Improved Pose-Controlled Animation: A Quantitative and Qualitative Analysis

Qinghui Xu[1, a], YanLin Wu[2], Yajun Yuan[2], Zongqi Ge[3], Khang Wen Goh[1, *]

[1]*Faculty of Data Science and Information Technology, INTI International University, Nilai 71800, Negeri Sembilan, Malaysia,*
[2]*School of Mathematics Education Management, INTI International University, Nilai 71800, Negeri Sembilan, Malaysia,*
[3]*University of East London Singapore Campus, 069542, Singapore,*
[a]*qinghui.xu.sg@gmail.com,*

Abstract:*Character animation, which aims to generate dynamic character videos from static images, has gained significant attention in recent years. Despite the advances in diffusion models, which have established themselves as the leading approach in visual generation tasks due to their strong generative capabilities, challenges remain in the domain of image-to-video synthesis, particularly in character animation. The preservation of temporal consistency and the retention of fine-grained character details across frames continue to pose significant obstacles. In this work, we propose a novel framework specifically designed for character animation, leveraging the potential of diffusion models. To address the challenge of maintaining intricate appearance details from the reference image, we introduce ReferenceNet, a network that integrates detailed features using spatial attention mechanisms. To enhance controllability and ensure smooth motion transitions, we present an efficient pose guide that directs the character's movements and incorporate an effective temporal modeling strategy to facilitate seamless inter-frame consistency. Our framework is capable of animating arbitrary characters by expanding the training data, outperforming existing image-to-video methods in character animation tasks. Experimental evaluations on benchmark image animation datasets demonstrate that our approach achieves state-of-the-art performance, setting a new standard for this domain.*

Keywords: *AI, Image recognition, video generation, diffusion model, dynamic video*

1. Introduction

Character animation, which involves transforming static character images into dynamic videos that follow specific pose sequences, holds immense potential across various domains, including online retail, entertainment, artistic production, and virtual characters. Since the introduction of Generative Adversarial Networks (GANs) [1, 11, 22], numerous studies have explored the challenges of image animation and pose transfer [7, 33, 37-39, 57, 61, 64]. Despite significant progress, issues such as local distortions, blurred details, semantic inconsistencies, and temporal instability continue to hinder the widespread adoption of these approaches in practical applications.

In recent years, diffusion models [14] have emerged as a powerful tool in generating high-quality images and videos. Researchers have begun exploring their potential for human image-to-video tasks,

leveraging the robust generative capabilities of pretrained diffusion model architectures. For instance, DreamPose [21] focuses on fashion image-to-video generation, enhancing Stable Diffusion [34] by integrating features from CLIP [31] and VAE [24] to synthesize realistic results. However, DreamPose requires fine-tuning on input samples for consistency, which impacts operational efficiency. Similarly, DisCo [47] investigates human dance generation by modifying Stable Diffusion and utilizing CLIP for character features and ControlNet [60] for background features. Despite these advancements, challenges remain in maintaining character details and addressing issues like inter-frame jitter.

Additionally, current research in character animation often targets specific tasks or datasets, limiting the ability to generalize across different applications. While text-to-video models [2, 19, 29, 32, 34, 36] have made significant strides in visual quality and diversity, methods extending this approach to image-to-video synthesis [8, 12, 48, 63] still struggle with capturing fine-grained details from the source image. These methods tend to offer greater diversity but lack the precision necessary for character animation, leading to inconsistencies in appearance and temporal variations in character details. Furthermore, when faced with large or complex movements, these approaches struggle to produce stable, continuous animations.

To address these issues, we introduce Animate Anyone, a novel method capable of converting character images into animated videos controlled by specific pose sequences. Our approach builds on the Stable Diffusion (SD) architecture, utilizing its network design and pretrained weights, and extends the denoising UNet [35] to handle multi-frame inputs. To preserve appearance consistency, we propose ReferenceNet, a specialized symmetrical UNet structure designed to capture spatial details from the reference image. At each layer of the UNet, features from ReferenceNet are integrated into the denoising UNet using spatial attention [46], allowing the model to effectively maintain consistency in appearance across frames.

For pose controllability, we develop a lightweight pose guider that efficiently incorporates pose signals into the denoising process. To ensure smooth temporal motion, we introduce a temporal layer that models relationships between frames, enabling stable and continuous transitions. This approach not only preserves high-resolution details but also generates animations with fluid motion across multiple frames.

