MIMAFace: Face Animation via Motion-Identity Modulated Appearance Feature Learning

Yue Han [*1], **Junwei Zhu** [*2], **Yuxiang Feng**[1], **Xiaozhong Ji**[2], **Keke He**[2], **Xiangtai Li**[3], **Zhucun Xue**[1], **Yong Liu**[†1]

[1]Zhejiang University [2]Tencent Youtu Lab [3] Nanyang Technological University

12432015@zju.edu.cn, yongliu@iipc.zju.edu.cn

Abstract

Current diffusion-based face animation methods generally adopt a ReferenceNet (a copy of U-Net) and a large amount of curated self-acquired data to learn appearance features, as robust appearance features are vital for ensuring temporal stability. However, when trained on public datasets, the results often exhibit a noticeable performance gap in both image quality and temporal consistency. To address this issue, we meticulously examine the essential appearance features in the facial animation tasks, which include motion-agnostic (e.g., clothing, background) and motion-related (e.g., facial details) texture components, along with high-level discriminative identity features. Drawing from this analysis, we introduce a Motion-Identity Modulated Appearance Learning Module (MIA) that modulates CLIP features at both motion and identity levels. Additionally, to tackle the semantic/ color discontinuities between clips, we design an Inter-clip Affinity Learning Module (ICA) to model temporal relationships across clips. Our method achieves precise facial motion control (i.e., expressions and gaze), faithful identity preservation, and generates animation videos that maintain both intra/inter-clip temporal consistency. Moreover, it easily adapts to various modalities of driving sources. Extensive experiments demonstrate the superiority of our method.

Introduction

Face Animation aims to generate a realistic talking head by animating a source face using the motion information of another person, *i.e.*, pose, expression, and gaze (Pei et al. 2024). It has diverse applications in virtual character creation for game production and video editing. Previous GAN-based methods (Siarohin et al. 2019a; Zhang et al. 2020b; Xu et al. 2022a; Nirkin, Keller, and Hassner 2019; Siarohin et al. 2019b; Ren et al. 2021; Yang et al. 2022; Yin et al. 2022; Bounareli et al. 2023; Zhang et al. 2023a; Hong et al. 2022; Tao et al. 2022; Hong and Xu 2023) most delivers results at resolution 256^2 and only support median pose variation, *i.e.*, less than 30 degrees), due to the lack of high-resolution and large pose dataset. Recent attempts leverage the powerful generation capability of pre-trained latent diffusion models to address these challenges. However, the high variance of noise in diffusion presents a new challenge in generating smooth videos.

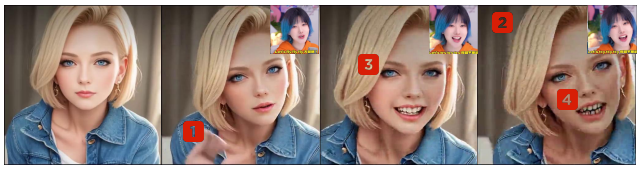

Figure 1: Typical failure cases for current diffusion-based face animation methods: **(1)/(2)** semantic/ color discontinuity across clips, **(3)** stiff expression, **(4)** quality degradation

The human-body animation methods (Hu et al. 2023; Xu et al. 2023b) identify that CLIP (Radford et al. 2021) fails to provide adequate appearance features, resulting in video flickering. To address this issue, ReferenceNet (a UNet copy) is proposed to supply multi-scale similar appearance features. Subsequent works in face animation adopt similar frameworks, integrating ReferenceNet with motion modules proposed in AnimateDiff (Guo et al. 2023) to ensure temporal stability and achieve remarkably favorable results. However, the quality of the results within this paradigm (Hu et al. 2023; Xu et al. 2023b; Tian et al. 2024; Chang et al. 2023; Wei, Yang, and Wang 2024) heavily depends on the quality and volume of the training data. When trained on publicly available datasets (Chang et al. 2023; Wei, Yang, and Wang 2024), the generated results often exhibit a noticeable gap in quality, as shown in fig. 1 (from released examples in AniPortrait). A question arises: *what appearance features are essential for producing a temporally stable, high-quality face animation video?* We argue that the necessary appearance features include motion-independent texture features (*i.e.*, clothing, background), motion-related texture features (*i.e.*, facial details), and high-level discriminative features (*i.e.*, identity). Based on this analysis, we propose the Motion-Identity Modulated Appearance Learning Module (MIA), which modulates CLIP features at both motion and identity levels. For motion modulation, we use 3DMM coefficients to modulate appearance features via cross-attention. This facilitates the generation of subtle facial textures, *e.g.*, wrinkles and muscle contractions, resulting from expressions. For identity modulation, we introduce an identity contrastive loss to compensate for high-level discriminative fea-

[*]These authors contributed equally.

[†]Corresponding author.

tures lost. This is because we find that the optimization objective, i.e., the denoise loss, encourages CLIP to focus more on the learning of low-level generative features while neglecting the learning of discriminative features. This necessitates that the model learns to preserve identity through training on data that encompasses a wider range of identities. The joint modulation of motion and identity makes appearance features learning easier, allowing us to achieve high-fidelity and temporally stable video results even when trained solely on public datasets.

Although adequate appearance features ensure intra-clip temporal consistency with temporal attention, the lack of modeling inter-clip relationships causes semantic/ color discontinuities between clips. Current solutions include fixing initial noise, frame interpolation, or temporal co-denoising, but noticeable discontinuities still exist. To address this issue, we introduce the Inter-clip Affinity Learning Module (ICA), which conditions preceding frames with augmentation strategies to bridge the gap between ground truth frames during training and generated frames during inference.

