Animate Anyone: Reproduction, Analysis and Extensions

Chen Shiyang
3035974432
whatever@connect.hku.hk

Li Ka Lam
3036058730
u3605873@connect.hku.hk

Chechneva Kseniia
3036465036
u3646503@connect.hku.hk

Abstract

This report presents a comprehensive reproduction and analysis of Animate Anyone [Hu et al., 2023], a state-of-the-art framework for pose-driven character animation, alongside an extension addressing its facial animation limitations. Using Moore Threads' unofficial implementation [2024], we evaluate the model's performance on standard benchmarks (UBC Fashion [Zablotskaia et al., 2019], TikTok [Siarohin et al., 2021]) and a custom dataset of diverse characters (humans, cartoons, robots). While our reproduction matches the original claims for in-distribution human videos—showing strong temporal consistency and detail preservation—it also exposes several key limitations: (1) facial expressions remain static or unnatural without explicit keypoints, (2) generalization falters on non-humanoid or stylized subjects, and (3) scale and pose mismatches cause artifacts like motion ghosting. To address the facial animation gap, we implement Moore's Face Reenactment module [2024], which allows independent control of head, eye, and mouth movements via OpenSeeFace. Results show improved lip sync and blink dynamics but still display lighting sensitivity and high-motion artifacts. Our work highlights discrepancies between the original paper's demonstrations and reproducible outcomes, emphasizing the need for transparency in training data and more extensive robustness testing.

1 Introduction

In the era of rapid AI advancement, a growing number of AI tools are being developed to enhance various aspects of human life. Among these innovations, image-to-video synthesis - which focuses on generating high-quality, realistic, and temporally coherent videos from images - has gained significant attention across various domains, including social media content creation, video game development, and medical imaging for scientific research. For this course project, our group chose to analyze, reproduce and extend the ideas presented in the paper titled "Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Character Animation" [Hu et al., 2023].

Animate Anyone is a state-of-the-art technique inspired by recent advances in diffusion models, which have demonstrated superior capabilities in high-quality image generation and have been adapted for video synthesis. However, previous image-to-video models - such as DeamPose and Disco - struggle with issues such as maintaining temporal coherence, ensuring consistent appearance, providing precise motion control, and generalizing to a wide range of characters. To address these limitations, the authors introduce three key components: ReferenceNet, Pose Guider, and a Temporal

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Layer. These modules, detailed in Section 3, are designed to enhance consistency, controllability, and robustness in character animation.

In this course project, we aim to critically assess whether the performance of the Animate Anyone model aligns with the results and claims reported in the original paper. We further evaluate its overall performance, focusing on robustness and generalizability. This report is organized as follows: Section 2 reviews related work in character animation. Section 3 details our reproduction pipeline, implementation process, and encountered challenges. Section 4 presents our reproduction results and analysis. Section 5 discusses potential extensions of the model, and Section 6 concludes the report.

2 Related Works

2.1 Animate Anyone

The proposed method, Animate Anyone [Hu et al., 2023], introduces a diffusion-based framework for high-fidelity image animation. By integrating temporal attention layers with a reference mechanism, the model synthesizes coherent animations from static images and pose sequences. Another key innovation would be its ability to preserve subject identity and fine details across frames, such as the clothing texture, addressing temporal inconsistency effectively. Evaluations highlight its robustness on diverse human subjects, though computational demands remain a limitation.

2.2 DreamPose

This model leverages diffusion models for fashion-centric animation, enabling pose-guided editing of apparel images [Karras et al., 2023]. Unlike Animate Anyone, which processes full-body motion, DreamPose introduces a lightweight adapter to fine-tune pretrained diffusion models on small datasets, prioritizing computational efficiency. While it is effective for fashion applications, its focus on static poses limits utility for dynamic animation tasks.

2.3 DisCo

DisCo separated motion into skeletal (i.e., pose) and non-skeletal (i.e., background, clothing dynamics) factors via a two-stage pipeline [Wang et al., 2023]. While this enables granular control over animation elements, its reliance on explicit decomposition introduces complexity which is absent in Animate Anyone's end-to-end approach. DisCo also requires dense motion annotations, limiting scalability of the model.

2.4 Bipartite Deformable Motion Model (BDMM)

This method utilizes a vector-quantized variational autoencoder (VQVAE) to separate motion into global and local components [Yu et al., 2023]. For instance, body trajectories are regarded as global movements while limb movements are considered local motions. This method excels in modelling complex motions, but struggles with preserving fine details, often producing blurry outputs which would be less effective compared to Animate Anyone's diffusion-based refinement.

