical Hidden Markov Models [2, 25, 45] to produce 3DMM parameter space or directly to 3D meshes from audio inputs. Karras et al. [30] propose a learning-based model from a small scale but high quality data, demonstrating a strong baseline at the cost of limited generalization. VOCA [10], on the other hand, is trained on multiple subjects, enabling further generalization. However, the generalization remains limited as it requires one-hot encoding of identities at inference time. Notably, MeshTalk [42], FaceFormer [18], and CodeTalker [58] adopt various strategies for speech-driven 3D facial animation to enable better generalization while maintaining high quality of generated motions. In another work, EMOTE [13] employs EMOCA [12] to generate pseudo-ground truth meshes of the MEAD [56] dataset, enabling it to generate speech-driven facial animation with various emotions style. Similarly, DiffPoseTalk [49] utilizes the pseudo-ground truth from HDTF [65] dataset to generate holistic 3D facial animation. FaceTalk [1] trains a latent-diffusion model on a custom 3D dataset generate volumetric 3D animation with diversity.

Table 1 underscores the significance of our method, the first to address the challenging task of synthesizing and editing diverse 3D facial animations holistically from a single audio input, which is an integral part of animation pipelines.

3 Method

Our goal is to synthesize and edit holistic 3D facial animation given input audio signal. In this context, holistic facial animation refers to facial motion *and* head motion which we model in two diffusion-based networks, motivated by the fact that the face motion is highly correlated to the speech signal, while the head motion is relatively less correlated and thus requires a longer context of information, hence, a different training scheme (and data). Following recent works [49,52], we train a denoising model θ that can reverse this noisy diffusion and estimate the original sample from a noised version guided by conditioning signal C. To add diversity for face motion synthesis, we employ Classifier-Free Guidance (CFG) [27] and calculate the output as a weighted sum of the conditional and unconditional prediction using the guidance scale s. In the following section, we discuss our facial and head motion generators in Section 3.1 and Section 3.2 respectively.

Audio Encoding: Similar to other state-of-the-art methods [10, 18, 53, 58], we adopt the pretrained Wav2Vec2.0 [2] to generate audio features from the raw audio signal. Wav2Vec2.0 uses a self-supervised learning approach to map audio to quantized feature vectors with 768 channels. We resample the encoder output via linear interpolation to match the sampling rate of the motion sequences (30fps for VOCAset [10]). A trainable linear layer is applied to project the feature vectors to 64 channels, resulting in a speech representation $\hat{A} \in \mathbb{R}^{N \times 64}$ for N frames.

3.1 Facial motion generator

Our diffusion-based facial motion generator takes an audio signal as input and produces a sequence of 3D vertex displacements w.r.t. a template mesh by iterative denoising, see Figure 2,. Let $x_0 \in \mathbb{R}^{N \times D \cdot 3}$ denote such a sequence of displacements, where N is the sequence length and D is the number of vertices in the template mesh. The input to our diffusion model parameterized by θ_f is a noisy vertex displacement sequence $x_t \in \mathbb{R}^{N \times D \cdot 3}$. The task is then to predict its noise-free counterpart $\hat{x}_0 = \theta_f(x_t, t, C)$, given diffusion step t and conditions C. As a first step, we employ a single fully connected layer as *Motion Encoder* to project x_t to a 64-dimensional latent space. We positionally encode the diffusion step t [46], map it to the latent space with a linear layer, and add it to the encoded x_t. We apply a series of 1D-convolution blocks to first reduce the temporal dimension of the activations, followed by an upsampling convolution block to restore the original temporal dimension. Each convolution block is followed by a condition block to incorporate the audio features. The condition blocks concatenate the input features with the audio and apply a dimension-preserving

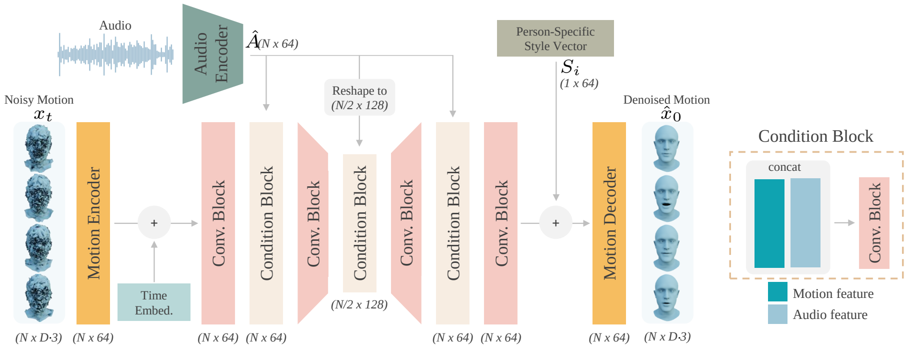


Fig. 2: Our method takes noised vertex displacements, denoted as x_t, and the diffusion time step embedding as inputs to predict a denoised sample $\hat{x}_0$, leveraging both the audio conditioning signal $\hat{A}$ and a person-specific feature vector S_i. Our approach employs wav2vec2.0 [2] for extracting audio features from the raw audio signal. The audio condition is injected into the network by concatenation through a series of convolutional blocks. Note that N corresponds to the frame count of the sequence and D to the number of vertices.

convolution. We add a person-specific feature vector $S_i \in \mathbb{R}^{1\times 64}$ to the output of the convolutional layers prior to applying the *Motion Decoder*. This procedure yields the final noise-free sample $\hat{x}_0$, as illustrated in Figure 2. Similar to the *Motion Encoder*, the *Motion Decoder* is a single fully connected layer. Note that in our formulation, the condition C represents the set of both the per-frame audio features $\hat{A}$ and the person-specific feature vector S_i.

In contrast to state-of-the-art methods on 3D facial animation synthesis that utilize transformer architectures [18, 49, 53, 58], we take inspiration from Pavllo et al. [38] and adopt a 1D-convolutional network as our backbone. Specifically, instead of infusing the condition through an attention mechanism, we use feature concatenation. In particular, our fully convolutional architecture without attention allows to sub-divide the input sequence into viseme-level motion segments (e.g., 30 frames) during training and generalize to sequences of arbitrary length at inference time. We empirically observed that these modifications to the architecture are essential for its effective training on the limited VOCA training dataset [10] (refer to Table 3), especially on the unconditional training setup. These modifications significantly enhance the model's performance during personalized fine-tuning on small, subject-specific datasets. Note that this strategy is not viable for transformer-based 3D facial animation baselines, since it struggles to capture any longer-term dependency beyond the predefined context length leading to context fragmentation [11] and subpar performance. This issue becomes even more pronounced in our training configuration, where we crop the sequences to only 30 frames. While auto-regressive motion synthesis could in theory mitigate this limitation, it would make the animation editing tasks, such as inbetweening distant motion frames, impossible.

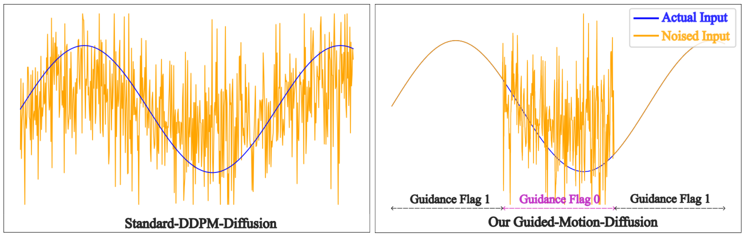


Fig. 3: Illustration of standard diffusion (left) and our modified guided-motion-diffusion (right), where in the forward diffusion process, part of the noisy input signal is replaced with the ground truth signal and a guidance flag of (0) and (1) is concatenated to the noisy and ground truth regions respectively.

Training Similar to [49, 52], we train our diffusion model to predict the ground truth vertex displacements x_0 from their noised counterparts x_t:

$$\mathcal{L}_{\text{simple}} = ||x_0 - \theta_f(x_t, t, C)||^2. \tag{1}$$

In comparison to predicting the applied noise which is common practice in related work [36, 43, 63], we empirically found that predicting the ground truth displacements yields better convergence in the unconditional and person-specific fine-tuning setup. Furthermore, we take inspiration from [10, 53] and add a velocity loss $\mathcal{L}_{\text{vel}}$ to improve temporal smoothness:

$$\mathcal{L}_{\text{vel}} = \frac{1}{N-1} \sum_{n=1}^{N} ||(x_{0,n} - x_{0,n-1}) - (\hat{x}_{0,n} - \hat{x}_{0,n-1})||^2, \tag{2}$$

where $x_{0,n}$ denotes the ground truth vertex displacements in frame n. Our final training objective is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{simple}} + \lambda_{\text{vel}} \cdot \mathcal{L}_{\text{vel}}. \tag{3}$$

We set $\lambda_{\text{vel}} = 10.0$ unless specified otherwise. Note that during training, we randomly set the audio condition C to 0 in 10% of the cases in order to enable unconditional synthesis at inference time.

