AnimateAnywhere: Context-Controllable Human Video Generation with ID-Consistent One-shot Learning

Hengyuan Liu
Institute of Information Engineering, Chinese Academy of Sciences
School of Cyber Security, University of Chinese Academy of Sciences
Beijing, China
liuhengyuan@iie.ac.cn

Xiaodong Chen
University of Science and Technology of China
Hefei, China
cxd1230@mail.ustc.edu.cn

Xinchen Liu
JD Explore Academy
Beijing, China
liuxinchen1@jd.com

Xiaoyan Gu
Institute of Information Engineering, Chinese Academy of Sciences
Beijing, China
guxiaoyan@iie.ac.cn

Wu Liu*
University of Science and Technology of China
Hefei, China
liuwu@ustc.edu.cn

Abstract

We demonstrate AnimateAnywhere, a personalized video generation framework that generates videos of a specific person with customized motions, scenes, and objects. Compared to existing approaches that animate a reference person image with a fixed background, AnimateAnywhere not only can preserve the consistency of the person but also can control the context like the scenes in the video. To achieve this goal, we first train a powerful base model using large-scale human images and videos with diverse scenes, poses, and captions to learn knowledge about contexts and human motions. Then, given a short video, we propose an ID-consistent one-shot learning method to obtain a personalized model by injecting the ID-related information into the pre-trained model. Finally, the user only needs to type in a text prompt to describe the expected scene/objects and select a reference motion, AnimateAnywhere can generate his/her video with the desired conditions.

CCS Concepts

• **Computing methodologies** → **Computer vision**.

Keywords

Customized Video Generation, One-shot Learning

ACM Reference Format:
Hengyuan Liu, Xiaodong Chen, Xinchen Liu, Xiaoyan Gu, and Wu Liu. 2024. AnimateAnywhere: Context-Controllable Human Video Generation with ID-Consistent One-shot Learning. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28–November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3688865.3689477

MM '24, October 28–November 1, 2024, Melbourne, VIC, Australia.

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3688865.3689477

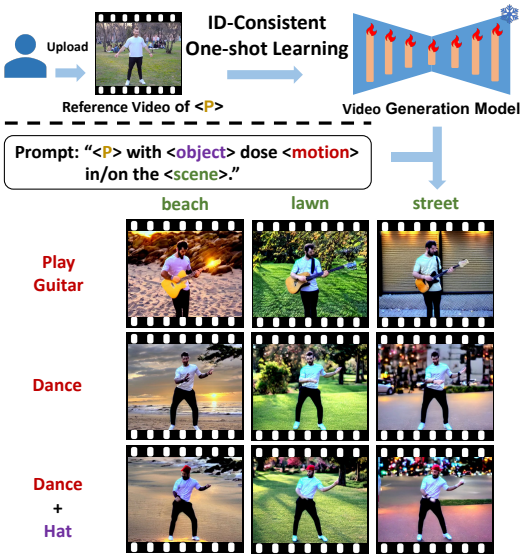

Figure 1: The illustration of ID-consistent one-shot learning.

1 Introduction

Human-centered video generation [10, 15, 17] is a challenging task in computer vision. Human image animation, owing to its adaptable and controllable characteristics, holds significant potential for application across various domains, including social media, movie industry, etc. This task aims to generate a video whose appearance is consistent with the reference image based on a sequence of motion signals (e.g. depth, pose, and mesh). The recent advent of diffusion models [2, 6, 14] has shown its superiority in this field.

