Face Animation with an Attribute-Guided Diffusion Model

Bohan Zeng[1]*, Xuhui Liu[1]*, Sicheng Gao[1]*, Boyu Liu[1], Hong Li[1]
Jianzhuang Liu[2], Baochang Zhang[1,3]†
[1]Beihang University
[2]Shenzhen Institutes of Advanced Technology, University of Chinese Academy of Sciences
[3]Zhongguancun Laboratory, Beijing, China

Figure 1. Comparison with SOTA face animation methods on reconstruction and reenactment tasks. **Left:** FOMM [42] often produces artifacts and over-smoothing textures, while FADM is capable of rectifying the distortions and enriching the fine-grained details. **Right:** Face vid2vid [49] suffers from typical unnatural facial details, while FADM prominently improves the overall visual quality.

Abstract

Face animation has achieved much progress in computer vision. However, prevailing GAN-based methods suffer from unnatural distortions and artifacts due to sophisticated motion deformation. In this paper, we propose a Face Animation framework with an attribute-guided Diffusion Model (FADM), which is the first work to exploit the superior modeling capacity of diffusion models for photo-realistic talking-head generation. To mitigate the uncontrollable synthesis effect of the diffusion model, we design an Attribute-Guided Conditioning Network (AGCN) to adaptively combine the coarse animation features and 3D face reconstruction results, which can incorporate appearance and motion conditions into the diffusion process. These specific designs help FADM rectify unnatural artifacts and distortions, and also enrich high-fidelity facial details through iterative diffusion refinements with accurate animation attributes. FADM can flexibly and effectively improve existing animation videos. Extensive experiments on widely used talking-head benchmarks validate the effectiveness of FADM over prior arts. The source code is available in https://github.com/zengbohan0217/FADM.

*These authors contributed equally.
†Corresponding Author: bczhang@buaa.edu.cn.

1. Introduction

Face animation, referring to the task of animating a still face with poses and expressions provided by a driving video, has drawn increasing attention due to its wide application scenarios, such as photography, online conferencing, social media, and video production. With the progress of generative models such as Generative Adversarial Networks (GANs), recent face animation methods have achieved impressive performance in synthesizing high-fidelity talking faces. However, they still suffer from undesirable artifacts and distortions in generated results.

Existing face animation methods are mostly based on GAN models, which mainly divide the generation process into warping and rendering. They utilize the difference of expressions and poses between the source image and the driving video to calculate the motion flow, which can guide the further warping process of the encoded source features. After that, the warped features are fed into a decoding module for rendering and synthesizing the final results. These methods can be roughly classified into three categories: model-free [4, 41, 42, 50], landmark-based [48, 54, 55] and 3D structure-based [7, 27, 49]. They obtain promising performances in preserving the identity and appearance of the source and generate relatively accurate motion from the driving video. However, due to the restricted ability of

adversarial learning on high-fidelity appearance reconstruction, these GAN-based methods focus more on face distributions but not much on facial details, and thus they might generate unnatural artifacts and distortions (see Fig. 1).

Recently, the great success of denoising diffusion probabilistic models (DMs) in computer vision, such as inpainting [5, 17, 28, 31, 38–40, 45], video synthesis [14, 18, 53, 57], and 3D points cloud modeling [32, 33, 59], indicates their superior capacity in generative tasks. They can model highly complex data distributions through a sequence of diffusion refinement steps. Based on optimizing a variant of the variational lower bound, diffusion models can effectively avoid the distortion problem encountered by GANs and generate high-fidelity facial details. However, existing DMs tend to encode images into arbitrarily high-variance latent spaces without specific attribute restrictions, which is unqualified for face animation that has explicit requirements on the facial appearance, pose, and expression.

In this paper, we enable face animation with an iterative denoising diffusion process to rectify the distortions and unnatural artifacts, while ensuring accurate animation attributes. We propose a Face Animation framework with an attribute-guided Diffusion Model (FADM) for photo-realistic talking-head generation. Specifically, we introduce a Coarse Generative Module (CGM) to obtain the preliminary animation results, which provide low-resolution features for the diffusion process. To mitigate the high variability of DMs, we design an Attribute-Guided Conditioning Network (AGCN) to incorporate appearance and motion conditions into the iterative refinement process. On the one hand, we utilize an encoder network to extract the appearance code from the driving frames and coarse features and introduce an MSE loss to align them. On the other hand, we leverage a 3D reconstruction module to predict the poses and expressions of the source and driving frames. Based on these, AGCN uses a Multi-Layer Perceptron (MLP) to assign different confidence values for the multi-resolution features, to adaptively adjust the expressed ratio of the coarse features and fuse them effectively as motion condition. Therefore, the diffusion refined process is well guided to synthesize accurate and fine-grained talking-head videos, as shown in Fig. 1. Moreover, it is worth noting that FADM can also be directly applied to improving the quality of existing animation videos as a flexible talking-head rectification tool. The contributions of this paper are summarized as follows:

