LeviTor: 3D Trajectory Oriented Image-to-Video Synthesis

Supplementary Material

Appendix

A. Comparison with more methods

This section compares our LeviTor with more recent methods SG-I2V [3] and MOFA-Video [4]. The qualitative comparison in Fig. S1 shows that these methods fail to follow complex trajectories or produce proper depth variation.

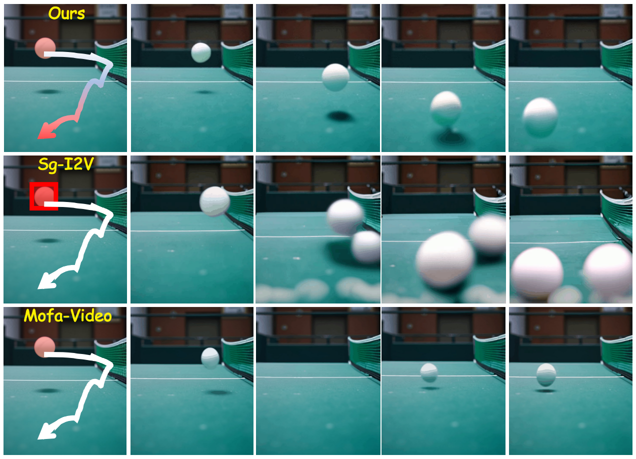

Figure S1. Qualitative comparison with SG-I2V and MOFA-Video.

B. More Ablations on the Number of Control Points for Inference

In this section, we show more examples of choosing different numbers of control points to generate videos with LeviTor. We conduct inference with our default number of control points and with more densely packed points, respectively. The results are shown in Fig. S2. It can be seen that with the default number of control points, our LeviTor can reasonably represent the state of fluid movement and human running. However, since the generation strictly follows the control points, the more control points used, the less space is left for our model to produce some non-rigid movements, resulting in the unreasonable results of waves floating in the air and people gliding on the road. This demonstrates that overly dense control points cannot generate non-rigid motion well. Thus, we implement LeviTor with multiple clustered points control rather than directly using object masks as the condition. In this way, users can flexibly adjust the number of control points as needed to generate both rigid and non-rigid motions.

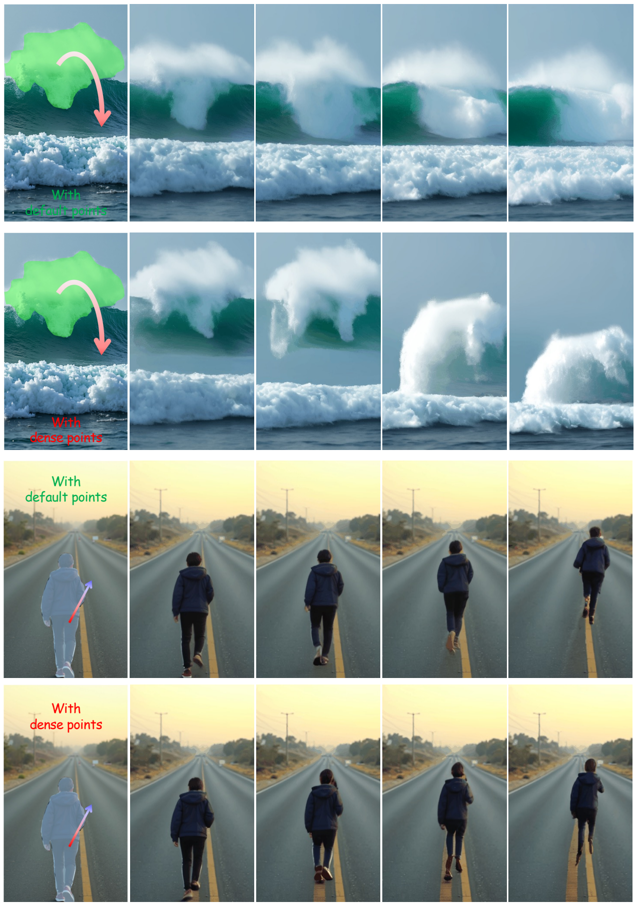

Figure S2. Ablation results on the Number of Control Points for Inference. We highly recommend viewing the visualization results for detailed video demonstrations.

Table S1. Quantitative comparison with Single-point Control on DAVIS [2].

Methods	FID ↓	FVD ↓	ObjMC ↓
Single-Point Control	30.91	253.73	38.21
Ours	**25.41**	**190.44**	**25.97**

C. Comparison with Single-point Control

One of our key motivations is to represent 3D motions by utilizing the clustering and dispersion of multiple points within object masks. Another more intuitive idea is whether we can represent 3D motion using 2D trajectories combined with depth information. That is, representing a 3D trajectory through a single 2D trajectory along with changes of depth values input by users. To validate this

