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 benchmarks for fashion video and human dance synthesis, 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, 10, 20], numerous studies have delved into the realms of image animation and pose transfer[6, 31, 35–37, 54, 57, 60]. 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[12] 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[19] focuses on fashion image-to-video synthesis, extending Stable Diffusion[32] and proposing an adaptar module to integrate CLIP[29] and VAE[22] features from images. However, DreamPose requires finetuning on input samples to ensure consistent results, leading to suboptimal operational efficiency. DisCo[45] explores human dance generation, similarly modifying Stable Diffusion, integrating character features through CLIP, and incorporating background features through ControlNet[56]. 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, 17, 27, 30, 32, 34], video generation (e.g., text-to-video, video editing)[4, 9, 11, 13–15, 21, 28, 38, 46, 49] has also achieved notable progress in terms of visual quality and diversity. Several studies extend text-to-video methodologies to image-to-video[7, 11, 46, 59]. 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[33] 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[44]. 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 two specific human video synthesis benchmarks (UBC fashion video dataset[55] and TikTok dataset[18]), 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, 17, 27, 30, 32, 34] have achieved significantly superior generation results, becoming the mainstream of research. To reduce computational complexity, Latent Diffusion Model[32] proposes denoising in the latent space, striking a balance between effectiveness and efficiency. ControlNet[56] and T2I-Adapter[25] 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[53] enables diffusion models to generate image results that incorporate the content specified by a given image prompt. ObjectStitch[40] and Paint-by-Example[50] leverage the CLIP[29] and propose diffusion-based image editing methods given image condition. TryonDiffusion[61] 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[9, 14, 15, 21, 24, 28, 38, 49, 51] explore the augmentation of inter-frame attention modeling

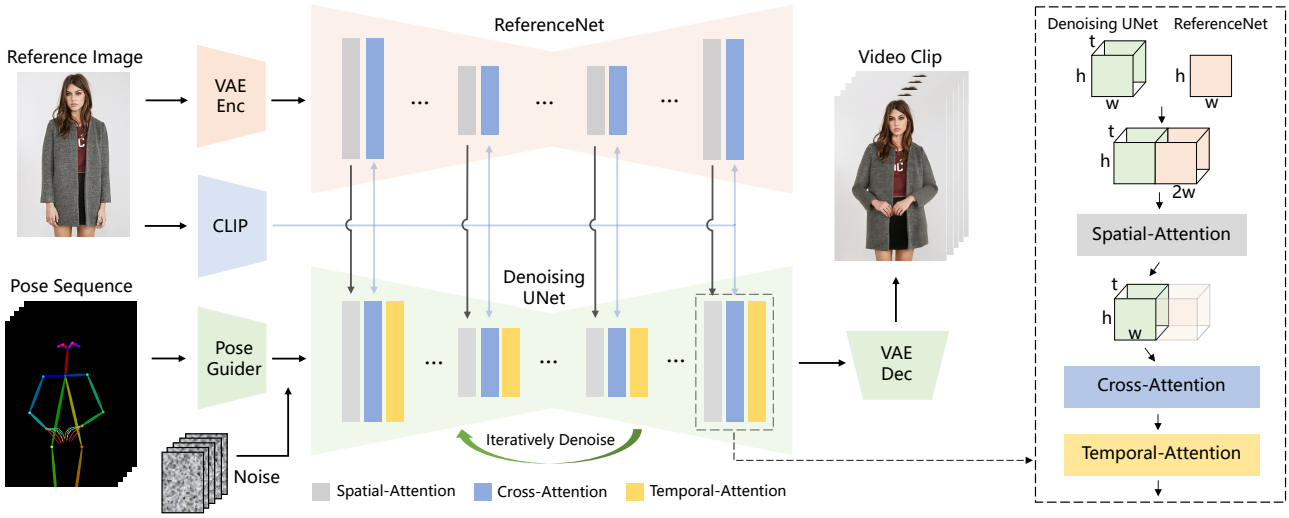

Figure 2. The overview of our method. The pose sequence is initially encoded using Pose Guider and fused with multi-frame noise, followed by the Denoising UNet conducting the denoising process for video generation. The computational block of the Denoising UNet consists of Spatial-Attention, Cross-Attention, and Temporal-Attention, as illustrated in the dashed box on the right. The integration of reference image involves two aspects. Firstly, detailed features are extracted through ReferenceNet and utilized for Spatial-Attention. Secondly, semantic features are extracted through the CLIP image encoder for Cross-Attention. Temporal-Attention operates in the temporal dimension. Finally, the VAE decoder decodes the result into a video clip.

