FD2Talk: Towards Generalized Talking Head Generation with Facial Decoupled Diffusion Model

Ziyu Yao
yaozy@stu.pku.edu.cn
Peking University
Beijing, China

Xuxin Cheng
chengxx@stu.pku.edu.cn
Peking University
Beijing, China

Zhiqi Huang*
zhiqihuang@pku.edu.cn
Peking University
Beijing, China

ABSTRACT

Talking head generation is a significant research topic that still faces numerous challenges. Previous works often adopt generative adversarial networks or regression models, which are plagued by generation quality and average facial shape problem. Although diffusion models show impressive generative ability, their exploration in talking head generation remains unsatisfactory. This is because they either solely use the diffusion model to obtain an intermediate representation and then employ another pre-trained renderer, or they overlook the feature decoupling of complex facial details, such as expressions, head poses and appearance textures. Therefore, we propose a **F**acial **D**ecoupled **D**iffusion model for **Talk**ing head generation called **FD2Talk**, which fully leverages the advantages of diffusion models and decouples the complex facial details through multi-stages. Specifically, we separate facial details into motion and appearance. In the initial phase, we design the Diffusion Transformer to accurately predict motion coefficients from raw audio. These motions are highly decoupled from appearance, making them easier for the network to learn compared to high-dimensional RGB images. Subsequently, in the second phase, we encode the reference image to capture appearance textures. The predicted facial and head motions and encoded appearance then serve as the conditions for the Diffusion UNet, guiding the frame generation. Benefiting from decoupling facial details and fully leveraging diffusion models, extensive experiments substantiate that our approach excels in enhancing image quality and generating more accurate and diverse results compared to previous state-of-the-art methods.

CCS CONCEPTS

• **Computing methodologies** → **Animation**.

KEYWORDS

Talking Head Generation, Diffusion Model, Video Generation

ACM Reference Format:
Ziyu Yao, Xuxin Cheng, and Zhiqi Huang. 2024. FD2Talk: Towards Generalized Talking Head Generation with Facial Decoupled Diffusion Model. In

*Corresponding author

MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3681238

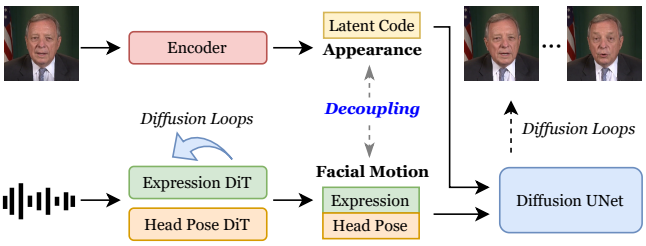

Figure 1: Our proposed FD2Talk leverages diffusion models to generate high-quality and diverse talking head videos. This framework decouples facial information into motion and appearance, thus maintaining motion plausibility, enhancing texture fidelity, and improving generalization.

Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28-November 1, 2024, Melbourne, VIC, Australia. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3664647.3681238

1 INTRODUCTION

Talking head generation is a task that creates a digital representation of a person's head and facial movements synchronized with the audio signal. This technology serves as a cornerstone with far-reaching applications, including virtual reality, augmented reality, and entertainment industries such as film production [17, 54, 55]. With the development of deep learning [11, 52, 53], it has recently attracted numerous researchers and achieved impressive results.

Prevailing methodologies for talking head generation can be broadly divided into two paradigms. One approach involves using a GAN-based framework [9, 12, 16, 22, 29, 43], which simultaneously optimizes a generator and a discriminator. However, due to the inherent flaws of GANs themselves and the suboptimal framework designs, this often results in unsatisfactory results, such as unnatural faces and inaccurate lip movements. The other approach utilizes regression models [5, 15, 18, 23, 46, 61] to map audio to facial movements, ensuring better temporal consistency. Nonetheless, regression-based methods encounter challenges in generating natural movements with individualized characteristics, leading to issues with average facial shapes and less diverse results.

Recently, the rise of diffusion models [21, 32, 39] has marked a new era in generative tasks. Due to their stable generation process and relative ease of training, diffusion models offer a promising avenue for the advancement of talking head technology. While some previous works [13, 36, 40] have attempted to apply diffusion models to talking head generation, their generated results still suffer from low image quality, unnaturalness, and insufficient lip synchronization. We analyze that there are two main issues in

current methods. **1)** Some approaches [27] apply diffusion models solely to predict facial intermediate representations, such as 3DMM coefficients. However, they still rely on pre-trained renderers for rendering the final faces, resulting in low-quality in the generated images. **2)** Other approaches [36, 40] directly generate faces through pixel-level denoising, globally conditioned on the audio and reference image. Nevertheless, they overlook the fact that faces contain rich information, such as expressions, poses, texture, etc. Previous methods couple these facial details, significantly complicating denoising generation and yielding unsatisfactory results.

To address the above issues, we propose the **F**acial **D**ecoupled **D**iffusion model for **Talk**ing head generation, named **FD2Talk**. Our FD2Talk leverages the generative advantages of diffusion models to generate high-quality, diverse and natural talking heads videos. As illustrated in Fig. 1, the proposed FD2Talk model is a multi-stage framework that decouples complex facial details into motion and appearance information. The first phase focuses on motion information generation, while the second phase is dedicated to driving frame synthesis. **1) Motion Generation.** Motion information, including lip movements, expressions, and head poses, is highly related to the given audio and is more decoupled from facial appearance, making it easier to learn. In the first stage, we design novel Diffusion Transformers to extract motion-only information, *i.e.*, 3DMM expression and head pose coefficients, from the raw audio. Through the denoising process, we generate natural and accurate motions, thereby enhancing the realism of our final outputs. Additionally, predicting the head pose coefficients at this stage enables us to produce more diverse motions compared to previous methods. **2) Frames Generation.** Moving on to the second stage, we first encode the reference image to capture appearance information, including human identity and texture characteristics. Combining this appearance information with the previously learned motion, we obtain a comprehensive facial representation related to the final RGB faces. Unlike previous methods that utilize a pre-trained face renderer to render final frames, we design a conditional Diffusion UNet and utilize motion and appearance as conditions to guide higher-quality and more natural animated frame generation.

Our two-stage approach not only maintains motion plausibility and accuracy, but also enhances texture fidelity. *Moreover, by focusing on generating appearance-independent information in the first stage, we can enhance the generalization ability of our FD2Talk.* This is because we can obtain pure motion coefficients from the audio signal without being influenced by the portrait domains. The contribution can be summarized as follows:

- Our proposed FD2Talk is a multi-stage framework that effectively decouples facial motion and appearance, enabling accurate motion modeling, superior texture synthesis, and improved generalization.
- Our approach fully leverages the generative power of diffusion models in both motion and frames generation stages, thus enhancing the quality of the results.
- Extensive experiments demonstrate that our method excels at generating accurate and realistic talking head videos, achieving state-of-the-art performance. By incorporating head pose modeling, our FD2Talk produces significantly more diverse results compared to previous methods.

