Supplementary Material for 'DragAnything'

Weijia Wu[1,2,3], Zhuang Li[1], Yuchao Gu[3], Rui Zhao[3], Yefei He[2], David Junhao Zhang[3], Mike Zheng Shou[()3], Yan Li[1], Tingting Gao[1], and Di Zhang[1]

[1] Kuaishou Technology, China
[2] Zhejiang University, China
[3] Show Lab, National University of Singapore, Singapore

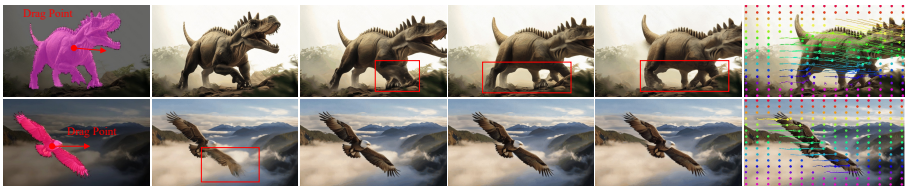

Fig. 1: Bad Case for DragAnything. DragAnything still has some bad cases, especially when controlling larger motions.

1 Supplementary Material

1.1 Discussion of Potential Negative Impact.

One potential negative impact is the possibility of reinforcing biases present in the training data, as the model learns from existing datasets that may contain societal biases. Additionally, there is a risk of the generated content being misused, leading to the creation of misleading or inappropriate visual materials. Furthermore, privacy concerns may arise, especially when generating videos that involve individuals without their explicit consent. As with any other video generation technology, there is a need for vigilance and responsible implementation to mitigate these potential negative impacts and ensure ethical use.

1.2 Limitation and Bad Case Analysis

Although our DragAnything has demonstrated promising performance, there are still some aspects that could be improved, which are common to current other trajectory-based video generation models: 1) Current trajectory-based motion control is limited to the 2D dimension and cannot handle motion in 3D scenes, such as controlling someone turning around or more precise body rotations. 2) Current models are constrained by the performance of the foundation model, Stable Video Diffusion [1], and cannot generate scenes with very large motions, as shown in Figure 1. It is obvious that in the first column of video frames, the legs of dinosaur don't adhere to real-world constraints. There are a few frames where there are five legs and some strange motions. A similar situation occurs

Fig. 2: More Visualization for DragAnything.

with the blurring of the wings of eagle in the second row. This could be due to excessive motion, exceeding the generation capabilities of the foundation model, resulting in a collapse in video quality. There are some potential solutions to address these two challenges. For the first challenge, a feasible approach is to incorporate depth information into the 2D trajectory, expanding it into 3D trajectory information, thereby enabling control of object motion in 3D space. As for the second challenge, it requires the development of a stronger foundation model to support larger and more robust motion generation capabilities. For example, leveraging the latest text-to-video foundation from OpenAI, SORA, undoubtedly has the potential to significantly enhance the quality of generated videos. In addition, we have provided more exquisite video cases in the supplementary materials for reference, as shown in Figure 2. For more visualizations in GIF format, please refer to `DragAnything.html` in the same directory. Simply click to open.

References

1. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)