



# FlashMotion: Few-Step Controllable Video Generation with Trajectory Guidance (— Supplementary Materials —)

Anonymous CVPR submission

Paper ID \*\*\*\*

## 1. Additional Ablation results

### 1.1. Quantitative Results

Here, we provide the complete quantitative results across all three benchmarks, including FlashBench, MagicBench [3], and DAVIS [6] in Table. 1 and Table. 2. All ablation studies are trained for 1K steps on 4 Nvidia A100 GPUs, with other training configurations kept consistent with FlashMotion Stage 3.

**Fast Adapter** To assess the importance of the *FastAdapter* training stage, we evaluate the performance of directly applying *SlowAdapter* to *FastGenerator* across all three benchmarks. As shown in Table 1, removing the *FastAdapter* stage results in a consistent decline in both video quality and trajectory accuracy across all benchmarks, underscoring the necessity of the additional *FastAdapter* training stage.

**Diffusion Loss** To evaluate the role of the diffusion loss, we remove it during training and measure performance across all benchmarks. As presented in Table 1, removing the diffusion loss leads to a noticeable drop in trajectory alignment for both adapter architectures. This shows that the diffusion loss is essential for maintaining trajectory consistency between generated motions and user-specified trajectories. Moreover, its removal also causes a degradation in both image and video quality.

**GAN Loss** We conduct an ablation study on the GAN loss, as summarized in Table 1. While removing the adversarial objectives slightly improves trajectory accuracy, it causes an approximately 90% reduction in both image and video quality, introducing severe blurring artifacts.

**Dynamic Diffusion Loss Scaling** We further validate the effectiveness of the proposed dynamic diffusion loss scaling strategy by fixing the loss scale to 1 during training. As shown in Table 1, disabling dynamic scaling leads to a clear decline in both image and video quality across all three benchmarks, again resulting in noticeable blurring artifacts.

**Discriminator Architecture** Finally, we assess the impact of different discriminator architectures, as shown in Table 2. Using only the *Video Cross-Attention* layer yields the lowest performance in both visual quality and trajectory accuracy. In contrast, incorporating the *Semantic Self-Attention* module enhances the model’s semantic understanding, improving visual realism, while the *Trajectory Cross-Attention* module strengthens trajectory control accuracy. Overall, our full discriminator architecture achieves the best results across all evaluation metrics and benchmarks.

### 1.2. More Qualitative Results

Detailed qualitative ablation results are presented in Fig.1, Fig.2, and Fig.3. As shown, directly applying *SlowAdapter* to *FastGenerator* produces pronounced artifacts—such as the color drift in Fig.1 and Fig.3, and the distorted object shapes in Fig.2. In addition, removing the diffusion loss during training markedly degrades trajectory fidelity: objects (e.g., the dog or the bus) drift away from the intended paths, and in the extreme case shown in Fig. 3, a single Spongebob is mistakenly duplicated into two. Finally, eliminating either the GAN loss or the dynamic scale strategy introduces severe blurring artifacts.

Table 1. Comprehensive ablation study of FlashMotion. We analyze both adapter variants (ResNet and ControlNet) by progressively removing key components — including the *FastAdapter* training stage, diffusion loss, GAN loss, and the dynamic diffusion loss scaling strategy. The results show that each component plays a crucial role in preserving high video quality and precise motion alignment.

Methods	FlashBench			MagicBench			DAVIS		
	FID(↓)	FVD(↓)	M/B IoU%(↑)	FID(↓)	FVD(↓)	M/B IoU%(↑)	FID(↓)	FVD(↓)	M/B IoU%(↑)
<b>Adapter Type: ResNet</b>									
Slow Adapter	22.75	168.46	49.79 / 56.62	21.59	162.93	60.24 / 67.23	52.01	992.26	36.33 / 51.37
w/o Diffusion Loss	18.87	161.07	52.04 / 58.04	21.95	162.31	63.14 / 69.02	55.28	983.91	37.22 / 52.47
w/o GAN Loss	22.74	206.75	<b>65.82 / 70.60</b>	30.51	167.91	<b>73.86 / 78.48</b>	66.46	1015.81	<b>47.13</b> / 62.58
w/o Dynamic Scale	26.32	210.93	65.54 / 69.77	21.90	167.00	73.60 / 78.15	73.12	998.85	47.01 / 60.12
<b>FlashMotion</b>	<b>15.81</b>	<b>108.96</b>	63.96 / 70.01	<b>14.16</b>	<b>109.20</b>	72.34 / 77.92	<b>50.58</b>	<b>786.42</b>	46.74 / <b>64.00</b>
<b>Adapter Type: ControlNet</b>									
Slow Adapter	19.44	171.83	62.72 / 69.38	21.19	161.80	70.20 / 76.54	46.42	875.37	50.52 / 70.83
w/o Diffusion Loss	21.21	172.04	55.91 / 61.59	22.36	176.01	66.25 / 71.82	49.27	882.81	42.46 / 59.01
w/o GAN Loss	28.82	265.46	<b>71.56</b> / 75.48	26.33	192.85	<b>78.26</b> / 82.15	75.42	1131.65	<b>55.87</b> / 68.59
w/o Dynamic Scale	19.93	155.55	70.46 / <b>75.89</b>	16.83	131.59	77.49 / <b>82.29</b>	61.47	958.22	55.51 / 70.13
<b>FlashMotion</b>	<b>14.35</b>	<b>96.08</b>	69.15 / 75.38	<b>12.49</b>	<b>99.30</b>	76.92 / 82.17	<b>45.66</b>	<b>690.13</b>	54.54 / <b>74.37</b>

Table 2. Ablation study on the discriminator architecture. VC denotes the *Video Cross-Attention* layer, SS denotes the *Semantic Self-Attention* layer, and TC denotes the *Trajectory Cross-Attention* layer. Results show that our discriminator design achieves the best overall performance across all benchmarks and metrics.

Methods	FlashBench			MagicBench			DAVIS		
	FID(↓)	FVD(↓)	M/B IoU%(↑)	FID(↓)	FVD(↓)	M/B IoU%(↑)	FID(↓)	FVD(↓)	M/B IoU%(↑)
<b>Adapter Type: ResNet</b>									
VC only	16.76	110.83	62.07 / 67.76	14.73	114.61	71.00 / 75.86	53.22	800.50	43.97 / 60.16
SS+VC	16.31	109.02	62.54 / 68.05	14.44	113.88	71.16 / 76.28	52.34	830.14	44.61 / 62.50
TC+VC	16.64	110.01	62.99 / 69.36	14.87	114.11	71.70 / 77.31	53.16	830.57	45.11 / 62.56
<b>FlashMotion</b>	<b>15.81</b>	<b>108.96</b>	<b>63.96 / 70.01</b>	<b>14.16</b>	<b>109.20</b>	<b>72.34 / 77.92</b>	<b>50.58</b>	<b>786.42</b>	<b>46.74 / 64.00</b>
<b>Adapter Type: ControlNet</b>									
VC only	15.56	115.72	63.04 / 71.73	13.71	120.22	75.78 / 81.33	49.39	798.79	51.48 / 69.00
SS+VC	15.37	99.24	65.84 / 72.35	13.42	101.58	75.35 / 81.06	46.24	711.82	53.33 / 71.99
TC+VC	15.70	101.06	68.78 / 73.85	13.96	105.49	76.48 / 82.15	48.50	758.96	53.91 / 72.90
<b>FlashMotion</b>	<b>14.35</b>	<b>96.08</b>	<b>69.15 / 75.38</b>	<b>12.49</b>	<b>99.30</b>	<b>76.92 / 82.17</b>	<b>45.66</b>	<b>690.13</b>	<b>54.54 / 74.37</b>

## 2. Additional Comparison results

### 2.1. Backbone Comparisons

As shown in Table 3, we present a comprehensive comparison of the backbone architectures used across different methods. The table summarizes the supported video length and spatial resolution, as well as the corresponding denoising latency and total parameter count. Notably, FlashMotion achieves the fastest denoising speed for both the ControlNet- and ResNet-based adapters, while also supporting the highest resolution and the longest generation length. Depending on their needs, users can flexibly choose between the ResNet or ControlNet variants of FlashMotion to balance generation speed, video quality, and trajectory accuracy.

