DANCER: DANCE ANIMATION VIA CONDITION ENHANCEMENT AND RENDERING WITH DIFFUSION MODEL

Yucheng Xing*, Jinxing Yin*, Xiaodong Liu*, Xin Wang

Stony Brook University, New York, USA
{yucheng.xing, jinxing.yin, xiaodong.liu, x.wang}@stonybrook.edu

*Equal contribution.

ABSTRACT

Recently, diffusion models have shown their impressive ability in visual generation tasks. Besides static images, more and more research attentions have been drawn to the generation of realistic videos. The video generation not only has a higher requirement for the quality, but also brings a challenge in ensuring the video continuity. Among all the video generation tasks, human-involved contents, such as human dancing, are even more difficult to generate due to the high degrees of freedom associated with human motions. In this paper, we propose a novel framework, named as ***DANCER (Dance ANimation via Condition Enhancement and Rendering with Diffusion Model)***, for realistic single-person dance synthesis based on the most recent stable video diffusion model. As the video generation is generally guided by a reference image and a video sequence, we introduce two important modules into our framework to fully benefit from the two inputs. More specifically, we design an ***Appearance Enhancement Module (AEM)*** to focus more on the details of the reference image during the generation, and extend the motion guidance through a ***Pose Rendering Module (PRM)*** to capture pose conditions from extra domains. To further improve the generation capability of our model, we also collect a large amount of video data from Internet, and generate a novel dataset ***TikTok-3K*** to enhance the model training. The effectiveness of the proposed model has been evaluated through extensive experiments on real-world datasets, where the performance of our model is superior to that of the state-of-the-art methods. All the data and codes will be released upon acceptance.

Keywords: Dance Animation, Diffusion Model, Dataset

1 Introduction

Creativity is one of the most important factors to measure the machine intelligence. In recent years, the emergence of various generative models, such as Variantional Auto-Encoders (VAEs) [Pinheiro Cinelli et al.(2021)], Generative Adversarial Networks (GANs) [Goodfellow et al.(2020)] and Diffusion Models (DMs) [Ho et al.(2020), Song et al.(2020)], has inspired more and more researchers to focus on the generation tasks, and the quality of the generative models reflect the creativity level of the machine learning algorithms. From images to videos, the objects to be generated get increasingly complex, benefiting from the complicated model structures and powerful computational resources. Among the generated contents, human-related activities, such as human dances, are particularly difficult due to the high degrees of freedom of human motions while they are also of vital value for building better human-assisted models in real applications.

Currently, there are several challenges that limit the generation performance of human dance videos. On one hand, due to the variety of changes in human body movements and the insufficient information provided in the reference image, details such as occluded limb distal segments and surrounding backgrounds may have been lost or distorted when transferred from the reference image to the corresponding video frames. On the other hand, besides the generation quality of a single video frame, the continuity of the adjacent frames is also of vital importance. Since the video frames are generated under the guidance of the prior pose sequence, the limited information and inherent jitters of the pose guidance will impact the smoothness of the generated videos.

In the human dance synthesis field, most current works rely on the generative models, such as GANs [Sarkar et al.(2021), Siarohin et al.(2019), Tian et al.(2021), Wang et al.(2020), Yoon et al.(2021)] and DMs [Ma et al.(2024), Feng et al.(2023), Karras et al.(2023), Wang et al.(2024b), Hu(2024), Chang et al.(2023), Wang et al.(2024a), Zhang et al.(2024), Zhai et al.(2024)]. Compared to GANs, DMs obtain a relatively better performance in general with a more stable training process. Existing DM-based methods usually integrate the reference image and the pose sequence into the diffusion process as extra conditions to make sure that the generated dance videos can match them. However, most of these methods overlook the subtle details present in reference images during generation and are limited by the lack of rich information in skeleton maps used for pose guidance.

To address the limitations in current models for human dance synthesis, we propose a novel framework named ***DANCER*** for high-quality generation of single-person dance videos. Our approach introduces two key modules: 1) a detail-aware *Appearance Enhancement Module (AEM)*, which captures subtle features from the reference image to enhance the visual quality of each generated frame; and 2) a *Pose Rendering Module (PRM)*, which enhances pose guidance to provide more detailed and accurate motion cues by enriching sparse skeleton maps with additional information extracted from the source reference video, such as segmentation maps, depth maps, and 3D parametric models. Together, AEM and PRM allow DANCER to generate dance videos with improved realism, visual fidelity, and temporal coherence. Another critical challenge in single-person dance generation is the limited size and diversity of existing datasets, which hampers the training effectiveness of deep learning models. To address this, we collect a wide variety of dance videos from the Internet and create a new large-scale dataset, **TikTok-3K**. This dataset significantly expands the available training data and is designed to support more robust and generalizable learning, which will ultimately benefit future research and development in the field of dance video generation. Our contributions can be summarized in the following four aspects:

- We propose a new SVD-based framework for the generation of single-person dance video, and verify the effectiveness of the framework through extensive experiments.
- We design a novel appearance enhancement module to focus more on the details during generation, which enhances the extracted appearance features on different levels.
- In order to improve the smoothness of the generated video, we augment the motion guidance and provide more pose references instead of only using skeleton maps, where motion conditions are preprocessed to remove jitters before being used in the generation to maintain the smoothness.
- We construct a new dataset comprising a large collection of human dance videos sourced and processed from the Internet. This dataset significantly extends the volume and diversity of existing resources in the field, providing a richer foundation for training and evaluating high-quality dance video generation models.

The remainder of this paper is organized as follows. In Sec. 2, we provide a brief review of the previous work related to human dance synthesis and diffusion models. In Sec. 3 and Sec. 4, we introduce the large dataset we newly create to facilitate the training of sophisticated network, and present in details our model designed for generating high quality video. We evaluate the effectiveness of the proposed framework through extensive experiments in Sec. 5. Finally, we conclude our work and point out some future research directions in Sec. 6.