We train our model on an internal dataset of 5,000 character video clips. Figure 1 showcases the animation results for various characters, demonstrating the effectiveness of our approach. Compared to previous methods, our method offers several advantages: it maintains both spatial and temporal consistency, produces high-definition videos without jitter or flickering, and can animate any character image, regardless of domain constraints. We evaluate our method on three distinct human video synthesis benchmarks—the UBC fashion video dataset [59], the TikTok dataset [20], and the Ted-Talk dataset [39]—and show that our method outperforms existing techniques. Additionally, when compared with other image-to-video methods trained on large-scale datasets, our approach demonstrates superior results in character animation. We believe that Animate Anyone has the potential to serve as a foundational tool for character video creation, paving the way for the development of more innovative and creative applications in the field.

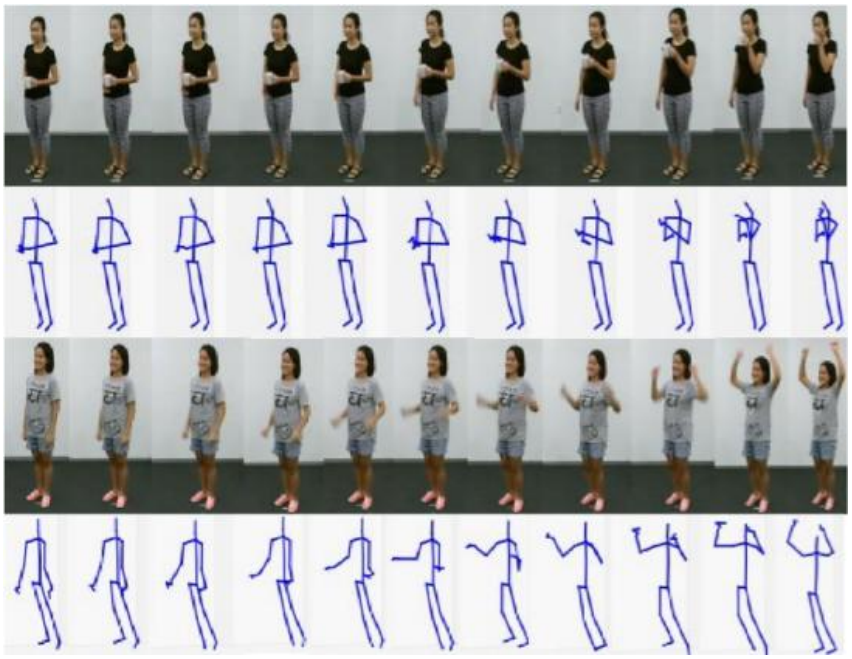

Figure 1. Consistent and Controllable Character Animation. Given a reference image (the leftmost image in each group), our method demonstrates the ability to animate a wide range of characters. It produces high-quality, temporally stable videos while preserving the visual consistency of the reference character's details.

2. Related Works

2.1 Diffusion Models for Image Generation

In recent years, diffusion models have emerged as the dominant approach in the field of text-to-image generation, offering superior results in terms of image quality and realism. These models rely on iterative denoising processes that enable them to synthesize high-fidelity images, making them a key tool in the image generation landscape. To enhance computational efficiency without sacrificing output quality, the Latent Diffusion Model (LDM) [34] proposes working in a lower-dimensional latent space, which reduces the computational load while still achieving high-quality results.

Recent advancements in controlling the generative process have led to the development of models such as ControlNet [60] and T2I-Adapter [27], which integrate additional encoding layers to provide more precise control over visual attributes, including pose, depth, and other fine-grained features. This has greatly expanded the range of controllable outputs, enabling targeted generation based on specific input conditions. For example, the IP-Adapter [56] enables diffusion models to generate images that adhere to specific criteria derived from a given input prompt, thus improving flexibility in content generation.

Additionally, several works have focused on enhancing image editing capabilities within the diffusion framework. ObjectStitch [42] and Paint-by-Example [53] leverage the CLIP [31] model to refine the generated images, guiding them with specific semantic conditions. Meanwhile, TryonDiffusion [65] applies these models to the virtual clothing try-on domain, introducing a Parallel-UNet structure to allow for realistic garment fitting and visualization.

2.2 Diffusion Models for Video Generation

Building on the success of diffusion models in image generation, there has been a significant push

towards adapting these methods for video synthesis. Many of these approaches extend the structure of text-to-image models by incorporating temporal components, which allow the models to generate coherent video sequences. Some studies [10, 16, 17, 23, 26, 30, 40, 52, 54] have explored incorporating inter-frame attention mechanisms into diffusion models to capture temporal dependencies between video frames. By introducing temporal layers into pre-existing text-to-image models, these methods seek to enable video generation while maintaining continuity across frames.