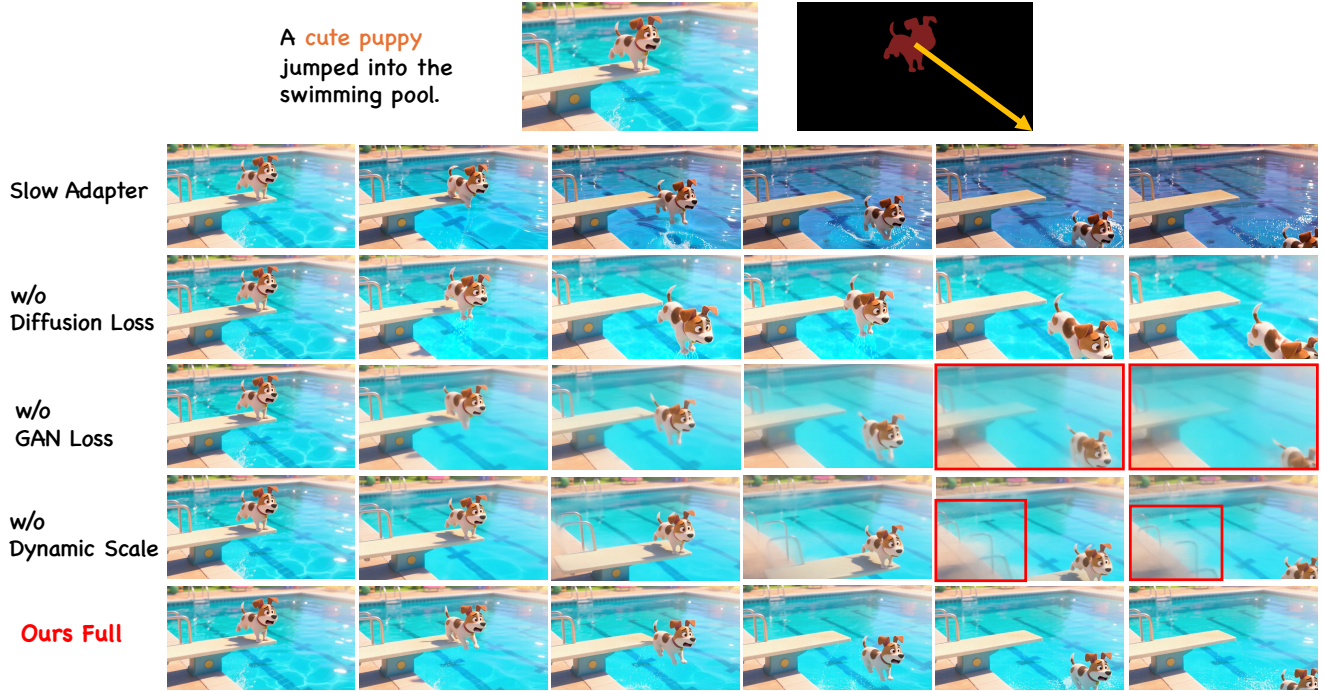


Figure 1. Additional ablation study results. Only our full method can generate videos with both high visual quality and trajectory accuracy.

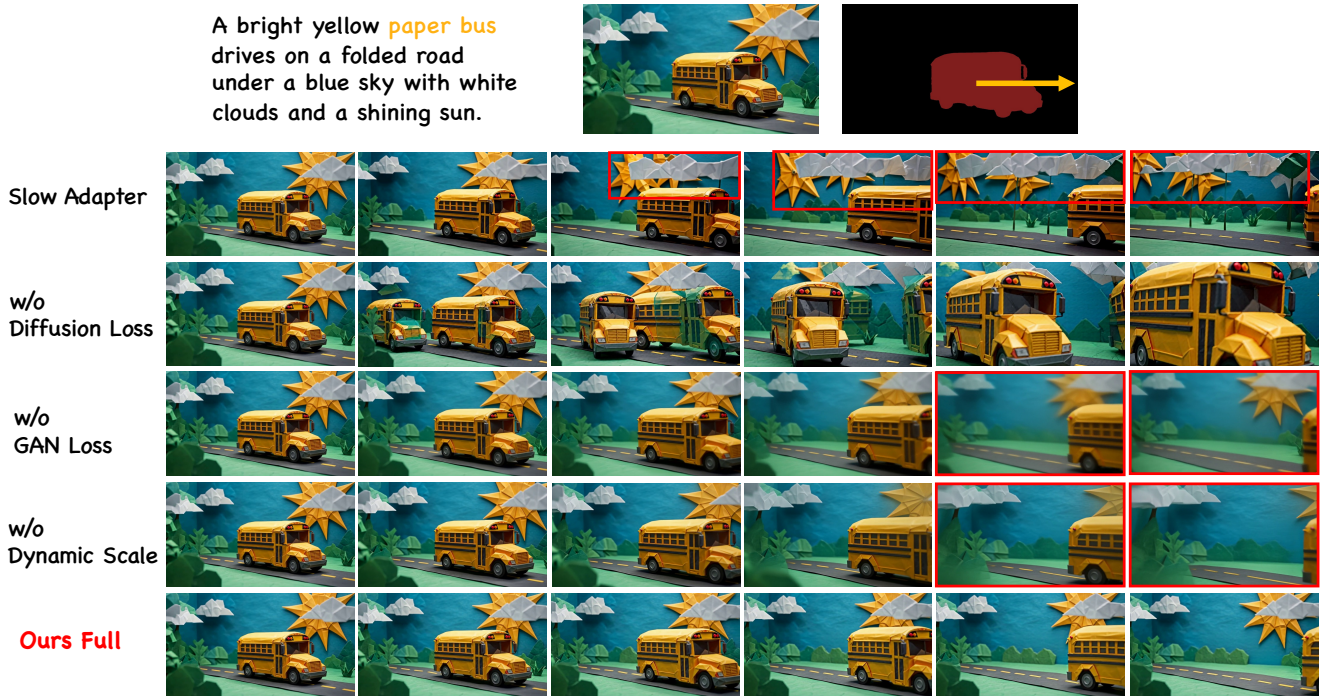


Figure 2. Additional ablation study results. Only our full method can generate videos with both high visual quality and trajectory accuracy.

## 2.2. Results Across Object Counts

Due to space limitations, the main paper only reports the overall quantitative comparison on FlashBench. Here, we present detailed evaluations under different numbers of controlled objects, covering cases from 1–5 to more than 5 foreground objects.

043

044

045





Figure 3. Additional ablation study results. Only our full method can generate videos with both high visual quality and trajectory accuracy.

Table 3. Comparison of model configurations and backbone architectures, including supported video length, spatial resolution, denoising latency, and total parameters. FlashMotion achieves the fastest denoising speed while supporting the highest resolution and longest generation length.