Our contributions can be summarized as follows:

- We identify potential issues in current diffuison-based face animation methods and carefully examine the essential appearance features. Based on the analysis, we introduce a Motion-Identity Modulated Appearance Learning Module (MIA) that modulates CLIP features at both motion and identity levels, making the appearance features learning more effective.
- We introduce the Inter-clip Affinity Learning Module (ICA) which models temporal relationships between clips to address the issue of semantic/color discontinuities.
- Our method achieves precise facial motion control (i.e. expressions and gaze), faithful identity preservation, and generates animation videos that maintain both intra/inter-clip temporal consistency. Moreover, it easily adapts to various modalities of driving sources. Extensive experiments demonstrate the superiority of our method.

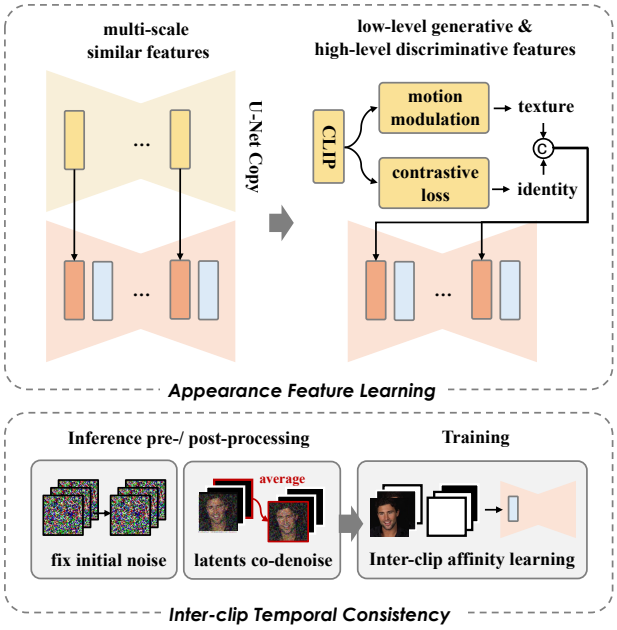

Figure 2: Comparison between our method and previous diffusion-based face animation methods in terms of appearance feature learning and inter-clip temporal consistency.

Related Works

GAN-based Face Animation

Methods (Siarohin et al. 2019a; Zhang et al. 2020b; Xu et al. 2022a; Nirkin, Keller, and Hassner 2019; Agarwal et al. 2023; Xu et al. 2022b; Yang et al. 2022; Bounareli et al. 2023; Zhang et al. 2023a; Gao et al. 2023; Bounareli et al. 2024) can be broadly divided into warping-based and 3DMM-based methods. *Warping-based methods* (Siarohin et al. 2019a; Zhao and Zhang 2022; Siarohin et al. 2019b; Hong et al. 2022)typically extract landmarks or region pairs to estimate flow fields and perform warping on the source appearance feature maps to transfer motions. Limited by the accuracy of the predicted flow field, these methods tend to produce blurry and distorted results when dealing with large motion variations.

3DMM-based methods use facial reconstruction coefficients or the render image from 3DMM as motion intermediate representation. Due to the inherent decoupling properties of coefficients, 3DMM-based methods, *e.g.*, PIRenderer (Ren et al. 2021), can freely control expressions and poses. Although 3DMM provides accurate structural references for facial regions, it lacks references for hair, teeth, and eye movement. Additionally, its coarse facial textures result in suboptimal generated outcomes. StyleHEAT (Yin et al. 2022) tackles this challenge by leveraging the powerful generation capabilities of StyleGAN2 to produce detailed textures and high-resolution frontal portraits. However, its generalization and large pose generation abilities are limited due to the constraint of the training dataset.

Diffusion-based Face Animation

To improve sample quality and generalization capability, diffusion models have gained popularity (Zeng et al. 2023; Peng et al. 2023; Hu et al. 2023; Xu et al. 2023b; Han et al. 2023). FADM (Zeng et al. 2023) combines the previous reenactment models with diffusion refinements, but the base model limits the driving accuracy. The human-body animation methods (Hu et al. 2023; Xu et al. 2023b) identify that CLIP fails to provide adequate appearance features, resulting in video flickering. To address this issue, ReferenceNet (a UNet copy) is proposed to supply multi-scale similar appearance features. Subsequent works in face animation adopt similar frameworks, integrating ReferenceNet with motion modules proposed in AnimateDiff (Guo et al. 2023) to ensure temporal stability and achieve remarkably favorable results. However, the quality of the results within this paradigm (Hu et al. 2023; Xu et al. 2023b; Tian et al. 2024; Chang et al. 2023; Wei, Yang, and Wang 2024) heav-

ily depends on the quality and volume of the training data. When trained on publicly available datasets (Chang et al. 2023; Wei, Yang, and Wang 2024), the generated results often exhibit a noticeable gap in quality. To address this issue, in this paper, we explore the effective way to learn robust appearance features.

