

# DreaMoving: A Human Video Generation Framework based on Diffusion Models

Mengyang Feng, Jinlin Liu, Kai Yu, Yuan Yao, Zheng Hui, Xiefan Guo, Xianhui Lin, Haolan Xue, Chen Shi, Xiaowen Li, Aojie Li, Xiaoyang Kang, Biwen Lei, Miaomiao Cui, Peiran Ren, Xuansong Xie  
Alibaba Group

{mengyang.fmy, ljl191782, jinmao.yk, ryan.yy, huizheng.hz, guoxiefan.gxf, xianhui.lxh, haolan.xhl, zhicheng.sc, lxw262398, liaojie.laj, kangxiaoyang.kxy, biwen.lbw, miaomiao.cmm, peiran.rpr, xingtong.xxs}@alibaba-inc.org

## Abstract

*In this paper, we present DreaMoving, a diffusion-based controllable video generation framework to produce high-quality customized human videos. Specifically, given target identity and posture sequences, DreaMoving can generate a video of the target identity moving or dancing anywhere driven by the posture sequences. To this end, we propose a Video ControlNet for motion-controlling and a Content Guider for identity preserving. The proposed model is easy to use and can be adapted to most stylized diffusion models to generate diverse results. The project page is available at <https://dreamoving.github.io/dreamoving>.*

## 1. Introduction

Recent text-to-video (T2V) models like Stable-Video-Diffusion<sup>1</sup> and Gen2<sup>2</sup> make breakthrough progress in video generation. However, it is still a challenge for human-centered content generation, especially character dance. The problem involves the lack of open-source human dance video datasets and the difficulty of obtaining the corresponding precise text description, making it a challenge to train a T2V model to generate videos with intraframe consistency, longer length, and diversity. Besides, personalization and controllability stand as paramount challenges in the realm of human-centric content generation, attracting substantial scholarly attention. Representative research like ControlNet [13] is proposed to control the structure in the conditional image generation, while DreamBooth [10] and LoRA [6] show the ability in appearance control through

images. However, these techniques often fail to offer precise control over motion patterns or necessitate hyperparameter fine-tuning specific to target identities, introducing an additional computational burden. Customized video generation is still under investigation and represents a largely uncharted territory. In this paper, we present a human dance video generation framework based on diffusion models (DM), named **DreaMoving**.

The rest of the paper is organized as follows. Sec. 2 presents a detailed description of how the DreaMoving is built. Sec. 3 presents some results generated by our method.

## 2. Architecture

DreaMoving is built upon Stable-Diffusion [9] models. As illustrated in Fig. 1, it consists of three main networks, including the Denoising U-Net, the Video ControlNet, and the Content Guider. Inspired by AnimateDiff [5], we insert motion blocks after each U-Net block in the Denoising U-Net and the ControlNet. The Video ControlNet and the Content Guider work as two plug-ins of the Denoising U-Net for controllable video generation. The former is responsible for motion-controlling while the latter is in charge of the appearance representation.

### 2.1. Data Collection and Preprocessing

To gain better performance in generating human videos, we collected around 1,000 high-quality videos of human dance from the Internet. As the training of the temporal module needs continuous frames without any transitions and special effects, we further split the video into clips and finally got around 6,000 short videos (8~10 seconds). For text description, we take Minigtpt-v2 [3] as the video captioner. Specifically, using the “grounding” version, the instruction is *[grounding] describe this frame in a detailed manner*. The generated caption of the centered frame in keyframes

<sup>1</sup><https://stability.ai/news/stable-video-diffusion-open-ai-video-model>