Method	Video Length	Video Resolution	Denoising Latency(s)	Total Params(B)	Base Model
LeviTor [8]	16	288×512	80.08	2.21	SVD [1]
DragAnything [9]	14	320×576	589.07	2.21	SVD [1]
SG-I2V [5]	14	576×1024	1277.15	1.52	SVD [1]
Tora [12]	49	480×720	691.13	6.32	CogVideoX [10]
MagicMotion [3]	49	480×720	1158.63	11.53	CogVideoX [10]
Wan+ResNet [7]	121	704×1280	333.00	5.02	Wan2.2 [7]
Wan+ControlNet [7]	121	704×1280	664.53	10.28	Wan2.2 [7]
<b>FlashMotion (ResNet)</b>	121	704×1280	11.72	5.02	Wan2.2 [7]
<b>FlashMotion (ControlNet)</b>	121	704×1280	24.44	10.28	Wan2.2 [7]

As shown in Table 4 and Table 5, the ControlNet variant of FlashMotion consistently surpasses all competing methods across all metrics, outperforming both multi-step and few-step baselines in terms of visual quality and trajectory accuracy. When using a ResNet-based trajectory adapter, FlashMotion also achieves better visual quality than the previous SOTA method MagicMotion [3], though it still falls slightly short in trajectory accuracy due to the limited parameter capacity.

### 2.3. More Qualitative Results

In this section, we present additional qualitative comparisons with previous methods. As illustrated in Figs. 4–10, FlashMotion accurately controls object trajectories and produces high-quality videos, whereas the other approaches exhibit notable artifacts and inconsistencies. For full video results, please refer to “Supplementary video.mp4” in the supplementary material.



Table 4. Quantitative comparison results on FlashBench for scenes containing 1, 2, and 3 controlled objects. The detailed evaluations show that FlashMotion with a ControlNet-based adapter consistently outperforms all competing methods across all metrics, while the ResNet-based adapter also delivers superior visual quality compared to prior work.

Methods	Obj_Num=1			Obj_Num=2			Obj_Num=3		
	FID(↓)	FVD(↓)	M/B IoU(↑)	FID(↓)	FVD(↓)	M/B IoU(↑)	FID(↓)	FVD(↓)	M/B IoU(↑)
<b>MultiSteps (50 Steps)</b>									
MagicMotion [3]	53.62	741.91	67.93/83.46	59.37	697.50	<u>61.05/73.47</u>	52.44	563.38	<u>66.13/72.92</u>
Wan2.2 (ResNet) [7]	49.01	599.93	61.10/76.34	56.19	582.42	51.49/62.07	57.39	566.53	50.06/56.75
Wan2.2 (ControlNet) [7]	50.04	594.54	66.07/83.98	<u>51.20</u>	591.56	59.64/73.18	49.49	547.90	62.64/70.01
DragAnything [9]	76.28	1076.20	62.70/74.88	91.08	1196.46	53.34/63.06	89.26	1099.45	54.01/57.55
SG-I2V [5]	70.20	984.94	64.09/76.45	78.93	926.79	47.16/57.04	73.08	891.52	48.31/54.25
Tora [12]	73.15	902.55	58.24/69.00	80.27	939.72	46.45/57.47	82.54	869.43	46.80/52.66
LeviTor [8]	128.25	1318.56	49.63/59.73	127.24	1124.07	38.09/44.82	131.60	1252.00	35.65/39.08
<b>FewSteps (4 Steps) — Adapter: ResNet</b>									
DMD [11]	64.71	709.74	55.34/74.30	63.28	687.09	45.21/59.62	64.03	636.34	43.08/53.14
GAN [2]	79.73	728.35	54.58/66.52	77.25	700.88	41.38/51.34	74.52	673.58	41.46/48.80
LCM [4]	58.97	875.26	64.61/80.06	72.26	1032.56	56.40/68.58	65.52	1033.12	53.52/59.67
<b>FlashMotion</b>	<u>46.64</u>	<u>509.36</u>	<u>68.02/84.86</u>	51.21	<u>497.62</u>	60.27/73.08	<u>44.41</u>	<u>433.60</u>	63.40/71.59
<b>FewSteps (4 Steps) — Adapter: ControlNet</b>									
DMD [11] / GAN [2]	OOM								
LCM [4]	61.13	851.48	62.83/76.15	76.41	929.77	56.68/66.79	69.65	831.79	57.86/63.51
<b>FlashMotion</b>	<b>44.97</b>	<b>465.86</b>	<b>68.44/84.51</b>	<b>46.16</b>	<b>437.18</b>	<b>63.87/76.99</b>	<b>42.20</b>	<b>422.16</b>	<b>66.45/73.91</b>

### 3. Case Studies

#### 3.1. Different Styles

As shown in Fig. 11, FlashMotion supports generating videos across diverse visual styles, including dreamlike realism, surreal miniature photography, 3D cartoon rendering, and Eastern ink-wash painting. To better demonstrate the model’s robustness and its ability to maintain consistent motion across challenging layouts, we deliberately choose vertically oriented images instead of horizontal ones. These examples collectively illustrate FlashMotion’s strong adaptability to various artistic domains while preserving coherent structure and motion.

#### 3.2. Camera Control

FlashMotion supports camera control operations such as zooming in and zooming out. As shown in Fig. 12, the camera motion can be adjusted by manipulating the bounding box size of the foreground object, such as the cup or the woman’s mask. Furthermore, as illustrated in Fig. 13, users can navigate scenes—like a bakery or a museum—by controlling the bounding boxes of objects such as the dinosaur, the mammoth, or the industrial mixer.

### 4. More Details on FlashBench

FlashBench comprises 600 videos, grouped into six categories based on the number of foreground objects (ranging from 1–5 and more than 5). To offer a more comprehensive analysis of the dataset, we further visualize the distributions of video lengths as shown in Fig. 14, demonstrating its support for evaluating long video generation.

Table 5. Quantitative comparison results on FlashBench for scenes containing 4, 5, and above 5 controlled objects. The detailed evaluations show that FlashMotion with a ControlNet-based adapter consistently outperforms all competing methods across all metrics, while the ResNet-based adapter also delivers superior visual quality compared to prior work.

