FaceTalk: Audio-Driven Motion Diffusion for Neural Parametric Head Models

Shivangi Aneja$^1$  Justus Thies$^{2,3}$  Angela Dai$^1$  Matthias Nießner$^1$

$^1$Technical University of Munich  $^2$MPI-IS, Tübingen  $^3$TU Darmstadt

Figure 1. Given an input speech signal, we propose a diffusion-based approach to synthesize high-quality and temporally consistent 3D motion sequences of high-fidelity human heads as neural parametric head models. Our method can generate a diverse set of expressions (including wrinkles and eye blinks) and the generated mouth motion is temporally synchronized with the given audio signal.

Abstract

We introduce FaceTalk$^1$, a novel generative approach designed for synthesizing high-fidelity 3D motion sequences of talking human heads from input audio signal. To capture the expressive, detailed nature of human heads, including hair, ears, and finer-scale eye movements, we propose to couple speech signal with the latent space of neural parametric head models to create high-fidelity, temporally coherent motion sequences. We propose a new latent diffusion model for this task, operating in the expression space of neural parametric head models, to synthesize audio-driven realistic head sequences. In the absence of a dataset with corresponding NPHM expressions to audio, we optimize for these correspondences to produce a dataset of temporally-optimized NPHM expressions fit to audio-video recordings of people talking. To the best of our knowledge, this is the first work to propose a generative approach for realistic and high-quality motion synthesis of volumetric human heads, representing a significant advancement in the field of audio-driven 3D animation. Notably, our approach stands out in its ability to generate plausible motion sequences that can produce high-fidelity head animation coupled with the NPHM shape space. Our experimental results substantiate the effectiveness of FaceTalk, consistently achieving superior and visually natural motion, encompassing diverse facial expressions and styles, outperforming existing methods by 75% in perceptual user study evaluation.

1. Introduction

Modeling 3D animation of humans has a wide range of applications in the realm of digital media, including animated movies, computer games, and virtual assistants. In recent years, there have been numerous works proposing generative approaches for motion synthesis of human bodies, enabling the animation of human skeletons conditioned on various signals such as action [3, 25, 46], language [1, 4, 36, 47, 64, 77] and music [2, 67]. While human faces are critical to synthesis of humans, generative synthesis of 3D faces in motion has focused on 3D morphable models (3DMMs) leveraging linear blendshapes [7, 40] to represent head motion and expression. Such models characterize a disentangled space of head shape and motion, but lack the capacity to comprehensively represent the complexity and fine-grained details of human face geometry in motion (e.g., hair, skin furrowing during motion, etc.).

Thus, we propose to represent animated head sequences...
with a volumetric 3D head representation, leveraging the expressive representation space of neural parametric head models (NPHMs) [23, 24]. NPHMs offer a flexible representation capable of handling complex and irregular facial expressions (e.g., blinking, skin creasing), along with a high-fidelity shape space including the head, hair, and ears, making them a much more suitable choice for face animation. We address the challenging task of creating an audio-conditional generative animation model for this volumetric representation.

We design the first transformer-based latent diffusion model for audio-driven head animation synthesis. Our diffusion model operates in the latent NPHM expression space to generate temporally coherent expressions consistent with an input audio signal represented with Wave2Vec 2.0 [6]. In the absence of paired audio-NPHM data, we optimize for corresponding NPHM expressions to fit to multi-view video recordings of people speaking, generating train supervision for our task. As NPHMs are designed for static (frame-by-frame) expressions without temporal consistency, we employ both geometric and temporal priors to produce temporally consistent optimized motion sequences. This enables training our audio-head diffusion model to synthesize realistic speech-conditioned 3D head sequences, capable of capturing high-frequency details like wrinkles and eye blinks present in the face region. Our method takes first step towards simplifying the task of high-fidelity facial motion generation of 3D faces for content creation applications.

In summary, we propose the first latent diffusion model for the creation of audio-conditioned animations of volumetric avatars. By producing volumetric head animation, our generative model is highly expressive yet efficient compared to existing 3D methods. We also demonstrate control over motion style, using classifier-free guidance to adjust the strength of the stylistic expression.