2 Related Works

Human dance synthesis, or human image animation, targets at creating a consecutive video frames containing dance-related contents based on a static image and a sequence of poses. Previously, GAN-based methods [Sarkar et al.(2021), Siarohin et al.(2019), Tian et al.(2021), Wang et al.(2020), Yoon et al.(2021)] mainly employ spatial warping functions to transform the reference image according to the expected video frames along with the predicted flow map extracted from the motion changes of the pose sequence. However, the performance of this branch of methods is often compromised with insufficient information caused by mode collapse. This limits their capability in dealing with local details, keeping temporal consistencies and recovering the missing parts from occlusion. Recently, with the emergence of Diffusion Model (DM) [Croitoru et al.(2023)], a much more powerful generative model superior to GANs on generation quality and training stability, DM-based methods [Ma et al.(2024), Feng et al.(2023), Karras et al.(2023), Wang et al.(2024b), Hu(2024), Chang et al.(2023), Wang et al.(2024a), Zhang et al.(2024), Zhai et al.(2024)] have also been proposed to generate human dance videos under the supervision of a given pose sequence. Specifically, the reference image and the pose sequence are used as two conditions during the frame generation process. In DreaMoving [Feng et al.(2023)], the reference image is directly encoded by CLIP [Radford et al.(2021)] to replace the original text prompts to guide the Stable Diffusion [Rombach et al.(2022a)]. However, since CLIP is trained to match high-level semantic features for text, it lacks the capability of capturing low-level details and is thus not enough to get a good reference condition. To solve this problem, DreamPose [Karras et al.(2023)] combines CLIP with VAE in their encoder, while DisCo [Wang et al.(2024b)] processes the foreground and the background of the reference image separately

with the help of ControlNet [Zhang et al.(2023)]. AnimateAnyone [Hu(2024)] further extends this idea and designs a novel ReferenceNet, which is also used in MagicPose [Chang et al.(2023)]. Besides, VividPose [Wang et al.(2024a)] observes the limitation of identity consistency in generated video and adds additional id controller with a reference face. As for the pose guidance, a skeleton sequence drawn from pose estimators, such as OpenPose [Cao et al.(2017)] or DW-Pose [Yang et al.(2023)], is commonly used in [Wang et al.(2024b), Hu(2024), Chang et al.(2023)]. However, the extracted skeletons might be inaccurate, and MimicMotion [Zhang et al.(2024)] proposes to use detection confidence to enhance the extracted skeleton map in the condition embedding. To ease the sparsity problem in these skeletons, MagicAnimate [Xu et al.(2024)] adopts DensePose [Güler et al.(2018)] while some works [Wang et al.(2024a), Zhu et al.(2024)] even utilize 3D parametric SMPL model [Loper et al.(2023)].

3 TikTok-3K Dataset

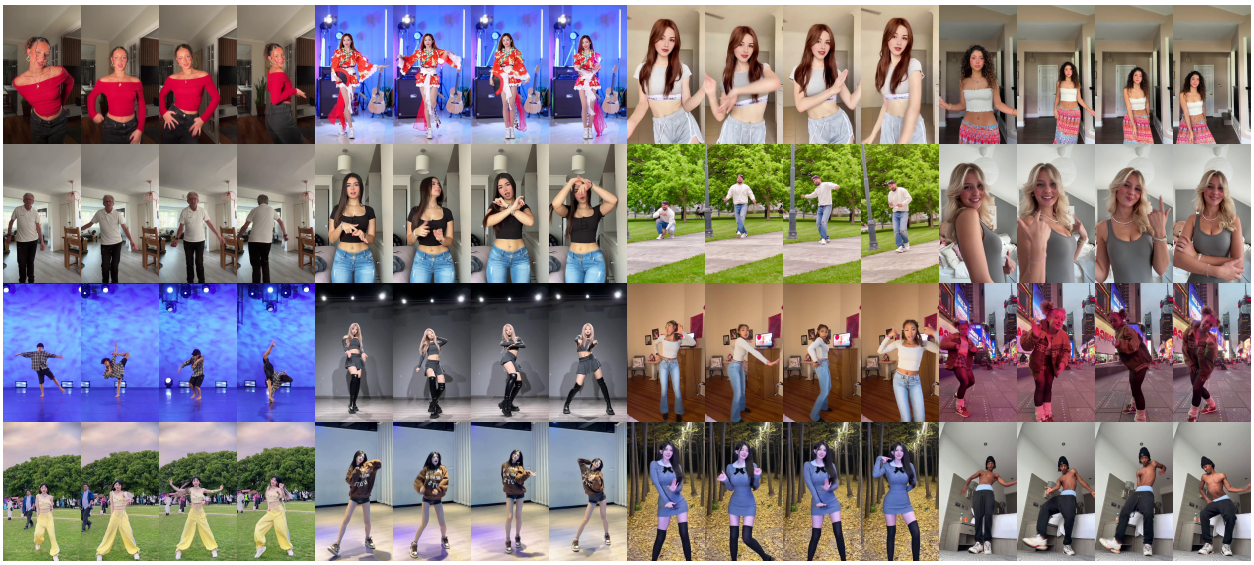

Figure 1: Illustration of video data in our *TikTok-3K* dataset.

Before diving into the details of our proposed model, we would like to first introduce our newly created large-scale single-dancer video dataset, *TikTok-3K*.

Modern machine learning models, particularly complex generative models with extensive learnable parameters, demand substantial amounts of high-quality training data to achieve a satisfactory performance. However, for the single-person dance generation, there exists only one commonly used dataset, $TikTok$ [Jafarian and Park(2021)]. It contains about 350 single-person dance videos, a very small number that limits its capacity for the training of diffusion models for video generation. The videos in $TikTok$ last for 10-15 seconds, with the sampling rate of 30 frames per second. Most videos record the faces and upper-bodies of dancers, and some have very low resolutions. The lack of large, high-resolution datasets with diverse video samples motivates us to curate TikTok-3K. Our dataset comprises approximately 3,000 carefully selected video clips with varied lengths, primarily ranging from 10 to 20 seconds, alongside some longer clips for learning long-range dependencies. These longer clips, typically recording the performance of professional dancers, feature complex and intricate motions.

TikTok-3K also offers greater variety across dance types, including genres from casual to classical, and expands on the environmental diversity. While prior datasets primarily consist of videos recorded in static indoor settings, TikTok-3K includes clips filmed in both static and dynamic outdoor environments, supporting model adaptability to different conditions. Additionally, our dataset encompasses a broad representation of dancers across races, genders, and age groups (Fig. 1), which helps prevent model overfitting to specific distributions—a limitation observed in previous generative models.