2.5 Temporal Convolutional Attention-based Network (TCAN)

TCAN combines temporal convolutional networks (TCNs) with attention mechanisms to capture both local and global motions, which is a sequence modeling framework specifically designed for tasks requiring long-range temporal dependencies [Hao et al., 2020], while Animate Anyone is a diffusion=based approach. TCAN's deterministic approach favors consistency while Animate Anyone favors diverse and high-fidelity outputs. Additionally, TCAN struggles with non-rigid deformations which indicate that it is lower in terms of scalability, due to its lightweight convolutional design.

3 Reproduction of the Original Method

3.1 Overview of the method

Animate Anyone is a framework builds on latent diffusion model (LDM) for character animation, which is designed to generate high-fidelity, temporally coherent videos with two inputs: a single reference image of the character, and a sequence of target motions. This architecture employs a two-stage training process: Firstly, deploy a base diffusion such as the Stable Diffusion model to learn spatial priors from image-pose pairs; Secondly, a temporal module is introduced during model fine-tuning process to enforce inter-frame coherence. The flow of the training strategy as well as the blocks of the model architecture are shown in Figure 1.

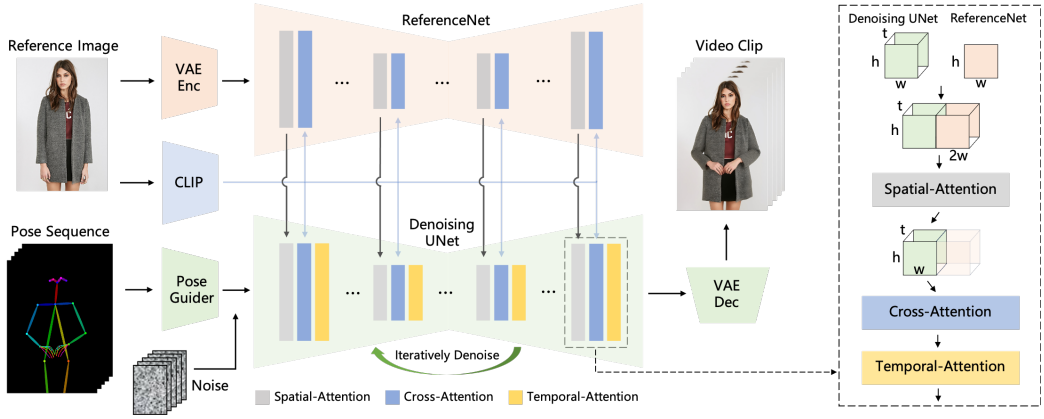

Figure 1: This serves as an overview of the Animate Anyone architecture.

There are three key components introduced in this model to solve common issues in image-to-video synthesis, including ReferenceNet which was invented to preserve intricate appearance features, Pose Guilder which was introduced to integrate pose control signals during the denoising process, and the temporal attention mechanism which utilizes a 3D U-Net that processes multiple frames simultaneously to ensure smooth motion transitions and long-range dependencies for periodic motions.

For the inference process of Animate Anyone, there are four steps in total: Firstly, the input image is encoded by the VAE encoder and CLIP respectively, while the pose sequence is encoded by the Pose Guilder. Secondly, iterative denoising is carried out, in which the diffusion model synthesizes individual frames conditioned on the extracted spatial, semantic and pose features. The first two steps are regarded as the first training stage of the model. Thirdly, the temporal module is involved in the denoising process in model fine-tuning to ensure inter-frame consistency. Lastly, the final latent sequence is decoded using VAE decoder to generate the result video.

3.2 Overview of our implementation

To investigate the performance of Animate Anyone, our group utilize an unofficial repository by Moore Treads [2024], as the official source code is not publicly available. This implementation, intended to approximate 80% of the original model's reported performance, presented several limitations, including background artifacts and suboptimal results due to scale mismatch between reference image and pose sequences.

Our experiments were conducted on Google Colab, following guild lines from the repository. We have utilized the tools and libraries required in the repository, including diffusers (0.24.0), transformers (4.30.2), along with open-clip-torch (2.20.0), and controlnet-aux (0.0.7). Resolving library version conflicts required considerable effort during the environment setup.