Methods	Obj_Num=4			Obj_Num=5			Obj_Num>5		
	FID(↓)	FVD(↓)	M/B IoU(↑)	FID(↓)	FVD(↓)	M/B IoU(↑)	FID(↓)	FVD(↓)	M/B IoU(↑)
<b>MultiSteps (50 Steps)</b>									
MagicMotion [3]	45.67	546.40	70.29/73.21	44.41	450.10	73.86/76.93	44.41	409.25	69.29/62.35
Wan2.2 (ResNet) [7]	61.69	575.65	50.89/53.93	52.04	476.04	55.56/56.98	41.59	453.60	44.31/41.03
Wan2.2 (ControlNet) [7]	49.25	503.03	66.15/68.65	43.58	409.57	70.70/70.94	37.11	406.06	67.27/61.05
DragAnything [9]	75.00	997.03	59.97/60.23	83.35	812.67	62.92/61.48	97.48	1006.25	56.95/49.51
SG-I2V [5]	64.83	861.49	50.87/55.46	65.91	713.41	54.21/55.83	66.22	828.14	36.75/35.52
Tora [12]	65.25	737.03	46.28/51.05	73.88	714.65	52.76/54.55	93.60	1073.05	37.98/36.89
LeviTor [8]	167.97	1774.66	35.10/34.02	185.75	2015.57	33.33/30.34	135.75	1287.67	24.23/23.41
<b>FewSteps (4 Steps) — Adapter: ResNet</b>									
DMD [11]	66.08	749.99	41.38/48.44	67.03	697.32	42.02/47.94	52.62	671.65	32.74/32.74
GAN [2]	69.55	571.76	45.87/50.43	65.83	500.22	48.67/52.49	59.83	584.86	31.00/30.78
LCM [4]	62.24	959.97	57.30/57.89	58.71	869.45	56.78/58.98	49.66	780.45	43.51/40.21
<b>FlashMotion</b>	<u>38.47</u>	<u>411.71</u>	66.58/67.87	<u>39.53</u>	<u>326.98</u>	68.67/79.78	<u>37.07</u>	<u>384.06</u>	56.92/52.02
<b>FewSteps (4 Steps) — Adapter: ControlNet</b>									
DMD [11] / GAN [2]	OOM								
LCM [4]	60.28	752.48	63.18/63.38	56.29	637.06	66.48/65.66	53.64	541.55	60.79/53.62
<b>FlashMotion</b>	<b>36.62</b>	<b>367.24</b>	<b>71.81/75.49</b>	<b>35.19</b>	<b>294.47</b>	<b>74.94/76.98</b>	<b>32.72</b>	<b>305.68</b>	<b>69.43/64.50</b>

A tiny hamster in a pistachio hat drives a bread-bulldozer, pushing rainbow sprinkles across the floor.

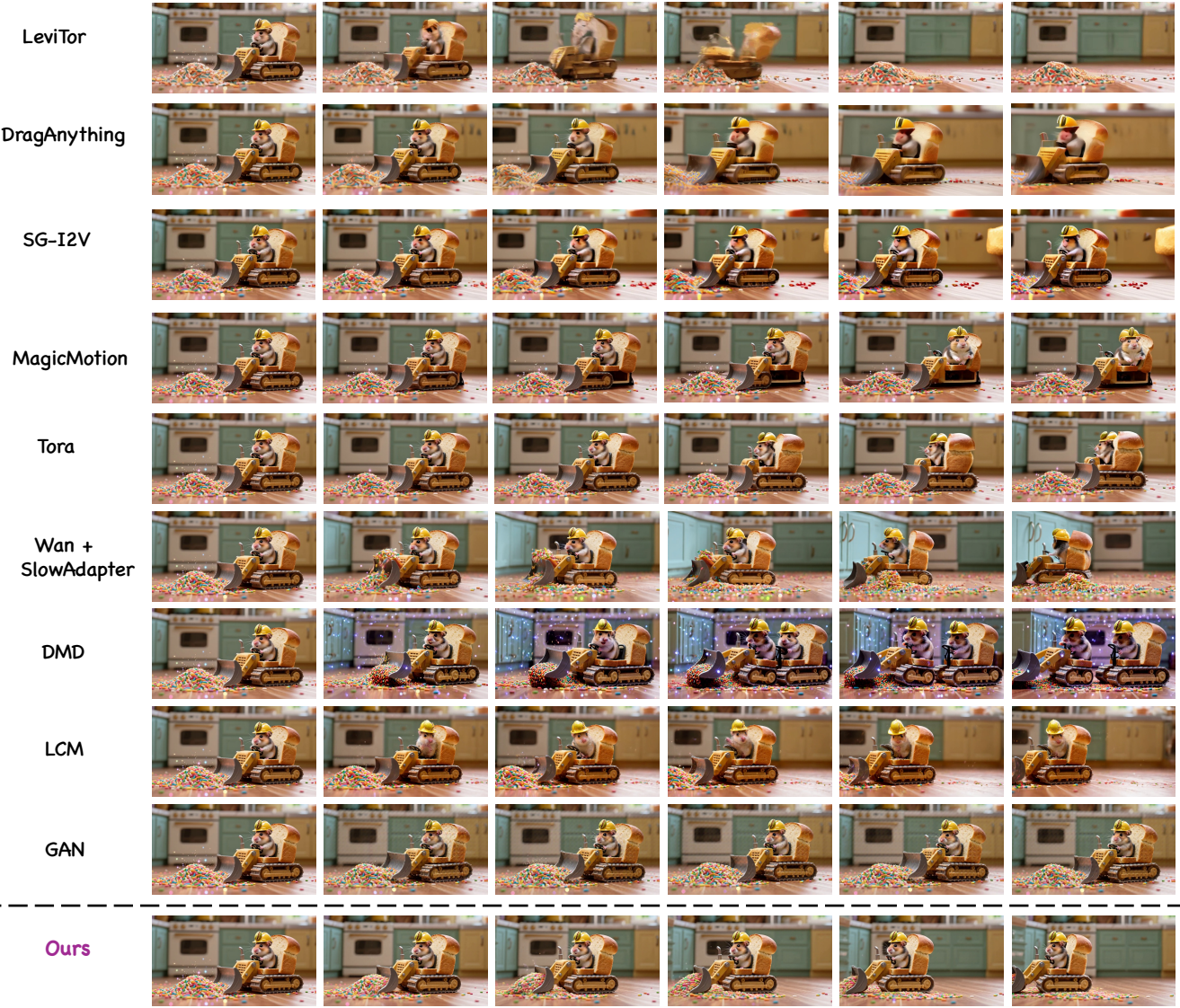


Figure 4. Qualitative Comparisons results with different methods.