on the foundation of text-to-image (T2I) models to achieve video generation. Some works turn pretrained T2I models into video generators by inserting temporal layers. Video LDM[4] proposes to first pretrain the model on images only and then train temporal layers on videos. AnimateDiff[11] presents a motion module trained on large video data which could be injected into most personalized T2I models without specific tuning. Our approach draws inspiration from such methods for temporal modeling.

Some studies extend text-to-video capabilities to image-to-video. VideoComposer[46] incorporates images into the diffusion input during training as a conditional control. AnimateDiff[11] performs weighted mixing of image latent and random noise during denoising. VideoCrafter[7] incorporates textual and visual features from CLIP as the input for cross-attention. However, these approaches still face challenges in achieving stable human video generation, and the exploration of incorporating image condition input remains an area requiring further investigation.

2.3. Diffusion Model for Human Image Animation

Image animation[6, 31, 35–37, 54, 57, 60] aims to generate images or videos based on one or more input images. In recent research, the superior generation quality and stable controllability offered by diffusion models have led to their integration into human image animation. PIDM[3] proposes texture diffusion blocks to inject the desired texture patterns into denoising for human pose transfer. LFDM[26] synthesizes an optical flow sequence in the latent space, warping the input image based on given conditions. LEO[47] represents motion as a sequence of flow maps and employs diffusion model to synthesize sequences of motion codes. DreamPose[19] utilizes pretrained Stable Diffusion model and proposes an adapter to model the CLIP and VAE image embeddings. DisCo[45] draws inspiration from ControlNet, decoupling the control of pose and background. Despite the incorporation of diffusion models to enhance generation quality, these methods still grapple with issues such as texture inconsistency and temporal instability in their results. Moreover, there is no method to investigate and demonstrate a more generalized capability in character animation.

3. Methods

We target at pose-guided image-to-video synthesis for character animation. Given a reference image describing the appearance of a character and a pose sequence, our model generates an animated video of the character. The pipeline of our method is illustrated in Fig. 2. In this section, we first provide a concise introduction to Stable Diffusion in Sec 3.1, which lays the foundational framework and network structure for our method. Then we provide a detailed explanation of the design specifics in Sec 3.1. Finally, we present the training process in Sec 3.3.

3.1. Preliminariy: Stable Diffusion

Our method is an extension of Stable Diffusion (SD), which is developed from Latent diffusion model (LDM). To

reduce the computational complexity of the model, it introduces to model feature distributions in the latent space. SD develops an autoencoder[22, 43] to establish the implicit representation of images, which consists of an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$. Given an image $\mathbf{x}$, the encoder first maps it to a latent representation: $\mathbf{z} = \mathcal{E}(\mathbf{x})$ and then the decoder reconstructs it: $\mathbf{x}_{recon} = \mathcal{D}(\mathbf{z})$.

SD learns to denoise a normally-distributed noise ϵ to realistic latent $\mathbf{z}$. During training, the image latent $\mathbf{z}$ is diffused in t timesteps to produce noise latent $\mathbf{z}_t$. And a denoising UNet is trained to predict the applied noise. The optimization process is defined as follow objective:

$$\mathbf{L} = \mathbb{E}_{\mathbf{z}_t, c, \epsilon, t}(||\epsilon - \epsilon_\theta(\mathbf{z}_t, c, t)||_2^2) \tag{1}$$

where ϵ_θ represents the function of the denoising UNet. c represents the embeddings of conditional information. In original SD, CLIP ViT-L/14[8] text encoder is applied to represent the text prompt as token embeddings for text-to-image generation. The denoising UNet consists of four downsample layers , one middle layer and four upsample layers. A typical block within a layer includes three types of computations: 2D convolution, self-attention[44], and cross-attention (terms as Res-Trans block). Cross-attention is conducted between text embedding and corresponding network feature.

