

# – Supplementary Material –

## Free-viewpoint Human Animation with Pose-correlated Reference Selection

Fa-Ting Hong<sup>1,2</sup>, Zhan Xu<sup>2,✉</sup>, Haiyang Liu<sup>2</sup>, Qinjie Lin<sup>3</sup>, Luchuan Song<sup>2</sup>,  
Zhixin Shu<sup>2</sup>, Yang Zhou<sup>2</sup>, Duygu Ceylan<sup>2</sup>, Dan Xu<sup>1,✉</sup>

<sup>1</sup>HKUST <sup>2</sup>Adobe Research <sup>3</sup>Northwestern University, USA

fhongac@connect.ust.hk, {zhaxu, zshu, yazhou, ceylan}@adobe.com, haiyangliu1997@gmail.com  
qinjielin2018@u.northwestern.edu, lsong11@ur.rochester.edu, danxu@ust.hk

### 1. Limitation And Future Work

Our method relies on the quality and diversity of reference images, which may limit its performance in real-world applications. In future work, we plan to replace it with a more robust model, such as DiT, to improve results. Additionally, our MSTed dataset primarily consists of "talk" type data, and we aim to expand it by incorporating a variety of data sources for better generalization and robustness.

### 2. Implementation

The training process consists of two steps: 1) *Image training step*: In this step, each video frame is preprocessed (e.g., sampling, resizing, and center-cropping) to a uniform resolution of  $512 \times 512$  pixels. This step involves 50,000 training iterations with a batch size of 32. 2) *Temporal training step*: In this step, we train only the motion module while freezing the other parameters. This step is conducted over 30,000 iterations with sequences of 12 frames and a batch size of 8, focusing on motion learning.

In both steps, the learning rate is set to  $1 \times 10^{-5}$ . We use eight NVIDIA A100 GPUs for training. During inference, users can input more than one reference image. To ensure continuity over extended sequences, we employ a temporal aggregation method that integrates results from separate batches, enabling the generation of longer video outputs.

During the training process, we select varying numbers of reference images in each training step. In our MSTed dataset, the maximum number of reference images used during training is 5, while in the DyHumanDataset [4], it is 8. This approach allows our model to accept different numbers of reference images during the testing stage. Moreover, we can accept more than 10 reference images during testing.

The structure of pose encoder in pose correlation module is the same as pose guider. The size of generated correlation map is 32. In different layer of UNet, we resize the correlation map to fit the size of reference feature map

and perform Eq. 3.

### 3. Metrics

In this work, we evaluate our method and compare it with other methods on both pixel-level and feature-level. We adopt five popular metrics: Peak Signal-to-Noise Ratio (PSNR), Motion-based Video Integrity Evaluation (MOVIE) [2], Learned Perceptual Image Patch Similarity (LPIPS), Fréchet Video Distance (FVD), and  $\mathcal{L}_1$  distance.

**Peak Signal-to-Noise Ratio (PSNR).** PSNR is used to measure the similarity between the generated frames and ground truth frames at the pixel level. It is based on the mean squared error (MSE). It can be formulated as:

$$\text{PSNR} = 10 \cdot \log_{10} \left( \frac{\text{MAX}_I^2}{\text{MSE}} \right) \quad (1)$$

Here,  $\text{MAX}_I$  is the maximum possible pixel value, and MSE is the average squared difference between the original and generated frames.

**Motion-based Video Integrity Evaluation (MOVIE).** The MOVIE metric can be utilized to evaluate both spatial and temporal differences between frames. It quantifies how well frames are interpolated and how smooth the transitions are. It can be formulated as:

$$\begin{aligned} \text{MOVIE} = & \frac{1}{M} \sum [(I_{gt}^{t+1} - I_{gt}^t) - (I_{gen}^{t+1} - I_{gen}^t)]^2 \\ & + \frac{1}{M} \sum (I_{gt}^t - I_{gen}^t)^2 \end{aligned} \quad (2)$$

where  $I_{gt}$  and  $I_{gen}$  represent the frames of ground truth videos and the generated videos, respectively. We can easily find that lower MOVIE values indicate better video quality.

**Learned Perceptual Image Patch Similarity (LPIPS).** LPIPS is designed to measure the similarity of two images in feature space, which is learned in a neural network. Specifically, it focuses on perceptual quality rather than just

pixel accuracy. Its formula is:

$$\text{LPIPS}(I_{gt}^t, I_{gen}^t) = \sum_l w_l \|\mathbf{F}_l(I_{gt}^t) - \mathbf{F}_l(I_{gen}^t)\|_2 \quad (3)$$

where  $\mathbf{F}_l$  is the feature map from the  $l$ -th layer of the pre-trained network, and  $w_l$  is a predefined weight for the layer. We can observe that smaller LPIPS values indicate better perceptual similarity. In this work, we utilize a pre-trained VGG [3] network as the feature extractor.

**Fréchet Video Distance (FVD).** FVD is an important metric for video generation tasks. It measures the similarity between generated videos and ground truth videos in feature space, considering both the average features and their variability over time:

$$\text{FVD} = \|\mu_{gen} - \mu_{gt}\|^2 + \text{Tr} \left( \Sigma_{gen} + \Sigma_{gt} - 2(\Sigma_{gen}\Sigma_{gt})^{1/2} \right) \quad (4)$$

where  $\mu_{gen}$  and  $\mu_{gt}$  are the means of the feature maps for generated videos and ground truth videos, respectively. The  $\Sigma_{gen}$  and  $\Sigma_{gt}$  are the corresponding covariances. From the above equation, lower FVD values indicate better realism and smoother motion in the generated videos. Here, we adopt the pre-trained I3D [1] network as the feature extractor.

## 4. Visualization

We provide video demos in the supplementary materials. In these video demos, we compare our method with other methods, and our method obviously achieves better results. Additionally, we found that when some of the reference images provide more details, the results can be even more realistic.

## References

- [1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. [2](#)
- [2] Kalpana Seshadrinathan and Alan Conrad Bovik. Motion tuned spatio-temporal quality assessment of natural videos. *IEEE TIP*, 19(2):335–350, 2009. [1](#)
- [3] Karen Simonyan. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. [2](#)
- [4] Xiaoyun Zheng, Liwei Liao, Xufeng Li, Jianbo Jiao, Rongjie Wang, Feng Gao, Shiqi Wang, and Ronggang Wang. Pkudymvhumans: A multi-view video benchmark for high-fidelity dynamic human modeling. *CVPR*, 2024. [1](#)