Audio-Driven Face Animation

Previous approaches (Prajwal et al. 2020; Zhou et al. 2020; Zhang et al. 2020a, 2021; Fan et al. 2022; Xing et al. 2023; Xu et al. 2023a; Garcia and Yousef 2023; Shen et al. 2022; Huang et al. 2023; Du et al. 2023; Tan, Ji, and Pan 2024; Ye et al. 2024; Wang et al. 2024) focus on learning models specific to individual speakers. Sadtalker (Zhang et al. 2023b) uses 3DMM as an intermediate representation for subject-agnostic reenactment. Via learning the 3D motion coefficients of the 3DMM model from audio, Sadtalker showcases robust generalization capabilities. However, it still struggles to accommodate significant motion variations and tends to produce blurry results. Recently, (Tian et al. 2024; Wei, Yang, and Wang 2024) have follow the 'ReferenceNet with motion module' paradigm and achieve significant advancements. (Tian et al. 2024) bypasses the need for intermediate 3D models or facial landmarks and achieve astonishing results. However, it heavily relies on the dataset, making it difficult to reproduce. (Wei, Yang, and Wang 2024) learns to map the audio to a 3D facial mesh and head pose, then projects these two elements into 2D keypoints as the intermediate motion representation. However, it also exhibits typical failure cases similar to methods using the same paradigm. In this paper, our primary goal is to explore how to effectively learn robust appearance features, not to focus on the learning of motion intermediate representations. Following the approach of (Yin et al. 2022; Wei, Yang, and Wang 2024), which can be easily adapted to multi-modal inputs, we support audio-driven face animation by mapping audio to the 3DMM coefficient space using pretrained audio-to-3DMM-coefficients encoders, *e.g.*, (Zhang et al. 2023b).

Temporal Consistency in Face Animation.

Current methods (Xu et al. 2023b; Chang et al. 2023; Tian et al. 2024; Wei, Yang, and Wang 2024) tend to focus more on modeling the temporal stability within clips while neglecting the preservation of temporal consistency between clips. Recent image animation (Zhang et al. 2023c) methods (typically generating only one clip) demonstrate that temporal attention is sufficient for ensuring intra-clip temporal stability. Therefore, we believe that the temporal stability within clips is mainly affected by the lack of adequate appearance features. Current solutions for inter-clip temporal stability include fixing initial noise, frame interpolation, or temporal co-denoising (Xu et al. 2023b), but noticeable discontinuities still persist. Unlike these pre/post-processing approaches during inference, we explicitly model the temporal relationships between clips during training.

Methods

Face Animation aims to create a lifelike talking head video by animating a source face $\boldsymbol{I}_S$ using motion information from another person. This motion information may include pose, expression, and gaze, which can be obtained from either a video sequence $\boldsymbol{I}_D^{1:N}$ or an audio sequence $\boldsymbol{A}_D^{1:N}$. The entire sequence has a length of $\boldsymbol{N}$ and is divided into several clips with the length of $\boldsymbol{F}$.

The current challenge in generating high-quality and smooth animation videos using diffusion-based methods can be summarized as the need for robust appearance features and ensuring temporal consistency within and between clips: ***1)*** Recent methods commonly use ReferenceNet for extracting appearance features, but this approach relies on learning from large amounts of curated data. We believe this is due to the lack of guidance on effective optimization directions. Thus, we propose MIA that modulates appearance features at both motion and identity levels to address this issue. The motion modulation encourages the model to learn facial texture details related to motion, while the identity modulation focuses on high-level discriminative features beyond low-level texture information guided by denoise loss. ***2)*** To ensure temporal consistency, the motion module is shown to maintain intra-clip temporal consistency effectively. However, the challenge of ensuring inter-clip consistency remains largely unexplored. Existing methods (Xu et al. 2023b; Wei, Yang, and Wang 2024) like frame interpolation, fixed initial noise, and temporal co-denoising have proven insufficient in resolving flickering. In this work, we propose ICA to address this issue by modeling relationships with preceding frames during training. The pipeline of the proposed MIMAFace is illustrated in fig. 3

Motion Intermediate Representation

We utilize the motion coefficients of 3DMM, *i.e.*, pose $\boldsymbol{\rho}$, expression $\boldsymbol{\beta}$, and the rendered image $\boldsymbol{I}_R$, as a unified motion intermediate representation (for details on 3DMM, please refer to the supplementary materials). The rendered image $\boldsymbol{I}_R$ is conditioned by concatenating with the noisy latent $\boldsymbol{z}_t$. The coefficients are used to modulate the appearance features to supplement more motion-related facial details. This approach offers several advantages: ***1)*** Highly disentangled coefficient space allows for convenient integration of multi-modal driving sources, mapping images/ audio to 3D coefficients with pre-trained models; ***2)*** Continuous coefficient space facilitates smooth video generation, in contrast to the spatial jittering characteristics of landmarks; ***3)*** The rendered image $\boldsymbol{I}_R$, similar to landmarks, provides coarse spatial guidance, while the coefficient encodes more information, compensating for the missing details about subtle facial expressions. Besides 3DMM, we employ an off-the-shelf gaze detector to extract gaze embedding $\boldsymbol{g}$.

Motion-Identity Modulated Appearance Learning Module (MIA)

The robust appearance features should include motion-independent texture features (*i.e.*, clothing, background), motion-related texture features (*i.e.*, facial details), and high-level discriminative features (*i.e.*, identity). However, under the guidance of the denoise loss, the appearance encoder tends to learn common textures more easily. This leads