Person-Specific Fine-tuning For capturing the speaking style of a subject that is not part of the training set, we require a short reference talking head video. The facial movements are extracted with the state-of-the-art monocular face tracker MICA [68]. We use the tracked meshes as pseudo ground truth and fine-tune the entire model to fit the expression distribution of the target subject using the training objective from Eq. (3).

3.2 Head-motion generator

Given an audio signal input, our head motion generator produces smooth and natural head motions $y_0 \in \mathbb{R}^{N \times 3}$, where N is the sequence length. We parameterize the head motion via the neck joint rotation in the FLAME model [33], where the rotation is represented via axis angle. Motivated by the head-motion editing issue mentioned in Section 1, we employ a modified guided motion diffusion for the head motion synthesis. The original guided motion diffusion(GMD) [31] was introduced to inject a spatial guidance signal into the full-body motion synthesis problem. We draw inspiration from it and modify the spatial guidance with an intra-sequence guidance injection, to highlight the relative importance of the different segments in the input signal. Specifically, as illustrated in Fig. 3, during the forward diffusion process, part of the noisy input is replaced with ground truth signals, and a corresponding guidance flag of 0 or 1 (ground truth signal) is concatenated. A denoising model parameterized by θ_h is trained to reverse this diffusion process by leveraging this additional information.

Similar to the facial motion generator, we employ a fully convolutional architecture as our backbone for the head-motion denoising model. Additionally, we introduce skip connections between the encoder and decoder layers, to aid the model in reproducing the ground truth signals. For the audio encoder, we use the pre-trained audio encoder from the facial motion synthesis pipeline, which is kept frozen during the head-motion training. The final diffusion formulation is represented as $\hat{y}_0 = \theta_h(y_t, t, C)$, given diffusion step t and audio conditions C.

Algorithm 1 Our GMD Training

```
repeat
    $y_0 \sim q(y_0)$
    $t \sim \text{Uniform}(\{1, \ldots, T\})$
    $\epsilon \sim \mathcal{N}(0, I)$
    $y_t = \sqrt{\bar{\alpha}_t} y_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$
    $\bar{y_t} = y_t \oplus (0)$
    $\bar{y_0} = y_0 \oplus (1)$
    $y_t = (1 - M_t) \odot \bar{y_t} + M_t \odot \bar{y_0}$
    grad desc. $\nabla_{\theta_h} \|y_0 - \theta_h(y_t, t, C)\|^2$
until converged
```

Algorithm 2 Our GMD Sampling

```
$y_T \sim \mathcal{N}(0, I)$
Input signal $Y_0$, if any
Imputation mask $M_0$, if any
$\bar{Y_0} = Y_0 \oplus (1)$
for $t = T, \ldots, 1$ do
    $\bar{y_t} = y_t \oplus (0)$
    $y_t = (1 - M_0) \odot \bar{y_t} + M_0 \odot \bar{Y_0}$
    $\hat{y_0} = \theta_h(y_t, t, C)$
    $\hat{y_0} = (1 - M_0) \odot \hat{y_0} + M_0 \odot Y_0$
    $\mu, \sigma \leftarrow \mu(y_t, \hat{y_0}), \sigma_t$
    $y_{t-1} \sim \mathcal{N}(\mu, \sigma)$
end for
return $y_0$
```

Training The complete training and sampling procedure of modified guided-motion diffusion is detailed in Algorithm 1 and Algorithm 2. In addition to the losses used in the facial motion generator Section 3.1, we add an additional guiding mask loss to enforce the model to faithfully reproduce the results in ground truth signal injected into the sequence. The guidance loss can be formulated as:

$$\mathcal{L}_{\text{mask}} = ||w_n \odot (y_{0,n} - \theta(y_{t,n}, t, C))||^2, \tag{4}$$

where n indicates the n^{th} frame in the sequence y_0 and w_n is the guidance weight, 1 for the ground truth frames and zero otherwise.

4 Dataset

We train our facial motion model on the VOCAset [10], since it provides high-quality, speech-aligned 3D face scan sequences. Following previous work [53], we

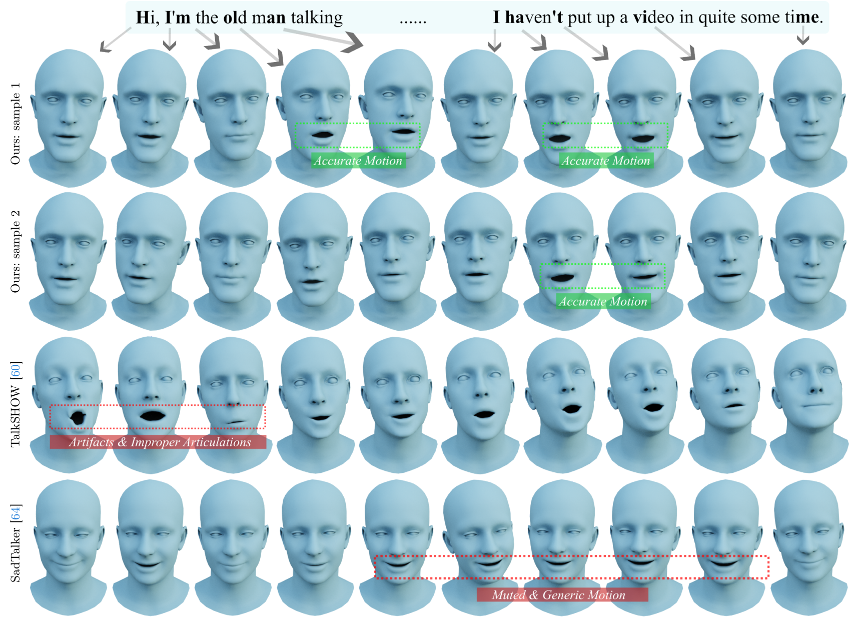


Fig. 4: Our qualitative comparison shows that our method outperforms the baseline in creating more accurate lip-synced facial animations with diverse head movements. Specifically, TalkSHOW produces animations with jittery artifacts, while SadTalker yields muted and generic animations.

use the train/val/test set split of 8, 2, 2 actors. All 40 sequences of the training actors are used during training. However, for the test and validation, only 20 sequences without overlap with the speech scripts of the training sequences are used. We evaluate person-specific fine-tuning on in-the-wild video sequences from Imitator [53]. The provided videos are 2 minutes long which we divide into 60/30/30 seconds for train/val/test respectively. To train our head-motion generator, we use the HDTF [65] dataset, as the VOCAset does not include large variations of head motion. Using the download and processing script provided by the authors, we extract 352 videos with 246 unique subjects and use the MICA tracker [68] to extract head poses. For our experiments, we split the dataset into 300/20/32 sequences for train/val/test accordingly. In this work, we employ the VOCAset, HDTF, and Imitator's in-the-wild dataset to train our method for generating and editing 3D facial animations with head-motion. This choice led us to exclude the Biwi dataset [19] from our study, as it lacks sequences that with full head model like FLAME [33], which is essential for synthesizing head motion effectively. More details on the dataset is provided in the suppl. material.

5 Results

We evaluate our method against state-of-the-art methods: SadTalker [64] and TalkShow [60] on the holistic 3D facial animation synthesis task and VOCA [10],

Method	Div$^{\mathbf{L}}$ ↑	Lip-Sync ↓	BA ↑	Div$^{\mathbf{H}}$ ↑
	Non-Personalized regression			
1 VOCA [10]	–	5.30	–	–
2 Faceformer [18]	–	2.85	–	–
3 Imitator [53]	–	1.95	–	–
4 CodeTalker [58]	1.40	2.55	–	–
5 Ours$_{s=0.5}$ (w/o sty)	**2.57**	**1.71**	–	–
	Non-Personalized diffusion			
6 FaceDiffuser [48]	0.05	**1.60**	–	–
7 Ours$_{s=1.0}$ (w/o sty)	**0.64**	1.62	–	–
	Personalized synthesis			
8 Imitator (w/ sty)	–	**1.35**	–	–
9 Ours$_{s=0.5}$ (w/ sty)	**1.57**	1.56	–	–
10 Ours$_{s=1.0}$ (w/ sty)	0.24	1.42	–	–
	Holistic 3D Facial animation syn.			
11 SadTalker [64]	1.59	4.01	0.285	0.004
12 TalkSHOW [60]	1.80	4.35	0.296	0.002
13 **Ours composite**	**2.57**	**1.71**	**0.338**	**0.007**

Table 2: Quantitative comparison: In general, our proposed method produces better holistic 3D facial animations with high-fidelity lip and head motions (refer row 11-13). In addition, ours allows for editing the animation unlike the baselines. On the regression and dission-based non-personalized facial motion synthesis task (row 1-7), our method produces more diverse and lip-synced samples than the baselines, except for FaceDiffuser, where we match the performance on *Lip-Sync* despite producing more diverse samples. Further, we see that our method is able to personalize facial motions on the level of Imitator [53], a method designed for personalization, while producing more diverse samples and allowing for motion editing using keyframes.

