FreeOrbit4D: Training-Free Arbitrary Camera Redirection for Monocular Videos via Foreground-Complete 4D Reconstruction

Wei Cao[1] Hao Zhang[1] Fengrui Tian[2] Yulun Wu[1]
Yingying Li[1] Shenlong Wang[1] Ning Yu[3] Yaoyao Liu[1]
[1]University of Illinois Urbana-Champaign [2]University of Pennsylvania [3]Eyeline Labs
{weicao3, haoz19, yulun5, yl101, shenlong, lyy}@illinois.edu tianfr@upenn.edu ning.yu@scanlinevfx.com

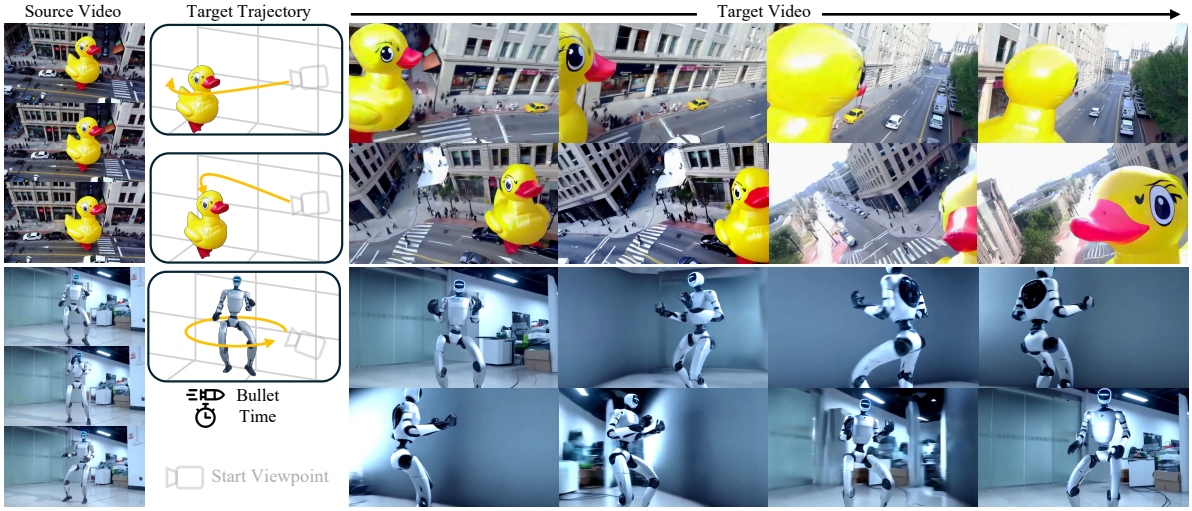

Figure 1. FreeOrbit4D enables **training-free** camera redirection from a single monocular video to arbitrary target trajectories. Given a source video and a target trajectory, our method produces redirected videos with faithful appearance and temporal coherence under large-angle camera motions, including bullet-time orbits.

Abstract

*Camera redirection aims to replay a dynamic scene from a single monocular video under a user-specified camera trajectory. Large-angle redirection is inherently ill-posed: a monocular video provides only partial observations of the underlying 4D world, and existing methods break down under large viewpoint changes due to missing visual grounding. We present **FreeOrbit4D**, a training-free framework that recovers a foreground-complete 4D proxy as structural grounding for video generation. We decouple foreground and background reconstructions: the monocular video is unprojected into a global scene space, while an object-centric multi-view diffusion model completes occluded foreground geometry in canonical object space. Dense pixel-synchronized 3D–3D correspondences align these representations into a unified proxy, whose depth scaffolds condition a video diffusion model for faithful novel-view synthesis. Extensive experiments show that FreeOrbit4D produces more faithful and temporally coherent redirected videos under challenging large-angle trajectories. **Project page:** https://freeorbit4d.vision.ischool.illinois.edu/*

1. Introduction

Camera redirection synthesizes novel video sequences from a source video along a user-specified camera trajectory [2, 7, 9, 33], with applications in autonomous driving [5, 36, 37], AR/VR [1], and bullet-time effects [6, 25]. However, this task is inherently ill-posed: a monocular video provides only partial observations of the underlying 4D world, and recovering complete geometry from this limited input remains challenging. Recent methods follow two paradigms: *implicit control* [2, 7, 9] encodes trajectories as learned embeddings but offers only soft controllability; *explicit warp-*

