LAMP: Language-Assisted Motion Planning for Controllable Video Generation

Supplementary Material

This document provides additional details and results complementing the main paper. Section 1 summarizes the statistics of our procedural dataset and visualizes the distribution of motion types. Section 2 presents extended quantitative evaluations on DataDoP and ET, including experiments with DSL-converted real data. Section 4 shows further qualitative examples, long-horizon generation, and multi-object motion results. Section 5 describes the user study setup. Section 7 provides a full specification of our motion domain-specific language (DSL). Section 6 discusses typical failure cases of the video generator and future opportunities for improving trajectory conditioning.

1. Procedural Dataset Statistics

Our procedural dataset consists of 200+K samples for camera (100K with free-form and 100K with relative motion) and 100K samples for object motion. Each object motion trajectory consists of 4 segments, where one motion primitive is assigned per segment, resulting in 400K segments in total. Each sample includes a DSL program, natural-language camera caption, and the corresponding 3D trajectory.

Fig. 8 provides a detailed distribution of free-form camera motions across 27 coarse translational motion categories (3 motion types × 3 directions) together with all rotational motion combinations. Fig. 7 shows the analogous distribution for object motions. As in real datasets [6, 36], the distribution is intentionally imbalanced in that common motions single-axis motions (e.g., forward, backward) appear more frequently than multi-axis composite motions.

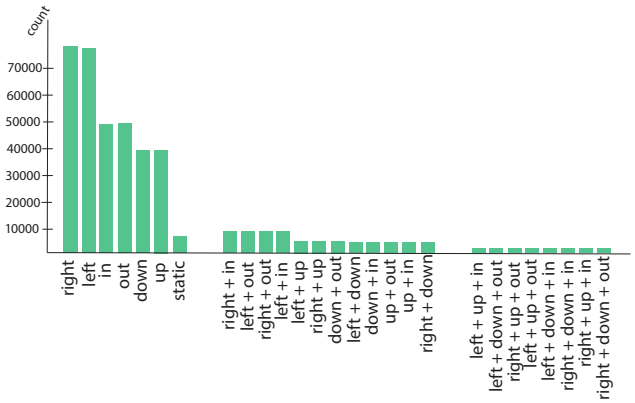

Figure 7. **Object motion distribution.** Frequency of single-axis and multi-axis object motions within our procedural dataset, capturing both dominant patterns and less common combinations.

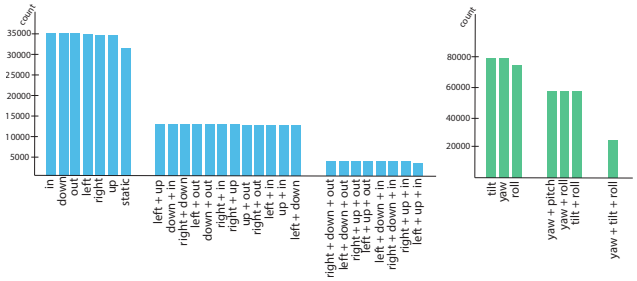

Figure 8. **Free-form camera motion distribution.** Frequency of single-axis and multi-axis motions for both translation (left) and rotation (right), illustrating the natural long-tailed structure of sampled camera behaviors.

2. Additional Quantitative Evaluations for Camera Motion

The main paper evaluates cross-dataset generalization by training LAMP solely on our procedural dataset and testing on DataDoP [36] and ET [6]. Here, we analyze the inverse setting: training or finetuning LAMP directly on these datasets after converting their trajectories to our DSL.

2.1. Training on DataDoP

DataDoP trajectories, which are extracted from real videos, are inherently noisy (see Fig. 9). After DSL conversion, we compute similarity between original and DSL-reconstructed trajectories, hiscard samples below a threshold, and remove static or near-static cases. This yields 10K usable samples. We evaluate four configurations: (i) DSL trajectories + origi-

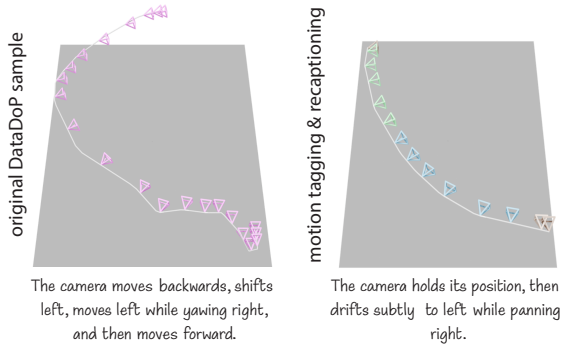

Figure 9. **Using real data for training.** Samples from real datasets (e.g., DataDoP) are converted to our DSL format via motion tagging and optionally re-captioned. The DSL conversion smooths the noisy trajectories extracted from real videos and improves alignment between textual descriptions and motion.