- We propose a Face Animation framework with an attribute-guided Diffusion Model (FADM) to rectify the distortions and unnatural artifacts, which can also enrich the facial details through an iterative diffusion refinement process.
- We design an Attribute-Guided Conditioning Network (AGCN) to adaptively extract appearance and motion conditions for the diffusion process and ensure the validity of generated results. Moreover, FADM can flexibly and effectively improve the quality of available animation videos.
- Extensive experiments are conducted to compare FADM with state-of-the-art methods. The results show that FADM generates overall best qualitative and quantitative results on widely used talking-head benchmarks, and genuinely achieves photo-realistic face animation.

2. Related Work

2.1. GAN-based Face Animation Models

Model-free methods [4, 41, 42, 49, 50, 52] learns the motion field for the deformation of face images in a self-supervised manner without additional facial priors. MonkeyNet [41] predicts sparse key-points to complete motion transfer. In particular, the First Order Motion Model (FOMM) [42] significantly improves the performance of face animation with a rigorous first-order mathematical model. Face vid2vid [49] extends FOMM by introducing 3D representations and achieves realistic face animation. Nevertheless, it has a considerable computational cost and performs poorly in expression transformation. Landmark-based methods [1, 10, 13, 48, 54, 55] utilize 2D facial landmarks as conditions for reenactment. However, these methods often cannot handle identity preservation well during the generation process. More recently, many 3D structure-based works [7, 23, 24, 27, 36, 47, 56] resort to the geometric prior of 3D faces and achieve impressive results on realistic talk-head synthesis. HeadGAN [7] takes the rendered 3D mesh as input and predicts the depth to deform the face, but it fails in expression transferring. Conditioned on the parameters of the 3D Morphable Model (3DMM) [3], StyleRig [46] and GIF [11] respectively employ pre-trained StyleGAN [21] and StyleGAN2 [22] to warp face images. PIRenderer [36] controls the face motions and predicts a flow field for deformation. And FNeVR [56], AD-NeRF [12] adopt the NeRF [34] to generate high-quality face animation.

Unfortunately, all of these methods rely on the GAN models to generate animated faces, which also brings unnatural distortions and artifacts. This paper enables current face animation methods with diffusion refinements to achieve high-fidelity face animation.

2.2. Diffusion Probabilistic Models

Diffusion Probabilistic Models (DMs) are first presented in [43] as a kind of generative models, which use a Markov chain to gradually add noise to obtain a latent variable, and then gradually transform the latent variable to obtain

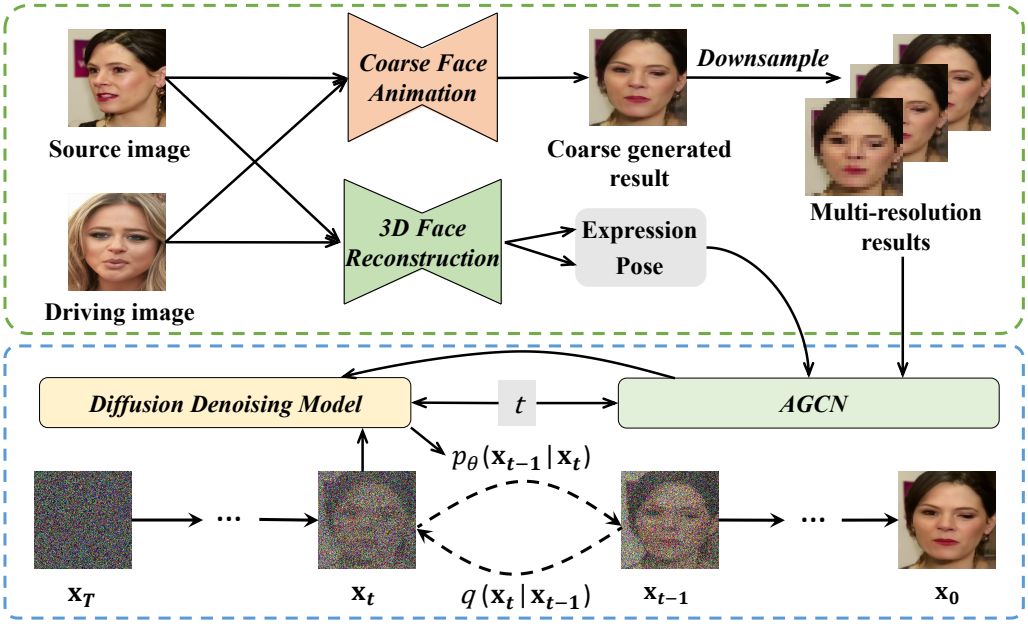

Figure 2. FADM framework. It consists of the coarse face animation generator, the encoder of the 3D face reconstruction model DECA [8], and the core diffusion rendering module (blue box). We first obtain the coarsely generated result and the facial expression and pose information, on which AGCN is then performed to estimate the appearance and motion conditions for further rendering by the diffusion model.

