3DiFACE: Synthesizing and Editing Holistic 3D Facial Animation

Balamurugan Thambiraja[1,3] Malte Prinzler[1,2,4] Sadegh Aliakbarian[5] Darren Cosker[5] Justus Thies[1,3]
[1]Max Planck Institute for Intelligent Systems [2]Max Planck ETH Center for Intelligent Systems
[3]Technical University of Darmstadt [4]ETH Zürich [5]Microsoft Mixed Reality & AI Lab, UK

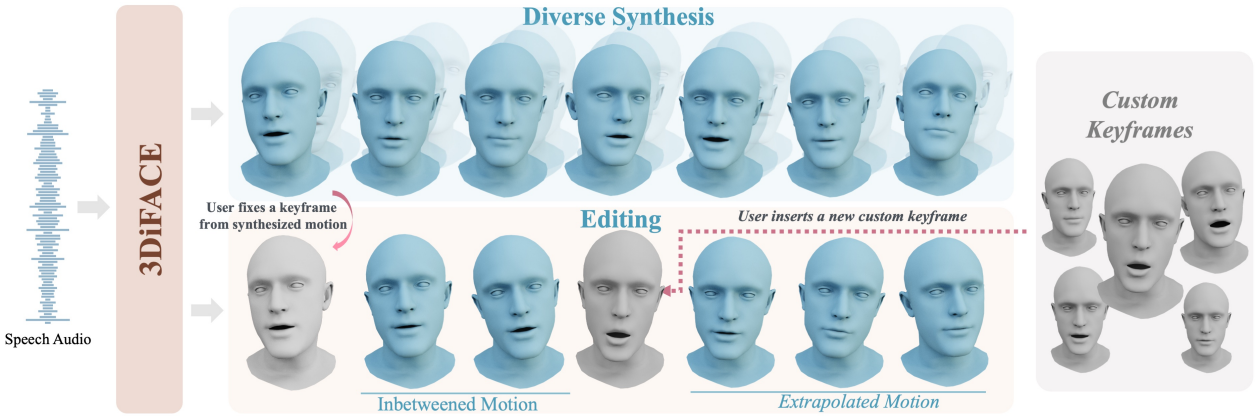

Figure 1. *3DiFACE* is a novel diffusion-based method for synthesizing holistic 3D facial animation from an audio input (top). In addition, users can seamlessly edit a synthesized or existing facial animation by defining part of the input as keyframes or by inserting new custom keyframes. These custom keyframes can be either manually created or sourced from an existing motion database.

Abstract

Creating personalized 3D animations with precise control and realistic head motions remains challenging for current speech-driven 3D facial animation methods. Editing *these animations is especially complex and time consuming, requires precise control and typically handled by highly skilled animators. Most existing works focus on controlling style or emotion of the synthesized animation and cannot edit/regenerate parts of an input animation. They also overlook the fact that multiple plausible lip and head movements can match the same audio input. To address these challenges, we present 3DiFACE, a novel method for holistic speech-driven 3D facial animation. Our approach produces* diverse plausible lip and head motions *for a single audio input and allows for* editing via keyframing and interpolation. *Specifically, we propose a fully-convolutional diffusion model that can leverage the viseme-level diversity in our training corpus. Additionally, we employ a speaking-style personalization and a novel sparsely-guided motion diffusion to enable precise control and editing. 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. Project page: https://balamuruganthambiraja.github.io/3DiFACE*

1. Introduction

Holistic 3D facial animation transforms digital figures into expressive characters, pivotal for compelling narratives in films and games. Artists craft these animations with great detail, using precise and iterative editing to ensure every glance and nod adds to the narrative. In this context, 'holistic' refers to 3D facial animation that includes both lip and head movements. Early works [5, 12] used procedural-rule based systems to map audio features with facial animation parameters, giving artists precise control and the ability to edit specific-parts of the animation sequence. However, this process is manual and labour intensive.

With advancements in machine learning, new learning-based methods have emerged that allow for quicker audio-driven facial animations [6, 16, 32, 34, 51, 53]. However, these methods mainly focus on controlling the emotion [9, 32] and style [6, 16, 34, 47, 51] of the animation. They cannot allow users to easily edit specific-parts of the animation sequence. I.e., if a user wants to edit the style of a part of an existing sequence, they have to generate a new sequence with desired style and then blend it with the original. Such an edit is often impractical, time-consuming and requires frame-by-frame manual inspection to ensure accurate lip-sync. Notably, diffusion-based facial motion synthesis methods [1, 4, 40, 42], where such an editing

could be considered as a byproduct of diffusion models, are not demonstrating this. Further, recent works [1, 4, 40] focus on showcasing diversity in eye-blinks and upper face motion, which have a weak (if any) correlation with the audio. However, the ability to lip-sync animations in varied but plausible ways is also essential, especially for the animation and movie dubbing. Because of these shortcomings of learning-based methods, procedural methods, though tedious and labor-intensive, still dominate 3D animation in movies and gaming, highlighting the need for a more efficient alternative.

Our goal is to achieve diverse motion synthesis with precise control from partial control signals like keyframes (also referred as Imputation signal). In this context, we face three main challenges: (i) Facial movements are highly person-specific. If the speaking style of the edited region isn't personalized and doesn't match the imputation signal, it results in unrealistic animations due to sudden style shifts between edited and unedited motion (see Figure 2). (ii) Diffusion models are known to require large training sets [36], yet the size of existing high-quality speech-to-3D-animation datasets is limited. Additionally, for personalization of speaking-style the model should be capable of fine-tuning on a short reference video ($1min$). (iii) Standard diffusion-based editing on head motion often ignores the imputation signal, as also observed in [25] (see Figure 2).