At inference, $\mathbf{z}_T$ is sampled from random Gaussian distribution with the initial timestep T and is progressively denoised and restored to $\mathbf{z}_0$ via deterministic sampling process (e.g. DDPM[12], DDIM[39]). In each iteration, the denoising UNet predicts the noise on the latent feature corresponding to each timestep t. Finally, $\mathbf{z}_0$ will be reconstructed by decoder $\mathcal{D}$ to obtain the generated image.

3.2. Network Architecture

Overview. Fig. 2 provides the overview of our method. The initial input to the network consists of multi-frame noise. The denoising UNet is configured based on the design of SD, employing the same framework and block units, and inherits the training weights from SD. Additionally, our method incorporates three crucial components: 1) ReferenceNet, encoding the appearance features of the character from the reference image; 2) Pose Guider, encoding motion control signals for achieving controllable character movements; 3) Temporal layer, encoding temporal relationship to ensure the continuity of character motion.

ReferenceNet. In text-to-video tasks, textual prompts articulate high-level semantics, necessitating only semantic relevance with the generated visual content. However, in image-to-video tasks, images encapsulate more low-level detailed features, demanding precise consistency in the generated results. In preceding studies focused on image-driven generation, most approaches[7, 19, 40, 45, 50, 53] employ the CLIP image encoder as a substitute for the text encoder in cross-attention. However, this design falls short of addressing issues related to detail consistency. One reason for this limitation is that the input to the CLIP image encoder comprises low-resolution (224×224) images, resulting in the loss of significant fine-grained detail information. Another factor is that CLIP is trained to match semantic features for text, emphasizing high-level feature matching, thereby leading to a deficit in detailed features within the feature encoding.

Hence, we devise a reference image feature extraction network named ReferenceNet. We adopt a framework identical to the denoising UNet for ReferenceNet, excluding the temporal layer. Similar to the denoising UNet, ReferenceNet inherits weights from the original SD, and weight update is conducted independently for each. Then we explain the integration method of features from ReferenceNet into the denoising UNet. Specifically, as shown in Fig. 2, we replace the self-attention layer with spatial-attention layer. Given a feature map $x_1 \in \mathbb{R}^{t\times h\times w\times c}$ from denoising UNet and $x_2 \in \mathbb{R}^{h\times w\times c}$ from ReferenceNet, we first copy x_2 by t times and concatenate it with x_1 along w dimension. Then we perform self-attention and extract the first half of the feature map as the output. This design offers two advantages: Firstly, ReferenceNet can leverage the pre-trained image feature modeling capabilities from the original SD, resulting in a well-initialized feature. Secondly, due to the essentially identical network structure and shared initialization weights between ReferenceNet and the denoising UNet, the denoising UNet can selectively learn features from ReferenceNet that are correlated in the same feature space. Additionally, cross-attention is employed using the CLIP image encoder. Leveraging the shared feature space with the text encoder, it provides semantic features of the reference image, serving as a beneficial initialization to expedite the entire network training process.

A comparable design is ControlNet[56], which introduces additional control features into the denoising UNet using zero convolution. However, control information, such as depth and edge, is spatially aligned with the target image, while the reference image and the target image are spatially related but not aligned. Consequently, ControlNet is not suitable for direct application. We will substantiate this in the subsequent experimental Section 4.4.

While ReferenceNet introduces a comparable number of parameters to the denoising UNet, in diffusion-based video generation, all video frames undergo denoising multiple times, whereas ReferenceNet only needs to extract features once throughout the entire process. Consequently, during inference, it does not lead to a substantial increase in computational overhead.

Pose Guider. ControlNet[56] demonstrates highly robust conditional generation capabilities beyond text. Different from these methods, as the denoising UNet needs to be fine-

tuned, we choose not to incorporate an additional control network to prevent a significant increase in computational complexity. Instead, we employ a lightweight Pose Guider. This Pose Guider utilizes four convolution layers (4×4 kernels, 2×2 strides, using 16,32,64,128 channels, similar to the condition encoder in [56]) to align the pose image with the same resolution as the noise latent. Subsequently, the processed pose image is added to the noise latent before being input into the denoising UNet. The Pose Guider is initialized with Gaussian weights, and in the final projection layer, we employ zero convolution.