To ensure the usability, all videos have been manually selected and preprocessed. This process is critical and introduced to make sure that each clip is relevant, without blank or extraneous content, and ready for training. For better comparison with prior datasets, we present a breakdown of frame distributions between TikTok-3K and the traditional TikTok dataset, visualized in Fig. 2.

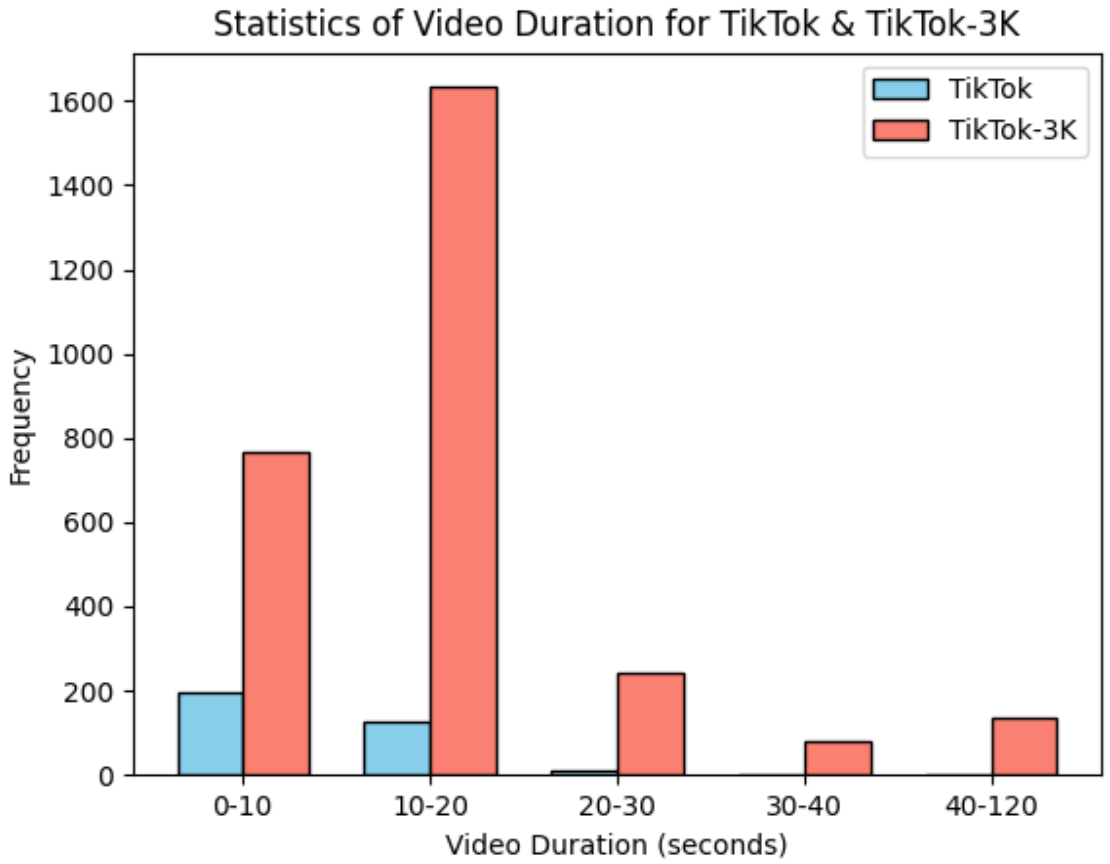

Figure 2: Comparison of frame distribution between *TikTok-3K* and *TikTok*.

4 Methodology

In this section, we will first give an overview of the framework of our *DANCER* model, and then introduce the two new modules, *Appearance Enhancement Module (AEM)* and *Pose Rendering Module (PRM)*, separately in detail.

4.1 Notations & Objective

Given an N-frame dance video $V_{src} = \{I_{src,0}, I_{src,1}, ..., I_{src,N}\}$ of the source performer src, where $I_{src,i}$ is the i^{th} frame within it, and a reference image I_{tgt} of the target person tgt, our goal is to design an effective model with the function $f(\cdot)$ to generate the corresponding N-frame video $V_{tgt} = \{I_{tgt,0}, I_{tgt,1}, ..., I_{tgt,N}\}$ whose subject is tgt, i.e.

$$f(V_{src}, I_{tgt}) := V_{tgt}. \tag{1}$$

4.2 Framework

As shown in Fig. 3, our model adopts a Latent-Diffusion [Rombach et al.(2022b)] framework, where video frames are first mapped into a low-dimensional latent space through a pre-trained VAE [Pinheiro Cinelli et al.(2021)]. The diffusion process is then carried out within this latent space to generate the required feature maps. Unlike image generation, video synthesis poses significantly higher computational demand due to the high-dimensional property of the video data. By operating in a compressed latent space, our approach effectively reduces the computational cost while maintaining the fidelity of the generated content.