the generated results through a learned iterative denoising process. Recently, DMs have achieved state-of-the-art results in various synthesis tasks, including image synthesis [16, 17, 39], speech synthesis [26, 29], and 3D cloud related tasks [32, 33, 59]. In image synthesis, [16] and [44] show the superior capability of diffusion models to generate high-quality images in many computer vision tasks. For example, [5, 9, 17, 28, 31, 38–40, 45] exhibit impressive performance on image super-resolution and in-painting. Particularly, the stable diffusion model [38] has been applied to various practical application scenarios, such as text-image generation, and has given rise to a wave of DMs. Likewise, the amazing performance of DMs in semantic segmentation [2], point cloud completion and generation [32, 33, 59], and video generation [14, 18, 53, 57] again demonstrates their excellent capabilities.

However, given that current DMs mostly have no strict requirements on the attributes of the generated results, the results often lie in an arbitrarily high-variance space. In contrast, face animation strictly demands animating the source with explicit poses and expressions provided by the driving video, while preserving the appearance of the source.

3. Method

Since GANs have limited capability to model complex facial structures and motion for our task, existing methods often suffer from essential distortions and unnatural artifacts. To address this problem, we propose the Face Animation framework with a Diffusion Model (FADM), which is comprised of: (1) a coarse generative module, (2) a 3D face reconstruction model, (3) an attribute-guided conditioning network, and (4) a diffusion rendering module. The overview of FADM is shown in Fig. 2. In this section, we describe the details of FADM and elaborate on how it rectifies current face animations with explicit and accurate attributes through a sequence of diffusion refinement steps.

3.1. Coarse Generative Module

In FADM, we first use a Coarse Generative Module (CGM) such and FOMM [42] or Face vid2vid [49] to generate coarse animation images. Given a source image s and the driving frames d, the objective of CGM is to deform s with the expression and pose information derived from d, while keeping the appearance information of s. It includes two steps: warping the source features first according to the expressions and poses of the driving frames, and then rendering warped features to obtain final animation images. Intuitively, the general process of generating the coarse animation results g can be conducted as:

$$g = G(\text{Warp}(s, exp_d, pose_d)), \tag{1}$$

where G represents the generative model, and exp_d and $pose_d$ denote the expressions and poses of the driving

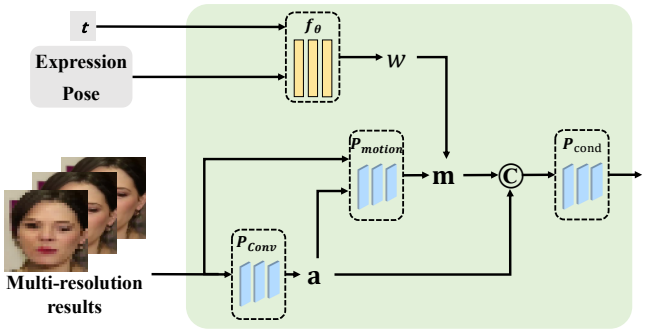

Figure 3. Architecture of AGCN during inference, where © indicates the channel-wise concatenation.

frames, respectively. Although the coarse results have a promising performance in preserving the appearance of s and transferring motion from d, there often exist undesirable distortions in the facial details and the background area beyond the face. To alleviate this problem, we leverage a diffusion process with explicit conditions to renovate the coarse results.

3.2. Diffusion Rendering Module

To handle the distortion problem caused by CGM, a diffusion rendering module (the blue box in Fig. 2) is designed in FADM to synthesize photo-realistic images through an iterative diffusion process from coarse to fine. Here we first give the preliminaries of DMs, and then describe this module for face animation.

3.2.1 Preliminaries of Diffusion Models

Following [16], we define the inference process p_θ of DMs, which denoises a normally distributed variable x_T to a target image x_0 as:

$$\begin{aligned} p_\theta(\mathbf{x}_{0:T}) =& p(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \\ p(\mathbf{x}_T) =& \mathcal{N}\left(\mathbf{x}_T \mid 0, I\right), \\ p_\theta(\mathbf{x}_{t-1}|x_t) =& \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)), \end{aligned} \tag{2}$$

where $\mathbf{x}_1, ..., \mathbf{x}_T$ are latent features with added noise, $p_\theta(\mathbf{x}_{0:T})$ represents the joint distribution which performs the image generation process and is defined as a Markov chain with learnable Gaussian transitions $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$. Inversed to the inference process, the forward process gradually adds Gaussian noise to $\mathbf{x}_0$ over T iterations, which can be expressed as:

$$\begin{aligned} q(\mathbf{x}_{1:T}|\mathbf{x}_0) =& \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \\ q(\mathbf{x}_t|\mathbf{x}_{t-1}) =& \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \end{aligned} \tag{3}$$

where $\beta_1, ..., \beta_T$ are the variance schedule. By optimizing the negative log-likelihood of the data distribution through the variational lower bound, the optimization objective can be interpreted as learning an equally weighted sequence of a denoising model $z_\theta(\mathbf{x}_t, t), t \in \{1, 2, ..., T\}$:

$$\mathcal{L}_\theta = E_{\mathbf{x}_0, z\sim\mathcal{N}(0,1), t}[\| z - z_\theta(\mathbf{x}_t, t) \|_2^2]. \tag{4}$$

3.2.2 Diffusion Rendering for Face Animation

In order to avoid the mismatch between the DM's high-variance encoding and the explicit attribute requirements of face animation, the DM needs to generate the rendering face frames in strict conformity with the appearance of the source, having the poses and expressions of the driving frames. Accordingly, we formulate the inference process of our diffusion rendering model as:

$$\begin{aligned} p_d(\mathbf{x}_{0:T}) =& p(\mathbf{x}_T)\prod_{t=1}^{T} p_d(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{a}, \mathbf{m}), \\ p_d(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{a}, \mathbf{m}) =& \mathcal{N}(\mathbf{x}_{t-1}; \mu_d(\mathbf{x}_t, t, \mathbf{a}, \mathbf{m}), \Sigma_d(\mathbf{x}_t, t, \mathbf{a}, \mathbf{m})), \end{aligned} \tag{5}$$

where $\mathbf{a}$ donates the appearance code of the source image, and $\mathbf{m}$ donates the motion condition derived from the variation between the source image and driving frames' poses and expressions. Consequently, the optimization objective of our diffusion rendering model is defined as a conditional diffusion loss $\mathcal{L}_d$:

$$\mathcal{L}_d = E_{\mathbf{x}_0, (\mathbf{a},\mathbf{m}), z\sim\mathcal{N}(0,1), t}[\| z - z_d(\mathbf{x}_t, t, (\mathbf{a}, \mathbf{m})) \|_2^2]. \tag{6}$$

To minimize $\mathcal{L}_d$ with $(\mathbf{a}, \mathbf{m})$, we design an Attribute-Guided Conditioning Network (AGCN) to extract appropriate appearance and motion conditions, and fuse them adaptively for navigating the diffusion process, as shown in Fig. 3. The appearance condition is used to provide the diffusion process with faithful characteristics of the source image, while the motion condition can impose restriction on the generated poses and expressions and dynamically modify the facial details.

Appearance Condition. Considering the current training style of using the same identity of the source and the driving frames, we note that the driving frames are the most appropriate conditions to provide faithful appearance information for the subsequent diffusion process, while only the coarse animation results are available during the inference process, which are the sub-optimal choice. Formally, we design a CNN encoder P_{Conv} to extract the appearance code $\mathbf{a}$:

$$\mathbf{a} = \begin{cases} P_{\text{Conv}}(\downarrow_* (d)), & \text{in training} \\ P_{\text{Conv}}(\downarrow_* (g)), & \text{in inference,} \end{cases} \tag{7}$$

where $\downarrow_*$ denotes the downsampling operation. As stated above, since the coarse animation results might involve unexpected interference with the appearance, we aim to alleviate the interference and guarantee the creditability of $\mathbf{a}$ provided during inference. Specifically, we respectively input

the driving frames and coarse animation results into P_{Conv}, and then align the appearance conditions predicted from d and g through an MSE loss $\mathcal{L}_{color}$ as:

$$\mathcal{L}_{color} = \text{MSE}(P_{\text{Conv}}(\downarrow_* (d)), P_{\text{Conv}}(\downarrow_* (g))). \quad (8)$$

Here d and g are both fixed features irrelevant to t. $\mathcal{L}_{color}$ works by facilitating P_{Conv} to extract the most valuable appearance information from g. Consequently, P_{Conv} is capable of providing faithful appearance conditions during the inference, which are comparable with those provided by d in the training process.

Motion Condition. Considering FADM aims to improve the quality of the coarse animation results, taking them as the motion condition seems to be an intuitively effective choice for the diffusion process. Nonetheless, they may also bring distortions in the diffusion process. Empirically, the coarse results suffer from extreme distortions when the motion changes dramatically between the source and the driving frames. In this case, features with a higher resolution tend to contain more distortions, while features with a lower resolution could weaken them. In other words, compared with high-resolution features, low-resolution features allow the diffusion process to synthesize richer facial details to compensate for the distortions. Based on this observation, we handle this problem in an adaptive way, seeking a balance in fusing multi-resolution coarse animation results as the motion condition, to alleviate the distortions and enrich the facial details on the basis of ensuring accurate animation attributes.

