# MoDiTalker: Motion-Disentangled Diffusion Model for High-Fidelity Talking Head Generation

**Seyeon Kim** [1, 2*], **Siyoon Jin** [1*], **Jihye Park** [1, 2*],
**Kihong Kim** [3], **Jiyoung Kim** [1], **Jisu Nam** [4], **Seungryong Kim** [4†]
[1]Korea University
[2]Samsung Electronics
[3]VIVE STUDIOS
[4]KAIST
{sey.kim, jye12.park}@samsung.com,
{siyun515, kplove01, jisunam, seungryong.kim }@kaist.ac.kr, hxngiee@gmail.com

## Abstract

Conventional GAN-based models for talking head generation often suffer from limited quality and unstable training. Recent approaches based on diffusion models have attempted to address these limitations and improve fidelity. However, they still face challenges, such as intensive sampling times and difficulties in maintaining temporal consistency due to the high stochasticity of diffusion models. To overcome these challenges, we propose a novel motion-disentangled diffusion model for high-quality talking head generation, called **MoDiTalker**. We introduce two modules: the **A**udio-**To**-**M**otion (AToM) module, designed to generate synchronized lip movements from audio, and the **M**otion-**To**-**V**ideo (MToV) module, designed to produce high-quality talking head videos based on the generated motions. AToM excels in capturing subtle lip movements by leveraging an audio attention mechanism. Additionally, MToV enhances temporal consistency by utilizing an efficient tri-plane representation. Our experiments on standard benchmarks demonstrate that our model outperforms existing GAN-based and diffusion-based models. We also provide comprehensive ablation studies and user study results.

**Code** — https://github.com/cvlab-kaist/MoDiTalker

## 1 Introduction

Audio-driven talking head generation (Zhou et al. 2021; Zhang et al. 2023; Wang et al. 2021a; Ji et al. 2021) aims to generate high-fidelity head videos with lip movements synchronized to given audio input. This task has been widely studied for many practical applications, including film production (Prajwal et al. 2020), video conferencing (Wang, Mallya, and Liu 2021) and digital avatars (He et al. 2023).

To solve this task, traditional GAN-based methods transform the audio embeddings into intermediate representations, such as dense motion fields (Yin et al. 2022; Zhang et al. 2023) or 2D/3D facial landmarks (Zhou et al. 2020; Wang et al. 2021a,b). Notably, recent works (Wang et al. 2021a,b; Zhang et al. 2023) have shown remarkable results by utilizing 3D facial models (Blanz and Vetter 2023), which effectively capture realistic motion. However, conventional approaches still inherit limitations of GANs, such as training instability (Brock, Donahue, and Simonyan 2018; Miyato et al. 2018) or mode collapse (Thanh-Tung and Tran 2020).

Meanwhile, diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020) have demonstrated superior performance in generation tasks, offering more stable training and enhanced fidelity compared to GANs. Inspired by these advancements, recent studies (Shen et al. 2023; Stypułkowski et al. 2023; Ma et al. 2023) propose diffusion-based frameworks for talking head generation. However, without explicit motion guidance, audio input alone is insufficient to capture the complex visual dynamics. Moreover, these models often rely on frame-by-frame generation, leading to suboptimal temporal consistency and increased inference time compared to GAN-based approaches (Prajwal et al. 2020; Wang, Mallya, and Liu 2021).

In this paper, we introduce a novel diffusion-based framework for talking head generation, dubbed **MoDiTalker**. We separate the generation process into two stages. First, we present **A**udio-**To**-**M**otion (**AToM**), a transformer-based diffusion model that generates motion sequences conditioned on audio input. Specifically, AToM predicts motion differences from an initial landmark for each frame. Our model design, which separately processes upper and lower facial landmarks, significantly contributes to achieving superior lip synchronization performance, both qualitatively and quantitatively. Second, we propose **M**otion-**To**-**V**ideo (**MToV**), a diffusion model that generates videos following facial motion sequences from AToM. Specifically, we leverage tri-plane representations to efficiently encode frame sequences. MoDiTalker achieves high-fidelity talking head generation with enhanced temporal consistency while significantly reducing inference time complexity compared to existing diffusion-based methods (Shen et al. 2023; Stypułkowski et al. 2023; Wei, Yang, and Wang 2024).

In experiments, our framework achieves state-of-the-art performance on HDTF dataset (Zhang et al. 2021), surpassing GAN-based (Prajwal et al. 2020; Zhou et al. 2021) and diffusion-based (Ma et al. 2023; Wei, Yang, and Wang 2024) approaches.

*Equal contribution
†Corresponding author

## 2 Related Work

### Audio-driven Talking Head Synthesis

Early audio-driven talking head generation methods (Prajwal et al. 2020; Liang et al. 2022; Zhou et al. 2021) relied on GANs to learn audio-to-lip translation. For instance, Wav2Lip (Prajwal et al. 2020) leverages a pre-trained SyncNet (Chung and Zisserman 2017) as a lip sync expert to enhance lip synchronization quality. However, they often produced blurry, unnatural results with limited facial expression and head pose control. Specifically, these methods struggled to capture the complex visual dynamics from audio input alone. To overcome this, subsequent approaches introduced intermediate 2D (Song et al. 2021; Lu, Chai, and Cao 2021; Chen et al. 2019; Zhou et al. 2020) or 3D (Zhang et al. 2021; Ren et al. 2021; Zhang et al. 2023) representations which utilize 3D Morphable Model (Booth et al. 2016) parameters, including expression, pose, and identity to better represent facial movements. While these methods can produce smoother and natural movements, they also unavoidably inherit the drawbacks of GANs, such as mode collapse or unstable training.