2. Related Work

Facial Animation. Our proposed method is the first work for audio-conditioned latent diffusion of volumetric head avatars. There is a large corpus of research works in the field of 2D audio-driven facial animation operating on RGB videos, synthesizing 2D sequences directly [10–13, 16, 27, 33, 34, 48, 56, 59, 61, 70, 71, 73, 79, 80]. However, these methods operate in pixel space and do not produce any geometric information. Another line of work also operating on RGB videos but using intermediate 3D representations are based on 3DMMs [19, 32, 57, 62, 66]. Although these methods generate RGB videos in the end, they use 3DMMs which produces very sparse geometric information.

Recent works based on radiance fields [22, 26, 37, 41, 55, 76] have also gained popularity due to their capability to model densities directly from images. These methods generate impressive RGB videos but the underlying geometry learned is highly imperfect (no segregation between background and facial region) as shown by Chan et al. [9]. Additionally, these methods require identity-specific training, and thus can not be used for content creation applications.

Learning animation of 3D meshes directly is much more promising, but only a handful of methods exist [14, 15, 21, 35, 44, 51, 65, 75]. A vast majority of these works model speech-conditioned animation for an artist-designed template mesh. Although these methods can match the facial motion with the speech signal, one limitation of these methods is their incapability to represent fine-scale details present in faces. Another downside of these methods is that these approaches learn a deterministic model producing no/muted motion in the upper region of the face, thus limiting them from being able to produce realistic motion. In this work, we solve these issues by proposing a generative model that can operate in the compact and detailed latent space of neural parametric head models, thus capable of representing fine-scale facial details and synthesizing a diverse set of expressions and speaking styles.

Diffusion Models for Generative Synthesis. In recent years, diffusion models have experienced a surge of interest as highly expressive and efficient generative models. These models have demonstrated strong performances as a generative model for a variety of domains such as images [17, 29, 49, 52, 53, 78], videos [8, 18, 20, 69], speech [30, 38, 39] and motion [2, 5, 36, 64, 67, 81].

2D facial animation has also seen some progress with diffusion models [56, 59]. DiffusedHeads [59] operates in pixel space, making the sampling process very slow. Although DiffTalk [56] learns the diffusion process in the latent space, the sampling is performed in an autoregressive fashion, limiting sampling efficiency. Moreover, these methods operate on RGB videos, and though they achieve high-quality results on a per-frame basis, consistency and coherence across different timesteps remain challenging due to temporal jitter. Hence, these methods can not be applied directly to the parametric head models. Concurrent to our work, DiffPoseTalk [60] and FaceDiffuser [58] propose a diffusion-based approach for animating 3D meshes from speech signal. However, these methods require additional conditions like style embedding, generate sequences in an autoregressive fashion coupled with diffusion denoising making them very slow and still they can not produce fine-scale facial details due to the use of 3DMMs which are limited in expressivity. Concurrent to ours, DPHM [63] also utilizes diffusion to regularize NPHM expressions from depth, operating on a frame-by-frame basis. In contrast, we synthesize the entire sequence simultaneously making it much faster and operate directly in the latent space of highly expressive NPHMs, producing high-fidelity and temporally consistent results.
Figure 2. Pipeline Overview. FaceTalk uses frozen Wave2Vec 2.0 [6] to extract audio embeddings from a speech signal. The diffusion timestamp is embedded using a timestamp embedder. The expression decoder employs a multi-head transformer decoder [68] with FiLM [45] layers, interleaved between Self-Attention, Cross Attention, and FeedForward layers, to incorporate diffusion timestamp. During training, the model is trained to denoise the noisy expression sequences from timestamp \( t \). At inference, FaceTalk denoises the gaussian noise sequence \( \{ \theta_{exp} \}_{t=1:T} \sim \mathcal{N}(0, I) \) iteratively until \( t = 0 \), yielding the estimated final sequence \( \{ \theta_{exp} \}_{t=1:T} \). These are then input to the frozen NPHM model, utilizing facial smoothing, and mesh sequences are extracted using MC [42].