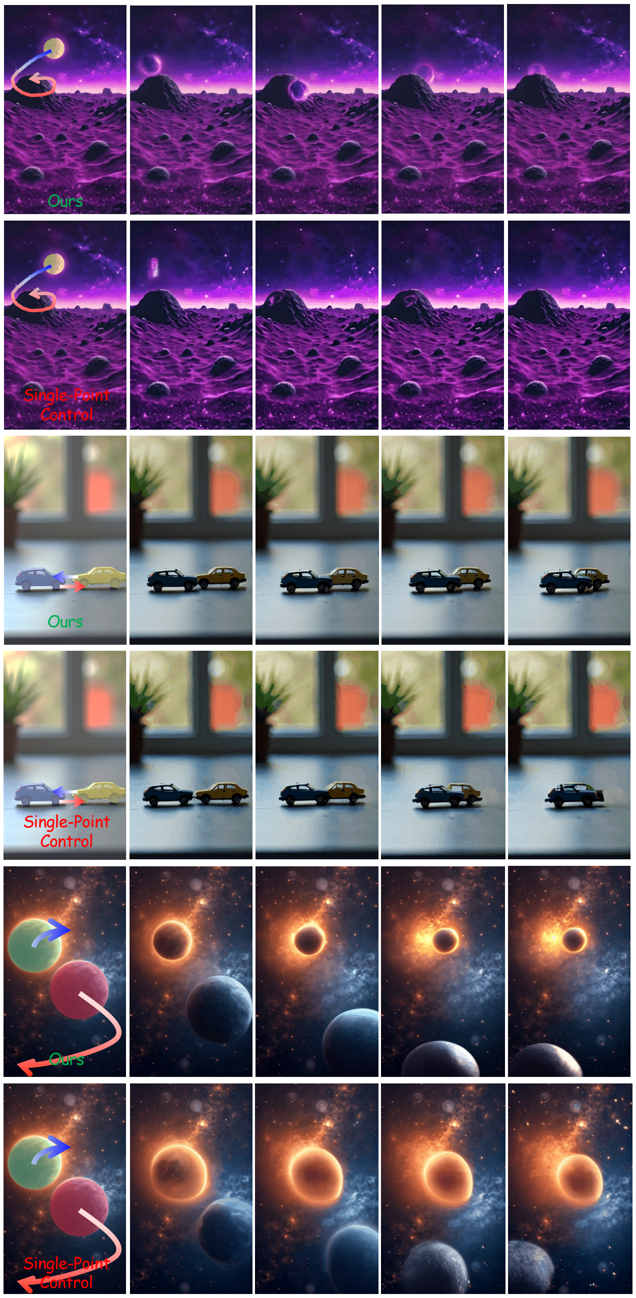

Figure S3. Comparison with Single-point Control model. We highly recommend viewing the visualization results for detailed video demonstrations.

idea, we use the center point of each object's mask as a control point and train the model with the value change of that point as the generation condition. We conduct both qualitative and quantitative analysis. Qualitative results in Fig. S3 show that such single-point control can not represent 3D motions well. The first two examples test the representation of occlusion. It can be observed that a single point with depth changes controlling struggles to accurately express occlusion, resulting in the disappearance of the purple light cluster and the deformation and merging of the cars. The third example tests the control of forward and backward movements. Compared to our `LeviTor`, single-point control is not very sensitive to size changes caused by forward and backward movement. Quantitative results in Tab. S1 also show the advantage of 3D motion representation with clustering and dispersion of multiple points. The ablation study in Tab. 2 of the main text indicates that the value of depth does not significantly affect the quality of the generated results. And results in this section show that 2D trajectories with depth value changes can not represent 3D motions. These conclusions both suggest that in our method, the clustering and dispersion of multiple control points are the key aspects of 3D motion representation, while depth information is generally used for moving objects in 3D space to obtain rendered object masks.

D. Bad Case Analysis

We, in this section, list some bad generation cases for analysis. Results shown in Fig. S4 indicate that our `LeviTor` has difficulties in reconstructing small faces and generating scenes with large motions. It may also confuse similar parts of objects. For example, in the first row of Fig. S4, the horse faces become blurry while walking, and the movement of their legs is also quite unnatural. Similarly, in Fig. S2, the movement of the person's feet while running also appears unnatural. In the second row, the elephant's front leg suddenly turns into a back one, and then a regenerated front leg appears. We attribute this phenomenon to the fact that the underlying video base model Stable Video Diffusion (SVD) [1] we apply is unable to reconstruct small faces and tends to produce artifacts when generating large-scale movements. We are going to enhance our model by integrating more advanced video-based models in the future, hoping to better capture deformable objects and complex dynamics to handle large-scale and non-rigid motions.

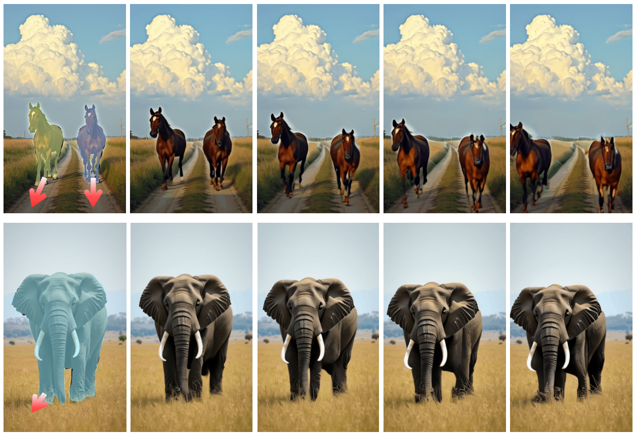

Figure S4. Bad Cases of `LeviTor`.

References

[1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *CoRR*, abs/2311.15127, 2023. 2

[2] Sergi Caelles, Jordi Pont-Tuset, Federico Perazzi, Alberto Montes, Kevis-Kokitsi Maninis, and Luc Van Gool. The 2019 DAVIS challenge on VOS: unsupervised multi-object segmentation. *CoRR*, abs/1905.00737, 2019. 1

[3] Koichi Namekata, Sherwin Bahmani, Ziyi Wu, Yash Kant, Igor Gilitschenski, and David B. Lindell. Sg-i2v: Self-guided trajectory control in image-to-video generation. *arXiv preprint arXiv:2411.04989*, 2024. 1

[4] Muyao Niu, Xiaodong Cun, Xintao Wang, Yong Zhang, Ying Shan, and Yinqiang Zheng. Mofa-video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model. *arXiv preprint arXiv:2405.20222*, 2024. 1