

UniAnimate-DiT: Human Image Animation with Large-Scale Video Diffusion Transformer

Xiang Wang¹ Shiwei Zhang² Longxiang Tang³ Yingya Zhang² Changxin Gao¹
Yuehuan Wang¹ Nong Sang¹

¹Key Laboratory of Image Processing and Intelligent Control,
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology
²Alibaba Group ³Tsinghua University

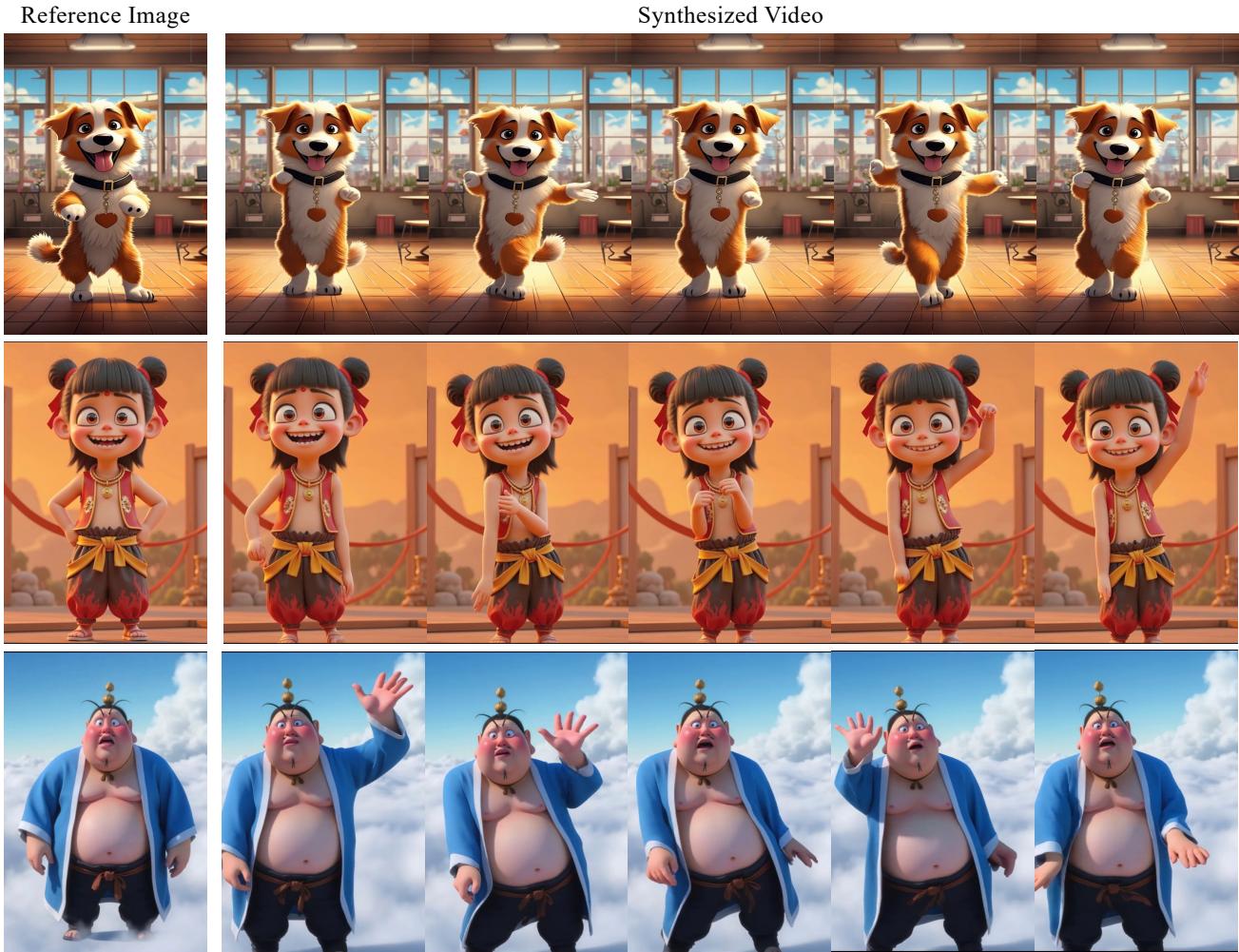


Figure 1. **Image animation examples** synthesized by the proposed UniAnimate-DiT with Wan2.1-I2V-14B [10] as the base model.