3. Preliminaries

NPHM. In contrast to traditional 3D morphable models, which remain limited in expressive detail, we employ the Neural Parametric Head Model (NPHM) representation [23, 24]. Similar to traditional 3DMMs, NPHM disentangles a human face into an identity and expression space; however, the shape and expression space are volumetric, enabling expressive capture of details such as fine-scale eye movements, hair, ears, etc. NPHM uses two auto-decoder style neural networks (a) Identity Network \( \{ I \} \) to represent the overall facial shape and (b) Expression Network \( \{ E \} \) to represent the facial movements such as jaw pose, wrinkles, eyelinks, etc., jointly denoted as \( F = \{ I, E \} \). The identity and expressions latent codes \( \{ \theta_{id}, \theta_{exp} \} \) define the facial shape and expression respectively. Unlike 3DMMs which predict a fixed-topology mesh that can not model fine surface details, NPHM can handle different hairstyles, wrinkles and complex facial expressions. NPHM represents identities with a signed distance field (SDF) in canonical space and models the expressions as deformations onto the facial region. Mathematically, it can be defined as:

\[
F(x_i, \theta_{id}, \theta_{exp}) \rightarrow s_i : \mathbb{R}^3 \times \mathbb{R}^{1344} \times \mathbb{R}^{200} \rightarrow \mathbb{R},
\]

where \( x_i \in \mathbb{R}^3 \) represents the query points, \( \theta_{id} \in \mathbb{R}^{1344} \) represents identity latent code, \( \theta_{exp} \in \mathbb{R}^{200} \) represents expression latent code and \( s_i \) denotes the predicted SDF. The mesh is then extracted as the zero-level iso-surface decision boundary using Marching Cubes [42].

Diffusion. Diffusion models are a class of generative models that consist of a forward and reverse process. The forward process converts the original structured data distribution into Gaussian noise, modeled following a fixed Markov chain as:

\[
q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}),
\]

where \( q(x_t|x_{t-1}) \) denotes the forward process adding white noise to the original data distribution \( x_0 \). Mathematically, the forward process can be written as:

\[
q(x_t|x_0) \sim \mathcal{N}(\sqrt{\tau_t}x_0, (1 - \tau_t)I),
\]

where \( \tau_t \in (0, 1) \) are constants following a decreasing cosine schedule such that when \( \tau_t \) tends to 0, we can approximate \( x_T \sim \mathcal{N}(0, I) \). The reverse diffusion process progressively denoises samples from \( x_T \sim \mathcal{N}(0, I) \) into samples from a learned distribution \( x_0 \). In our experiments, we learn a generative model \( G_\theta \) to reverse the forward diffusion process by learning to estimate \( G_\theta(x_t, t, c) = \hat{x} \approx x \), where \( \hat{x} \) refers to predicting the cleaned samples directly, \( x_t \) refers to noisy input at diffusion timestamp \( t \). Given a conditional signal \( c \), we optimize model parameters \( \theta \) for all diffusion timestamps \( t \) using the following objective:

\[
L_\theta = E_{x,t}[||x - G_\theta(x_t, t, c)||_2^2].
\]

4. Method

FaceTalk performs high-fidelity and temporally consistent generative synthesis of motion sequences of heads, condii-
toned on audio signal. In order to characterize complex face motions and fine-scale movements, we synthesize realistic heads in the latent expression space of a parametric model for volumetric head representations, i.e., neural parametric head models (NPHMs) [23]. We, thus, develop a speech-conditioned latent diffusion model to synthesize temporally coherent head expression sequences that are coupled with the NPHM shape space to produce complex, realistic head animations of different identities. An overview of our approach is illustrated in Fig. 2.

**Audio Encoding.** We employ a state-of-the-art pre-trained speech model Wave2Vec 2.0 [6] to encode the audio signal. Specifically, we first use the audio feature extractor made up of temporal convolution layers (TCN) to extract audio feature vectors \( \{a_i\}_{i=1}^{N} \) from the raw waveform. This is followed by the Frequency Interpolation layer that aligns the input audio signal \( \{a_i\}_{i=1}^{N} \) (captured at frequency \( f_a = 16\text{kHz} \)) with our dataset \( \{a_i\}_{i=1}^{N} \) (captures at framerate \( f_c = 24\text{Hz} \)). Finally, a stacked multi-layer transformer encoder network processes these resampled features and outputs the processed and aligned audio feature vectors. The audio encoder is initialized with the pre-trained wav2vec 2.0 weights, followed by a feedforward layer to project the aligned audio features into the latent space of our expression decoder model. These aligned audio features are denoted as:

\[
A^{1:N} = \{a_i\}_{i=1}^{N}.
\]

These audio features \( A^{1:N} \) are then fed to the cross-attention layers of the expression decoder via the expression-audio alignment mask \( M \) to learn speech-conditioned expression features during training.

