ReMask-Animate: Refined Character Image Animation Using Mask-Guided Adapters

Xunzhi Xiang[1,2*], Haiwei Xue[1,2,3*], Zonghong Dai[2*], Di Wang[1], Minglei Li[2], Ye Yue[1], Fei Ma[1†], Weijiang Yu[4†], Heng Chang[3†], Fei Richard Yu[5,6]

[1]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen, China
[2]01AI, Beijing, China
[3]Tsinghua University, Shenzhen, Guangdong, China
[4]Sun Yat-sen University, Guangzhou, Guangdong, China
[5]Shenzhen University, Shenzhen, Guangdong, China
[6]Carleton University, Canada
mafei@gml.ac.cn, {xiangxunzhi999, weijiangyu8}@gmail.com, changh17@tsinghua.org.cn

Abstract

Pose-controlled human video generation is of significant interest and finds extensive applications in areas such as automated advertising and content creation on social media platforms. While existing methods employing pose sequences and reference images for human image animation have exhibited notable performance, they tend to encounter issues such as specific region blurring, background sharpening, and decreased identity consistency. In this paper, we introduce ReMask-Animate, which utilizes masks as additional priors to guide the model's local visual attention to specific areas, thereby alleviating feature confusion between different regions of the image. Three distinct mask-guided adapters are designed for cross-condition regional fusion of hand and face pose features, mitigating feature confusion between the foreground and background, and enhancing the visual consistency of character identity. Moreover, these lightweight adapters introduce minimal computational overhead and can be seamlessly integrated into specific layers of the backbone architecture. Extensive experiments show that our method outperforms state-of-the-art methods on five metrics in public datasets. Additionally, qualitative evaluations highlight a significant improvement in the quality of generated videos, demonstrating our approach's superiority.

Introduction

Human image animation involves transforming a static character image into a dynamic video sequence, driven by a different conditioning pose sequence, while preserving the character's fidelity (Chan et al. 2019; Wang et al. 2021; Zhang et al. 2022; Milis et al. 2023). This technique is crucial in the realm of digital humans, where creating realistic animated representations is essential. Consequently, it has gained significant attention and become a key area of research. Methods for generating human motion videos driven by pose sequences can be broadly categorized into two approaches: GAN-based and diffusion-based frameworks.

Figure 1: Zero-shot dance video generation with ReMask-Animate. Our method is trained solely on the TikTok dataset, achieving high-quality, temporally consistent videos and demonstrating remarkable cross-ID transfer capability.

GAN-based methods (Siarohin et al. 2019b) typically use a warping function to adjust the reference image to align with the target pose. Subsequently, GAN models (Jang et al. 2024) are employed to reconstruct and generate any missing or occluded body parts. However, due to their dependence on image-warping modules, the generation of these occluded parts in the reference images often lacks stability. Addition-

*These authors contributed equally.
†Corresponding author.

ally, GAN-based methods (Zhou et al. 2019) struggle to generalize to images that differ from the training data, exhibiting poor diversity and transferability. Diffusion models (Dhariwal and Nichol 2021; Rombach et al. 2022), which learn to denoise data by reversing a gradual noising process, have emerged as powerful visual priors for various downstream tasks. By leveraging their remarkable generative capabilities under different conditions, related works such as MagicAnimate (Xu et al. 2023), and MagicPose (Chang et al. 2023) can enhance the controlled and efficient generation of human animation videos. These methods explore the extension of various types of pose conditions, such as dense poses that capture the overall body shape and sparse poses that specify the positions of key body joints. However, existing methods rely on a single modality of pose conditions, either dense or sparse, missing the opportunity to combine them for more comprehensive and precise control over the generated animation. Integrating dense poses for overall body shape guidance with sparse poses for fine-grained control of specific body parts can potentially lead to more realistic and customizable human animation video generation. Moreover, while these methods can generate visually plausible animations, they are still limited by issues such as misalignment or blurriness in body parts, poor quality of generated complex backgrounds, and inconsistent character identity.

In this work, we propose ReMask-Animate, a novel human image animation framework that uses masks as priors to guide the model's visual attention to particular regions and mitigate feature confusion between different parts of the image. As shown in Figure 3, ReMask-Animate includes three lightweight mask-guided adapters: (1) Mask-guided P-Adapter utilizes gated units to dynamically adjust the contribution of different modality pose conditions across various regions of the human body for enhanced pose control; (2) Mask-guided S-Adapter guides the model to generate foreground and background layouts using masks and separates the calculation of foreground and background from reference features to alleviate feature confusion; (3) Mask-guided C-Adapter injects identity features into the foreground region, enhancing identity consistency while reducing coupling with background features. Compared with current methods, our method achieves significant improvements in both qualitative and quantitative experiments. The main contributions of this work are as follows:

- We propose a Mask-guided Human-Centric framework, ReMask-Animate, which exclusively utilizes open-source datasets to achieve character image animation and significantly enhances the quality of visual generation.
- We design three distinct Mask-guided Adapters by introducing additional hand, face, and background masks as priors, guiding DenoiseNet's attention to specific regions. This approach reduces the difficulty of generating layouts using only pose conditions and mitigates the visual quality degradation caused by feature confusion.
- We conduct extensive quantitative and qualitative evaluations using public datasets, and the experimental results demonstrate the superiority of our method.