2 RELATED WORKS

Audio-Driven Talking Head Generation. Previous methods have attempted to utilize generative adversarial networks [4, 6, 29, 41, 44, 45, 59, 60] and regression models, such as RNN [42], LSTM [18, 46, 61] and Transformer [1, 15] to synthesis talking head videos based on audio signals. Among GAN-based methods, [29] proposed a novel lip-synchronization network that generates talking head videos with accurate lip movements across different identities by learning from a powerful lip-sync discriminator. [59] disentangled person identity and speech information through adversarial learning, leading to improved talking head generation. [44] introduced a temporal GAN with three discriminators focused on achieving detailed frames, audio-visual synchronization, and realistic expressions, capable of generating lifelike talking head videos. On the other hand, in regression-based methods, [18] adopts LSTM for better temporal consistency using explicit and implicit keypoints as the intermediate representation. Additionally, [15] proposed a Transformer-based autoregressive model that encodes long-term audio context and autoregressively predicts a sequence of animated 3D face meshes. Despite significant progress, the unrealistic results in GAN-based generation and the average facial shape problem in regression-based models remain unresolved.

Diffusion Models for Talking Head Generation. Diffusion models have demonstrated the remarkable ability across multiple generative tasks, such as image generation [30, 34, 35], image inpainting [24, 50, 51], and video generation [3, 20, 25]. Recently, some studies [13, 36, 40] have delved into using diffusion models for talking head generation. However, these studies still face challenges in producing natural and accurate faces. On one hand, they [27] generate intermediate representations using diffusion models but rely on pre-trained face renderers for synthesizing the final frames. On the other hand, they [36, 40] globally utilize audio features to condition the generation of faces, which couples the complex facial motion and appearance. To fully leverage the advantages of the diffusion model and disentangle the complex facial information, we utilize the diffusion model in both motion generation and frame generation, thereby achieving better performance.

3 METHOD

Given a reference image $\mathcal{I} \in \mathbb{R}^{3\times H\times W}$ and a corresponding audio input, our model is designed to synthesize a realistic talking head video $\mathcal{V} \in \mathbb{R}^{3\times F\times H\times W}$ with lip movements synchronized with the audio signal. Here, the symbols F, H, and W denote the frame numbers, frame height and frame width respectively.

Our FD2Talk framework consists of two stages that decouple facial information into motion and appearance, thus enhancing the modeling of facial representation. We employ powerful diffusion models in both stages, making FD2Talk a fully diffusion-based approach that produces high-quality talking head results. Specifically, we start by using Diffusion Transformers to predict expressions and pose motions from the audio input. In the subsequent stage, we utilize a Diffusion UNet to generate final RGB images, conditioned on the previously predicted motion information along with appearance texture information extracted from a reference image.

3.1 Preliminary Knowledge

3.1.1 3D Morphable Model. To generate high-quality talking heads, we integrate 3D information into our method, specifically employing the 3D Morphable Model (3DMM) [10] to decouple the facial representation from a given face image. This allows us to describe the 3D face space (3D mesh) using Principal Component Analysis:

$$\mathbf{S} = \mathbf{S}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \bar{\mathbf{S}} + \mathbf{B}_{id}\boldsymbol{\alpha} + \mathbf{B}_{exp}\boldsymbol{\beta}. \tag{1}$$

Here, $\mathbf{S} \in \mathbb{R}^{3N}$ (where N represents the number of vertices of a face, and 3 represents the axes x, y, and z) denotes a 3D face, while $\bar{\mathbf{S}}$ is the mean shape. $\boldsymbol{\alpha} \in \mathbb{R}^{D_\alpha}$ and $\boldsymbol{\beta} \in \mathbb{R}^{D_\beta}$ represent the predicted coefficients of identity and expression, respectively. $\mathbf{B}_{id}$ and $\mathbf{B}_{exp}$ are the PCA bases of identity and expression. Moreover, rotation coefficients $\boldsymbol{r} \in SO(3)$ and translation coefficients $\boldsymbol{t} \in \mathbb{R}^3$ represent the head rotation and translation, respectively, collectively constituting the facial pose coefficients $\boldsymbol{p} = [\boldsymbol{r}, \boldsymbol{t}]$.

3.1.2 Diffusion Model. Diffusion models are formulated as time-conditional denoising networks that learn the reverse process of a Markov Chain with a length T. Specifically, starting from the clean signal $\boldsymbol{x}_0$, the process of adding noise can be denoted as follows:

$$\boldsymbol{x}_t = \sqrt{\overline{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\overline{\alpha}_t}\epsilon_t. \tag{2}$$

Here, $\epsilon_t \sim \mathcal{N}(0,1)$ denotes random Gaussian noise, while $\overline{\alpha}_t$ represents the hyper-parameter for the diffusion process. $\boldsymbol{x}_t$ refers to the noisy feature at step t, where $t \in [1, \ldots, T]$. During inference, the T-step denoising process progressively denoise random Gaussian noise $\mathcal{N}(0,1)$ to estimate the clean signal $\boldsymbol{x}_0$. In our work, all diffusion-based models are designed to predict signal itself rather than noise. Thus, the overall goal can be described as follows:

$$L := \mathbb{E}_{\boldsymbol{x}_0, t}\left[\|\boldsymbol{x}_0 - \theta(\boldsymbol{x}_t, t, \boldsymbol{c})\|_2^2\right], \tag{3}$$

where θ represents the diffusion model and c represents conditional guiding. We utilize the $L2$ error between the estimated signal and the ground truth $\boldsymbol{x}_0$.

3.2 Motion Generation with Diffusion Transformers

Early diffusion-based methods [36, 40] globally utilize audio signals as a condition for the pixel-level denoising process. However, this approach combines motion and appearance, making it challenging for overall training convergence. In contrast, in the first stage, our method focuses on generating motion-only information from the audio signal, specifically 3DMM expression and head pose coefficients. These coefficients exclusively represent facial and head motion, which are highly decoupled from the appearance textures and greatly influence lip synchronization and motion diversity. Furthermore, compared to high-dimensional RGB faces, low-dimensional 3DMM coefficients are considerably easier for the model to learn.

To ensure smooth continuity between different frame motions and fully leverage the diffusion models, we introduce sequence-to-sequence Diffusion Transformers for generating both expression and pose coefficients. Meanwhile, to effectively address the one-to-many mapping problem and accurately predict lip movements and diverse head poses, we decouple the prediction of expression and pose coefficients using an Expression Transformer θ_{exp} and a Pose Transformer θ_{pose}, which is illustrated in Fig. 2.

