Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Character Animation

Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, Liefeng Bo
Institute for Intelligent Computing, Alibaba Group
{hooks.hl, zimu.gx, futian.zp, xisheng.sk, zhangbang.zb, liefeng.bo}@alibaba-inc.com
https://humanaigc.github.io/animate-anyone/

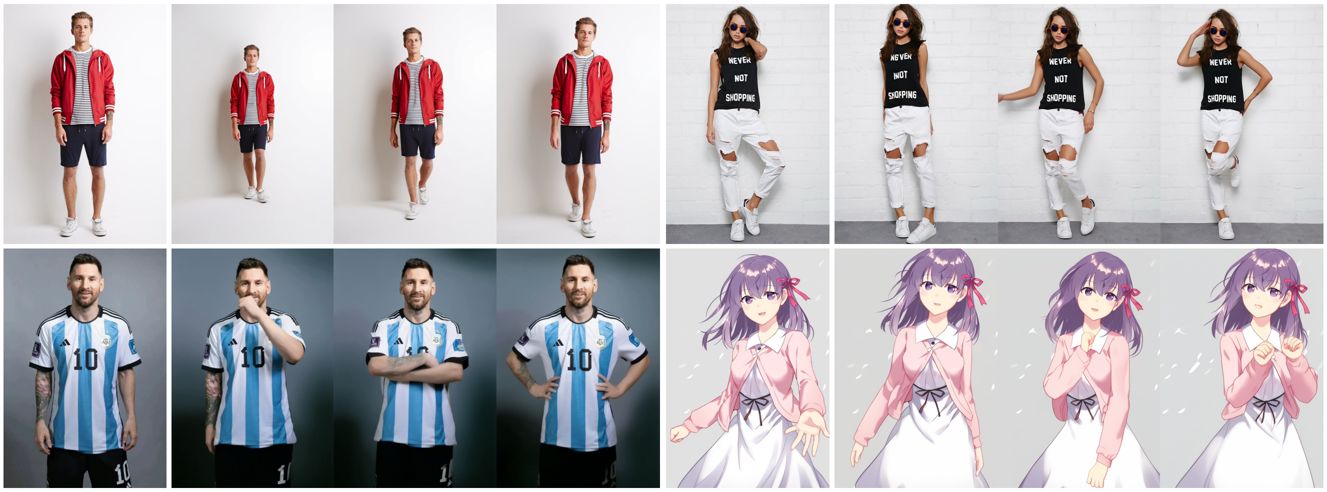

Figure 1. Consistent and controllable character animation results given reference image (the leftmost image in each group) . Our approach is capable of animating arbitrary characters, generating clear and temporally stable video results while maintaining consistency with the appearance details of the reference character.

Abstract

Character Animation aims to generating character videos from still images through driving signals. Currently, diffusion models have become the mainstream in visual generation research, owing to their robust generative capabilities. However, challenges persist in the realm of image-to-video, especially in character animation, where temporally maintaining consistency with detailed information from character remains a formidable problem. In this paper, we leverage the power of diffusion models and propose a novel framework tailored for character animation. To preserve consistency of intricate appearance features from reference image, we design ReferenceNet to merge detail features via spatial attention. To ensure controllability and continuity, we introduce an efficient pose guider to direct character's movements and employ an effective temporal modeling approach to ensure smooth inter-frame transitions between video frames. By expanding the training data, our approach can animate arbitrary characters, yielding superior results in character animation compared to other image-to-video methods. Furthermore, we evaluate our method on image animation benchmarks, achieving state-of-the-art results.

1. Introduction

Character Animation is a task to animate source character images into realistic videos according to desired posture sequences, which has many potential applications such as online retail, entertainment videos, artistic creation and virtual character. Beginning with the advent of GANs[1, 11, 22], numerous studies have delved into the realms of image animation and pose transfer[7, 33, 37–39, 57, 61, 64]. However, the generated images or videos still exhibit issues such as local distortion, blurred details, semantic inconsistency, and temporal instability, which impede the widespread application of these methods.