Temporal Layer. Numerous studies have suggested incorporating supplementary temporal layers into text-to-image (T2I) models to capture the temporal dependencies among video frames. This design facilitates the transfer of pretrained image generation capabilities from the base T2I model. Adhering to this principle, our temporal layer is integrated after the spatial-attention and cross-attention components within the Res-Trans block. The design of the temporal layer was inspired by AnimateDiff[11]. Specifically, for a feature map $x\in\mathbb{R}^{b\times t\times h\times w\times c}$, we first reshape it to $x\in\mathbb{R}^{(b\times h\times w)\times t\times c}$, and then perform temporal attention, which refers to self-attention along the dimension t. The feature from temporal layer is incorporated into the original feature through a residual connection. This design aligns with the two-stage training approach that we will describe in the following subsection. The temporal layer is exclusively applied within the Res-Trans blocks of the denoising UNet. For ReferenceNet, it computes features for a single reference image and does not engage in temporal modeling. Due to the controllability of continuous character movement achieved by the Pose Guider, experiments demonstrate that the temporal layer ensures temporal smoothness and continuity of appearance details, obviating the need for intricate motion modeling.

3.3. Training Strategy

The training process is divided into two stages. In the first stage, training is performed using individual video frames. Within the denoising UNet, we temporarily exclude the temporal layer and the model takes single-frame noise as input. The ReferenceNet and Pose Guider are also trained during this stage. The reference image is randomly selected from the entire video clip. We initialize the model of the denoising UNet and ReferenceNet based on the pretrained weights from SD. The Pose Guider is initialized using Gaussian weights, except for the final projection layer, which utilizes zero convolution. The weights of the VAE's Encoder and Decoder, as well as the CLIP image encoder, are all kept fixed. The optimization objective in this stage is to enable the model to generate high-quality animated images under the condition of a given reference image and target pose. In the second stage, we introduce the temporal layer into the previously trained model and initialize it using pretrained weights from AnimateDiff[11]. The input for the model consists of a 24-frames video clip. During this stage, we only train the temporal layer while fixing the weights of the rest of the network.

4. Experiments

4.1. Implementations

To demonstrate the applicability of our approach in animating various characters, we collect 5K character video clips (2-10 seconds long) from the internet to train our model. We employ DWPose[52] to extract the pose sequence of characters in the video, including the body and hands, rendering it as pose skeleton images following OpenPose[5]. Experiments are conducted on 4 NVIDIA A100 GPUs. In the first training stage, individual video frames are sampled, resized, and center-cropped to a resolution of 768×768. Training is conducted for 30,000 steps with a batch size of 64. In the second training stage, we train the temporal layer for 10,000 steps with 24-frame video sequences and a batch size of 4. Both learning rates are set to 1e-5. During inference, we rescale the length of the driving pose skeleton to approximate the length of the character's skeleton in the reference image and use a DDIM sampler for 20 denoising steps. We adopt the temporal aggregation method in [41], connecting results from different batches to generate long videos. For fair comparison with other image animation methods, we also train our model on two specific benchmarks (UBC fashion video dataset[55] and TikTok dataset[18]) without using additional data, as will be discussed in Section 4.3.

4.2. Qualitative Results

Fig. 3 demonstrates that our method can animate arbitrary characters, including full-body human figures, half-length portraits, cartoon characters, and humanoid characters. Our approach is capable of generating high-definition and realistic character details. It maintains temporal consistency with the reference images even under substantial motion and exhibits temporal continuity between frames. More video results are available in the supplementary material.

4.3. Comparisons

To demonstrate the effectiveness of our approach compared to other image animation methods, we evaluate its performance in two specific benchmarks: fashion video synthesis and human dance generation. For quantitative assessment of image-level quality, SSIM[48], PSNR[16] and LPIPS[58] are employed. Video-level evaluation uses FVD[42] metrics.