Method	Div$^{\mathbf{L}}$ ↑	Lip-Sync ↓
	(a) Design choices	
1 Ours (concat + win30)	**2.57**	**1.71**
2 Ours (attn + win30)	0	3.21
3 Ours (FF arch + win30)	0	3.49
4 Ours (concat + no win)	0	1.98
	(b) Person-specific Fine-tuning	
5 Ours (∼ 5s)	29.95	4.89
6 Ours (∼ 30s)	0.18	1.81
7 Ours (∼ 60s)	0.67	1.69
8 Ours (∼100s)	1.57	1.56
	(c) Audio noise ablation	
9 Ours (high noise)	6.41	2.56
10 Ours (med. noise)	2.54	1.97
11 Ours (low noise)	1.85	1.78
	(d) GMD ablation	
Method	**BA ↑**	**Div$^{\mathbf{H}}$ ↑**
12 Ours w. In mask	0.368	0.008
13 Ours w. KF mask	0.308	0.008
14 Ours w/o. mask	0.338	0.007

Table 3: Ablation study: (a) Design choices: We show that the combination of a fully convolutional architecture without attention or transformer and viseme-level windowing is critical for the training the model on the VOCAset [10] to produce diverse samples. (b) Fine-tuning: Further, we show that 30s of video suffice to perform person-specific fine-tuning while 100s further improve all scores (row 6-9). (c) Noise: Row 10-12 illustrates the robustness of method w.r.t. medium and low audio noise levels. (d) GMD ablation: shows the performance of our method w.r.t the imputation signal, our method produce better metrics compared to the baselines irrespective of the imputation signal.

Faceformer [18], CodeTalker [58], EMOTE [13], FaceDiffuser [48] and Imitator [53] on facial motion synthesis task. Figure 4 presents the qualitative comparison on holistic 3D motion synthesis, where our method produces more accurate lip-synced facial animations with diverse head movements. A qualitative comparison to the facial motion synthesis baselines on a test sequence from the VOCAset is shown in Figure 5, where our method produces expressive facial animations that matches the speaking style of the target subject. Additional qualitative results are shown in the suppl. video.

Quantitative Comparison: In Table 2, we present a quantitative evaluation based on the following metrics: *Lip-Sync* measures the lip synchronization using Dynamic Time Warping to compute the temporal similarity [53]. Diversity metric Div^L and Div^H proposed by Ren et al. [41] measures the diversity of lip motion and head motion generated from the same audio. Similar to DiffPoseTalk [49], we employ a modified beat alignment *BA* to measure the synchronization of the head movement beats.

In Table 2 (rows 11-13), *holistic 3D facial animation synthesis* is evaluated, and it is evident that ours significantly surpasses the baselines, particularly, in terms of lip-sync accuracy and beat alignment, offering greater diversity.

For the *facial motion synthesis task without head motion*, we quantitatively compare our method on 3 different setups namely, non-personalized regression and diffusion and personalized synthesis. On the non-personalized regression setup, we find that our method with the guidance scale of $s = 0.5$, outperforms the baseline on synthesis diversity (by over 80%) and *Lip-Sync* (see row 1 and 5 of Table 2). On the non-personalized diffusion, our model with the guidance scale $s = 1.0$ matches FaceDiffuser [48] on *Lip-Sync*, while significantly outperforming it on diversity. Note that we can use the guidance scale parameter to freely trade synthesis diversity for lip-sync accuracy, which is not possible in FaceDiffuser [48]. Figure 8 indicate that adjusting the guidance scale parameter s within $0.3 \leq s \leq 1.0$, effectively enhances the lip-motion diversity, surpassing all baseline methods, while only slightly diminishing lip-sync accuracy. Finally, on the *personalization synthesis* setup, we achieve higher synthesis diversity compared to Imitator [53] and match the performance closely on *Lip-Sync*. Note that Imitator is a deterministic model that does not allow for diverse lip-motion synthesis and facial motion editing.

User Study: We conducted A/B user studies to assess our method's perceptual performance. For the facial motion synthesis task, we compare our method on a *high diversity* ($s = 0.5$) and *high fidelity* ($s = 1.0$) setup. In the high fidelity setup, we outperform the baselines in terms of both expressiveness and lip-synchronization. Even on a extreme the high diversity setup, we outperform CodeTalker [58] and perform closely to FaceDiffuser [48], which produces high-fidelity samples at the expense of diversity. Furthermore, we assessed how well our personalized model preserves speaking styles by comparing it to Imitator, the sole baseline model that offers personalization capabilities. To this end, the users rated the similarity based on a reference video and the synthesized videos of the VOCA test set, where 55% of the users preferred our method. Finally, we evaluated our method on the holistic 3D facial animation task. In Table 4 (row 6-8) we can see that our method generates significantly better facial animations with plausible facial motion and natural head motion. For more details on the user-study, please refer to our supplemental material.

Motion Editing: We show motion editing using keyframes in Figure 6. In this application, we selectively replace the predicted denoised vertex-displacement sequences $\hat{x}_0$ with ground truth values during the denoising process. This is similar in spirit to well-established diffusion-based image inpainting methods [35]. We additionally show unconditional motion synthesis and editing results in the supplemental material. As can be seen in Figure 6, the personalization of the motion synthesis is important to match the talking style, preventing an abrupt style change. In Figure 7, we visualize the trajectory of the edited head-motions.

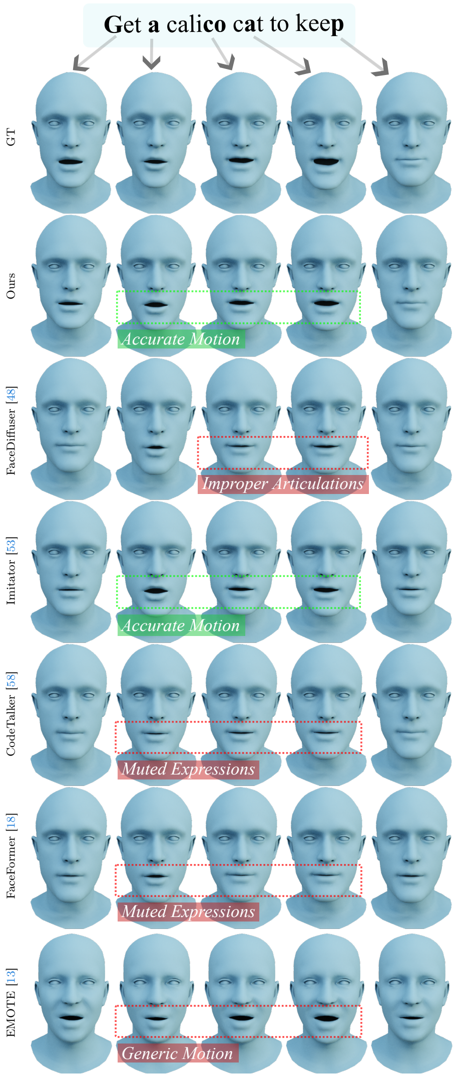


Fig. 5: Qualitative comparison of the facial motion synthesis model. Only our method and Imitator produce expressive motions that match the target speaking style. While Imitator synthesizes similarly convincing animations, its outputs are not diverse (see Table 2) and it cannot be used for animation editing.

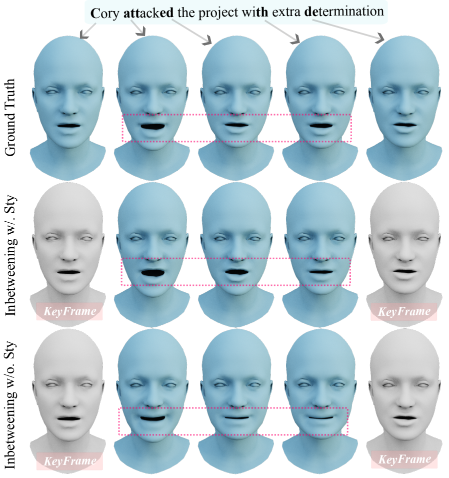


Fig. 6: Qualitative evaluation of the importance of person-specific finetuning for motion editing. As highlighted in purple, without finetuning we observe an abrupt change in speaking style between the keyframes and the generated motion, thus rendering the results unrealistic.