Specifically, we first utilize the downsampling operation to process the coarse animation result into three coarse animation features with different resolutions. Meanwhile, we exploit the advanced 3D face reconstruction model DECA [8] to extract the facial poses $pose$ and expressions exp from the source and the driving frames, and concatenate them as the motion state. Then, an MLP f_θ is introduced as a motion measuring function to model the changing amplitude of the motion between the source image and the driving frames, so as to obtain the motion weight w. The process (Fig. 3) is represented as :

$$w = f_\theta(\text{Concat}(exp_s, pose_s) - \text{Concat}(exp_d, pose_d), t), \quad (9)$$

where t is an arbitrary timestep of the diffusion process.

With the initial weight, we assign a larger value to the features with a lower resolution when the motion changes drastically. If the motion does not change much, the features with a higher resolution should be allocated greater weights for guaranteeing high-fidelity generation. Formally, we calculate the motion condition $\mathbf{m}$ by:

$$\begin{aligned} w_i &= \frac{K-i}{K} \cdot \exp(w-\alpha) + \frac{i}{K} \cdot \exp(-w+\alpha), \\ \mathbf{m} &= \sum_{i=1}^{K} w_i \cdot P_{motion}(g_i, \mathbf{a}), \end{aligned} \quad (10)$$

where α is a hyper-parameter, g_i and w_i denote the coarse generated images and the motion weights for refined images with different resolutions, respectively, K is the number of resolutions, and P_{motion} is a CNN to generate the motion conditions in different resolutions. Note that we set $\alpha = 0.3$ in this paper.

Lastly, we introduce a CNN P_{cond} to fuse the appearance condition $\mathbf{a}$ and the motion condition $\mathbf{m}$ together. In general, the objective of our diffusion model is rewritten as:

$$\mathcal{L}_d = E_{\mathbf{x}_0,(\mathbf{a},\mathbf{m}),z\sim\mathcal{N}(0,1),t}[\| z - z_d(\mathbf{x}_t, t, P_{cond}(\mathbf{a}, \mathbf{m}, t)) \|_2^2], \quad (11)$$

where z_d denotes the diffusion denoising model. Following [44], we employ an U-net architecture as the denoising model which is optimized to iteratively remove the noise, and synthesize high-fidelity target faces in 100 timesteps.

It is worth noting that our FADM can be directly used to improve the visual quality of existing animated videos by setting the first frame as the source image, bringing great convenience in practice. Related analysis and visual results are provided in the supplementary materials.

4. Experiments

4.1. Implementation Details

Datasets. We evaluate the performance of FADM on three datasets: VoxCeleb [35], VoxCeleb2 [6], and CelebA [30]. VoxCeleb contains about 100,000 videos covering 1,251 different speakers. VoxCeleb2 has more than 1M videos of different celebrities. CelebA consists of 200,000 images of 10,000 different persons with different genders and multi-age groups. Note that we use the images from CelebA as the source images to evaluate the performance on the reenactment task. Following FOMM [42], we preprocess the data by cropping faces from the videos and resizing them to 256×256.

Training Details. We use the pretrained FOMM [42] or Face vid2vid [49] to generate the coarse face animation results and then train FADM for about 100 epochs with the images from the videos repeating 75 times per epoch. We adopt the Adam [25] optimizer with learning rate $\eta = 2 \times 10^{-4}$, $\gamma_1 = 0.5$ and $\gamma_2 = 0.9$. Furthermore, we use four 24GB NVIDIA 3090 GPUs for training.

Evalution Metrics. The evaluation metrics include: (1) $\mathcal{L}_1$, PSNR, and SSIM [51] (2) LPIPS [58] and FID [15] (3) Average Keypoint Distance (AKD) and Average Euclidean Distance (AED) as set in [42]; (4) identity preservation cosine similarity (CSIM) [20, 37] calculated by CircularFace [20]

Table 1. Quantitative comparison of same-identity reconstruction on VoxCeleb [35].