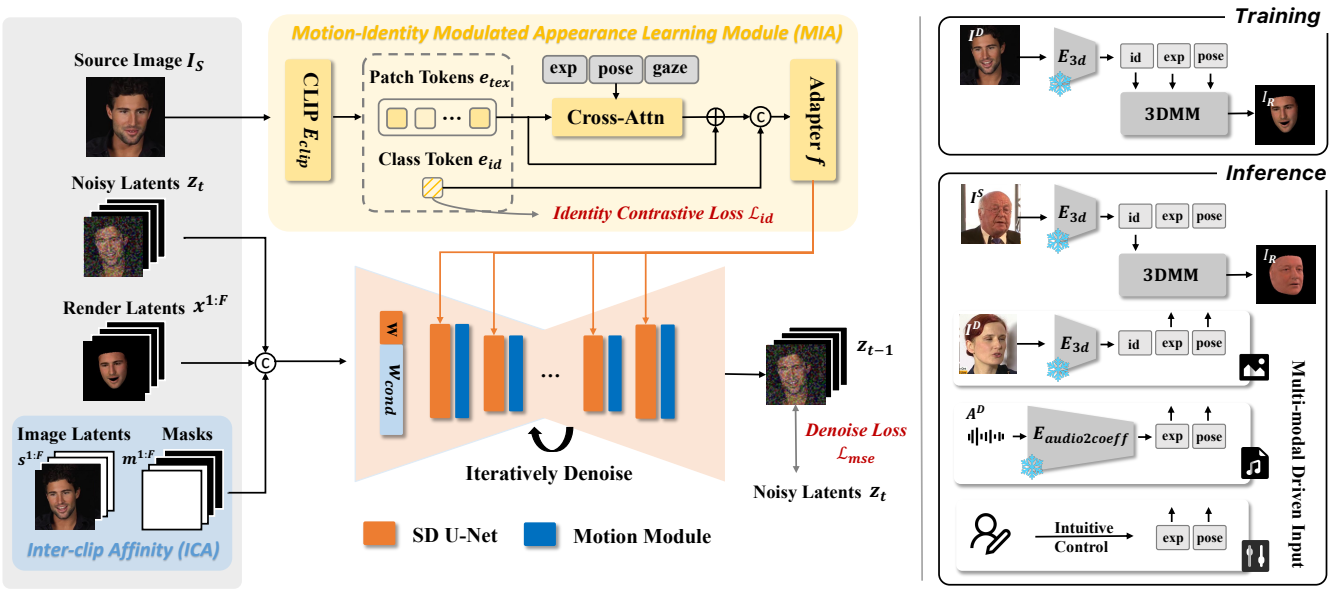

Figure 3: **Pipeline of the proposed MIMAFace**, which consists of: ***1)*** **Motion-Identity Modulated Appearance Learning Module (MIA)** and ***2)*** **Inter-clip Affinity Learning Module (ICA)**. MIA modulates the appearance features at both motion and identity levels. The source image $\boldsymbol{I}_S$ is passed to CLIP $\boldsymbol{E}_{clip}$ to obtain patch tokens $\boldsymbol{e}_{tex}$ and a class token $\boldsymbol{e}_{id}$, which capture the texture and identity, respectively. $\boldsymbol{e}_{tex}$ are then modulated with motion coefficients $\boldsymbol{\rho}, \boldsymbol{\beta}, \boldsymbol{g}$ via cross attention. $\boldsymbol{e}_{id}$ is used to calculate the identity contrastive loss $\mathcal{L}_{id}$. The modulated $\boldsymbol{e}_{tex}$ are concatenated with $\boldsymbol{e}_{id}$ to form the conditioning appearance features. ICA ensures inter-clip temporal consistency by conditioning image latent $\boldsymbol{s}^{1:F}$ (of ground truth during training and denoised ones during inference) and indicating masks $\boldsymbol{m}^{1:F}$ with the added condition module $\boldsymbol{W}_{cond}$. Additionally, we employ 3DMM coefficients $\boldsymbol{\rho}, \boldsymbol{\beta}$ and rendered images $\boldsymbol{I}_R$ as intermediate representations for motion. The 3DMM coefficients can adapt to various modalities of driving inputs, *i.e.*, images, audio, and manual modifications.

to two issues: 1) Although the appearance interacts with the motion intermediate representation through SD cross-attention, it is difficult for landmarks/render to express subtle facial movements. 2) The optimization target for generation encourages the model to prioritize the learning of low-level features, neglecting high-level discriminative features. This can lead to slow convergence of the model. As a result, to improve the effectiveness and robustness of appearance features, we introduce modulation at both the motion and identity levels. Specifically, in this paper, we still choose CLIP as the appearance encoder to validate our hypothesis. The source image is passed through the CLIP vision encoder $\boldsymbol{E}_{clip}$, yielding patch tokens $\boldsymbol{e}_{tex}$ to capture low-level textures, and a class token $\boldsymbol{e}_{id}$ to represent high-level identity.

Motion Modulation To enable the model to generate more vivid and subtle facial movements, we introduce additional 3DMM motion parameters $\boldsymbol{\rho}$, $\boldsymbol{\beta}$ and gaze embedding $\boldsymbol{g}$ to modulate the texture tokens $\boldsymbol{e}_{tex}$ via simple cross-attention with a residual connection to prevent information loss:

$$\begin{aligned} \boldsymbol{e}'_{tex} &= \boldsymbol{e}_{tex} + \text{Cross-Attn}\left(\boldsymbol{e}_{tex}, [\boldsymbol{\rho}; \boldsymbol{\beta}; \boldsymbol{g}]\right), \\ \text{Cross-Attn}\left(\boldsymbol{e}_{tex}, [\boldsymbol{\rho}; \boldsymbol{\beta}; \boldsymbol{g}]\right) &= \text{softmax}\left(\frac{\boldsymbol{e}_{tex} \cdot [\boldsymbol{\rho}; \boldsymbol{\beta}; \boldsymbol{g}]^T}{\sqrt{d_k}}\right)[\boldsymbol{\rho}; \boldsymbol{\beta}; \boldsymbol{g}], \end{aligned} \tag{1}$$

where d_k is the dimensionality of the key. This process can be understood as adding more motion information to refine the coarse spatial structure determined by the render. It can also be seen as aligning the facial appearance features to the same motion as the render, making the cross-attention feature interaction within SD more effective.

