Make Pixels Dance: High-Dynamic Video Generation (Supplementary Material)

Yan Zeng* Guoqiang Wei* Jiani Zheng
Jiaxin Zou Yang Wei Yuchen Zhang Hang Li

ByteDance Research

* Equal Contribution

{zengyan.yanne, weiguoqiang.9, lihang.lh}@bytedance.com

https://makepixelsdance.github.io

1. More Experimental Details

1.1. Implementation Details

We enhance the quality of the WebVid-10M dataset. Our process involves filtering out videos that are almost static by assessing their optical flow values. For this, we employ VideoFlow [3], a tool capable of estimating bi-directional optical flows across multiple frames. This approach differs from traditional methods that typically estimate optical flow between just two frames. Additionally, we exclude videos with an aesthetic score below 3.95, as determined by the Improved Aesthetic Predictor available on GitHub[1]. Following these filtration steps, our final dataset comprises approximately 9 million video clips from the initial 10 million samples in Webvid. The self-collected videos are also filtered by the above process.

We first train the model at resolution of 256×256, with batch size of 192 on 32 A100 GPUs for 200K iterations, which is utilized for quantitative evaluations. This model is then finetuned to two additional model variants at higher resolutions, 320×320 and 448×256 pixels, based on this pre-trained model. We fine-tune the models for 50k steps. Higher resolutions help to preserve visual details given the first frame instructions. All quantitative results are evaluated with the 256×256 model. In inference, generating a 16-frame video clip with 256×256 resolution takes about 7.5 seconds on A100 GPU through DDIM sampler [4] with 50 denoising steps.

1.2. Human Evaluation

To make a fire human evaluation comparison with Gen2 and PiKa, three generated videos per prompt are displayed with a random order. Users are asked to sort three videos in terms of two orthogonal aspects. A fixed mask is applied over all videos to cover the watermark information. The averaged ratings across all prompts are taken as the final human evaluation result per method.

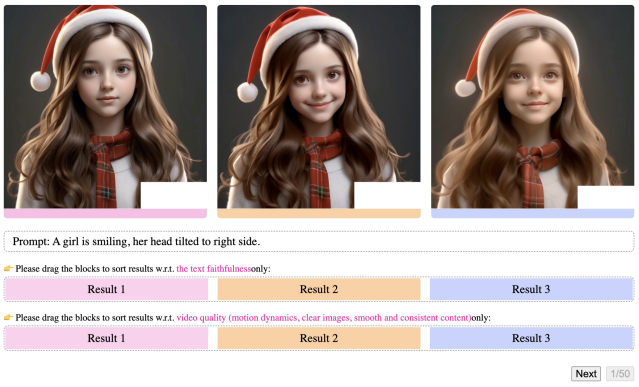

Figure 1. Human evaluation interface.

Figure 2. Qualitative results of long video generation on UCF-101. Each frame of the first three rows is selected with a frame interval of 16. The first two rows are copied from the LVDM paper [2].

[1] https://github.com/christophschuhmann/improved-aesthetic-predictor

2. More Experimental Results

Visualization of Long Video Generation In Figure 2, we compare the long video generation on UCF-101 with TATS-AR [1] and LVDM-AR [2]. PixelDance demonstrates superior generation performance with minor quality degradation problem. Additionally, we provide a two-minute video with 448×256 resolution generated by PixelDance in the attachment, where the major character holds consistent cross diverse scenes, including the real-world and science fiction scenarios.

More Visualization Results Video generation examples conditioned on text, first frame instructions are illustrated in Figure 3. Video generation examples conditioned on text, first frame and last frame instructions are illustrated in Figure 4.

References

[1] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *European Conference on Computer Vision*, pages 102–118. Springer, 2022.

[2] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022.

[3] Xiaoyu Shi, Zhaoyang Huang, Weikang Bian, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Videoflow: Exploiting temporal cues for multi-frame optical flow estimation. *arXiv preprint arXiv:2303.08340*, 2023.

[4] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Figure 3. Video generation conditioned on text and first frame instructions.

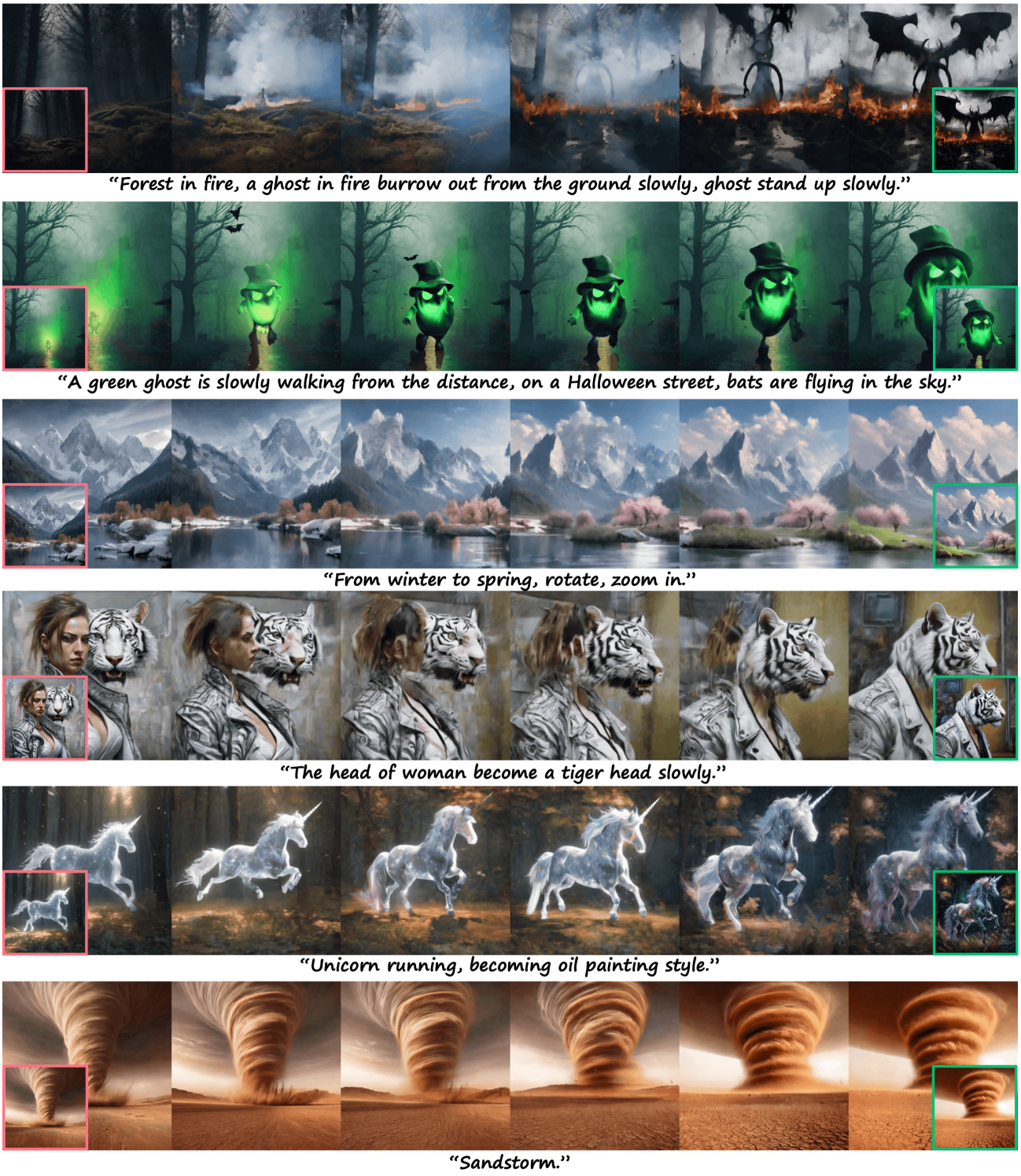

Figure 4. Video generation conditioned on text, first frame and last frame instructions.