Regarding data, the official Animate Anyone model was trained using an internal dataset with 5K character video clips which is not publicly accessible. Therefore, we employed two widely-used human video synthesis benchmarks - UBC fashion video data [Zablotskaia et al., 2019] and TikTok

dataset [Siarohin et al., 2021] - as well as a custom dataset we created to evaluate generalizability and robustness. Further details on these datasets are provided in Section 4.

3.3 Challenges encountered in reproduction

As mentioned in the Animate Anyone paper, the model was trained using a massive database of videos clips and experiments are conducted on 4 NVIDIA A100 GPUs. In contrast, our resources were more limited—even with a Google Colab Pro subscription, GPU availability was insufficient for efficient reproduction. For example, we frequently encountered interruptions due to running out of GPU memory during inference.

4 Experimental Results & Analysis

In this section, we will evaluate results derived from Moore Theads' unofficial implementation [2024] on the datasets mentioned in Section 3. We find that while the model achieves impressive detail preservation and temporal consistency on in-distribution human videos, its generalization ability is overstated in the original paper. In particular, the method struggles with realistic facial animation, adapting to different character types, and handling mismatches between the reference and driving pose sizes—limitations that become especially obvious when testing on out-of-distribution or stylized subjects. Additionally, our findings suggest potential bias in the original results, possibly due to overlap between training and test data or selective example reporting. The lack of access to official model weights and training data, along with resource constraints, further affects the reproducibility and robustness of our evaluation.

4.1 Reproduction vs. Original

4.1.1 Experimental Results

UBC Fashion Dataset Using the pre-trained model from Moore Threads' Animate Anyone implementation 2024, we reproduced experiments on the UBC Fashion dataset [Zablotskaia et al., 2019]. As shown in Figure 2, our qualitative results are consistent with the claims of the original paper: clothing details, such as structural elements and pattern designs, are effectively preserved and maintain consistency throughout the video sequence.

TikTok Dataset Testing on the TikTok dataset [Siarohin et al., 2021], we observed that the generated outputs exhibited strong temporal consistency, with smooth transitions and coherence from frame to frame. Distortion, jitter, and flicker were minimal, and any unnatural artifacts remained within acceptable limits, not affecting the overall perceptual quality. These results further support the model's effectiveness in maintaining visual stability and realism on human-centric, in-distribution benchmarks (see Figure 3).

4.1.2 Experimental Analysis

Overall, our reproduction results are consistent with the claims presented in the original paper, demonstrating a strong capacity to maintain both temporal and detail consistency. However, several limitations and concerns were identified.

Facial Expression Limitation A notable discrepancy concerns the quality of facial animation. In the official demos, the generated videos demostrate diverse and natural facial expressions—such as eye blinking, subtle mouth movements, and even realistic hair motion and shadow with different level of depth — despite the absence of explicit facial keypoints in the input pose sequence (See Apendix 1). In contrast, when we reproduced these scenarios by intentionally removing facial points from the pose skeleton (to match the OpenPose format), our outputs consistently resulted in blurred and featureless faces, lacking any realistic dynamics or expressions (See Figure 4). Even when facial keypoints were provided using DWPose, the reproduced facial expressions remained largely static, aside from occasional unnatural angular changes. (See Figure 5). This is in contrast to the diverse and dynamic facial animations shown in the official demonstrations.

