MagicMotion: Controllable Video Generation with Dense-to-Sparse Trajectory Guidance (— Supplementary Materials —)

Quanhao Li[1*] Zhen Xing[1*] Rui Wang[1] Hui Zhang[1] Qi Dai[2] Zuxuan Wu[1†]
[1] Fudan University [2] Microsoft Research Asia
https://quanhaol.github.io/magicmotion-site/

A. Ablations on Latent Segment Loss

We present qualitative results from the ablation study on Latent Segment Loss. As shown in Fig. 1, removing the Latent Segment Loss significantly reduces the model's ability to capture object shapes, resulting in less accurate trajectory control. For instance, without this loss, the woman's arm in the generated video appears incomplete.

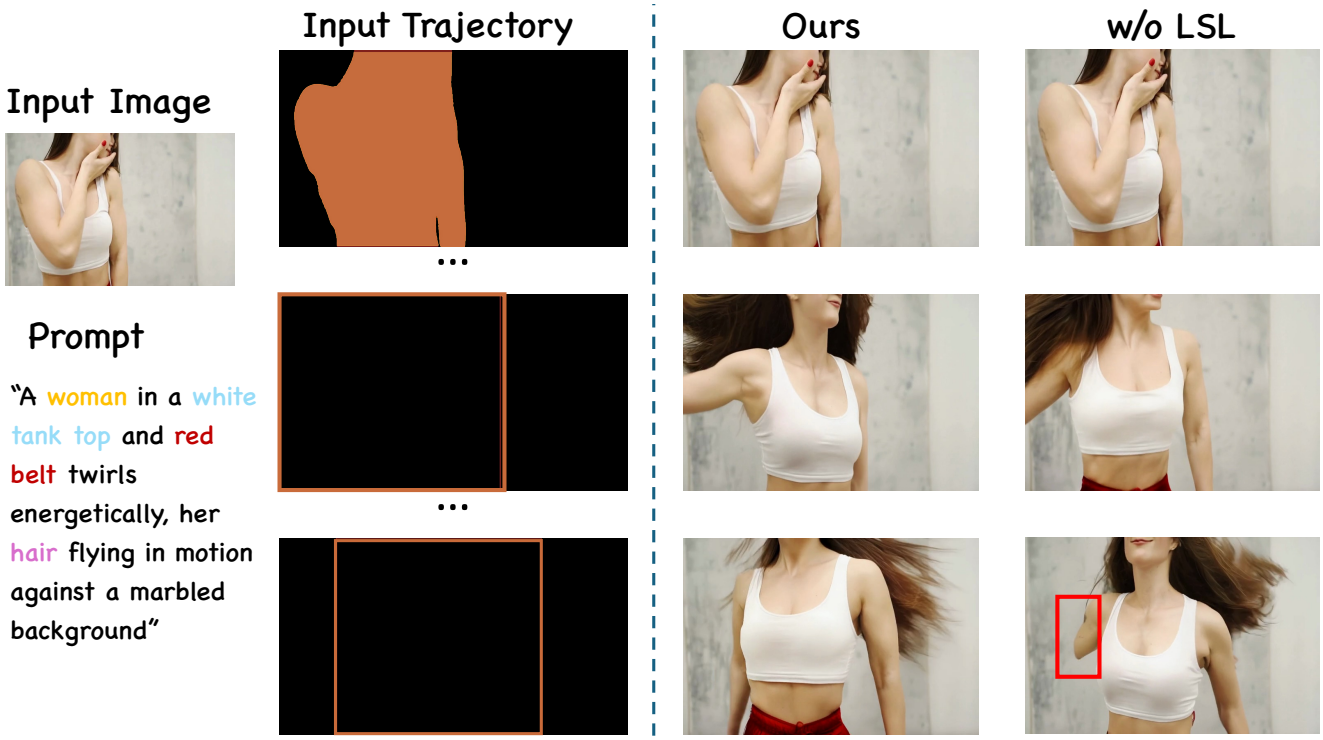

Figure 1. Ablation Study on latent segment loss. Without it, the generated arms appear partially missing.

B. Additional Experiments

We conducted additional experiments using MagicMotion under various task settings, including camera motion control and video editing. We also generate videos by applying different motion trajectories to a single input image.