Fashion Video Synthesis. Fashion Video Synthesis aims to turn fashion photographs into realistic, animated videos

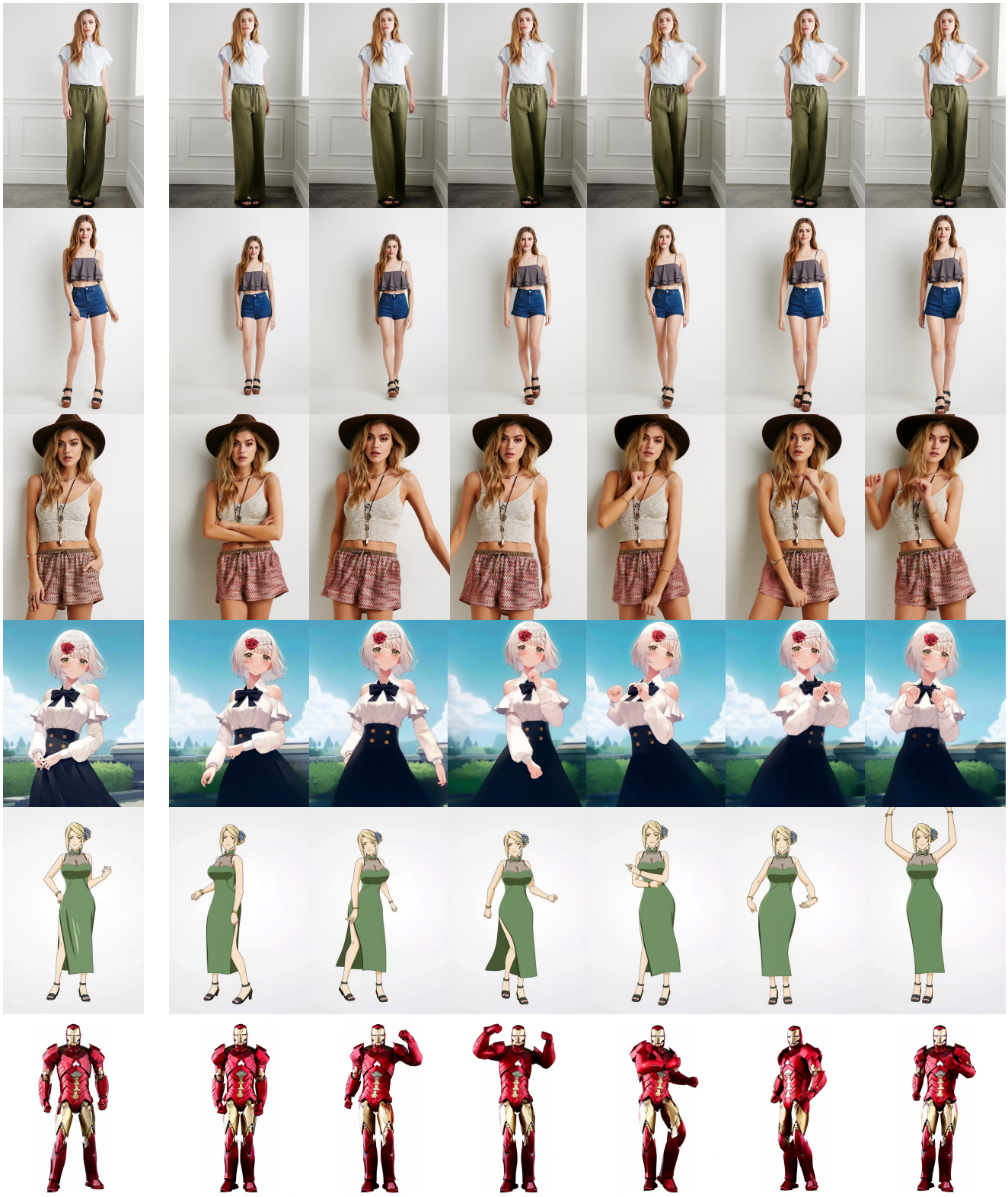

Figure 3. Qualitative Results. Given a reference image (the leftmost image), our approach demonstrates the ability to animate diverse characters, encompassing full-body human figures, half-length portraits, cartoon characters, and humanoid figures. The illustration showcases results with clear, consistent details, and continuous motion. More video results are available in supplementary material.