<sup>2</sup><https://research.runwayml.com/gen2>

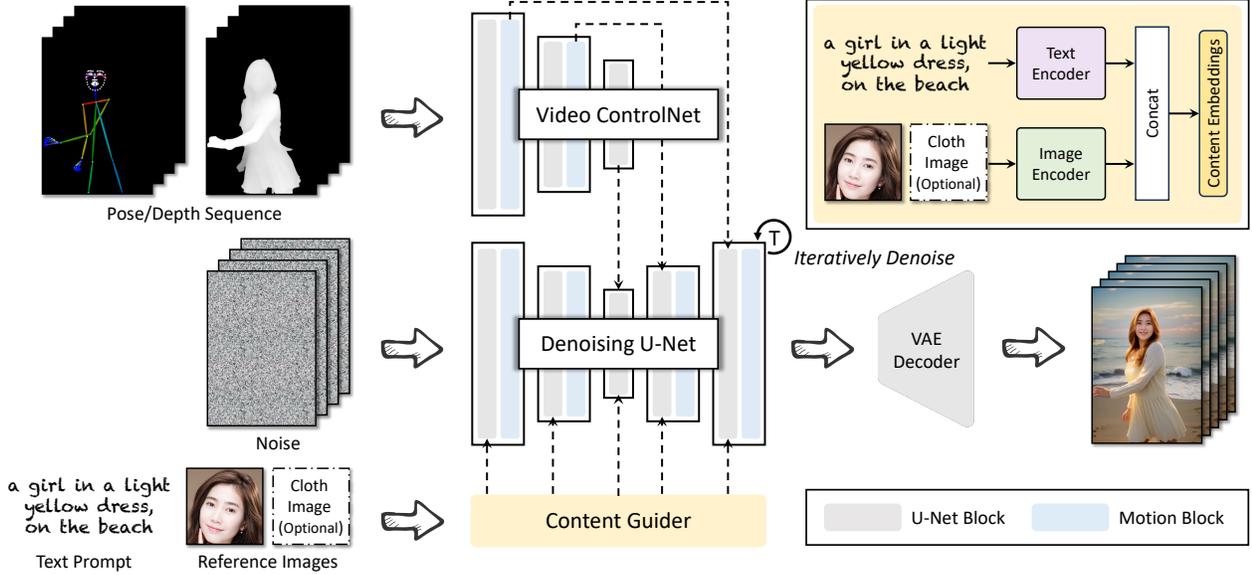


Figure 1. The overview of DreaMoving. The Video ControlNet is the image ControlNet [13] injected with motion blocks after each U-Net block. The Video ControlNet processes the control sequence (pose or depth) to additional temporal residuals. The Denoising U-Net is a derived Stable-Diffusion [9] U-Net with motion blocks for video generation. The Content Guider transfers the input text prompts and appearance expressions, such as the human face (the cloth is optional), to content embeddings for cross attention.

represents the whole video clip’s description, mainly describing the content of the subject and background faithfully.

## 2.2. Motion Block

To improve the temporal consistency and motion fidelity, we integrate motion blocks into both the Denoising U-Net and ControlNet. The motion block is extended from the AnimateDiff [5], and we enlarge the temporal sequence length to 64. We first initialize the weights of motion blocks from the AnimateDiff (*mm\_sd\_v15.ckpt*) and fine-tune them on the private human dance video data.

## 2.3. Content Guider

The Content Guider is designed to control the content of the generated video, including the appearance of human and the background. One simple way is to describe the human appearance and background with a text prompt, such as ‘a girl in a white dress, on the beach’. However, it is hard to describe a personalized human appearance for a normal user. Even by complex prompt engineering, the model may not give the desired output.

Inspired by IP-Adapter [12], we propose to utilize the image prompt for precise human appearance guidance and the text prompt for background generation. Specifically, a face image is used to encode the facial features through an image encoder, and a cloth/body image is optionally involved to encode the body features. The text and human appearance features are concatenated as the final content em-

beddings. The content embeddings are then sent to cross-attention layers for human appearance and background representations as described in IP-Adapter [12]. Given the query features  $Z$ , the text features  $c_t$ , the face features  $c_f$ , and the cloth features  $c_c$ , the output of cross-attention  $Z'$  can be defined by the following equation:

$$Z' = \text{softmax} \left( \frac{QK_t^T}{\sqrt{d}} \right) V_t + \alpha_f \text{softmax} \left( \frac{QK_f^T}{\sqrt{d}} \right) V_f + \alpha_c \text{softmax} \left( \frac{QK_c^T}{\sqrt{d}} \right) V_c, \quad (1)$$

where,  $Q = ZW_q$ ,  $K_t = c_t W_k^t$ , and  $V_t = c_t W_v^t$  are the query, key, and values matrices from the text features,  $K_f = c_f W_k^f$ , and  $V_f = c_f W_v^f$  are the key, and values matrices from the face features, and  $K_c = c_c W_k^c$ , and  $V_c = c_c W_v^c$  are the key, and values matrices from the cloth features.  $\alpha_f$  and  $\alpha_c$  are the weights factor.