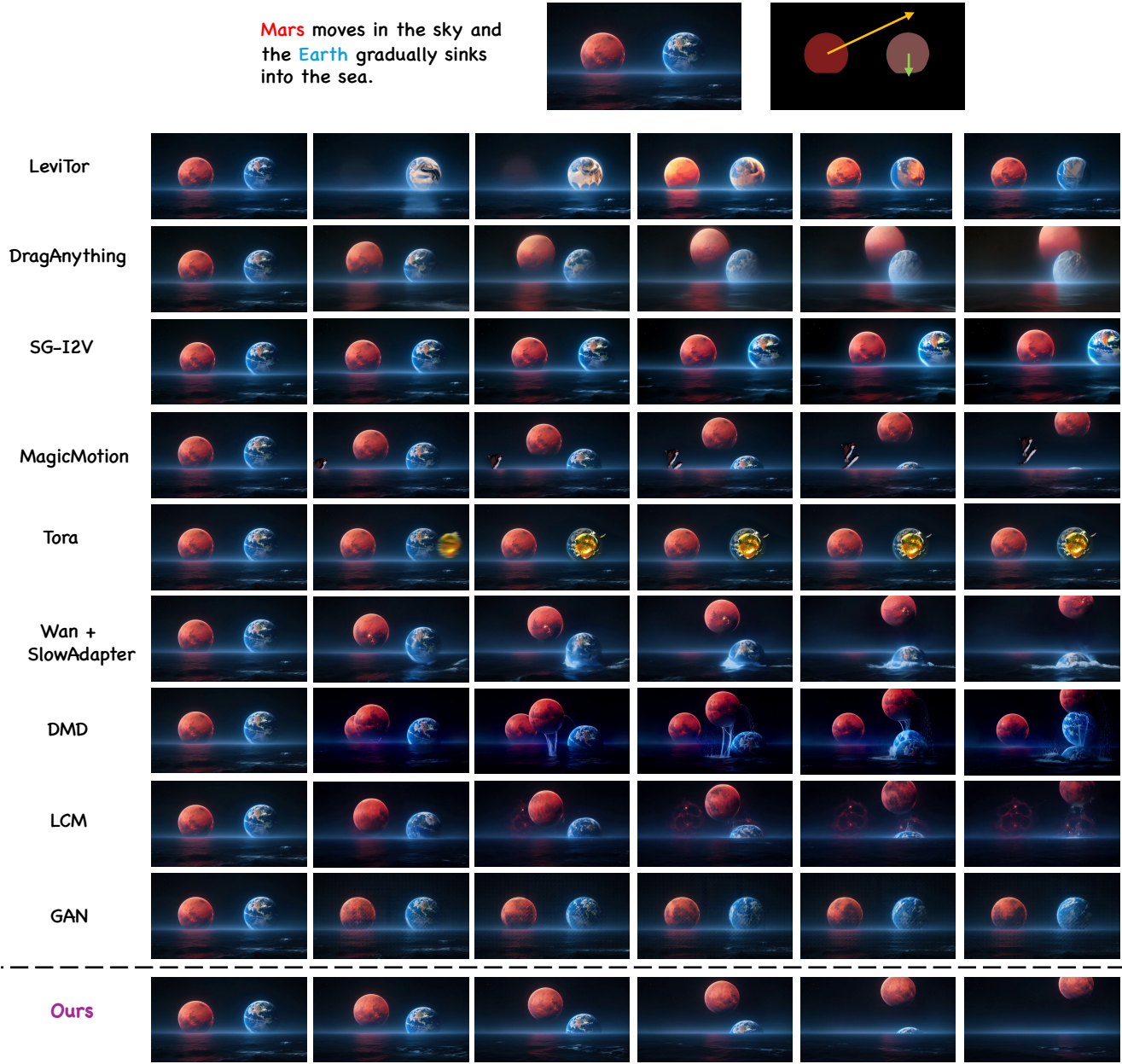


Figure 5. Qualitative Comparisons results with different methods.

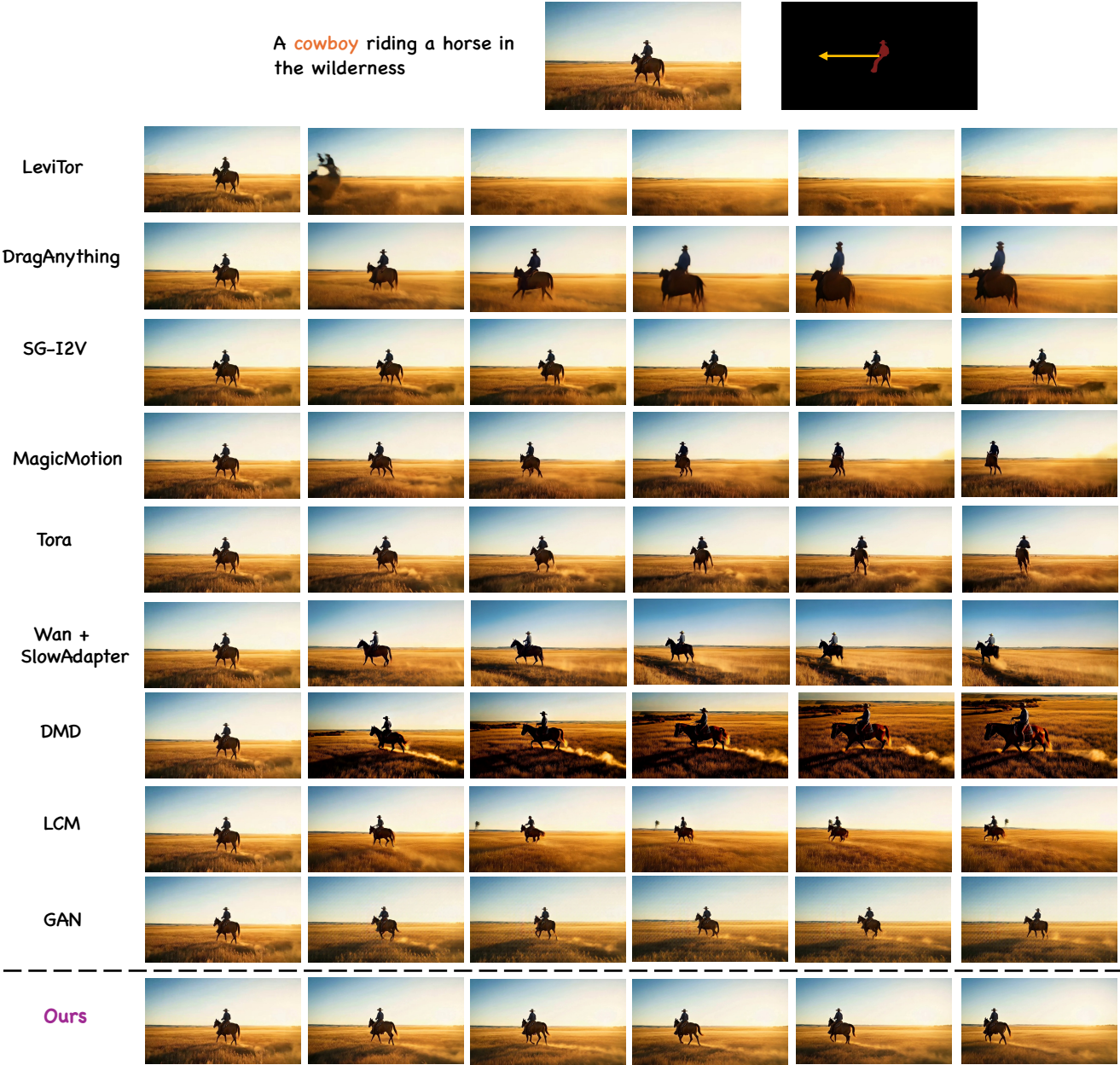


Figure 6. Qualitative Comparisons results with different methods.