Figure 4. Qualitative comparison for fashion video synthesis. DreamPose and BDMM exhibit shortcomings in preserving fine-textured details of clothing, whereas our method excels in maintaining exceptional detail features. More comparison could be found in supplementary material.

	SSIM ↑	PSNR ↑	LPIPS ↓	FVD ↓
MRAA[37]	0.749	-	0.212	253.6
TPSMM[60]	0.746	-	0.213	247.5
BDMM[54]	0.918	24.07	0.048	148.3
DreamPose[19]	0.885	-	0.068	238.7
DreamPose*	0.879	34.75	0.111	279.6
Ours	**0.931**	**38.49**	**0.044**	**81.6**

Table 1. Quantitative comparison for fashion video synthesis. "Dreampose*" denotes the result without sample finetuning.

using a driving pose sequence. Experiments are conducted on the UBC fashion video dataset, which consists of 500 training and 100 testing videos, each containing roughly 350 frames. The quantitative comparison is shown in Tab. 1. Our result outperforms other methods, particularly exhibiting a significant lead in video metric. Qualitative comparison is shown in Fig. 4. For fair comparison, we obtain results of DreamPose without sample finetuning using its open-source code. In the domain of fashion videos, there is a stringent requirement for fine-grained clothing details. However, the videos generated by DreamPose and BDMM fail to maintain the consistency of clothing details and exhibit noticeable errors in terms of color and fine structural elements. In contrast, our method produces results that effectively preserves the consistency of clothing details.

Human Dance Generation. Human Dance Generation

Figure 5. Qualitative comparison between DisCo and our method. DisCo's generated results display problems such as pose control errors, color inaccuracy, and inconsistent details. In contrast, our method demonstrates significant improvements in addressing these issues.

	SSIM ↑	PSNR ↑	LPIPS ↓	FVD ↓
FOMM[35]	0.648	29.01	0.335	405.2
MRAA[37]	0.672	29.39	0.296	284.8
TPSMM[60]	0.673	29.18	0.299	306.1
Disco[45]	0.668	29.03	0.292	292.8
Ours	**0.718**	**29.56**	**0.285**	**171.9**

Table 2. Quantitative comparison for human dance generation.

focuses on animating images in real-world dance scenarios. We utilize the TikTok dataset, comprising 340 training and 100 testing single human dancing videos (10-15 seconds long). Following DisCo's dataset partitioning and utilizing an identical test set (10 TikTok-style videos), we conduct a quantitative comparison presented in Tab. 2, and our method achieves the best results. For enhanced generalization, DisCo incorporates human attribute pre-training, utilizing a large number of image pairs for model pre-training. In contrast, our training is exclusively conducted on the TikTok dataset, yielding results superior to DisCo. We present qualitative comparison with DisCo in Fig. 5. Given the scene's complexity, DisCo's pipeline involves an extra step with SAM[23] to generate the human foreground mask. Conversely, our approach showcases that, even without explicit learning of human mask, the model can grasp the foreground-background relationship from the subject's motion, negating the necessity for prior human segmentation. Additionally, during intricate dance sequences, our model stands out in maintaining visual continuity throughout the motion and exhibits enhanced robustness in handling diverse character appearances.

General Image-to-Video Methods. Currently, numerous

Figure 6. Qualitative comparison with image-to-video methods. Current image-to-video methods struggle to generate substantial character movements and face challenges in maintaining long-term consistency with character image features.

studies propose video diffusion models with strong generative capabilities based on large-scale training data. We select two of the most well-known and effective image-to-video methods for comparison: AnimateDiff[11] and Gen-2[9]. As these two methods do not perform pose control, we only compare their ability to maintain the appearance fidelity to the reference image. As depicted in Fig. 6, current image-to-video methods face challenges in generating substantial character movements and struggle to maintain long-term appearance consistency in videos, thus hindering effective support for consistent character animation.

4.4. Ablation study

To demonstrate the effectiveness of the ReferenceNet design, we explore alternative designs, including (1) using only the CLIP image encoder to represent reference image features without integrating ReferenceNet, (2) initially fine-tuning SD and subsequently training ControlNet with the reference image. (3) integrating the above two designs. Experiments are conducted on the UBC fashion video dataset. As shown in Fig. 7, visualizations illustrate that ReferenceNet outperforms the other three designs. Solely relying on CLIP features as reference image features can preserve image similarity but fails to fully transfer details. ControlNet does not enhance results as its features lack spatial correspondence, rendering it inapplicable. Quantitative results are also presented in Tab. 3, demonstrating the superiority of our design.

Figure 7. Ablation study of different design. Only our ReferenceNet ensures the consistent preservation of details in the character's appearance.

	SSIM ↑	PSNR ↑	LPIPS ↓	FVD ↓
CLIP	0.897	36.09	0.089	208.5
ControlNet	0.892	35.89	0.105	213.9
CLIP+ControlNet	0.898	36.03	0.086	205.4
Ours	**0.931**	**38.49**	**0.044**	**81.6**

Table 3. Quantitative comparison for ablation study.

5. Limitations

The limitations of our method can be summarized in three aspects: First, similar to many visual generation models, our model may struggle to generate highly stable results for hand movements, sometimes leading to distortions and motion blur. Second, since images provide information from only one perspective, generating unseen parts during character movement is an ill-posed problem which encounters potential instability. Third, due to the utilization of DDPM, our model exhibits a lower operational efficiency compared to non-diffusion-model-based methods.

6. Conclusion

In this paper, we present *Animate Anyone*, a character animation framework capable of transforming character photographs into animated videos controlled by a desired pose sequence while ensuring consistent appearance and temporal stability. We propose ReferenceNet, which genuinely preserves intricate character appearances and we also achieve efficient pose controllability and temporal continuity. Our approach not only applies to general character animation but also outperforms existing methods in specific benchmarks. *Animate Anyone* serves as a foundational method, with the potential for future extension into various image-to-video applications.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon 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-image diffusion models with an ensemble of expert denoisers. *arXiv preprint 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/CVF Conference 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] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 5

[6] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5933–5942, 2019. 1, 3

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

[8] 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 image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 4

[9] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023. 2, 8

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1

[11] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2, 3, 5, 8

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

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

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

[15] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 2

[16] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 5

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

[18] 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. 2, 5

[19] 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. 2, 3, 4, 7

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1

[21] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15954–15964, 2023. 2

[22] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 2, 4

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 7

[24] 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

[25] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2

[26] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin 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

[27] 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 *International Conference on Machine Learning*, 2021. 2

[28] 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

[29] 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

[30] 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

[31] Yurui Ren, Ge Li, Shan Liu, and Thomas H Li. Deep spatial transformation for pose-guided person image generation and animation. *IEEE Transactions on Image Processing*, 29: 8622–8635, 2020. 1, 3

[32] 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. 2

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2

[34] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[35] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in neural information processing systems*, 32, 2019. 1, 3, 7

[36] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2377–2386, 2019.

[37] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13653–13662, 2021. 1, 3, 7

[38] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 2

[39] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 4

[40] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Object compositing with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18310–18319, 2023. 2, 4

[41] 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. 5

[42] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 5

[43] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 4

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 4

[45] 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. 2, 3, 4, 7

[46] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2, 3

[47] Yaohui Wang, Xin Ma, Xinyuan Chen, Antitza Dantcheva, Bo Dai, and Yu Qiao. Leo: Generative latent image animator for human video synthesis. *arXiv preprint arXiv:2305.03989*, 2023. 3

[48] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[49] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu

Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 2

[50] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023. 2, 4

[51] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. *arXiv preprint arXiv:2306.07954*, 2023. 2

[52] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023. 5

[53] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2, 4

[54] 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, 3, 7

[55] 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. 2, 5

[56] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 4, 5

[57] Pengze Zhang, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Exploring dual-task correlation for pose guided person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7713–7722, 2022. 1, 3

[58] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[59] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 2

[60] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3657–3666, 2022. 1, 3, 7

[61] Luyang Zhu, Dawei Yang, Tyler Zhu, Fitsum Reda, William Chan, Chitwan Saharia, Mohammad Norouzi, and Ira Kemelmacher-Shlizerman. Tryondiffusion: A tale of two unets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4606–4615, 2023. 2