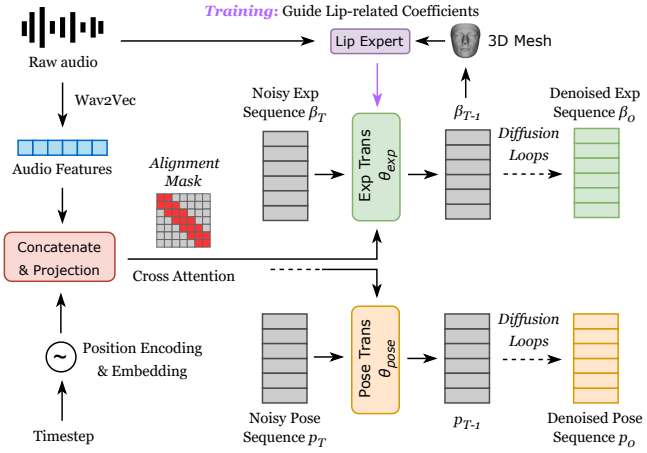

Figure 2: Pipeline of the motion generation. We decouple the motion into expression and head poses, both of which are predicted by our designed DiTs. The audio guides the generation through cross-attention layers, utilizing an alignment mask to ensure accurate lip movements. Furthermore, the pre-trained lip expert also enhances the lip synchronization.

Specifically, we initialize the noisy expression sequence $\boldsymbol{\beta}_T \in \mathbb{R}^{F\times D_\beta}$ and noisy pose sequence $\boldsymbol{p}_T \in \mathbb{R}^{F\times D_p}$ from random Gaussian noise $\mathcal{N}(0,1)$, where F represents the number of frames aligned with the final video. We then denoise the $\boldsymbol{\beta}_T$ and $\boldsymbol{p}_T$ conditioned on audio features through T loops to estimate denoised sequence $\boldsymbol{\beta}_0$ and $\boldsymbol{p}_0$. Here, the length of audio clip $\boldsymbol{A}$ is aligned with F, and we adopt the state-of-the-art self-supervised pre-trained speech model, Wav2Vec 2.0 [2], to extract the audio features.

Taking θ_{pose} as an example, at each timestep t, we concatenate the embedding from the timestep and audio features to obtain the condition c. We then project c to an intermediate representation $\tau(c) \in \mathbb{R}^{F\times D_\tau}$ using a linear layer. Then, $\tau(c)$ is fused into θ_{pose} via the cross-attention layer, where the query ($\mathbf{Q}$) is derived from $\boldsymbol{p}_t$, while the key ($\mathbf{K}$) and value ($\mathbf{V}$) are obtained from $\tau(c)$. Meanwhile, we design an alignment mask $\mathcal{M}$ to ensure the consistency of generated coefficients and the audio signal, so that $\tau(c)$ for the i^{th} timestamp attends to $\boldsymbol{p}_t$ at the j^{th} timestamp only if $j-k \le i \le j+k$. For FD2Talk, we empirically set $k = 3$. The $\mathcal{M}$ can be denoted as:

$$\mathcal{M} = \begin{cases} True, & \text{if } j-k \le i \le j+k \\ False, & \text{otherwise} \end{cases} \tag{4}$$

In our diffusion process, we directly estimate the original signal. Therefore, after L-layer Pose Transformer, we obtain $\tilde{\boldsymbol{p}}_0$. Subsequently, we can calculate the single-step denoising result $\boldsymbol{p}_{t-1}$:

$$\boldsymbol{p}_{t-1} = \sqrt{\overline{\alpha}_{t-1}}\tilde{\boldsymbol{p}}_0 + \frac{\sqrt{1-\overline{a}_{t-1}-\sigma_t^2}}{\sqrt{1-\overline{a}_t}}(\boldsymbol{p}_t - \sqrt{\overline{a}_t}\tilde{\boldsymbol{p}}_0) + \sigma_t\epsilon, \tag{5}$$

where σ_t is the Gaussian covariance at the t^{th} timestep.

The Exp Transformer θ_{exp} and Pose Transformer θ_{pose} share the same architecture, and the denoising process for $\boldsymbol{\beta}_t$ is identical to that for $\boldsymbol{p}_t$. Therefore, after T iterations, we obtain $\boldsymbol{\beta}_0$ and $\boldsymbol{p}_0$ as the final values for the expression and pose coefficients.

3.3 Frame Generation with Diffusion UNet

Previous methods primarily employ pre-trained face renderers [31, 47] to generate final RGB faces, whose performance sets an upper bound on talking face generation. Therefore, we design a conditional Diffusion UNet θ_{unet} to generate the final frame based on previously predicted 3DMM coefficients, aiming to utilize the diffusion models to achieve diverse and realistic faces generation.

To reduces computational overhead and accelerates convergence, we introduce a pair of encoder $\mathcal{E}$ and decoder $\mathcal{D}$ [32] to transition the frame generation into the latent space. Suppose the downsampling factor is $f = H/h = W/w$, then we can encode the reference image $\mathcal{I}$ into the reference latent code $x = \mathcal{E}(\mathcal{I}) \in \mathbb{R}^{d\times h\times w}$.

As shown in Fig. 3, we initialize the noisy latent image $\boldsymbol{J}_T \in \mathbb{R}^{d\times h\times w}$ from $\mathcal{N}(0,1)$, then we progressively denoise it conditioned on both reference latent code x and 3DMM coefficients $\boldsymbol{\beta}_0$ and $\boldsymbol{p}_0$. Here, x encompasses the appearance texture of the reference image, while $\boldsymbol{\beta}_0$ and $\boldsymbol{p}_0$ includes the driving facial and head motions.

An intuitive approach is to directly concatenate the x, $\boldsymbol{\beta}_0$ and $\boldsymbol{p}_0$ to obtain the conditions. However, we observe that this operation leads to difficulties in training convergence, because there exists a gap between the image domain and the motion coefficients domain. To address the impact of domain gap, we use two cross-attention layers to introduce these two conditions respectively. Specifically, both the encoder and decoder of Diffusion UNet consist of two cross-attention layers, denoted as ϕ_1 and ϕ_2. The coefficients $\boldsymbol{\beta}_0$ and $\boldsymbol{p}_0$ are concatenated, following with a linear projection, to form the condition for ϕ_1. The calculation of ϕ_1 can be defined as:

$$\boldsymbol{m}_1 = \phi_1(\{\boldsymbol{\beta}_0, \boldsymbol{p}_0\}, \boldsymbol{J}_t), \tag{6}$$

where the query (**Q**) is from $\boldsymbol{J}_t$, and the key (**K**) and value (**V**) are from the condition $\{\boldsymbol{\beta}_0, \boldsymbol{p}_0\}$. Then, in the second layer ϕ_2, we utilize the reference latent code x as the condition to guide this process:

$$\boldsymbol{m}_2 = \phi_2(x, \boldsymbol{m}_1), \tag{7}$$

where the query (**Q**) is from $\boldsymbol{m}_1$, and the key (**K**) and value (**V**) are derived from x. Here, the x is reshaped into sequence, and positional encoding is also introduced. This decoupling of conditions enhances the denoising stability, leading to higher-quality results.