Camera Motion Control. As shown in Fig. 2, MagicMotion enables precise control over camera motion, allowing for operations such as rotation, zoom, and pan. In the first row of Fig. 2, we enclose oranges within bounding boxes and apply

Figure 2. Camera motion controlled results. By setting specific trajectory conditions, MagicMotion can control camera movements.

rotation to the boxes. This results in a video with a simulated camera rotation effect. In the second and third rows, we adjust the size of the foreground object's bounding box to control its perceived distance from the camera, effectively achieving zoom-in and zoom-out effects. In the last two rows, we shift the bounding box to the left and downward, creating the effect of the camera moving in the opposite direction.

Video Editing. As shown in Fig. 3, MagicMotion can be applied to video editing to generate high-quality videos. Specifically, we first use FLUX [3] to edit the first frame of the original video, which serves as the input for MagicMotion. Then, we extract the segment mask of the original video and use it as trajectory guidance for the MagicMotion Stage1. Using this approach, we transform a black swan into a diamond swan, make the camel walk in a majestic palace, and turn a hiking backpacker into an astronaut.

Same input image with different trajectories Extensive experiments have demonstrated that MagicMotion enables objects to move along specified trajectories, generating high-quality videos. To further showcase the capabilities of MagicMotion, we use stage2 of MagicMotion to animate objects from the same input image along different motion trajectories. As shown in Fig. 4, MagicMotion successfully animates two bears, two fish, and the moon, each following their designated paths.

C. Latent Segment Masks

In this section, we provide a more detailed demonstration of Latent Segmentation Masks. Specifically, we use MagicMotion Stage 3 to predict the latent segmentation masks for each frame based on sparse bounding box conditions. As shown in Fig. 5, MagicMotion accurately predicts the Latent Segmentation Masks throughout dynamic scenes, such as a man gradually standing up to face a robot and a boy's head slowly sinking into the water. This holds true for frames where only the bounding box trajectory is available and even for frames where no trajectory information is provided at all.

D. Additional Comparison results

Baseline Comparisons. As shown in Table 1, we provide a comparison of the backbones used by each method, along with the supported video generation length and resolution.

Figure 3. Video Editing Results. We use FLUX [3] to edit the first-frame image and MagicMotion Stage1 to move the foreground objects following the trajectory of the origin video.

Quantitative Comparisons on different object number Due to space constraints, we only included radar charts in the main text to compare the performance of different methods in controlling varying numbers of objects on MagicBench. Here, we provide the specific quantitative results. As shown in Table 2, Table 3, and Table 4, MagicMotion consistently outperforms other methods across all metrics by a significant margin, especially when the number of moving objects is large. This demonstrates that other methods exhibit poorer performance when controlling a larger number of objects.

More qualitative comparison results. In this section, we provide additional qualitative comparison results with previous works. As shown in Fig.6, Fig.7, Fig.8, Fig.9, and Fig. 10, MagicMotion accurately controls object trajectories and generates high-quality videos, while other methods exhibit significant defects. For fully rendered videos, we refer the reader to

Figure 4. MagicMotion can generate videos using the same input image and different trajectories (marked by red arrows).

	Resolution	Length	Base Model
Motion-I2V [6]	320*512	16	AnimateDiff [2]
ImageConductor [4]	256*384	16	AnimateDiff [2]
DragAnything [9]	320*576	25	SVD [1]
LeViTor [8]	288*512	16	SVD [1]
DragNUWA [11]	320*576	14	SVD [1]
SG-I2V [5]	576*1024	14	SVD [1]
Tora [12]	480*720	49	CogVideoX [10]
MagicMotion-CogVideoX	480*720	49	CogVideoX [10]
MagicMotion-Wan1.3B	480*832	81	Wan [7]

Table 1. Comparisons on each method's backbone.

"`Supplementary video.mp4`" in supplementary material.