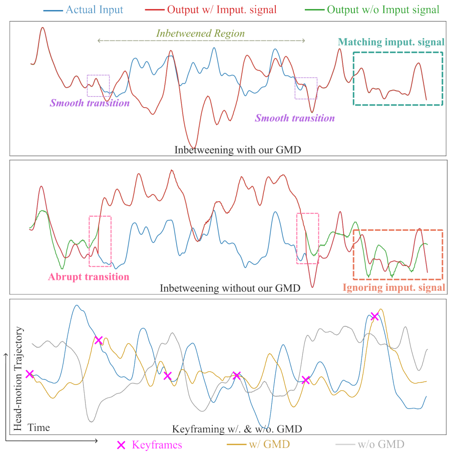


Fig. 7: Impact of our Guided motion diffusion (GMD) on the head-motion editing. Without GMD the model cannot faithfully reproduce the imputation signal resulting in a jittery transition between the edited and unedited regions. Additionally, without GMD the model ignores the sparse keyframes completely.

We can see that without our proposed guided motion diffusion there is a sharp transition between the edited and unedited head-motion, as a result the final motion will have jitter artifacts which look unrealistic.

Ablation: In the following, we will address important questions regarding our design choices and robustness.

• *Is a 1D-convolutional U-net architecture the right choice?* As discussed in Section 3 using our proposed architecture instead of the transformer-based architecture from Faceformer or the attention-based Unet architectures used in [36] results in significantly better performance on both the *Lip-Sync* and diversity (refer to Table 3 rows 1-3).

• *What is the effect of viseme-level window-based training?* Table 3 row 1 vs 4 shows that without window-based training the performance worsens in terms of both lip-sync and diversity. Further, in the suppl. video, we demonstrate the ability of our method to generate 20 *sec* long motion compared to the baselines, despite being trained only on 1 *sec* segments.

• *How much data do we need for person-specific fine-tuning?* Table 3 rows 5-8 indicate 30 and 60 seconds of data are sufficient for good results, 100 seconds yield the best lip-sync and diversity Div^L.

• *How robust is it to noisy input audio ?* As reported in Table 3 rows 9-11, our method produces robust high-quality facial animations for low(36db) and medium(24db) noise levels.

• *Does the guided-motion-diffusion help to generate diverse motion?* On Table 3 rows 12-14, we analyze the effect of our keyframe(KF mask) and inbetweening(In mask) based guidance on the synthesis quality. Our findings demonstrate that employing guided-motion-diffusion improves the diversity of head-motion with minimal impact on overall quality, while offering additional editing capabilities.

6 Discussion

Our proposed method excels in synthesizing and editing diverse holistic 3D facial animations based on speech. Similar to Imitator, for personalization, our method depends on the appearance and quality of the face tracker. However, through qualitative results, we demonstrate that our method is able to personalize from both high-quality motion capture sequence from VOCAset and from monocular head trackers applied to in-the-wild videos. In this work, we employ a modified guided motion diffusion [31] to tackle the boundary artifact [7] and sparse-signal neglect in the head-motion synthesis and editing. In contrast to the head motion, boundary artifacts are imperceptible for facial motion. For face motion editing, style-personalization is the critical contribution to enable seamless editing. One key capability of our method is that it offers control to animators and creators via keyframes, which can be additionally extended to an explicit natural language based condition to control the synthesis, which we leave for future works.

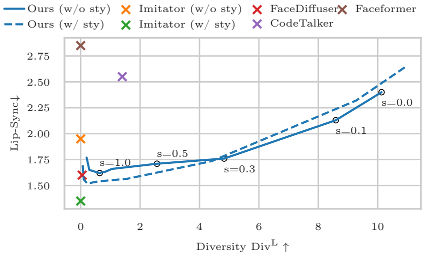


Fig. 8: We investigate the impact of the classifier-free-guidance scale s [27] using the 'Lip-sync' and Div^L metrics on the non-personalized facial motion synthesis task. Lower guidance values yield animations with significantly more diverse motion but inferior lip-sync quality. Conversely, higher guidance values result in high-quality animation with reduced diversity. We observe a similar trend in our perceptual evaluation. We find that the guidance scale s is an effective tool to increase synthesis diversity beyond all baselines with only a small loss of lip-sync accuracy for $0.3 \leq s \leq 1.0$.

Method	**Exprs** (%)	**Lip-sync** (%)
	High-Fidelity (Ours $s = 1.0$)	
1 Ours vs Imitator [53]	65.72	69.47
2 Ours vs Faceformer [18]	73.28	71.43
3 Ours vs FaceDiffuser [48]	67.85	66.71
	High-diversity (Ours $s = 0.5$)	
4 Ours vs CodeTalker [58]	53.64	53.80
5 Ours vs FaceDiffuser [48]	40.84	41.55
	Holistic synthesis	
Method	**Face Motion** (%)	**Head motion** (%)
6 Ours vs Ground Truth	78.57	55.41
7 Ours vs SadTalker [64]	88.13	86.43
8 Ours vs TalkShow [60]	90.77	87.96

Table 4: User study. Our method produces accurate facial movement with expressiveness and lip-sync. Even on a extreme high-diversity setup with $s = 0.5$, our method outperforms the CodeTalker and pars sightly in comparison to FaceDiffuser, which offers negligible diversity. In addition, our method produces consistently better holistic 3D facial animation compared to the baselines. Similar to Imitator [53], we evaluate the person-specific speaking-style similarity against [53], where 55% of users favored our method for better style-similarity.

7 Conclusion

With 3DiFACE we present the first method that can both generate and edit diverse holistic 3D facial animations from speech input. Employing classifier-free guidance provides us with an effective tool to balance synthesis diversity and accuracy allowing us to generate animations with unprecedented diversity. Through personalization, we can extract person-specific speaking styles from short ($\sim 100s$) videos which significantly improves performance. Further, we demonstrate the ability to edit both facial and head motion through keyframes. We are convinced that these properties make 3DiFACE a powerful tool for content creators and are excited to see future applications.

8 Additional Evaluation

8.1 Is it possible to unconditionally synthesize and edit motion?

While unconditional motion synthesis has been extensively applied in the motion synthesis domain [40,52], to the best of our knowledge, its application in 3D facial animation synthesis remains widely unexplored. The significance of an unconstrained facial motion synthesis method cannot be overstated. It holds substantial potential for various applications, such as animating background characters in movies and games. Additionally, it enables targeted editing of specific facial elements—such as eye blinks and eyebrow motions—since these non-verbal facial expressions often exhibit weak or no correlation with audio features. Moreover, an unconditional model serves as a valuable motion before downstream tasks, extending its utility beyond synthesis and editing applications. Our demonstration of unconditional synthesis and editing are showcased in Figure 9, underscoring the potential and versatility of such unconstrained models for 3D facial animation synthesis.

8.2 Is the performance of motion editing consistent across varying degrees of imputation signal?

We evaluate the performance of motion inbetweening with respect to the input data. To this end, we preserve 5%, 10%, 20%, and 50% of the starting and ending frames, and then perform inbetweening for the intermediate motion sequences. Furthermore, we assess the robustness of the inbetweening by randomly inserting keyframes at different rates: 1KF/sec, 2KF/sec, and 3KF/sec. For facial motion, these evaluations are conducted for all sequences of the test subject 024 from the VOCAset [10], and the resulting metrics are presented in Table 5. Similarly for head motion, these evaluations are conducted in the HDTF [65] test set. For the Table 5, it is evident that as the imputation signal strength increases, the synthesis's fidelity improves while its diversity decreases. This is because there are fewer frames available for generating varied samples, and the model must align closely with the frames provided by the imputation signal. This demonstrates the efficacy and robustness of our method in synthesizing and editing holistic 3D facial animations.

8.3 Why standard diffusion-based head motion-editing fails?

In this section, we analyze standard diffusion fails for the head motion editing task. An Illustration of a conditional inbetweening sequence across various diffusion steps t is shown in the Figure 10 Observing the left column of the figure, we notice that the model without our guided motion diffusion primarily concentrates on generating a valid sample from the distribution, given the condition. As a result, it starts to overlook the imputation signal in the low noise regime, focusing instead on refining the sequence to produce an improved sample from the distribution. This approach leads to jittery transitions when the imputation signal is replaced to generate the final inbetweened sample at the end of the sampling. Throughout its training, the diffusion model was only trained to generate a valid sample from the distribution, not to align with any part of the imputation signal.