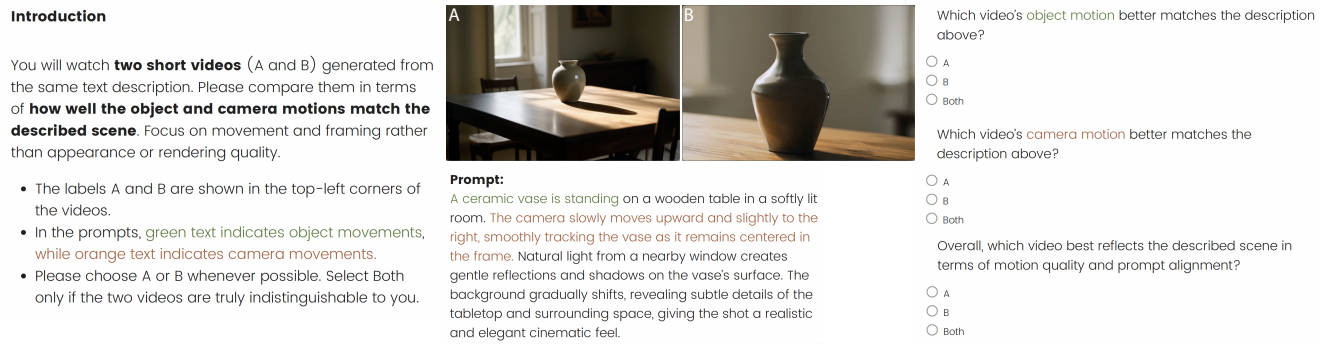

Figure 10. **User-study interface.** Given a prompt and a pair of videos, users compare the outputs along three criteria.

nal captions, (ii) DSL + regenerated captions, (iii) Finetuning (*ft*), and (iv) Training from scratch (*tr*).

Table 5 shows that while additional training data yields small gains, the improvements remain marginal due to the limited data volume. Importantly, the results confirm that our procedural dataset is already diverse and enables strong cross-dataset generalization.

In Table 5, the results for the first five baselines are directly sourced from the original GenDoP publication, while the subsequent rows represent our experimental results obtained using the official implementation. A critical revision in our evaluation involves the F1-score computation[1]. The original GenDoP framework evaluates F1-scores on trajectories reconstructed by CLaTr. However, this introduces a dependency on CLaTr's reconstruction quality, which serves as a bottleneck given its limited training on DataDoP. Aligning with the rationale established in ET, we consider direct evaluation to be more indicative of the model's true generative performance. Consequently, we revised the metric to evaluate the direction and rotation of the predicted trajectories directly. This refinement eliminates the reconstruction bias and yields a significant improvement in the observed F1-scores.

Table 5. **Camera trajectory evaluation on the DataDoP dataset.** LAMP achieves performance comparable to DataDoP-trained baselines despite no dataset-specific training.

Model	Data	Revised F1-Score	F1-Score	CLaTr	Coverage	FID
CCD	pretrained	-	0.297	5.29	0.332	357.822
ET	pretrained	-	0.330	2.46	0.020	609.906
Director3D	pretrained	-	0.058	0.00	0.171	542.385
Director3D	DataDoP	-	0.391	31.69	0.839	31.979
GenDoP	DataDoP	-	**0.400**	**36.18**	**0.872**	**22.714**
GenDoP (exp)	DataDoP	0.360	0.383	35.91	**0.853**	**48.123**
Ours	pretrained	0.763	0.380	36.29	0.794	66.86
Ours	(ft) w/ DataDoP org cap	0.613	0.390	29.44	0.834	85.46
Ours	(ft) w/ DataDoP	**0.776**	**0.390**	**36.52**	0.779	67.24
Ours	(tr) w/ DataDoP org cap	0.616	0.385	29.13	0.835	91.51
Ours	(tr) w/ DataDoP	**0.776**	**0.400**	35.48	0.805	71.52

[1]We have verified the necessity of this revision with the authors of GenDoP.