One such effort, VideoLDM [4], first pretrains the model on image data and subsequently fine-tunes it by introducing temporal layers for video generation. This method demonstrates how pretrained models can be extended to handle video sequences, while also addressing the need for efficient training. Similarly, AnimateDiff [12] introduces a motion module that can be integrated into a variety of text-to-image models, enabling them to generate dynamic video content without requiring extensive fine-tuning.

A related approach is the extension of text-to-image models to image-to-video generation, which has been explored in various studies. For instance, VideoComposer [48] introduces images as conditional inputs during training, allowing for the generation of videos that remain consistent with the original images. Similarly, AnimateDiff [12] uses a weighted mixture of image latent representations and random noise to generate more controlled video sequences. VideoCrafter [8] incorporates both textual and visual features from CLIP [31], using cross-attention mechanisms to better integrate these inputs into video synthesis. While these methods have made significant strides in video generation, challenges remain in achieving stable, high-quality human video synthesis, particularly when incorporating image conditions into the process.

2.3 Diffusion Models for Human Image Animation

Human image animation, the task of generating animated sequences or videos from static images, has been a major focus of recent research. Leveraging the strengths of diffusion models in image generation, several approaches have been developed to address the unique challenges of human pose transfer and animation. These methods aim to synthesize realistic human movements while maintaining visual consistency across frames.

For example, PIDM [3] introduces a texture diffusion block that allows for the incorporation of specific texture patterns during denoising, improving the quality of human pose transfer. Similarly, LFDM [28] synthesizes optical flow sequences in the latent space, allowing the model to warp input images according to motion parameters specified by the user. LEO [49] proposes representing human motion as a sequence of flow maps, which the diffusion model uses to generate smooth motion transitions.

Another notable approach, DreamPose [21], builds on the pretrained Stable Diffusion model and proposes an adapter that leverages CLIP and VAE embeddings to enhance pose transfer capabilities. DisCo [47] takes inspiration from ControlNet, decoupling the control of pose and background, which allows for greater flexibility in generating complex animations.

Despite these advances, the integration of diffusion models into human image animation still faces several challenges. These include issues such as texture inconsistency across frames and the difficulty in maintaining temporal stability, leading to visual artifacts like jitter and flickering. Furthermore, the generalization of these models across various human animation tasks, while preserving fine details, remains an open research problem. Current methods still struggle to handle the complex relationships

between pose, appearance, and temporal continuity in human image animation, suggesting the need for more robust models that can handle diverse animation tasks while maintaining high quality.

3. Methods

This work focuses on pose-guided image-to-video synthesis for character animation. Given a reference image that describes the appearance of a character, along with a sequence of poses, our model generates an animated video of the character. The overall pipeline of our approach is depicted in Figure 2. In this section, we first provide a brief overview of Stable Diffusion in Section 3.1, which serves as the foundational framework for our method. Section 3.2 details the specific design choices and architectural components of our model, and Section 3.3 outlines the training strategy employed for optimization.

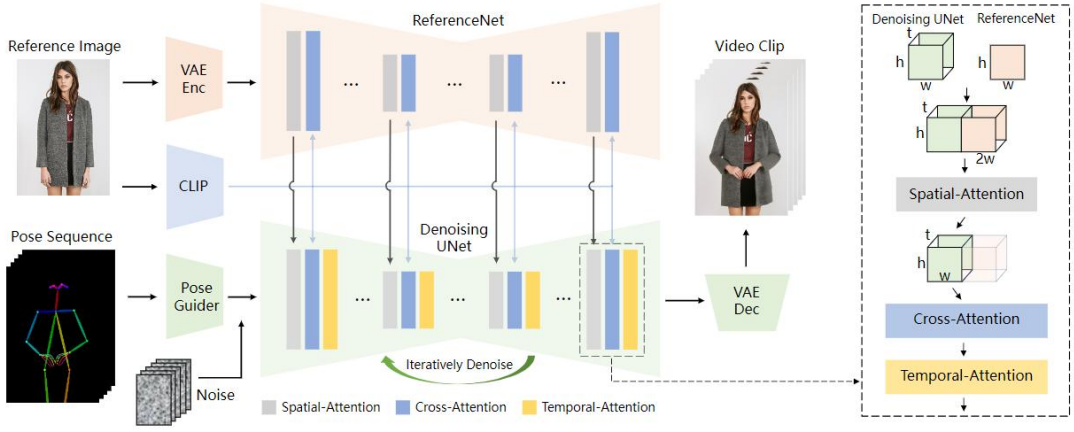

Figure 2. Overview of Our Method. The proposed method begins by encoding the pose sequence using the Pose Guider, which is then fused with multi-frame noise. The Denoising UNet performs the denoising process to generate video frames, with its computational blocks comprising Spatial-Attention, Cross-Attention, and Temporal-Attention, as depicted in the dashed box on the right. The integration of the reference image involves two components: (1) extracting detailed features via ReferenceNet for Spatial-Attention, and (2) extracting semantic features via the CLIP image encoder for Cross-Attention. Temporal-Attention operates across the temporal dimension to ensure smooth motion transitions. Finally, the VAE decoder reconstructs the latent representations into video clips.