To circumvent the challenges inherent in GANs, recent studies (Stypułkowski et al. 2023; Shen et al. 2023; Wei, Yang, and Wang 2024; Tian et al. 2024; Xu et al. 2024) have explored diffusion-based frameworks for talking head video generation. Some works (Stypułkowski et al. 2023; Shen et al. 2023) inject audio features into model layers to control lip motions, leveraging the strengths of diffusion models to generate high-quality images. However, these methods often struggle to produce realistic lip movements due to the direct injection of audio embeddings without explicit motion guidance. To address this, AniPortrait (Wei, Yang, and Wang 2024) incorporates 2D motion conditions from audio through Pose Guider (Hu 2024), while EMO (Tian et al. 2024) implicitly integrates motion information. However, these methods still suffer from slow sampling times and temporal inconsistencies due to their frame-by-frame generation process. Meanwhile, some works (Peng et al. 2024; Cho et al. 2024) have employed tri-plane representation for 3D talking heads, enabling greater flexibility in head movement and viewing angles. However, these methods often require training a separate model for each individual, limiting their generalizability.

We introduce a novel diffusion-based method that leverages intermediate representations as motion guidance to improve lip movement accuracy. Our approach represents video data using a tri-plane structure, ensuring temporal smoothness and preserving identity in generated videos.

### Video Diffusion Model

Several studies (Blattmann et al. 2023; Wu et al. 2023; Khachatryan et al. 2023) propose to fine-tune the pre-trained text-to-image diffusion model for video generation. Although they exhibit promising results with minimal computational complexity, they often encounter challenges of temporal consistency at high frame rates. To alleviate this, recent methodologies (Yu et al. 2023; Hu, Chen, and Luo 2023; Wang et al. 2023) train the diffusion model from scratch using video datasets. While these works exhibit better temporal consistency and video quality, they struggle with memory inefficiency and controllability of video generation. To overcome these challenges, several methods focus on a low-dimensional latent space (Yu et al. 2023), and an autoencoder module (Hu, Chen, and Luo 2023) to minimize computational complexity, while utilizing flow maps as a motion condition (Wang et al. 2023) for conditional video generation. In this paper, we propose an efficient conditional video diffusion model, tailored for talking head generation.

## 3 Methodology

### Preliminary

Diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020) approximate the data distribution by reconstructing the data sample from pure Gaussian noise through a gradual denoising process. Latent diffusion models (Rombach et al. 2022) perform this in the latent space. In the forward diffusion process, $z_0$ is gradually noisified into $z_t$ at time step $t \in \{1, \dots, T\}$. The forward diffusion process is formulated with pre-defined variance $\beta_t$ such that

$$z_t = \sqrt{\alpha_t} z_0 + \sqrt{1-\alpha_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

where $\alpha_t = \prod_{i=1}^{t}(1-\beta_i)$.

For the reverse process, the neural network $\mathcal{F}_\theta(z_t, t)$ is trained to reconstruct $z_0$ at any time step $t$. The objective function is defined by the Mean Squared Error as follows:

$$\mathcal{L} = \mathbb{E}_{z,t,\epsilon\sim\mathcal{N}(0,1)}\left[\|z_0 - \mathcal{F}_\theta(z_t, t)\|_2^2\right]. \tag{2}$$

During sampling time, the neural network $\mathcal{F}_\theta(z_t, t)$ predicts the denoised latent $\hat{z}_{0,t}$ and converts it to $\hat{z}_{t-1}$ by the reparameterization trick (Kingma and Welling 2013), following:

$$\hat{z}_{t-1} = \sqrt{\alpha_{t-1}}\mathcal{F}_\theta(z_t, t) + \frac{\sqrt{1-\alpha_{t-1}-\sigma_t^2}}{\sqrt{1-\alpha_t}}(z_t - \sqrt{\alpha_t}\mathcal{F}_\theta(z_t, t)) + \sigma_t\epsilon, \tag{3}$$

where $\sigma_t$ is the covariance of the Gaussian distribution.

By iteratively sampling starting from $z_T \sim \mathcal{N}(0, I)$ during time step $t \in \{T, \dots, 1\}$, we obtain the new data sample $\hat{z}_0$ from the desired distribution. $\hat{z}_0$ is then decoded to an RGB image $\hat{x}_0$ by the pre-trained decoder.

### Overview

Our objective is to generate a lip-synced video, $\hat{X} = \{\hat{x}^1, \dots, \hat{x}^N\}$, given audio, $A$, and a target identity video, $X = \{x^1, \dots, x^N\}$, where $N$ represents the number of frames. To achieve this, MoDiTalker comprises two distinct diffusion models: Audio-to-Motion (AToM) and Motion-to-Video (MToV). AToM generates facial motion sequences, represented by landmarks $L = \{l^1, \dots, l^N\}$, from input audio $A$, incorporating the speaker's facial characteristics from $x_{id}$. We designate the first frame of $X$ as $x_{id}$. Subsequently, MToV generates a realistic video, $\hat{X}$, synchronized with the motion sequences, $L$, produced by AToM. MToV employs the upper half of $X$ as pose frames $X_P = \{x_P^1, \dots, x_P^N\}$ and and prioritizes generating lip movements synchronized with audio. For providing identity condition, MToV utilizes stacked $x_{id}$ as identity frames $X_I$. The overall architecture of MoDiTalker is illustrated in Fig. 1.