However, current human image animation methods [4, 7, 16], often train an image encoder to extract appearance information from a single image and discards the text condition, which significantly reduces the flexibility and freedom of video generation. More importantly, a single image from a single viewpoint only provides partial reference character information for the generation of complex and changeable human video and is easily limited to its lack of dynamic information. It is difficult for the model without sufficient

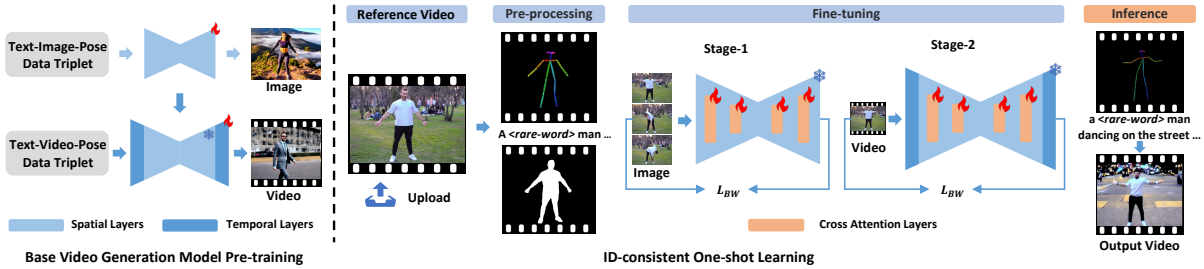

Figure 2: An overview of our AnimateAnywhere pipeline. The base video generation model is initialized with the pre-trained text-to-image model [12] and integrates temporal layers [5] (left panel). We conduct two stages of fine-tuning on the reference video for ID-consistent one-shot learning (right panel). First, we extract masks and poses from the reference video and use GPT-4o [1] to generate a text description. Then, we fine-tune the model to map "a <rare-word> man" to the character. During inference, users can customize the video's context and poses using prompts and pose videos.

prior knowledge to accurately predict the overall appearance of the reference character, which leads to artifacts or ID inconsistencies.

To better preserve the reference character's details and allow greater freedom in video generation, we propose a training-based approach that fine-tunes a base video generation model on a user-provided reference video, without needing extensive image encoder training. This method captures the character's appearance more comprehensively while retaining text-based control over the video's context. Specifically, we introduce a base video generation model conditioned on text and pose sequences [3], along with a two-stage fine-tuning strategy for ID-consistent one-shot learning. As shown in Fig. 1, the user can generate an ID-consistent video with customized context and poses by prompt and pose video.

2 Methodology

Our goal is to provide an ID-consistent video generation scheme conditional on text and poses through one-shot learning. To this end, we first design a text and pose guided video generation model in Sec. 2.1, as the left panel in Fig. 2; Besides, we propose a two-stage fine-tuning strategy to learn the appearance concepts of the reference character in Sec. 2.2, as the right panel in Fig. 2.

2.1 Base Video Generation Model

To improve pose alignment, we concatenate the pose latent directly with the noisy latent on the channel dimension, following the HumanSD [9]. This approach provides stronger conditioning for denoising and is more parameter-efficient than ControlNet-based [18] methods. To reduce computational costs and leverage existing image generation models, we use a two-stage training strategy. First, we train a text-to-image model on text-image-pose triplets for text and pose-guided image generation. Then, we integrate temporal layers and train on text-video-pose triplets for temporal modeling.

2.2 ID-Consistent One-shot Learning

To enhance ID consistency in video generation, we propose a two-stage one-shot learning strategy that avoids extensive image encoder training while maintaining text control over the video context. First, we extract poses and text captions from the reference video to build a fine-tuning dataset, using captions with a unique identifier and character class (e.g., "A <special*-new*> man"), similar to DreamBooth [13]. In the first stage, we treat all video frames as images and fine-tune an intermediate image generation model to learn the character's appearance. In the second stage, we integrate trained temporal layers the model obtained in the first stage and continue fine-tuning on video data. Only the cross-attention layers are trained, keeping other parameters frozen to preserve the model's prior knowledge. This two-stage fine-tuning approach allows for more effective learning of appearance details compared to direct one-step video fine-tuning.