Related work

Diffusion Models for Image/Video Generation

Compared to GAN-based generative models, diffusion models have the advantages of high generation quality, stable training, and greater diversity. The initial latent diffusion model is applied in the field of image generation. However, text-to-image generation models based on diffusion, guided by textual prompts (Radford et al. 2021), often lack precision in spatial arrangement and structural details. To enhance control over visual generation, ControlNet (Zhang, Rao, and Agrawala 2023) employs zero convolutions and copy networks to control the semantics and structure of generated images. Furthermore, IP-Adapter introduces advanced visual semantic control through a dual-layer cross-attention mechanism (Ye et al. 2023). Due to their outstanding generative capabilities, diffusion models have recently been extended to video generation. Animatediff (Guo et al. 2024) extends image generation to video generation by integrating a tunable motion module, allowing adaptation to various personalized styles. Several studies (Ma et al. 2024) propose replacing self-attention mechanisms in generative models with inter-frame attention mechanisms to ensure temporal consistency in generated videos. VideoCrafter (Chen et al. 2023) combines textual and visual conditions to achieve both text-to-video and image-to-video generation.

Pose Guidance for Human Image Generation

High-quality human animation generation requires precise pose guidance. OpenPose (Cao et al. 2021), an open-source system for multi-person 2D pose detection, captures keypoints of the body, feet, hands, and face. DWPose (Yang et al. 2023), which enhances OpenPose, delivers more accurate hand and facial keypoint detections. DensePose (Güler, Neverova, and Kokkinos 2018) maps RGB images to a surface-based model to establish dense correspondences. Meanwhile, SMPL (Loper et al. 2015; Cai et al. 2023) introduces a 3D parametric model tailored to various body shapes, enriching capabilities for human animation. In this study, we integrate SMPL and DWPose to refine pose control, which surpasses traditional single-modality methods which rely on either keypoints or parametric models alone.

Diffusion Models for Human Animation

Animating human images is a significant challenge in video generation, which involves transforming static images into smooth, dynamic videos. Recently, the advanced generative capabilities of diffusion models have opened up new avenues for human animation. Disco (Wang et al. 2023) introduces a pose and background ControlNet to control character movements and ensure background consistency. MagicPose (Chang et al. 2023) leverages an appearance control model to improve generation quality. MagicAnimate (Xu et al. 2023) adopts a similar strategy but replaces OpenPose with DensePose for enhanced accuracy. AnimateAnyone (Hu et al. 2023) integrates an additional ReferenceNet to extract fine-grained features from reference images. This promotes character ID consistency and background stability.

Preliminary

Latent Diffusion Models (LDM)

LDM reduce computational resource consumption by transforming features from pixel space to latent space for the denoising process , while maintaining high generation quality. Specifically, the encoder of a variational autoencoder (VAE) compresses the pixel space image x to a latent space feature $z = \mathcal{E}(x)$, while the decoder reconstructs the generated latent space feature z' back into the image $x' = \mathcal{D}(z')$. The denoising process is an iterative Markov process that progressively denoises the initial Gaussian noise $z_r \sim \mathcal{N}(0, I)$ into the target latent space features z_o. The single-step iterative process involves predicting the noise of the forward process and subtracting it. LDM uses a U-Net or Transformer to estimate the noise, with the loss function of

$$L_{LDM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z^t, t, c)\|_2^2\right], \quad (1)$$

where $\epsilon_\theta(\cdot)$ is the network for predicting noise. ϵ represents the actual noise. c denotes the conditional information. t represents the timestep of the denoising process and z^t represents the intermediate result of denoising at timestep t.

Attention in Diffusion Models

The denoising network adopts a U-Net architecture, with each block containing self-attention (SA) and cross-attention (CA) layers. Specific editing tasks adjust attention calculations to modify the generated results. Research (Cao et al. 2023) demonstrates both self-attention and cross-attention layers are crucial for the layout, style, and content of the generated images. Besides, the visualization of the attention maps focuses on the critical human regions, determining the final position and region of the generated character. Inspired by this, we propose Mask-guided Adapters in both self-attention and cross-attention layers to guide the model's attention to visual features in specific regions and reduce the difficulty of learning the generated layout without significantly altering the model weights or inference speed.

Methods

Figure 3 illustrates the overview of our proposed pipeline. Given a reference image and pose sequences, the primary objective is to generate character video clips with a plausible and naturalistic motion while accurately preserving the visual appearance of the reference image.