Identity Modulation To prevent the model from neglecting the learning of high-level features, we introduce a discriminative objective. For faces, identity is the most intuitive high-level feature. Therefore, we incorporate an *Identity Contrastive Loss* $\mathcal{L}_{\text{id}}$ during training, as illustrated in fig. 4. Specifically, we augment the source image $\boldsymbol{I}_S$ through photometric transformation, altering the RGB channels by shifting pixel colors to new values. This includes techniques such as grayscaling, color jittering, various filtering methods (like edge enhancement, blurring, and sharpening), lighting perturbation, noise addition, vignetting, and contrast adjustment. Subsequently, the source image $\boldsymbol{I}_S$ and the augmented image $\boldsymbol{I}_{aug}$ are each processed through $\boldsymbol{E}_{clip}$ to obtain a positive ID token pair. During the training process, we maintain an ID token memory bank $\mathcal{M} = \{\boldsymbol{e}_{id}\}$ to store ID tokens. Consequently, an ID token belonging to the same person but with a different structure will produce another type of positive ID token pair. Explicitly combining the same identity samples across pixel and structural variations enhances the model's generalization capability and robustness.

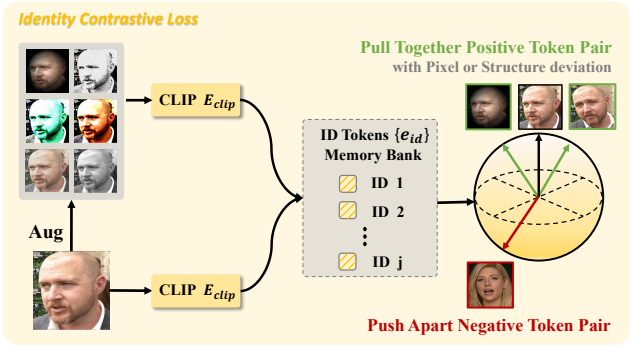

Figure 4: **Illustration of our Identity Contrastive Loss.** We apply photometric data augmentation on the source image and maintain an ID token memory bank to store ID tokens. By pulling together positive token pairs with variances in pixels or structure and pushing apart the negative token pairs, the loss encourages the appearance encoder to capture high-level discriminative features.

The Identity Contrastive Loss $\mathcal{L}_{id}$ is given by:

$$\mathcal{L}_{id} = -\log\left(\frac{\exp\left(\text{sim}\left(z_i, z_i^+\right)\right)}{\exp\left(\text{sim}\left(z_i, z_i^+\right)\right) + \sum_{j=1}^{N}\exp\left(\text{sim}\left(z_i, z_j^-\right)\right)}\right), \tag{2}$$

where z_i represents the token of the identity token, z_i^+ and z_j^- represent the token of the same and different identity, respectively. $\text{sim}\left(z_i, z_j\right)$ denotes the cosine similarity between two tokens. Finally, the modulated patch tokens $\boldsymbol{e}'_{tex}$ and the class token $\boldsymbol{e}_{id}$ are concatenated and go through an adapter $\boldsymbol{f}$ constructed by a 3-layer transformer to obtain the final appearance condition.

Inter-clip Affinity learning Module (ICA)

Although temporal consistency within video clips is learned and preserved through temporal attention layers, variations in lighting, color, and semantics may still occur between different clips. Previous methods (Xu et al. 2023b; Wei, Yang, and Wang 2024) commonly apply noise pre-processing or denoised latent post-processing during inference without explicitly learning temporal correspondence.

Our goal is to enable the model to smoothly generate the next video clip by referring to the last few frames in the previous clip. This problem can be reformulated as follows: When the model is conditioned on the preceding frames, we aim for it to reconstruct these frames, thereby maintaining temporal consistency with the preceding frames. This is because the temporal consistency of subsequent frames within the video clip can be ensured by temporal attention. During training, given a video clip of length F, the initial k frames are ground truth image latent $\boldsymbol{s}_{gt}^{1:k}$, and the remaining frames $\boldsymbol{s}^{k+1:F}$ are padded with zeros. Meanwhile, one-channel masks $\boldsymbol{m}^{1:F}$ are used to indicate whether the model should reconstruct the given image latent correspondingly.

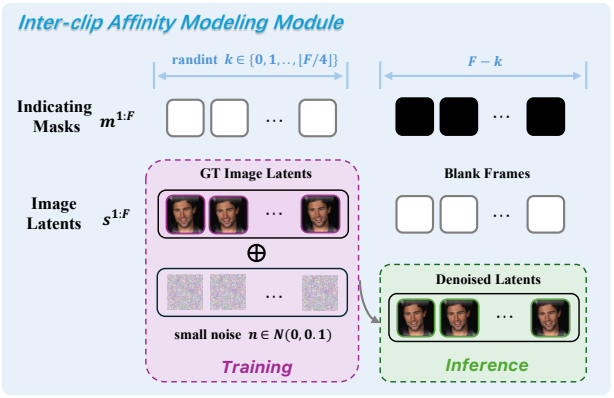

Figure 5: **Illustration of our Inter-clip Affinity Learning Module.** The model learns inter-clip temporal consistency by conditioning the image latent of the preceding frames and using masks to indicate whether reconstruction is required.

$$\boldsymbol{s}^{1:k} = \boldsymbol{s}_{gt}^{1:k}, \quad \boldsymbol{s}^{k+1:F} = 0, \quad \boldsymbol{m}^{1:k} = 1, \quad \boldsymbol{m}^{k+1:F} = 0, \\ where \quad k \sim Uniform\{0, 1, \ldots, \lfloor F/4 \rfloor\}, \tag{3}$$

This approach differs from using denoised latent codes $\boldsymbol{s}_d^{1:F}$ during inference, which we find leads to a decrease in video quality. To bridge this domain gap, we introduce a small amount of Gaussian noise to the ground truth image latent during training, simulating the denoised latent:

$$\boldsymbol{s}_{gt,\,\text{noise}}^{1:k} = \boldsymbol{s}_{gt}^{1:k} + \mathcal{N}\left(0, \sigma^2\right), \tag{4}$$

Here, $\mathcal{N}\left(0, \sigma^2\right)$ represents Gaussian noise with mean 0 and standard deviation σ (set to 0.1 here). Inspired by inter-frame affinity in PIA, we condition the image latent $\boldsymbol{s}^{1:F}$ in the input of the U-Net using a lightweight learnable single-layer convolution $\boldsymbol{W}_{cond}$ as the condition module, without compromising the original functionality, as shown in fig. 5.