Similar to that in the first stage, at each diffusion timestep t, we predict $\tilde{\boldsymbol{J}}_0$ from $\boldsymbol{J}_t$, and then calculate the corresponding $\boldsymbol{J}_{t-1}$ using the Eq. (5). After T iterations, this process generates the accurate denoised latent image $\boldsymbol{J}_0$. The reference latent code and the denoised latent image are further concatenated as the input of decoder $\mathcal{D}$, allowing us to generate the RGB image $\mathcal{V}_i$, which serves as each frame for the talking head video $\mathcal{V} = \{\mathcal{V}_i\}_1^F$. Moreover, as we denoise in the latent space, we can easily extend to higher-resolution talking head synthesis by adjusting the downsampling factor f, thereby further enhancing our generation quality.

3.4 Training Strategies

Our training process consists of two stages. In the first stage, we train the Exp Transformer and Pose Transformer to generate accurate expression and pose coefficients. Using these accurate coefficients as a foundation, we then train the Diffusion UNet in the second stage to generate natural and diverse RGB frames.

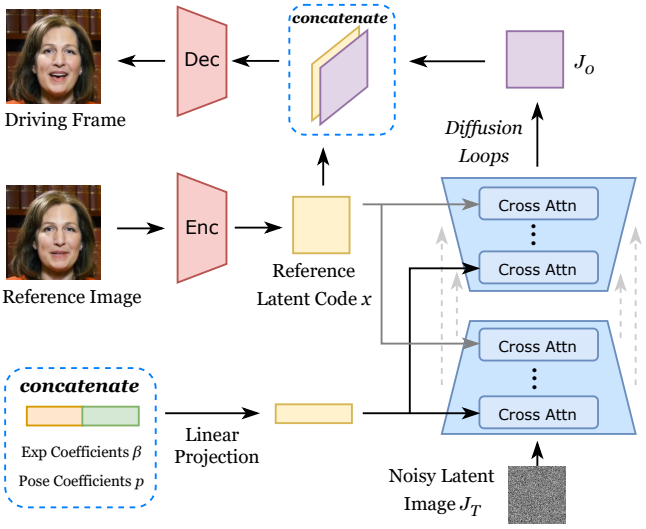

Figure 3: Pipeline of the frame generation. The facial appearance extracted from the reference image and the predicted motion coefficients are fused within the Diffusion UNet using distinct cross-attention layers to prevent interference.

3.4.1 Motion Generation Stage. In the first stage, we randomly extract a video clip along with the corresponding audio clip $\boldsymbol{A}$ from the training set. We utilize the Deep3d [10] method to generate the expression coefficient sequence $\boldsymbol{\beta}_0$ and pose coefficient sequence $\boldsymbol{p}_0$ from this video clip. $\boldsymbol{\beta}_0$ and $\boldsymbol{p}_0$ also serve as the ground truths. Then, our Exp Transformer θ_{exp} and Pose Transformer θ_{pose} can be trained using the tuples $(\boldsymbol{\beta}_0, t, A)$ and $(\boldsymbol{p}_0, t, A)$, respectively.

For the Exp Transformer θ_{exp}, by adding random Gaussian noise, the $\boldsymbol{\beta}_0$ can become $\boldsymbol{\beta}_t$ using Eq. (5). The θ_{exp} estimates $\tilde{\boldsymbol{\beta}}_{\boldsymbol{0}} = \theta_{exp}(\boldsymbol{\beta}_t, t, A)$, and the objective can be defined as follows:

$$\mathcal{L}_{exp} = \mathbb{E}_{\boldsymbol{\beta}_0, t}\left[\left\|\boldsymbol{\beta}_0 - \theta_{exp}(\boldsymbol{\beta}_t, t, A)\right\|_2^2\right]. \tag{8}$$

Similar to the θ_{exp}, the objective of the Pose Transformer θ_{pose} is:

$$\mathcal{L}_{pose} = \mathbb{E}_{\boldsymbol{p}_0, t}\left[\left\|\boldsymbol{p}_0 - \theta_{pose}(\boldsymbol{p}_t, t, A)\right\|_2^2\right]. \tag{9}$$

While the random noise introduced in the diffusion model can effectively facilitate the diverse generation, it also leads to inaccurate mouth shape generation to some extent. Therefore, we utilize a pre-trained lip expert [29] to guide this denoising process and generate more accurate mouth shape. Specifically, we first obtain the identity coefficients from the reference image, and then calculate the 3D meshes using these identity coefficients along with the predicted expression coefficients $\tilde{\boldsymbol{\beta}}_{\boldsymbol{0}}$ via Eq. (1). From these 3D meshes, we select vertices in the mouth area to represent lip motion [26]. The pre-trained lip expert calculates the cosine similarity between mouth motion embedding v and audio embedding a as follows:

$$P_{sync} = \frac{v \times a}{max(\|v\|_2 \times \|a\|_2, \epsilon)}, \tag{10}$$

where ϵ is a small number for avoiding the division-by-zero. Then, the θ_{exp} minimizes the synchronous loss as follows:

$$\mathcal{L}_{sync} = -\log(P_{sync}). \tag{11}$$

Table 1: Comparison with the state-of-the-art methods on HDTF and VoxCeleb dataset. The best results are highlighted in bold, and the second best is underlined. Our FD2Talk surpasses previous methods in motion diversity and image quality, as well as offering competitive lip synchronization performance. The data presented in the table are in the order of *HDTF / VoxCeleb*.