Mask-guided P-Adapter

Different pose conditions emphasize varying aspects of human morphology. For instance, DWPose offers accurate detection for hands and faces, while SMPL excels in capturing body bends and shapes. As shown in Figure 5, different pose conditions can complement each other, but they may also conflict or misalign in certain regions. Therefore, it is essential to perform both intra-condition feature encoding, which obtains corresponding semantic information for each condition, and inter-condition feature fusion, which mitigates feature conflicts between conditions Therefore, we design a gated adapter based on hands and face masks to adjust the weighting of different pose features in specific regions.

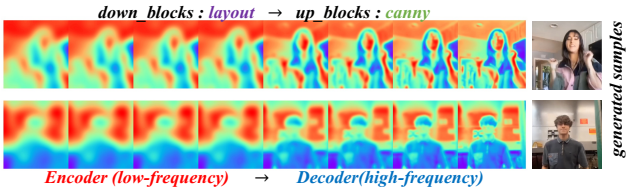

Figure 2: Attention in diffusion models.

The P-Adapter is designed to integrate the pose signals into the initial noise of the denoising process. Therefore, **it only needs to be applied to the initial input layer** of the U-Net. This process can be described as follows:

$$S = G([f_s, f_m]), \quad (2)$$

$$\mathcal{O}_p = \sum_{i=f,h,b} \text{Conv}_i(f_s \cdot S + f_m \cdot (1 - S)) \cdot \mathcal{M}_i, \quad (3)$$

where $\mathcal{M}$ denotes the corresponding mask, and i indicates the index of different masks, including face, hand, and body. G represents the gating unit, S represents the output weights of the gating unit, and f_s, and f_m represent sparse skeletal features and dense mesh features.

Mask-guided S-Adapter

Leveraging fine-grained semantic features in ReferenceNet significantly enhances the visual quality and consistency of generated images and videos. However, this enhancement comes at the cost of a substantial increase in token count during self-attention computations. This increased complexity hinders the model's ability to effectively learn structural and spatial layouts, indirectly leading to feature confusion across different regions. Utilizing foreground and background mask priors presents a direct solution for guiding the model in structural and layout generation. This approach facilitates dynamic adjustment of visual focus across various regions based on mask priors, effectively mitigating feature confusion. However, As illustrated in Figure 2, in the initial layers of the U-Net, the internal structure and layout within the self-attention mechanism remain underdeveloped. Applying complete control across all layers may lead to the model's excessive reliance on mask priors, compromising the effectiveness of pose conditioning. To mitigate this issue, **we introduce mutual control specifically within the decoder of the U-Net only**. Initially, we incorporate an additional self-attention layer bypass to facilitate supplementary feature computations. Subsequently, the outputs from these self-attention layers undergo refinement using a lightweight convolution layer to distinctly emphasize foreground and background features, respectively. This approach can be formalized as follows:

$$\text{Mask-SA}(z_t, z_{ref}) = \text{Softmax}\left(\frac{Q_s(K_s)^\top}{\sqrt{d}}\right) V_s \cdot \mathcal{M}_i, \quad (4)$$

$$O_s = \sum_{i=b,f} \text{Conv}_i(\text{Mask-SA}(z_t, z_{ref})), \quad (5)$$

where $Q_s = W_q z_t$, $K_s = W_k[z_t, z_{ref}]$, $V_s = W_v[z_t, z_{ref}]$, W_q, W_k, and W_v are learnable weight matrices of projection