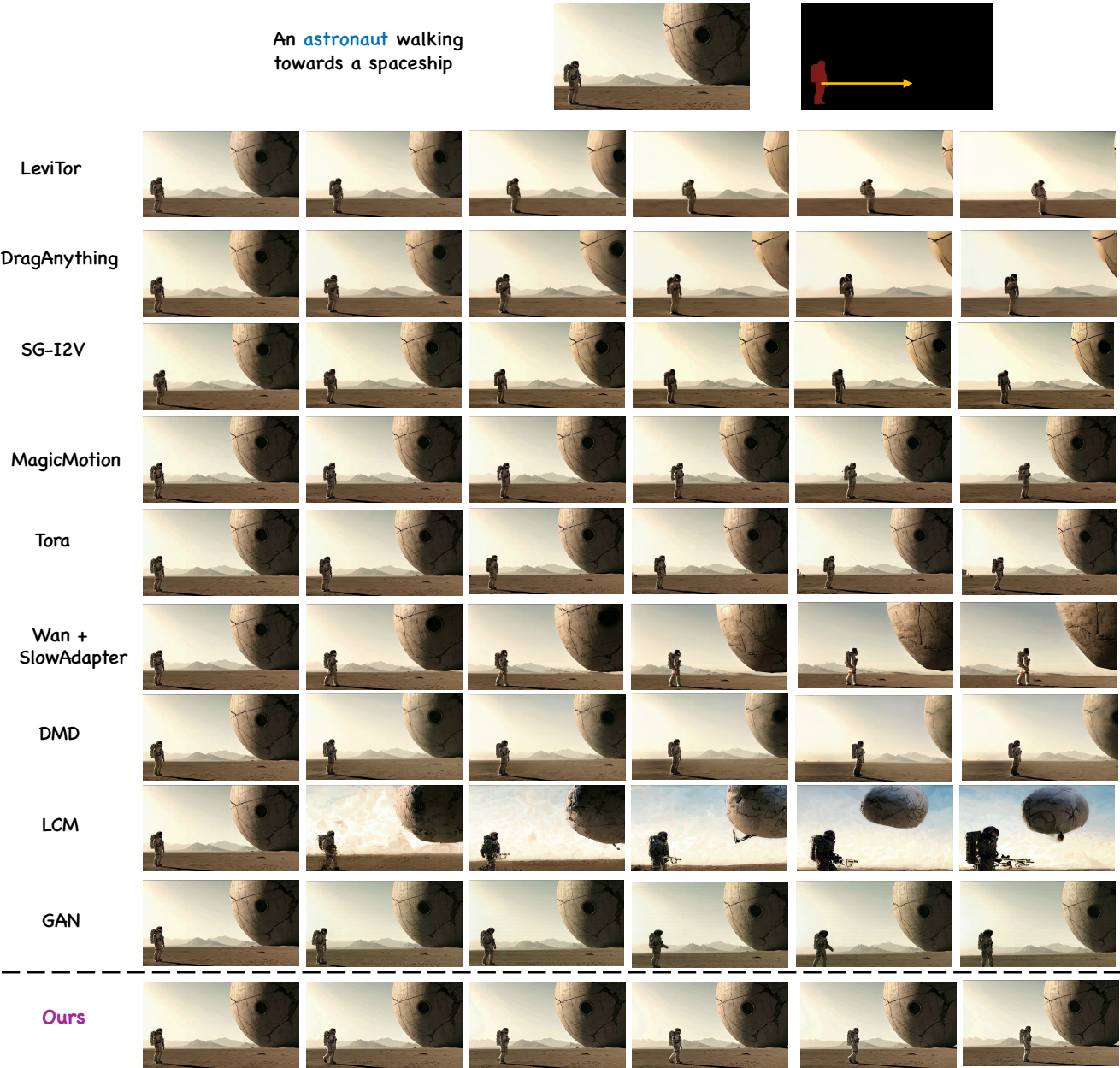


Figure 7. Qualitative Comparisons results with different methods.



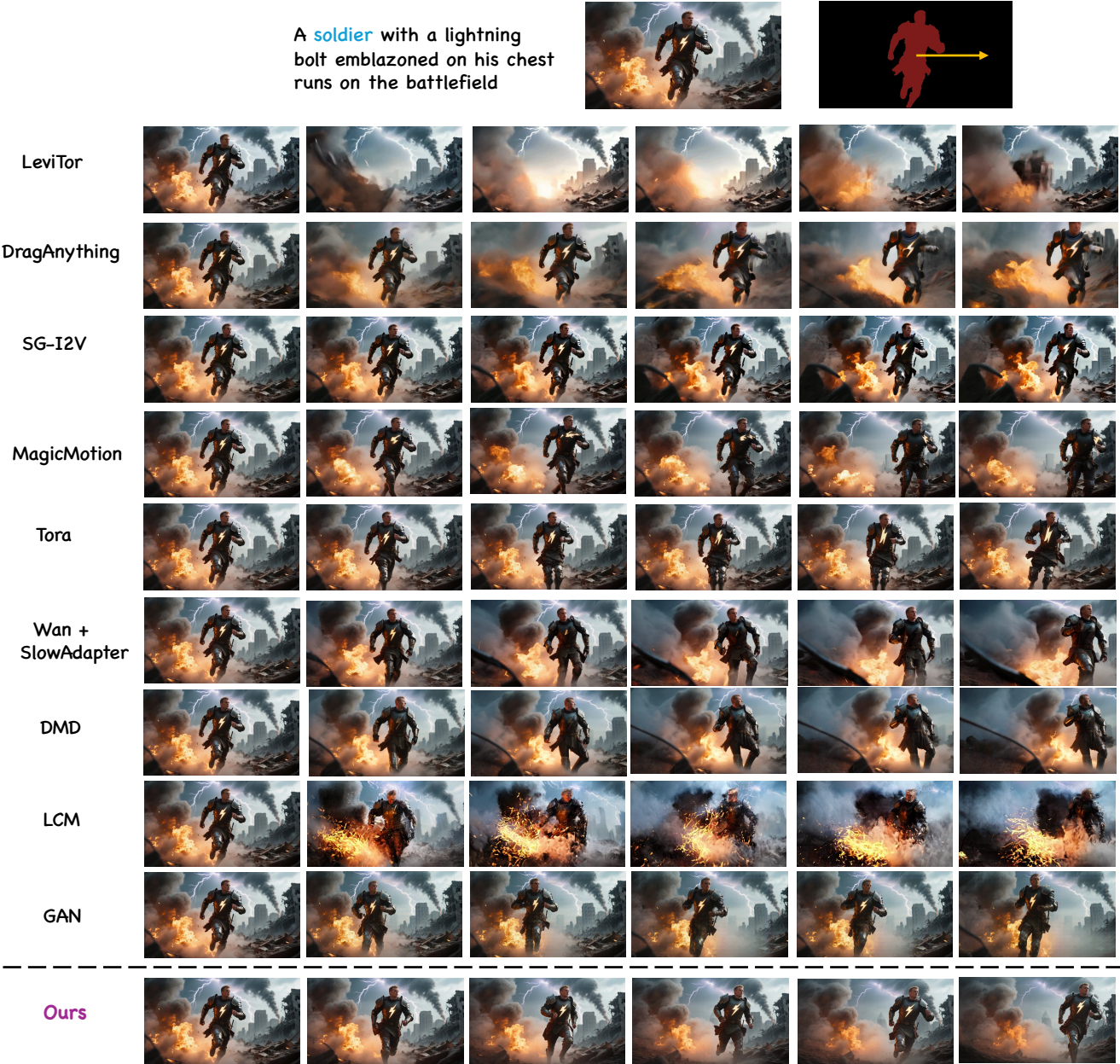


Figure 8. Qualitative Comparisons results with different methods.