## 2.4. Model Training

### 2.4.1 Content Guider Training

The Content Guider serves as an independent module for base diffusion models. Once trained, it can be generalized to other customized diffusion models. We trained the Content Guider based on SD v1.5 and used OpenCLIP ViT-H14 [7] as the image encoder as [12]. To better preserve the identity

of the reference face, we employ the Arcface [4] model to extract the facial correlated features as a supplement to the clip features. We collect the human data from LAION-2B, then detect and filter images without faces. During training, the data are randomly cropped and resized to  $512 \times 512$ . Content Guider is trained on a single machine with 8 V100 GPUs for 100k steps, batch size is set to 16 for each GPU, AdamW optimizer [8] is used with a fixed learning rate of  $1e - 4$  and weight decay of  $1e - 2$ .

### 2.4.2 Long-Frame Pretraining

We first conduct a warming-up training stage to extend the sequence length in the motion module from 16 to 64 on the validation set (5k video clips) of WebVid-10M [1]. We only train the motion module of the Denoising U-Net and freeze the weights of the rest of the network. No ControlNet and image guidance are involved in this stage. The learning rate is set to  $1e - 4$  and the resolution is  $256 \times 256$  (resize & center crop). The training is stopped after 10k steps with a batch size of 1.

### 2.4.3 Video ControlNet Training

After the long-frame pretraining, we train the Video ControlNet with the Denoising U-Net by unfreezing both the motion blocks and U-Net blocks in the Video ControlNet and fixing the Denoising U-Net. The weights of motion blocks in Video ControlNet are initialized from the long-frame pretraining stage. In this stage, we train the network on the collected 6k human dance video data. No image guidance is involved in this stage. The human pose or depth is extracted as the input of the Video ControlNet using DW-Pose [11] or ZoeDepth [2], respectively. The learning rate is set to  $1e - 4$  and the resolution is  $352 \times 352$ . The training is stopped after 25k steps with a batch size of 1.

### 2.4.4 Expression Fine-tuning

To gain better performance in human expression generation, we further fine-tune the motion blocks in Denoising U-Net by training with the Video ControlNet on the collected 6k human dance video data. In this stage, the whole Video ControlNet and the U-Net blocks in the Denoising U-Net are locked, and only the weights of the motion blocks in the Denoising U-Net are updated. The learning rate is set to  $5e - 5$  and the resolution is  $512 \times 512$ . The training is stopped after 20k steps with a batch size of 1.

## 2.5. Model Inference

At the inference stage, the inputs are composed of the text prompt, the reference images, and the pose or depth sequence. The control scale of the Video ControlNet is set to 1.0 for pose or depth only. Our method also supports

the form of multi-controlnet, and the depth and pose Video ControlNets can be used simultaneously. The strength of the face/body guidance is also controllable in the Content Guider by adjusting the  $\alpha_f$  and  $\alpha_c$  in Eqn. 1. The content is fully controlled by the text prompt if  $\alpha_f = \alpha_c = 0$ .

## 3. Results

DreaMoving can generate high-quality and fidelity videos given guidance sequence and simple content description (text prompt only, image prompt only, or text-and-image prompts) as input. In Fig. 2, we show the result with text prompt only. To keep the face identity, the user can input the face image to the Content Guider to generate a video of some specific person (demonstrated in Fig. 3). Moreover, the user can define both the face content and clothes content, as exhibited in Fig. 4. We further test the generalization of the proposed method on images of unseen domains. In Fig. 5, we run DreaMoving using unseen stylized images. Our method is able to generate videos in accordance with the style and content in the input image.