Methods	Lip Synchronization		Motion Diversity		Image Quality		
	LSE-C ↑	SyncNet ↑	Diversity ↑	Beat Align ↑	FID ↓	PSNR ↑	SSIM ↑
Ground Truth	8.32 / 6.29	7.99 / 5.73	0.256 / 0.307	0.276 / 0.319	—	—	—
Wav2Lip [29]	**10.08** / **8.13**	**8.06** / **6.40**	N./A. / N./A.	N./A. / N./A.	22.67 / 23.85	32.33 / 35.19	0.740 / 0.653
MakeItTalk [61]	4.89 / 2.96	3.72 / 2.67	0.238 / 0.260	0.221 / 0.252	28.96 / 31.77	17.95 / 21.08	0.623 / 0.529
SadTalker [57]	6.11 / 4.51	5.19 / 4.88	0.275 / 0.319	0.296 / 0.328	23.76 / 24.19	35.78 / 37.90	0.746 / 0.690
DiffTalk [36]	6.06 / 4.38	4.98 / 4.67	0.235 / 0.258	0.226 / 0.253	23.99 / 24.06	36.51 / 36.17	0.721 / 0.686
DreamTalk [27]	6.93 / 4.76	5.46 / 4.90	0.236 / 0.257	0.213 / 0.249	24.30 / 23.61	32.82 / 33.16	0.738 / 0.692
Ours	7.29 / 5.16	6.63 / 5.66	**0.338** / **0.359**	**0.336** / **0.377**	**20.96** / **21.89**	**38.89** / **39.95**	**0.779** / **0.756**

Overall, the first stage optimizes the following loss:

$$\mathcal{L}_{first} = \lambda_{exp}\mathcal{L}_{exp} + \lambda_{pose}\mathcal{L}_{pose} + \lambda_{sync}\mathcal{L}_{sync}, \tag{12}$$

where λ_{exp}, λ_{pose} and λ_{sync} are the weight factors to control the three losses in the same numeric scale.

3.4.2 Frame Generation Stage. We utilize the pre-trained [14] encoder $\mathcal{E}$ and decoder $\mathcal{D}$ as the foundation for learning in the latent space. Given that the input channel for the decoder in our method is $2 \times d$, we opt to substitute the first convolution layer of the decoder. Subsequently, we fine-tune both the encoder and decoder using frames from the training set. Specifically, in each iteration, we randomly select two frames F_1 and F_2 from a single video and then calculate the reconstruction loss as follows:

$$\mathcal{L}_{rec} = \|F_2 - \mathcal{D}([\mathcal{E}(F_1), \mathcal{E}(F_2)])\|_2^2. \tag{13}$$

Meanwhile, we introduce the perceptual loss [56] to enforce $\mathcal{E}$ and $\mathcal{D}$ to accurately reconstruct the frames in the image space:

$$\mathcal{L}_{per} = \|\phi(F_2) - \phi(\mathcal{D}([\mathcal{E}(F_1), \mathcal{E}(F_2)]))\|_1, \tag{14}$$

where ϕ represents the perceptual feature extractor [56]. Then the overall objective of encoder $\mathcal{E}$ and decoder $\mathcal{D}$ can be defined as:

$$\mathcal{L}_{e\&d} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{per}\mathcal{L}_{per}, \tag{15}$$

where λ_{rec} and λ_{per} control the numeric scales.

During the training of Diffusion UNet θ_{unet}, we randomly extract a video clip along with its corresponding audio clip. The first frame from this video clip serves as the reference image $\mathcal{I}$. Utilizing the trained encoder $\mathcal{E}$, we obtain the reference latent code x, as well as the ground truths for each latent image $\boldsymbol{J}_0$. Subsequently, we employ the trained Exp Transformer and Pose Transformer to acquire the $\tilde{\boldsymbol{\beta}}_0 \in \mathbb{R}^{F \times D_\beta}$ and $\tilde{\boldsymbol{p}}_0 \in \mathbb{R}^{F \times D_p}$. Different from the sequence-to-sequence Diffusion Transformer in the first stage, our Diffusion UNet generate each RGB frames one by one, so we extract the coefficient $\tilde{\boldsymbol{\beta}} \in \mathbb{R}^{D_\beta}$ and $\tilde{\boldsymbol{p}} \in \mathbb{R}^{D_p}$ for each frame. The training of our Diffusion UNet is facilitated by a tuple denoted as $(\boldsymbol{J}_0, t, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{p}}, x)$. Specifically, we add the random Gaussian noise on $\boldsymbol{J}_0$ to obtain the noisy latent image $\boldsymbol{J}_t$ at the t-th timestep. We then optimize θ_{unet} using the following objective function:

$$\mathcal{L}_{second} = \mathbb{E}_{\boldsymbol{J}_0, t}\left[\left\|\boldsymbol{J}_0 - \theta_{unet}(\boldsymbol{J}_t, t, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{p}}, x)\right\|_2^2\right]. \tag{16}$$

4 EXPERIMENTS

4.1 Experimental Setup

Datasets. We use HDTF [58] and VFHQ [49] datasets to train our FD2Talk. HDTF is a large in-the-wild high-resolution and high-quality audio-visual dataset that consists of about 362 different videos spanning 15.8 hours. The resolution of the face region in the video generally reaches 512×512. VFHQ is a large-scale video face dataset, which contains over 16000 high-fidelity clips of diverse interview scenarios. However, since VFHQ lacks audio components, it is exclusively utilized during the second phase of training. All videos are clipped into small fragments and cropped [37] to obtain the face region. Then we use Deep3d [10], a single-image face reconstruction method, to recover the facial image and extract the relevant coefficients. Both HDTF and VFHQ are split 70% as the training set, 10% as the validation set, and 20% as the testing set. Moreover, we introduce VoxCeleb [28] to further evaluate our method, which contains over $100k$ videos of 1251 subjects.

Implementation Detail. We train the model on video frames with 256×256 resolution. In the first stage, the 6-layer Exp Transformer and Pose Transformer are trained with a batch size of 1 and a generated sequence length of 25. In the second stage, we first fine-tune the pre-trained [14] encoder and decoder, and then train the Diffusion UNet with a batch size of 32, and the resolution of the latent image is 64×64. The two-stage framework is trained with the Adam [8] optimizer separately and can be inferred in an end-to-end fashion. The diffusion step is set to 1000 and 50 during training and inference, respectively. Our two-stage model is trained for approximately 8 and 32 hours using 8 NVIDIA 3090 GPUs.

Baselines. We compare our method with several previous methods of audio-driven talking head generation, including Wav2Lip [29], MakeItTalk [61], SadTalker [57], DiffTalk [36], and DreamTalk [27]. We provide a reference image and audio signal as input for all methods. Note that Wav2Lip requires additional videos to offer head pose information, so we also fixed the head pose of our method for a fair comparison in quantitative evaluation.