**Expression Encoding.** We train our expression decoder network on optimized NPHM expression sequences \( \{\theta_{exp}\}^{1:N} \) (obtained in Section 5), where \( N \) refers to number of frames in a sequence. We train a diffusion-based stacked multi-layer transformer [68] decoder network to synthesize facial expressions in the latent space of the NPHM model. During training, following forward diffusion (Eq. 3) we add noise for a randomly sampled diffusion timestamp \( t \sim \text{Uniform}(0, T) \) to create noisy expression codes \( \{\theta_{exp}\}^{1:N}_t \). These noisy expression codes are then projected to the latent space of our model via a linear layer, followed by a stack of transformer decoder blocks, and then projected back to the original NPHM space using another linear layer. To embed diffusion timestamp in the latent space of our model, we apply sinusoidal embedding and process it through a three-layer MLP. Next, to fuse diffusion timestamp into the model, we use a one-layer FiLM (feature-wise linear modulation) network [45] between the multi-head self-attention, multi-head cross-attention and feedforward layers of the transformer decoder block, which is critical for the model to produce high-quality expressions as we show in the results. We leverage a look-ahead binary target mask \( T \in \mathbb{R}^{N \times N} \) in the multi-head self-attention layers to prevent the model from peeking into the future expression codes. Mathematically, it can be written as:

\[
T_{ij} = \begin{cases} 
  \text{True} & \text{if } i \leq j \\
  \text{False} & \text{else}
\end{cases}
\]

where \( T_{ij} \) refers to the \((i, j)^{th}\) element of the matrix and \( 1 \leq i, j \leq N \). The audio features \( A^{1:N} \) are fused into the network via the multi-head cross-attention layers. We leverage the expression-audio alignment mask \( M \) to fuse the audio features into the network to ensure that mouth poses learned by the expression codes are consistent with the speech signal. The binary mask \( M \in \mathbb{R}^{N \times N} \) is Kronecker delta function \( \delta_{ij} \) such that the audio features for \( t^{th} \) timestamp attend to expression features at the \( j^{th} \) timestamp if and only if \( i = j \). Mathematically,

\[
M = \delta_{ij} = \begin{cases} 
  \text{True} & \text{if } i = j \\
  \text{False} & \text{if } i \neq j
\end{cases}
\]

We demonstrate in the results (Section 6) that this alignment is crucial for learning audio-consistent expression codes. As a by-product, this further enables the model to generalize to arbitrary long audio signals during inference which we also show in our results. Our model is trained with the diffusion loss in Eq. 4, with input \( x = \{\theta_{exp}\}^{1:N} \), and conditioning \( c = A^{1:N} \).

**Expression Augmentation.** Unlike other domains (like text-to-motion) where the synthesized motion can vary dramatically, speech-driven 3D facial animation is more constrained, requiring precise mouth alignment with the audio signal, making it prone to overfitting. To alleviate this, we propose a novel augmentation strategy capable of producing diverse expressions for a given audio signal. Our key insight is to augment the dataset to generate different expression codes for the same speech signal by randomly amplify and suppressing its magnitude. Specifically, we randomly sample modulation factor \( r \) within the bounds \([a, b] \) as:

\[
r \sim \text{Uniform}(a, b),
\]

and scale the expression codes as \( \{r \times \theta_{exp}\}^{1:N} \). We show in the results this augmentation strategy helps the model synthesize diverse expressions.