Abstract

This report presents UniAnimate-DiT, an advanced project that leverages the cutting-edge and powerful ca-

pabilities of the open-source Wan2.1 model for consistent human image animation. Specifically, to preserve the robust generative capabilities of the original Wan2.1 model, we implement Low-Rank Adaptation (LoRA) technique to

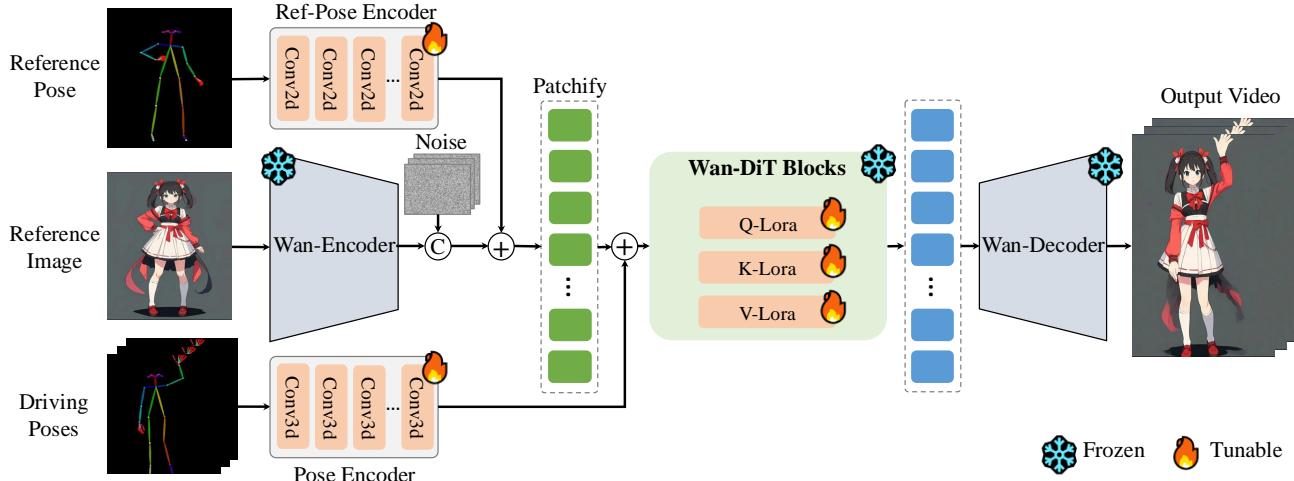


Figure 2. The overall architecture of the proposed UniAnimate-DiT base on Wan2.1 model.

fine-tune a minimal set of parameters, significantly reducing training memory overhead. A lightweight pose encoder consisting of multiple stacked 3D convolutional layers is designed to encode motion information of driving poses. Furthermore, we adopt a simple concatenation operation to integrate the reference appearance into the model and incorporate the pose information of the reference image for enhanced pose alignment. Experimental results show that our approach achieves visually appealing and temporally consistent high-fidelity animations. Trained on 480p (832x480) videos, UniAnimate-DiT demonstrates strong generalization capabilities to seamlessly upscale to 720P (1280x720) during inference. The training and inference code is publicly available at <https://github.com/ali-vilab/UniAnimate-DiT>.

1. Introduction

Human image animation has undergone significant advances with the convergence of generative modeling techniques [7, 8, 14], especially with the rise of diffusion models [2–6, 12, 13, 15, 16]. This task aims to enable the transformation of a static reference image into dynamic video sequences that depict lifelike, temporally consistent movements that adhere to the guidance of driving poses.

Traditional methods [4, 9, 12, 13] in this domain often leverage a 3D-UNet base model to generate videos, struggling with temporal coherence and realism. For example, UniAnimate [12] presents a unified framework based on a 3D-UNet TF-T2V [11] model to encode reference appearance and motion movement. The animation performance may be constrained by the capability of the base model. This motivates a shift towards more advanced video generative models. The transition to a more advanced Diffusion

Transformer (DiT)-based model, such as Wan2.1 [10], provides a potential direction to enhance the quality of generated videos.

To this end, this report aims to fulfill this gap and presents UniAnimate-DiT, a simple but effective framework based on Wan2.1 for consistent human image animation. Specifically, we employ LoRAs to finetune a smaller set of model parameters, reducing training memory overhead while maintaining the original model’s generative potency. A lightweight pose encoder consisting of multiple stacked 3D convolutional layers is used to encode driving motion information. In addition, the reference pose information is also incorporated to enhance appearance alignment. Qualitative experimental results (Fig. 1) show that our approach achieves visually appealing and temporally consistent high-fidelity animations. Despite being trained at 480P (832x480) video resolution, UniAnimate-DiT has the ability to seamlessly upscale to 720P (1280x720) during inference.