2.2. Training on ET

We perform a similar analysis with the ET dataset. After filtering and DSL conversion, we obtain $21k$ samples from the ET dataset for additional training. Since the ET dataset lacks rotational camera motion, limiting potential gains.

Table 6 reports results on the pure and mixed splits. Similar to our observations with DataDoP evaluations, DSL-converted ET data yields minor improvements when finetunes, yet LAMP trained purely on our procedural data already achieves the best overall generalization, further demonstrating the strength of our controlled motion design.

Table 6. **Camera trajectory results on the ET dataset.** The ET benchmark includes a simpler *pure* and a harder *mixed* split. LAMP attains the highest F1-scores on both, demonstrating strong generalization across (unseen) motion complexity levels.

Model	Pure Split		Mixed Split	
	F1 Score	CLaTr Score	F1 Score	CLaTr Score
CCD	0.27	3.21	0.17	6.26
MDM	0.76	21.26	0.34	18.32
ET - DirB	0.86	23.10	0.39	20.78
ET - DirC	0.80	21.49	0.48	21.95
Ours	0.976	**35.10**	0.769	36.59
Ours (ft w/ ET org cap)	0.666	30.66	0.446	28.06
Ours (ft w/ ET)	0.978	35.02	**0.779**	**36.95**
Ours (tr w/ ET)	**0.980**	**35.10**	0.755	36.37
Ours (tr w/ DataDoP)	0.967	34.90	0.747	35.71

3. Universality of DSL Output

Since LAMP produces explicit 6-DoF camera and object trajectories, it provides a **model-agnostic interface** that does not require any modification to the video generator. We integrated the same DSL-derived motion signal with off-the-shelf video generation frameworks: *text-to-video* (CameraCtrl [11]), *image-to-video* (EPiC [32]), and *video-to-video* (ReCamMaster [2]). As shown in Fig. 11, the identical DSL program successfully drives all three pretrained backbones.

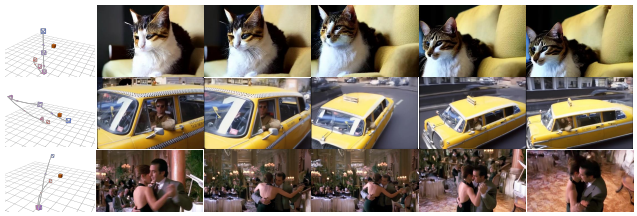

Figure 11. **DSL w/ CameraCtrl, EPiC, and ReCamMaster**. DSL applied to multiple off-the-shelf models (T2V, I2V, V2V, respectively).

4. Additional Qualitative Results

We provide extensive qualitative examples on the project page. Beyond reproducing results in the main paper, we report:

- **Multi-object motion:** LAMP can generate trajectories for multiple objects independently, followed by camera motion relative to one of them. After manually aligning initial object positions, VACE can synthesize videos conditioned on the resulting 3D paths. Automating relational multi-object trajectory generation is left for future work. Automating this step would allow LLMs to define complex temporal and spatial scene layouts and interactions directly. Since an MLLM call (Qwen2.5-VL-7B) typically takes less than 5 seconds and trajectory rendering requires only 2–3 seconds, users can validate motion in near real-time. This allows for rapid iteration before committing to the final video synthesis, which takes around 6 minutes for 81 frames with VACE-1.3B (on NVIDIA A40 48GB).
- **Long-horizon motion generation:** We prompt LAMP sequentially across temporal segments. DSL-generated trajectories are concatenated by initializing each segment with the previous endpoint. VACE, which is limited to short clips, is run iteratively using the last frame as conditioning. We expect that the visual quality of the results will improve as the video models become capable of generating longer video sequences. While we anticipate that future video models will achieve higher visual fidelity for longer sequences, explicit trajectory control will remain essential. Such control is critical for orchestrating complex, multi-stage movements and ensuring precise spatiotemporal planning that simple text prompts cannot guarantee.

5. User Study Details

As illustrated in Fig. 10, each question in the user study presents a text prompt and two generated videos (ours vs. a baseline), and three evaluation questions. Video order is randomized per trial, and users may choose either video or select both when undecided.

6. Limitations

Failure Rate. We evaluated the validity of the DSL tags generated by LAMP to assess their structural and semantic integrity. The results demonstrate that LAMP is empirically robust: across four datasets (10k captions each), it shows a **0.11% failure rate**, with only 3 motion type confusions and 8 invalid tags per 10k samples. This highlights the stability of the DSL and deterministic grounding.

Video Adherence to Motion Conditioning Controls. In our work, we use VACE [13] as a pre-trained video generator and convert the motion trajectories to intermediate control videos. While effective, this conditioning mechanism sometimes falls short in adhering to the conditioning signals as shown in the supplementary video. Finetuning a video generator to be conditioned on directly the 3D motion trajectories can mitigate this limitation in the future.

7. Motion Domain Specific Language

In Table 7-10, we provide the details of the motion domain specific language we propose. Our DSL consists of four motion primitives, each with a corresponding list of motion modifiers that parameterize that motion type. In particular, for each motion primitive supported, we provide a list of the motion modifiers along with their description, syntax, the possible set of values, and the default value.

Table 7. Free form behaviour (free_form) and its modifiers

Modifier	Description	Key	Possible Values	Default
Lateral	Lateral translation	t_x	{far_left, left, near_left, no, near_right, right, far_right}	no
Vertical	Vertical translation	t_y	{far_down, down, near_down, no, near_up, up, far_up}	no
Depth	Depth translation	t_z	{far_in, in, near_in, no, near_out, out, far_out}	no
Yaw	Yaw in degrees	yaw	{-180, -170, -160, ..., -100, -90, -85, -80, ..., -10, -5, 0, 5, 10, ..., 80, 85, 90, 100, ..., 160, 170, 180}	0
Pitch	Pitch in degrees	pitch	{-180, -170, -160, ..., -100, -90, -85, -80, ..., -10, -5, 0, 5, 10, ..., 80, 85, 90, 100, ..., 160, 170, 180}	0
Roll	Roll in degrees	roll	{-180, -170, -160, ..., -100, -90, -85, -80, ..., -10, -5, 0, 5, 10, ..., 80, 85, 90, 100, ..., 160, 170, 180}	0

Table 8. Orbit track behaviour (orbit_track) and its modifiers

Modifier	Description	Key	Possible Values	Default
Dutch	Camera roll angle	dutch	{-45, -30, -15, 0, 15, 30, 45}	0
Easing	Defines the camera acceleration curve	ease	{in, out, in_out, out_in, linear}	linear
Jitter	Small, random vibrations, simulating a handheld effect	jitter	{low, high, none}	none
Vertical Angle	Camera's vertical perspective relative to the object	ver	{aerial, low-angle, none}	none
Framing Offset	Offsets the object's position within the frame	object	{left, right, none}	none
Orbit Plane	The primary axis around which the camera orbits the object	plane_axis	{x, y, z}	y
Orbit Degrees	Total angular distance	deg	{30, 45, 60, 90, 180, 270, 360}	90
Direction	Direction of rotation	dir	{cw, ccw}	cw
Spiral Dolly	Combines the orbit motion with a simultaneous camera movement towards object, creating a spiral	spiral	{in_0.1, in_0.3, in_0.5, out_0.1, out_0.3, out_0.5, no}	no

Table 9. Tail track behaviour (`tail_track`) and its modifiers

Modifier	Description	Key	Possible Values	Default
Dutch	Camera roll angle	`dutch`	{-45, -30, -15, 0, 15, 30, 45}	0
Easing	Defines the camera acceleration curve	`ease`	{in, out, in_out, out_in, linear}	linear
Jitter	Small, random vibrations, simulating a handheld effect	`jitter`	{low, high, none}	none
Vertical Angle	Camera's vertical perspective relative to the object	`ver`	{aerial, low-angle, none}	none
Framing Offset	Offsets the object's position within the frame	`object`	{left, right, none}	none
Follow style	Responsiveness of the camera (strict vs delayed)	`follow_style`	{hard, soft, lazy}	hard
Follow axis	Axis or axes the camera uses to follow the object	`follow_axis`	{x, y, z, full}	full
Amplitude	Scales the camera's travel distance relative to the object	`amp`	{x_0.5, x_0.8, x_1.2, x_1.5, y_0.5, y_0.8, y_1.2, y_1.5, z_0.5, z_0.8, z_1.2, z_1.5, all_0.5, all_0.8, all_1.2, all_1.5, no}	no
Static Dolly	Moves the camera toward or away from the object	`dolly`	{in_0.1, in_0.3, in_0.5, out_0.1, out_0.3, out_0.5, no}	no
Mirror	Produces symmetrical camera motion	`mirror_axis`	{x, y, no}	no
Look at	Disables or enforces orientation towards the objects	`dont_look`	{dont_look, none}	none
Lead	Positions the camera ahead of the object's motion direction,	`lead`	{lead, none}	none