To tackle these challenges, we propose a diffusion-based method for speech-driven holistic 3D facial animation synthesis and editing. Specifically, to address the aforementioned challenges (i) and (ii), we employ a fully convolutional 1D U-net architecture that can be trained on the small VOCAset [6] and fine-tuned using a short $1min$ reference video of the target subject. The fully convolutional design of our method allows to sub-divide the input sequence into $1sec$ viseme-level motion segments during training and generalize to sequences of arbitrary length at inference time. Specifically, we leverage the viseme-level motion diversity present in the dataset to train our method and generate diverse sequence-level samples for a single audio input at inference. For head motion editing (iii), we introduce a novel Sparsely-Guided Diffusion (SGDiff) approach. This method involves replacing part of the noisy sequence with ground truth data and enforcing the model to precisely replicate the samples within the ground truth region. This approach prevents the model from ignoring the sparse imputation signal, resulting in smoother and more 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. Further, we demonstrate the importance of our architecture design choices, data-efficiency and robustness in detailed ablation studies.

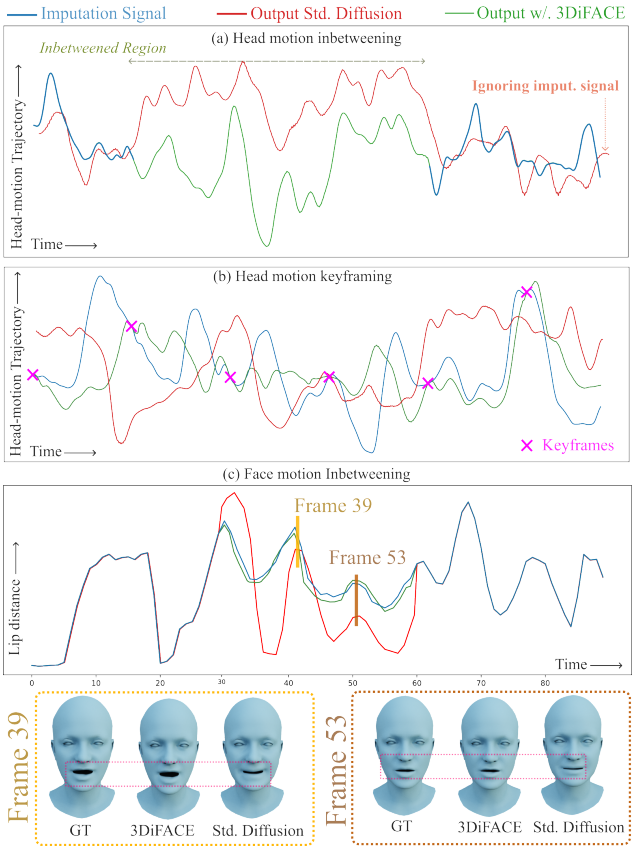

Figure 2. Illustration of holistic 3D facial motion editing with and without *3DiFACE*. Head motion editing is shown in (a) and (b), where one can see that standard diffusion is ignoring the imputation signal. Facial motion editing (c) shows the unrealistic style-shifts for classical diffusion, refer *Frame 39* and *Frame 53*.

In summary, our contributions are twofold:

- a fully convolutional speech-driven diffusion model that leverages viseme-level diversity to synthesize diverse holistic 3D facial animations of arbitrary length.
- employing personalization and sparsely-guided diffusion, we demonstrate explicit 3D facial animation editing, including seamless motion interpolation and keyframing.

2. Related Work

Speech-Driven 3D Facial Animation: Methods for 3D facial animation synthesis can be classified into 2 categories namely procedural and learning-based methods. *Procedural methods:* Earlier works on 3D facial animation used procedural rules [10, 12, 15, 23] to map audio to facial rigs. The methods offer artists precise control and ability to modify part of the animation. Despite being manual and labor-intensive, they remain the standard for 3D animation in movies and gaming. *Learning-based methods:* using motion-captured 3D audio visual datasets like

VOCA [6] or BIWI [17], statistical approaches [3, 6, 9, 16, 22, 24, 31, 34, 39, 41, 45, 47, 48] learn to animate 3D meshes or blendshapes from audio inputs. They produce facial animation with high lip-synchronization. However they struggle to model head-motion and the many-to-many mapping between audio and facial animations. Concurrent to our work, several methods [1, 4, 40, 42] employ Diffusion-probabilistic models [20] to learn and synthesize 3D facial animation with diversity. DiffPoseTalk [42] and Media2Face [58] are closest to our 3D facial animation method with head-motion synthesis. DiffPoseTalk [42] synthesizes personalized 3D facial animation with head motion, however, it cannot offer any editing capabilities and the synthesis quality is limited by the tracker. Media2Face [58] utilizes an in-house 4D corpus over 200 subjects to learn a motion prior and subsequently builds a large in-the-wild training corpus with text annotations to train a diffusion model with various global control signals. In contrast to the above works, our approach focuses on synthesizing personalized holistic facial animation, that can be precisely controlled using user-defined key-frames, thus, bridging the fine control from procedural methods with diverse multi-modal synthesis from learning-based methods.

Concurrent works [29, 43, 44] have focused on improving the generalization to different language, removing topological constraints and synthesizing animation with authentic laughter. As this is not the focus of our work, we suggest interested readers to check out the respective works.