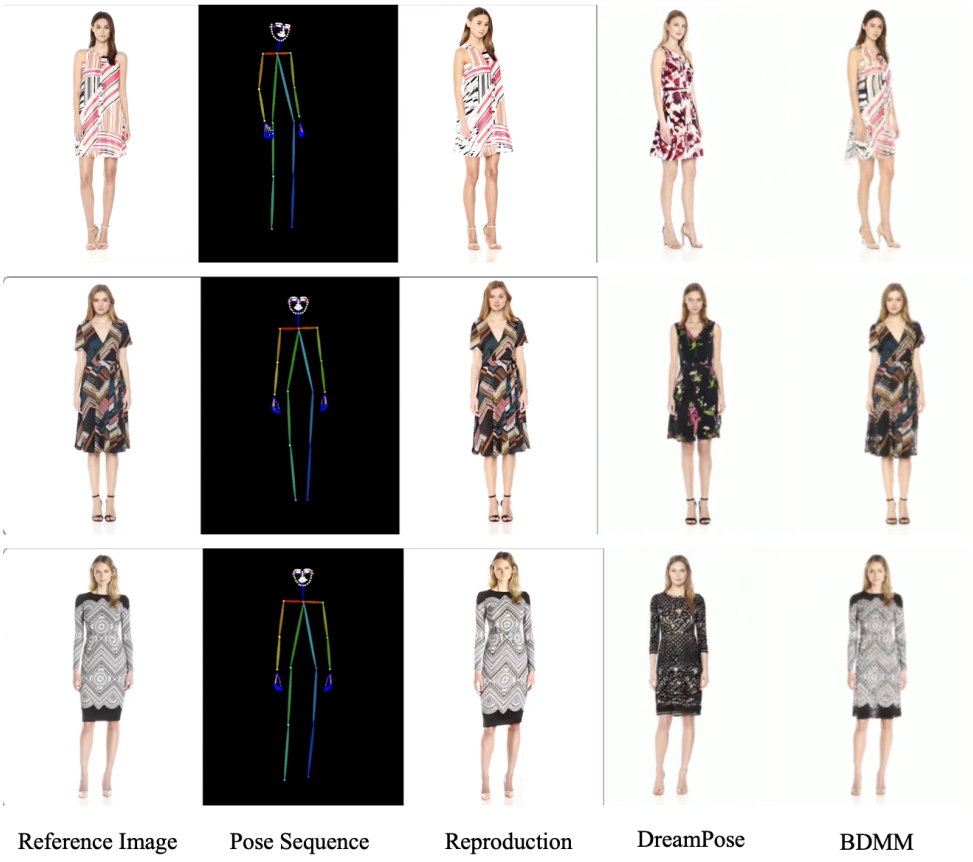

Figure 2: Qualitative comparison on the UBC Fashion dataset. Our reproduced results demonstrate strong detail preservation and temporal consistency, closely matching the original paper's claims. Competing methods (DreamPose, BDMM) fail to accurately capture clothing details and maintain high resolution over the video sequence.

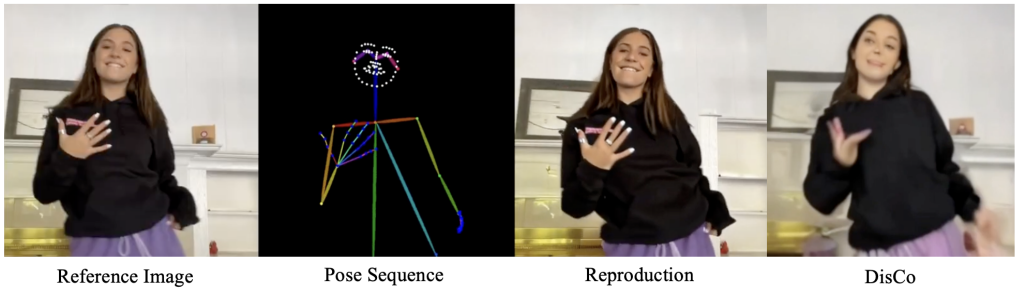

Figure 3: Qualitative comparison on the TikTok dataset. Our reproduction achieves smooth transitions and strong temporal consistency, with minimal distortion or flicker. In contrast, DisCo outputs display pronounced flickering, unstable frame-to-frame transitions, and frequent visual artifacts, resulting in noticeably lower video quality.

Figure 4: When facial keypoints are removed, our reproductions produce blurred, expressionless faces with no realistic dynamics.

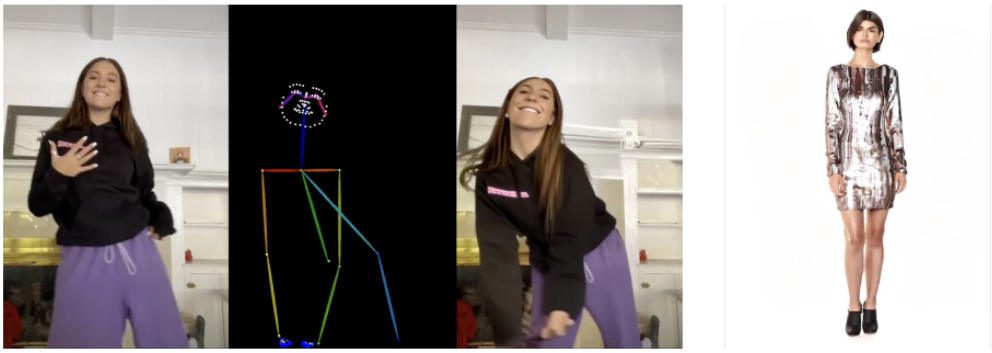

Figure 5: When facial keypoints are included, the generated faces remain mostly static and closely resemble the reference image, with occasional unnatural and abrupt changes in facial orientation.