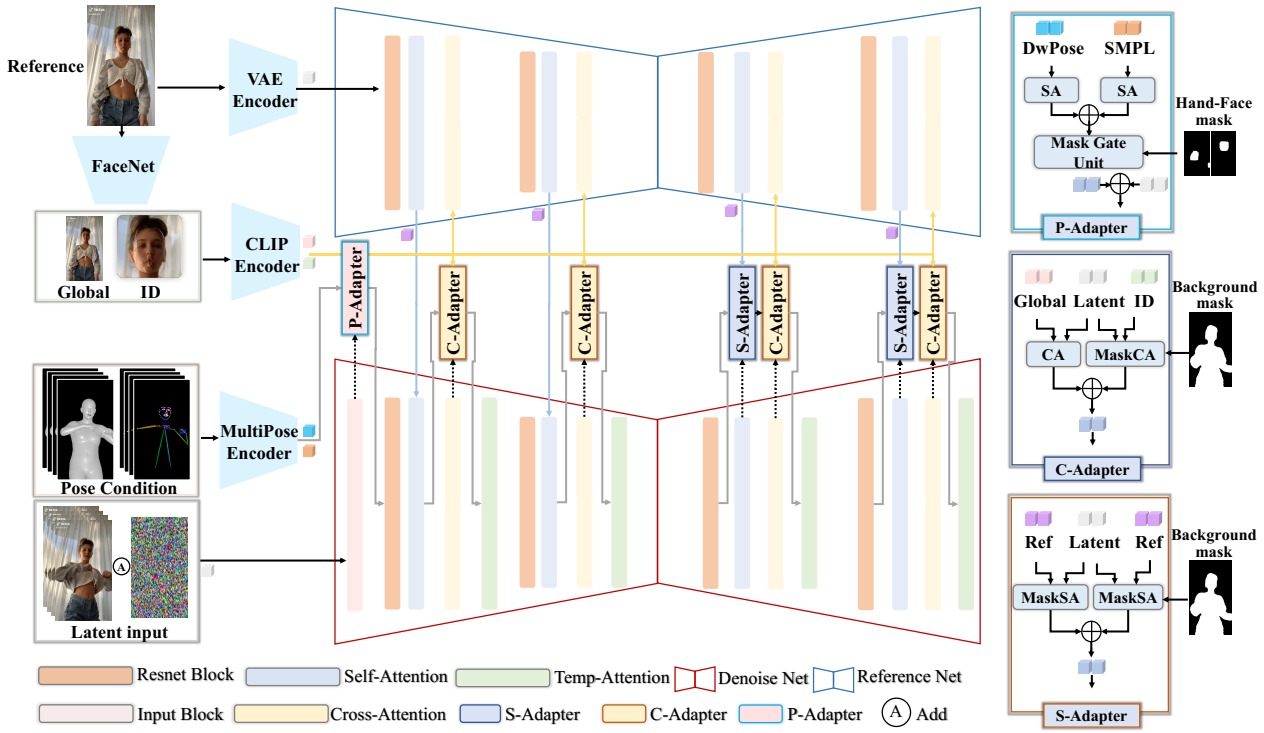

Figure 3: The overview architecture of ReMask-Animate. Given a reference image and its corresponding pose sequence, the P-adapter is utilized to modulate the weights of local features across various pose conditions, thereby improving pose accuracy. Following this, foreground and background mask priors are employed to guide DenoiseNet in layout generation, with separate processing of foreground and background features to minimize feature confusion across regions. To address the challenge of reduced character identity consistency in the generated images or videos, the C-Adapter leverages the foreground mask as a prior to accurately inject face features into the corresponding regions.

layers. [] denotes concatenation along the width dimension. b, f represent the background and foreground. z_t and z_{ref} represent the normalized latent feature of the self-attention layers in the denoising net and the reference net.

Mask-guided C-Adapter

In human image animation, preserving the appearance details of the reference image, such as character identity, clothing texture, and background, is essential. Recent techniques commonly employ ReferenceNet to encode the fine-grained features of the entire reference human image. However, these approaches often struggle in maintaining character identity consistency. Accordingly, we propose a Mask-Guided C-Adapter, which hierarchically computes cross-attention by introducing facial features and using a foreground mask to inject them into targeted regions. Leveraging global features from CLIP, it preserves high-frequency details in non-facial areas and **integrates seamlessly across all layers**, enhancing facial generation quality without affecting other regions. First, we use CLIP to extract the global feature f_g from the reference image for U-Net's cross-attention. This process can be mathematically represented as follows:

$$\text{CA}(z_t, f_g) = \text{Softmax}\left(\frac{Q_c(K_c)^\top}{\sqrt{d}}\right)V_c, \tag{6}$$

where $Q_c = W_q z_t, K_c = W_k f_g, V_c = W_v f_g$. Next, we use a pre-trained FaceNet to identify and crop the facial region from the reference image, obtaining a 224×224 facial image. This pure facial image is then input into CLIP to extract only facial features f_{id}, which are injected into the C-Adapter to enhance facial quality as follows:

$$\text{Mask-CA}(z_t, f_{id}) = \text{Softmax}\left(\frac{Q_c(K_c')^\top}{\sqrt{d}}\right)V_c' \cdot \mathcal{M}_f, \tag{7}$$

where $K_c' = W_k' f_{id}$, $V_c' = W_v' f_{id}$, M_f represents the foreground mask scaled to the feature size. Hence, the final formulation of the C-Adapter is defined as follows:

$$O_c = \text{CA}(z_t, f_g) + \text{Mask-CA}(z_t, f_{id}). \tag{8}$$

Mask Generate Process

The hand and face masks are used to adjust the weighting of conditioned pose features within specific regions. As illustrated in Figure 4, we apply dilated convolutions to the skeletal keypoints to generate these masks. The foreground-background masks play a crucial role in determining the final layout of the generated content. During training, we utilize the masks provided by the dataset. However, during inference, these masks are not available beforehand. To address this issue, we adopt a self-evolution approach to approximate the scenario. Initially, a rectangular mask is used