Evaluation Metrics. To evaluate the superiority of our proposed method, we consider three aspects: 1) Lip synchronization is assessed using two metrics: LSE-C [29] and SyncNet [7]. LSE-C measures the confidence score of perceptual differences in mouth shape from Wav2Lip, while the SyncNet score assesses the audio-visual

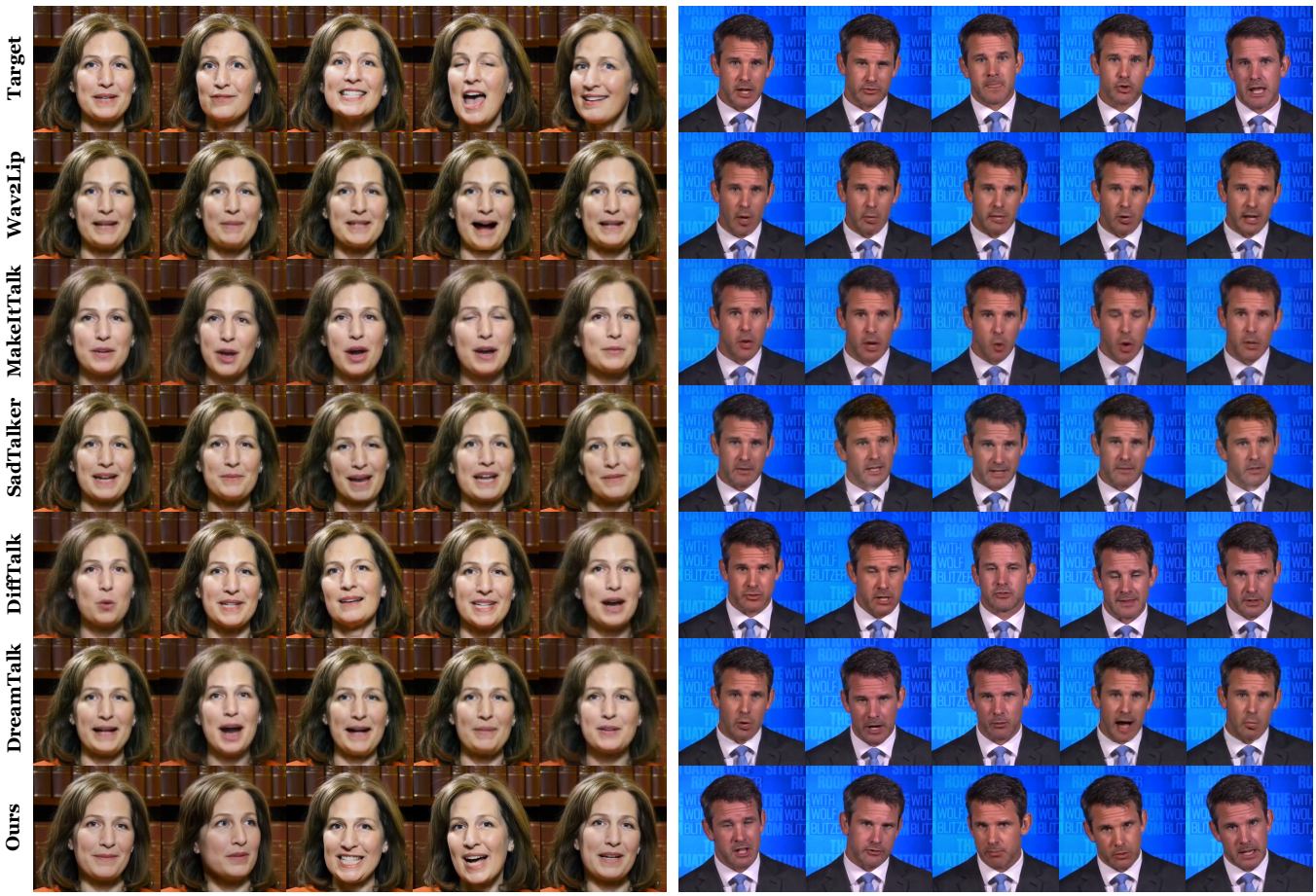

Figure 4: Qualitative comparison with several state-of-the-art methods. Our FD2Talk achieves superior lip synchronization compared to previous methods while preserving naturalness and high image quality. By leveraging diffusion models for predicting head motion, our generated results also exhibit enhanced motion diversity.

synchronization quality. 2) Motion diversity is evaluated by extracting head motion feature embeddings using Hopenet [33] and calculating their standard deviations. Additionally, we use the Beat Align Score [38] to measure alignment between the audio and generated head motions. 3) Generated image quality is assessed using widely recognized metrics: FID [19], PSNR, and SSIM [48].

4.2 Qualitative Comparison

We compare our method with previous state-of-the-art methods qualitatively. The results are visualized in Fig. 4. While Wav2Lip can generate accurate lip movements, it falls short in producing high-quality images due to blurriness issues in the mouth region. Moreover, Wav2Lip focuses solely on animating the lips, neglecting other facial areas and resulting in a lack of motion diversity. MakeItTalk and SadTalker attempt to address some weaknesses of Wav2Lip, such as enhancing motion diversity. However, they still struggle to synthesize detailed facial features like apple cheeks and teeth due to generative limitations in GANs and regression models. For diffusion-based methods, DiffTalk combines appearance and motion during denoising, leading to inaccurate lip movement generation. DreamTalk, on the other hand, neglects head pose modeling and still relies on pre-trained render models, resulting in synthesized results with unreasonable head poses and slightly distorted facial regions. In contrast, our FD2Talk fully leverages powerful diffusion models in both stages and effectively separates appearance and motion information. These operations result in accurate lip movements, diverse head poses, and high-quality, lifelike talking head videos.

4.3 Quantitative Comparison

We further quantitatively analyze the comparison between FD2Talk and previous state-of-the-art methods in lip synchronization, motion diversity, and image quality, on HDTF and VoxCeleb datasets.

Our approach surpasses MakeItTalk, SadTalker, DiffTalk, and DreamTalk in terms of lip synchronization. We attribute this improvement to the alignment mask used during cross-attention in the Exp and Pose Transformer. This mask enables the predicted coefficients consistency with the corresponding audio signal. Additionally, the accurate lip movements are further enhanced by the lip

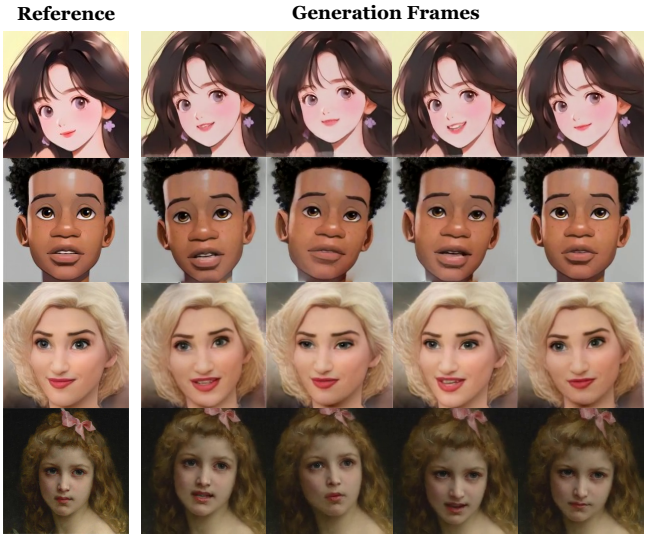

Figure 5: Our FD2Talk demonstrates strong generalization when applied to out-of-domain portraits. We generate each talking head video using the same audio but different portrait domains, which significantly diverge from training data.