2. Method

The overall architecture of UniAnimate-DiT is displayed in Fig. 2, which is designed to adapt the human image animation task by incorporating the advanced video DiT model. The framework comprises several key components: 1) Video Diffusion Transformer Model (Wan2.1): The video DiT model serves as the primary generative engine, providing a robust mechanism for high-quality output generation. 2) LoRA Fine-tuning: Low-Rank Adaptation (LoRA) is an efficient parameter tuning technique that reduces the number of trainable parameters. Through LoRA, we fine-tune a limited number of parameters, reducing memory overhead and enhancing adaptability without sacrificing performance. 3) Pose Encoder: This lightweight

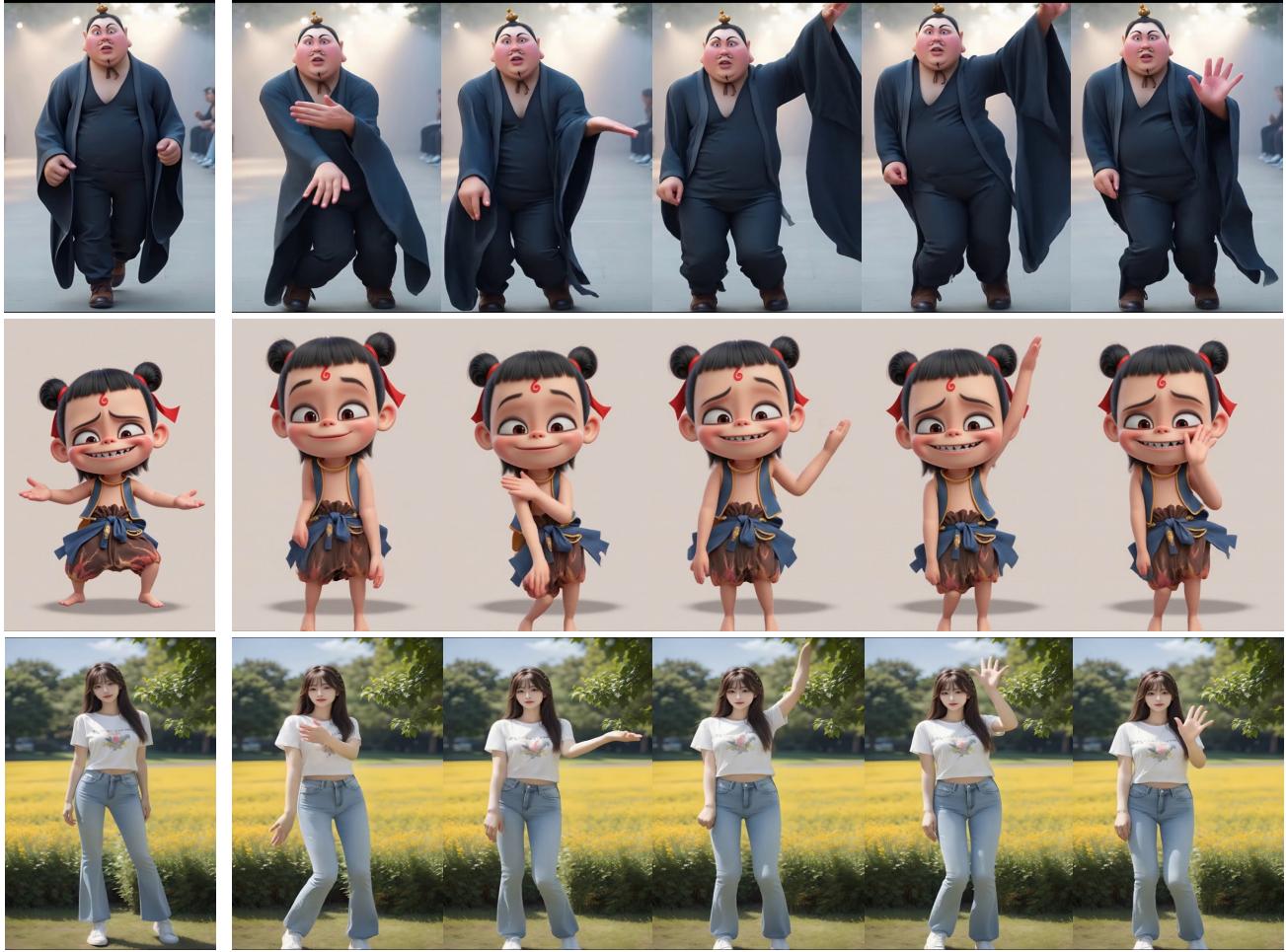


Figure 3. Video cases synthesized by the proposed UniAnimate-DiT.

module consists of multiple stacked 3D convolutional layers designed to extract temporal and spatial features of driving poses effectively. 4) Ref-Pose Encoder: the reference pose information is also incorporated by stacked 2D convolutional layers to enhance appearance alignment. Reference pose information is incorporated by summing it with a noisy latent vector.