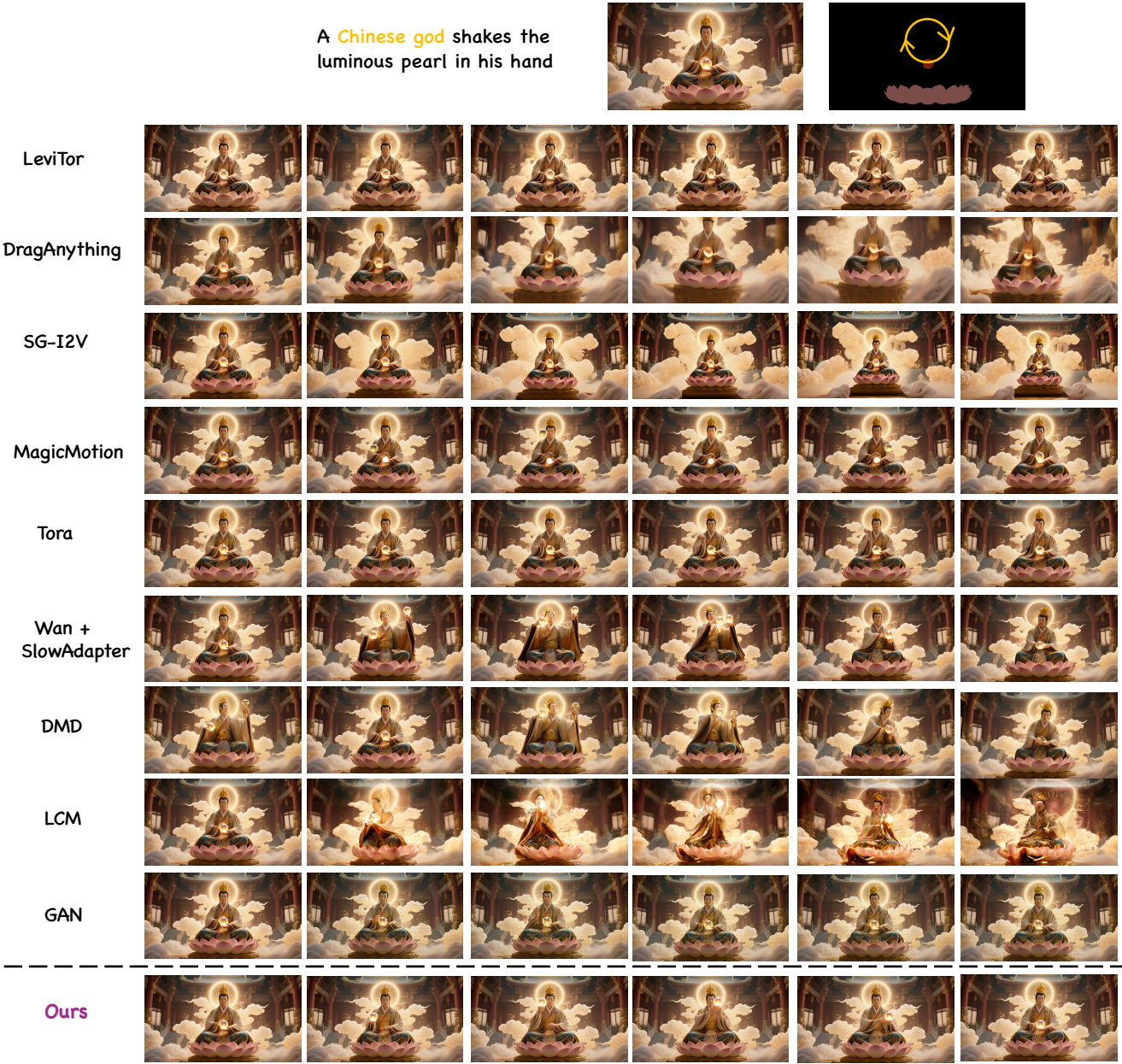


Figure 9. Qualitative Comparisons results with different methods.



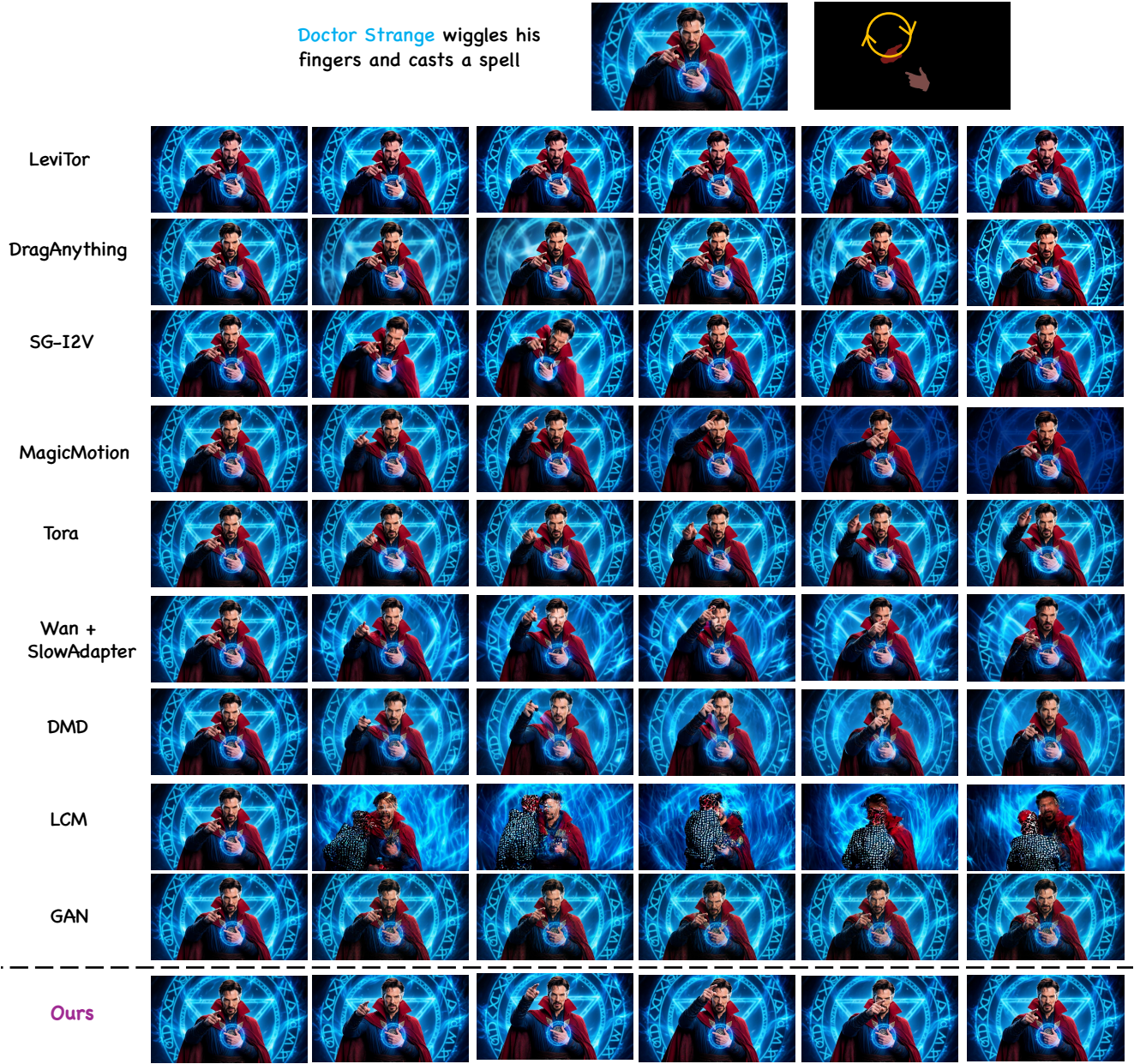


Figure 10. Qualitative Comparisons results with different methods.



Dreamlike  
Realism



Surreal  
Photography



Cartoon



Ink Painting



Figure 11. FlashMotion supports generating videos of different styles.