Observing this, we employed a modified guided motion diffusion to incorporate guidance signals during training. This adjustment teaches the diffusion to maintain the imputation signal while still producing a sample from the distribution. Unlike the initial approach, our guided motion diffusion model utilizes the

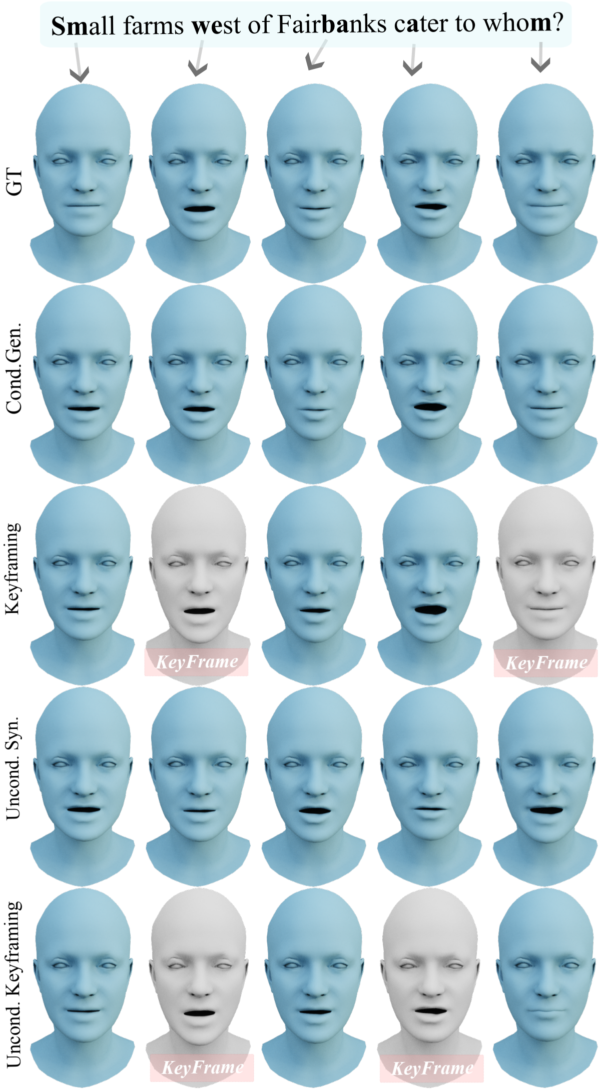


Fig. 9: Qualitative illustration of facial motion inbetweening using our conditional and unconditional model. In rows 2 and 3, we showcase a sequence synthesized conditionally and subsequently refined using keyframes. In Row 4 (Uncond. Syn.), we present our unconditional synthesis results. As observed from the results, our model can unconditionally synthesize facial animations that appear plausible. Further, in row 5 (Uncond. Keyframing), we see that our method can unconditionally inbetween facial animation while preserving the speaking style of the target actor. This progression demonstrates our model's capabilities: from conditional synthesis and keyframe-based editing to unconditional synthesis and editing, while maintaining the actor's speaking style.

Method	Div^{L} ↑	Lip-Sync ↓	BA ↓	Div^{H} ↓
1 Ours (synthesis)	1.35	1.4	0.338	0.007
2 Ours (Ip 5%)	1.27	1.17	0.341	0.007
3 Ours (Ip 10%)	1.24	1.15	0.352	0.006
4 Ours (Ip 20%)	1.15	1.01	0.358	0.005
5 Ours (Ip 50%)	0.9	0.68	0.403	0.004
6 Ours (1KF/sec)	1.26	1.28	0.321	0.006
7 Ours (2KF/sec)	1.14	1.2	0.347	0.006
8 Ours (3KF/sec)	1.05	1.1	0.365	0.005

Table 5: We quantitatively evaluate our facial motion editing capability on all the test sequences of the subject 024 in the VOCAset [10] and head motion editing on the test sequences of the HDTF [65]. To this end, first, we preserve 5%, 10%, 20%, and 50% of the starting and ending frames, and then perform inbetweening for the intermediate motion sequences. In addition, we assess the robustness of the inbetweening by randomly inserting keyframes (KF): 1KF/sec, 2KF/sec, and 3KF/sec. From the metrics, we can see that the synthesis quality increases significantly with the addition of more keyframes, which is a clear indication that the model matches the ground truth and produces realistic motion. For animators and artists, this means that they can insert any number of keyframes they want and have fine-grained control over the motion synthesis. Note that keyframes could also stem from previously generated motion sequences using our method (iterative refinement).

guidance flag and imputation signal to generate samples from the distribution that accurately reproduce the imputation and align with the condition.

9 Implementation

In the section, we provide more details on the diffusion model, dataset, baselines, and metrics respectively.

9.1 Preliminaries

Denoising Diffusion Probabilistic Models: Our method is based on the diffusion framework of Sohl et al. [46], where a training sample x_0 gradually transforms into white noise through the addition of Gaussian noise across T steps. This transformation is mathematically represented as:

$$x_t \sim q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I), t = 1...T, \tag{5}$$

where β_t is following a predefined variance schedule.

Following recent work [49, 52], we train a denoising model θ that can reverse this noisy diffusion and estimate the original sample x_0 from a noised version x_t,

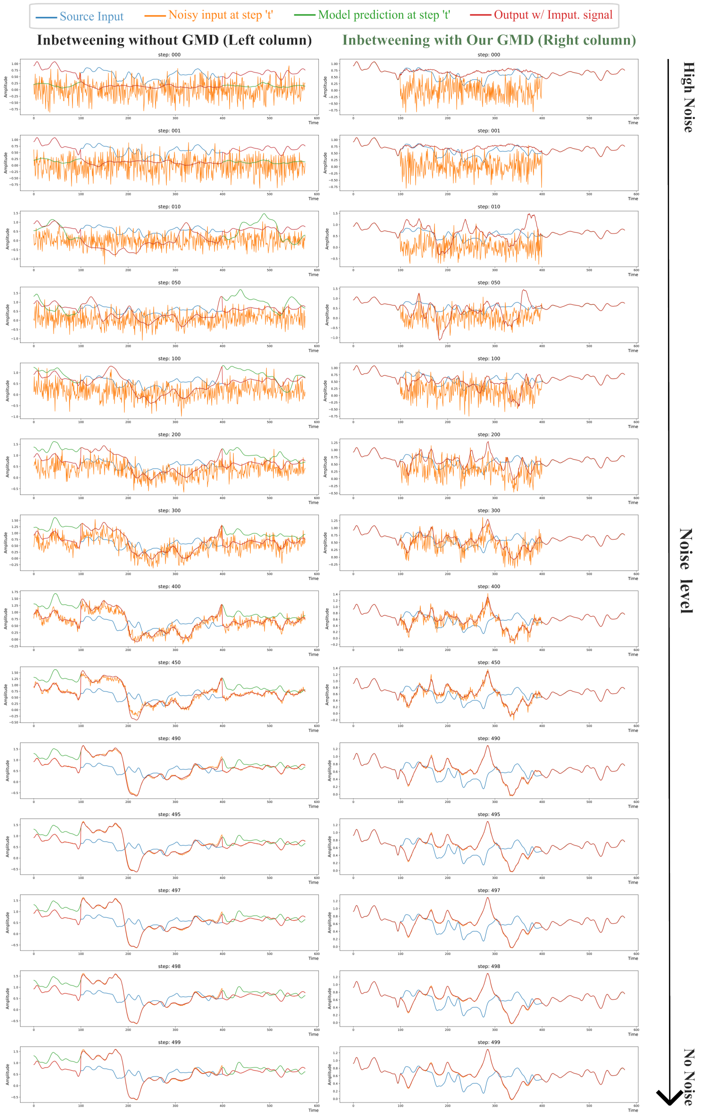


Fig. 10: Illustration of a conditional inbetweening sequence across various diffusion steps t. Observing the left column of the figure, we notice that the model without our guided motion diffusion(GMD) primarily concentrates on generating a valid sample from the distribution, given the condition. As a result, it starts to overlook the imputation signal in the low noise regime, focusing instead on refining the sequence to produce an improved sample from the distribution. This approach leads to jittery transitions when the imputation signal is replaced to generate the final inbetweened sample at the end of the sampling. In contrast, our guided motion diffusion model utilizes the guidance flag and imputation signal to generate samples from the distribution that accurately reproduce the imputation and align with the condition.

guided by: $\hat{x}_0 = \theta(x_t, t, C)$. With θ being the neural network and C representing additional conditions. The reverse diffusion is achieved through:

$$q(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \sqrt{\bar{\alpha}_{t-1}}\theta(x_t, t, C), (1 - \bar{\alpha}_{t-1})I\right),$$

where $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{k=1}^{t} \alpha_k$.

To generate new samples, we start from random noise x_T and apply iterative denoising until reaching $t = 0$. We introduce diversity in generation using Classifier-Free Guidance (CFG) [27] by combining conditional and unconditional predictions of the network, controlled by a guidance scale s:

$$\theta_s(x_t, t, C) := \theta(x_t, t, \emptyset) + s \cdot [\theta(x_t, t, C) - \theta(x_t, t, \emptyset)],$$

adjusting s to balance between diversity and adherence to conditions. Following [26], the inverse diffusion process is then given through:

$$q(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \sqrt{\bar{\alpha}_{t-1}}\theta(x_t, t, C), (1 - \bar{\alpha}_{t-1})I\right), \tag{6}$$

with $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{k=1}^{t} \alpha_k$. For generating new samples, we randomly sample x_T from a Gaussian distribution and iteratively denoise it until $t = 0$ is reached.

