Consistent and Controllable Image Animation with Motion Diffusion Models

Supplementary Material

1. Four methods for dynamics degree control

We delve into four methods for evaluating motion intensity in videos. Our study begins by randomly selecting 1,000 videos from the training dataset. For each video, we extract a 16-frame segment starting from the first frame, using different frame intervals. The frame intervals used are 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, and 25. Generally, videos with a frame interval of 3 tend to exhibit lower overall motion intensity compared to those with a frame interval of 7, and so on.

Subplots (a), (c), and (d) in Fig. 1 illustrate the overall motion intensity of the video groups estimated using the mean absolute difference between frames, the mean structural similarity index (SSIM), and the mean multi-scale structural similarity index (MS-SSIM), respectively. The vertical axis in these subplots represents the similarity of the videos, and theoretically, the similarity should gradually decrease as the frame interval increases. Subplot (b) shows the dynamic degree estimated using the RAFT optical flow estimator [5], with the vertical axis representing the dynamic degree. Theoretically, this value should increase as the frame interval grows.

As shown in Fig. 1, the mean absolute difference between frames, the mean MS-SSIM, and the dynamic degree estimated through optical flow can all accurately reflect the motion intensity of the videos. However, in terms of computational efficiency, the average time required to estimate the motion intensity of a video using these three methods is 0.03 seconds, 0.07 seconds, and 1.17 seconds, respectively. Given this, we exclude the optical flow-based method. Additionally, considering that the MS-SSIM-based method more accurately reflects changes in motion as perceived by human vision, while the mean absolute difference method is overly sensitive to noise or changes in non-essential details (for example, for videos with high color contrast but low actual motion intensity, the mean absolute difference might be exaggerated), we ultimately choose the MS-SSIM-based method to estimate the motion intensity of the videos.

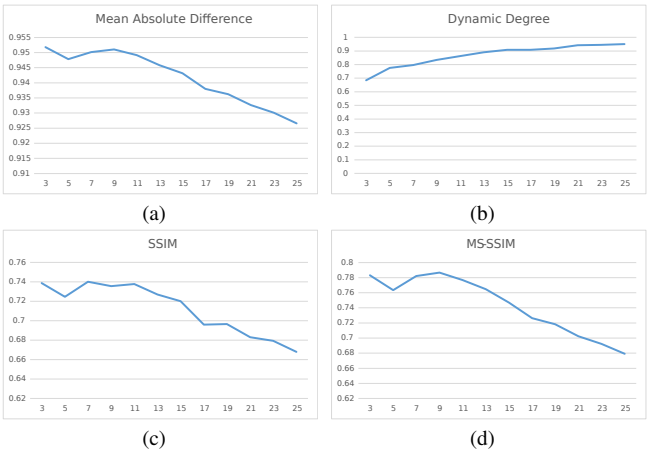

Figure 1. The performance differences between four methods.

2. Intuition behind the edge effect of FFT

FFT assumes that the input signal is periodic. However, due to the finite length of real-world signals or discontinuities at their boundaries, "spectral leakage" occurs. This phenomenon refers to the spread of energy, originally concentrated at specific frequencies, to other frequencies in the spectrum. When the signal is truncated or windowed under non-ideal conditions, the energy distribution manifests as sidelobes across the frequency domain [1, 2, 4].

During noise refinement, we extract low-frequency components from the input static image. These low-frequency components typically capture macroscopic information, such as overall brightness variations and color distribution. However, when FFT is used for this extraction, spectral leakage may cause the low-frequency energy to become dispersed or weakened. This can negatively impact the color tonal consistency of the generated video.

3. Limitations and discussions

Our model is based on the pre-trained LaVie [6] model and is further trained on similar datasets. This means that the performance of our model is, to some extent, limited by the inherent characteristics of LaVie. For example, the resolution of the current video generation is constrained by LaVie, fixed at 320 x 512. Recently, the technological development trend in the field of video generation has clearly shifted towards Transformer-based architectures, gradually replacing the traditional UNet architecture. This shift is mainly due to the more effective scalability of the model parameters of Transformers. In light of this, our future plans include adopting Transformer-based architectures, such as Latte [3], to further validate and optimize our model.

References

[1] Fredric J Harris. On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, 66(1):51–83, 1978. 1

[2] Gerhard Heinzel, Albrecht Rüdiger, and Roland Schilling. Spectrum and spectral density estimation by the discrete fourier transform (dft), including a comprehensive list of window functions and some new at-top windows. 2002. 1

[3] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024. 1

[4] John G Proakis. *Digital signal processing: principles algorithms and applications*. Pearson Education India, 2001. 1

[5] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, pages 402–419. Springer, 2020. 1

[6] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, pages 1–20, 2024. 1