Discussions. The pose encoder is a critical component that enables accurate representation of human poses and movement dynamics. In our setting, the encoder consists of seven layers of 3D convolutions. Through experiments, we observe that a lower number of layers (e.g., four layers) resulted in a limited receptive field, which hinders the model’s ability to control the generated animation accurately. Thus, a deeper architecture improves the model’s understanding of temporal contexts and enhances motion control. Initially, the driven pose features were concatenated directly with the noisy latent vector (which is 16-dimensional), resulting in ineffective control over the generated animations due to its limited feature representation capacity. Recognizing this

limitation, we try to inject reference pose information into the model at a more meaningful feature level. Finally, we choose to integrate pose information into the patchified tokens (5120-dimension). This adjustment provides a significantly richer representation, improving the model’s ability to learn and control detailed pose characteristics during the generation phase.

Long video generation. Our UniAnimate-DiT also supports long video generation by applying the overlapped slide window strategy [1]. The first frame feature after the Wan-VAE only represents one frame. To improve the consistency between each window, the first two frame features of the subsequent windows are discarded.

3. Experiments

3.1. Dataset and setup

We collect a video dataset that contains about 10K human dance videos to train our UniAnimate-DiT. The dataset used for training consists of diverse human images

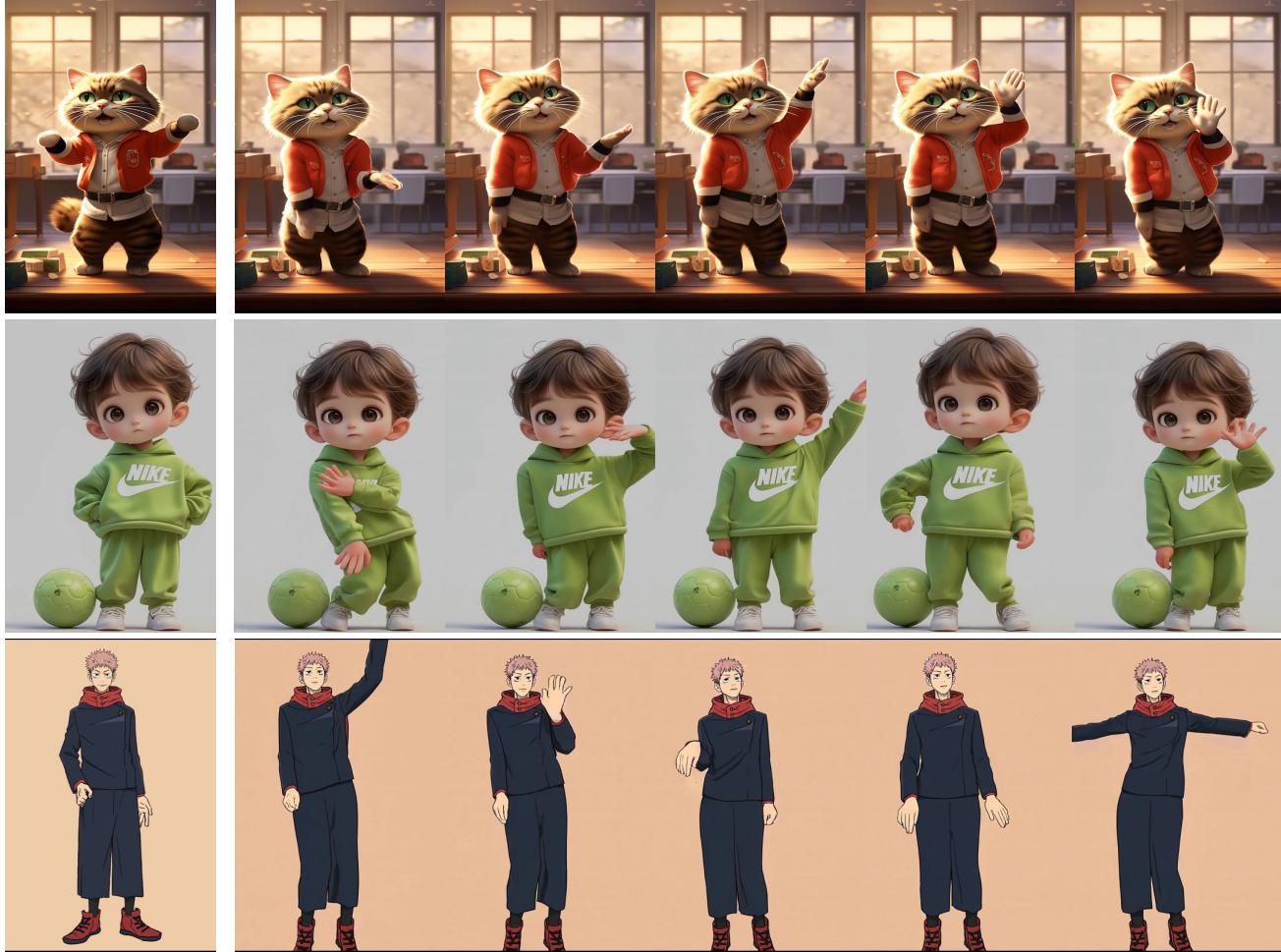


Figure 4. Video cases synthesized by the proposed UniAnimate-DiT.

annotated with key poses and spans a wide range of actions and lighting conditions, allowing the model to learn a comprehensive set of animation dynamics. In our experiments, 6 Nvidia A800 GPUs are leveraged. The model is trained at 832x480 resolution due to GPU memory restriction. Please refer to our open source code for more training and inference details.

3.2. Qualitative evaluation

An important feature of our framework is its capability to generalize to large resolutions during inference. While training occurs at 480P, our model can upscale outputs to 720P effortlessly. This scalability broadens the practical applications of our technology for higher-resolution video generation. In Figs. 3 and 4, we show the qualitative evaluations of our method. The generated videos reveal a high degree of fidelity and continuity, demonstrating lifelike movements. These results indicate the effectiveness of the proposed UniAnimate-DiT.