3D Holistic Facial motion Editing: We define facial motion editing as the task of explicitly editing an input sequence, by defining keyframes and regenerating selected parts of it. 3D animation artists can manually define these key-frames (e.g. from a database, or an already generated sequence) and refine them further to get desired expressions and poses at specific points in the motion sequence.

Procedural animation methods [10, 12, 15, 23] allow this precise control via modifying the animation curves in the interested parts of the sequence. However as mentioned in the previous section, they are labour intensive and limited in terms of animation styles. Learning-based methods offer to control style [1, 6, 16, 31, 34, 40, 42, 47, 51] and emotion [4, 9] during animation synthesis, however, they cannot regenerate or control part of the sequence. Further, diffusion-based facial motion synthesis methods [1, 4, 40, 42], where editing might be considered as a byproduct of diffusion models, are not demonstrating this. Design choices such as auto-regressive mechanisms [1, 4, 40, 42], self-attention with look-ahead masks [1, 42], and lack of speaking style personalization [1, 40] prevent these models from effectively editing facial motions. Concurrent work, Media2face [58] demonstrate the ability to locally edit facial animation in one of the sequence in the paper. However, several questions regarding the effectiveness, head-motion editing and adaptability to new subject remain open. In this work, we propose a diffusion-based holistic facial animation method that can synthesize 3D holistic animation from input audio and allowing to locally edit the generated or existing animation sequence. Further, through detailed experiments, we demonstrate the effectiveness of our method, adaptability to new subjects and ability to edit head-motions.

Diffusion Guidance: Diffusion models have significantly advanced generating images [19, 20, 35], videos [13, 18, 21, 37], audio [14, 26, 49], and motion [7, 25, 46, 50]. Control in these models is implemented through several methods: Classifier-guidance [11] uses a classifier's gradient, while classifier-free guidance [19] balances quality and diversity with conditional and unconditional models. ControlNet [54] employs a trainable copy of a diffusion model for processing conditions, and inpainting methods [25] generate consistent outputs from partial data. Guided motion diffusion(GMD) [25] a full-body motion synthesis method, is the work closest to ours. GMD support spatial guidance, cannot offer keyframing and control on sparse pose.

3. Preliminaries

Denoising Diffusion Probabilistic Models: Our method is based on the diffusion framework of Sohl et al. [38], where a training sample x_0 gradually transforms into white noise through the addition of Gaussian noise across T steps. Following Tevet et al. [46], we train a denoising model θ that can reverse the forward diffusion and estimate the original sample x_0 from a noised version x_t, conditioned on input C: $\hat{x}_0 = \theta(x_t, t, C)$. To generate new samples, we start from random noise x_T and apply iterative denoising until reaching $t = 0$. To improve the diversity of the samples during inference, we employ Classifier-Free Guidance (CFG) [19] and calculate the output as a weighted sum of the conditional and unconditional prediction:

$$\theta(x_t, t, C) := \theta(x_t, t, \emptyset) + s \cdot [\theta(x_t, t, C) - \theta(x_t, t, \emptyset)], \tag{1}$$

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

Audio Encoding: Similar to [6, 16, 47, 51], we adopt the pretrained Wav2Vec2.0 [2] to generate audio features from the 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 [6]). 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.

4. Method

Our goal is to synthesize and edit holistic 3D facial animation given an input audio signal. In this context, 'holistic' refers to animation with facial *and* head motion, which we model in two diffusion-based networks, see Figure 3. This is motivated by the fact that the facial motion is highly correlated to the speech signal, while the correlation w.r.t. the head motion is weaker and, thus, requires a longer context of information, hence, a different training scheme (and data). The two diffusion models for facial (θ_f) and head (θ_h) motion are conditioned on the encoded audio signal $\hat{A}$ using a pretrained Wav2Vec2.0 [2] as explained in Section 3. We leverage convolutional architectures for the denoising models which we describe in the following.

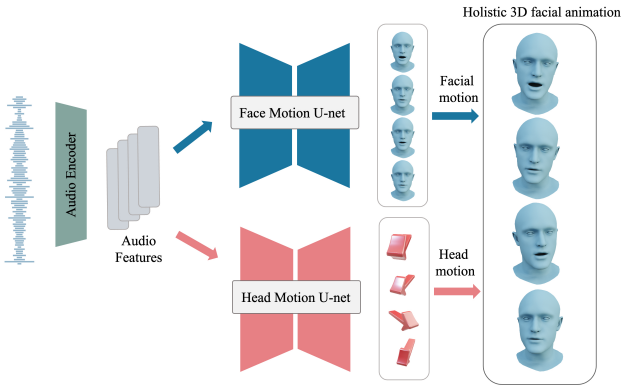

Figure 3. Overview of our method. We employ two diffusion-based motion generators with shared audio encoder to model 3D facial and head motion separately.

4.1. 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 4. 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 $\hat{x}_0 = \theta_f(x_t, t, C_f)$, given diffusion step t and conditions C_f. Note that in our formulation, the condition C_f represents the set of both the per-frame audio features $\hat{A}$ and person-specific feature S_i.