In recent years, diffusion models[14] have showcased their superiority in producing high-quality images and videos. Researchers have begun exploring human image-to-video tasks by leveraging the architecture of diffusion models and their pretrained robust generative capabilities. DreamPose[21] focuses on fashion image-to-video synthesis, extending Stable Diffusion[34] and proposing an adaptar module to integrate CLIP[31] and VAE[24] features from images. However, DreamPose requires finetuning on input samples to ensure consistent results, leading to suboptimal operational efficiency. DisCo[47] explores human dance generation, similarly modifying Stable Diffusion, integrating character features through CLIP, and incorporating background features through ControlNet[60]. However, it exhibits deficiencies in preserving character details and suffers from inter-frame jitter issues.

Furthermore, current research on character animation predominantly focuses on specific tasks and benchmarks, resulting in a limited generalization capability. Recently, benefiting from advancements in text-to-image research[2, 19, 29, 32, 34, 36], video generation (e.g., text-to-video, video editing)[4, 10, 12, 15–17, 23, 30, 40, 48, 52] has also achieved notable progress in terms of visual quality and diversity. Several studies extend text-to-video methodologies to image-to-video[8, 12, 48, 63]. However, these methods fall short of capturing intricate details from images, providing more diversity but lacking precision, particularly when applied to character animation, leading to temporal variations in the fine-grained details of the character's appearance. Moreover, when dealing with substantial character movements, these approaches struggle to generate a consistently stable and continuous process. Currently, there is no observed character animation method that simultaneously achieves generalizability and consistency.

In this paper, we present *Animate Anyone*, a method capable of transforming character images into animated videos controlled by desired pose sequences. We inherit the network design and pretrained weights from Stable Diffusion (SD) and modify the denoising UNet[35] to accommodate multi-frame inputs. To address the challenge of maintaining appearance consistency, we introduce ReferenceNet, specifically designed as a symmetrical UNet structure to capture spatial details of the reference image. At each corresponding layer of the UNet blocks, we integrate features from ReferenceNet into the denoising UNet using spatial-attention[46]. This architecture enables the model to comprehensively learn the relationship with the reference image in a consistent feature space, which significantly contributes to the improvement of appearance details preservation. To ensure pose controllability, we devise a lightweight pose guider to efficiently integrate pose control signals into the denoising process. For temporal stability, we introduce temporal layer to model relationships across multiple frames, which preserves high-resolution details in visual quality while simulating a continuous and smooth temporal motion process.

Our model is trained on an internal dataset of 5K character video clips. Fig. 1 shows the animation results for various characters. Compared to previous methods, our approach presents several notable advantages. Firstly, it effectively maintains the spatial and temporal consistency of character appearance in videos. Secondly, it produces high-definition videos without issues such as temporal jitter or flickering. Thirdly, it is capable of animating any character image into a video, unconstrained by specific domains. We evaluate our method on three specific human video synthesis benchmarks (UBC fashion video dataset[59], TikTok dataset[20] and Ted-talk dataset[39]), using only the corresponding training datasets for each benchmark in the experiments. Our approach achieves state-of-the-art results. We also compare our method with general image-to-video approaches trained on large-scale data and our approach demonstrates superior capabilities in character animation. We envision that *Animate Anyone* could serve as a foundational solution for character video creation, inspiring the development of more innovative and creative applications.

2. Related Works

2.1. Diffusion Model for Image Generation

In text-to-image research, diffusion-based methods[2, 19, 29, 32, 34, 36] have achieved significantly superior generation results, becoming the mainstream of research. To reduce computational complexity, Latent Diffusion Model[34] proposes denoising in the latent space, striking a balance between effectiveness and efficiency. ControlNet[60] and T2I-Adapter[27] delve into the controllability of visual generation by incorporating additional encoding layers, facilitating controlled generation under various conditions such as pose, mask, edge and depth. Some studies further investigate image generation under given image conditions. IP-Adapter[56] enables diffusion models to generate image results that incorporate the content specified by a given image prompt. ObjectStitch[42] and Paint-by-Example[53] leverage the CLIP[31] and propose diffusion-based image editing methods given image condition. TryonDiffusion[65] applies diffusion models to the virtual apparel try-on task and introduces the Parallel-UNet structure.

2.2. Diffusion Model for Video Generation

With the success of diffusion models in text-to-image applications, research in text-to-video has extensively drawn inspiration from text-to-image models in terms of model structure. Many studies[10, 16, 17, 23, 26, 30, 40, 52, 54] explore the augmentation of inter-frame attention modeling