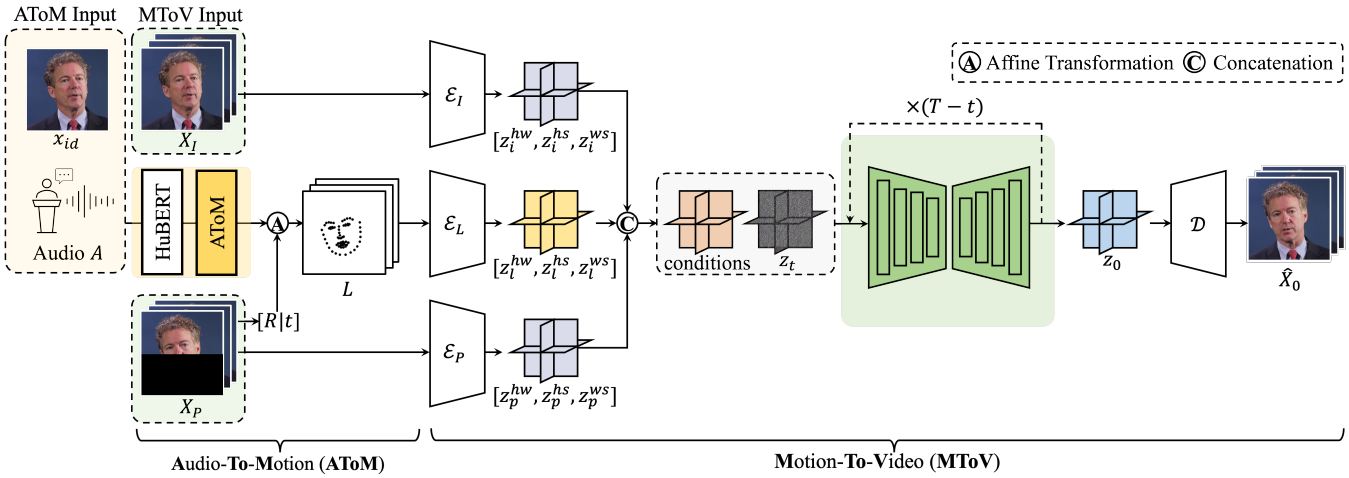



Figure 1: **Overall network architecture of MoDiTalker.** Our framework comprises two components: AToM generates lip-synchronized facial landmarks from an identity frame, $x_{id}$, and audio input, $A$. MToV produces high-fidelity talking head videos, $\hat{X}_0$, using synthesized facial landmarks $L$, from AToM, along with identity frames $X_I$, and pose frames $X_P$.

## Audio-to-Motion (AToM) Model

Given audio input $A$ and a single identity frame $x_{id}$, AToM aims to generate facial landmark sequences $L$, that accurately reflect both the audio content of $A$ and the facial characteristics of $x_{id}$. We extract facial landmarks $l_{id}$ from the identity frame $x_{id}$ to provide facial characteristics. For long video generation, the last generated motion frame becomes the new $l_{id}$ for subsequent sequences. As illustrated in Fig. 2(a), we introduce a transformer-based diffusion model specialized for lip-synced facial motion generation, building upon the (Tseng, Castellon, and Liu 2023).

**Architectural details.** We design AToM to learn the residuals $\Delta L = \{\Delta l^1, \ldots, \Delta l^N\}$ from $l_{\text{id}}$. This design helps to reduce unwanted subtle movements of keypoints, as shown in Tab. 2. Specifically, we extract the initial facial landmark $l_{\text{id}}$ from the identity frame $x_{\text{id}}$ using a 3D Facial Morphable Model (3DMM) (Blanz and Vetter 2023). It is then passed through a trainable landmark encoder to produce the landmark embedding $F_L$. Note that in the training phase, we use the frontalized facial landmarks to ensure the model focuses on lip movements rather than head pose.

We also extract the audio embedding $F_A = \{f^1, \ldots, f^N\}$ from the audio $A$, using HuBERT (Hsu et al. 2021) and pass through subsequent audio encoder. Both $F_L$ and $F_A$ are then integrated into the cross-attention modules of the diffusion model, by concatenating with timestep embeddings. Benefiting from $l_{\text{id}}$, which provides the speaker-specific facial structure, the model is enabled to focus on speaker-agnostic representations (e.g., lip motions) from $F_A$. Both landmark encoder and audio encoder are composed of simple transformer encoders sharing the same architecture.

Moreover, to improve lip synchronization quality, we design AToM to disentangle lip-related and lip-unrelated facial landmarks, as illustrated in Fig. 2(b). We divide the facial landmarks into upper-half (lip-unrelated) and lower-half (lip-related) segments, then design the network to process them separately and subsequently merge them. The audio embedding $F_A$ is exclusively injected into the cross-attention module of the lip-related block, while the landmark embedding $F_L$ is conditioned after the two blocks are merged. This architectural choice enables the model to focus solely on lip movements, preserving the other facial regions, and ultimately enhancing lip sync quality. Finally, under the conditions of $F_L$ and $F_A$, the Gaussian noise $\Delta L_T \sim \mathcal{N}(0, I)$ is iteratively denoised through AToM to generate the denoised residual landmark sequences $\Delta \hat{L}_0$. The final facial motion sequences $L$ are obtained by adding $\Delta \hat{L}_0$ to the initial landmark $l_{\text{id}}$.

**Training.** In training, the model is trained to reconstruct $\Delta L$, using the initial landmark embedding $F_L$ for speaker-specific facial structures and the audio embedding $F_A$ for lip movements. Therefore, we redefine Eq. 2 for the objective function of training AToM, $\mathcal{F}_{\text{AToM}}$, as follows:

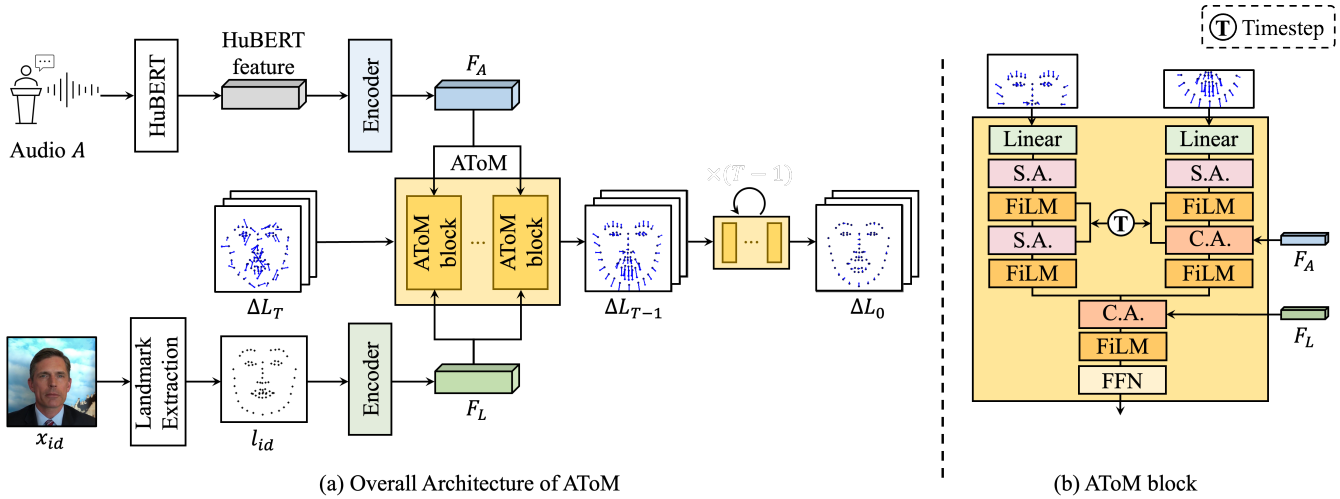
$$\mathcal{L}_{\text{AToM}} = \mathbb{E}_{\Delta L, t, \epsilon, F_L, F_A}\left[\|\Delta L_0 - \mathcal{F}_{\text{AToM}}(\Delta L_t, t; F_L, F_A)\|_2^2\right]. \quad (4)$$

Here, $\Delta L_0$ is defined as the difference between facial landmarks from ground-truth frames and those from an identity frame, and $\Delta L_t$ is the noised version of $\Delta L_0$ at time step $t$.

## Motion-to-Video (MToV) Model

MToV aims to generate realistic talking head videos $\hat{X}$, mirroring desired identities and audio inputs. To achieve this, we use three conditions: facial landmark sequences $L$, pose frames $X_P$, and identity frames $X_I$.

Previous diffusion-based methods (Stypułkowski et al. 2024; Shen et al. 2023) generate videos in a frame-by-frame manner, resulting in sub-optimal temporal consistency and identity preservation. To tackle this, as illustrated in Fig. 1,




(a) Overall Architecture of AToM (b) AToM block

Figure 2: **Overview of the Audio-to-Motion (AToM):** (a) AToM is a transformer-based diffusion model that learns the residual between the initial landmark, $l_{\text{id}}$, and the landmark sequence, conditioned on the audio embedding $F_A$ and the initial landmark embedding $F_L$. In addition, (b) we design AToM block to process lip-unrelated (upper-half) and lip-related (lower-half) landmarks, allowing the model to focus more on generating lip-related movements while preserving the facial shape of the speaker. S.A. and C.A. represent Self Attention and Cross Attention, respectively. FiLM stands for Feature-wise Linear Modulation.

we introduce a video diffusion model that generates multiple frames simultaneously based on (Yu et al. 2023) with given conditions. However, conditioning the video diffusion model traditionally requires 4-dimensional representations (Ho et al. 2022), which pose significant computational challenges during training and inference. To overcome this limitation, we adopt tri-plane representations (Yu et al. 2023; Ho et al. 2022) to effectively employ a video diffusion model. Visualizations of each feature are shown in Fig. 3. The frontal planes primarily capture spatial information, while the side planes capture motion information across the temporal axis.

**Architectural details.** AToM synchronizes landmark sequences $L$ with audio and identity. We leverage head poses from the pose frames $X_P$ to achieve facial landmarks alignment. We achieve this by applying an affine transformation (Wang, Mallya, and Liu 2021) to the frontalized facial landmarks from AToM, aligning them with the desired head pose. Detailed process is explained in the Appendix 1.2.

These sequences are then encoded into tri-plane representations $Z_L = \{z_l^{hw}, z_l^{hs}, z_l^{ws}\}$ by the landmark encoder $\mathcal{E}_L$, where $z_l^{hw} \in \mathbb{R}^{c\times h\times w}$, $z_l^{hs} \in \mathbb{R}^{c\times h\times s}$, and $z_l^{ws} \in \mathbb{R}^{c\times w\times s}$. Here, $c$, $h$ and $w$ denote the embedded channel, height, and width, respectively. Specifically, $z_l^{hw}$ provides the model with speaker-specific facial structures, while $z_l^{hs}$ and $z_l^{ws}$ encode the temporal relationships between frames.

Simultaneously, following previous works (Prajwal et al. 2020; Shen et al. 2023), we condition the model on pose frames $X_P \in \mathbb{R}^{S\times C\times H\times W}$, corresponding to the upper part of the desired video $X$. $X_P$ is then encoded into tri-plane representations $Z_P = \{z_p^{hw}, z_p^{hs}, z_p^{ws}\}$ through the pose encoder $\mathcal{E}_P$. With pose information provided, the model no longer needs to account for pose and can focus solely on generating lip movements while preserving other facial regions, ultimately improving lip synchronization. Furthermore, to ensure temporal consistency in extended video generation and provide identity cues, we utilize previous video clips as identity frames $X_I$, following (Prajwal et al. 2020). These frames are processed through the identity encoder $\mathcal{E}_I$, resulting in tri-plane representations $Z_I = \{z_i^{hw}, z_i^{hs}, z_i^{ws}\}$. Finally, we concatenate $Z_L$, $Z_P$, and $Z_I$ along the channel axis. Given the conditions of $Z_L$, $Z_P$, and $Z_I$, the pure Gaussian noise $z_T$ is gradually denoised through the MToV process, generating the denoised video latent $\hat{z}_0$. This is then decoded back into the RGB space, $\hat{X}_0$, by a pre-trained decoder. MToV can generate high-fidelity talking head videos that exhibit enhanced temporal consistency and identity preservation, all while requiring significantly less computational time compared to other diffusion-based models.