In contrast to state-of-the-art methods on 3D facial animation synthesis that utilize transformer architectures [16, 42, 47, 51], we adopt a 1D-convolution network inspired from Pavllo et al. [30]. Specifically, we replace the commonly used attention-based condition injection with feature concatenation. Our fully convolutional architecture, free from attention, allows us to sub-divide the input sequence into viseme-level motion segments (e.g., 30 frames) during training and to generalize to sequences of arbitrary length at inference time. We empirically observed that these modifications to the architecture are critical to train on the limited VOCA training dataset [6], especially in the unconditional training setup (see Table 3, row 1-4). 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 [8] and subpar performance (see Table 3, row 3). This issue becomes even more pronounced in our training setting, where the sequences have only 30 frames. While auto-regressive motion synthesis could in theory mitigate this limitation, it would make the animation editing tasks, such as motion inbetweeing, impossible.

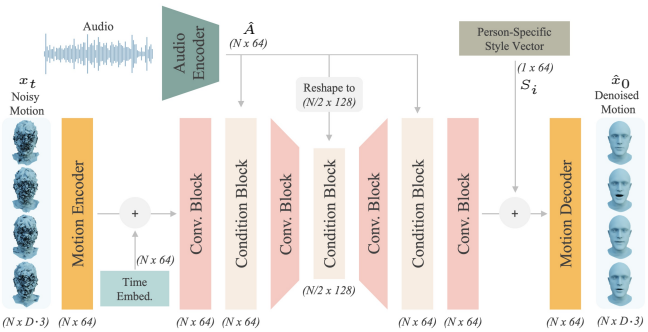

Figure 4. Our facial motion generator takes noised vertex displacements, denoted as x_t, and the diffusion time step embedding as inputs to predict a denoised sample $\hat{x}_0$, leveraging both the audio features signal $\hat{A}$ and a person-specific feature vector S_i. Note that N corresponds to the frame count of the sequence and D to the number of vertices.

4.1.1 Training

We train our model to predict the vertex displacements x_0 from their noised counterparts x_t on VOCAset [6]:

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

In contrast to predicting the applied noise which is common practice [28, 36, 55], we empirically found that predicting the ground truth displacements yields better convergence in the unconditional and person-specific fine-tuning case. To improve temporal smoothness [47], we add a velocity loss:

$$\mathcal{L}_{\text{vel}} = \frac{1}{N-1} \sum_{n=1}^{N} ||(x_{0,n} - x_{0,n-1}) - (\hat{x}_{0,n} - \hat{x}_{0,n-1})||^2, \tag{3}$$

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

$$\mathcal{L}_{\text{face}} = \mathcal{L}_{\text{simple}} + \lambda_{\text{vel}} \cdot \mathcal{L}_{\text{vel}}, \tag{4}$$

with $\lambda_{\text{vel}} = 10.0$. Note that during training, we randomly set the audio features $\hat{A}$ in the condition C_f to 0 for 10% of the time to enable unconditional synthesis at inference time.

4.1.2 Person-specific fine-tuning

As described in the introduction, personalization of speaking-style is indispensable for facial motion editing. For capturing the speaking style of a subject that is not part of the training set, we require a short reference video. The facial movements are reconstructed with the state-of-the-art monocular face tracker MICA [59]. We use the reconstructed 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. (4).

4.2. Head-motion generator with sparse guidance

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 [27], where the rotation is represented via axis angle. Motivated by the head-motion editing issue mentioned in the introduction, we introduce a sparsely-guided diffusion (SGDiff) for the head motion synthesis. Specifically, in addition to the audio features, we inject an intra-sequence guidance to highlight the relative importance of the different segments in the input signal. As illustrated in Figure 5, during the forward diffusion process, part of the noisy input (y_t) 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 sparse ground truth signals. For the audio encoding, 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 for diffusion step t is $\hat{y}_0 = \theta_h(y_t, t, \hat{A})$.

4.2.1 Training

The complete training procedure of the sparsely-guided diffusion is detailed in Algorithm 1. In addition to the losses used in the facial motion generator (Section 4.1.1), we add an additional guiding mask loss to enforce the model to faithfully reproduce the results of the ground truth signal injected into the sequence. The guidance loss is:

$$\mathcal{L}_{\text{mask}} = ||w_{0,n} \odot (y_{0,n} - \theta_h(y_{t,n}, t, \hat{A}))||^2, \qquad (5)$$

where n indicates the n^{th} frame in sequence y_0, $w_{0,n}$ is the guidance weight, 1 for ground truth frames and zero otherwise and $\odot$ is the Hadamard product.

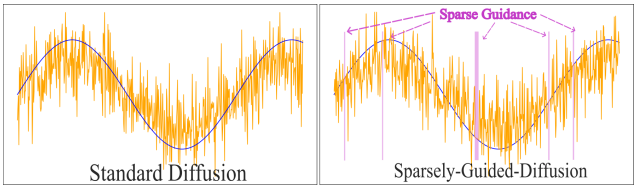

Figure 5. Illustration of standard diffusion (left) and our sparsely-guided diffusion (right), where in the forward diffusion process, part of the noisy input signal is replaced with the ground truth signal and a guidance flag of (0) and (1) is concatenated to the noisy and ground truth regions respectively.