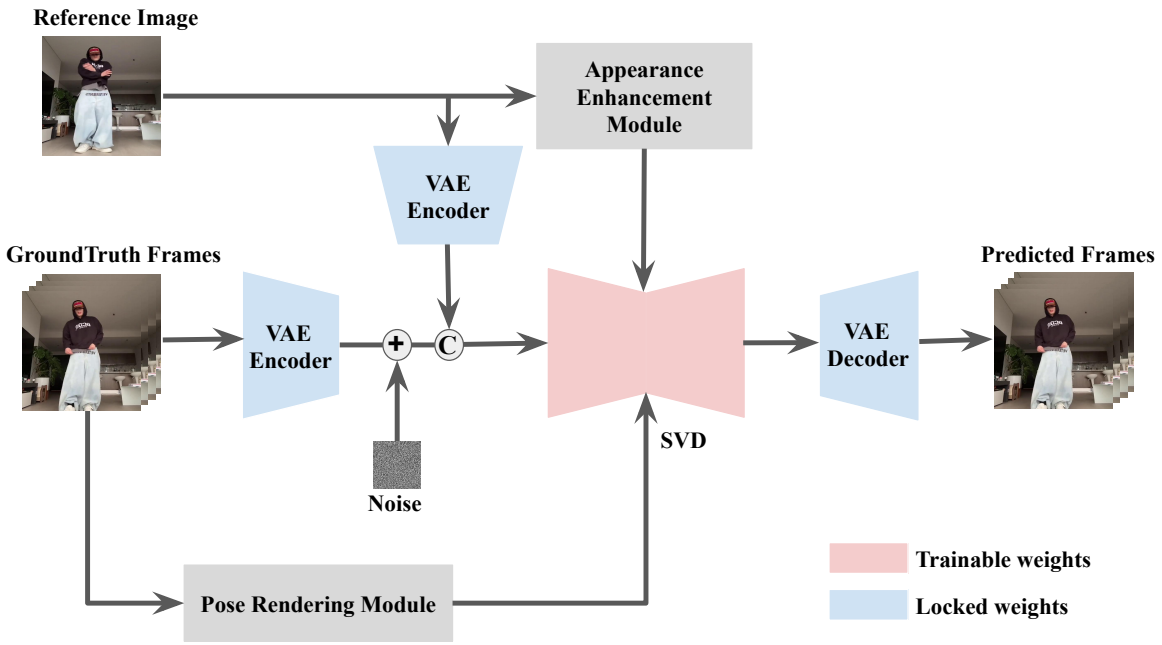

Figure 3: Overall framework of our DANCER model.

Specifically, given an image data I_{in}, the encoder $\mathbf{E}_{vae}(\cdot)$ within the VAE will extract an abstract feature K, which can be perfectly restored to a pixel-level image I_{out} through the VAE decoder $\mathbf{D}_{vae}(\cdot)$ accordingly, with $I_{in} = I_{out}$, i.e.

$$I_{in} = I_{out} = \mathbf{D}_{vae}(K) = \mathbf{D}_{vae}(\mathbf{E}_{vae}(I_{in})). \tag{2}$$

Therefore, to obtain the i^{th} frame $I_{tgt,i}$ of the expected dance video, we only need to make sure that the corresponding latent feature $K_{tgt,i}$ can be well generated through an conditional diffusion model $G(\cdot)$. The diffusion process is guided by both conditions c_{app} and c_{pose}, which are provided by the target reference image I_{tgt} and the source dance video V_{src} through our newly designed *Appearance Enhancement Module (AEM,* $\mathbf{F}_{AEM}(\cdot)$*)* and *Pose Rendering Module (PRM,* $\mathbf{F}_{PRM}(\cdot)$*)* separately. The details of the two modules will be discussed later in this section 4.3 and 4.4.

For the diffusion model $\mathbf{G}(\cdot)$, we choose to use the Stable Video Diffusion (SVD) [Blattmann et al.(2023)], a powerful image-to-video diffusion model trained on a large-scale video dataset. Compared to previous video generation models, SVD has shown superior performance on both generation quality and continuity, by adding additional temporal cross-attention layers right after the spatial convolutional layers within the U-Net [Ronneberger et al.(2015)] to handle the temporal consistency among generated latent features of adjacent frames. Following the formulations in [Blattmann et al.(2023)], the generation process of the target latent features $S_{tgt,t} = \{K_{tgt,0,t}, K_{tgt,1,t}, ..., K_{tgt,N,t}\}$ in the $(T-t)^{th}$ diffusion step can be expressed as

$$\begin{aligned} S_{tgt,t} &= \mathbf{G}(S_{tgt,t+1}, c_{app}, c_{pose}, t), \\ c_{app} &= \mathbf{F}_{AEM}(I_{tgt}), \\ c_{pose} &= \mathbf{F}_{PRM}(V_{src}), \end{aligned} \tag{3}$$

where T is the total number of diffusion steps. To better integrate the information of the target reference image I_{tgt}, instead of setting the initial input $S_{tgt,T}$ of the SVD as pure Gaussian noise, we repeat the latent feature K_{tgt} of the reference image N times and concatenate them with the random noise. Gaussian noise is added on top of the repeated K_{tgt} to make sure that the generated frames are similar to but not totally the same as the reference image, so

$$\begin{aligned} S_{tgt,T} &= \{K_{tgt,0,T}, K_{tgt,1,T}, ..., K_{tgt,N,T}\}, \\ &= \{Concat[\epsilon_{src,0}, K_{tgt} + \epsilon_{tgt,0}], \\ &\quad Concat[\epsilon_{src,1}, K_{tgt} + \epsilon_{tgt,1}], \\ &\quad ..., \\ &\quad Concat[\epsilon_{src,N}, K_{tgt} + \epsilon_{tgt,N}]\}, \end{aligned} \tag{4}$$

where $K_{tgt} = \mathbf{E}_{vae}(I_{tgt})$, $\{\epsilon_{src,0}, \epsilon_{src,1}, ..., \epsilon_{src,N}\}$ and $\{\epsilon_{tgt,0}, \epsilon_{tgt,1}, ..., \epsilon_{tgt,N}\}$ are Gaussian noise.

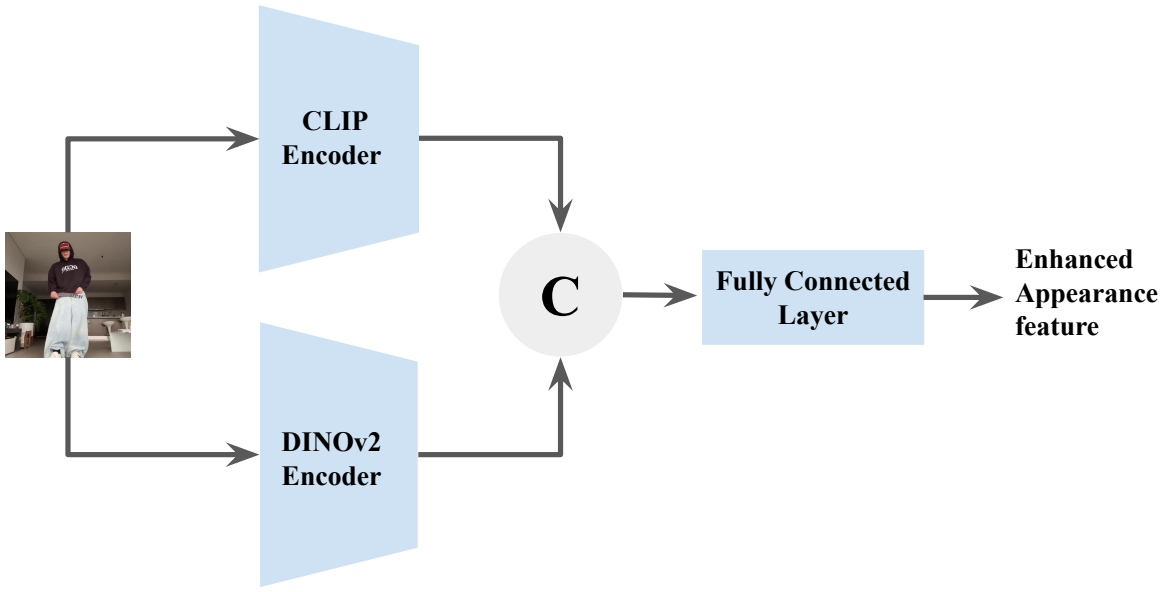

Figure 4: Illustration of our newly proposed Appearance Enhancement Module.

