SteerX: Creating Any Camera-Free 3D and 4D Scenes with Geometric Steering

Byeongjun Park[1,2*] Hyojun Go[2*] Hyelin Nam[2] Byung-Hoon Kim[2,3]
Hyungjin Chung[2†] Changick Kim[1†]

[1] KAIST [2] EverEx [3] Yonsei University

https://byeongjun-park.github.io/SteerX/

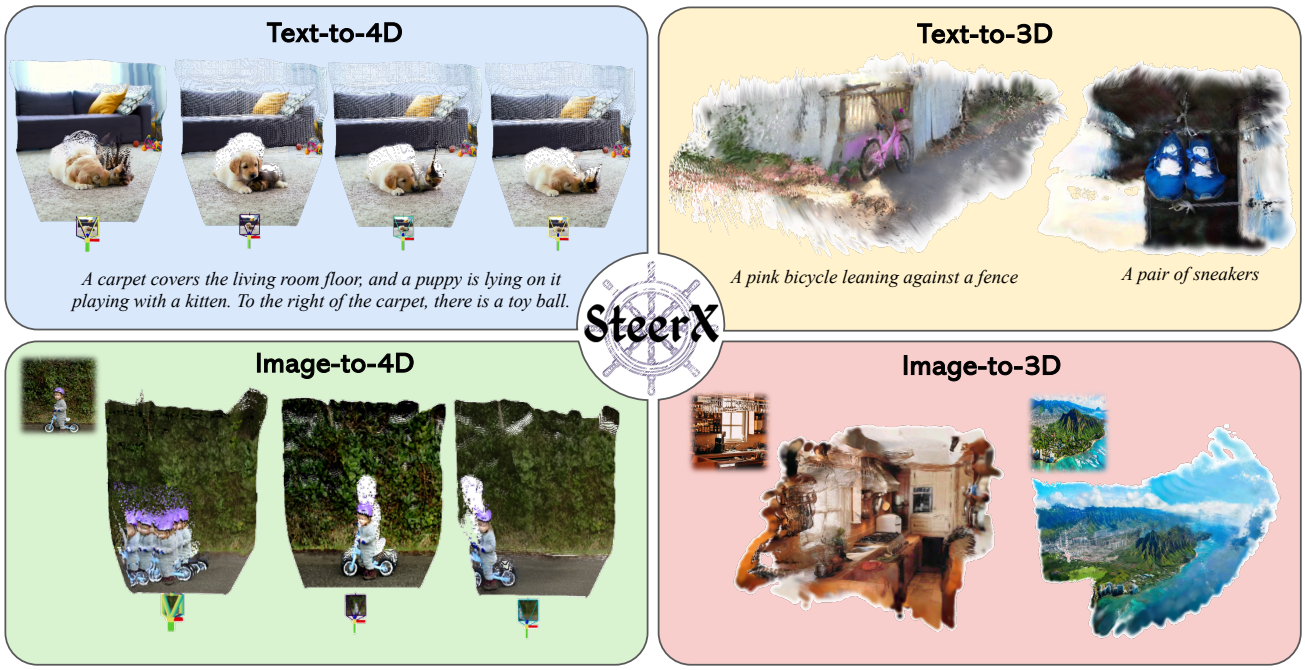


Figure 1. **SteerX** is a zero-shot inference-time steering approach that seamlessly integrates video generative models [12, 16, 21, 23, 29] and feed-forward scene reconstruction models [12, 22, 31], enabling any 3D and 4D scene generation without explicit camera conditions.