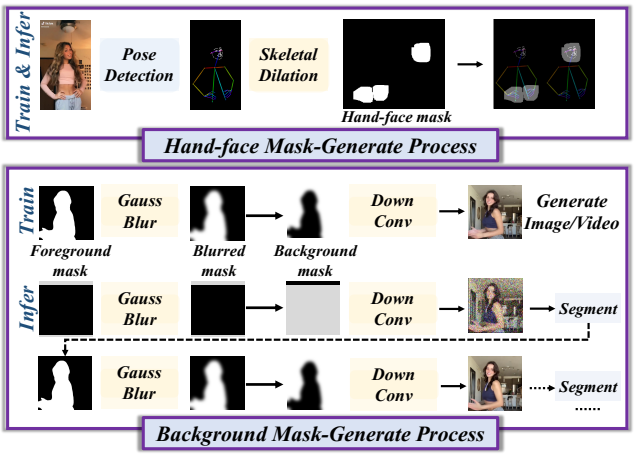

Figure 4: The pipeline of the mask generation process. The first line employs dilated convolutions to generate hand and face masks, while the second line utilizes a recursive self-evolution algorithm to produce background masks.

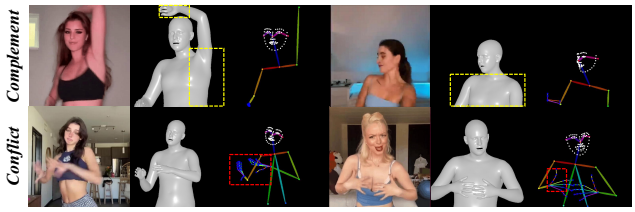

Figure 5: Complementarity and conflict between different pose conditions are rigorously depicted. The yellow dashed box effectively demonstrates how complementarity enhances interactions, while the red dashed box highlights conflicts, pinpointing areas where discrepancies arise.

to guide the generation process, which is later refined by a pre-trained segmentation network applied to the generated image or video. This iterative process updates and improves the accuracy of the masks over N cycles.

Experiments

Experimental Setup

Datasets. The TikTok dataset comprises 350 dance videos, each ranging from 10 to 15 seconds, predominantly capturing facial expressions and upper body movements. In contrast, the Fashion dataset is characterized by a minimalistic, pure white background and limited motion variation, including 500 training videos and 100 testing videos, with each video containing approximately 350 frames. For each video, we extract frames at a rate of 30 fps and apply DW-Pose (Yang et al. 2023) and SMPL (Cai et al. 2023) to each frame to obtain the corresponding pose sequence.

Training Strategy. Our model extends previous work by employing a two-stage training strategy. During both stages, we freeze the CLIP (Radford et al. 2021) image encoder and VAE (van den Oord, Vinyals, and Kavukcuoglu 2017). In the initial stage, DenoiseNet, ReferenceNet, and multiple-condition pose encoders are trained to synchronize their spatial generative capabilities. In the subsequent stage, we introduce the motion module to generate video, while maintaining all other weights frozen.

Implementation Details. We train our model using 4 NVIDIA A800 GPUs in a two-stage process. In the first stage, we randomly center-crop the input images to 768×768, use a batch size of 4, and train the model for 60,000 steps with a learning rate of 0.0001. In the second stage, we randomly center-crop video frames to 512×512, use a batch size of 1, and train for an additional 20,000 steps while maintaining the same learning rate.

Quantitative Comparison. Our evaluation metrics adhere to research standards. To assess the quality of individual frames, we consistently employ conventional image metrics, including L1, PSNR (Horé and Ziou 2010), SSIM (Wang et al. 2004), and LPIPS (Zhang et al. 2018). For video evaluation, we utilize the FID-VID (Balaji et al. 2019) metric, generating samples by concatenating every 16 frames.

Comparisons

Baselines. We compare our model with several state-of-the-art character animation methods, which can be categorized into two main types: (1) GAN-based methods, such as MRAA and TPSMM; and (2) advanced latent diffusion methods extended to videos, including DreamPose, Disco, MagicAnimate, MagicPose, and AnimateAnyone.

Evaluation on TikTok dataset. Table 1 presents a comparative analysis of experimental results on the TikTok dataset. Our approach consistently outperforms existing techniques across four critical metrics: SSIM, L1, PSNR, and FID-VID, while ranking second in the LPIPS metric. Notably, our model demonstrates significant advancements, particularly in single-frame SSIM and video-based FID-VID metrics, with an improvement of 0.067 in SSIM and 2.24 in FID-VID over leading models like MagicAnimate and AnimateAnyone. Figure 6 provides a visual comparison of the synthesized results, further highlighting the advantages of our approach. The first and fifth columns showcase our approach's exceptional ability to render intricate hand regions, while the third and fourth columns emphasize the model's robust pose control, accurately capturing complex motions such as arm bending. The sixth and eighth columns underscore the method's realistic portrayal of human body structure, with no observable errors like extra hands or misaligned limbs. These findings clearly demonstrate the advantages of our approach in advancing human animation.