4.3 Appearance Enhancement Module (AEM)

Following the pipeline of text-guided conditional generation, image-to-video diffusion models in the early stage mainly use CLIP [Radford et al.(2021)] as their image condition encoder. However, as indicated in [Karras et al.(2023)], CLIP is designed as a multi-modal vision and language model to connect texts and images, where the image encoder is designed to extract features that can better fit the text. Besides, CLIP focuses on abstracting the high-level information within the image. As a result, most of earlier studies on video generation only rely on the CLIP-based image condition, without paying attention to the appearance details in the target reference image. The reference image details are of vital importance when generating target video frames. Therefore, we can often observe the missing of details or the distortion in the dance videos generated by current works, especially in the facial area, limb distal segments of the dancers, and the background area around them.

To solve the problem, we propose a novel module, Appearance Enhancement Module (AEM, $\mathbf{F}_{AEM}(\cdot)$), to better extract the details of the reference image so they can be applied to enhance the extracted appearance condition c_{app}. Specifically, as shown in Fig. 4, our AEM is composed of two main parts. Besides the commonly used CLIP encoder $\mathbf{E}_{clip}(\cdot)$, we also incorporate another ViT [Han et al.(2021)]-based encoder, and a DINO-v2 [Oquab et al.(2023)] ($\mathbf{E}_{dino}(\cdot)$), into our module. Different from CLIP, DINO-v2 is designed for the image segmentation task, which means it can handle the low-level information better and can be exploited to capture more details from the reference image I_{tgt}. In other words, these two components can provide information in two aspects:

$$\begin{aligned} c_h &= \mathbf{E}_{clip}(I_{tgt}), \\ c_l &= \mathbf{E}_{dino}(I_{tgt}). \end{aligned} \tag{5}$$

By combining the extracted features from both encoders, we can ensure that both high-level semantic features c_h and the low-level details c_l from the target reference image can be simultaneously captured by our model. With the equal distribution of effort on the low-level details and the high-level abstract information, our AEM can provide an enhanced appearance condition c_{app} through the effective combination of c_h and c_l using a fusion module $\mathbf{C}(\cdot)$, which is implemented with Fully-Connected Layers, and the total process within our AEM can be formulated as

$$\begin{aligned} c_{app} &= \mathbf{F}_{AEM}(I_{tgt}) \\ &= \mathbf{C}(Concat[c_h, c_l]) \\ &= \mathbf{C}(Concat[\mathbf{E}_{clip}(I_{tgt}), \mathbf{E}_{dino}(I_{tgt})]). \end{aligned} \tag{6}$$

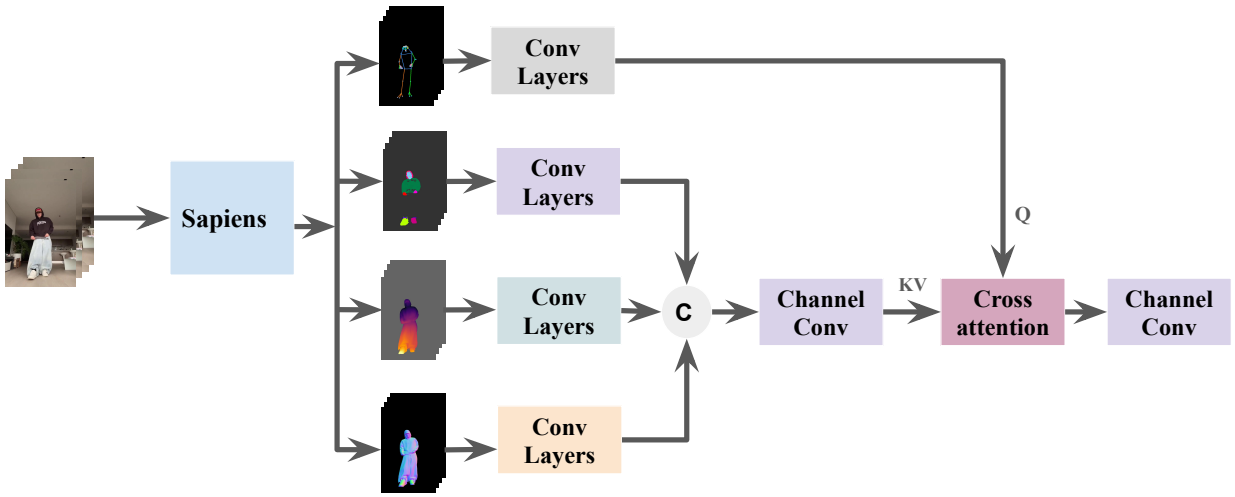

Figure 5: Illustration of our newly proposed Pose Rendering Module.

4.4 Pose Rendering Module (PRM)

Another limitation observed in prior methods is the insufficient quality of pose guidance. Most existing approaches rely heavily on skeleton maps extracted from source dance videos to guide generation. However, inaccuracies in skeleton detection can significantly degrade generation performance. Furthermore, even dense skeletons [Yang et al.(2023)] fail to fully capture the richness of motion present in the source video, let alone the sparse skeletons commonly used by many existing models. As a result, the generated videos often lack precision and realism in pose representation and movement continuity.

In order to capture the motion information in the source video V_{src} to the greatest extent, we introduce a more effective Pose Rendering Module (PRM, $\mathbf{F}_{PRM}(\cdot)$), as shown in Fig. 5, to provide the motion condition c_{pose} in Eq. (3) to control the generation process. Specifically, we adopt **Sapiens** [16], a most recent model, as our pose extractor. Pre-trained on a large-scale dataset of human images, it has shown impressive capability on several human-centric tasks. For each frame $I_{src,i}$ in V_{src}, Sapiens model $\mathbf{E}_{sapiens}(\cdot)$ can provide pose information from different aspects, including a body-part segmentation map $P^{seg}_{src,i}$, a depth map $P^{dep}_{src,i}$, a normal map $P^{norm}_{src,i}$ as well as a skeleton map $P^{ske}_{src,i}$, and these components can complement each other to more effectively guide the video generation. In formula, the pose representation $P_{src,i} = \{P^{ske}_{src,i}, P^{seg}_{src,i}, P^{dep}_{src,i}, P^{norm}_{src,i}\}$ for the i^{th} frame $I_{src,i}$ in the source video can be expressed as

$$P_{src,i} = \mathbf{E}_{sapiens}(I_{src,i}). \tag{7}$$

Different from [Zhu et al.(2024)], where a single model is used to extract features from source inputs of all domains, we design separate convolutional blocks $\mathbf{Conv}_{ske}(\cdot)$, $\mathbf{Conv}_{seg}(\cdot)$, $\mathbf{Conv}_{dep}(\cdot)$, $\mathbf{Conv}_{norm}(\cdot)$ for different pose maps within the pose representation $P_{src,i}$. By doing so, we can guarantee that each encoder best fits the corresponding pose map. Incorporating the extracted features from augmented pose maps, we can obtain enriched pose guidance from different perspective, and the final motion condition $c_{pose} = \{c_{pos,0}, c_{pos,1}, ..., c_{pos,N}\}$ participating in the diffusion process in Eq. (3) can be formulated as

$$\begin{aligned}
c_{pos,i} &= \mathbf{H}_2(\mathbf{A}(c_{ske}, c_{aug})) \\
&= \mathbf{H}_2(\mathbf{A}(c_{ske}, \mathbf{H}_1(Concat[c_{seg}, c_{dep}, c_{norm}]))), \\
c_{ske} &= \mathbf{Conv}_{ske}(P^{ske}_{src,i}), \\
c_{seg} &= \mathbf{Conv}_{seg}(P^{seg}_{src,i}), \\
c_{dep} &= \mathbf{Conv}_{dep}(P^{dep}_{src,i}), \\
c_{norm} &= \mathbf{Conv}_{norm}(P^{norm}_{src,i}),
\end{aligned} \tag{8}$$

where $\mathbf{A}(\cdot)$ is the cross-attention function, $\mathbf{H}_1(\cdot)$ and $\mathbf{H}_2(\cdot)$ are channel-wise convolutional block to further reduce the channel dimension of $c_{pos,i}$.

5 Experiments

5.1 Experimental Settings

Table 1: Overall performance of different models on the *TikTok* testset. † represents the results measured by ourselves by rerunning the corresponding models. For each metric, the best result is in **bold** and the second best is underlined.

	Image-Level					Video-Level	
	FID (↓)	SSIM (↑)	LPIPS (↓)	PSNR (↑)	L1 (↓)	FID-VID (↓)	FVD (↓)
DISCO [Wang et al.(2024b)]	28.31	0.674	0.285	16.68	3.69E-04	55.17	267.75
MagicPose [Chang et al.(2023)]	25.50	0.752	0.292	29.53	8.10E-05	46.30	/
VividPose [Wang et al.(2024a)]	31.89	0.758	0.261	29.83	6.89E-05	18.81	152.97
AnimateAnyone [Hu(2024)]	/	0.718	0.285	29.56	/	/	171.90
MimicMotion† [Zhang et al.(2024)]	33.65	0.740	0.296	29.08	9.07E-05	19.31	206.46
DANCER (w/o TikTok-3K)	26.56	0.802	0.249	29.13	7.25E-05	16.58	140.64
DANCER (w/ TikTok-3K)	**25.27**	**0.803**	**0.247**	**30.03**	**6.48E-05**	**13.66**	**129.45**

5.1.1 Evaluation Metrics

To better evaluate the quality of generated dance videos, following outlines proposed in DisCo [Wang et al.(2024b)], we adopt metrics from two aspects: image generation quality and video generation quality. For the first part, we test frame-wise FID [Heusel et al.(2017)], SSIM [Wang et al.(2004)], LPIPS [Zhang et al.(2018)], PSNR [Hore and Ziou(2010)] and L1; while for the second part, we report FID-VID [Balaji et al.(2019)] and FVD [Unterthiner et al.(2018)] for every consecutive 16 frames.

5.1.2 Implementation Details

We adopt the pre-trained stable video diffusion image-to-video model. During training, we train the weights of U-Net [Ronneberger et al.(2015)] and our augmented pose module, keeping other parts frozen. All experiments are conducted on 8 NVIDIA H100 with batch-size 1 and learning rate 1e-5. Each batch includes 1 reference frame, 8 randomly sampled consecutive ground-truth frames and corresponding extracted pose images.

5.2 Overall Performance

In order to quantitatively evaluate our model, we conduct experiments on *TikTok* [Jafarian and Park(2021)] dataset and compare the performance with some other baselines, including DISCO [Wang et al.(2024b)], MagicPose [Chang et al.(2023)], VividPose [Wang et al.(2024a)], AnimateAnyone [Hu(2024)] and MimicMotion [Zhang et al.(2024)]. During experiments, we conduct the same preprocessing as done in [Wang et al.(2024b)] to data for fair comparison. The overall results are provided in Table. 1. We can see that our proposed model is superior to the state-of-the-art methods on the single-person dance synthesis task with obviously better SSIM, LPIPS, FID-VID and FVD. Specifically, our SSIM surpasses Vividpose by 5.8%, LPIPS by 4.6%, FID-VID by 12% and FVD by 8%. The improvement is reflected from two aspects. First, according to the image-level metrics, our proposed model can generate high-quality video frames, which is attributed to the attention from our model to the details of reference images. Second, the videos generated by our model have a better continuity, which is demonstrated through its improvement of video-level measurement.

Table 2: Performance comparison among our model with different appearance encoders on *TikTok* dataset. For each metric, the best result is in **bold**.

AEM		Image-Level					Video-Level	
CLIP	DINO-v2	FID (↓)	SSIM (↑)	LPIPS (↓)	PSNR (↑)	L1 (↓)	FID-VID (↓)	FVD (↓)
✓	✗	32.49	0.7693	0.2874	**29.37**	8.27E-5	22.34	167.06
✓	✓	**31.53**	**0.7782**	**0.2604**	29.12	**7.91E-5**	**21.95**	**159.33**

Table 3: Performance comparison between augmented pose guidance and traditional skeleton map on *TikTok* dataset. For each metric, the best result is in **bold**.