Suspected Overfitting or Implicit Biases We suspect that some of the visually impressive results shown in the original paper may be influenced by overfitting or dataset bias. The reference images and driving videos likely come from the same dataset, often sharing similar scene composition and pose scales. Without access to the exact training set, it is unclear whether the reported results include train-test overlap or if "cherry-picked" examples were used. In our reproduction—using only publicly available pre-trained weights and strictly following the pose input protocol—we were unable to achieve results as good as those shown in the paper. This raises questions about the generalization ability of the method and whether similar test and training data contributed to the strong official results.

4.2 Robustness Evaluation

So far, the results in the original paper have been reported only on standard public benchmarks, with the potential for significant overlap between training and testing data. To better assess generalization and robustness, we constructed a custom dataset and evaluated the model on this new data without any additional pre-training.

4.2.1 Customized Dataset Summary

To test the method's generalization, we created a customized dataset with 12 videos covering a wide range of movements, from half-length to full-length actions. We also gathered 20 reference images featuring diverse subjects, including real humans, humanoid characters, cartoons, robots, 3D models, and artistic styles (See Appendix 2). This diverse collection was designed to challenge the model with both familiar and novel scenarios, providing a more thorough evaluation of its robustness.

Figure 6: Representative results showing the model's limitations: the robot (left) and stylized human (second) exhibit distorted or unstable outputs due to unreliable pose estimation, while only the right two in-distribution human figures yield usable animations. This demonstrates the model's reliance on human-like inputs and its difficulty generalizing to non-humanoid or highly stylized characters.

4.2.2 Experimental Results

Qualitative results are shown in Figure 6. Our analysis shows that the model is ill-suited for animating non-human or highly stylized characters. In such cases, pose sequences generated by DWPose are frequently unreliable, leading to severe temporal flickering and unrecognizable outputs. Only sequences involving human figures yield usable results, and even then, the model's performance is highly sensitive to the alignment of pose and scale between the reference and driving inputs. Mismatches in figure size or pose configuration routinely result in "motion ghosting," unstable backgrounds, and various visual artifacts. Attempts to animate non-humanoid figures, such as robots, consistently fail to reconstruct facial and body features, producing outputs that are distorted and lacking in semantic coherence.

4.2.3 Experimental Analysis

Several factors may explain these limitations. First, using Moore Threads' unofficial codebase and pre-trained weights [2024] means we don't know how diverse or high-quality the original training data was. Our results suggest the model hasn't been exposed to a wide variety of samples, which may limit its generalization. Due to resource and time constraints, we couldn't retrain the model or run large-scale quantitative tests. Producing enough outputs for thorough evaluation was computationally expensive, and inconsistent output frame sizes made direct comparisons challenging.

5 Extensions

One of the primary shortcomings of the original Animate Anyone [Hu et al., 2023] framework was its limited ability to generate accurate facial movements. While the method excelled at full-body pose transfer, subtle facial expressions - such as lip synchronization, eye blinks, and nuanced emotional cues - often appeared stiff or unnatural. To resolve this in, in this section Moore-AnimateAnyone code [2024] is implemented. Authors introduced a dedicated Face Reenactment module, significantly enhancing facial animation quality through landmark disentanglement, specialized architecture modifications, and optimized training protocols.

5.1 Key differences

Face Reenactment allows precise control of facial expressions in the original image using facial landmarks from the video driver. Unlike the basic solution, which focused on controlling the pose of the whole body, the new module specializes in detailed reproduction of facial expressions, including lip, eyebrow, and eye movements.

The key improvement was the separation of facial landmarks into two independent components: head movements (including eye blinking) and mouth movements. This approach provides more natural animation, allowing, for example, to accurately synchronize articulation with speech or reproduce complex emotions. To extract landmarks, the OpenSeeFace library is used, which tracks 68 key points of the face according to the FACS standard. This allows analyzing even subtle changes in facial expressions, such as a slight smile or a raised eyebrow.

Technically, the system was improved by introducing additional Pose Guiders to control head movements and to control mouth expressions. ReferenceNet, inherited from the original project, was adapted to work with facial landmarks, which improved the preservation of skin textures and other details. The output video is generated in 512 by 512 resolution, which provides a high level of detail. Datasets with various facial expressions are used to train the model: conversations, singing, expressions of emotions. Video preprocessing includes landmark extraction.