To add diversity, we employ Classifier-Free Guidance (CFG) [27] and calculate the output as a weighted sum of the conditional and unconditional prediction:

$$\theta_s(x_t, t, C) := \theta(x_t, t, \emptyset) + s \cdot [\theta(x_t, t, C) - \theta(x_t, t, \emptyset)], \tag{7}$$

where s is the guidance scale and $\theta(x_t, t, \emptyset)$ denotes the unconditional prediction in which we set the audio conditions to zero. Note that while CFG is typically used with a guidance scale > 1 to enhance alignment with the condition, we set it to values < 1 (0.5 unless specified otherwise) to increase diversity.

9.2 Dataset

VOCA: We train our facial motion model on the VOCAset [10] since it provides high-quality, speech-aligned 3D face scan sequences. It consists of 12 actors (6 female and 6 male) with 40 sequences each with a length of 3-5 seconds, resampled at 30fps. Following previous work [53], we use the train/val/test set split of $8, 2, 2$ actors. All 40 sequences of the training actors are used during training. However, for the test and validation, only 20 sequences that do not overlap with the speech scripts of the training sequences are used. For the style adaption experiment, we split the 40 sequences of the test actors to $18, 2, 20$ for train/val/test sets. The test sequences of the experiments w/ and w/o style adaptation are identical, allowing a direct comparison of the scores in the quantitative comparison in the main paper (Table 2).

In-the-wild dataset: We evaluate person-specific fine-tuning on in-the-wild video sequences from Imitator [53]. The provided videos are 2 minutes long which we divide into 60/30/30 seconds for train/val/test respectively. Similar to Imitator, we employ the MICA tracker [68] to extract the face motion tracking for the personalization step.

HDTF: We train our head-motion generator on the HDTF [65] dataset. The High-definition Talking Face Dataset (HDTF) is a large in-the-wild audio-visual dataset for talking face generation. It consists of about 362 different high-resolution (720P or 1080P) YouTube videos of 15.8 hours in total. Using the download and processing script provided by the authors, we extracted 352 videos with 246 unique subjects. We additionally crop the video to 30 seconds long and use them for extracting head-poses using the MICA tracker [68], which provides head poses as global axis rotation. For our experiments, we split the dataset into 300/20/32 sequences for train/val/test accordingly.

Discussion: In this work, we employ the VOCAset, HDTF, and Imitator's in-the-wild dataset to train our method for generating and editing 3D facial animations with head motion. The motivation of generating and editing holistic 3D facial animation made both BIWI [19] and BEAT [34] incompatible for our study, both BIWI and BEAT are in different model spaces compared to the existing Face trackers like [12,20,68], which is a key necessity for personalization and subsequently face motion editing. Such a problem could in theory be addressed by converting the meshes provided in the dataset to our target FLAME model space by optimization-based fitting using pre-defined correspondence between the source and target mesh space. However, for BIWI a combination of weaker noisy surface reconstruction provided in the dataset and not fully completed face model makes this fitting challenging and reduces the quality of the fitted meshes further. Similarly, for BEAT the dependence on ARKit which produces improper lip-closures and not fully completed face model, reduces the realism of the reconstructed sequences. As studied in [53], lip-closures are paramount in conveying realism for the generated sequences.

9.3 Baselines

Holistic 3D motion synthesis: For TalkShow [60], we use the pre-trained model provided in their official repository and extract the predicted facial and head motion parameters for our evaluation. For SadTalker [64], we use the pre-trained model provided in the repository to generate 2D talking face videos and use the MICA tracker [68] to the face and head motion.

Facial Motion Synthesis: For VOCA [10], Faceformer [18], Imitator [53] and FaceDiffuser [48], we use the pre-trained model provided in the official repositories. For CodeTalker [58], we adapt the official implementation to add the functionality of generating diverse motion. Especially, we re-train the audio-conditioned codebook sampling (stage 02) to randomly sample a code from the top 'm' closest codes instead of always using the closest code. This process is in spirit close to training the language-based models, where a new diverse text sequence is generated by sampling the 2nd or 3rd closest language token over the token with maximum probability. By adapting this method, we ensure that CodeTalker can generate diverse samples for a given audio input. For EMOTE [13], we request the authors to run their method on the VOCAset [10] and use it for the qualitative and perceptual user study.

9.4 Training Details

Facial Motion Synthesis: We train our method using ADAM [32] with a learning rate of *1e-4* for 140K iterations with a batch size of 64. Our diffusion framework is based on the Gaussian diffusion from Nichol *et al.* [37], we set the diffusion step to 500 for our experiments. During training, we randomly crop the sequences to the length of 30 frames. Our lightweight architecture enables us to train our model on a single Nvidia Quadro P6000 32GB within 30 hours. The lightweight architecture is also critical for person-specific style adaptation with a short reference video. For person-specific speaking style, we use the same training setup as from the generalized setting, except that we only train it for 30K iterations. For evaluating the best checkpoint, we fix the guidance scale $s = 0.99$ and evaluate all the saved checkpoints on the validation set. Further, we fix the best checkpoint and vary the guidance scale from *s=0, 0.1 ... to 1.0* with an increment of 0.1 and find the best guidance factor. From our experiment, we found the guidance scale of 0.5 balances the lip-synchronization and diversity and provides the best results.

Head Motion Synthesis Similar to the Facial motion synthesis pipeline, we train our method using ADAM [32] with a learning rate of *1e-4* for 100K iterations with a batch size of 64. During training the sequences are randomly cropped to 300 frames long. For our inbetweening and keyframing-based Guided motion model training, we randomly sample a mask of arbitrary length for imputation signal, using which the noisy input is replaced with the ground truth imputation signal.

9.5 Inference

Our method takes 3.15 sec to produce 1 sec (30 frame) of facial motion and 1.04 sec to produce 1 sec (30 frame) of head motion on a single Nvidia GeForce RTX 3090 24GB, compared to 5.78 sec for the concurrent method FaceDiffuser [48]. In total, our method takes 4.19 sec to produce 1 sec (30 frame) of holistic 3D facial animation, compared to 6.78 sec for TalkSHOW [60].

9.6 Metrics

Lip-Sync measures the lip synchronization using Dynamic Time Warping to compute the temporal similarity [53].

Diversity metric introduced by Ren et al. [41] measures the diversity of 3D motions for the same text input. We employ this metric and propose Div^L and Div^H to measure the diversity of lip motion and head motions generated from the same audio. Given a set of generated 3D facial or head motions with N sequences generated from the same audio condition. The diversity can be formalized as:

$$Diversity = \frac{1}{L} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \|m_i - m_j\|_2 \tag{8}$$

Where m_i represents the i-th motion and L is the total number of possible combinations in the generated motion set.

Beat alignment (BA) : Similar to DiffPoseTalk [49], we employ a modified beat alignment BA to measure the synchronization of the head movement beats between the predicted and ground truth motion, where we calculate the average temporal distance between beat in predicted head movement its closest ground truth beat as the Beat Align Score.

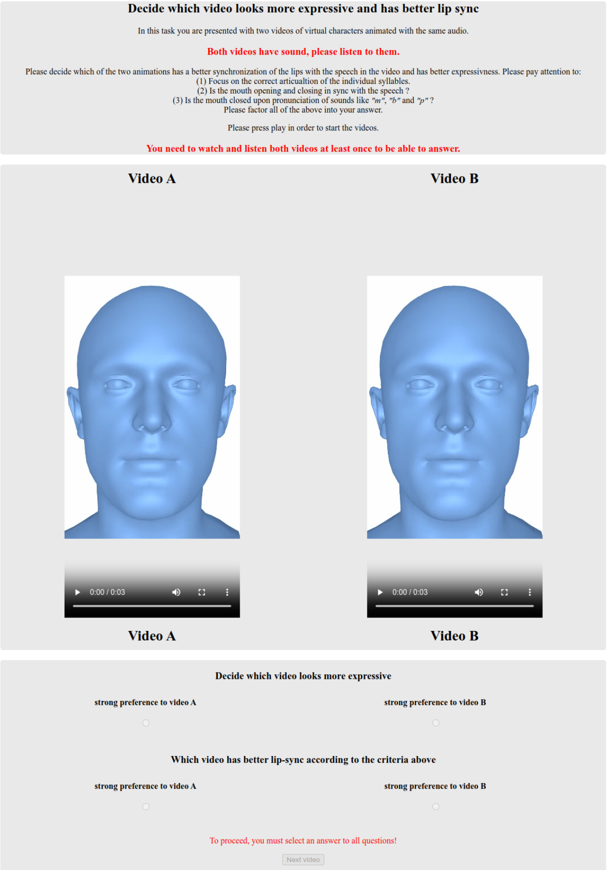
$$\text{Beat Align Score} = \frac{1}{|\mathbf{B}_g|} \sum_{t_g \in \mathbf{B}_g} \exp\left(-\frac{\min_{t_p \in \mathbf{B}_p} \|t^p - t^g\|_2^2}{2\sigma^2}\right), \tag{9}$$

Where $\mathbf{B}_g$ and $\mathbf{B}_p$ record the time of the beats in the ground truth and predicted head motion respectively, while σ is the normalized parameter which is set to be 3 in our experiment.