** Sampling.** FaceTalk uses a diffusion-based framework to learn to synthesize NPHM expression sequences of \( N \) frames \( \{\theta_{exp}\}^{1:N} \in \mathbb{R}^{N \times 200} \). At each of the denoising timestep \( t \), FaceTalk predicts the denoised sample and noises it back to timestamp \( t - 1 \):

\[
\{\theta_{exp}\}^{1:N}_{t-1} = G_\theta(\{\theta_{exp}\}^{1:N}_t, A^{1:N}, t-1),
\]
terminating when it reaches \( t = 0 \). We train our model using classifier-free guidance by randomly replacing the audio with null conditioning \( A^{1:N} = \Phi \) during training with 25% probability. During inference, we can use the weighted combination of conditional and unconditionally generated samples:

\[
\{\theta^e_{exp}\}_i^{1:N} = w_i \{\theta^u_{exp}\}_i^{1:N} + (1 - w_i) \{\theta^id_{exp}\}_i^{1:N},
\]

where \( \theta^e_{exp} \) and \( \theta^id_{exp} \) refers to conditionally and unconditionally generated samples. To amplify the audio conditioning, we can use guidance strength \( w > 1 \).

**Sequence Generation.** The expression codes predicted above \( \{\hat{\theta}^e_{exp}\}_i^{1:N} \) are then passed to the pretrained NPHM expression mapper \( \{\mathcal{E}\} \) to obtain expression deformations \( \{\delta^e_{exp}\}_i^{1:N} \). We then apply smoothing based on a Gaussian kernel to these deformations to remove the unwanted wiggle in the head and neck regions and to ensure that generating sequences are free from flickering artifacts. Specifically, we first define a control center \( \mathcal{C} \) on the mouth region in the canonical space as:

\[
\mathcal{C} = [c_x, c_y, c_z].
\]

Next, using a 3D gaussian kernel with standard deviation \( \Sigma = [\sigma_x, \sigma_y, \sigma_z] \), centered at \( \mathcal{C} \), for each point sampled from 3D grid \( \mathcal{P} \in \mathbb{R}^{|X| \times |Y| \times |Z|} \) as \( \mathcal{P}_{xyz} = [p_x, p_y, p_z] \), we compute its distance from \( \mathcal{C} \) as:

\[
d_{xyz} = \sqrt{\frac{(p_x - c_x)^2}{\sigma^2_x} + \frac{(p_y - c_y)^2}{\sigma^2_y} + \frac{(p_z - c_z)^2}{\sigma^2_z}}.
\]

Based on this distance, smoothing weights are obtained from the Gaussian kernel with min-max normalization as:

\[
w_{xyz} = \frac{1}{2\pi\sigma_x\sigma_y\sigma_z} e^{-\left(\frac{1}{2}d^2_{xyz}\right)}
\]

\[
w^k_{xyz} = \frac{w^k_{xyz} - \min(w^1_{xyz})}{\max(w^1_{xyz}) - \min(w^1_{xyz})},
\]

where \( M = |X| \times |Y| \times |Z| \). The obtained expression deformations \( \{\delta^e_{exp}\}_i^{1:N} \) are then multiplied with these smoothing weights \( \{w\}_i^{1:N} = \{(w^1, w^2, \ldots, w^M)\}_i^{1:N} \) to obtain smooth expression deformations as:

\[
\{\delta^e_{exp}\}_i^{1:N} = \{\delta^e_{exp}\}_i^{1:N} \times \{w\}_i^{1:N}.
\]

These smoothed expression deformations \( \{\delta^e_{exp}\}_i^{1:N} \), along with identity \( \theta^id \) and predicted expression codes \( \{\hat{\theta}^e_{exp}\}_i^{1:N} \) are passed to the NPHM identity MLP \( \{\mathcal{I}\} \) to obtain smoothed SDF, from which meshes are extracted using MC [42].