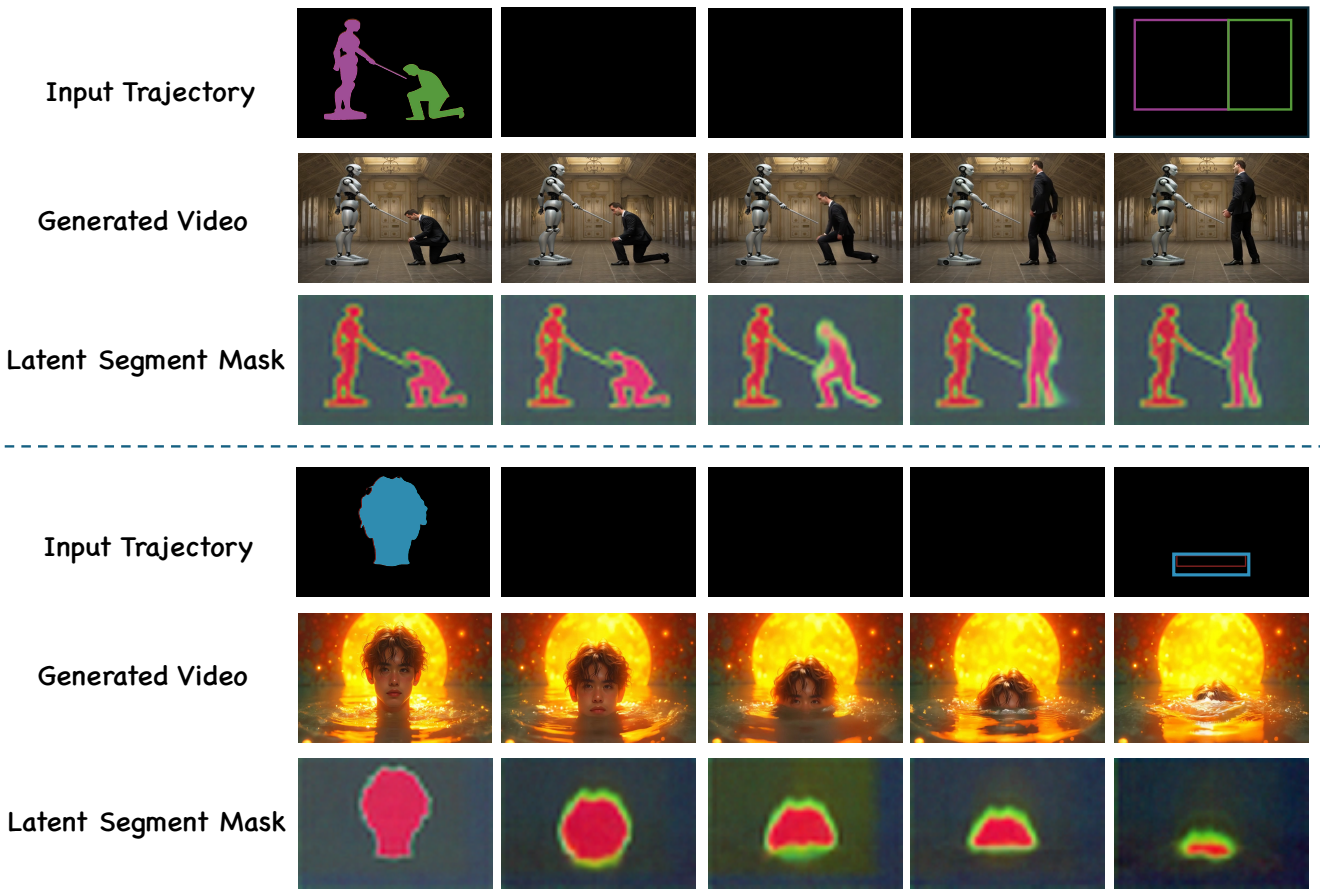

Figure 5. Latent Segment Masks visualization. MagicMotion can predict out latent segment masks of each frame even when only provided with sparse bounding boxes guidance.