Method	$\mathcal{L}_1 \downarrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	AKD $\downarrow$	AED $\downarrow$
Bilayer [54]	0.1753	0.5733	12.802	0.3201	13.83	0.0564
PIRender [36]	0.0574	0.2225	21.154	0.6564	2.249	0.0321
FOMM [42]	0.0451	0.1479	23.422	0.7521	1.456	0.0247
Face vid2vid [49]	0.0456	0.1395	23.279	0.7487	1.615	0.0258
DaGAN [19]	0.0468	0.1465	23.449	0.7564	1.546	0.0257
FADM	**0.0402**	**0.1379**	**24.434**	**0.7841**	**1.392**	**0.0241**

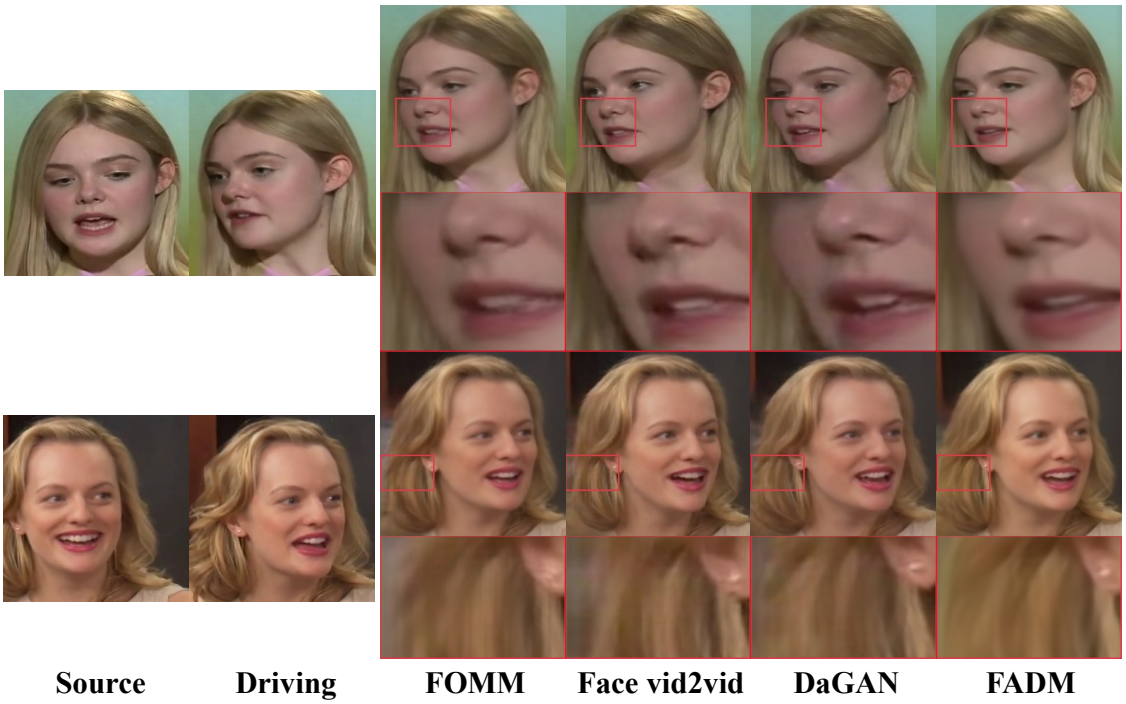

Figure 4. Qualitative comparison with SOTA methods on the reconstruction task. Evidently, our FADM can produce more fine-grained details, and is effective to rectify the unnatural parts.

4.2. Comparison with State-of-the-Art Methods

Methods. We compare our FADM with five state-of-the-art methods: FOMM [42], Face vid2vid [49], Bilayer [54], DaGAN [19], and PIRenderer [36]. For Bilayer, FOMM, DaGAN, and PIRenderer, we use their official pre-trained models for evaluation, while for Face vid2vid, we adopt a widely recognized unofficial model due to the absence of the official code. All of these models are pre-trained on VoxCeleb.

Same-Identity Reconstruction. We conduct quantitative and qualitative comparisons of the same-identity reconstruction task on the VoxCeleb dataset, where FOMM [42] is used to generate the coarse face animation results. In order to accelerate the inference, we randomly select several short-time videos from the testing dataset of VoxCeleb for quantitative comparison, and the selected video list is available in the supplementary materials. As shown in Table 1, compared to other SOTA methods, it is evident that FADM achieves the best performance in all metrics, especially the reconstruction faithfulness $\mathcal{L}_1$ and the visual quality LPIPS, demonstrating the superiority of our FADM in generating high-fidelity face animation. Moreover, we show the visual results of FADM and existing SOTA methods in Fig. 4. FADM exhibits overall better quality with more fine details than other methods.

Figure 5. Visual comparison with SOTA methods on the reenactment task. **(a)** Results on VoxCeleb [35]. **(b)** Results on VoxCeleb2 [6]. On both testing datasets, our FADM can produce more identity-preserving and photo-realistic results compared with other methods. Particularly, compared with the most outstanding existing method Face vid2vid, FADM can effectively alleviate the unnatural artifacts on the generated face images.

Table 2. Quantitative comparison for cross-identity reenactment on the testing datasets of VoxCeleb [35], VoxCeleb2 [6], and CelebA [30].

Method	VoxCeleb		VoxCeleb2		CelebA	
	FID↓	CSIM ↑	FID↓	CSIM ↑	FID↓	CSIM ↑
FOMM	106.9	0.5491	**138.1**	0.5228	96.29	0.5410
Face vid2vid	**106.6**	0.6447	148.6	0.6290	93.44	0.6218
DaGAN	110.3	0.5305	139.6	0.4932	96.47	0.4983
FADM	**106.6**	**0.6598**	151.7	**0.6320**	**86.55**	**0.6366**

Cross-Identity Reenactment. We validate the effectiveness of FADM on the testing datasets of VoxCeleb, VoxCeleb2, and CelebA for the cross-identity reenactment task, where Face vid2vid [49] is used to generate the coarse animation results. Specifically, we randomly select 10 source images and 14 driving videos with different identities from the testing datasets of VoxCeleb and VoxCeleb2 to form various groups. We also use the driving videos of VoxCeleb to animate 10 randomly selected source images from the testing dataset of CelebA. Table 2 shows the quantitative results of the reenactment task. Our FADM outperforms other SOTA methods on the testing datasets of VoxCeleb and CelebA. On VoxCeleb2, FADM exhibits better identity preservation capability, but performs not well in terms of FID. In fact, FID measures the similarity of the data distributions extracted from the two groups. The data quality of VoxCeleb2 is quite poor, including low resolution (the original image size is 224×224) and blurred textures, while FADM tends to generate fine-detailed images, resulting in mismatching between them. Therefore, we think that our FID result on VoxCeleb2 is reasonable.

To prominently show the effectiveness of FADM, we specially select several samples in which the source image and driving videos come from different genders or age groups, and visualize their results in Fig. 5. As we can observe, when the appearance and motion change dramatically, existing SOTA methods may encounter severe distortions or artifacts. In contrast, our FADM can effectively rectify these distortions and enrich the facial details, while ensuring faithful appearance and motion, thereby generating photo-realistic animation results.

Figure 6. Visualization of the ablation study. Without (w/o) the proposed appearance condition, the animation model performs poorly in synthesizing realistic facial areas (eyes, mouth, and hair). We also specially mark the noteworthy areas in the right part, demonstrating the effectiveness of the designed motion condition in modifying the facial details.

Table 3. Ablation study for same-identity reconstruction on VoxCeleb [35].

Method	$\mathcal{L}_1 \downarrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	AKD $\downarrow$	AED $\downarrow$
w/o color condition	0.0428	0.1774	23.835	0.7701	1.488	0.0263
w/o motion condition	0.0407	0.1408	24.110	0.7818	1.400	0.0243
FADM	**0.0402**	**0.1379**	**24.434**	**0.7841**	**1.392**	**0.0241**

4.3. Ablation Study

We conduct comprehensive experiments on the same-identity reconstruction task to demonstrate the effectiveness of the appearance and motion condition mechanism in AGCN, and elaborate why the designed AGCN is the relatively optimal choice for face animation diffusion model over other possible designs.

Appearance Condition. In AGCN, we employ an encoder P_{Conv}, which is optimized by $\mathcal{L}_{color}$, to extract the appearance condition from the driving frames during training, and take the coarse animation results as the appearance motion in the inference process. Here we construct another possible design as a comparison model: using the coarse animation results as appearance condition in both training and inference. As shown in the left part of Fig. 6, directly using the coarse animation results may fail to synthesize realistic facial regions, Moreover, we show the quantitative comparison in Table 3, in which AGCN obtains better results, illustrating the rationality of our appearance condition.

Motion Condition. We further evaluate the effectiveness of the motion condition in AGCN. Specifically, we adopt a comparison model directly using the coarse generated results as the motion condition without the dynamical adjustment according to the 3D reconstruction results in AGCN. The visualization is shown in the right part of Fig. 6. We can observe that with the designed motion condition, FADM can enrich more fine-grained details, while the comparison model are prone to generate over-smoothing results. Table 3 also illustrates that the motion condition in AGCN is effective to improve the quality of generated results.

5. Conclusion

In this paper, we propose an attribute-guided Diffusion Model for face animation (FADM), which introduces the iterative diffusion steps to improve the quality of the animation results. To guarantee the generated facial attributes, including appearance and motion, and meet the requirements of face animation, we design an Attribute-Guided Conditioning Network (AGCN) to extract faithful appearance and motion conditions for the subsequent diffusion process. Based on the coarse generated results and disentangled poses and expressions predicted by the advanced 3D face reconstruction module, AGCN adaptively modulates the appearance and motion conditions and navigates the diffusion denoising model to synthesize ideal facial details and rectify the distortions. Moreover, FADM can be directly used to improve the quality of talking-head videos without extra modulation. Extensive experiments demonstrate that our FADM achieves new state-of-the-art performance.

6. Acknowledgements

This work was supported by National Natural Science Foundation of China under Grant 62076016, Beijing Natural Science Foundation L223024.

References

[1] Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F Cohen. Bringing portraits to life. *TOG*, 2017. 2

[2] Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *ICLR*, 2022. 3

[3] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, 1999. 2

[4] Egor Burkov, Igor Pasechnik, Artur Grigorev, and Victor Lempitsky. Neural head reenactment with latent pose descriptors. In *CVPR*, 2020. 1, 2

[5] Hyungjin Chung and Jong Chul Ye. Score-based diffusion models for accelerated mri. In *MedIA*, 2022. 2, 3

[6] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018. 5, 7

[7] Michail Christos Doukas, Stefanos Zafeiriou, and Viktoriia Sharmanska. Headgan: One-shot neural head synthesis and editing. In *ICCV*, 2021. 1, 2

[8] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *TOG*, 2021. 3, 5

[9] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution. *arXiv preprint arXiv:2303.16491*, 2023. 3

[10] Jiahao Geng, Tianjia Shao, Youyi Zheng, Yanlin Weng, and Kun Zhou. Warp-guided gans for single-photo facial animation. *TOG*, 2018. 2

[11] Partha Ghosh, Pravir Singh Gupta, Roy Uziel, Anurag Ranjan, Michael J Black, and Timo Bolkart. Gif: Generative interpretable faces. In *3DV*, 2020. 2

[12] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *CVPR*, 2021. 2

[13] Sungjoo Ha, Martin Kersner, Beomsu Kim, Seokjun Seo, and Dongyoung Kim. Marionette: Few-shot face reenactment preserving identity of unseen targets. In *AAAI*, 2020. 2

[14] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. *arXiv:2205.11495*, 2022. 2, 3

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 5

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3, 4

[17] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. In *JMLR*, 2022. 2, 3

[18] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022. 2, 3

[19] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *CVPR*, 2022. 6

[20] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *CVPR*, 2020. 5

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2

[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 2

[23] Hyeongwoo Kim, Mohamed Elgharib, Michael Zollhöfer, Hans-Peter Seidel, Thabo Beeler, Christian Richardt, and Christian Theobalt. Neural style-preserving visual dubbing. *TOG*, 2019. 2

[24] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *TOG*, 2018. 2

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[26] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *ICLR*, 2020. 3

[27] Mohammad Rami Koujan, Michail Christos Doukas, Anastasios Roussos, and Stefanos Zafeiriou. Head2head: Video-based neural head synthesis. In *FG*, 2020. 1, 2

[28] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 2022. 2, 3

[29] Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, Peng Liu, and Zhou Zhao. Diffsinger: Diffusion acoustic model for singing voice synthesis. *arXiv:2105.02446*, 2021. 3

[30] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 2018. 5, 7

[31] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022. 2, 3

[32] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *CVPR*, 2021. 2, 3

[33] Zhaoyang Lyu, Zhifeng Kong, Xudong Xu, Liang Pan, and Dahua Lin. A conditional point diffusion-refinement paradigm for 3d point cloud completion. In *ICLR*, 2022. 2, 3

[34] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2

[35] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*, 2017. 5, 6, 7, 8

[36] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *ICCV*, 2021. 2, 6

[37] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *CVPR*, 2021. 5

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3

[39] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *SIGGRAPH*, 2022. 2, 3

[40] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *TPAMI*, 2022. 2, 3

[41] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *CVPR*, 2019. 1, 2

[42] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *NeurIPS*, 2019. 1, 2, 3, 5, 6

[43] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 2

[44] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2020. 3, 5

[45] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. In *ICLR*, 2022. 2, 3

[46] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *CVPR*, 2020. 2

[47] Qiulin Wang, Lu Zhang, and Bo Li. Safa: Structure aware face animation. In *3DV*, 2021. 2

[48] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Bryan Catanzaro, and Jan Kautz. Few-shot video-to-video synthesis. In *NeurIPS*, 2019. 1, 2

[49] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*, 2021. 1, 2, 3, 5, 6, 7

[50] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. In *ICLR*, 2022. 1, 2

[51] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 5

[52] Olivia Wiles, A Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *ECCV*, 2018. 2

[53] Ruihan Yang, Prakhar Srivastava, and Stephan Mandt. Diffusion probabilistic modeling for video generation. *arXiv:2203.09481*, 2022. 2, 3

[54] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *ECCV*, 2020. 1, 2, 6

[55] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *ICCV*, 2019. 1, 2

[56] Bohan Zeng, Boyu Liu, Hong Li, Xuhui Liu, Jianzhuang Liu, Dapeng Chen, Wei Peng, and Baochang Zhang. Fnevr: Neural volume rendering for face animation. In *NeurIPS*, 2022. 2

[57] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv:2208.15001*, 2022. 2, 3

[58] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[59] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *ICCV*, 2021. 2, 3