Figure 7: Improved facial animation and remaining artifacts. While facial characteristics are enhanced, issues such as background artifacts and motion-induced distortions—like the three-legged effect—persist in certain outputs.

5.2 Examples and results

Facial characteristics have improved, as shown in Figure 7. Despite the progress, the current version still has some limitations. Artifacts on a uniform background or slight shaking with subtle movements are possible, as illustrated in Figure 7, where the girl appears to have three legs instead of two.

6 Conclusion

Our reproduction of Animate Anyone confirms the model performs well in full body pose transfer and maintains temporal coherence for human subjects. However, we observed several limitations in facial animation and generalization to other character types. The lack of access to the original training data and high computational requirements also raise concerns about reproducibility. Additionally, the model struggles with non-human characters, which challenges the versatility claimed by the original paper. By integrating Moore's Face Reenactment, we demonstrate actionable improvements for expressive facial control, though challenges like jitter and occlusion sensitivity remain.

References

Hongyan Hao, Yan Wang, Siqiao Xue, Yudi Xia, Jian Zhao, and Furao Shen. Temporal convolutional attention-based network for sequence modeling. *arXiv preprint arXiv:2002.12530*, 2020. https://arxiv.org/abs/2002.12530.

Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023. https://arxiv.org/abs/2311.17117.

Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. *arXiv preprint arXiv:2304.06025*, 2023. https://arxiv.org/abs/2304.06025.

MooreThreads. Github - moorethreads/moore-animateanyone: Character animation (animateanyone, face reenactment), 2024. https://github.com/MooreThreads/Moore-AnimateAnyone/tree/master.

Aliaksandr Siarohin, Oliver J. Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. *arXiv preprint arXiv:2104.11280*, 2021. https://arxiv.org/abs/2104.11280.

Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for realistic human dance generation. *arXiv preprint arXiv:2307.00040*, 2023. https://arxiv.org/abs/2307.00040.

Wing-Yin Yu, Lai-Man Po, Ray C.C. Cheung, Yuzhi Zhao, Yu Xue, and Kun Li. Bidirectionally deformable motion modulation for video-based human pose transfer. *arXiv preprint arXiv:2307.07754*, 2023. https://arxiv.org/abs/2307.07754.

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.

Appendix

A Qualitative examples from the official Animate Anyone demos

Despite the absence of facial keypoints in the input, the model generates highly realistic and expressive results. As shown in the following examples, subtle details such as hair motion, facial expression changes, eye-blinking, and nuanced shadow depth are all well-captured, resulting in smooth and visually consistent outputs.