Experiment

Datasets. We train our model on VoxCeleb2 (Nagrani, Chung, and Zisserman 2017). We evaluate the VoxCeleb1 test set following the sampling strategy of the PIRenderer (Ren et al. 2021). We test on FFHQ (Karras, Laine, and Aila 2019) to verify the generalization ability.

Metrics. We use PSNR and LPIPS (Zhang et al. 2018) to evaluate the reconstruction quality for same-identity reenactment. Expression, pose, and gaze accuracy are assessed by calculating the average Euclidean distance of the corresponding embedding between the generated and driving faces. These three embeddings are derived through the respective estimator. Identity preservation cosine similarity (CSIM) is calculated by (Huang et al. 2020). FID is used to evaluate the realism of the generated faces.

Training Details. Our training process is divided into two stages. In the first stage, we train the image-driven model. We begin training from the StableDiffusion v1-5 model and OpenAI clip-vit-large-patch14 vision model. Our models are

Figure 6: **Same-identity and cross-identity reenactment results on Voxceleb1 test set.**

trained for 30k steps on 4 NVIDIA A100 GPUs, with a constant learning rate of 1e-5 and a batch size of 32. To facilitate classifier-free guidance sampling, we train the model without appearance condition on 10 of the instances. In the second stage, we train the video-driven model. We freeze the SD U-Net and the appearance encoder, and train the condition module $\boldsymbol{W}_{cond}$, along with the motion module. Our models are trained for 30k steps on 4 NVIDIA A100 GPUs, with a constant learning rate of 1e-5, a batch size of 8 and clip sequence 12.

Comparison with State-of-the-Art Methods

Methods. For image-driven reenactment, we compare our method with GAN-based methods, including FOMM (Siarohin et al. 2019b), Face-vid2vid (Wang, Mallya, and Liu 2021), PIRenderer (Ren et al. 2021), TPSM (Zhao and Zhang 2022), DAM (Tao et al. 2022) and diffusion-based methods, including FADM (Zeng et al. 2023) and AniPortrait (Wei, Yang, and Wang 2024). Except AniPortrait, all of these models are trained on VoxCeleb1. For video-driven reenactment, we compare with MCNet (Hong and Xu 2023).

Qualitative Results. In fig. 6, previous GAN-based methods tend to produce blurry results with noticeable artifacts. While FADM refines the results of FOMM or Face-vid2vid using diffusion, enhancing the generation quality, this approach inherently inherits the motion deviations from these methods. In contrast, our method dlivers results of both high fidelity and precise control simultaneously. Compared to previous outcomes that appear relatively smooth, it is worth noting the nuanced skin wrinkles and light-shadow variations caused by expressions in our method, which make the generated results appear much more lifelike.

Quantitative Results. In table 1, according to FADM, the data quality of VoxCeleb is relatively subpar, characterized by low resolution and blurred textures. Diffusion-based methods tend to generate fine-detailed images, leading to a mismatch between them. Consequently, we face noticeable disadvantages in pixel-level metrics (PSNR and LPIPS), and distribution-level metrics (FID). However, in terms of semantic-level metrics, including identity similarity (CSIM) and motion accuracy (expression, pose, and gaze), we exhibit clear advantages.

Handling Unseen Identities. In fig. 8, we evaluste on FFHQ to verify the generalization capability of our approach. Whereas previous methods yield blurry results, our approach consistently delivers high-fidelity generation.

Video-driven Reenactment

MCNet is a method designed specifically for videos and fail when driven by a single image or a short video. Thus, we compare our results exclusively with theirs in the context of video generation in fig. 9. MCNet relies on a memory mechanism to maintain clarity in appearance under continuous large motion changes in videos. However, if the drive image and source image are initially misaligned, the errors in pose and expression will gradually accumulate, leading to a decline in generated quality and the appearance of noticeable

Table 1: **Quantitative evaluations among current popular methods on Voxceleb1 test set.**

Methods	Same-Identity							Cross-Identity				
	PSNR↑	LPIPS↓	Exp↓	Pose↓	Gaze↓	CSIM↑	FID↓	Exp↓	Pose↓	Gaze↓	CSIM↑	FID↓
FOMM (Siarohin et al. 2019b)	22.38	0.1405	2.77	0.0261	0.0554	0.8328	26.69	6.28	0.0638	0.0959	0.5642	42.76
PIRenderer (Ren et al. 2021)	21.00	0.1468	2.94	0.0496	0.0800	0.7997	25.48	5.84	0.0752	0.0977	0.5659	**35.99**
Face-vid2vid (Wang, Mallya, and Liu 2021)	22.63	**0.1243**	2.84	0.0283	0.0850	0.8383	**25.36**	6.68	0.0847	0.1220	0.6328	39.87
TPSM (Zhao and Zhang 2022)	23.24	0.1442	2.58	0.0224	0.0538	0.8277	33.63	6.10	0.0535	0.0900	0.5836	50.43
DAM (Tao et al. 2022)	**23.37**	0.1550	2.81	0.0263	0.0628	0.8333	36.40	6.31	0.0626	0.0967	0.5534	54.13
FADM (Zeng et al. 2023)	22.36	0.1425	2.95	0.0303	0.0879	0.8352	31.70	6.71	0.0821	0.1242	0.6522	42.22
Ours	18.64	0.1907	**2.21**	**0.0195**	**0.0482**	**0.8412**	60.91	**5.03**	**0.0503**	**0.0614**	**0.6778**	64.49