Algorithm 1 Our SGDiff Training

repeat
  $y_0 \sim q(y_0)$ # sample from train distribution
  $t \sim \text{Uniform}(\{1, \dots, T\})$
  $\epsilon \sim \mathcal{N}(0, I)$
  $y_t = \sqrt{\bar{\alpha}_t} y_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$ # $\bar{\alpha}_t$ denotes diffusion noise schedule
  $\bar{y_t} = y_t \oplus (0)$ # $\oplus$ = concatenation operation
  $\bar{y_0} = y_0 \oplus (1)$
  $y_t = (1 - M_t) \odot \bar{y_t} + M_t \odot \bar{y_0}$ # Guidance injection
  grad desc. $\nabla_{\theta_h} \left\| y_0 - \theta_h(y_t, t, \hat{A}) \right\|^2$
until converged
M_t = random imputation mask

Algorithm 2 Our SGDiff Sampling

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

4.3. Sampling and editing of a 3D facial animation

Following standard diffusion methods [20, 46], we generate new samples by starting from random noise (X_T for facial motion and Y_T for head motion) drawn from Gaussian noise and iteratively denoise for T steps using the respective denoising models θ_f and θ_h. For control and part-regeneration, we replace the corresponding noisy sample (x_t or y_t) with the imputation signal at each time t before denoising. For facial motion editing, we personalize θ_f to the target subject using the steps detailed in Section 4.1.2 before iterative denoising. For head motion editing, we follow the procedure in Algorithm 2, replacing the input (y_t) with the imputation signal and adding a guidance flag.

5. Dataset

We train our facial motion model on VOCAset [6], since it provides high-quality, speech-aligned 3D face scan sequences. Following previous works [16, 47, 51], 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 without overlap with the speech scripts of the training sequences are used. We evaluate person-specific fine-tuning on in-the-wild videos from Imitator [47]. 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 [57] dataset, as the VOCAset does not include 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 [59] to extract head poses. For our experiments, we split the dataset into 300/20/32 sequences for train/val/test accordingly. 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 [17] from our study, as it lacks sequences with full head model like FLAME [27], which is essential for synthesizing head motion effectively. For more details refer the supplemental document.

6. Results

We evaluate our method against state-of-the-art methods: SadTalker [56] and TalkShow [52] on the holistic 3D facial animation synthesis task and VOCA [6], Faceformer [16], CodeTalker [51], EMOTE [9], FaceDiffuser [40] and Imitator [47] on facial motion synthesis task. Figure 6 presents the qualitative comparison on holistic 3D motion synthesis, where our method produces more accurate lip-synced facial animations with diverse head movements. Additional qualitative results are shown in the suppl. material and video.

Quantitative Comparison: In Table 1, 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 [47]. Diversity metric Div^L and Div^H proposed by Ren et al. [33] measures the diversity of lip motion and head motion generated from the same audio. Similar to DiffPoseTalk [42], we employ a modified beat alignment *BA* to measure the synchronization of the head movement beats. Please refer the suppl. material for detailed information about the metrics.

From Table 1 (rows 1-3), we see that our method significantly surpasses the baselines in *holistic 3D facial animation synthesis*, particularly in terms of lip-sync accuracy and beat alignment, while offering greater diversity.

For the *facial motion synthesis task without head motion*, we quantitatively compare our method on three different setups namely, non-personalized regression and diffusion and personalized synthesis. In the non-personalized regression and diffusion setup, our method outperforms the baselines, except for FaceDiffuser, where we match the performance on *Lip-Sync* despite producing more diverse samples (refer to Table 1 rows 4-10). Note that we can adjust the guidance scale parameter to control synthesis diversity and lip-sync accuracy, which FaceDiffuser [40] cannot do. Please refer the supplemental document for a more detailed study on the impact of guidance scale s. Finally, in the *personalization synthesis* setup, we achieve higher synthesis diversity compared to Imitator [47] and match the performance closely in terms of *Lip-Sync* (refer to Table 1 rows 11-13). Note that Imitator is a deterministic model that does not allow for diverse lip-motion synthesis and facial motion editing.

	Method	$\mathbf{Div^L}\uparrow$	Lip-Sync $\downarrow$	BA $\uparrow$	$\mathbf{Div^H}\uparrow$
		Holistic 3D Facial animation syn.			
1	SadTalker [56]	1.59	4.01	0.285	0.004
2	TalkSHOW [52]	1.80	4.35	0.296	0.002
3	**Ours composite**	**2.57**	**1.71**	**0.338**	**0.007**
		Non-Personalized regression			
4	VOCA [6]	–	5.30	–	–
5	Faceformer [16]	–	2.85	–	–
6	Imitator [47]	–	1.95	–	–
7	CodeTalker [51]	1.40	2.55	–	–
8	$\text{Ours}_{s=0.5}$ (w/o sty)	**2.57**	**1.71**	–	–
		Non-Personalized diffusion			
9	FaceDiffuser [40]	0.05	**1.60**	–	–
10	$\text{Ours}_{s=1.0}$ (w/o sty)	**0.64**	1.62	–	–
		Personalized synthesis			
11	Imitator (w/ sty)	–	**1.35**	–	–
12	$\text{Ours}_{s=0.5}$ (w/ sty)	**1.57**	1.56	–	–
13	$\text{Ours}_{s=1.0}$ (w/ sty)	0.24	1.42	–	–

Table 1. Quantitative comparison: Our proposed method produces better holistic 3D facial animations with high-fidelity lip and head motions (refer row 1-3). On the non-personalized regression and diffusion facial motion synthesis task (row 4-10), our method produces outperforms the baselines, except for FaceDiffuser, where we match the performance on *Lip-Sync* despite producing more diverse samples. Finally, our method is able to personalize facial motions on the level of Imitator [47], while producing more diverse samples and allowing for motion editing using keyframes.