Abstract

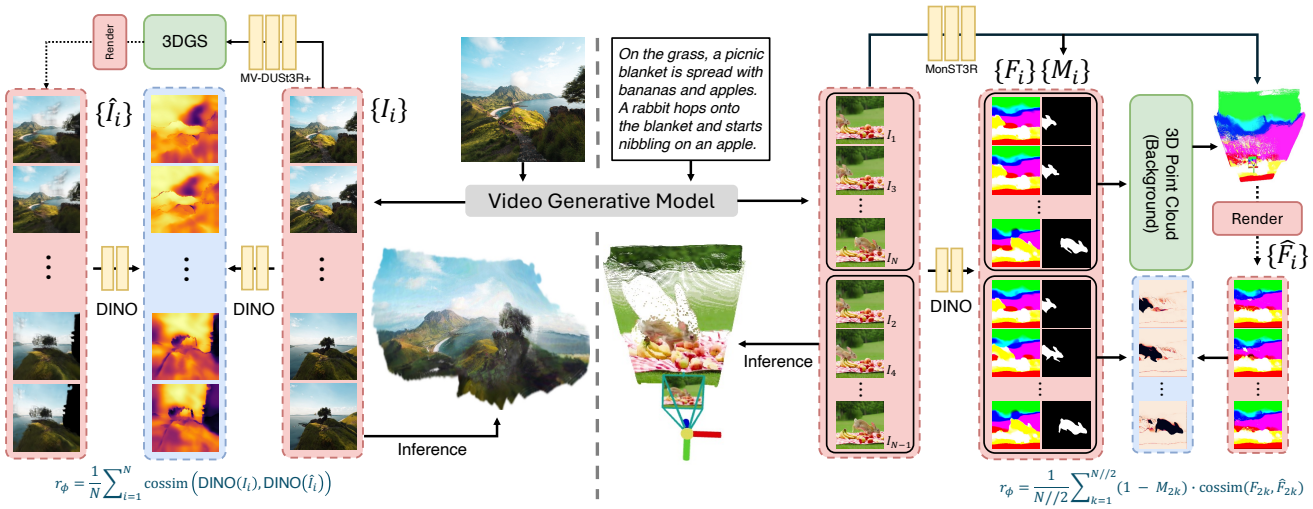
Recent progress in 3D/4D scene generation emphasizes the importance of physical alignment throughout video generation and scene reconstruction. However, existing methods improve the alignment separately at each stage, making it difficult to manage subtle misalignments arising from another stage. Here, we present SteerX, a zero-shot inference-time steering method that unifies scene reconstruction into the generation process, tilting data distributions toward better geometric alignment. To this end, we introduce two geometric reward functions for 3D/4D scene generation by using pose-free feed-forward scene reconstruction models. Through extensive experiments, we demonstrate the effectiveness of SteerX in improving 3D/4D scene generation.

1. Introduction

Generating 3D and 4D scenes from images or text prompts has attracted significant attention due to its potential applications in AR/VR and robotics [6, 24, 28]. This progress is largely driven by the advancement of generative models [4, 13, 16, 23, 29] and neural scene representations [15, 19, 26]. Generative models learn the underlying distribution of large-scale video data, and neural scene representations lift these distributions into structured 3D or 4D spaces.

To generate geometrically aligned 3D and 4D scenes, previous works handle physical alignment separately in either video generation [2, 3, 11, 14, 21, 25, 27, 32] or scene reconstruction [12, 17]. This makes it difficult to address cross-stage misalignments, as inconsistencies in one stage may not be fully corrected in the other. We observe that achieving precise alignment remains an ongoing challenge due to the indistinct link between the two stages.

*Equal contribution, †Corresponding author



(a) GS-MEt3R (b) Dyn-MEt3R

Figure 2. **An overview of geometric rewards.** Our reward functions assess the geometric consistency of intermediate generated video frames by computing the feature similarity of upscaled DINO features. (a) GS-MEt3R evaluates feature similarity between the original video frames and their corresponding rendered images from 3DGS. (b) Dyn-MEt3R focuses on background regions by unprojecting background features from half of the video frames and reprojecting them onto the remaining frames to compute feature similarity.

In this work, we introduce SteerX, a zero-shot inference-time steering method that seamlessly integrates video generation and scene reconstruction, generating geometrically aligned high-quality 3D and 4D scenes. To this end, we propose two geometric reward functions tailored for 3D and 4D scene generation. They evaluate geometric consistency across multiple video frames by incorporating advanced pose-free feed-forward scene reconstruction methods such as MV-DUSt3R+ [22] and MonST3R [31]. These reconstruction methods lift intermediate generated video frames during the reverse sampling process into 3D and 4D spaces. The reconstructed scenes are then projected back into the original image space for consistency evaluation.

To guide the generation process toward geometrically plausible outputs, we formulate a steering algorithm based on sequential Monte Carlo (SMC) [8]. Built on SMC-based guided sampling, SteerX offers a generalizable framework that can pair *any* generative video model with *any* 3D reconstruction method, enabling diverse generation tasks, including Image-to-3D, Image-to-4D, Text-to-3D, and Text-to-4D. Through extensive experiments with various video generative models [12, 16, 21, 29], we demonstrate the effectiveness and broad applicability of our approach.

2. Methods

Here, we present SteerX, which unifies feed-forward scene reconstruction models into the video generation process, iteratively tilting data distributions towards geometrically aligned samples. In Section 2.1, we define two geometric reward functions to evaluate geometric consistency in generated multi-view images and dynamic videos, respectively. In Section 2.2, we detail our SMC-based steering algorithm.

2.1. Geometric Rewards

Our geometric rewards build upon MEt3R [1], which measures feature similarity in overlapping regions between image pairs. It evaluates multiple images by averaging feature similarity across all image pairs, but the cost grows quadratically. To address this, as shown in Fig. 2, we introduce two geometric reward functions, GS-MEt3R and Dyn-MEt3R, which measure global geometric consistency across multiple video frames and mitigate computational bottlenecks.

3D Scene Reward. Recent methods [22, 30] introduce a mapping function f_ϕ that directly reconstructs 3DGS and camera poses from N views $\{I_i \in \mathbb{R}^{H\times W\times 3}\}_{i=1}^N$ as:

$$f_\phi : \{I_i\}_{i=1}^N \rightarrow \begin{cases} \{\boldsymbol{\mu}_j, \mathbf{o}_j, \boldsymbol{\Sigma}_j, \mathbf{c}_j\}_{j=1}^{N\times H\times W} \\ \{P_i\}_{i=1}^N \end{cases}, \quad (1)$$

where the 3D scene is represented as Gaussian parameters, including position $\boldsymbol{\mu}$, volume density $\mathbf{o}$, covariance $\boldsymbol{\Sigma}$, and color $\mathbf{c}$. Then, we produce images $\{\hat{I}_i\}_{i=1}^N$ by rendering the scene with estimated camera poses $\{P_i\}_{i=1}^N$, ensuring they correspond to the same viewpoints as the input images. Finally, GS-MEt3R is measured by computing the cosine similarity between upscaled DINO [5, 10] features of input images $\{F_i\}_{i=1}^N$ and rendered images $\{\hat{F}_i\}_{i=1}^N$ as:

$$r_\phi = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{H}\sum_{k=1}^{W} \frac{F_i^{jk}\cdot \hat{F}_i^{jk}}{||F_i^{jk}||\cdot||\hat{F}_i^{jk}||}. \quad (2)$$

4D Scene Reward. While the 3D scene reward function is based on 3DGS, feed-forward dynamic scene reconstruction with Gaussians remains underexplored, making it diffi-

Algorithm 1 SteerX (v-prediction)

Required: $\mathbf{v}$-parametrized diffusion model $\mathbf{v}_\theta$, reward function r_ϕ, number of particles k, and initial noise $\{\mathbf{x}_T^i\}_{i=1}^k \sim \mathcal{N}(0, I)$.

Sampling:

for $t \in \{T-1, \dots, 0\}$ **do**
  for $i \in \{1 \dots k\}$ **do**
    $\hat{\mathbf{x}}_0^i \leftarrow \sqrt{\bar{\alpha}_{t+1}}\mathbf{x}_{t+1}^i - \sqrt{1-\bar{\alpha}_{t+1}}\mathbf{v}_\theta(\mathbf{x}_{t+1}^i)$
    $\mathbf{x}_t^i \leftarrow$ *dpm-solver*$(\hat{\mathbf{x}}_0^i, \mathbf{x}_{t+1}^i)$
    $\mathbf{s}_t^i \leftarrow r_\phi(\hat{\mathbf{x}}_0^i)$ ▷ *Intermediate rewards*
    $G_t^i \leftarrow \exp(\lambda \max_{j=t}^T(\mathbf{s}_j^i))$ ▷ *Potential*
  end for
  $\{\mathbf{x}_t^i\}_{i=1}^k \sim$ Multinomial$(\{\mathbf{x}_t^i, G_t^i\}_{i=1}^k)$ ▷ *Resample*
end for
$l \leftarrow \arg\max_{i \in \{1,\dots,k\}} r_\phi(\mathbf{x}_0^i)$
return $\mathbf{x}_0^l$

cult to directly apply 3DGS-based rewards. Instead, we employ 3D point cloud representations, where MonST3R [31] reconstructs it with point maps $\{X_i\}_{i=1}^N$, binary dynamic masks $\{M_i\}_{i=1}^N$, and camera poses $\{P_i\}_{i=1}^N$ as:

$$f_\phi : \{I_i\}_{i=1}^N \to \{X_i, M_i, P_i\}_{i=1}^N, \tag{3}$$

where we leverage these time-varying point clouds as 4D scene representations and design a reward function for evaluating geometric consistency in dynamic videos.

Since dynamic masks are produced in the camera pose estimation process to retain only high-confidence points, a well-reconstructed 4D scene should effectively filter out dynamic objects while preserving geometric consistency in the background regions. Therefore, we evaluate the consistency only for background regions of video frames, which are not filtered out by the dynamic mask. To this end, we first split N video frames into two subsets: $\mathcal{I}_{src} = \{I_1, I_3, \dots, I_N\}$ and $\mathcal{I}_{tgt} = \{I_2, I_4, \dots, I_{N-1}\}$. Then, we unproject the upsampled DINO features of background regions in $\mathcal{I}_{src}$ into 3D space using MonST3R. Finally, we reproject them onto the viewpoint of $\mathcal{I}_{tgt}$, where the rendered features $\hat{\mathcal{F}}_{tgt} = \{\hat{F}_1, \hat{F}_3, \dots, \hat{F}_N\}$ are used to compute the feature similarity with background regions in $\mathcal{I}_{tgt}$ as:

$$r_i = \sum_{j=1}^{H}\sum_{k=1}^{W}(1 - M_i^{jk})\frac{F_i^{jk} \cdot \hat{F}_i^{jk}}{||F_i^{jk}|| \cdot ||\hat{F}_i^{jk}||}, \tag{4}$$

$$r_\phi = \frac{1}{(N//2)}\sum_{i=1}^{N//2} r_i. \tag{5}$$

2.2. Geometric Steering

By using the rewards defined in Section 2.1 and target distribution $\tilde{p}_\theta$, SMC operates with the three following steps:

1. (**Proposal**) For each particle i, sample from the proposal distribution $\mathbf{x}_t^i \sim q_t(\mathbf{x}_t|\mathbf{x}_{t+1}^i)$
2. (**Weighting**) Compute weights from reward-based potentials $\omega_t^i = \frac{p_\theta(\mathbf{x}_t^i|\mathbf{x}_{t+1}^i)}{q_t(\mathbf{x}_t^i|\mathbf{x}_{t+1}^i)} G_t(\mathbf{x}_{T:t}^i)$
3. (**Resampling**) Draw new particles from the multinomial distribution $\{\mathbf{x}_t^i\}_{i=1}^k \sim$ Multinomial$(\{\mathbf{x}_t^i, G_t^i\}_{i=1}^k)$

Two choices should be made: the potential G_t, and the proposal distribution q_t. For the potential, we use max potential

$$G_t(\mathbf{x}_{T:t})^i := \exp\left(\lambda \max_{j=t}^{T}[r_\phi(\hat{\mathbf{x}}_0)]\right), \tag{6}$$

with

$$G_0(\mathbf{x}_{T:0}) := \exp\left(\lambda r_\phi(\mathbf{x}_0)\right)\left(\prod_{t=1}^{T} G_t(\mathbf{x}_{T:t})\right)^{-1}, \tag{7}$$

such that the particle with the highest reward is preferred. Notice that we use the Tweedie estimate $\hat{\mathbf{x}}_0 = \mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]$ [7, 9, 20] in intermediate steps to avoid full reverse sampling. For the proposal kernel, to save computation, we leverage DPM-solver++ [18], which approximates the true sampling trajectory limited to small neural function evaluation (NFE). These choices lead to SteerX, as shown in Alg. 1.

3. Experimental Results

In this section, we conduct extensive experiments to verify the scalability and effectiveness of SteerX across various video generative models in four scene generation scenarios: Text-to-4D, Image-to-4D, Text-to-3D, and Image-to-3D.

3.1. Experimental Setup

We evaluate SteerX with $k = 4$ particles and utilize HunyuanVideo [16], CogVideoX [29], and DimensionX [21] for Text-to-4D, Image-to-4D, and Image-to-3D, respectively. For Text-to-3D, we utilize the video generation and scene reconstruction models proposed in SplatFlow [12]. We compare against the best-of-N approach, where the particles are generated independently, and the one with the highest reward is selected. We include a baseline ($k = 1$) that generates a video and directly reconstructs the scene.

3.2. Main Results

Qualitative results. Figures 3 to 6 illustrate the examples of various generation tasks for the baseline, best-of-N (BoN), and our SteerX. While the BoN approach often fails to maintain 3D consistency and tends to struggle with generalizing to diverse text prompts, our SteerX effectively captures object motions, camera movements, and generates realistic 3D/4D scenes. This highlights the effectiveness and scalability of SteerX, ensuring that generated video frames are optimally structured for precise scene reconstruction.

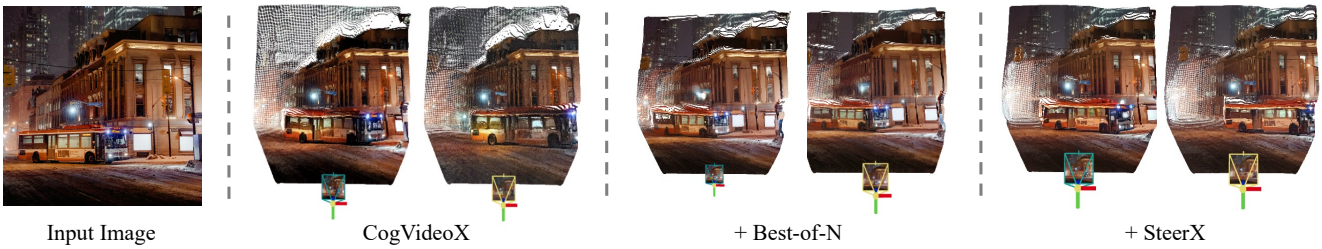


Figure 3. **Qualitative results in Image-to-4D.** SteerX naturally lifts object motion into 4D spaces, while preserving geometric alignments.

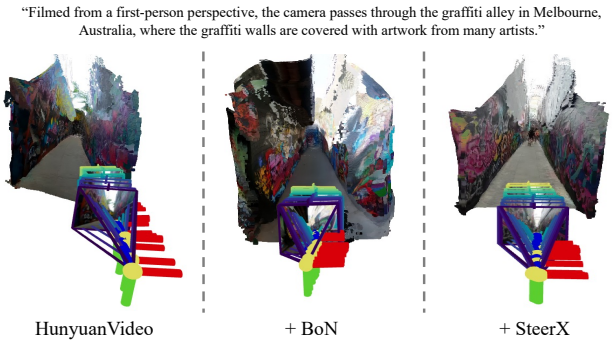


Figure 4. **Qualitative results in Text-to-4D.**

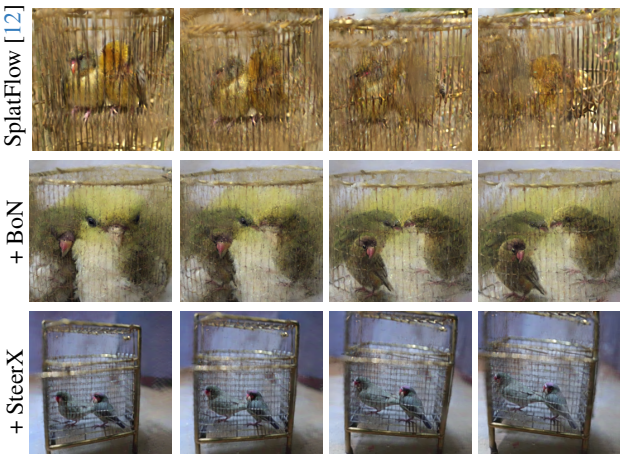


Figure 5. **Qualitative results in Text-to-3D.** SteerX improves the visual quality and textual alignment, verifying its compatibility.

Figure 6. **Qualitative results in Image-to-3D.**

Quantitative results. Table 1 shows that our SteerX outperforms the baseline and BoN approach for all generation scenarios. Notably, SteerX can be applied to the multi-view rectified flow model and the 3DGS decoder of SplatFlow [12] via GS-MEt3R, showing its applicability across any video generation and scene reconstruction models.

Method (Text-to-4D)	Aesthetic↑	Subject↑	Temporal↑	Dyn-MEt3R↑
HunyuanVideo [16]	0.549	0.967	0.241	0.911
+ Best-of-N	0.551	0.978	0.239	0.931
+ SteerX	**0.555**	**0.980**	**0.243**	**0.964**
Method (Image-to-4D)	Aesthetic↑	Subject↑	Dynamic↑	Dyn-MEt3R↑
CogVideoX [29]	0.592	0.945	0.158	0.880
+ Best-of-N	0.594	0.944	0.143	0.901
+ SteerX	**0.596**	**0.957**	**0.170**	**0.909**
Method (Text-to-3D)	BRISQUE↓	NIQE↓	CLIPScore↑	GS-MEt3R↑
SplatFlow [12]	23.4	4.84	32.7	0.727
+ Best-of-N	17.2	4.41	32.3	0.768
+ SteerX	**13.1**	**4.30**	**33.4**	**0.775**
Method (Image-to-3D)	BRISQUE↓	NIQE↓	CLIP-I↑	GS-MEt3R↑
DimensionX [21]	37.3	4.25	82.4	0.708
+ Best-of-N	29.8	4.33	83.2	0.745
+ SteerX	**29.7**	**4.24**	**83.7**	**0.749**

Table 1. **Quantitative results in various scene generation tasks.**

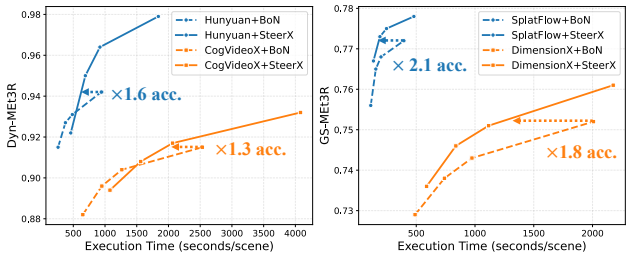


Figure 7. **Scalability analysis** with $k = 2, 3, 4, 8$. We use 100 randomly selected samples in VBench-I2V for Image-to-3D/4D.

Inference-time scalability. Figure 7 presents the execution time versus reward values for all generation tasks. