HunyuanPortrait: Implicit Condition Control for Enhanced Portrait Animation

Supplementary Material

A. Benchmark Metrics Details

We employ both qualitative and quantitative analyses to assess the generated video quality and motion accuracy of our portrait animation results. In evaluating self-reenactment, we consider multiple metrics: the Peak Signal-to-Noise Ratio (PSNR), the Structural Similarity Index (SSIM) [10], and the Learned Perceptual Image Patch Similarity (LPIPS) [12]. Specifically, for the LPIPS metric, we apply the AlexNet-based perceptual similarity measure LPIPS [12] to gauge the perceptual similarity between the generated animated images and the driving images. We also utilize the Fréchet Inception Distance (FID) [5] to assess image quality, the Fréchet Video Distance (FVD) [9] to evaluate temporal consistency, and the Landmark Mean Distances (LMD) to measure the accuracy of generated facial expressions. The landmarks are extracted using Mediapipe [6]. We compute the average Euclidean distance between the facial landmarks [6] of the reference and generated frames. Lower values indicate better geometric accuracy. FID [5] is utilized to measure the similarity in feature distribution between generated and real images, employing Inception-v3 features. Lower scores indicate better perceptual quality. Additionally, the FVD is used to evaluate temporal coherence through features extracted from a pretrained network [9]. For cross-reenactment, we utilize the ArcFace Score [1] as the identity (ID) similarity metric between the generated frames and the reference image. For Average Expression Distance (AED) [8] and Average Pose Distance (APD) [8], we calculates the Manhattan distance of expression and pose parameters from SMIRK [7], with lower values indicating better expression and pose simiarity.

B. Discussions, Limitations and Future work

In this section, we make comparisons of our method on ID modules with X-Portrait [11] and DiffPortrait3D [4]. The ID modules share the common objective of preserving the identity of the reference portrait during the generation of new animations or views. However, each method employs distinct strategies to achieve this goal, leading to differences in implementation, training, and application focus.

Our method utilizes a fine-grained appearance extractor coupled with an ID-aware multi-scale adapter (IMAdapter). This design enables detailed modeling of identity and background information from the reference image, ensuring high-fidelity identity preservation in the generated animations. The IMAdapter incorporates multi-scale convolutions and cross-attention mechanisms, which enhance the model's ability to capture intricate identity features and maintain their consistency across different frames and contexts. In contrast, X-Portrait's ID module extracts appearance and background features from a single reference image, which are then concatenated into the UNet's transformer blocks. This straightforward approach ensures consistent identity representation across generated frames but may not capture fine-grained details as effectively as our multi-scale architecture. X-Portrait focuses on expressive portrait animation, aiming to transfer facial expressions and head poses from driving videos to the reference portrait while maintaining identity similarity.

DiffPortrait3D adopts a different strategy by injecting appearance context from the reference image into the self-attention layers of a frozen UNet. This method effectively preserves identity across various rendering views but is primarily designed for 3D view synthesis rather than full animation. It leverages the generative power of pre-trained diffusion models to synthesize 3D-consistent novel views from as few as a single portrait.

For future work, we believe that exploring the synthesis of images from unknown perspectives to enhance identity retention capability is crucial. The current method has limitations in maintaining identity consistency from unknown perspectives after significant head rotation. The ID preservation design of DiffPortrait3D presents a potential improvement; however, it is still constrained by specific angles and cannot achieve the generation of unknown 360-degree views. We assert that enhancing the identity retention capability from unknown perspectives is a vital and feasible direction for improving the method proposed in this paper.

Besides, there are still several limitations with our method. Currently, our methodology is restricted to generating only the head and shoulder portions of portraits. We try to apply our method for generating full-body portraits that include hands; however, the results are unsatisfactory. The generated images of hands occasionally exhibit deformities and blurriness. This limitation stems from the inadequate representation of hand regions within the dataset, which hinders our ability to accurately render hand details. Future work can enhance our method by incorporating data that includes hand movements and improving the representation of hand features, thereby extending our approach to encompass full-body movements. Additionally, our approach is limited by the inherent constraints of the diffusion model. The significant computational costs impede the real-time applicability of our methods. Future work can accelerate the generation process through model distillation.

C. More Implementation Details

C.1. Appearance Extractor

For the input of the appearance extractor, we first resize the reference image to 256x256. We use the DiNOv2-Large with 4 register tokens as the appearance extractor, and fix the weights of its backbone during training. In the implementation of the IMAdapter, we first reduce the dimensionality of the features to 384 using linear layers. Subsequently, we employ convolutional kernels of varying scales (e.g., $1 \times 1, 3 \times 3, 5 \times 5$) for parallel processing and fuse these with ID features through a multi-head cross-attention mechanism. We set the number of heads for the multi-head cross-attention to 8.

C.2. Motion Extractor

The motion extractor comprises a total of six blocks within the network, utilizing consistent network configurations. The dimensionality of the latent features is set at 768, with the attention mechanism employing eight heads. The activation function implemented is the Sigmoid Linear Unit (SiLU) [3]. The motion encoder is borrowed from MegaPortrait [2]. We upgrade the architecture of the motion encoder from ResNet-18 to ResNet-50 to facilitate the pretraining of the motion encoder. After completing pretraining, the weights of this motion encoder is fixed and utilized for fine-tuning SVD.

D. More Visualizations

As shown in Figure 1, we present additional visualization results under the self-reenactment and cross-reenactment setting to better demonstrate the decoupling capability of our method in terms of appearance and motion. In order to demonstrate the robust generalization of our method, we selected a variety of images and videos from different style domains, including Civitai [1], Bilibili [2], and VFHQ [3], for demonstration purposes. To avoid copyright issues with driving videos, we first use a source image along with the original video to obtain the generated result. As illustrated in Figure 2, Figure 3 and Figure 4, we then utilize the generated results as the driving videos to animate other videos.

E. Ethics Consideration

E.1. User Study Details

In the user study, a total of 120 experienced participants are invited to take part. For each participant, we paid compensation that exceeded the local average hourly wage. We employ three metrics: Facial Movement, Video Quality, and Temporal Smoothness. For each metric, participants are presented with a video rated on a five-point scale. The grading options available to participants were as follows: Very Good (5), Good (4), Average (3), Poor (2), and Very Poor (1). As illustrated in Figure 5, the online evaluations are conducted using a well-structured website questionnaire. The questionnaire provides a comprehensive guideline, along with several example videos at the beginning. These example videos are not included in the rating but serve to illustrate the quality of video generation and ensure consistency in the rating criteria among participants.

E.2. Societal Impacts and Responsible AI

Our focus is on advancing the visual effects of virtual AI avatars to enhance their effectiveness for beneficial applications. It is essential to clarify that our research objectives are not intended to deceive or mislead. Like other content creation methods, our approach is not immune to potential misuse. We firmly oppose any misuse that could result in the creation of deceptive or harmful content through the impersonation of real individuals. Despite the risks of misuse, it's important to highlight the substantial positive outcomes of our technology. These include promoting educational fairness, aiding those with communication difficulties, and offering companionship or therapeutic aid. The significance of our research is underscored by its potential to assist those in need. We are committed to the ethical progression of AI, with the goal of fostering human welfare. The output videos generated by our method still retain identifiable traces of the actual individuals they are based on. To mitigate the potential for abuse, we are developing a neural network-powered tool designed to differentiate between genuine and synthetic videos, which includes our synthetic talking face videos in the training dataset. We will keep the community updated on any advancements in our models.

[1] https://civitai.com
[2] https://bilibili.com
[3] https://liangbinxie.github.io/projects/vfhq

References

[1] 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. 1

[2] Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2663–2671, 2022. 2

[3] Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107:3–11, 2018. 2

[4] Yuming Gu, Hongyi Xu, You Xie, Guoxian Song, Yichun Shi, Di Chang, Jing Yang, and Linjie Luo. Diffportrait3d: Controllable diffusion for zero-shot portrait view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer*

Figure 1. More visualizations of self-reenactment and cross-reenactment.

Figure 2. More visualizations of animated portraits.

Figure 3. More visualizations of animated portraits.

Figure 4. More visualizations of animated portraits.