**5. Dataset Creation**

To train our latent diffusion model, we require temporally consistent NPHM expression codes as well as their corresponding audio. To this end, we leverage multi-view recordings of NeRSemble [37] captured with 16 cameras of talking human faces paired with corresponding audio signals. Notably, these sequences are captured for faces already present in the identity space of the NPHM model. Thus, for a given subject, only the expression codes corresponding to the multi-view sequences have to be estimated. Specifically, we first extract 3D pointclouds \( \{\mathcal{P}_i\}_i^{N} \) from individual frames using COLMAP [54], where \( N \) refers to the number of frames in a sequence, with each \( \mathcal{P}_i \in \mathbb{R}^{K \times 3} \) consisting of \( K \) points. Given the pretrained NPHM model, for the known facial shape \( \theta^id \), we then optimize the expression codes \( \{\theta^e_{exp}\}_i^{1:N} \) to match the pointclouds. In particular, at every iteration, we randomly sample a subset \( S_i \) of 5000 points from \( \mathcal{P}_i \) as \( S_i \sim \text{Uniform}(5000, \mathcal{P}_i) \), and then feed the sampled points \( S_i \) identity code \( \theta^id \) and learnable expression code \( \theta^e_{exp} \) into the expression decoder \( \{\mathcal{E}\} \) to obtain the expression deformation \( \delta^e_i \) for the sampled points \( S_i \). The deformation \( \delta^e_i \) is then added to \( S_i \) to obtain deformed points \( D_i = S_i + \delta^e_i \). These deformed points \( D_i \) are finally fed to the identity decoder \( \{\mathcal{I}\} \) along with latent codes \( \theta^id \) and \( \theta^e_{exp} \) to obtain the SDF. The expression
codes are then optimized using the loss $L_{total}$ which is illustrated in Fig. 3. Naively optimizing the expression codes on a per-frame basis shows flickering artifacts. To mitigate this, we optimize expressions in groups of $n = 10$ frames in a sliding window fashion (with an overlap of 2 frames between adjacent windows) using an additional temporal regularization $L_{temp}$. To further prevent the expression codes from deviating too much from the distribution already learned by the NPHM model, we employ additional L2 expression regularization $L_{exp}$. We minimize the objective $L_{total}$ defined as:

$$L_{total} = \lambda_{sdf} L_{sdf} + \lambda_{temp} L_{temp} + \lambda_{reg} L_{reg}. \quad (16)$$

The SDF loss $L_{sdf}$, temporal regularization $L_{temp}$, and expression regularization $L_{exp}$ can be defined as:

$$L_{sdf} = \sum_{i=1}^{n} \frac{1}{|S|} \sum_{k=1}^{|S|} \|\text{SDF}(S^k_i)\|_1 \quad (17)$$

$$L_{temp} = \sum_{i=1}^{n} \|\theta_{exp}^{i+1} - \theta_{exp}^i\|_2, \quad L_{exp} = \sum_{i=1}^{n} \|\theta_{exp}^i\|_{\epsilon}, \quad (18)$$

where $||\cdot||_1, ||\cdot||_2$ and $||\cdot||_{\epsilon}$ refers to L1, L2 and Huber loss respectively, and $n$ refers to number of frames simultaneously optimized. More details regarding aligning the coordinate system of NeRSemble dataset [37] with NPHM can be found in the supplemental document. Once optimized, next to generate mesh sequences, we uniformly sample points from a 3D grid $P \in \mathbb{R}^{|X| \times |Y| \times |Z|}$ such that $P_{xyz} = [p_x, p_y, p_z]$ is a point in 3d space. These points, along with the identity $\theta_{id}$ and expression codes $\{\theta_{exp}^i\}_{i=1}^N$ are passed to the pretrained NPHM model outputting an SDF; meshes are then extracted with [42]. In total, our FaceTalk dataset consists of 1000 sequences, an order of magnitude larger than the existing datasets [14].

6. Results

We evaluate FaceTalk on the task audio-driven motion synthesis and compare it against state-of-the-art methods. 

**Evaluation Metrics.** Unlike template-based methods that rely on LVE (Lip Vertex Error), our method produces SDFs for each frame; extracted mesh topologies thus vary across frames, making LVE inapplicable for evaluation. Thus, we employ LSE-D (Lip Sync Error Distance) [48] for quantitative evaluation of lip synchronization. For diversity, we compute FID and KID, as well as diversity scores [50]. As FID/KID may not reflect all quality considerations, when samples are limited, we include two established quality metrics: (1) FIQA (Face Image Quality Assessment) [43] (2) VQA (Video Quality Assessment) [74]. FIQA quantifies image quality for face recognition and similarity to in-the-wild real faces and VQA measures overall video qual-

![Figure 4. Qualitative comparison for audio-driven face animation. Our approach maintains high fidelity while demonstrating rich mouth and nasolabial movements. In particular, we demonstrate more accurate lip articulation, precisely synchronized to phonetic movements.](image-url)
port is strong wine
we can easily be applied to diverse identities with complex geometry.