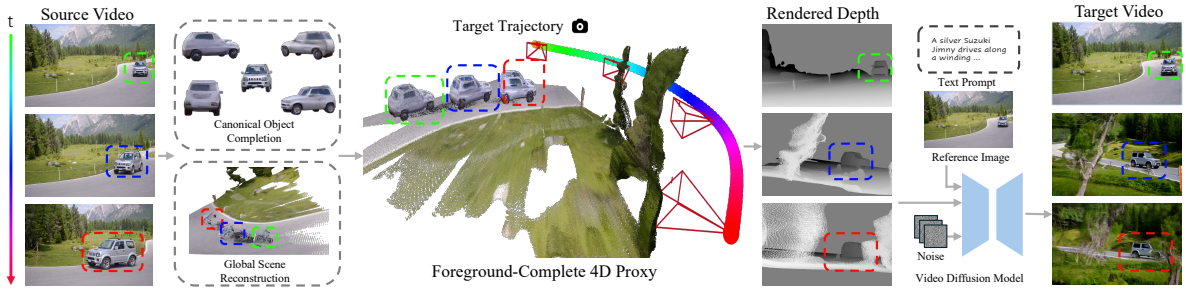

Figure 2. **Overview of FreeOrbit4D.** Our framework constructs a foreground-complete 4D proxy through two branches: *Global Scene Reconstruction* recovers background and partial foreground, while *Canonical Object Completion* reconstructs complete foreground via multi-view synthesis. After alignment, depth maps rendered from the proxy condition a video diffusion model for novel-view synthesis.

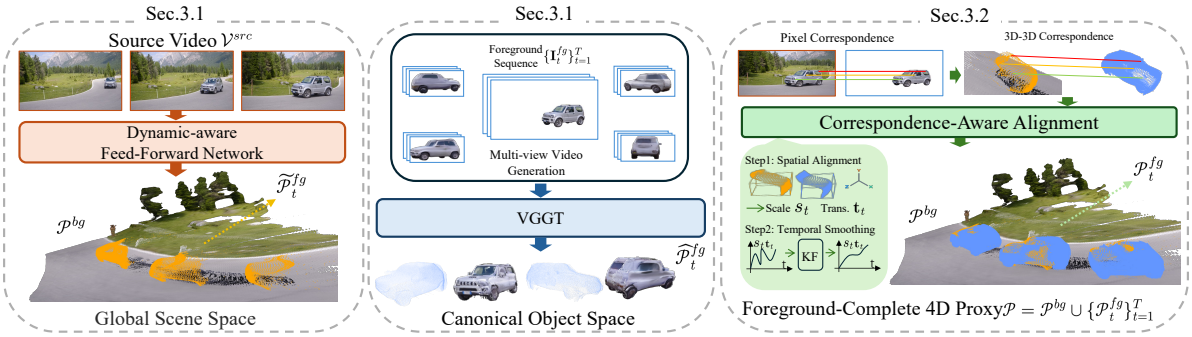

Figure 3. **Decoupled 4D reconstruction and alignment pipeline. Left:** A dynamic-aware feed-forward network lifts $\mathcal{V}^{src}$ into *global scene space*, producing static background $\mathcal{P}^{bg}$ and geometrically incomplete foreground $\widetilde{\mathcal{P}}_t^{fg}$. **Middle:** An object-centric video diffusion model synthesizes multi-view images, from which VGGT reconstructs complete foreground geometry $\widehat{\mathcal{P}}_t^{fg}$ in *canonical object space*. **Right:** Dense 3D-3D correspondences enable per-frame spatial alignment $(s_t, \mathbf{t}_t)$ followed by Kalman filtering, yielding the unified foreground-complete 4D proxy $\mathcal{P} = \mathcal{P}^{bg} \cup \{\mathcal{P}_t^{fg}\}_{t=1}^T$.

ing [10, 23, 32] warps observed pixels via depth but leaves occluded regions unfilled. Neither achieves both precise control and complete visibility for large-angle redirection.

We propose FreeOrbit4D, a training-free framework for arbitrary camera redirection via foreground-complete 4D reconstruction. We decouple the problem into global scene lifting (in world space) and foreground geometry completion (in canonical object space), then unify them through dense pixel-synchronized 3D–3D correspondences. The resulting 4D proxy is distilled into depth maps that condition a video diffusion model [28] for faithful large-angle view synthesis. Our contributions:

- A training-free method for foreground-complete 4D reconstruction via global scene lifting, object completion, and 3D correspondences.
- A camera redirection framework for large-angle view synthesis with strong spatio-temporal consistency.
- State-of-the-art results validated by experiments and a user study; the 4D proxy further enables edit propagation and data generation.

2. Related Work

Camera-controlled video generation. Recent methods encode camera trajectories as learned embeddings [2, 7, 9, 17] or geometric representations such as Plücker coordinates [7, 34] and PRope [11, 16] to control video diffusion models [3, 15]. However, these implicit approaches offer only soft controllability and often fail to follow prescribed trajectories due to the absence of explicit geometric constraints.

Reconstruction-grounded 4D generation. Another line of work [4, 23, 24, 26, 32, 33, 35] reconstructs 4D representations from monocular video via NeRF [18, 25] or Gaussian Splatting [14], renders partial views, and relies on diffusion models to complete missing regions. However, the underlying geometry is limited to observed surfaces, making large-angle redirection unreliable. Our method addresses this by constructing a foreground-complete 4D proxy via correspondence-aware alignment.

Table 1. **Quantitative comparison and user study.** VBench for perceptual quality, DINO/CLIP-SIM for similarity, FID-V/FVD-V for distributional fidelity, and user ratings (1–5). **Bold**: best; underline: second-best.

Method	VBench ↑						Similarity & Fidelity				User Study		
	Subject Consis.	BG Consis.	Motion Smooth.	Overall Consis.	Aesth. Qual.	Imaging Qual.	DINO-SIM (↑)	CLIP-SIM (↑)	FID-V (↓ $\times 10^2$)	FVD-V (↓ $\times 10^3$)	Overall (↑)	Motion (↑)	Stab. (↑)
ReCamMaster	0.84	0.92	**0.98**	0.16	0.39	43	0.37	0.75	2.6	3.9	2.0	2.5	2.0
TrajectoryCrafter	0.80	0.91	0.94	0.19	0.47	53	0.47	0.79	2.0	3.6	2.8	3.2	2.9
EX-4D	0.76	0.89	0.94	0.16	0.42	46	0.28	0.69	3.2	3.8	2.0	2.5	2.0
GEN3C	0.79	0.88	0.95	0.18	0.42	49	0.43	0.75	2.3	**3.3**	2.4	3.5	2.3
Ours	**0.88**	**0.94**	0.96	**0.24**	**0.52**	**64**	**0.65**	**0.84**	**1.7**	3.6	**4.6**	**4.5**	**4.5**

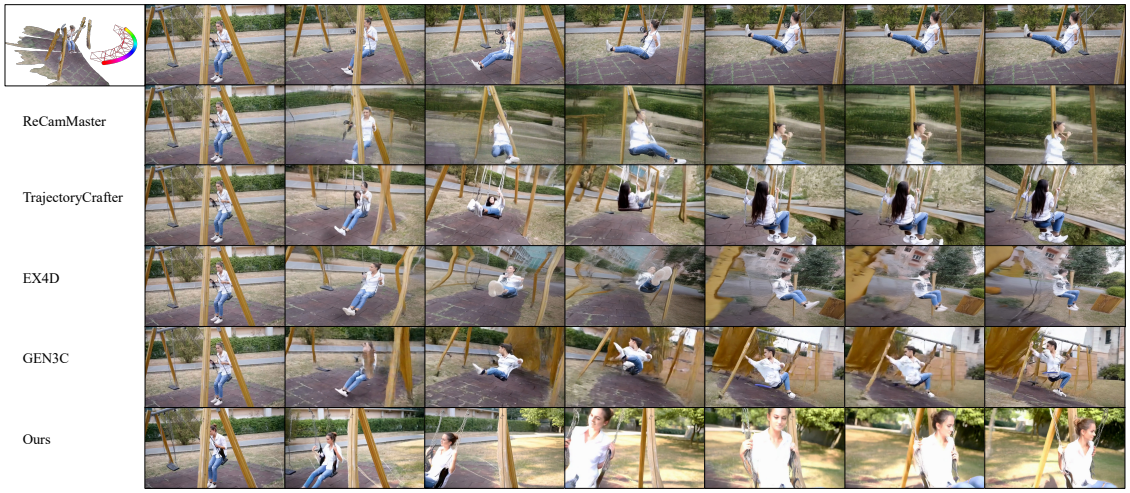

Figure 4. **Qualitative comparison on "Swing".** The top-left inset shows our reconstructed 4D proxy and target trajectory. Existing methods exhibit structural disintegration [2, 10] or geometric warping and semantic drift [23, 32]. Our method generates sharp details and stable geometry by anchoring pixels to the foreground-complete proxy.