Figure 6: The visualization results of: 1) Utilizing a single DiT to predict expressions and head poses jointly; 2) Concatenating the two conditions of UNet; and 3) Our full FD2Talk model. We can observe that using a single DiT makes the results less diverse and synchronized, while concatenating two conditions leads to distorted and unnatural faces.

synchronization loss with a well-pretrained lip expert. It is worth noting that although Wav2Lip achieves the highest lip accuracy, it neglects the overall naturalness and diversity of the results.

When considering the three metrics of image quality, *i.e.*, FID, PSNR, and SSIM, our approach significantly outperforms previous methods, which can be attributed to two aspects: 1) Our method maximizes the potential of diffusion models to generate more natural results compared to previous works using GANs, regression models, or partial diffusion models. 2) We disentangle complex facial information through two stages, enabling accurate motion prediction, and the creation of natural, high-fidelity appearance textures, ultimately resulting in superior and high-quality results.

Moreover, our work surpasses previous methods in the diversity of head motions and achieves the best performance in Diversity and Beat Align Score. This achievement is attributed to our Pose Transformer, which predicts the head pose coefficients through the denoising process. The introduced random noise facilitates the generation of richer and more diverse pose results compared to previous methods.

4.4 Generalization Performance

We also test the generalization of our FD2Talk model for out-of-domain portraits. As demonstrated in Fig. 5, whether the provided faces are paintings, cartoon portraits, or oil paintings, our FD2Talk can animate them using audio signals, ensuring lip synchronization while preserving the appearance details of the reference face with high fidelity, thus enhancing image quality. Moreover, the generated results include rich head poses, demonstrating excellent motion diversity as expected. This generalization ability stems from the decoupling of facial representation. In the first stage, we focus on generating appearance-independent motion information, which is solely linked to the audio signal and remains robust across various portrait domains.

4.5 Ablation Studies

4.5.1 Decoupling Expressions and Head Poses. In the first stage, we decouple the Diffusion Transformers for the prediction of expressions and poses to address the one-to-many mapping issue. We compare this approach with a baseline where a Diffusion Transformer is used to jointly predict expression and pose coefficients. As shown in Tab. 2 and Fig. 6, this baseline exhibits a noticeable decrease in lip synchronization and motion diversity. This is because lip movements are heavily influenced by facial expressions but have little correlation with head pose. On the other hand, motion diversity is closely related to predicted pose coefficients. Jointly learning these coefficients leads to mutual interference and makes training more challenging. Therefore, we choose to decouple the prediction of expression and pose coefficients using Exp and Pose Transformers, respectively.

4.5.2 Conditions of Diffusion UNet. In the second stage, the predicted motion information and encoded appearance texture are passed through distinct cross-attention layers to guide the Diffusion UNet. We verify its effectiveness by comparing it with a baseline where we directly concatenate these two conditions and guide the denoising process. As demonstrated in Tab. 2 and Fig. 6, concatenating motion and appearance leads to a decrease in each metric, particularly image quality, as we can observe the distortion in faces. We analyze that appearance textures constitute image-domain information, which is much higher than coefficient-domain motion. Therefore, decoupling them using two distinct cross-attention layers can significantly enhance the robustness of overall diffusion models and ensure convergence.

Table 2: Ablation studies on the 1) Decoupling of Diffusion Transformers and 2) Conditions of the Diffusion UNet.

Settings	Lip Synchronization		Motion Diversity		Image Quality		
	LSE-C ↑	SyncNet ↑	Diversity ↑	Beat Align ↑	FID ↓	PSNR ↑	SSIM ↑
Single Diffusion Transformer	3.11	3.08	0.193	0.189	22.06	37.66	0.761
Concatenate UNet Conditions	4.79	4.66	0.249	0.251	30.79	29.91	0.523
Full (Our FD2Talk)	**7.31**	**6.26**	**0.322**	**0.331**	**21.32**	**38.10**	**0.776**

Table 3: Ablation studies of lip synchronization. w/o alignment: We remove the alignment mask in DiTs. w/o $\mathcal{L}_{sync}$: We eliminate the constraint from the pre-trained lip expert.

Settings	Lip Synchronization	
	LSE-C ↑	SyncNet ↑
w/o alignment	4.66	3.97
w/o $\mathcal{L}_{sync}$	5.35	4.63
Full (Our FD2Talk)	**7.31**	**6.26**

Table 4: User studies results.

Methods	Lip Sync	Motion Diversity	Image Quality
Wav2Lip	24.9%	1.2%	2.1%
MakeItTalk	3.6%	2.7%	3.5%
SadTalker	16.8%	23.6%	17.8%
DiffTalk	12.9%	8.1%	16.5%
DreamTalk	15.2%	10.6%	8.5%
Ours	**26.6%**	**53.8%**	**51.6%**

4.5.3 Ablation Studies of Lip Synchronization. In FD2Talk, we ensure lip synchronization from two aspects: **1) Aligning the audio and motions during cross-attention.** When we integrate audio features into the network, an alignment mask $\mathcal{M}$ is designed to ensure the consistency of generated coefficients and audio. To assess its significance, we conduct an experiment by removing the $\mathcal{M}$. As indicated in Tab. 3, the absence of $\mathcal{M}$ notably affects lip synchronization. Our analysis demonstrates that without $\mathcal{M}$, the motion generation in each timestamp is misled by audio from other unrelated timestamps. **2) Guided with the pre-trained lip expert.** During the training of Exp Transformer, we utilize a pre-trained lip expert to constrain the lip-related coefficients using $\mathcal{L}_{sync}$. Here, we remove it to compare the effectiveness of $\mathcal{L}_{sync}$. As shown in Tab. 3, when the model is trained without $\mathcal{L}_{sync}$, lip synchronization significantly drops. We attribute this to the fact that the coefficients are generated through a denoising process, which means introduced random noise may lead to inaccurate lip shapes. Fig. 7 also shows that utilizing the alignment mask and training with $\mathcal{L}_{sync}$ result in much better lip synchronization for the generated faces.

4.6 User Studies

We conduct user studies with 20 participants to evaluate the performance of all methods. We generate 30 test videos covering different genders, ages, styles, and expressions. For each method, participants are required to choose the best one based on three metrics: 1) lip synchronization, 2) head motion diversity, and 3) overall image quality. As demonstrated in Tab. 4, our work outperforms previous methods across all aspects, particularly in motion diversity and image quality. We attribute this to the decoupling of motion and appearance, as well as adopting diffusion models to generate higher-quality frames.

Figure 7: Comparison of 1) w/o AM: without alignment mask, 2) w/o LE: training without lip expert, and 3) full FD2Talk. We can notice that both the alignment mask and pre-trained lip expert can enhance lip synchronization of our model.