Given a speech signal, our method can further generate diversity in expression for the same identity. Note the difference in speaking style (intensity of mouth opening) as well as eyeblinks/frowning in the upper face area.

Implementation Details. For dataset creation, we optimize NPHM expressions for 500 iterations in a group of $n = 10$ frames with a step size of 0.001 for iterations $\leq 300$, and 0.0001 otherwise. We use $\lambda_{sd} = 10$, $\lambda_{temp} = 0.1$ and $\lambda_{reg} = 0.0025$ for Eq. 16 during optimization. To train our diffusion model, we randomly clip sequences to 2 seconds (48 frames) for efficient minibatching. We train with Adam optimizer with a learning rate of 0.0001. For diffusion, we use 1000 noising/denoising timestamps, a cosine schedule to add noise to input sequences. We apply data augmentation to the expression codes by uniformly modulating them.

Baseline Comparisons. This is the first work to perform speech-conditioned synthesis for volumetric head motion sequences. We compare with state-of-the-art template-based methods in Fig 4 and Tab. 1. Specifically, we compare against speech-driven animation methods [14, 21, 51, 75], personalized methods [65], as well as recent emotion-based methods [15, 44]. Our approach more accurately captures subtle mouth movements compared to less expressive baselines, resulting in better FID/KID. These scores were evaluated only on mouth region crops to avoid bias towards rest of facial geometry. Our method consistently achieves better lip-audio synchronization while also representing fine-scale facial details like creasing in the nasolabial folds with wider mouth motions. This is confirmed by our perceptual user study in Fig 6.

Generative Synthesis. Our generated expression codes are identity-agnostic, and are easily transferred to different identities, as shown in Fig. 5. Additionally, we can synthesize diverse expressions per-identity for a given audio.

<table>
<thead>
<tr>
<th>Method</th>
<th>LSE-D ↓</th>
<th>FID ↓</th>
<th>KID ↓</th>
<th>FIQA ↑</th>
<th>VQA ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOCA [14]</td>
<td>13.6191</td>
<td>239.043</td>
<td>0.280</td>
<td>33.36</td>
<td>0.4251</td>
</tr>
<tr>
<td>MeshTalk [51]</td>
<td>12.9607</td>
<td>220.172</td>
<td>0.254</td>
<td>36.14</td>
<td>0.4855</td>
</tr>
<tr>
<td>FaceFormer [21]</td>
<td>11.9848</td>
<td>215.274</td>
<td>0.222</td>
<td>38.24</td>
<td>0.5227</td>
</tr>
<tr>
<td>CodeTalker [75]</td>
<td>11.8054</td>
<td>208.064</td>
<td>0.207</td>
<td>38.38</td>
<td>0.5274</td>
</tr>
<tr>
<td>Imitator [65]</td>
<td>11.6119</td>
<td>208.479</td>
<td>0.212</td>
<td>38.44</td>
<td>0.5384</td>
</tr>
<tr>
<td>EMOTE [15]</td>
<td>11.7192</td>
<td>201.311</td>
<td>0.201</td>
<td>31.80</td>
<td>0.5037</td>
</tr>
<tr>
<td>Ours</td>
<td>11.2737</td>
<td>40.692</td>
<td>0.009</td>
<td>45.75</td>
<td>0.6145</td>
</tr>
</tbody>
</table>