Figure 5. **Applications. Top:** Appearance editing—a single edited reference frame (e.g., zebra pattern or anime style) is propagated consistently across all novel viewpoints via our foreground-complete proxy. **Bottom:** Geometry editing—manipulating the point cloud (scaling or compositing) produces plausible redirected videos from modified 4D geometry.

3. Method

Given a monocular source video $\mathcal{V}^{src}$ and a target camera trajectory, our goal is to synthesize a visually faithful and temporally consistent target video $\mathcal{V}^{tgt}$ from the target viewpoints. As shown in Fig. 2, our framework has three stages: decoupled 4D reconstruction (Sec. 3.1), correspondence-aware alignment (Sec. 3.2), and geometry-conditioned video synthesis (Sec. 3.3).

3.1. Decoupled 4D Reconstruction

Since static scenes and moving objects exhibit fundamentally different geometric characteristics, we decouple their reconstructions into two complementary coordinate spaces.

Global Scene Reconstruction. We adopt a temporally-aware feed-forward model [29, 38] that processes $\mathcal{V}^{src}$ and predicts temporally consistent point maps $\widetilde{\mathbf{P}}_t \in \mathbb{R}^{H\times W\times 3}$ in a unified global coordinate system. Using semantic masks $\mathbf{M}_t$ from SAM2 [22], we separate background and

foreground:

$$\mathcal{P}^{bg} = \bigcup_{t=1}^{T} \{\widetilde{\mathbf{P}}_t(\mathbf{u}) \mid \mathbf{M}_t(\mathbf{u}) = 0\},$$
$$\widetilde{\mathcal{P}}_t^{fg} = \{\widetilde{\mathbf{P}}_t(\mathbf{u}) \mid \mathbf{M}_t(\mathbf{u}) = 1\}, \tag{1}$$

where $\widetilde{\mathcal{P}}_t^{fg}$ captures only visible foreground surfaces from the source viewpoint.

Canonical Object Completion. To complete occluded foreground geometry, we extract the masked foreground sequence $\{\mathbf{I}_t^{fg}\}_{t=1}^T$ and feed it into a multi-view video diffusion model [31], which synthesizes four novel-view videos at 90° azimuthal intervals $\{\mathbf{I}_t^{(k)}\}_{t=1}^T$ $(k = 1, \ldots, 4)$. VGGT [29] then reconstructs multi-view point maps from all five views per frame:

$$\widehat{\mathbf{P}}_t^{fg} = \Phi_{\text{VGGT}}\left(\mathbf{I}_t^{fg}, \mathbf{I}_t^{(1)}, \ldots, \mathbf{I}_t^{(4)}\right) \in \mathbb{R}^{5\times H\times W\times 3}. \tag{2}$$

After filtering background points using SAM2 masks and color thresholding, we obtain the complete canonical foreground:

$$\widehat{\mathcal{P}}_t^{fg} = \{\widehat{\mathbf{P}}_t(\mathbf{u}) \mid \mathbf{M}_t(\mathbf{u}) = 1\} \cup \bigcup_{k=1}^{4} \{\widehat{\mathbf{P}}_t^{(k)}(\mathbf{u}) \mid \mathbf{M}_t^{(k)}(\mathbf{u}) = 1\}. \tag{3}$$

3.2. Correspondence-Aware Alignment

We now align $\widehat{\mathcal{P}}_t^{fg}$ (canonical space) to $\widetilde{\mathcal{P}}_t^{fg}$ (global space). Since both $\widetilde{\mathbf{P}}_t$ and $\widehat{\mathbf{P}}_t$ originate from the same source image $\mathbf{I}_t$, pixels at the same coordinate $\mathbf{u}$ correspond to the same surface point, yielding dense 3D–3D correspondences:

$$\mathcal{C}_t = \{(\widehat{\mathbf{P}}_t(\mathbf{u}), \widetilde{\mathbf{P}}_t(\mathbf{u})) \mid \mathbf{M}_t(\mathbf{u}) = 1\}. \tag{4}$$