Figure 7: **Ablation study of different components on the VoxCeleb test set.**

Figure 8: **Evaluation on FFHQ.** Our method exhibits outstanding generalization capabilities. Even when dealing with unseen identities, it maintains consistent identity preservation. Moreover, it ensures intricate facial textures and precise expressions.

artifacts. In contrast, our approach can produce more accurate motion.

Ablation Studies

We perform ablation studies in fig. 7 to verify the effectiveness of the Motion-Identity Modulated Appearance Learning Module (MIA).

Motion Modulation. If we remove the render image and rely solely on 3DMM coefficients for motion control, it results in a dramatic decrease in the precision of the generated pose and expression. because the model loses a direct reference to the spatial position, making the learning process more challenging. If we remove the 3DMM coefficients and rely solely on the render image for motion control, it results in a moderate decrease in the precision of the generated pose and expression.

Identity Modulation. Without incorporating the identity contrastive loss, the model, during finetuning, is inclined by the denoise loss to capture low-level features, leading to a decline in identity discrimination capability.

Figure 9: **Video-driven Reenactment Results**.

Conclusion

This work proposes a novel MIMAFace to address the limitations of current diffusion-based face animation methods. We introduce a Motion-Identity Modulated Appearance Learning Module (MIA) that modulates CLIP features at both motion and identity levels, and an Inter-clip Affinity Learning Module (ICA) to model temporal relationships across clips. Our method ensures precise facial motion control and faithful identity preservation and generates animation videos with intra/inter-clip temporal consistency. Extensive experiments demonstrate the effectiveness of the proposed method.

Limitation and Future Works. Integrating large language models and face-specific SD can enhance performance and application value. However, since our approach can generate realistic images that could be used for facial forgery, regulatory constraints are necessary to mitigate this risk.

References

Agarwal, M.; Mukhopadhyay, R.; Namboodiri, V. P.; and Jawahar, C. 2023. Audio-visual face reenactment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5178–5187.

Bounareli, S.; Tzelepis, C.; Argyriou, V.; Patras, I.; and Tzimiropoulos, G. 2023. HyperReenact: one-shot reenactment via jointly learning to refine and retarget faces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7149–7159.

Bounareli, S.; Tzelepis, C.; Argyriou, V.; Patras, I.; and Tzimiropoulos, G. 2024. One-Shot Neural Face Reenactment via Finding Directions in GAN's Latent Space. *International Journal of Computer Vision*, 1–31.

Chang, D.; Shi, Y.; Gao, Q.; Fu, J.; Xu, H.; Song, G.; Yan, Q.; Yang, X.; and Soleymani, M. 2023. Magicdance: Realistic human dance video generation with motions & facial expressions transfer. *arXiv preprint arXiv:2311.12052*.

Du, C.; Chen, Q.; He, T.; Tan, X.; Chen, X.; Yu, K.; Zhao, S.; and Bian, J. 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.

Fan, Y.; Lin, Z.; Saito, J.; Wang, W.; and Komura, T. 2022. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18770–18780.

Gao, Y.; Zhou, Y.; Wang, J.; Li, X.; Ming, X.; and Lu, Y. 2023. High-fidelity and freely controllable talking head video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5609–5619.

Garcia, M. B.; and Yousef, A. M. F. 2023. Cognitive and affective effects of teachers' annotations and talking heads on asynchronous video lectures in a web development course. *Research and Practice in Technology Enhanced Learning*, 18: 020–020.

Guo, Y.; Yang, C.; Rao, A.; Wang, Y.; Qiao, Y.; Lin, D.; and Dai, B. 2023. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*.

Han, Y.; Zhang, J.; Zhu, J.; Li, X.; Ge, Y.; Li, W.; Wang, C.; Liu, Y.; Liu, X.; and Tai, Y. 2023. A Generalist FaceX via Learning Unified Facial Representation. *arXiv preprint arXiv:2401.00551*.

Hong, F.-T.; and Xu, D. 2023. Implicit Identity Representation Conditioned Memory Compensation Network for Talking Head video Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23062–23072.

Hong, F.-T.; Zhang, L.; Shen, L.; and Xu, D. 2022. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3397–3406.

Hu, L.; Gao, X.; Zhang, P.; Sun, K.; Zhang, B.; and Bo, L. 2023. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*.

Huang, R.; Lai, P.; Qin, Y.; and Li, G. 2023. Parametric implicit face representation for audio-driven facial reenactment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12759–12768.

Huang, Y.; Wang, Y.; Tai, Y.; Liu, X.; Shen, P.; Li, S.; Li, J.; and Huang, F. 2020. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5901–5910.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4401–4410.

Nagrani, A.; Chung, J. S.; and Zisserman, A. 2017. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*.

Nirkin, Y.; Keller, Y.; and Hassner, T. 2019. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7184–7193.

Pei, G.; Zhang, J.; Hu, M.; Zhai, G.; Wang, C.; Zhang, Z.; Yang, J.; Shen, C.; and Tao, D. 2024. Deepfake Generation and Detection: A Benchmark and Survey. *arXiv preprint arXiv:2403.17881*.