4. Conclusion

UniAnimate-DiT establishes a significant advancement in human image animation by effectively combining the capabilities of large-scale video DiTs with efficient fine-tuning techniques. Our open-source implementation provides a valuable resource for researchers and developers in the field, further enabling the evolution of realistic animation technologies.

Acknowledgment

This work is supported by the National Natural Science Foundation of China under grants U22B2053 and 623B2039, and Alibaba Group through Alibaba Research Intern Program.

References

- [1] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled im-

- age generation. *Proceedings of Machine Learning Research*, 202:1737–1752, 2023. 3
- [2] 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 *CVPR*, pages 22563–22575, 2023. 2
- [3] 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.
- [4] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *CVPR*, pages 8153–8163, 2024. 2
- [5] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion video synthesis with stable diffusion. In *ICCV*, pages 22680–22690, 2023.
- [6] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanyvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 2
- [7] Minghui Lin, Xiang Wang, Yishan Wang, Shu Wang, Fengqi Dai, Pengxiang Ding, Cunxiang Wang, Zhengrong Zuo, Nong Sang, Siteng Huang, et al. Exploring the evolution of physics cognition in video generation: A survey. *arXiv preprint arXiv:2503.21765*, 2025. 2
- [8] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023. 2
- [9] Shuai Tan, Biao Gong, Xiang Wang, Shiwei Zhang, Dandan Zheng, Ruobing Zheng, Kecheng Zheng, Jingdong Chen, and Ming Yang. Animate-x: Universal character image animation with enhanced motion representation. In *ICLR*, 2025. 2
- [10] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1, 2
- [11] Xiang Wang, Shiwei Zhang, Hangjie Yuan, Zhiwu Qing, Biao Gong, Yingya Zhang, Yujun Shen, Changxin Gao, and Nong Sang. A recipe for scaling up text-to-video generation with text-free videos. In *CVPR*, 2024. 2
- [12] Xiang Wang, Shiwei Zhang, Changxin Gao, Jiayu Wang, Xiaoli Zhou, Yingya Zhang, Luxin Yan, and Nong Sang. Unianimate: Taming unified video diffusion models for consistent human image animation. *Science China Information Sciences*, 2025. 2
- [13] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *CVPR*, pages 1481–1490, 2024. 2
- [14] Ceyuan Yang, Zhe Wang, Xinge Zhu, Chen Huang, Jianping Shi, and Dahua Lin. Pose guided human video generation. In *ECCV*, pages 201–216, 2018. 2
- [15] 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
- [16] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *ECCV*, 2024. 2