Discussion The $L2$-based vertex error metrics employed in previous studies [18, 53, 58] are not apt for our task due to its preference for solutions that are close to the mean of the dataset, which penalizes the diversity present in our predictions.

9.7 Perceptual Study

We conducted A/B user studies to assess our method's perceptual performance. First, we conducted a study to evaluate the holistic motion synthesis based on the naturalness of the facial and head motion to the input audio. For this, we sampled 10 sequences from the test set of the HDTF and 10 external audio from YouTube and synthesized holistic 3D facial motion using our method and the baselines [60, 64] resulting in a total of 60 A/B comparisons including ground truth. For extracting the ground truth for the YouTube sequences, similar to the HDTF dataset processing we utilize the MICA tracker [68] to extract the facial and head motion. Though the head motion extracted using MICA is smooth and natural, the tracker might produce jittery or under-articulated facial motion based on the extremity of head poses in the sequence, this underlines the finding that 78% of the users prefer the facial motion predicted by our method over the ground truth (refer to Table 4 User study in the main paper). Second for facial motion synthesis, we sample 20 sequences combined from the VOCAset test set and the in-the-wild sequences from Imitator, resulting in 100 A/B comparisons across five baselines. On Amazon Mechanical Turk(AMT), we divided the A/B comparisons into 5 HITs (Human Intelligence Task), each with 25 individual assignments. For each HIT, users select their preference for a method based on expressiveness and lip-synchronization. Finally, we evaluated the speaking style preservation of our personalized model in comparison to Imitator. To this end, the AMT users rated the similarity based on a reference video and the synthesized videos of the VOCA test set. Figure 11 illustrates an example interface in our user-study.



Decide which video looks more expressive and has better lip sync

In this task you are presented with two videos of virtual characters animated with the same audio.

Both videos have sound, please listen to them.

Please decide which of the two animations has a better synchronization of the lips with the speech in the video and has better expressivness. Please pay attention to:
(1) Focus on the correct articualtion of the individual syllables.
(2) Is the mouth opening and closing in sync with the speech ?
(3) Is the mouth closed upon pronunciation of sounds like *"m"*, *"b"* and *"p"* ?
Please factor all of the above into your answer.

Please press play in order to start the videos.

You need to watch and listen both videos at least once to be able to answer.

Video A **Video B**

Video A **Video B**

Decide which video looks more expressive

strong preference to video A ○ **strong preference to video B** ○

Which video has better lip-sync according to the criteria above

strong preference to video A ○ **strong preference to video B** ○

To proceed, you must select an answer to all questions!

Fig. 11: Example of the interface employed for our user-study.

10 Broader Impact

We introduce a method for realistic facial animation synthesis and editing that matches the speaking style of any given target actor. These animations hold promise for driving virtual avatars in AR or VR settings, especially, in immersive communication technologies. Yet, it is essential to acknowledge the potential pitfalls of such advancements, notably in the realm of 'DeepFakes.' By employing voice cloning techniques, our method can generate 3D facial animations that drive digital avatar methods like [3, 22, 23, 28, 67], which could be abused for identity theft, cyberbullying, and various criminal activities. Advocating for transparent research practices, we strive to illuminate the risks associated with technology misuse. Sharing our implementation aims to foster research in digital multimedia forensics, particularly in developing synthesis methods crucial for training data utilized in spotting forgeries [44].

References

1. Aneja, S., Thies, J., Dai, A., Nießner, M.: Facetalk: Audio-driven motion diffusion for neural parametric head models (2023) 2, 4
2. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020) 4, 5, 6
3. Bharadwaj, S., Zheng, Y., Hilliges, O., Black, M.J., Abrevaya, V.F.: FLARE: Fast learning of animatable and relightable mesh avatars. ACM Transactions on Graphics **42**, 15 (Dec 2023). https://doi.org/https://doi.org/10.1145/3618401 23
4. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th annual conference on Computer graphics and interactive techniques. pp. 187–194 (1999) 4
5. Cao, Y., Tien, W.C., Faloutsos, P., Pighin, F.: Expressive speech-driven facial animation. ACM Trans. Graph. **24**(4), 1283–1302 (oct 2005). https://doi.org/10.1145/1095878.1095881 4
6. Chen, P., Wei, X., Lu, M., Zhu, Y., Yao, N., Xiao, X., Chen, H.: Diffusiontalker: Personalization and acceleration for speech-driven 3d face diffuser (2023) 2, 4
7. Chung, H., Sim, B., Ryu, D., Ye, J.C.: Improving diffusion models for inverse problems using manifold constraints. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) Advances in Neural Information Processing Systems (2022) 2, 13
8. Chung, J.S., Jamaludin, A., Zisserman, A.: You said that? arXiv preprint arXiv:1705.02966 (2017) 4
9. Cohen, M.M., Clark, R., Massaro, D.W.: Animated speech: research progress and applications. In: Proc. Auditory-Visual Speech Processing. p. 200 (2001) 2
10. Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., Black, M.J.: Capture, Learning, and Synthesis of 3D Speaking Styles. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10093–10103. IEEE, Long Beach, CA, USA (Jun 2019). https://doi.org/10.1109/CVPR.2019.01034 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 17, 19, 20
11. Dai, Z., Yang, Z., Yang, Y., Carbonell, J.G., Le, Q.V., Salakhutdinov, R.: Transformer-xl: Attentive language models beyond a fixed-length context. ArXiv **abs/1901.02860** (2019), https://api.semanticscholar.org/CorpusID:57759363 6
12. Danecek, R., Black, M.J., Bolkart, T.: EMOCA: Emotion driven monocular face capture and animation. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20311–20322 (2022) 4, 20
13. Daněček, R., Chhatre, K., Tripathi, S., Wen, Y., Black, M., Bolkart, T.: Emotional speech-driven animation with content-emotion disentanglement. ACM (Dec 2023). https://doi.org/10.1145/3610548.3618183 2, 3, 4, 10, 12, 20
14. De Martino, J.M., Pini Magalhães, L., Violaro, F.: Facial animation based on context-dependent visemes. Computers & Graphics **30**(6), 971–980 (Dec 2006). https://doi.org/10.1016/j.cag.2006.08.017 4
15. Edwards, P., Landreth, C., Fiume, E., Singh, K.: Jali: an animator-centric viseme model for expressive lip synchronization. ACM Transactions on graphics (TOG) **35**(4), 1–11 (2016) 2, 4

16. Egger, B., Smith, W.A., Tewari, A., Wuhrer, S., Zollhoefer, M., Beeler, T., Bernard, F., Bolkart, T., Kortylewski, A., Romdhani, S., et al.: 3d morphable face models—past, present, and future. ACM Transactions on Graphics (TOG) **39**(5), 1–38 (2020) 4
17. Ezzat, T., Poggio, T.: MikeTalk: a talking facial display based on morphing visemes. In: Proceedings Computer Animation '98 (Cat. No.98EX169). pp. 96–102. IEEE Comput. Soc, Philadelphia, PA, USA (1998). https://doi.org/10.1109/CA.1998.681913 4
18. Fan, Y., Lin, Z., Saito, J., Wang, W., Komura, T.: Faceformer: Speech-driven 3d facial animation with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18770–18780 (2022) 2, 4, 5, 6, 10, 12, 14, 20, 22
19. Fanelli, G., Gall, J., Romsdorfer, H., Weise, T., Gool, L.V.: A 3-d audio-visual corpus of affective communication. IEEE Transactions on Multimedia **12**(6), 591 – 598 (October 2010) 9, 20
20. Filntisis, P.P., Retsinas, G., Paraperas-Papantoniou, F., Katsamanis, A., Roussos, A., Maragos, P.: Visual speech-aware perceptual 3d facial expression reconstruction from videos (2022) 20
21. Gafni, G., Thies, J., Zollhöfer, M., Nießner, M.: Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. CoRR **abs/2012.03065** (2020), https://arxiv.org/abs/2012.03065 4
22. Gafni, G., Thies, J., Zollhöfer, M., Nießner, M.: Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8649–8658 (June 2021) 23
23. Grassal, P.W., Prinzler, M., Leistner, T., Rother, C., Nießner, M., Thies, J.: Neural head avatars from monocular rgb videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18653–18664 (2022) 23
24. Guo, Y., Chen, K., Liang, S., Liu, Y., Bao, H., Zhang, J.: Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 4
25. Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., Y. Ng, A.: DeepSpeech: Scaling up end-to-end speech recognition (12 2014) 4
26. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020) 19
27. Ho, J., Salimans, T.: Classifier-free diffusion guidance (2022) 3, 5, 14, 19
28. Kabadayi, B., Zielonka, W., Bhatnagar, B.L., Pons-Moll, G., Thies, J.: Gan-avatar: Controllable personalized gan-based human head avatar. In: International Conference on 3D Vision (3DV) (March 2024) 23
29. Kalberer, G., Van Gool, L.: Face animation based on observed 3D speech dynamics. In: Proceedings Computer Animation 2001. Fourteenth Conference on Computer Animation (Cat. No.01TH8596). pp. 20–251. IEEE Comput. Soc, Seoul, South Korea (2001). https://doi.org/10.1109/CA.2001.982373 4
30. Karras, T., Aila, T., Laine, S., Herva, A., Lehtinen, J.: Audio-driven facial animation by joint end-to-end learning of pose and emotion. ACM Transactions on Graphics **36**(4), 1–12 (Jul 2017). https://doi.org/10.1145/3072959.3073658 4
31. Karunratanakul, K., Preechakul, K., Suwajanakorn, S., Tang, S.: Guided motion diffusion for controllable human motion synthesis. In: Proceedings of the

IEEE/CVF International Conference on Computer Vision. pp. 2151–2162 (2023) 2, 3, 8, 13

32. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2017) 21
33. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4D scans. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **36**(6) (2017) 7, 9
34. Liu, H., Zhu, Z., Iwamoto, N., Peng, Y., Li, Z., Zhou, Y., Bozkurt, E., Zheng, B.: Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. arXiv preprint arXiv:2203.05297 (2022) 20
35. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Gool, L.V.: Repaint: Inpainting using denoising diffusion probabilistic models (2022) 11
36. Ma, J., Bai, S., Zhou, C.: Pretrained diffusion models for unified human motion synthesis. arXiv preprint arXiv:2212.02837 (2022) 7, 13
37. Nichol, A., Dhariwal, P.: Improved denoising diffusion probabilistic models (2021) 21
38. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3d human pose estimation in video with temporal convolutions and semi-supervised training (2019) 6
39. Peng, Z., Wu, H., Song, Z., Xu, H., Zhu, X., He, J., Liu, H., Fan, Z.: Emotalk: Speech-driven emotional disentanglement for 3d face animation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20687–20697 (2023) 2
40. Raab, S., Leibovitch, I., Li, P., Aberman, K., Sorkine-Hornung, O., Cohen-Or, D.: Modi: Unconditional motion synthesis from diverse data (2022) 14
41. Ren, Z., Pan, Z., Zhou, X., Kang, L.: Diffusion motion: Generate text-guided 3d human motion by diffusion model (2023) 10, 21
42. Richard, A., Zollhofer, M., Wen, Y., de la Torre, F., Sheikh, Y.: MeshTalk: 3D Face Animation from Speech using Cross-Modality Disentanglement. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1153–1162. IEEE, Montreal, QC, Canada (Oct 2021). https://doi.org/10.1109/ICCV48922.2021.00121 2, 4
43. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022) 2, 7
44. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. ICCV 2019 (2019) 23
45. Schneider, S., Baevski, A., Collobert, R., Auli, M.: wav2vec: Unsupervised pretraining for speech recognition. In: Kubin, G., Kacic, Z. (eds.) Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019. pp. 3465–3469. ISCA (2019). https://doi.org/10.21437/Interspeech.2019-1873, https://doi.org/10.21437/Interspeech.2019-1873 4
46. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning. pp. 2256–2265. PMLR (2015) 5, 17
47. Song, L., Wu, W., Qian, C., He, R., Loy, C.C.: Everybody's talkin': Let me talk as you want. IEEE Transactions on Information Forensics and Security **17**, 585–598 (2022) 4
48. Stan, S., Haque, K.I., Yumak, Z.: Facediffuser: Speech-driven 3d facial animation synthesis using diffusion. In: ACM SIGGRAPH Conference on Motion, Interaction

and Games (MIG '23), November 15–17, 2023, Rennes, France. ACM, New York, NY, USA (2023). https://doi.org/10.1145/3623264.3624447 2, 4, 10, 11, 12, 14, 20, 21

49. Sun, Z., Lv, T., Ye, S., Lin, M.G., Sheng, J., Wen, Y.H., Yu, M., jin Liu, Y.: Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models (2023) 2, 4, 5, 6, 7, 11, 17, 22
50. Suwajanakorn, S., Seitz, S.M., Kemelmacher-Shlizerman, I.: Synthesizing obama: learning lip sync from audio. ACM Transactions on Graphics (ToG) **36**(4), 1–13 (2017) 4
51. Taylor, S.L., Kim, T., Yue, Y., Mahler, M., Krahe, J., Rodriguez, A.G., Hodgins, J.K., Matthews, I.A.: A deep learning approach for generalized speech animation. ACM Trans. Graph. **36**(4), 93:1–93:11 (2017). https://doi.org/10.1145/3072959.3073699, https://doi.org/10.1145/3072959.3073699 4
52. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., Bermano, A.H.: Human motion diffusion model. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=SJ1kSyO2jwu 5, 7, 14, 17
53. Thambiraja, B., Habibie, I., Aliakbarian, S., Cosker, D., Theobalt, C., Thies, J.: Imitator: Personalized speech-driven 3d facial animation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20621–20631 (October 2023) 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 19, 20, 21, 22
54. Thies, J., Elgharib, M., Tewari, A., Theobalt, C., Nießner, M.: Neural voice puppetry: Audio-driven facial reenactment. ECCV 2020 (2020) 4
55. Vougioukas, K., Petridis, S., Pantic, M.: Realistic speech-driven facial animation with gans. International Journal of Computer Vision **128**(5), 1398–1413 (2020) 4
56. Wang, K., Wu, Q., Song, L., Yang, Z., Wu, W., Qian, C., He, R., Qiao, Y., Loy, C.C.: Mead: A large-scale audio-visual dataset for emotional talking-face generation. In: ECCV (2020) 4
57. Wang, S., Li, L., Ding, Y., Fan, C., Yu, X.: Audio2head: Audio-driven one-shot talking-head generation with natural head motion. In: International Joint Conference on Artificial Intelligence. IJCAI (2021) 4
58. Xing, J., Xia, M., Zhang, Y., Cun, X., Wang, J., Wong, T.T.: Codetalker: Speech-driven 3d facial animation with discrete motion prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12780–12790 (2023) 2, 4, 5, 6, 10, 11, 12, 14, 20, 22
59. Yao, S., Zhong, R., Yan, Y., Zhai, G., Yang, X.: Dfa-nerf: Personalized talking head generation via disentangled face attributes neural rendering. arXiv preprint arXiv:2201.00791 (2022) 4
60. Yi, H., Liang, H., Liu, Y., Cao, Q., Wen, Y., Bolkart, T., Tao, D., Black, M.J.: Generating holistic 3d human motion from speech. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 469–480 (June 2023) 4, 9, 10, 14, 20, 21, 22
61. Yi, R., Ye, Z., Zhang, J., Bao, H., Liu, Y.J.: Audio-driven talking face video generation with learning-based personalized head pose. arXiv preprint arXiv:2002.10137 (2020) 4
62. Zhang, C., Ni, S., Fan, Z., Li, H., Zeng, M., Budagavi, M., Guo, X.: 3d talking face with personalized pose dynamics. IEEE Transactions on Visualization and Computer Graphics (2021) 2
63. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. arXiv preprint arXiv:2208.15001 (2022) 7

64. Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., Shan, Y., Wang, F.: Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. arXiv preprint arXiv:2211.12194 (2022) 4, 9, 10, 14, 20, 22
65. Zhang, Z., Li, L., Ding, Y., Fan, C.: Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3661–3670 (2021) 4, 9, 15, 17, 20
66. Zhou, Y., Han, X., Shechtman, E., Echevarria, J., Kalogerakis, E., Li, D.: Makelttalk: speaker-aware talking-head animation. ACM Transactions on Graphics (TOG) **39**(6), 1–15 (2020) 4
67. Zielonka, W., Bolkart, T., Thies, J.: Instant volumetric head avatars. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4574–4584 (2022) 23
68. Zielonka, W., Bolkart, T., Thies, J.: Towards metrical reconstruction of human faces. ECCV (2022). `https://doi.org/10.48550/ARXIV.2204.06607` 7, 9, 19, 20, 22