3.1. Preliminary: Stable Diffusion

Our method is an extension of Stable Diffusion (SD), which itself is derived from the Latent Diffusion Model (LDM). SD aims to reduce the computational complexity of image generation by performing feature distribution modeling in a lower-dimensional latent space.

The core of SD is its autoencoder [24,45], which includes an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$. The encoder maps an image x to a latent representation $\mathrm{z} = \mathcal{E}(\mathrm{x})$, and the decoder reconstructs the image from this latent representation $\hat{\mathrm{x}} = \mathcal{D}(\mathrm{z})$. In the SD framework, a noise vector is progressively denoised through multiple timesteps to recover the original image. The process starts by generating a noisy latent representation z_t at timestep t, and a denoising U-Net model is trained to predict the noise added at each timestep. The objective function for the training process is defined as:

$$\mathcal{L} = \mathbb{E}_{x,\epsilon,t}\left(\left|\epsilon - \hat{\epsilon}_\theta(z_t, c, t)\right|\right)$$

where $\hat{\epsilon}_\theta$ denotes the denoising function, c represents the conditional embeddings (e.g., text prompts), and t is the timestep. The denoising U-Net architecture typically includes four downsampling layers, a middle layer, and four upsampling layers. Each block within these layers incorporates 2D convolutions, self-attention [46], and cross-attention (often referred to as a Res-Trans block). Cross-attention is used to fuse the text embeddings with the visual features in the network. During inference, a random latent vector z_0 is sampled from a Gaussian distribution, and denoising proceeds in a deterministic fashion through methods like DDPM [14] or DDIM [41]. The U-Net model iteratively predicts the noise at each timestep, and the final latent representation z_0 is decoded back into the image space.

3.2. Network Architecture

The architecture of our model is depicted in Figure 2. The network receives multi-frame noisy inputs, and the denoising U-Net is based on the design of Stable Diffusion. It uses the same structural components and shares pre-trained weights from SD. In addition, we introduce three critical components to enhance the generation process: 1) ReferenceNet, which encodes the appearance features of the character from the reference image, 2) Pose Guider, which encodes the motion control signals for producing controllable character movements, and 3) Temporal Layer, which captures temporal dependencies to ensure continuity of character motion across frames.

3.2.1. ReferenceNet

In contrast to text-to-image models, which only require semantic relevance between the input text and generated images, image-to-video tasks demand precise consistency in visual details. Traditional image-driven methods, such as those using CLIP image encoders [8, 21, 42, 47, 53, 56], suffer from detail inconsistency because the CLIP image encoder processes low-resolution images (224x224), which leads to the loss of fine-grained features. Furthermore, CLIP was trained to match semantic features for text-to-image generation, not to preserve the detailed structures of images.

To address these limitations, we propose ReferenceNet, an image feature extraction network specifically designed to capture the fine-grained details of the reference image. ReferenceNet shares the same architecture as the denoising U-Net, but without the temporal layer. It also inherits the pre-trained weights from the original SD model, and its parameters are updated independently during training.

The features from ReferenceNet are integrated into the denoising U-Net by replacing the self-attention layer with a spatial-attention layer. Specifically, the feature maps c_i^R from the denoising U-Net and z_e^R from ReferenceNet are concatenated along the spatial dimension w ,and a self-attention mechanism is applied to the concatenated map. This design allows the U-Net to selectively learn features from ReferenceNet that align with the spatial features of the target image. In this way, ReferenceNet acts as a reference model, ensuring that highlevel features from the source image are preserved in the output.

3.2.2. Pose Guider

In contrast to ControlNet [60], which introduces control features such as depth and edges into the denoising U-Net through zero convolutions, we employ a lightweight Pose Guider to control the motion of the character. The Pose Guider comprises four convolutional layers with 4x4 kernels and 2x2

strides, with increasing channel sizes (16, 32, 64, 128). It aligns the pose image with the resolution of the latent noisy input and injects this processed pose information directly into the denoising U-Net. This approach maintains computational efficiency while offering high flexibility in controlling the character's movement.