User Study: We conducted A/B user studies to assess our method's perceptual performance. From Table 2 (row 1-2), we see that our method outperforms the baselines on the holistic 3D facial animation synthesis. For the facial motion synthesis task, we compare our method in 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 in the high diversity setup, we outperform CodeTalker [51] and perform closely to FaceDiffuser [40], which trades fidelity for

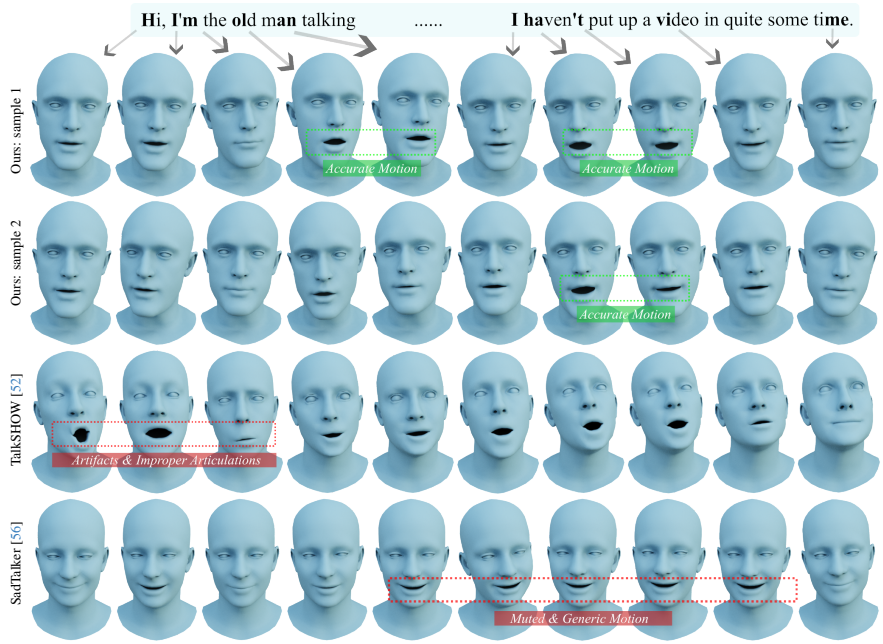

Figure 6. Qualitative comparison: Our method outperforms the baseline in creating more accurate lip-synced facial animations with diverse head movements. Specifically, TalkSHOW produces animations with jittery artifacts, while SadTalker yields muted and generic animations.

		Holistic synthesis	
	Method	**Face Motion** (%)	**Head motion** (%)
1	Ours vs SadTalker [56]	88.13	86.43
2	Ours vs TalkShow [52]	90.77	87.96
	Method	**Exprs** (%)	**Lip-sync** (%)
		High-Fidelity (Ours $s = 1.0$)	
3	Ours vs Imitator [47]	65.72	69.47
4	Ours vs Faceformer [16]	73.28	71.43
5	Ours vs FaceDiffuser [40]	67.85	66.71
		High-diversity (Ours $s = 0.5$)	
6	Ours vs CodeTalker [51]	53.64	53.80
7	Ours vs FaceDiffuser [40]	40.84	41.55

Table 2. *User study on holistic motion synthesis and facial motion synthesis task:* Compared to the baselines, our method produces consistently better holistic 3D facial animation in both low diversity($s = 0.5$) and high-fidelity setup($s = 1.0$). Similar to Imitator [47], we evaluate the person-specific speaking-style similarity against [47], where 55% of users favored our method.

diversity. Furthermore, we assessed style-personalization similar to Imitator [47]. The users rated the style-similarity based on a reference video and the synthesized videos on the VOCA test set, where 55% of the users preferred our method. For more details on the user-study, see suppl. mat.

Motion Editing: To demonstrate head-motion editing, we perform both keyframing and inbetweening on head-motion data and present the results in the Figure 2 (a) and (b). From which we infer that our sparsely-guided motion diffusion matches the imputation signal in both, the keyframing and inbetweening scenario and produces realistic motion in the edited/re-generated part of the sequence. Similarly, for facial motion editing, our method is able to match the speaking-style of the target imputation signal and produce smoothly edited sequences (refer Figure 2 (c)). We show additional examples and a unconditional motion synthesis and editing results in the suppl. mat. and video.

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 4, using our proposed architecture instead of the transformer-based architecture from Faceformer (FF) or the attention-based Unet architectures used in [28] 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 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.

• *Does the sparsely-guided motion-diffusion help to generate diverse motion?* In Table 3 rows 12-14, we analyze the effect of our keyframe (KF mask) and inbe-

	Method	$\mathbf{Div^L}\uparrow$	Lip-Sync $\downarrow$
		(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 ($\sim$ 5s)	29.95	4.89
6	Ours ($\sim$ 30s)	0.18	1.81
7	Ours ($\sim$ 60s)	0.67	1.69
8	Ours ($\sim$100s)	1.57	1.56
		(c) GMD ablation	
	Method	**BA** $\uparrow$	$\mathbf{Div^H}\uparrow$
9	Ours w. In mask	0.368	0.008
10	Ours w. KF mask	0.308	0.008
11	Ours w/o. mask	0.338	0.007

Table 3. Ablation study: (a) *Design choices* study on the VOCAset [6] shows the importance of a fully convolutional architecture and viseme-level training. (b) *Fine-tuning Data requirement:* 30s of video suffice to perform person-specific fine-tuning while 100s further improve all scores (row 5-9). (c) *GMD ablation:* shows the performance w.r.t the keyframing(sparse) and inbetweeen(dense) based imputation signal.