PRM		Image-Level					Video-Level	
Skeleton	Sapiens	FID (↓)	SSIM (↑)	LPIPS (↓)	PSNR (↑)	L1 (↓)	FID-VID (↓)	FVD (↓)
✓	✗	32.49	0.7693	0.2874	**29.37**	8.27E-5	22.34	167.06
✓	✓	**29.97**	**0.7728**	**0.2639**	29.31	**7.34E-5**	**19.03**	**155.02**

5.3 Ablation Study

To further demonstrate the effectiveness of our design, we separately study each component through extensive experiments, including the *Appearance Enhancement Module*, the *Pose Rendering Module* and the effect of additional training data brought by our novel dataset.

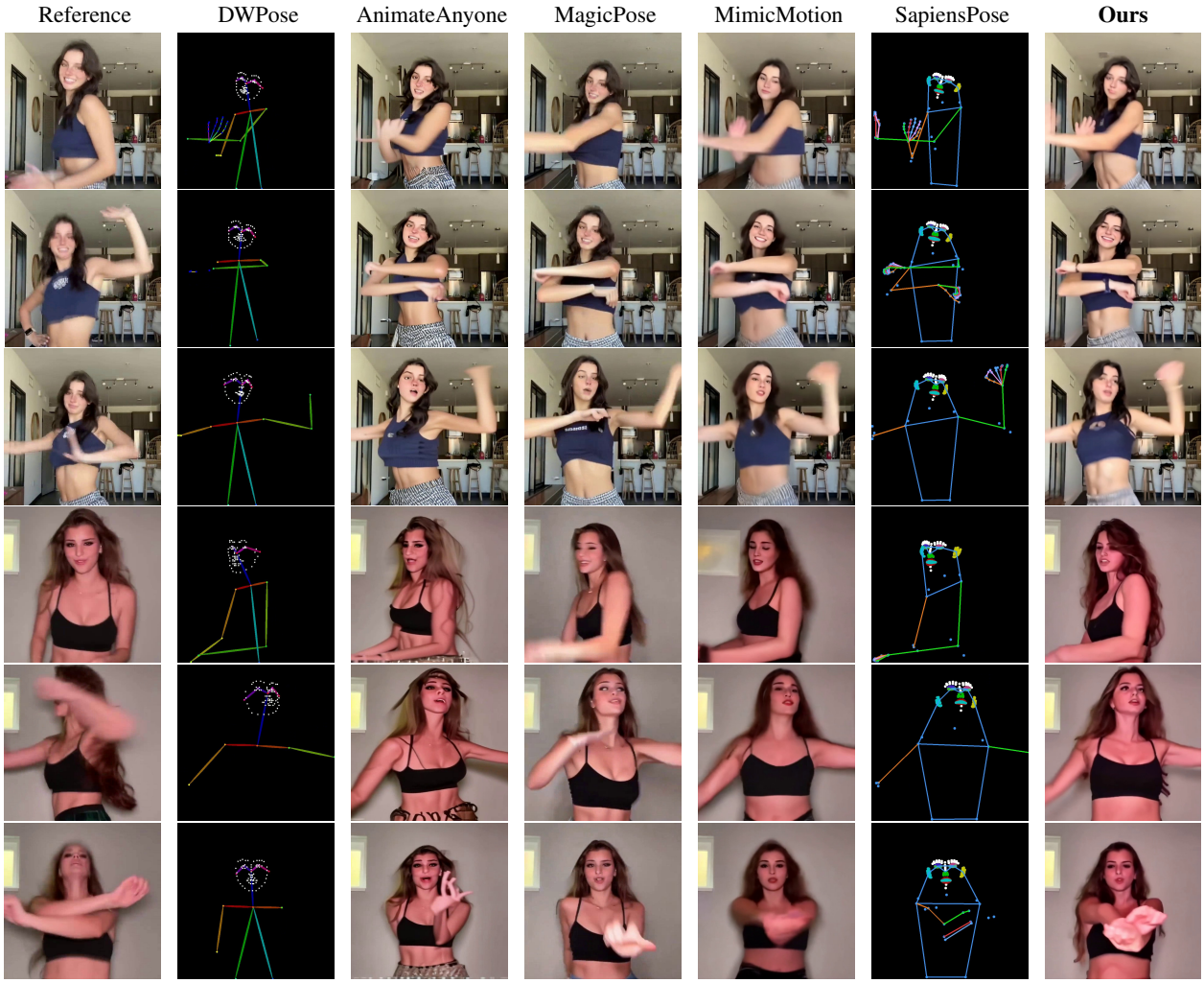

Figure 6: Illustration of the comparison among generated video samples using different methods.

- **Appearance Enhancement Module:** In this part, we compare the performance of our framework using the newly designed appearance enhancement module (AEM) with the same framework only containing CLIP encoder. As shown in Table. 2, the effectiveness of our novel module has been proved through its better performance compared to others, which confirmed the importance of feature extraction in different levels

and the extra attention on details. A larger improvement is observed in the evaluation of the overall video performance, while the continuity is an important factor to consider. This is as expected, as a complete video involves more details and requires a method that is able to comprehensively consider all the information.

- **Pose Rendering Module:** Another design that we want to evaluate is the Pose Rendering Module (PRM) in our model. As discussed in Sec. 4.4, compared to only using skeleton maps, augmented pose guidance can provide extra information for reference. According to Table. 3, the improvement brought by the augmented pose guidance in our pose rendering module is obvious, especially in FID metric.
- **Additional Training Data:** To further demonstrate the impact of our newly proposed large-scale dataset, we also conduct extra experiments to compare the model performances with and without using our *TikTok-3K* for further training. From the Table. 1, we can see that the continuing training on the new dataset obviously empowers our model, and the performance improvement is larger for the video sequence generation as it requires more data in the training. Since our *TikTok-3K* contains much larger number of data samples, much longer video duration as well as much more diversity on dance contents, it can greatly improve the performance of machine learning algorithms that rely on data. The big dataset we create not only helps improve the performance of our model, but will also provide benefits to future methods exploring video related fields.