Evaluation on Fashion dataset. Figure 8 presents qualitative results on the Fashion dataset. Due to the dataset's uniform white background and limited pose variations, our method achieves performance comparable to other methods while better-preserving ID consistency during rotations.

Ablation Study

To validate the effectiveness of the proposed modules, we conduct extensive ablation studies by testing three model

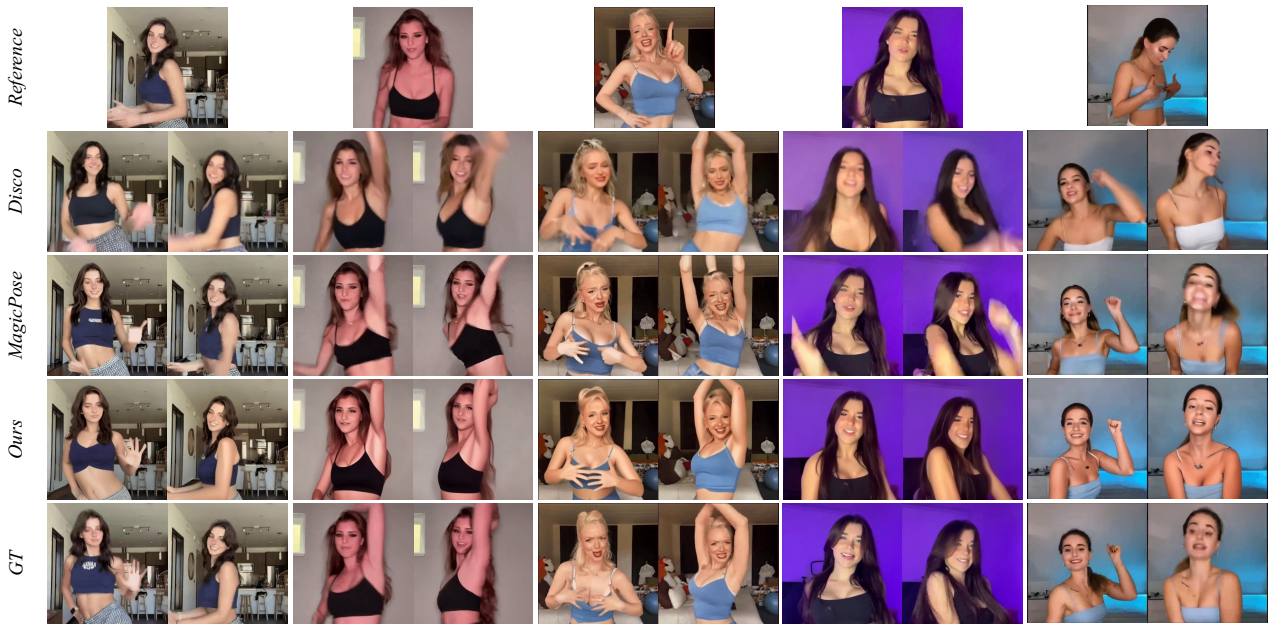

Figure 6: Qualitative comparison with existing state-of-the-art methods on the TikTok dataset.

Method	L1 ↓	PSNR* ↑	SSIM ↑	LPIPS ↓	FID-VID ↓
FOMM (Siarohin et al. 2019a)	3.61E-04	17.26	0.648	0.335	90.09
MRAA (Siarohin et al. 2021)	3.21E-04	18.41	0.672	0.296	66.36
TPSMM (Zhao and Zhang 2022)	3.23E-04	18.32	0.673	0.299	72.55
DreamPose (Karras et al. 2023)	6.88E-04	12.82	0.511	0.442	80.51
DisCo (Wang et al. 2023)	3.78E-04	16.55	0.668	0.292	61.41
MagicAnimate (Xu et al. 2023)	3.13E-04	-	0.714	0.239	21.75
MagicPose (Chang et al. 2023)	0.81E-04	17.33	0.752	0.292	46.30
Animate Anyone (Hu et al. 2023)	-	-	0.718	0.285	-
Ours	0.72E-04	18.61	0.785	0.261	19.51

Table 1: Performance comparison of different methods on TikTok dataset. * indicates that the correct result is obtained using a method that prevents numerical overflow.

variants, each excluding one of the P-Adapter, S-Adapter, and C-Adapter. As demonstrated in Table 2, our full method, which includes all components, consistently outperforms the other variants across all evaluation metrics. This finding underscores the substantial positive impact of these modules in enhancing the quality of generated animations.

Efficiency Analysis. In our study, we thoroughly assess the efficiency of the proposed adapters—Mask-guided P-Adapter, S-Adapter, and C-Adapter—with 8.17 million, 9.24 million, and 9.95 million trainable parameters, respectively. Their integration modestly increases computational demand, with memory usage rising by 0, 3.06, and 1.04 and training time extending by 0, 0.88, and 0.34 hours. Despite this, our approach significantly enhances the base model's overall performance with minimal computational overhead.