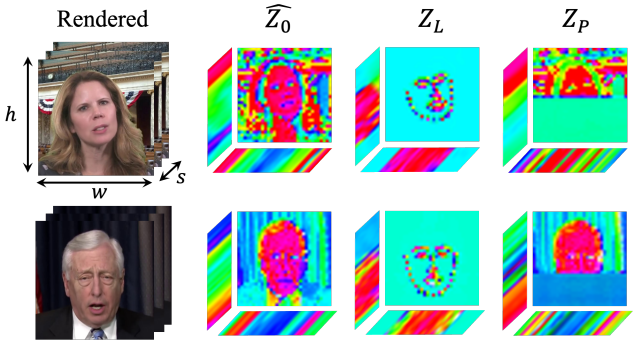



Figure 3: **Visualization of the tri-plane features.** The sequence shows a rendered frames followed by its denoised latents and corresponding condition latents extracted from motion and pose frames. Features are visualized using PCA.

**Training encoders.** MToV includes three encoders: the landmark encoder $\mathcal{E}_L$, identity encoder $\mathcal{E}_I$, and pose encoder $\mathcal{E}_P$. All encoders compress cubic-like 4D video inputs $V \in \mathbb{R}^{S\times C\times H\times W}$ into image-like three 2D latents $\in \mathbb{R}^{c\times h\times w}$, where $S$, $C$, $H$, and $W$ represent sequence, channel, height, and width, while $c$, $h$, and $w$ represent embedded channel, height, and width, respectively. For simplicity, we denote $\mathcal{E}$ to include all three encoders $\mathcal{E}_L$, $\mathcal{E}_I$, and $\mathcal{E}_P$. The encoder $\mathcal{E}$ is trained using an autoencoder (Bertasius, Wang, and Torresani 2021), where $\mathcal{E}$ embeds $V$ in the RGB space into the latent space, and the corresponding decoder $\mathcal{D}$ reconstructs the latent space back into $\hat{V}$ in the RGB space.

In the training phase, we use two different losses: pixel-level reconstruction loss (Zhao et al. 2016) and the perceptual loss (Zhang et al. 2018), which enforce the encoder $\mathcal{E}$ and decoder $\mathcal{D}$ to accurately embed and reconstruct the given inputs in both image space and feature space. The total objective function is formulated as:

$$\mathcal{L}_{\text{encoder}} = \lambda_1 \mathbb{E}_V \left[ \|V - \hat{V}\|_1 \right] + \lambda_2 \mathbb{E}_V \left[ \|\phi(V) - \phi(\hat{V})\|_1 \right], \tag{5}$$

where $\phi$ denotes the perceptual feature extractor (Zhang et al. 2018) and hyperparameter $\lambda_1 = \lambda_2 = 1$.

**Training diffusion model.** In training, the video diffusion model learns to reconstruct talking head videos $\hat{X}$, using landmark embeddings $Z_L$, pose embeddings $Z_P$, and identity embeddings $Z_I$ as conditions. We can redefine Eq. 2 for training MToV, $\mathcal{F}_{\text{MToV}}$, as follows:

$$\mathcal{L}_{\text{MToV}} = \mathbb{E}_{X,t,\epsilon,Z_L,Z_I,Z_P} \left[ \|X_0 - \mathcal{F}_{\text{MToV}}(X_t, t; Z_L, Z_I, Z_P)\|_2^2 \right], \tag{6}$$

where $X_0$ is the ground-truth talking head video, and $X_t$ is the noised $X_0$ at time step $t$.

# 4 Experiments

## Experimental Settings

**Implementation details.** For all experiments, we used single NVIDIA RTX 3090 GPU. For AToM, we train the model for 300k iterations with a learning rate of 1e-4. For MToV, we train the model for 600k iterations with a learning rate of 1e-4. To alleviate jittering, we employed a blending technique using Gaussian blur, as described in (Chen et al. 2020). Additional implementation details are provided in the Appendix 1.

**Datasets.** We used the LRS3-TED (Afouras, Chung, and Zisserman 2018) and HDTF (Zhang et al. 2021) datasets to train our AToM and MToV models, respectively. The LRS3-TED dataset comprises 400 hours of TED videos, providing a large lip-reading corpus. For AToM, we extracted video frames at 25 fps and audio at a 16,000 Hz sampling rate. For MToV, we randomly selected 312 videos from the HDTF dataset for training, using remaining 98 videos for testing.

**Evaluation metrics.** We evaluated using metrics predominantly used in this field. We used **FID** (↓) (Seitzer 2020) and **PSNR** (↑) (Heusel et al. 2017) to assess the image fidelity. Also, we used **CPBD** (↑) (Narvekar and Karam 2011) to evaluate the sharpness of the generated frames, **LPIPS** (↓) (Zhang et al. 2018) to measure the visual resemblance, and **CSIM** (↑) (Deng et al. 2019) to examine the identity preservation. For lip-sync quality, we utilized SyncNet (Chung and Zisserman 2017) scores: **LSE-D** (↓), measuring the distance between lip and audio features. , and **LSE-C** (↑), measuring the confidence score between them. We computed **LMD** (↓) (Chen et al. 2018) which measures the accuracy of generating lip movements.