3.2.3. Temporal Layer

To ensure temporal consistency across video frames, we introduce a Temporal Layer, which incorporates temporal attention into the model. Previous works have demonstrated the importance of modeling temporal dependencies in text-to-image models to facilitate videc generation. Our temporal layer is integrated into the Res-Trans blocks of the denoising UNet, specifically after the spatial-attention and cross-attention components.lt reshapes the feature map $\mathrm{c} \in \mathbb{R}^{\mathrm{b}\times h\times \mathrm{w}\times \mathrm{c}}$ to $\mathrm{c} \in \mathbb{R}^{(\mathrm{b}\times h\times \mathrm{w})\times \mathrm{t}\times \mathrm{c}}$ and performs self-attention along the temporal dimension t. The output of the temporal layer is then fused with the original feature map through a residual connection. The temporal layer captures the motion dynamics of the character across frames, ensuring smooth transitions and maintaining consistency in the appearance details. The Pose Guiden already ensures continuous character movement, so the temporal layer primarily serves to enforce temporal smoothness, eliminating the need for complex motion modeling

3.3. Training Strategy

Our training procedure follows a two-stage process. In the first stage, we train the model using individual video frames. The denoising U-Net is initialized with pre-trained weights from Stable Diffusion, and the ReferenceNet and Pose Guider are trained alongside it. During this phase, the temporal layer is excluded from the model, and the input consists of single-frame noise. The optimization objective is to generate high-quality images that match the reference image and the target pose. The reference image is randomly selected from a video clip to guide the generation process.

In the second stage, we introduce the Temporal Layer into the model, initialized with pre-trained weights from AnimateDiff [12]. During this phase, we use video clips consisting of 24 frames as input, and only the temporal layer is trained, with the weights of the denoising U-Net, ReferenceNet, and Pose Guider kept fixed. This stage ensures that the model learns to maintain temporal consistency across frames while preserving the fine-grained details of the character's appearance.

Table 1: Quantitative Comparison for Fashion Video Synthesis

Method	**SSIM ↑**	**PSNR ↑**	**LPIPS ↓**	**FVD ↓**
MRAA [35]	0.749	24.07	0.212	253.6
TPSMM [62]	0.746	34.75	0.213	247.5
BDMM [54]	0.918	36.01	0.048	148.3
DreamPose [45]	0.885	38.49	0.068	238.7
DreamPose*	0.879	38.49	0.095	279.6
SD-2V	0.894	0.931	0.111	175.4

Ours	**0.932**	**38.47**	**0.041**	**81.5**

Table 2: Quantitative Comparison for Human Dance Generation

Method	**SSIM ↑**	**PSNR ↑**	**LPIPS ↓**	**FVD ↓**
FOMM [45]	0.648	29.01	0.335	405.2
MRAA [31]	0.672	29.39	0.296	284.8
PSMM [47]	0.673	29.18	0.299	306.1
Disco [35]	0.668	29.03	0.292	292.8
SD-2V	0.67	29.11	0.295	225.5
Ours	**0.720**	**29.51**	**0.289**	**171.1**

Table 3: Quantitative Comparison on Ted-Talk Dataset

Method	**SSIM ↑**	**PSNR ↑**	**LPIPS ↓**	**FVD ↓**
MRAA [31]	0.826	33.86	0.16	82.8
TPSMM [58]	0.83	33.81	0.157	80.7
Disco [43]	0.754	31.25	0.193	223.5
SD-2V	0.773	32.11	0.179	158.3
Ours	**0.829**	**33.89**	**0.157**	**80.6**

Table 4: Quantitative Comparison for Image Condition Modeling

Method	**SSIM ↑**	**PSNR ↑**	**LPIPS ↓**	**FVD ↓**
CLIP Image Encoder Only	0.89	36.29	0.105	173.2
Fine-tuning SD and Training ControlNet	0.902	37.12	0.098	161.3
Integration of Both	0.931	38.49	0.044	81.6
ReferenceNet (Our Design)	**0.930**	**38.48**	**0.043**	**81.4**

Table 5: Quantitative Results for ReferenceNet Design

Method	**SSIM ↑**	**PSNR ↑**	**LPIPS ↓**	**FVD ↓**
Replacing UNet (SD Weights) with ResNet (ImageNet Weights)	0.897	36.79	0.11	189.2
Replacing Spatial-Attention with Feature Concatenation	0.912	37.85	0.089	134.1

Our Design (ReferenceNet)	**0.929**	**38.61**	**0.067**	**83.6**

Table 6: Quantitative Results for Temporal Modeling

Method	**SSIM ↑**	**PSNR ↑**	**LPIPS ↓**	**FVD ↓**
No Temporal Layer (Concatenating Images Temporally)	0.876	35.71	0.128	275.4
No Two-Stage Training	0.902	37.11	0.101	204.5
Two-Stage Training (Our Approach)	**0.920**	**38.01**	**0.067**	**83.6**

4. Experimental result verification

4.1. Implementations

To evaluate the effectiveness of our approach for animating various characters, we collected a dataset of 5,000 character video clips from the internet for model training. We used DWPose [55] to extract pose sequences of characters in the video, including both body and hand poses, and rendered these as pose skeleton images following the OpenPose method [6].