	Method	$\mathbf{Div^L}\uparrow$	Lip-Sync $\downarrow$	BA $\uparrow$	$\mathbf{Div^H}\uparrow$
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 4. Evaluation of editing on subject 024 in VOCAset [6] and HDTF [57] by varying the imputation signal. From the table, we observe significant improvements in synthesis quality with increase in imputation signal, indicating that the model closely matches the imputation signal and produces realistic motion.

tweening (In mask) based guidance on the synthesis quality. It demonstrates that employing sparsely-guided motion-diffusion improves the diversity of head-motion with minimal impact on overall quality, while offering additional editing capabilities.

• *Is the performance of motion editing consistent across varying degrees of imputation signal?* We evaluate the robustness of motion editing with respect to the imputation signal in both the inbetweening and keyframing scenario. 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. Further, we randomly insert keyframes at different rates: 1KF/sec, 2KF/sec, and 3KF/sec and fill the motion with our method. For facial motion, these evaluations are conducted for all sequences of the test subject 024 from the VOCAset [6], and the resulting metrics are presented in Table 4. Similarly for head motion, these evaluations are conducted in the HDTF [57] test set. For the Table 4, it is evident that as the imputation signal strength increases, the synthesis's fidelity improves, indicating that model matches the imputation signal and produces realistic motion. This allows animators to insert any number of keyframes for fine-grained control.

7. Discussion

Our proposed method excels in synthesizing and editing diverse holistic 3D facial animations based on speech. Similar to [47], for personalization, our method depends on the 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 monocular head trackers applied to in-the-wild videos. In this work, we employ a sparsely-guided motion diffusion to tackle the imputation signal neglect in the head-motion synthesis and editing. In contrast to head motion, for face motion editing, style-personalization is the critical contribution to enable seamless editing. For completeness, we include an experiment evaluating the personalization of head-motion and sparse-guidance for facial motion training in the supplemental documenet. 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.

8. Conclusion

With 3DiFACE, we presented a method that can both generate and edit diverse holistic 3D facial animations from speech input. Through detailed experiments, we demonstrated precise control and ability to edit parts of an animation sequence. Our work combines the precise control from procedural methods with the diverse multi-modal synthesis from learning-based methods. We believe that these properties will make 3DiFACE a powerful tool that can reduce production time and costs, making high-quality animations more accessible, helping create lifelike avatars for movies, games, and VR, expanding creative possibilities.

Acknowledgements: This project has received funding from the Mesh Labs, Microsoft, UK. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting B. Thambiraja. We would like to thank A. Cseke, T. Alexiadis, T. McConnell, for conducting the user studies.

References

[1] Shivangi Aneja, Justus Thies, Angela Dai, and Matthias Nießner. Facetalk: Audio-driven motion diffusion for neural parametric head models, 2023. 1, 2, 3

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 3, 4

[3] Yong Cao, Wen C. Tien, Petros Faloutsos, and Frédéric Pighin. Expressive speech-driven facial animation. *ACM Trans. Graph.*, 24(4):1283–1302, 2005. 3

[4] Peng Chen, Xiaobao Wei, Ming Lu, Yitong Zhu, Naiming Yao, Xingyu Xiao, and Hui Chen. Diffusiontalker: Personalization and acceleration for speech-driven 3d face diffuser, 2023. 1, 2, 3

[5] Michael M. Cohen, Rashid Clark, and Dominic W. Massaro. Animated speech: research progress and applications. In *Proc. Auditory-Visual Speech Processing*, page 200, 2001. 1

[6] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J. Black. Capture, Learning, and Synthesis of 3D Speaking Styles. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10093–10103, Long Beach, CA, USA, 2019. IEEE. 1, 2, 3, 4, 6, 8

[7] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[8] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *ArXiv*, abs/1901.02860, 2019. 4

[9] Radek Daněček, Kiran Chhatre, Shashank Tripathi, Yandong Wen, Michael Black, and Timo Bolkart. Emotional speech-driven animation with content-emotion disentanglement. ACM, 2023. 1, 3, 6

[10] José Mario De Martino, Léo Pini Magalhães, and Fábio Violaro. Facial animation based on context-dependent visemes. *Computers & Graphics*, 30(6):971–980, 2006. 2, 3

[11] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *CoRR*, abs/2105.05233, 2021. 3

[12] Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. Jali: an animator-centric viseme model for expressive lip synchronization. *ACM Transactions on graphics (TOG)*, 35(4):1–11, 2016. 1, 2, 3

[13] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models, 2023. 3

[14] Zach Evans, CJ Carr, Josiah Taylor, Scott H. Hawley, and Jordi Pons. Fast timing-conditioned latent audio diffusion, 2024. 3

[15] T. Ezzat and T. Poggio. MikeTalk: a talking facial display based on morphing visemes. In *Proceedings Computer Animation '98 (Cat. No.98EX169)*, pages 96–102, Philadelphia, PA, USA, 1998. IEEE Comput. Soc. 2, 3

[16] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18770–18780, 2022. 1, 3, 4, 6, 7

[17] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. Van Gool. A 3-d audio-visual corpus of affective communication. *IEEE Transactions on Multimedia*, 12(6):591 – 598, 2010. 3, 6

[18] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *International Conference on Learning Representations*, 2024. 3

[19] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. 3

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3, 5

[21] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022. 3

[22] Ke Gong Keze Wang Tianshui Chen Haojie Li Haifeng Zeng Wenxiong Kang Hui Fu, Zeqing Wang. Mimic: Speaking style disentanglement for speech-driven 3d facial animation. In *The 38th Annual AAAI Conference on Artificial Intelligence (AAAI)*, 2024. 3