Table 1. In comparison to state of the art, FaceTalk more accurately matches the audio, while producing high perceptual fidelity.
Figure 6. User study comparison with baselines. We measure preference for (1) Overall Animation Quality, (2) Lip Synchronization and (3) Facial Realism. FaceTalk results are overwhelmingly preferred over the best baseline methods on all these aspects.

```
<table>
<thead>
<tr>
<th>Our Method</th>
<th>LSE-D ↓</th>
<th>Diversity ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o expr. aug.</td>
<td>11.6229</td>
<td>1.61e-8</td>
</tr>
<tr>
<td>w/o facial smoothing</td>
<td>11.3488</td>
<td>0.34</td>
</tr>
<tr>
<td>w/o FiLM layer</td>
<td>12.029</td>
<td>0.006</td>
</tr>
<tr>
<td>w/o expr.-audio align.</td>
<td>13.720</td>
<td>5.17e-8</td>
</tr>
<tr>
<td>w/o diffusion</td>
<td>11.3217</td>
<td>0.000</td>
</tr>
<tr>
<td>Full (Ours)</td>
<td>11.2737</td>
<td>0.34</td>
</tr>
</tbody>
</table>
```

Table 2. Ablation over model design. Expression augmentation improves diversity, while facial smoothing alleviates inter-frame jitter. Using FiLM conditioning achieves accurate mouth pose and increased diversity. Without expression-audio alignment, the model ignores the audio signal. Without diffusion, the model fails to generate diverse results. Our full model with all components achieves the best results.

**Architecture Ablations.** We ablate our model design choices in Tab 2, with visual results shown in Fig 7, and refer to the supplemental for video demonstration. Our proposed expression augmentation helps the model to synthesize diverse expressions. Facial smoothing removes unwanted wobbliness in the head and neck region, thereby reducing the inter-frame jitter. FiLM conditioning ensures that expressions generated by our method are more pronounced and that expressions accurately match the audio signal. The expression-audio alignment mask is necessary to match the synthesized expression codes with the audio signal. Our diffusion training ensures that expressions generated by our method are high-quality and diverse.

**Limitations.** While FaceTalk effectively synthesizes high-fidelity and audio-synchronized facial expressions, it still has limitations. For instance, the use of a diffusion model requires multiple denoising steps during inference, limiting its real-time application. We believe that this could be improved by investigating efficient sampling techniques [72]. Currently, our method specializes in synthesizing only the expression codes. For holistic 3D facial animation, we need to extend its capability to synthesize facial identities. In the future, we would like to generate diverse identities aligned with the gender inferred directly from the audio.

**7. Conclusion**

In this work, we have introduced FaceTalk, a new approach to synthesize animations of realistic volumetric human heads from audio. We introduced the first latent diffusion model for audio-driven head animation, producing significantly higher fidelity results than existing methods. FaceTalk leverages a parametric head model producing volumetric head representations to generate expressive, fine-grained motions such as eye blinks, skin creasing, etc. We also demonstrate the applicability of our method for other conditioning signals such as facial landmarks. We believe this is an important first step towards enabling the animation of highly detailed 3D face models, which can enable many new possibilities for content creation and digital avatars.

**8. Acknowledgments**

This work was supported by the ERC Starting Grant Scan2CAD (804724), the Bavarian State Ministry of Science and the Arts and coordinated by the Bavarian Research Institute for Digital Transformation (bidt), the German Research Foundation (DFG) Grant “Making Machine Learning on Static and Dynamic 3D Data Practical,” the German Research Foundation (DFG) Research Unit “Learning and Simulation in Visual Computing,” and Sony Semiconductor Solutions Corporation. We would like to thank Simon Giebenhain and Tobias Kirschstein for their help with dataset.
References


[38] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis, 2021.
[56] Shuai Shen, Wenliang Zhao, Zibin Meng, Wanhu Li, Zheng Zhu, Jie Zhou, and Jiwen Lu. DiffTalk: Crafting diffusion
models for generalized audio-driven portraits animation. In CVPR, 2023. 2
[57] Linsen Song, Wayne Wu, Chen Qian, Ran He, and Chen Change Loy. Everybody’s talkin’: Let me talk as you want, 2020. 2
[61] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing lip sync from audio. ACM Trans. Graph., 2017. 2
[70] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. End-to-end speech-driven facial animation with temporal gans, 2018. 2
[71] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans, 2019. 2
[78] Lvmi Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 2
[80] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: Speaker-aware talking-head animation. ACM Trans. Graph., 39(6), 2020. 2