Monocular lifting cannot determine absolute depth scale, which may vary across frames, causing per-point inconsistencies in $\widetilde{\mathcal{P}}_t^{fg}$. We therefore use $\widetilde{\mathcal{P}}_t^{fg}$ only to determine the global placement (position and scale) while preserving the geometry of $\widehat{\mathcal{P}}_t^{fg}$: $\mathcal{P}_t^{fg} = s_t\,\widehat{\mathcal{P}}_t^{fg} + \mathbf{t}_t$, where $(s_t, \mathbf{t}_t)$ are estimated from $\mathcal{C}_t$. To compensate for frame-to-frame depth inconsistency, we smooth the centroid trajectory using a bidirectional Kalman filter with a constant-velocity motion model, yielding the unified foreground-complete 4D proxy $\mathcal{P} = \mathcal{P}^{bg} \cup \{\mathcal{P}_t^{fg}\}_{t=1}^T$.

3.3. Geometry-conditioned Video Synthesis

Given the 4D proxy $\mathcal{P}$ and a target trajectory $\{\boldsymbol{\pi}_t^{tgt}\}_{t=1}^T$, we render depth scaffolds and synthesize the output:

$$\mathcal{V}^{tgt} = \Phi_{\text{VDM}}\left(\mathbf{I}_1,\ \left\{\text{Render}\left(\mathcal{P},\ \boldsymbol{\pi}_t^{tgt}\right)\right\}_{t=1}^{T},\ \mathbf{c}\right), \tag{5}$$

where Φ_{VDM} is a depth-conditioned video diffusion model [13, 28], $\mathbf{I}_1$ is the first source frame as appearance reference, and $\mathbf{c}$ is a text prompt. Depth maps encode the 3D layout from the target viewpoint, enforcing cross-view and temporal consistency.

Table 2. **Ablation study.** We progressively add multi-view generation (MVG) and Kalman filter smoothing (KF) to evaluate each component's contribution. **Bold**: best.

Method	DINO-SIM (↑)	CLIP-SIM (↑)	FID-V (↓ $\times 10^2$)	FVD-V (↓ $\times 10^3$)
Baseline	0.58	0.81	1.9	4.1
+ MVG	0.60	0.82	1.9	4.1
+ KF (Full)	**0.65**	**0.84**	**1.7**	**3.6**

4. Experiments

Setup. Our pipeline uses PAGE-4D [38], SAM2 [22], SV4D2.0 [31], VGGT [29], and Wan2.2-VACE [13, 28] on a single A40 GPU. We evaluate on DAVIS [20] and online videos with 120°–180° target rotations, reporting VBench [12], DINO/CLIP-SIM [19, 21], FID/FVD-V [8, 27], and a 20-participant user study. We compare with ReCamMaster [2], TrajectoryCrafter [32], EX-4D [10], and GEN3C [23].

Results. As shown in Fig. 4, baselines suffer from structural disintegration or geometric drift under large viewpoint changes. Table 1 shows our method ranks first on 5/6 VBench dimensions, with the best DINO-SIM (0.65), CLIP-SIM (0.84), and FID-V (1.7×10^2). In the user study, we outperform all baselines across all axes, with particularly large margins in motion accuracy (4.5 vs. 3.5).

4.1. Applications

Our explicit 4D representation enables two applications beyond camera redirection (Fig. 5). **Appearance propagation**: edits to a single reference frame (e.g., color, style transfer via [30]) propagate to all viewpoints via consistent depth scaffolds. **4D geometry manipulation**: directly modifying the point cloud (scaling, compositing) yields plausible redirected videos.

4.2. Ablation Study

Tab. 2 evaluates each component. Removing multi-view generation (MVG) degrades all metrics, showing geometry completion is critical. Disabling Kalman filtering (KF) further reduces performance, validating temporal smoothing.

5. Conclusion

We present **FreeOrbit4D**, a training-free framework for camera redirection from monocular video. Built on a foreground-complete 4D proxy as geometric scaffolds, our method achieves state-of-the-art fidelity, temporal coherence, and camera control under large viewpoint changes, and further enables applications like appearance propagation and scene manipulation.

Acknowledgements

This research is supported by the National Artificial Intelligence Research Resource (NAIRR) Pilot under award NAIRR250199. Computational resources are also provided by Delta and DeltaAI at the National Center for Supercomputing Applications (NCSA) through ACCESS allocations CIS250012, CIS250816, and CIS251188. S. W. is also supported by NSF Awards #2525287, #2404385, #2414227, #2340254, #2312102, and #2331878, and research grants from IBM, Meta, NVIDIA, and Intel.