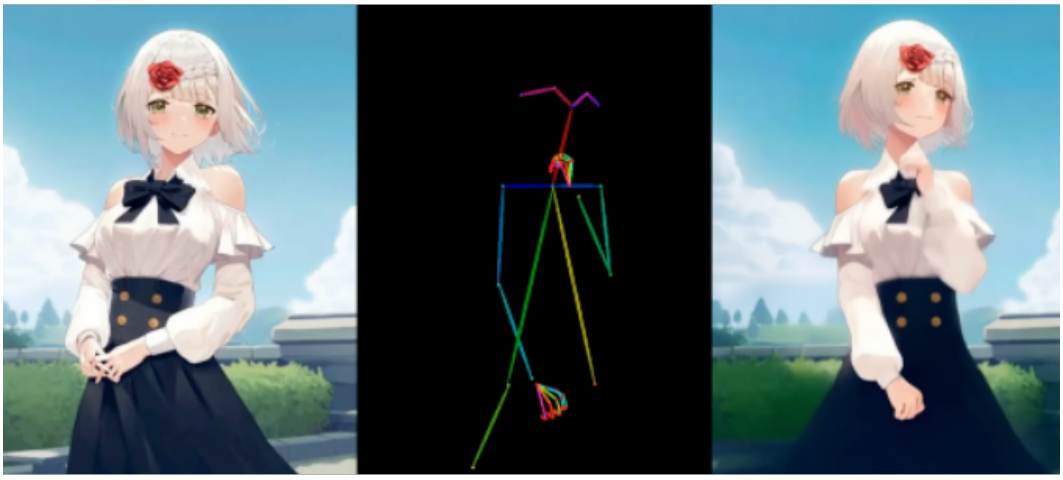

Figure 8: Natural head movements and changes in face orientation

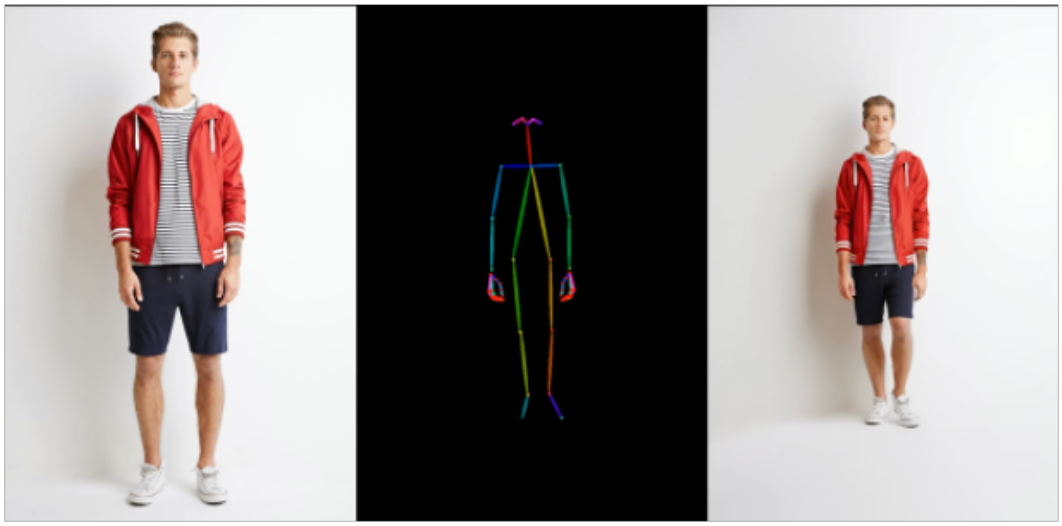

Figure 9: Smooth variations in shadow depth as the figure steps forward

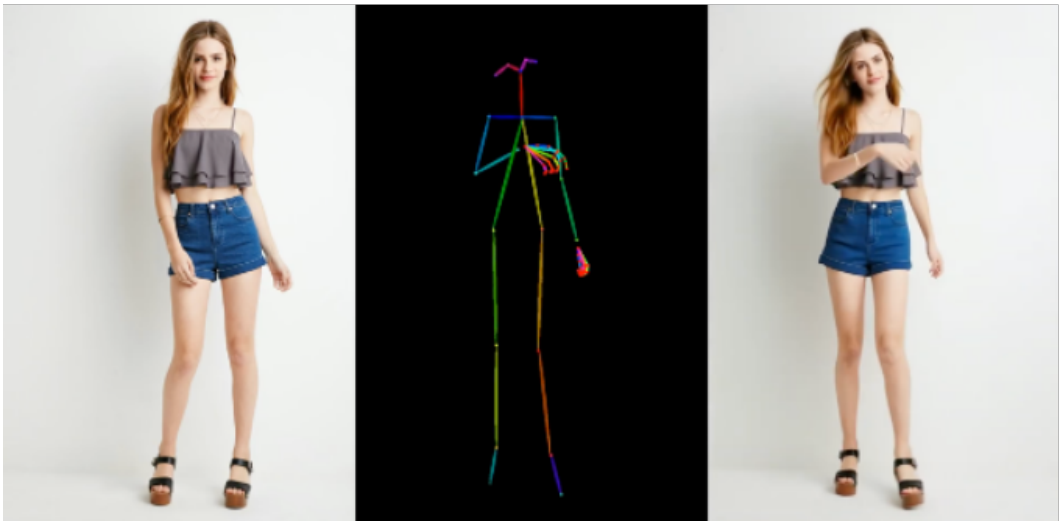

Figure 10: Realistic hair motion corresponding to changes in pose

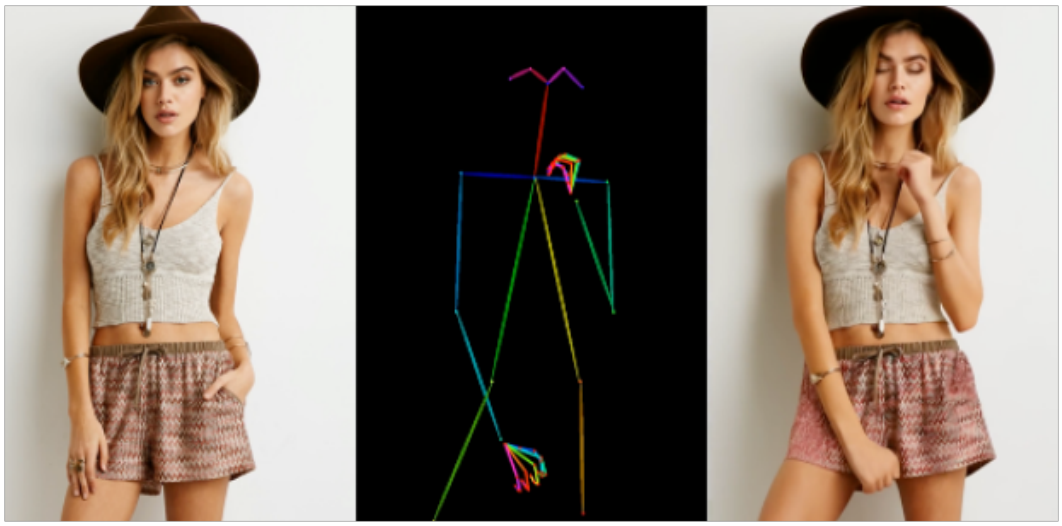

Figure 11: Natural and convincing eye-blinking

B Customized dataset examples

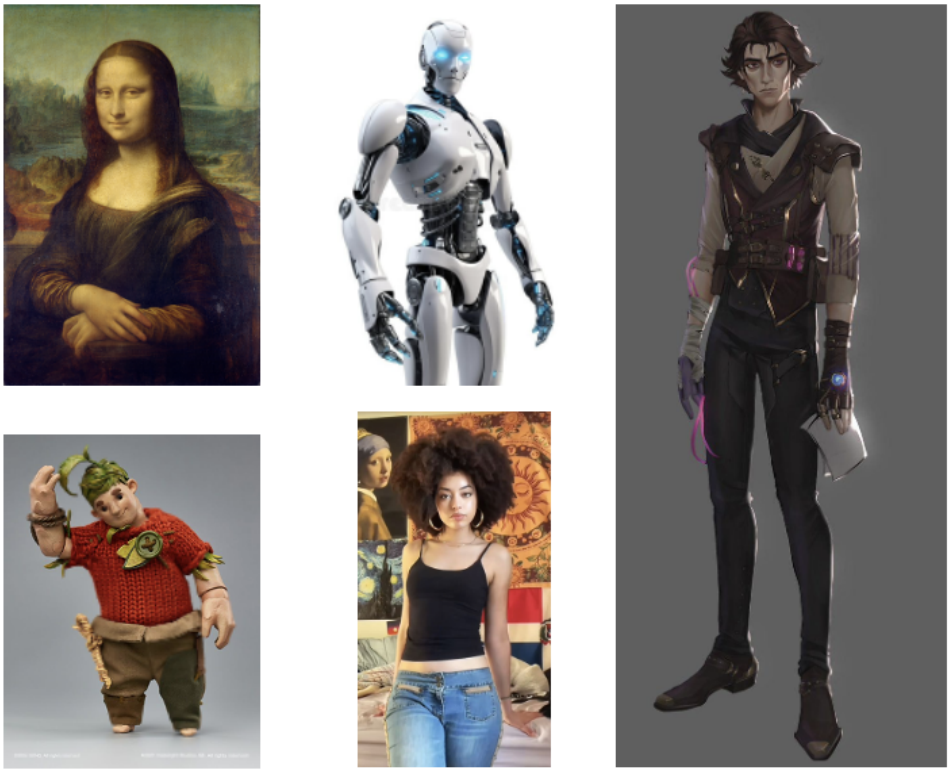

Figure 12: Example reference images from our custom evaluation set, featuring diverse subjects—including real humans, robots, cartoons, 3D models, and artistic styles—used to test the model's generalization and robustness.

C Work split table

Team member	**Work distribution**
All members	Paper & code understanding, try to implement the code via different methods
Li Ka Lam (Leader)	1. Schedule meetings and regular progress checking with members 2. Explain paper details and model architecture in the presentation 3. Responsible for Section 1, 2, 3 of the report.
Chen Shiyang	1. Reproduction of videos and experiment analysis 2. Present the reproduction results in the presentation 3. Responsible for Section 4 of the report
Chechneva Kseniia	1. Conduct research of potential extensions of Animate Anyone 2. Discuss the extensions in the presentation 3. Responsible for Abstract, Section 5 and 6 of the report

Table 1: Team Member Work Distribution