5.4 Qualitative Study

To illustrate our results more effectively, we provide generated samples for comparison. As shown in Fig. 6, the dance videos produced using our method, DANCER, exhibit superior visual quality compared to previous works, featuring clearer details. Specifically, the pose images obtained from Sapiens capture clearer body shapes, finer details, and more accurate orientation than those derived from DWPose, as evident in the first and last rows of Fig. 6. This enhanced body rendering improves coherence with the reference object, significantly enhancing fidelity. Moreover, fine-grained details such as abdominal muscles, facial expressions, and hair shapes are well preserved, rather than being overly smoothed. These improvements are attributable to our intrinsic encoder design, which effectively synthesizes low-level structural features with high-level semantic features. Consequently, our results align more naturally with the reference, offering a more visually consistent outcome compared to prior approaches.

6 Conclusion

In this paper, we propose a novel framework for human dance synthesis based on the novel generation of conditions to guide the operation of video diffusion models. To better leverage pose references, beyond traditional skeleton maps, we design a pose rendering module that incorporates diverse motion cues, including segmentation maps, depth maps, and normal maps. As another important component of our proposed framework, we design a new appearance enhancement module to extract finer visual details from the reference image, such as facial features, limb distal segments and the background areas around the dancer. The effectiveness of our design has been proved on real-world datasets through extensive experiments. To further support model training and evaluation, we compile and plan to release a large-scale dance video dataset, which we believe will benefit the broader research community. Although we achieve some promising results, we also see that potential challenges remain in this field. In the future, we will extend our current work to more complex scenarios like multi-person dance or general human activity generation.

References

[Balaji et al.(2019)] Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. 2019. Conditional GAN with Discriminative Filter Generation for Text-to-Video Synthesis.. In *IJCAI*, Vol. 1. 2.

[Blattmann et al.(2023)] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127* (2023).

[Cao et al.(2017)] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7291–7299.

[Chang et al.(2023)] Di Chang, Yichun Shi, Quankai Gao, Hongyi Xu, Jessica Fu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. 2023. MagicPose: Realistic Human Poses and Facial Expressions Retargeting with Identity-aware Diffusion. In *Forty-first International Conference on Machine Learning*.

[Croitoru et al.(2023)] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 9 (2023), 10850–10869.

[Feng et al.(2023)] Mengyang Feng, Jinlin Liu, Kai Yu, Yuan Yao, Zheng Hui, Xiefan Guo, Xianhui Lin, Haolan Xue, Chen Shi, Xiaowen Li, et al. 2023. Dreamoving: A human dance video generation framework based on diffusion models. *arXiv preprint arXiv:2312.05107* (2023).

[Goodfellow et al.(2020)] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.

[Güler et al.(2018)] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. 2018. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7297–7306.

[Han et al.(2021)] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. 2021. Transformer in transformer. *Advances in neural information processing systems* 34 (2021), 15908–15919.

[Heusel et al.(2017)] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

[Ho et al.(2020)] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

[Hore and Ziou(2010)] Alain Hore and Djemel Ziou. 2010. Image quality metrics: PSNR vs. SSIM. In *2010 20th international conference on pattern recognition*. IEEE, 2366–2369.

[Hu(2024)] Li Hu. 2024. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8153–8163.

[Jafarian and Park(2021)] Yasamin Jafarian and Hyun Soo Park. 2021. 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*. 12753–12762.

[Karras et al.(2023)] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. 2023. Dreampose: Fashion image-to-video synthesis via stable diffusion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 22623–22633.

[Loper et al.(2023)] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2023. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 851–866.

[Ma et al.(2024)] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. 2024. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 4117–4125.

[Oquab et al.(2023)] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).

[Pinheiro Cinelli et al.(2021)] Lucas Pinheiro Cinelli, Matheus Araújo Marins, Eduardo Antúnio Barros da Silva, and Sérgio Lima Netto. 2021. Variational autoencoder. In *Variational Methods for Machine Learning with Applications to Deep Networks*. Springer, 111–149.

[Radford et al.(2021)] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. `https://proceedings.mlr.press/v139/radford21a.html`

[Rombach et al.(2022a)] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022a. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10684–10695.

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

[Ronneberger et al.(2015)] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. 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*. Springer, 234–241.

[Sarkar et al.(2021)] Kripasindhu Sarkar, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. 2021. Humangan: A generative model of human images. In *2021 International Conference on 3D Vision (3DV)*. IEEE, 258–267.

[Siarohin et al.(2019)] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First order motion model for image animation. *Advances in neural information processing systems* 32 (2019).

[Song et al.(2020)] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

[Tian et al.(2021)] Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N Metaxas, and Sergey Tulyakov. 2021. A good image generator is what you need for high-resolution video synthesis. *arXiv preprint arXiv:2104.15069* (2021).

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

[Wang et al.(2024a)] Qilin Wang, Zhengkai Jiang, Chengming Xu, Jiangning Zhang, Yabiao Wang, Xinyi Zhang, Yun Cao, Weijian Cao, Chengjie Wang, and Yanwei Fu. 2024a. VividPose: Advancing Stable Video Diffusion for Realistic Human Image Animation. *arXiv preprint arXiv:2405.18156* (2024).

[Wang et al.(2024b)] Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. 2024b. Disco: Disentangled control for realistic human dance generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9326–9336.

[Wang et al.(2020)] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. 2020. G3AN: Disentangling appearance and motion for video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5264–5273.

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

[Xu et al.(2024)] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. 2024. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1481–1490.

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

[Yoon et al.(2021)] Jae Shin Yoon, Lingjie Liu, Vladislav Golyanik, Kripasindhu Sarkar, Hyun Soo Park, and Christian Theobalt. 2021. Pose-guided human animation from a single image in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15039–15048.

[Zhai et al.(2024)] Yuanhao Zhai, Kevin Lin, Linjie Li, Chung-Ching Lin, Jianfeng Wang, Zhengyuan Yang, David Doermann, Junsong Yuan, Zicheng Liu, and Lijuan Wang. 2024. IDOL: Unified Dual-Modal Latent Diffusion for Human-Centric Joint Video-Depth Generation. In *Proceedings of the European Conference on Computer Vision*.

[Zhang et al.(2023)] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models.

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

[Zhang et al.(2024)] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. 2024. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680* (2024).

[Zhu et al.(2024)] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. 2024. Champ: Controllable and consistent human image animation with 3d parametric guidance. *arXiv preprint arXiv:2403.14781* (2024).