The experiments were conducted on 4 NVIDIA A100 GPUs. In the first training stage, individual video frames were sampled, resized, and center-cropped to a resolution of 768×768 pixels. The model was trained for 30,000 steps using a batch size of 64. In the second training stage, the temporal layer was trained for 10,000 steps using 24-frame video sequences with a batch size of 4. Both stages employed a learning rate of 1e-5.

During inference, the length of the driving pose skeleton was rescaled to approximate the length of the character's skeleton in the reference image. A DDIM sampler was used for 20 denoising steps. To generate long video sequences, we adopted the temporal aggregation method from [43], which connects results from different batches.

For a fair comparison with other methods, we also trained our model on three established benchmarks: the UBC fashion video dataset [59], the TikTok dataset [20], and the Ted-Talk dataset [39], without using additional data, as discussed in Section

4.2. Qualitative Results

Figure 3 demonstrates that our method can animate a variety of character types, including full-body human figures, half-length portraits, cartoon characters, and humanoid figures. The approach is capable of generating high-definition, realistic character details while maintaining temporal consistency with the reference images, even under significant motion. Additionally, the model exhibits smooth temporal continuity between frames, effectively preserving the dynamics of character movements.

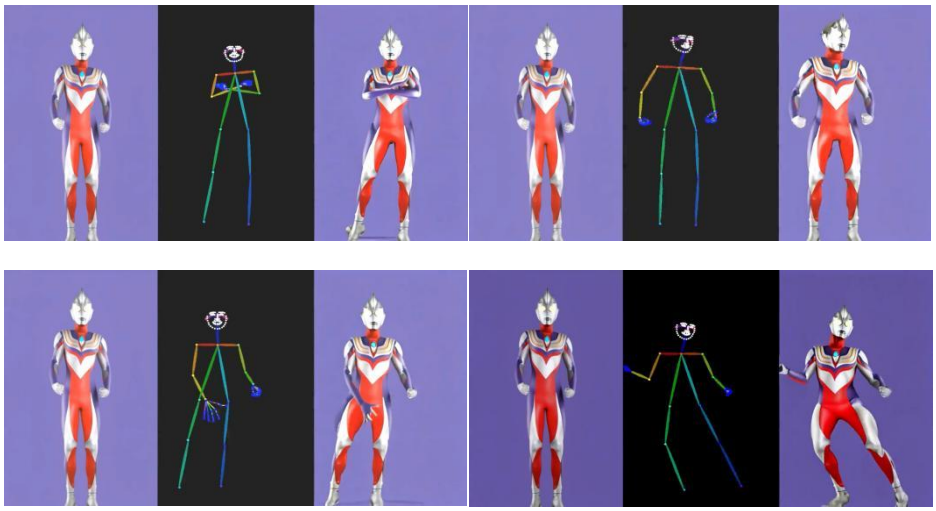

Figure 3. Qualitative Results. Given a reference image (the leftmost image in each group), our approach demonstrates the capability to animate a variety of character types, including full-body human figures, half-length portraits, cartoon characters, and humanoid figures. The figure illustrates the results, highlighting clear and consistent character details, as well as continuous motion between frames

Figure 4. Qualitative Comparison for Fashion Video Synthesis. While other methods struggle to preserve the fine-textured details of clothing, our approach stands out by effectively maintaining high-quality, detailed features throughout the video.

Figure 5. Qualitative Comparison between DisCo and Our Method. DisCo shows issues such as errors in pose control, color inconsistencies, and loss of detail. In contrast, our method significantly improves upon these aspects, delivering more accurate and consistent results.

Figure 6. Qualitative Comparison on the Ted-Talk Dataset. Our model generates more accurate and clearer results, demonstrating superior performance compared to other methods.

Figure 7. Qualitative Comparison with Image-to-Video Methods. These methods face difficulties in generating substantial character movements and struggle with maintaining long-term appearance consistency.

4.3. Comparisons

To evaluate the performance of our approach, we conducted quantitative comparisons using three specific benchmarks: fashion video synthesis, human dance generation, and talking gesture generation. We also compared our method against a baseline that integrates Stable Diffusion, ControlNet, IP-Adapter [56], and AnimateDiff, referred to as SD-2V.