Method	MagicBench (Object Number = 1)				MagicBench (Object Number = 2)			
	FID(↓)	FVD(↓)	Mask_IoU(↑)	Box_IoU(↑)	FID(↓)	FVD(↓)	Mask_IoU(↑)	Box_IoU(↑)
Motion-I2V [6]	78.3245	660.8655	0.6057	0.7142	90.4269	867.6057	0.5078	0.5938
ImageConductor [4]	106.1569	674.4987	0.5706	0.6974	109.2959	879.8890	0.4786	0.5579
DragAnything [9]	72.2494	884.6453	0.6706	0.8088	85.6511	940.4941	0.6148	0.7232
LeViTor [8]	80.9714	492.4725	0.5536	0.7057	94.0106	542.8533	0.4352	0.5364
DragNUWA [11]	76.3963	610.0368	0.6699	0.7769	81.2921	624.3985	0.6033	0.6809
SG-I2V [5]	60.8617	547.8107	0.7144	0.8668	71.8699	619.2791	0.6315	0.7378
Tora [12]	60.2737	805.0145	0.6468	0.7776	73.6499	795.8535	0.5509	0.6584
Ours (Stage1-Wan1.3B)	48.4211	461.1011	0.8518	0.9186	49.3071	438.2138	0.8143	0.8161
Ours (Stage2-Wan1.3B)	51.2964	555.1390	0.6693	0.8163	55.5581	569.0549	0.6318	0.7454
Ours (Stage1-CogVideoX)	**42.3523**	**473.2179**	**0.9359**	**0.9607**	**48.6525**	**428.3430**	**0.9080**	**0.9097**
Ours (Stage2-CogVideoX)	46.1080	564.1036	0.7363	0.9017	52.7296	550.5857	0.6931	0.8256

Table 2. Quantitative Comparison results on MagicBench with moving objects number equals to 1 / 2.

E. Additional Ablation Results.

Here, we provide additional qualitative comparison results from the ablation study. As shown in Fig. 11, not using MagicData for training results in the generation of a woman with an extra hand. Not using the Progressive Training Procedure results in significant defects, such as a dancing woman showing severe issues when turning, with a second face appearing where her hair should be. Additionally, without the Latent Segment Loss, the woman's lipstick is distorted into a rectangular shape.

Method	MagicBench (Object Number = 3)				MagicBench (Object Number = 4)			
	FID(↓)	FVD(↓)	Mask_IoU(↑)	Box_IoU(↑)	FID(↓)	FVD(↓)	Mask_IoU(↑)	Box_IoU(↑)
Motion-I2V [6]	89.3195	842.6530	0.5366	0.6076	83.2024	744.5470	0.6018	0.6484
ImageConductor [4]	114.7662	927.3884	0.5169	0.5578	110.8037	832.4498	0.5679	0.5417
DragAnything [9]	85.1437	925.2795	0.6332	0.6625	71.8865	901.7427	0.6946	0.7148
LeViTor [8]	93.3111	607.7522	0.3809	0.4671	90.4561	688.8164	0.3555	0.4044
DragNUWA [11]	79.3941	642.4184	0.6420	0.7012	71.0587	512.5130	0.7085	0.7250
SG-I2V [5]	70.0600	520.2733	0.6531	0.7068	56.9318	460.1303	0.7145	0.7423
Tora [12]	66.6571	742.4080	0.5926	0.6356	64.0669	779.0798	0.6226	0.6312
Ours (Stage1-Wan1.3B)	51.7821	458.4122	0.8236	0.8122	47.8972	463.4390	0.8598	0.8288
Ours (Stage2-Wan1.3B)	56.9454	568.4777	0.6632	0.7089	50.0364	548.9445	0.7112	0.7252
Ours (Stage1)	**42.9636**	**421.0036**	**0.9107**	**0.8797**	**37.6524**	**396.4754**	**0.9231**	**0.8896**
Ours (Stage2)	45.4721	440.2373	0.7562	0.8097	37.4172	442.0640	0.7998	0.8253

Table 3. Quantitative Comparison results on MagicBench with moving objects number equals to 3 / 4.

Method	MagicBench (Object Number = 5)				MagicBench (Object Number >5)			
	FID(↓)	FVD(↓)	Mask_IoU(↑)	Box_IoU(↑)	FID(↓)	FVD(↓)	Mask_IoU(↑)	Box_IoU(↑)
Motion-I2V [6]	84.2231	582.0029	0.6295	0.6267	79.9077	923.2557	0.4899	0.4511
ImageConductor [4]	117.8054	857.3489	0.5180	0.4737	106.3168	963.6862	0.4536	0.3442
DragAnything [9]	79.7015	710.5812	0.7050	0.7011	87.1509	719.2442	0.6534	0.6045
LeViTor [8]	90.8332	578.5567	0.3913	0.4281	92.5265	763.3157	0.2812	0.2768
DragNUWA [11]	77.2094	435.9205	0.7253	0.6988	78.5956	549.7680	0.6638	0.5709
SG-I2V [5]	65.2560	453.1147	0.7431	0.7367	95.7569	596.4075	0.6616	0.6211
Tora [12]	67.3827	709.1618	0.6228	0.6111	77.2571	907.9254	0.4976	0.4866
Ours (Stage1-Wan1.3B)	40.8559	384.3586	0.8645	0.8112	39.6606	**395.2139**	0.8136	0.6726
Ours (Stage2-Wan1.3B)	46.7040	491.7079	0.7461	0.7210	42.0005	454.2931	0.6941	0.5861
Ours (Stage1)	**39.4636**	374.6467	**0.9155**	**0.8600**	39.2044	449.3122	**0.9012**	**0.7653**
Ours (Stage2)	40.0656	**350.5010**	0.8106	0.8123	**37.1917**	396.0661	0.8004	0.7124

Table 4. Quantitative Comparison results on MagicBench with moving objects number equals to 5 / above 5.

F. More Details on MagicData

Here, we provide some detailed statistical information about MagicData. On average, each video in MagicData contains 346 frames, with a typical height of 999 pixels and a width of 1503 pixels. For a more comprehensive understanding of the distribution and variability across the dataset, please refer to Fig. 13, which visualizes the detailed distribution of video frame counts, heights, and widths. During training, these videos are resized to 48 frames and converted to a 720p resolution.

G. More Details on MagicBench

For evaluation purposes, all videos in MagicBench are sampled to 49 frames and resized to a resolution of 720p. MagicBench is categorized into 6 classes based on the number of annotated foreground objects. Below, we provide one video example for each category, offering a more intuitive understanding of MagicBench.

Figure 6. Qualitative Comparisons Results. MagicMotion successfully control the cat jumping over the bowl, while all other methods exhibit significant defects.

Figure 7. Qualitative Comparisons Results. MagicMotion successfully control the witch flying over the input trajectory, while all other methods exhibit significant defects.

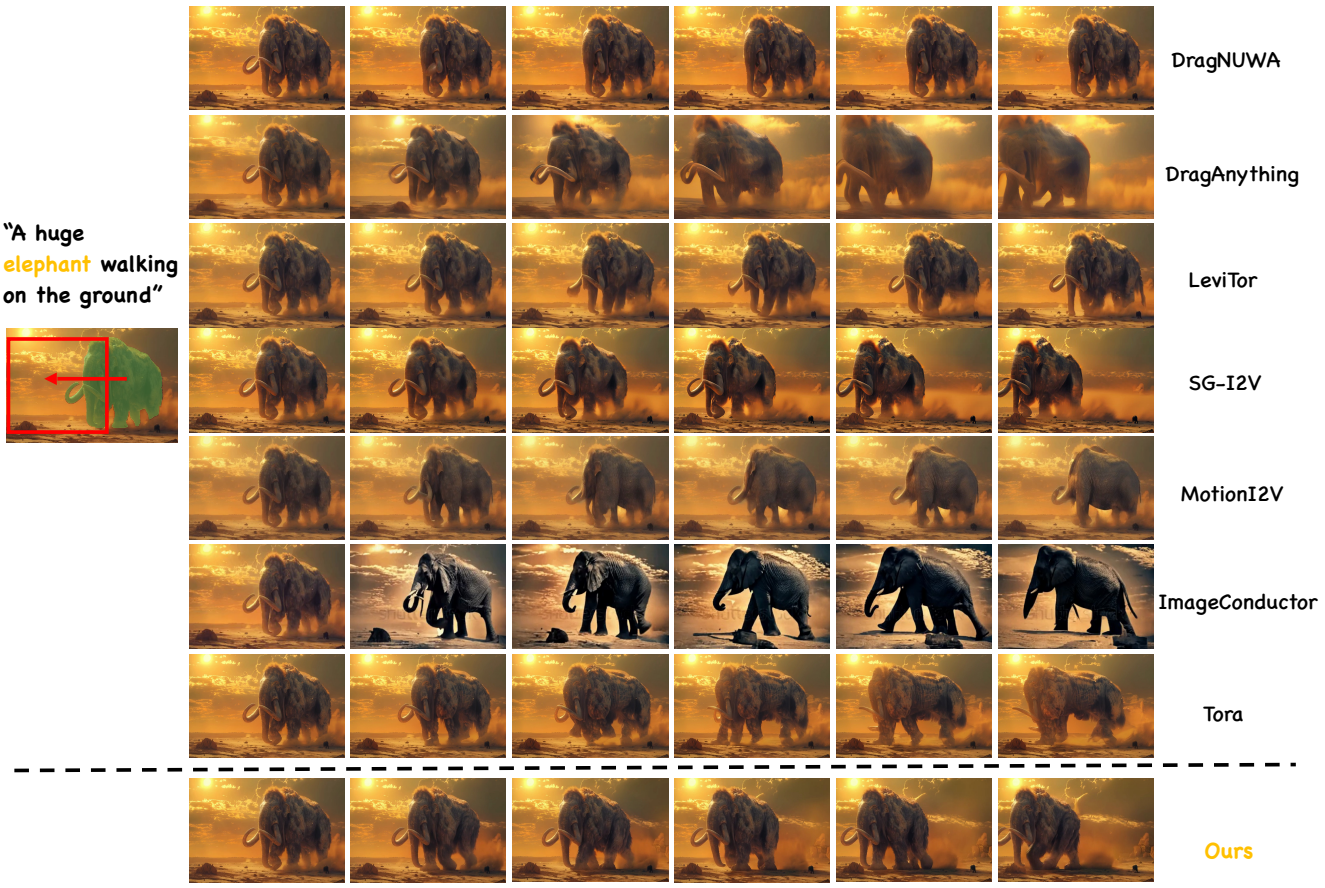

Figure 8. Qualitative Comparisons Results. MagicMotion successfully control the elephant walking along the input trajectory, while all other methods exhibit significant defects.

Figure 9. Qualitative Comparisons Results. MagicMotion successfully control the robot moving along the input trajectory, while all other methods exhibit significant defects.

Figure 10. Qualitative Comparisons Results. MagicMotion successfully control the tiger's head moving along the input trajectory, while all other methods exhibit significant defects.