Peng, X.; Zhu, J.; Jiang, B.; Tai, Y.; Luo, D.; Zhang, J.; Lin, W.; Jin, T.; Wang, C.; and Ji, R. 2023. PortraitBooth: A Versatile Portrait Model for Fast Identity-preserved Personalization. *arXiv preprint arXiv:2312.06354*.

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

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ren, Y.; Li, G.; Chen, Y.; Li, T. H.; and Liu, S. 2021. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13759–13768.

Shen, S.; Li, W.; Zhu, Z.; Duan, Y.; Zhou, J.; and Lu, J. 2022. Learning dynamic facial radiance fields for few-shot talking head synthesis. In *European Conference on Computer Vision*, 666–682. Springer.

Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2019a. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2377–2386.

Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2019b. First order motion model for image animation. *Advances in neural information processing systems*, 32.

Tan, S.; Ji, B.; and Pan, Y. 2024. FlowVQTalker: High-Quality Emotional Talking Face Generation through Normalizing Flow and Quantization. *arXiv preprint arXiv:2403.06375*.

Tao, J.; Wang, B.; Xu, B.; Ge, T.; Jiang, Y.; Li, W.; and Duan, L. 2022. Structure-aware motion transfer with deformable anchor model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3637–3646.

Tian, L.; Wang, Q.; Zhang, B.; and Bo, L. 2024. EMO: Emote Portrait Alive-Generating Expressive Portrait Videos with Audio2Video Diffusion Model under Weak Conditions. *arXiv preprint arXiv:2402.17485*.

Wang, S.; Ma, Y.; Ding, Y.; Hu, Z.; Fan, C.; Lv, T.; Deng, Z.; and Yu, X. 2024. StyleTalk++: A Unified Framework for Controlling the Speaking Styles of Talking Heads. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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

Wei, H.; Yang, Z.; and Wang, Z. 2024. AniPortrait: Audio-Driven Synthesis of Photorealistic Portrait Animation. *arXiv preprint arXiv:2403.17694*.

Xing, J.; Xia, M.; Zhang, Y.; Cun, X.; Wang, J.; and Wong, T.-T. 2023. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12780–12790.

Xu, C.; Zhang, J.; Han, Y.; Tian, G.; Zeng, X.; Tai, Y.; Wang, Y.; Wang, C.; and Liu, Y. 2022a. Designing one unified framework for high-fidelity face reenactment and swapping. In *European Conference on Computer Vision*, 54–71. Springer.

Xu, C.; Zhang, J.; Hua, M.; He, Q.; Yi, Z.; and Liu, Y. 2022b. Region-aware face swapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7632–7641.

Xu, C.; Zhu, S.; Zhu, J.; Huang, T.; Zhang, J.; Tai, Y.; and Liu, Y. 2023a. Multimodal-driven talking face generation via a unified diffusion-based generator. *CoRR (2023)*, 1–14.

Xu, Z.; Zhang, J.; Liew, J. H.; Yan, H.; Liu, J.-W.; Zhang, C.; Feng, J.; and Shou, M. Z. 2023b. Magicanimate: Temporally consistent human image animation using diffusion model. *arXiv preprint arXiv:2311.16498*.

Yang, K.; Chen, K.; Guo, D.; Zhang, S.-H.; Guo, Y.-C.; and Zhang, W. 2022. Face2Face ρ: Real-Time High-Resolution One-Shot Face Reenactment. In *European conference on computer vision*, 55–71. Springer.

Ye, Z.; Zhong, T.; Ren, Y.; Yang, J.; Li, W.; Huang, J.; Jiang, Z.; He, J.; Huang, R.; Liu, J.; Zhang, C.; Yin, X.; MA, Z.; and Zhao, Z. 2024. Real3D-Portrait: One-shot Realistic 3D Talking Portrait Synthesis. In *The Twelfth International Conference on Learning Representations*.

Yin, F.; Zhang, Y.; Cun, X.; Cao, M.; Fan, Y.; Wang, X.; Bai, Q.; Wu, B.; Wang, J.; and Yang, Y. 2022. Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In *European conference on computer vision*, 85–101. Springer.

Zeng, B.; Liu, X.; Gao, S.; Liu, B.; Li, H.; Liu, J.; and Zhang, B. 2023. Face Animation with an Attribute-Guided Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 628–637.

Zhang, B.; Qi, C.; Zhang, P.; Zhang, B.; Wu, H.; Chen, D.; Chen, Q.; Wang, Y.; and Wen, F. 2023a. Metaportrait: Identity-preserving talking head generation with fast personalized adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22096–22105.

Zhang, J.; Liu, L.; Xue, Z.; and Liu, Y. 2020a. Apb2face: Audio-guided face reenactment with auxiliary pose and blink signals. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4402–4406. IEEE.

Zhang, J.; Zeng, X.; Wang, M.; Pan, Y.; Liu, L.; Liu, Y.; Ding, Y.; and Fan, C. 2020b. Freenet: Multi-identity face reenactment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5326–5335.

Zhang, J.; Zeng, X.; Xu, C.; and Liu, Y. 2021. Real-time audio-guided multi-face reenactment. *IEEE Signal Processing Letters*, 29: 1–5.

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

Zhang, W.; Cun, X.; Wang, X.; Zhang, Y.; Shen, X.; Guo, Y.; Shan, Y.; and Wang, F. 2023b. 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.

Zhang, Y.; Xing, Z.; Zeng, Y.; Fang, Y.; and Chen, K. 2023c. Pia: Your personalized image animator via plug-and-play modules in text-to-image models. *arXiv preprint arXiv:2312.13964*.

Zhao, J.; and Zhang, H. 2022. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3657–3666.

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