[23] G.A. Kalberer and L. Van Gool. Face animation based on observed 3D speech dynamics. In *Proceedings Computer Animation 2001. Fourteenth Conference on Computer Animation (Cat. No.01TH8596)*, pages 20–251, Seoul, South Korea, 2001. IEEE Comput. Soc. 2, 3

[24] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics*, 36(4):1–12, 2017. 3

[25] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2151–2162, 2023. 2, 3

[26] Jean-Marie Lemercier, Julius Richter, Simon Welker, Eloi Moliner, Vesa Välimäki, and Timo Gerkmann. Diffusion models for audio restoration, 2024. 3

[27] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 5, 6

[28] Jianxin Ma, Shuai Bai, and Chang Zhou. Pretrained diffusion models for unified human motion synthesis. *arXiv preprint arXiv:2212.02837*, 2022. 4, 7

[29] Federico Nocentini, Thomas Besnier, Claudio Ferrari, Sylvain Arguillere, Stefano Berretti, and Mohamed Daoudi. Scantalk: 3d talking heads from unregistered scans, 2024. 3

[30] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training, 2019. 4

[31] Ziqiao Peng, Yihao Luo, Yue Shi, Hao Xu, Xiangyu Zhu, Hongyan Liu, Jun He, and Zhaoxin Fan. Selftalk: A self-supervised commutative training diagram to comprehend 3d talking faces. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 5292–5301, 2023. 3

[32] Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20687–20697, 2023. 1

[33] Zhiyuan Ren, Zhihong Pan, Xin Zhou, and Le Kang. Diffusion motion: Generate text-guided 3d human motion by diffusion model, 2023. 6

[34] Alexander Richard, Michael Zollhofer, Yandong Wen, Fernando de la Torre, and Yaser Sheikh. MeshTalk: 3D Face Animation from Speech using Cross-Modality Disentanglement. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1153–1162, Montreal, QC, Canada, 2021. IEEE. 1, 3

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 3

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 4

[37] Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *CVPR*, 2023. 3

[38] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 3

[39] Wenfeng Song, Xuan Wang, Shi Zheng, Shuai Li, Aimin Hao, and Xia Hou. Talkingstyle: Personalized speech-driven 3d facial animation with style preservation. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–12, 2024. 3

[40] Stefan Stan, Kazi Injamamul Haque, and Zerrin Yumak. 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*, New York, NY, USA, 2023. ACM. 1, 2, 3, 6, 7

[41] Yasheng Sun, Wenqing Chu, Hang Zhou, Kaisiyuan Wang, and Hideki Koike. Avi-talking: Learning audio-visual instructions for expressive 3d talking face generation. *IEEE Access*, 12:57288–57301, 2024. 3

[42] Zhiyao Sun, Tian Lv, Sheng Ye, Matthieu Gaetan Lin, Jenny Sheng, Yu-Hui Wen, Minjing Yu, and Yong jin Liu. Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models, 2023. 1, 3, 4, 6

[43] Kim Sung-Bin, Lee Chae-Yeon, Gihun Son, Oh Hyun-Bin, Janghoon Ju, Suekyeong Nam, and Tae-Hyun Oh. Multitalk: Enhancing 3d talking head generation across languages with multilingual video dataset. *arXiv preprint arXiv:2406.14272*, 2024. 3

[44] Kim Sung-Bin, Lee Hyun, Da Hye Hong, Suekyeong Nam, Janghoon Ju, and Tae-Hyun Oh. Laughtalk: Expressive 3d talking head generation with laughter, 2024. 3

[45] Sarah L. Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica K. Hodgins, and Iain A. Matthews. A deep learning approach for generalized speech animation. *ACM Trans. Graph.*, 36(4): 93:1–93:11, 2017. 3

[46] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 3, 5

[47] Balamurugan Thambiraja, Ikhsanul Habibie, Sadegh Aliakbarian, Darren Cosker, Christian Theobalt, and Justus Thies. Imitator: Personalized speech-driven 3d facial animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20621–20631, 2023. 1, 3, 4, 6, 7, 8

[48] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. *ECCV 2020*, 2020. 3

[49] Yuancheng Wang, Zeqian Ju, Xu Tan, Lei He, Zhizheng Wu, Jiang Bian, and Sheng Zhao. Audit: Audio editing by following instructions with latent diffusion models, 2023. 3

[50] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[51] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12780–12790, 2023. 1, 3, 4, 6, 7

[52] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 469–480, 2023. 6, 7

[53] Chenxu Zhang, Saifeng Ni, Zhipeng Fan, Hongbo Li, Ming Zeng, Madhukar Budagavi, and Xiaohu Guo. 3d talking face with personalized pose dynamics. *IEEE Transactions on Visualization and Computer Graphics*, 2021. 1

[54] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 3

[55] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 4

[56] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. *arXiv preprint arXiv:2211.12194*, 2022. 6, 7

[57] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. 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*, pages 3661–3670, 2021. 6, 8

[58] Qingcheng Zhao, Pengyu Long, Qixuan Zhang, Dafei Qin, Han Liang, Longwen Zhang, Yingliang Zhang, Jingyi Yu, and Lan Xu. Media2face: Co-speech facial animation generation with multi-modality guidance, 2024. 3

[59] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. ECCV, 2022. 5, 6