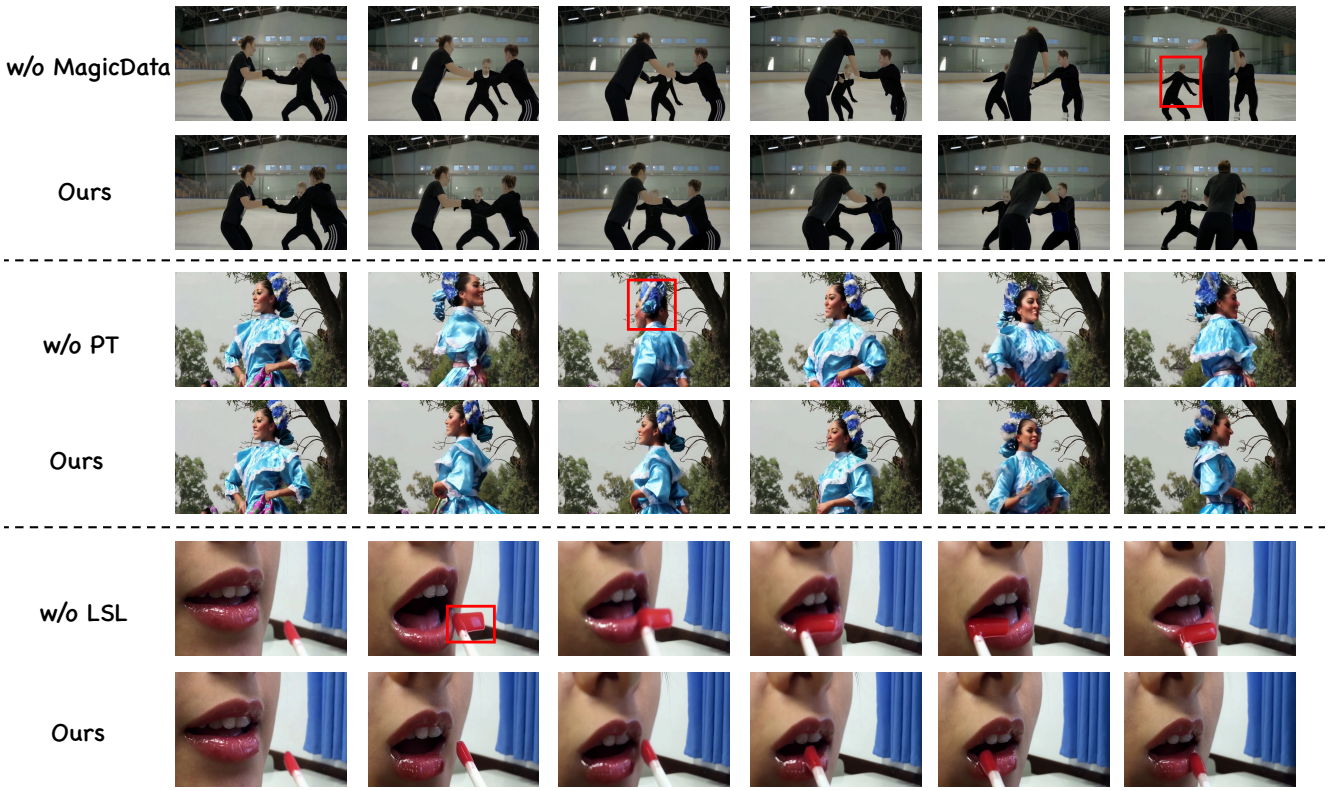

Figure 11. Additional Ablation results.

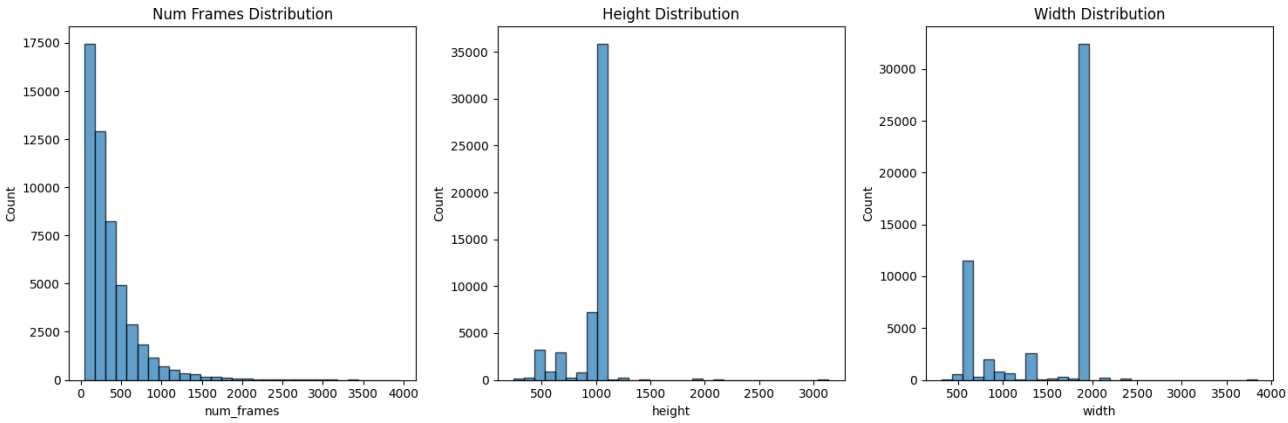

Figure 12. Detail information on MagicData.

Figure 13. MagicBench visualization. We provide one video as a visual example for each object number category..

References

[1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 4

[2] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *International Conference on Learning Representations*, 2024. 4

[3] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 2, 3

[4] Yaowei Li, Xintao Wang, Zhaoyang Zhang, Zhouxia Wang, Ziyang Yuan, Liangbin Xie, Yuexian Zou, and Ying Shan. Image conductor: Precision control for interactive video synthesis, 2024. 4, 5, 6

[5] 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. In *The Thirteenth International Conference on Learning Representations*, 2025. 4, 5, 6

[6] Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, et al. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. *SIGGRAPH 2024*, 2024. 4, 5, 6

[7] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 4

[8] Hanlin Wang, Hao Ouyang, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Qifeng Chen, Yujun Shen, and Limin Wang. Levitor: 3d trajectory oriented image-to-video synthesis. 2024. 4, 5, 6

[9] Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. Draganything: Motion control for anything using entity representation, 2024. 4, 5, 6

[10] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 4

[11] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023. 4, 5, 6

[12] Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation, 2024. 4, 5, 6