Fashion Video Synthesis

Experiments were conducted on the UBC fashion video dataset, and the quantitative comparison is presented in Table 1. Our method outperforms other approaches, especially in terms of video metrics. Notably, our method shows significant improvements in terms of SSIM, PSNR, LPIPS, and FVD. Qualitative results, shown in Figure 4, highlight that other methods struggle to maintain the consistency of clothing details, resulting in noticeable errors in color and structure. In contrast, our approach effectively preserves fine-grained clothing details, which is crucial in fashion video synthesis.

Human Dance Generation

For human dance generation, we conducted experiments on the TikTok dataset. Table 2 presents the quantitative comparison, where our method demonstrates superior performance in terms of SSIM, PSNR, LPIPS, and FVD. This demonstrates the model's ability to generalize well, even without incorporating pre-trained human attribute data, unlike DisCo, which uses a large set of images for pre-training.

Qualitative comparisons are shown in Figure 5, where our method excels at maintaining visual continuity in intricate dance sequences and handles diverse character appearances with greater robustness.

Talking Gesture Generation

We evaluated our method on the Ted-Talk dataset for talking gesture generation. As shown in Figure 6 and Table 3, our method significantly outperforms DisCo and SD-2V, achieving better results using only pose information. In contrast, other methods relying on ground truth (GT) images as driving signals (such as MRAA and TPSMM) perform worse, especially on more intricate datasets like UBC (with detailed clothing textures) and TikTok (with complex human movements).

General Image-to-Video Methods

To assess the ability of general image-to-video methods, we compared our approach with two popular methods: AnimateDiff [12] and Gen2 [10]. These methods do not use pose control, so the comparison focuses on their ability to maintain the appearance fidelity of reference images. As shown in Figure 7, both AnimateDiff and Gen2 face difficulties in generating substantial character movements and maintaining long-term appearance consistency across frames, which limits their ability to support consistent character animation over time.

4.4. Ablation Study

To investigate the effectiveness of specific design choices in our method, we conducted an ablation study on the UBC fashion video dataset, exploring the following alternatives:

1.Using only the CLIP image encoder to represent reference image features, without integrating ReferenceNet.

2.First fine-tuning Stable Diffusion and then training ControlNet with the reference image.

3.Combining the above two approaches.

The results, presented in Figure 8 and Table 4, demonstrate that the ReferenceNet design outperforms the alternatives. Relying solely on CLIP features for reference image representation preserves image similarity but fails to transfer fine details effectively. Meanwhile, ControlNet alone does not enhance the results, as its features lack the necessary spatial correspondence, rendering it ineffective.

We also conducted experiments to evaluate the effectiveness of the ReferenceNet design by replacing UNet (SD weights) with ResNet (ImageNet weights), and replacing spatial-attention with feature concatenation. The results, shown in Table 5, demonstrate that utilizing SD weights and spatial-attention is crucial for optimal performance, as these elements improve the integration of conditioning information during the generation process.

In terms of temporal modeling, we assessed two alternatives: 1) omitting the temporal layer and concatenating images temporally to create videos, and 2) skipping the two-stage training process and training the entire network simultaneously. The quantitative results in Table 6 indicate that omitting the temporal layer results in noticeable texture sticking and inter-frame jitter, which significantly reduces FVD scores. Additionally, skipping the two-stage training process leads to a decline in image quality, as the network tends to focus on overall temporal coherence at the expense of fine details in individual frames. The two-stage training method ensures both high-quality video frames and temporal smoothness.

5. Discussion and Conclusion

The proposed model, Animate Anyone, demonstrates the ability to transform static character images into animated videos guided by specific pose sequences. While effective, it faces limitations such as challenges in stabilizing hand movements, occasional distortions or motion blur, and difficulties in generating unseen parts of characters due to the single-view nature of input images. Additionally, the utilization of DDPM introduces lower operational efficiency compared to non-diffusion-based methods. Despite these limitations, the framework's ReferenceNet ensures intricate detail preservation, efficient pose control, and temporal continuity, outperforming existing approaches. However, its potential misuse for creating manipulated videos raises ethical concerns, which can be mitigated through face anti-spoofing techniques.

References:

[1] Martin Arjovsky, Soumith Chintala, and Leon Bottou. Wasserstein generative adversarial networks. In International conference on machine learning, pages 214–223.PMLR, 2017. 1

[2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat,Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila,Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-imagediffusion models with an ensemble of expert denoisers. arXivpreprint arXiv:2211.01324, 2022. 2

[3] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal,Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah,and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. In Proceedings of the IEEE/CVFConference on Computer Vision and Pattern Recognition, pages 5968–5976, 2023. 3

[4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis.Align your latents: High-resolution video synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 22563–22575, 2023. 2, 3

[5] Zinelabidine Boulkenafet, Jukka Komulainen, and AbdenourHadid. Face anti-spoofing based on color texture analysis.In 2015 IEEE international conference on image processing(ICIP), pages 2636–2640. IEEE, 2015. 8

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh.Realtime multi-person 2d pose estimation using part affinityfields. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7291–7299, 2017. 5

[7] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei AEfros. Everybody dance now. In Proceedings of theIEEE/CVF international conference on computer vision,pages 5933–5942, 2019. 1, 3

[8] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang,Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu,Qifeng Chen, Xintao Wang, et al. Videocrafter1: Opendiffusion models for high-quality video generation. arXivpreprint arXiv:2310.19512, 2023. 2, 3, 4

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov,Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner,Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An imageis worth 16x16 words: Transformers for image recognitionat scale. In 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021. 4

[10] Patrick Esser, Johnathan Chiu, Parmida Atighehchian,Jonathan Granskog, and Anastasis Germanidis. Structureand content-guided video synthesis with diffusion models.In Proceedings of the IEEE/CVF International Conferenceon Computer Vision, pages 7346–7356, 2023. 2, 7

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, BingXu, David Warde-Farley, Sherjil Ozair, Aaron Courville, andYoshua Bengio. Generative adversarial nets. Advances inneural information processing systems, 27, 2014. 1

[12] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, YuQiao, Dahua Lin, and Bo Dai. Animatediff: Animate yourpersonalized text-to-image diffusion models without specifictuning. arXiv preprint arXiv:2307.04725, 2023. 2, 3, 5, 7

[13] Hsuan-I Ho, Lixin Xue, Jie Song, and Otmar Hilliges. Learning locally editable virtual humans. In Proceedings ofthe IEEE/CVF Conference on Computer Vision and PatternRecognition, pages 21024–21035, 2023. 3

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. Advances in neural informationprocessing systems, 33:6840–6851, 2020. 2, 4

[15] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang,Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, BenPoole, Mohammad Norouzi, David J Fleet, et al. Imagenvideo: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303, 2022. 2

[16] Jonathan Ho, Tim Salimans, Alexey A. Gritsenko, WilliamChan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In NeurIPS, 2022. 2

[17] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, andJie Tang. Cogvideo: Large-scale pretraining for text-to-videogeneration via transformers. In The Eleventh InternationalConference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net, 2023. 2

[18] Alain Hore and Djemel Ziou. Image quality metrics: Psnrvs. ssim. In 2010 20th international conference on patternrecognition, pages 2366–2369. IEEE, 2010. 5

[19] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao,and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. In InternationalConference on Machine Learning, 2023. 2

[20] Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social mediadance videos. In Proceedings of the IEEE/CVF Conferenceon Computer Vision

and Pattern Recognition, pages 12753–12762, 2021. 2, 5
[21] Johanna Karras, Aleksander Holynski, Ting-Chun Wang,and Ira Kemelmacher-Shlizerman. Dreampose: Fashionvideo synthesis with stable diffusion. In Proceedings of the IEEE/CVF International Conference on Computer Vision,pages 22680–22690, 2023. 2, 3, 4, 6
[22] Tero Karras, Samuli Laine, and Timo Aila. A style-basedgenerator architecture for generative adversarial networks.In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 4401–4410, 2019. 1
[23] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, ShantNavasardyan, and Humphrey Shi. Text2video-zero: Textto-image diffusion models are zero-shot video generators. InProceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 15954–15964, 2023. 2
[24] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In 2nd International Conference on LearningRepresentations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, 2014. 2, 4
[25] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, andShenghua Gao. Liquid warping gan: A unified frameworkfor human motion imitation, appearance transfer and novel view synthesis. In Proceedings of the IEEE/CVF international conference on computer vision, pages 5904–5913, 2019. 3
[26] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. Follow your pose:Pose-guided text-to-video generation using pose-free videos.arXiv preprint arXiv:2304.01186, 2023. 2
[27] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learningadapters to dig out more controllable ability for text-to-imagediffusion models. arXiv preprint arXiv:2302.08453, 2023. 2
[28] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, andMartin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18444–18455, 2023. 3
[29] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In Interna tional Conference on Machine Learning, 2021. 2
[30] Chenyang QI, Xiaodong Cun, Yong Zhang, Chenyang Lei,Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 15932–15942, 2023. 2
[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PMLR, 2021. 2
[32] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu,and Mark Chen. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 1(2):3, 2022. 2