References

[1] Ronald T Azuma. A survey of augmented reality. *Presence: teleoperators & virtual environments*, 6(4):355–385, 1997. 1

[2] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. In *ICCV*, 2025. 1, 2, 3, 4

[3] 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. 2

[4] Wei Cao, Chang Luo, Biao Zhang, Matthias Nießner, and Jiapeng Tang. Motion2vecsets: 4D latent vector set diffusion for non-rigid shape reconstruction and tracking. In *CVPR*, 2024. 2

[5] Wei Cao, Marcel Hallgarten, Tianyu Li, Daniel Dauner, Xunjiang Gu, Caojun Wang, Yakov Miron, Marco Aiello, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. Pseudo-simulation for autonomous driving. In *CoRL*, 2025. 1

[6] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *ICCV*, 2021. 1

[7] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. In *ICLR*, 2025. 1, 2

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 4

[9] Chen Hou and Zhibo Chen. Training-free camera control for video generation. In *ICLR*, 2025. 1, 2

[10] Tao Hu, Haoyang Peng, Xiao Liu, and Yuewen Ma. Ex-4d: Extreme viewpoint 4d video synthesis via depth watertight mesh. *arXiv preprint arXiv:2506.05554*, 2025. 2, 3, 4

[11] Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson WH Lau, Wangmeng Zuo, and Chunchao Guo. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *arXiv preprint arXiv:2506.04225*, 2025. 2

[12] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024. 4

[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. 4

[14] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *TOG*, 2023. 2

[15] 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. 2

[16] Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. In *NeurIPS*, 2025. 2

[17] 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

[18] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis, 2020. 2

[19] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4

[20] 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. 4

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 4

[22] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. In *ICLR*, 2025. 3, 4

[23] 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, 3, 4

[24] Jiapeng Tang, Wei Cao, Biao Zhang, Chang Luo, Yaoyao Liu, and Matthias Nießner. Motion2VecSets: Non-rigid shape reconstruction and tracking with 4D latent set diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. 2

[25] Fengrui Tian, Shaoyi Du, and Yueqi Duan. MonoNeRF: Learning a generalizable dynamic radiance field from monocular videos. In *ICCV*, 2023. 1, 2

[26] Fengrui Tian, Tianjiao Ding, Jinqi Luo, Hancheng Min, and René Vidal. Voyaging into perpetual dynamic scenes from a single view. In *ICCV*, 2025. 2

[27] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. In *ICLR*, 2019. 4

[28] 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. 2, 4

[29] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025. 3, 4

[30] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 4

[31] Chun-Han Yao, Yiming Xie, Vikram Voleti, Huaizu Jiang, and Varun Jampani. SV4D2.0: Enhancing spatio-temporal consistency in multi-view video diffusion for high-quality 4d generation. In *ICCV*, 2025. 4

[32] Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *ICCV*, 2025. 2, 3, 4

[33] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 1, 2

[34] Songchun Zhang, Huiyao Xu, Sitong Guo, Zhongwei Xie, Pengwei Liu, Hujun Bao, Weiwei Xu, and Changqing Zou. Spatialcrafter: Unleashing the imagination of video diffusion models for scene reconstruction from limited observations. In *ICCV*, 2025. 2

[35] Yanran Zhang, Ziyi Wang, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Joint 3d geometry reconstruction and motion generation for 4d synthesis from a single image. *arXiv preprint arXiv:2512.05044*, 2025. 2

[36] Hongkuan Zhou, Wei Cao, Aifen Sui, and Zhenshan Bing. What matters to enhance traffic rule compliance of imitation learning for end-to-end autonomous driving. *arXiv preprint arXiv:2309.07808*, 2023. 1

[37] Hongkuan Zhou, Stefan Schimid, Yicong Li, Lavdim Halilaj, Xiangtong Yao, and Wei Cao. Predicting the road ahead: A knowledge graph based foundation model for scene understanding in autonomous driving. In *European Semantic Web Conference*, 2025. 1

[38] Kaichen Zhou, Yuhan Wang, Grace Chen, Xinhai Chang, Gaspard Beaudouin, Fangneng Zhan, Paul Pu Liang, and Mengyu Wang. Page-4d: Disentangled pose and geometry estimation for 4d perception. *arXiv preprint arXiv:2510.17568*, 2025. 3, 4