Direct fine-tuning often causes overfitting because the model struggles to decouple the character from the background. To address this, we propose **background weakening loss** and **mask-guided attention**. Using an off-the-shelf model [11], we obtain foreground and background masks M_{fore} and M_{back} for each frame. We then reduce the influence of the background in the target loss, enhancing the focus on learning the character.

$$L_{BW} = \mathbb{E}\left(\left\|M_{fore} * (\epsilon - \epsilon_\theta)\right\|_2^2\right) + \mathbb{E}\left(\left\|\alpha * M_{back} * (\epsilon - \epsilon_\theta)\right\|_2^2\right) \quad (1)$$

where ϵ_θ represents the function of the denoising UNet, α set as 0.001, indicates the degree of weakening of background loss.

To reduce the mutual influence between character and background during fine-tuning, we implement **mask-guided attention** by assigning weights to the original attention based on the obtained mask. This guides each part to focus on itself, enabling the model to distinguish between the character and background, thus improving decoupling learning of the character in the reference video.

Figure 3: Qualitative comparisons with VideoBooth.

3 System Implementation

The base video generation model is trained for 20k steps on image data and 14k steps on video data using 8 NVIDIA A100 GPUs. For one-shot learning, we resize the training data to 512, enabling completion with a single NVIDIA 4090, while incorporating regularization data to prevent overfitting. We preserve character details better than image encoder-based methods [8], as shown in Fig. 3.

This work is sponsored by NO.XDB0690302.

References

[1] 2024. GPT-4o. https://openai.com/index/hello-gpt-4o/.

[2] Ankan Kumar Bhunia, Salman H. Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. 2023. Person Image Synthesis via Denoising Diffusion Model. In *CVPR*. IEEE, 5968–5976.

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. In *CVPR*. IEEE Computer Society, 1302–1310.

[4] Di Chang, Yichun Shi, Quankai Gao, Jessica Fu, Hongyi Xu, Guoxian Song, Qing Yan, Xiao Yang, and Mohammad Soleymani. 2023. MagicDance: Realistic Human Dance Video Generation with Motions & Facial Expressions Transfer. *CoRR* abs/2311.12052 (2023).

[5] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. 2023. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. *CoRR* abs/2307.04725 (2023).

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *NeurIPS*.

[7] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. 2023. Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Character Animation. *CoRR* abs/2311.17117 (2023).

[8] Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. 2023. VideoBooth: Diffusion-based Video Generation with Image Prompts. *CoRR* abs/2312.00777 (2023).

[9] Xuan Ju, Ailing Zeng, Chenchen Zhao, Jianan Wang, Lei Zhang, and Qiang Xu. 2023. HumanSD: A Native Skeleton-Guided Diffusion Model for Human Image Generation. In *ICCV*. IEEE, 15942–15952.

[10] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. 2023. DreamPose: Fashion Image-to-Video Synthesis via Stable Diffusion. In *ICCV*. IEEE, 22623–22633.

[11] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. 2024. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. *CoRR* abs/2401.14159 (2024).

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*. IEEE, 10674–10685.

[13] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *CVPR*. IEEE, 22500–22510.

[14] Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. 2023. Disco: Disentangled control for realistic human dance generation. *CoRR* abs/2307.00040 (2023).

[15] Yaohui Wang, Piotr Bilinski, François Brémond, and Antitza Dantcheva. 2020. G3AN: Disentangling Appearance and Motion for Video Generation. In *CVPR*. Computer Vision Foundation / IEEE, 5263–5272.

[16] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. 2023. MagicAnimate: Temporally Consistent Human Image Animation using Diffusion Model. *CoRR* abs/2311.16498 (2023).

[17] Quanwei Yang, Xinchen Liu, Wu Liu, Hongtao Xie, Xiaoyan Gu, Lingyun Yu, and Yongdong Zhang. 2022. REMOT: A Region-to-Whole Framework for Realistic Human Motion Transfer. In *ACM Multimedia*. ACM, 1128–1137.

[18] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *ICCV*. IEEE, 3813–3824.