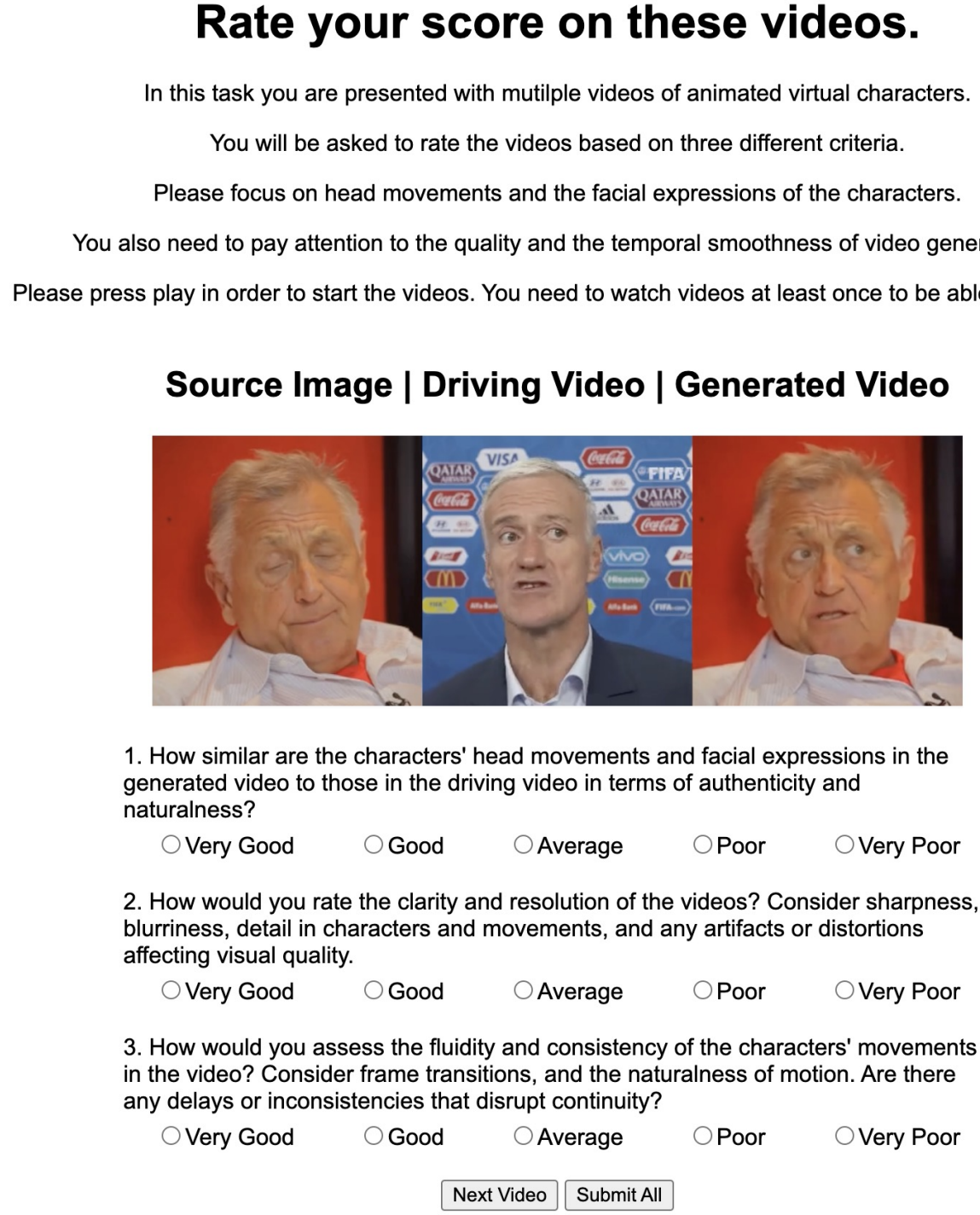

Figure 5. The screenshots of user study website for participants.

Vision and Pattern Recognition, pages 10456–10465, 2024. 1

[5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 1

[6] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Medi-

apipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 1

[7] George Retsinas, Panagiotis P Filntisis, Radek Danecek, Victoria F Abrevaya, Anastasios Roussos, Timo Bolkart, and Petros Maragos. 3d facial expressions through analysis-by-neural-synthesis. In *CVPR*, 2024. 1

[8] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Advances in Neural Information Processing Systems*, 2019. 1

[9] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019. 1

[10] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 1

[11] You Xie, Hongyi Xu, Guoxian Song, Chao Wang, Yichun Shi, and Linjie Luo. X-portrait: Expressive portrait animation with hierarchical motion attention. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 1

[12] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 1