Table 10. Rotation track behaviour (`rotation_track`) and its modifiers

Modifier	Description	Key	Possible Values	Default
Dutch	Camera roll angle	`dutch`	`{-45, -30, -15, 0, 15, 30, 45}`	`0`
Easing	Defines the camera acceleration curve	`ease`	{`in`, `out`, `in_out`, `out_in`, `linear`}	`linear`
Jitter	Adds small, random vibrations, simulating a handheld effect	`jitter`	{`low`, `high`, `none`}	`none`
Vertical Angle	Camera's vertical perspective relative to the object	`ver`	{`aerial`, `low-angle`, `none`}	`none`
Framing Offset	Offsets the object's position within the frame	`object`	{`left`, `right`, `none`}	`none`
Rotation axis	Determines the rotation axis or axes	`rot_axis`	`{pan, tilt, full}`	`full`
Local Dolly	Controls the distance of the camera to the object during tracking	`push`	`{in_0.1, in_0.3, in_0.5, out_0.1, out_0.3, out_0.5, no}`	`no`
Local Offset	Shifts the camera's look-at point relative to the target	`local_offset`	`{x_-0.3, x_-0.1, x_0.1, x_0.3, y_-0.3, y_-0.1, y_0.1, y_0.3, no}`	`no`
World moves 1	If enabled, the camera rotates while compensating for world-space motion to maintain focus on the target. In this mode, local modifiers are not used.		`{truck_right_{amount}, truck_left_{amount}, pedestal_up_{amount}, pedestal_down_{amount}, goes_in_{amount}, goes_out_{amount}`	`none`
World moves 2	Similar to *World moves 1* but defines the world-space motion in the second half of the shot		`{truck_right_{amount}, truck_left_{amount}, pedestal_up_{amount}, pedestal_down_{amount}, goes_in_{amount}, goes_out_{amount}`	`none`

References

[1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025. 2

[2] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, and Di Zhang. ReCamMaster: Camera-controlled generative rendering from a single video. In *ICCV*, 2025. 2

[3] David Bordwell and Kristin Thompson. *Film Art: An Introduction*. McGraw-Hill Education, 12th edition, 2020. 4

[4] Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *CVPR*, 2025. 2

[5] Norman Christie, Jean-Marie Normand, and Patrick Olivier. Camera control in computer graphics. *Computer Graphics Forum*, 27(8):2197–2218, 2008. 4

[6] Robin Courant, Nicolas Dufour, Xi Wang, Marc Christie, and Vicky Kalogeiton. E.T. the exceptional trajectories: Text-to-camera-trajectory generation with character awareness. In *ECCV*, 2024. 1, 2, 3, 5, 6

[7] Wanquan Feng, Tianhao Qi, Jiawei Liu, Mingzhen Sun, Pengqi Tu, Tianxiang Ma, Fei Dai, Songtao Zhao, Siyu Zhou, and Qian He. I2VControl: Disentangled and unified video motion synthesis control. In *ICCV*, 2025. 2

[8] Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Carl Doersch, Yusuf Aytar, Michael Rubinstein, Chen Sun, Oliver Wang, Andrew Owens, and Deqing Sun. Motion prompting: Controlling video generation with motion trajectories. In *CVPR*, 2024. 2

[9] Ahmet Berke Gokmen, Yigit Ekin, Bahri Batuhan Bilecen, and Aysegul Dundar. RoPECraft: Training-free motion transfer with trajectory-guided rope optimization on diffusion transformers. In *Adv. Neural Inform. Process. Syst.*, 2025. 2

[10] Google DeepMind. Veo-2. `https://deepmind.google/technologies/veo/veo-2/`, 2025. Accessed October 7, 2025. 2

[11] Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. CameraCtrl II: Dynamic scene exploration via camera-controlled video diffusion models. *ArXiv preprint arXiv:2503.10592*, 2025. 1, 2

[12] Hongda Jiang, Xi Wang, Marc Christie, Libin Liu, and Baoquan Chen. Cinematographic camera diffusion model. In *EG*, 2024. 3, 6

[13] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-one video creation and editing. In *ICCV*, 2025. 1, 5, 3

[14] Wonjoon Jin, Qi Dai, Chong Luo, Seung-Hwan Baek, and Sunghyun Cho. FloVD: Optical flow meets video diffusion model for enhanced camera-controlled video synthesis. In *CVPR*, 2025. 2

[15] Steven D. Katz. *Film Directing: Shot by Shot – Visualizing from Concept to Screen*. Michael Wiese Productions, 1991. 4

[16] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 1

[17] Teng Li, Guangcong Zheng, Rui Jiang, Shuigen Zhan, Tao Wu, Yehao Lu, Yining Lin, Chuanyun Deng, Yepan Xiong, Min Chen, Lin Cheng, and Xi Li. RealCam-I2V: Real-world image-to-video generation with interactive complex camera control. In *ICCV*, 2025. 2

[18] Xinyang Li, Zhangyu Lai, Linning Xu, Yansong Qu, Liujuan Cao, Shengchuan Zhang, Bo Dai, and Rongrong Ji. Director3d: Real-world camera trajectory and 3d scene generation from text. In *NeurIPS*, 2024. 3, 6

[19] Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. VideoDirectorGPT: Consistent multi-scene video generation via llm-guided planning. In *COLM*, 2024. 2, 3

[20] Zhiqiu Lin, Siyuan Cen, Daniel Jiang, Jay Karhade, Hewei Wang, Chancharik Mitra, Yu Tong Tiffany Ling, Yuhan Huang, Sifan Liu, Mingyu Chen, Rushikesh Zawar, Xue Bai, Yilun Du, Chuang Gan, and Deva Ramanan. Towards understanding camera motions in any video. In *Adv. Neural Inform. Process. Syst.*, 2025. 3, 6, 8

[21] Matheus Lino and Norman Christie. Intuitive and efficient camera control with the toric space. *ACM Transactions on Graphics (SIGGRAPH)*, 34(4):82:1–82:12, 2015. 4

[22] Yawen Luo, Jianhong Bai, Xiaoyu Shi, Menghan Xia, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, and Tianfan Xue. CamCloneMaster: Enabling reference-based camera control for video generation. *ArXiv preprint arXiv:2506.03140*, 2025. 2

[23] OpenAI. Introducing GPT-5. https://openai.com/index/introducing-gpt-5/, 2025. Accessed: 2025-11-12. 7

[24] OpenAI. Sora: Generating videos from text, 2025. Accessed October 7, 2025. 2

[25] Zirui Pan, Xin Wang, Yipeng Zhang, Hong Chen, Kwan Man Cheng, Yaofei Wu, and Wenwu Zhu. Modular-Cam: Modular dynamic camera-view video generation with llm. In *AAAI*, 2025. 3

[26] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. GEN3C: 3d-informed world-consistent video generation with precise camera control. In *CVPR*, 2025. 2

[27] Quanjian Song, Zhihang Lin, Zhanpeng Zeng, Ziyue Zhang, Liujuan Cao, and Rongrong Ji. LightMotion: A light and tuning-free method for simulating camera motion in video generation. *ArXiv preprint arXiv:2503.06508*, 2025. 2

[28] Qwen Team. Qwen2.5-vl, 2025. 2

[29] 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. 1, 2, 6

[30] Qinghe Wang, Yawen Luo, Xiaoyu Shi, Xu Jia, Huchuan Lu, Tianfan Xue, Xintao Wang, Pengfei Wan, Di Zhang, and Kun Gai. CineMaster: A 3d-aware and controllable framework for cinematic text-to-video generation. In *SIGGRAPH*, 2025. 2

[31] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 6

[32] Zun Wang, Jaemin Cho, Jialu Li, Han Lin, Jaehong Yoon, Yue Zhang, and Mohit Bansal. EPiC: Efficient video camera control learning with precise anchor-video guidance. *ArXiv preprint arXiv:2505.21876*, 2025. 1, 2

[33] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *CVPR*, pages 4210–4220, 2023. 2

[34] 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. 1

[35] Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. TrajectoryCrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *ICCV*, 2025. 2

[36] Mengchen Zhang, Tong Wu, Jing Tan, Ziwei Liu, Gordon Wetzstein, and Dahua Lin. GenDoP: Auto-regressive camera trajectory generation as a director of photography. In *ICCV*, 2025. 1, 2, 3, 5, 6

[37] Sixiao Zheng, Zimian Peng, Yanpeng Zhou, Yi Zhu, Hang Xu, Xiangru Huang, and Yanwei Fu. VidCRAFT3: Camera, object, and lighting control for image-to-video generation. *ArXiv preprint arXiv:2502.07531*, 2025. 2

[38] Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. *ArXiv preprint arXiv:2503.14489*, 2025. 2