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

Supplementary Material

1. Video Results

We provide a video demo to show the results of our method, including:

Comparisons with Other Methods. We present visual comparisons of the video generation results on: 1) Fashion Video Synthesis task between our method and DreamPose[2], BDMM[6]. 2) Human Dance Generation task between our method with DisCo[5].

Character Animation for Various Cahracters. We present the animation results of our method given various character images, including full-body human figures, half-length portraits, cartoon characters, and humanoid figures. The video results demonstrate that our method achieves excellent clarity, appearance consistency and temporal stability in the generated videos.

2. Details of the Network

As shown in Fig. 1, we adopt the UNet structure from Stable Diffusion[3] to construct our network architecture. ReferenceNet employs a similar architecture to the denoising UNet. The Res-Trans Block comprises ResNet layers and Transformer layers. The Res Block includes only ResNet layers. The Transformer layer in ReferenceNet contains both self-attention and cross-attention, while in the denoising UNet, the self-attention is replaced with spatial-attention and augmented with temporal-attention.

3. The Usage of Pose Sequence

During training, we extract the pose sequence of the character from video frames and render it into spatially aligned images. In conventional image animation benchmarks[1, 7], the pose in the test set is also extracted from character videos, ensuring spatial alignment. However, in more generalized scenarios where a corresponding skeleton sequence for the reference character is not provided, the source pose and the target character may not be perfectly aligned (different skeletal proportions, positions, and resolutions). In such cases, we employ an approximate retargeting operation, as illustrated in Fig. 2. Firstly, we rescale the bone lengths of the source pose, following a general principle where we compute the length ratios for each limb bone between reference character and source pose. Then, using the shoulder center as the base point, we recalculate the new positions of target skeleton points based on the scaling factors. Finally, we align the entire skeleton to the same position as the character in the reference image

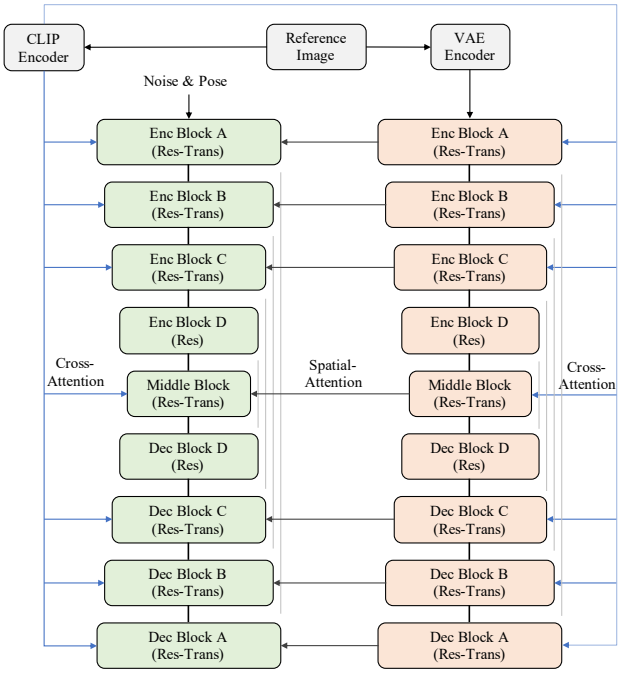

Figure 1. Detail of the network.

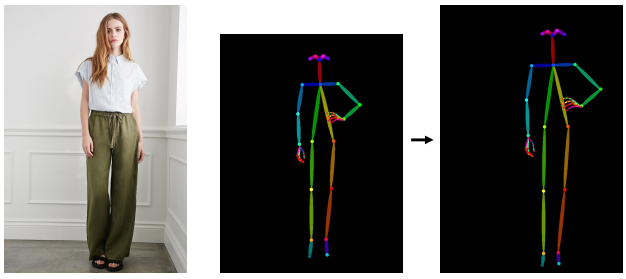

Figure 2. Retargeting pose before inference.

(using the feet as a reference point for full-body characters or the shoulder center for half-body characters).

It is important to mention that this method provides an approximation of the target skeleton and requires the source pose and target character to be in a frontal view with extended limbs. In our paper, as this aspect is not the primary focus of our discussion, we cannot guarantee its applicability to any reference image and source pose. In our testing samples, the method has demonstrated effectiveness. Further research is needed for a more universal solution.

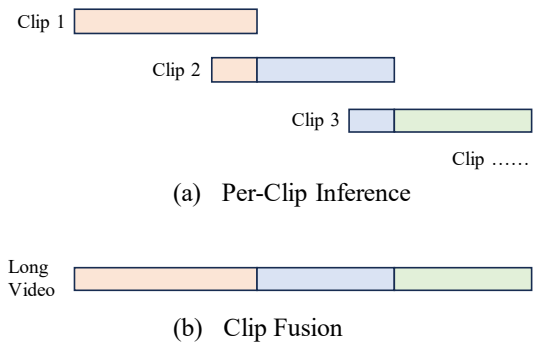

Figure 3. Inter-clip fusion at inference.

4. Long Video Generation

Previous diffusion-based video generation often focused on generating short video clips, resulting in discontinuities when creating generating multiple clips within a single video. Thanks to ReferenceNet and Pose Guider, the model can significantly control the appearance and motion of characters across the generation of multiple clips. Inspired by [4], we introduce inter-clip fusion to ensure continuity of details between clips, shown in Fig. 3. Specifically, we retain the results of each step in the denoising process for the last 4 frames of each video clip. When inferring the next video clip, we select the previously retained 4 frames and the subsequent 20 frames as input. Before each denoising step, we overlay the intermediate results of the last 4 frames from the previous clip onto the current first 4 frames, and so on for the subsequent frames. In this way, our model can generate videos of arbitrary length while maintaining content consistency. It should be noted that this method still cannot entirely eliminate slight jitter between clips, especially in the presence of complex backgrounds and character appearances. Generating long videos is a significant research direction, but it is not the primary focus of our work. Therefore, we directly apply the approach, and a more comprehensive solution awaits further investigation in future research.

References

[1] Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12753–12762, 2021. 1

[2] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion video synthesis with stable diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22680–22690, 2023. 1

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[4] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 448–458, 2023. 2

[5] Tan Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for referring human dance generation in real world. *arXiv preprint arXiv:2307.00040*, 2023. 1

[6] Wing-Yin Yu, Lai-Man Po, Ray CC Cheung, Yuzhi Zhao, Yu Xue, and Kun Li. Bidirectionally deformable motion modulation for video-based human pose transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7502–7512, 2023. 1

[7] Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. Dwnet: Dense warp-based network for pose-guided human video generation. *arXiv preprint arXiv:1910.09139*, 2019. 1