Mask-guided P-Adapter. Different modalities of pose conditions can sometimes complement each other by compensating for their respective shortcomings, but in some cases, they may also conflict. As illustrated in Figure 5, the first row shows that SMPL features can compensate for DWPose's deficiencies in controlling arm bending for certain specific poses. The second row demonstrates that SMPL features result in inaccurate hand pose modeling, which conflicts with DWPose. Figure 7a qualitatively demonstrates the results of the model variations and the complete model. The complete method shows higher quality in generating hands

Method	PSNR*	SSIM ↑	LPIPS ↓	FID-VID ↓
w/o P-A	18.48	0.771	0.275	22.77
w/o S-A	18.11	0.735	0.290	25.10
w/o C-A	18.25	0.769	0.271	23.10
Ours	**18.61**	**0.785**	**0.261**	**19.51**

Table 2: Quantitative ablation results on TikTok dataset.

(a) Qualitative ablation of P-Adapter (b) Qualitative ablation of S-Adapter (c) Qualitative ablation of C-Adapter

Figure 7: Qualitative ablations of Mask-guided Adapters

Figure 8: Qualitative comparison for fashion video synthesis

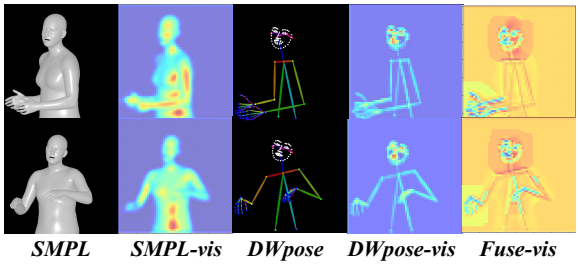

Figure 9: Visualization of Features of Pose Condition

and effectively reduces the issues of limb and body overlap misalignment. Directly merging two types of pose features without leveraging a P-adapter can induce conflicts in specific conditioned regions, significantly degrading the generation quality of human body parts in those areas. Additionally, Figure 9 visualizes the features derived from different pose modalities. DWPose features are focused on hand and face constraints, while SMPL features prioritize body shape constraints. The fused features effectively combine the strengths of both modalities by adaptively adjusting feature weights in specific regions.

Mask-guided S-Adapter. We perform a qualitative analysis of character animations with intricate backgrounds, as illustrated in Figure 7b, to validate the efficacy of the S-Adapter. Our full method excels in generating text on clothing, meticulously preserving text clarity and positioning, while avoiding the issue of background elements being redundantly repeated in other regions. This outcome highlights the substantial positive influence of the S-Adapter.

Method	PSNR*	SSIM ↑	LPIPS ↓	FID-VID ↓
w/o Ske	18.47	0.765	0.278	20.61
w/o Mesh	18.54	0.777	0.270	20.13

Table 3: Quantitative results under different pose conditions.

Model	Params (M)	GPU Mem (G)	Costs (H)
Full	887.91	65.01	27.21
w/o P-A	-8.17	-	-
w/o S-A	-9.24	-3.06	-0.88
w/o C-A	-9.95	-1.04	-0.34

Table 4: Efficiency analysis of different modules.

Mask-guided C-Adapter. We conduct a comparison experiment of character animation with facial effects to validate the positive impact of the Mask-guided C-Adapter. As illustrated in Figure 7c, the Mask-guided C-Adapter effectively preserves facial identity during character movements, such as head turns, ensuring the generated images/videos maintain high fidelity and consistent visual quality.

Conclusion

In this paper, we propose a pose-controllable character animation framework called ReMask-Animate. This innovative framework leverages diverse region-mask priors to adjust the model's visual attention to specific areas, effectively directing the generation of layouts for images and videos. Our method seamlessly integrates three types of mask-guided adapters into the PoseNet, SA, and CA layers. These adapters effectively reduce feature confounding, ensuring pose accuracy and high visual quality. Experimental evaluations demonstrate the effectiveness and quality of our method in generating character images and animations.

References

Balaji, Y.; Min, M. R.; Bai, B.; Chellappa, R.; and Graf, H. P. 2019. Conditional GAN with Discriminative Filter Generation for Text-to-Video Synthesis. In Kraus, S., ed., *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, 1995–2001. ijcai.org.

Cai, Z.; Yin, W.; Zeng, A.; Wei, C.; Sun, Q.; Yanjun, W.; Pang, H. E.; Mei, H.; Zhang, M.; Zhang, L.; Loy, C. C.; Yang, L.; and Liu, Z. 2023. SMPLer-X: Scaling Up Expressive Human Pose and Shape Estimation. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Cao, M.; Wang, X.; Qi, Z.; Shan, Y.; Qie, X.; and Zheng, Y. 2023. MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 22503–22513. IEEE.

Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.; and Sheikh, Y. 2021. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(1): 172–186.

Chan, C.; Ginosar, S.; Zhou, T.; and Efros, A. A. 2019. Everybody Dance Now. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 5932–5941. IEEE.

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. *CoRR*, abs/2311.12052.

Chen, H.; Xia, M.; He, Y.; Zhang, Y.; Cun, X.; Yang, S.; Xing, J.; Liu, Y.; Chen, Q.; Wang, X.; Weng, C.; and Shan, Y. 2023. VideoCrafter1: Open Diffusion Models for High-Quality Video Generation. *CoRR*, abs/2310.19512.

Dhariwal, P.; and Nichol, A. Q. 2021. Diffusion Models Beat GANs on Image Synthesis. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y. N.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 8780–8794.

Güler, R. A.; Neverova, N.; and Kokkinos, I. 2018. DensePose: Dense Human Pose Estimation in the Wild. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 7297–7306. Computer Vision Foundation / IEEE Computer Society.

Guo, Y.; Yang, C.; Rao, A.; Liang, Z.; Wang, Y.; Qiao, Y.; Agrawala, M.; Lin, D.; and Dai, B. 2024. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Horé, A.; and Ziou, D. 2010. Image Quality Metrics: PSNR vs. SSIM. In *20th International Conference on Pattern Recognition, ICPR 2010, Istanbul, Turkey, 23-26 August 2010*, 2366–2369. IEEE Computer Society.

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. *CoRR*, abs/2311.17117.

Jang, Y.; Kim, J.; Ahn, J.; Kwak, D.; Yang, H.; Ju, Y.; Kim, I.; Kim, B.; and Chung, J. S. 2024. Faces that Speak: Jointly Synthesising Talking Face and Speech from Text. *CoRR*, abs/2405.10272.

Karras, J.; Holynski, A.; Wang, T.; and Kemelmacher-Shlizerman, I. 2023. DreamPose: Fashion Image-to-Video Synthesis via Stable Diffusion. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 22623–22633. IEEE.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6): 248:1–248:16.

Ma, Y.; He, Y.; Cun, X.; Wang, X.; Chen, S.; Li, X.; and Chen, Q. 2024. Follow Your Pose: Pose-Guided Text-to-Video Generation Using Pose-Free Videos. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, 4117–4125. AAAI Press.

Milis, G.; Filntisis, P. P.; Roussos, A.; and Maragos, P. 2023. Neural Text to Articulate Talk: Deep Text to Audiovisual Speech Synthesis achieving both Auditory and Photorealism. *CoRR*, abs/2312.06613.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 10674–10685. IEEE.

Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2019a. First Order Motion Model for Image Animation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 7135–7145.

Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2019b. First Order Motion Model for Image Animation. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Siarohin, A.; Woodford, O. J.; Ren, J.; Chai, M.; and Tulyakov, S. 2021. Motion Representations for Articulated Animation. In *IEEE Conference on Computer Vision and*

Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, 13653–13662. Computer Vision Foundation / IEEE.

van den Oord, A.; Vinyals, O.; and Kavukcuoglu, K. 2017. Neural Discrete Representation Learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 6306–6315.

Wang, T.; Li, L.; Lin, K.; Zhai, Y.; Lin, C.-C.; Yang, Z.; Zhang, H.; Liu, Z.; and Wang, L. 2023. Disco: Disentangled control for realistic human dance generation. *arXiv preprint arXiv:2307.00040*.

Wang, T.-C.; Mallya, A.; Liu, M.-Y.; and xxx. 2021. One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4): 600–612.

Xu, Z.; Zhang, J.; Liew, J. H.; Yan, H.; Liu, J.; Zhang, C.; Feng, J.; and Shou, M. Z. 2023. MagicAnimate: Temporally Consistent Human Image Animation using Diffusion Model. *CoRR*, abs/2311.16498.

Yang, Z.; Zeng, A.; Yuan, C.; and Li, Y. 2023. Effective Whole-body Pose Estimation with Two-stages Distillation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023 - Workshops, Paris, France, October 2-6, 2023*, 4212–4222. IEEE.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. *CoRR*, abs/2308.06721.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 3813–3824. IEEE.

Zhang, P.; Yang, L.; Lai, J.; and Xie, X. 2022. Exploring Dual-task Correlation for Pose Guided Person Image Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 7703–7712. IEEE.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 586–595. Computer Vision Foundation / IEEE Computer Society.

Zhao, J.; and Zhang, H. 2022. Thin-Plate Spline Motion Model for Image Animation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 3647–3656. IEEE.

Zhou, Y.; Wang, Z.; Fang, C.; Bui, T.; and Berg, T. L. 2019. Dance Dance Generation: Motion Transfer for Internet Videos. In *2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*, 1208–1216. IEEE.