## References

- [1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021. 3
- [2] Shariq Farooq Bhat, Reiner Birkel, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth, 2023. 3
- [3] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 1
- [4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 3
- [5] 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. 1, 2
- [6] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 1
- [7] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hananeh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 2

a girl with short hair wearing black clothes in the room



A cheerleader wearing red and golden uniform on the football field



a woman with long hair wearing white suit and pants in the street



Figure 2. The results of DreaMoving with text prompt as input.



a girl, smiling, dancing in a French town, wearing long light blue dress



a girl, smiling, in the park with golden leaves in autumn wearing coat with long sleeve



a girl, smiling, standing on beach, wearing light yellow dress with long sleeves



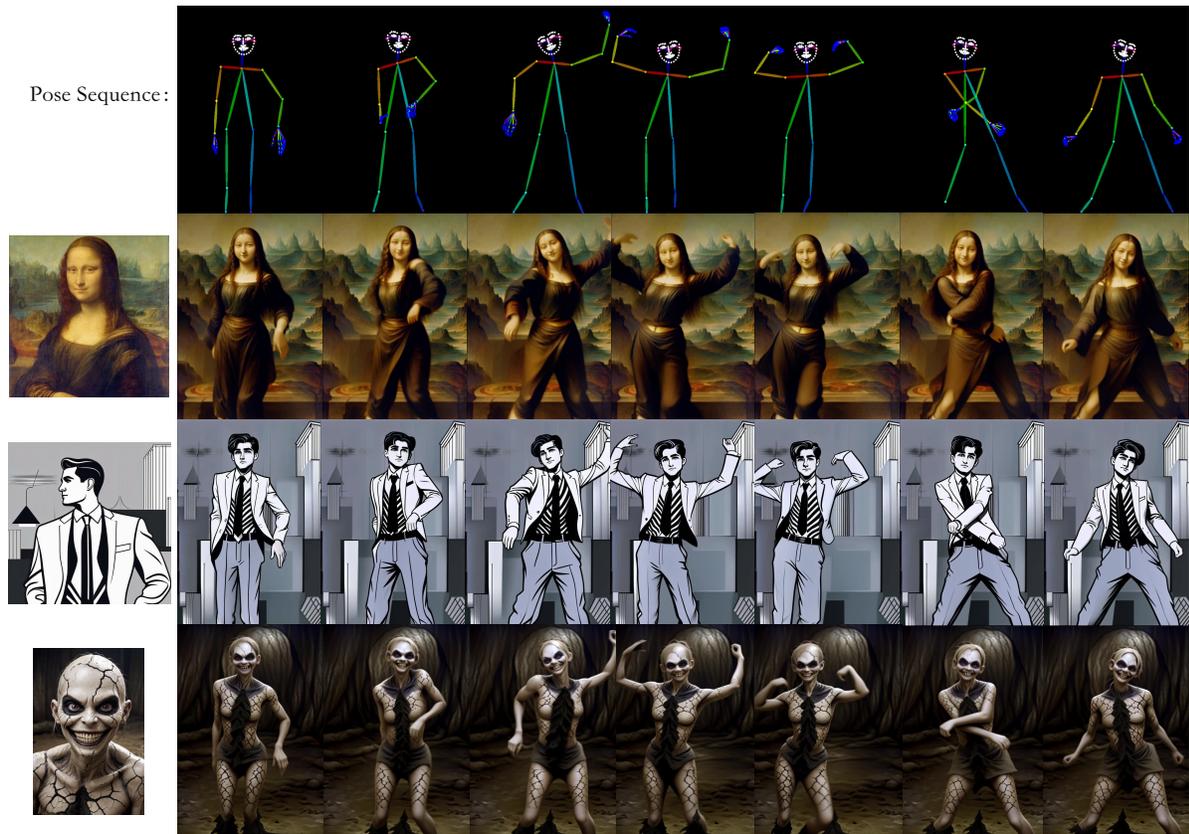
a man, dancing in front of Pyramids of Egypt, wearing a suit with a blue tie



Figure 3. The results of DreaMoving with text prompt and face image as inputs.



Figure 4. The results of DreaMoving with face and cloth images as inputs.



- [8] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3
- [9] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1, 2
- [10] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. 2022. 1
- [11] 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. 3
- [12] Hu Ye, Jun Zhang, Sibio Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. 2023. 2
- [13] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 1, 2