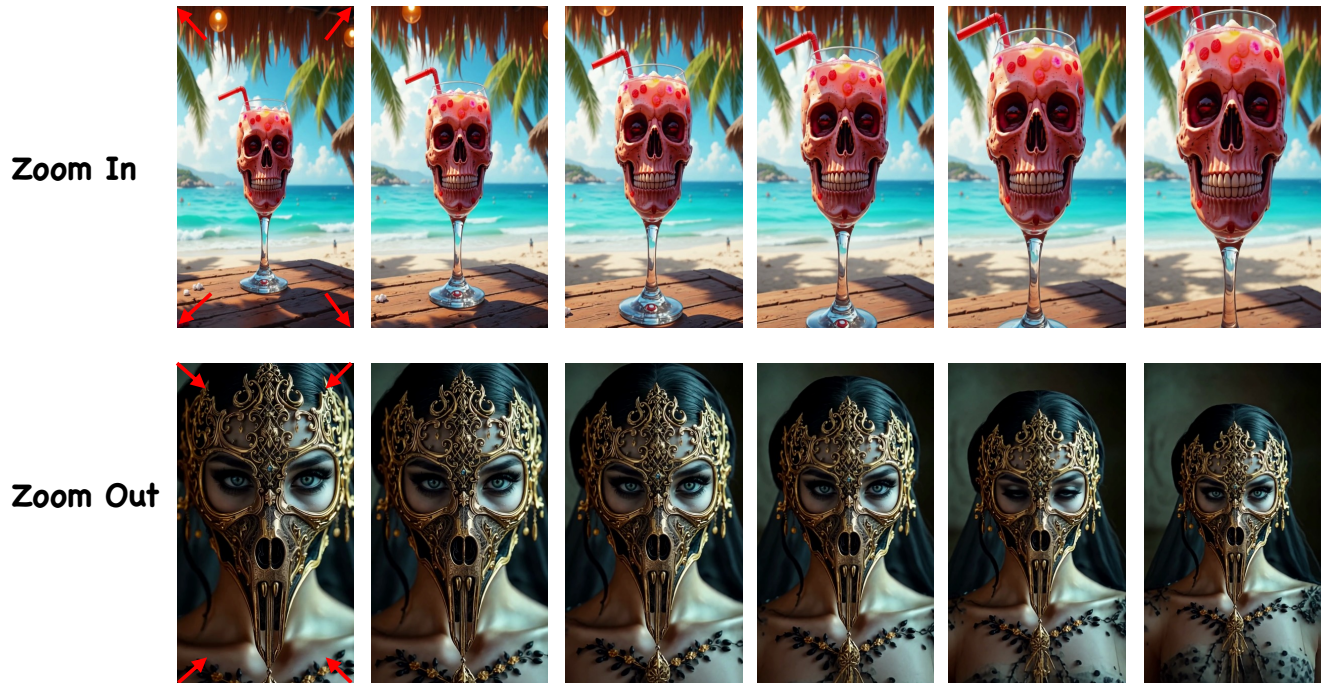


Figure 12. FlashMotion enables controllable camera movements, such as zooming in or out, by adjusting the bounding box size of the foreground object (e.g., the cup or the woman’s mask).

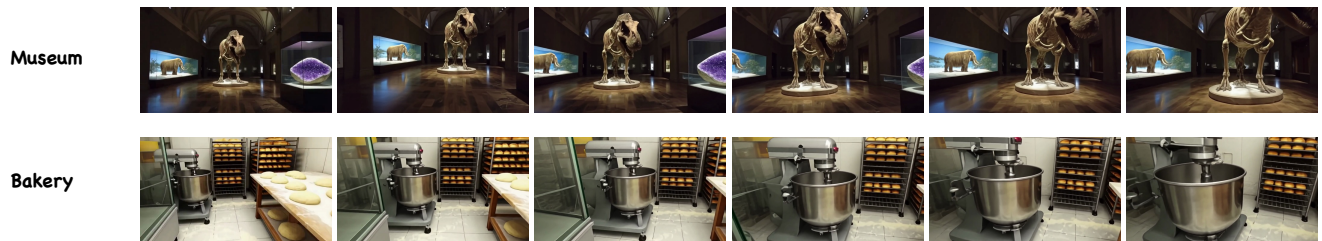


Figure 13. FlashMotion supports scene navigation in various environments—such as a bakery or a museum—by manipulating the bounding boxes of key objects, including the dinosaur, the mammoth, and the industrial mixer.

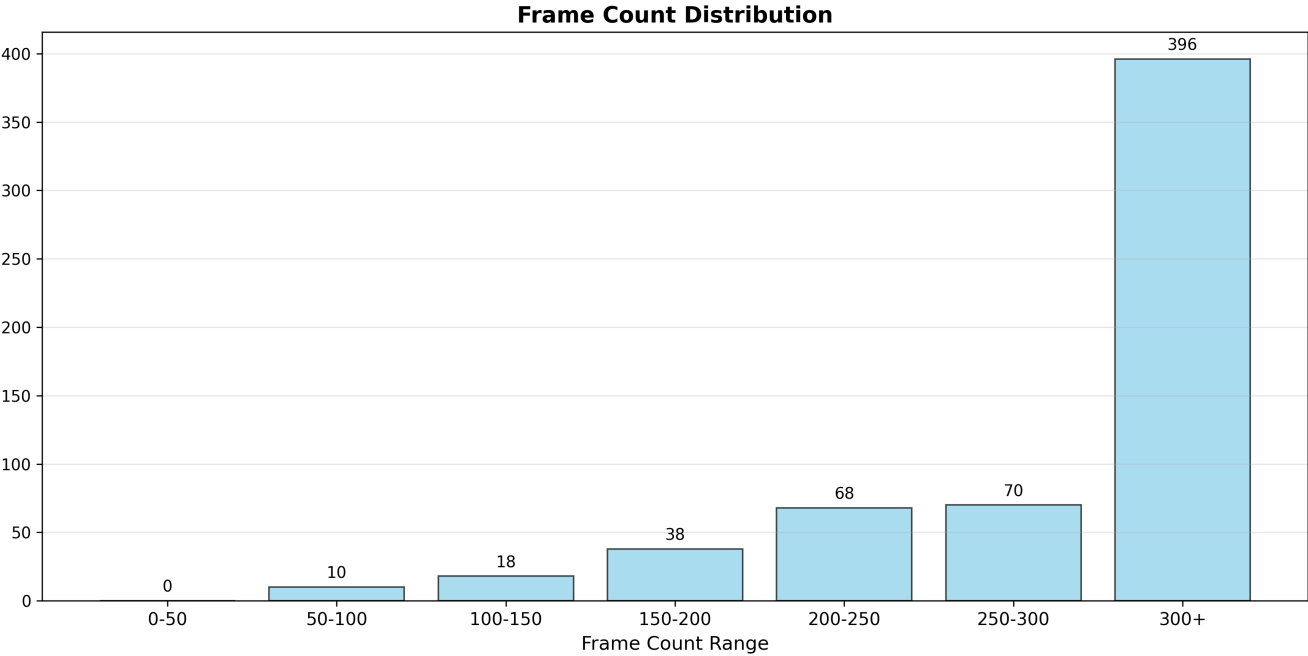


Figure 14. Distribution of video frame counts in FlashBench, demonstrating its support for evaluating long video generation.



## 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. 4 071 072 073 074
- [2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 2014. 5, 6 075 076
- [3] Quanhao Li, Zhen Xing, Rui Wang, Hui Zhang, Qi Dai, and Zuxuan Wu. Magicmotion: Controllable video generation with dense-to-sparse trajectory guidance. In *ICCV*, 2025. 1, 4, 5, 6 077 078
- [4] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 5, 6 079 080
- [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 *ICLR*, 2025. 4, 5, 6 081 082
- [6] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016. 1 083 084
- [7] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwei 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, Wenten 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, 5, 6 085 086 087 088 089 090 091 092
- [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. In *CVPR*, pages 12490–12500, 2025. 4, 5, 6 093 094
- [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. In *ECCV*. Springer, 2024. 4, 5, 6 095 096
- [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. 4 097 098 099
- [11] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. In *NeurIPS*, 2024. 5, 6 100 101
- [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. In *CVPR*, 2025. 4, 5, 6 102 103