## Qualitative Results

In Fig. 4, we present results under self-driven and cross-driven settings. While GAN-based methods (Prajwal et al. 2020; Zhou et al. 2021) exhibit sub-optimal lip-sync quality, they suffer from blurry outputs and jitterings. DreamTalk (Ma et al. 2023) falls short in identity preservation and temporal consistency due to its frame-by-frame generation. This is particularly evident in cross-driven scenarios where the model struggles to separate speaker-specific and speaker-agnostic representations due to audio embedding injection via cross-attention. AniPortrait (Wei, Yang, and Wang 2024) contains the unrealistic results with exaggerated lip movements and additional background artifacts due to its teeth inpainting process. In contrast, MoDiTalker excels in generating high-quality videos while preserving identity, temporal consistency, and demonstrating superior generalization. This is attributed to its distinct training process, which extracts intermediate facial landmarks from audio to produce the final video. While 3D coordinate tri-plane hash representation of SyncTalk (Peng et al. 2024) helps preserve identity, which is audio-independent, it compromises lip-sync accuracy and results in jittering and background artifacts. In contrast, our tri-plane representation leverages the temporal axis for smooth frame transitions and enhanced temporal consistency, without sacrificing lip-sync accuracy. Additional comparisons with other previous GAN-based works (Zhou et al. 2020; Wang et al. 2021a) and diffusion-based works (Stypułkowski et al. 2024; Shen et al. 2023) and analyses on the unseen data (Cao et al. 2014) are available in Appendix 3.

## Quantitative Results

In Tab. 1, we present a quantitative comparison with existing approaches (Prajwal et al. 2020; Zhou et al. 2021; Ma et al. 2023; Wei, Yang, and Wang 2024; Peng et al. 2024) using the HDTF dataset (Zhang et al. 2021). MoDiTalker significantly outperforms current state-of-the-art methods across all video quality metrics and the LMD score. Notably, our model also achieves competitive performance in the LSE-D score without utilizing SyncNet in our training, while previous works (Prajwal et al. 2020; Ma et al. 2023) incorporate SyncNet loss during their training phase. This demonstrates the superior generalizability of our method. Notably, AToM surpasses the state-of-the-art landmark generative model, GeneFace (Ye et al. 2023), in both lip synchronization and identity preservation, as shown in quantitative comparison in Tab. 2 and qualitative comparisons in Appendix 5.2. This


Figure 4: We compare our method to state-of-the-art audio-driven methods under both *self-driven* and *cross-driven* settings. It showcases lip shape variations corresponding to specific phonemes in the words “holidays", “and", “time", “enjoying", “life", “teachings", “doctor" and “junior". For a more detailed comparison, please refer to our supplementary video.

is further analyzed in Ablation Study, and detailed in Tab. 2, Tab. 3 and Appendix 4.

## Ablation Study

**Component analysis of AToM** In Tab. 2, we aim to show the effectiveness of each component of AToM. **(I)** presents a baseline diffusion model which generates facial landmark sequences $L$ using only the audio condition $F_A$. **(II)** represents residual prediction, where the model estimates residuals $\Delta L$ rather than directly estimating $L$. Compared to **(I)** and **(II)**, residual prediction significantly enhances lip-sync scores, benefiting from enhanced initialization and training stability. **(III)** incorporates initial landmark embedding $F_L$ as a condition, which remarkably boosts LMD scores and the accuracy of the facial shape in the desired identity $x_{\mathrm{id}}$. Lastly, **(IV)** presents the disentangled framework detailed in the right part of Fig. 2, which separates lip-related and unrelated segments. With disentangled attention, AToM concentrates on lip movements while preserving other facial segments, leading to its highest performance. Combining these, compared to GeneFace (Ye et al. 2023), AToM achieves superior performance in LMD and Lip-sync scores. This is further evident in the qualitative comparison of AToM with GeneFace in Appendix 5.2.

| | Video Quality | | | | | Lip Sync. | |
|---|---|---|---|---|---|---|---|
| | FID ↓ | CPBD ↑ | PSNR ↑ | LPIPS ↓ | CSIM ↑ | LMD ↓ | LSE-D ↓ |
| Real Video | 0.00 | 0.43 | - | - | 1.00 | 0.00 | 6.98 |
| Wav2Lip (Prajwal et al. 2020) | <u>16.34</u> | 0.35 | <u>34.81</u> | <u>0.03</u> | <u>0.90</u> | <u>1.43</u> | **5.86** |
| PC-AVS (Zhou et al. 2021) | 117.85 | 0.29 | 28.23 | 0.38 | 0.35 | 9.36 | <u>7.06</u> |
| DreamTalk (Ma et al. 2023) | 36.98 | 0.43 | 29.59 | 0.20 | 0.72 | 2.56 | 8.37 |
| AniPortrait (Wei, Yang, and Wang 2024) | 37.07 | <u>0.44</u> | 33.86 | 0.09 | 0.75 | 1.92 | 10.36 |
| SyncTalk (Peng et al. 2024) | 26.69 | 0.37 | 30.95 | 0.13 | 0.66 | 3.58 | 10.57 |
| **MoDiTalker** | **14.15** | **0.46** | **35.82** | **0.01** | **0.92** | **1.38** | 9.15 |

Table 1: **Quantitative comparison with the state-of-the-art method on HDTF dataset.** We conduct a quantitative comparison to evaluate performance in terms of video quality and lip synchronization accuracy. A more comprehensive comparison with additional models is provided in the Appendix.

| | Component | LMD ↓ | LSE-C ↑ |
|---|---|---|---|
| | GeneFace (Ye et al. 2023) | 1.41 | 0.339 |
| **(I)** | Baseline | 1.63 | 0.316 |
| **(II)** | (I) + Residual Prediction | 1.47 | 0.386 |
| **(III)** | (II) + Landmark Embedding | 1.31 | 0.385 |
| **(IV)** | (III) + Disentangled Attention (**Ours**) | **1.26** | **0.390** |

Table 2: Component analysis on Audio-to-Motion model.

| | Component | CSIM ↑ | CPBD ↑ | LMD ↓ |
|---|---|---|---|---|
| **(I)** | Baseline | - | - | - |
| **(II)** | (I) + Identity Frames | 0.87 | 0.42 | 11.18 |
| **(III)** | (II) + Pose Frames | 0.91 | 0.44 | 5.23 |
| **(IV)** | (III) + Audio Condition | 0.89 | 0.43 | 2.67 |
| **(V)** | (III) + Facial Landmarks (**Ours**) | **0.92** | **0.46** | **1.38** |

Table 3: Component analysis on Motion-To-Video model.

**Component analysis of MToV** We also evaluate different configurations of MToV, as presented in Tab. 3. This ablation study highlights the significance of different conditions. **(I)** presents the results of the baseline model, which learns the general data distribution without any conditions, but fails to capture the desired lip synchronization, identity, and head pose. **(II)** shows the results of introducing identity frames as a condition, which ensures the model preserves the same identity throughout the generated video. Moreover, **(III)** displays the effect of incorporating pose frames as additional conditions. These extra conditions not only help focus the generation on the mouth region but also accurately capture the head pose, which notably improves both video quality and lip-sync metrics. The introduction of audio embedding through cross-attention layers shown in **(IV)** roughly aligns with the lip movements but may lack precision. **(V)** overcomes the insufficient precision by leveraging facial landmarks from AToM and significantly enhances lip-sync accuracy, demonstrating the effectiveness of leveraging intermediate facial motion to generate high-fidelity videos.

**Computational complexity.** We conducted an efficiency comparison with prior diffusion-based models. Although several studies (Shen et al. 2023; Stypułkowski et al. 2023; Wei, Yang, and Wang 2024) have successfully employed diffusion models to generate talking head videos, their frame-by-frame generation approach poses a significant time-inefficiency. For instance, DiffTalk, Diffused Heads, and AniPortrait require 1003, 716 and 63 seconds, respectively, to produce a 5-second video at 25 fps and a resolution of 256. In contrast, our model needs only 23 seconds, demonstrating a significant efficiency advantage. Our method is 43 times faster than DiffTalk, 31 times faster than Diffused Heads and 2.7 times fater than AniPortrait. This speed gain is especially valuable for producing longer videos, making our approach much more practical for real-world use.

## 5 Conclusion

In this paper, we present MoDiTalker, a novel motion-disentangled diffusion model for generating talking head videos. MoDiTalker addresses the limitations of previous GAN-based methods, such as mode collapse and sub-optimal performance, by incorporating AToM, a module for high-fidelity lip movement generation from audio and identity inputs. Furthermore, MoDiTalker overcomes the temporal inconsistency and high computational cost of prior diffusion-based models through MToV, an efficient video diffusion model that utilizes tri-plane representations as conditions. Our model demonstrates state-of-the-art performance in both qualitative and quantitative evaluations, as well as in user studies. Moreover, we significantly reduce computational demands compared to earlier diffusion-based approaches. Comprehensive ablation studies and user feedback validate the effectiveness of our approach.

## Acknowledgements

This research was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2019-II190075, RS-2024-00509279) and the Culture, Sports, and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism (RS-2024-00345025, RS-2023-00266509, RS-2024-00333068), and National Research Foundation of Korea (RS-2024-00346597).

## References


Afouras, T.; Chung, J. S.; and Zisserman, A. 2018. LRS3-TED: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is Space-Time Attention All You Need for Video Understanding? In *International Conference on Machine Learning*, 813–824. PMLR.

Blanz, V.; and Vetter, T. 2023. A morphable model for the synthesis of 3D faces. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 157–164.

Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S. W.; Fidler, S.; and Kreis, K. 2023. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22563–22575.

Booth, J.; Roussos, A.; Zafeiriou, S.; Ponniah, A.; and Dunaway, D. 2016. A 3d morphable model learnt from 10,000 faces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5543–5552.

Brock, A.; Donahue, J.; and Simonyan, K. 2018. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Cao, H.; Cooper, D. G.; Keutmann, M. K.; Gur, R. C.; Nenkova, A.; and Verma, R. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4): 377–390.

Chen, L.; Li, Z.; Maddox, R. K.; Duan, Z.; and Xu, C. 2018. Lip movements generation at a glance. In *Proceedings of the European conference on computer vision (ECCV)*, 520–535.

Chen, L.; Maddox, R. K.; Duan, Z.; and Xu, C. 2019. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7832–7841.

Chen, R.; Chen, X.; Ni, B.; and Ge, Y. 2020. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2003–2011.

Cho, K.; Lee, J.; Yoon, H.; Hong, Y.; Ko, J.; Ahn, S.; and Kim, S. 2024. GaussianTalker: Real-Time High-Fidelity Talking Head Synthesis with Audio-Driven 3D Gaussian Splatting. *arXiv preprint arXiv:2404.16012*.

Chung, J. S.; and Zisserman, A. 2017. Out of time: automated lip sync in the wild. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, 251–263. Springer.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4690–4699.

He, T.; Guo, J.; Yu, R.; Wang, Y.; Zhu, J.; An, K.; Li, L.; Tan, X.; Wang, C.; Hu, H.; et al. 2023. GAIA: Zero-shot Talking Avatar Generation. *arXiv preprint arXiv:2311.15230*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022. Video diffusion models. *arXiv:2204.03458*.

Hsu, W.-N.; Bolte, B.; Tsai, Y.-H. H.; Lakhotia, K.; Salakhutdinov, R.; and Mohamed, A. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 3451–3460.

Hu, L. 2024. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8153–8163.

Hu, Y.; Chen, Z.; and Luo, C. 2023. LaMD: Latent Motion Diffusion for Video Generation. *arXiv preprint arXiv:2304.11603*.

Ji, X.; Zhou, H.; Wang, K.; Wu, W.; Loy, C. C.; Cao, X.; and Xu, F. 2021. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14080–14089.

Khachatryan, L.; Movsisyan, A.; Tadevosyan, V.; Henschel, R.; Wang, Z.; Navasardyan, S.; and Shi, H. 2023. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Liang, B.; Pan, Y.; Guo, Z.; Zhou, H.; Hong, Z.; Han, X.; Han, J.; Liu, J.; Ding, E.; and Wang, J. 2022. Expressive talking head generation with granular audio-visual control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3387–3396.

Lu, Y.; Chai, J.; and Cao, X. 2021. Live speech portraits: real-time photorealistic talking-head animation. *ACM Transactions on Graphics (TOG)*, 40(6): 1–17.

Ma, Y.; Zhang, S.; Wang, J.; Wang, X.; Zhang, Y.; and Deng, Z. 2023. Dreamtalk: When expressive talking head generation meets diffusion probabilistic models. *arXiv preprint arXiv:2312.09767*.

Miyato, T.; Kataoka, T.; Koyama, M.; and Yoshida, Y. 2018. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*.

Narvekar, N. D.; and Karam, L. J. 2011. A no-reference image blur metric based on the cumulative probability of blur detection (CPBD). *IEEE Transactions on Image Processing*, 20(9): 2678–2683.

Peng, Z.; Hu, W.; Shi, Y.; Zhu, X.; Zhang, X.; Zhao, H.; He, J.; Liu, H.; and Fan, Z. 2024. Synctalk: The devil is in the synchronization for talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 666–676.

Prajwal, K.; Mukhopadhyay, R.; Namboodiri, V. P.; and Jawahar, C. 2020. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, 484–492.

Ren, Y.; Li, G.; Chen, Y.; Li, T. H.; and Liu, S. 2021. PIRenderer: Controllable Portrait Image Generation via Semantic Neural Rendering. arXiv:2109.08379.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Seitzer, M. 2020. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid. Version 0.3.0.

Shen, S.; Zhao, W.; Meng, Z.; Li, W.; Zhu, Z.; Zhou, J.; and Lu, J. 2023. DiffTalk: Crafting Diffusion Models for Generalized Audio-Driven Portraits Animation. arXiv:2301.03786.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, L.; Wu, W.; Fu, C.; Qian, C.; Loy, C. C.; and He, R. 2021. Everything's Talkin': Pareidolia Face Reenactment. *arXiv preprint arXiv:2104.03061*.

Stypułkowski, M.; Vougioukas, K.; He, S.; Zięba, M.; Petridis, S.; and Pantic, M. 2024. Diffused heads: Diffusion models beat gans on talking-face generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5091–5100.

Stypułkowski, M.; Vougioukas, K.; He, S.; Zięba, M.; Petridis, S.; and Pantic, M. 2023. Diffused Heads: Diffusion Models Beat GANs on Talking-Face Generation. arXiv:2301.03396.

Thanh-Tung, H.; and Tran, T. 2020. Catastrophic forgetting and mode collapse in GANs. In *2020 international joint conference on neural networks (ijcnn)*, 1–10. IEEE.

Tian, L.; Wang, Q.; Zhang, B.; and Bo, L. 2024. Emo: Emote portrait alive-generating expressive portrait videos with audio2video diffusion model under weak conditions. *arXiv preprint arXiv:2402.17485*.

Tseng, J.; Castellon, R.; and Liu, K. 2023. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 448–458.

Wang, S.; Li, L.; Ding, Y.; Fan, C.; and Yu, X. 2021a. Audio2Head: Audio-driven One-shot Talking-head Generation with Natural Head Motion. arXiv:2107.09293.

Wang, S.; Li, L.; Ding, Y.; and Yu, X. 2021b. One-shot Talking Face Generation from Single-speaker Audio-Visual Correlation Learning. arXiv:2112.02749.

Wang, T.-C.; Mallya, A.; and Liu, M.-Y. 2021. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10039–10049.

Wang, Y.; Ma, X.; Chen, X.; Dantcheva, A.; Dai, B.; and Qiao, Y. 2023. LEO: Generative Latent Image Animator for Human Video Synthesis. arXiv:2305.03989.

Wei, H.; Yang, Z.; and Wang, Z. 2024. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*.

Wu, J. Z.; Ge, Y.; Wang, X.; Lei, S. W.; Gu, Y.; Shi, Y.; Hsu, W.; Shan, Y.; Qie, X.; and Shou, M. Z. 2023. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7623–7633.

Xu, S.; Chen, G.; Guo, Y.-X.; Yang, J.; Li, C.; Zang, Z.; Zhang, Y.; Tong, X.; and Guo, B. 2024. Vasa-1: Lifelike audio-driven talking faces generated in real time. *arXiv preprint arXiv:2404.10667*.

Ye, Z.; Jiang, Z.; Ren, Y.; Liu, J.; He, J.; and Zhao, Z. 2023. Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. *arXiv preprint arXiv:2301.13430*.

Yin, F.; Zhang, Y.; Cun, X.; Cao, M.; Fan, Y.; Wang, X.; Bai, Q.; Wu, B.; Wang, J.; and Yang, Y. 2022. StyleHEAT: One-Shot High-Resolution Editable Talking Face Generation via Pre-trained StyleGAN. arXiv:2203.04036.

Yu, S.; Sohn, K.; Kim, S.; and Shin, J. 2023. Video probabilistic diffusion models in projected latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18456–18466.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, W.; Cun, X.; Wang, X.; Zhang, Y.; Shen, X.; Guo, Y.; Shan, Y.; and Wang, F. 2023. SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation. arXiv:2211.12194.

Zhang, Z.; Li, L.; Ding, Y.; and Fan, C. 2021. 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*, 3661–3670.

Zhao, H.; Gallo, O.; Frosio, I.; and Kautz, J. 2016. Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging*, 3(1): 47–57.

Zhou, H.; Sun, Y.; Wu, W.; Loy, C. C.; Wang, X.; and Liu, Z. 2021. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4176–4186.

Zhou, Y.; Han, X.; Shechtman, E.; Echevarria, J.; Kalogerakis, E.; and Li, D. 2020. Makelttalk: speaker-aware talking-head animation. *ACM Transactions On Graphics (TOG)*, 39(6): 1–15.