5 CONCLUSION

Talking head generation is an important research topic that still faces great challenges. Considering the issues of previous works, such as reliance on generative adversarial networks (GANs), regression models, and partial diffusion models, and neglecting the disentangling of complex facial representation, we propose a novel facial decoupled diffusion model, called FD2Talk, to generate high-quality, natural, and diverse results. Our FD2Talk fully leverages the strong generative ability of diffusion models and decouples the high-dimensional facial information into motion and appearance. We firstly utilize Diffusion Transformers to predict the accurate 3DMM expression and head pose coefficients from the audio signal, which serves as the decoupled motion-only information. Then these motion coefficients are fused into the Diffusion UNet, along with the appearance texture extracted from the reference image, to guide the generation of final RGB frames. Extensive experiments demonstrate that our approach surpasses previous methods in generating more accurate lip movements and yielding higher-quality and more diverse results.

REFERENCES

[1] Shivangi Aneja, Justus Thies, Angela Dai, and Matthias Nießner. 2023. Facetalk: Audio-driven motion diffusion for neural parametric head models. *arXiv preprint arXiv:2312.08459* (2023).

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

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127* (2023).

[4] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. 2020. Talking-head generation with rhythmic head motion. In *European Conference on Computer Vision*. Springer, 35–51.

[5] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. 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.

[6] Sen Chen, Zhilei Liu, Jiaxing Liu, Zhengxiang Yan, and Longbiao Wang. 2021. Talking head generation with audio and speech related facial action units. *arXiv preprint arXiv:2110.09951* (2021).

[7] Joon Son Chung and Andrew Zisserman. 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*. Springer, 251–263.

[8] Kingma Da. 2014. A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[9] Dipanjan Das, Sandika Biswas, Sanjana Sinha, and Brojeshwar Bhowmick. 2020. Speech-driven facial animation using cascaded gans for learning of motion and texture. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*. Springer, 408–424.

[10] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. 2019. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 0–0.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[12] Michail Christos Doukas, Stefanos Zafeiriou, and Viktoriia Sharmanska. 2021. Headgan: One-shot neural head synthesis and editing. In *Proceedings of the IEEE/CVF International conference on Computer Vision*. 14398–14407.

[13] Chenpeng Du, Qi Chen, Tianyu He, Xu Tan, Xie Chen, Kai Yu, Sheng Zhao, and Jiang Bian. 2023. Dae-talker: High fidelity speech-driven talking face generation with diffusion autoencoder. In *Proceedings of the 31st ACM International Conference on Multimedia*. 4281–4289.

[14] Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12873–12883.

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

[16] Kuangxiao Gu, Yuqian Zhou, and Thomas Huang. 2020. Flnet: Landmark driven fetching and learning network for faithful talking facial animation synthesis. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 10861–10868.

[17] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. 2021. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5784–5794.

[18] Siddharth Gururani, Arun Mallya, Ting-Chun Wang, Rafael Valle, and Ming-Yu Liu. 2023. Space: Speech-driven portrait animation with controllable expression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 20914–20923.

[19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

[20] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303* (2022).

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

[22] Prajwal KR, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and CV Jawahar. 2019. Towards automatic face-to-face translation. In *Proceedings of the 27th ACM international conference on multimedia*. 1428–1436.

[23] Yuanxun Lu, Jinxiang Chai, and Xun Cao. 2021. Live speech portraits: real-time photorealistic talking-head animation. *ACM Transactions on Graphics (TOG)* 40, 6 (2021), 1–17.

[24] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11461–11471.

[25] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. 2023. Videofusion: Decomposed diffusion models for high-quality video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10209–10218.

[26] Yifeng Ma, Suzhen Wang, Zhipeng Hu, Changjie Fan, Tangjie Lv, Yu Ding, Zhidong Deng, and Xin Yu. 2023. Styletalk: One-shot talking head generation with controllable speaking styles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 1896–1904.

[27] Yifeng Ma, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yingya Zhang, and Zhidong Deng. 2023. Dreamtalk: When expressive talking head generation meets diffusion probabilistic models. *arXiv preprint arXiv:2312.09767* (2023).

[28] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612* (2017).

[29] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 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.

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

[31] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. 2021. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13759–13768.

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[33] Nataniel Ruiz, Eunji Chong, and James M Rehg. 2018. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2074–2083.

[34] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22500–22510.

[35] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems* 35 (2022), 36479–36494.

[36] Shuai Shen, Wenliang Zhao, Zibin Meng, Wanhua Li, Zheng Zhu, Jie Zhou, and Jiwen Lu. 2023. Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1982–1991.

[37] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First order motion model for image animation. *Advances in neural information processing systems* 32 (2019).

[38] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. 2022. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11050–11059.

[39] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

[40] Michał Stypułkowski, Konstantinos Vougioukas, Sen He, Maciej Zięba, Stavros Petridis, and Maja Pantic. 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.

[41] Yasheng Sun, Hang Zhou, Ziwei Liu, and Hideki Koike. 2021. Speech2Talking-Face: Inferring and Driving a Face with Synchronized Audio-Visual Representation.. In *IJCAI*, Vol. 2. 4.

[42] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)* 36, 4 (2017), 1–13.

[43] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. 2020. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6142–6151.

[44] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 2020. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision* 128, 5 (2020), 1398–1413.

[45] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. 2020. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on*

Computer Vision. Springer, 700–717.

[46] S Wang, L Li, Y Ding, C Fan, and X Yu. 2021. Audio2Head: Audio-driven One-shot Talking-head Generation with Natural Head Motion. In *International Joint Conference on Artificial Intelligence*. IJCAI.

[47] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. 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.

[48] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

[49] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. 2022. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 657–666.

[50] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. 2023. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22428–22437.

[51] Shiyuan Yang, Xiaodong Chen, and Jing Liao. 2023. Uni-paint: A unified framework for multimodal image inpainting with pretrained diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*. 3190–3199.

[52] Jiawei Yao, Chuming Li, Keqiang Sun, Yingjie Cai, Hao Li, Wanli Ouyang, and Hongsheng Li. 2023. Ndc-scene: Boost monocular 3d semantic scene completion in normalized device coordinates space. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, 9421–9431.

[53] Jiawei Yao, Qi Qian, and Juhua Hu. 2024. Multi-modal proxy learning towards personalized visual multiple clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14066–14075.

[54] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. 2020. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*. Springer, 524–540.

[55] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. 2019. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9459–9468.

[56] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 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.

[57] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. 2023. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8652–8661.

[58] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. 2021. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3661–3670.

[59] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. 2019. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 9299–9306.

[60] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. 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.

[61] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. 2020. Makelttalk: speaker-aware talking-head animation. *ACM Transactions On Graphics (TOG)* 39, 6 (2020), 1–15.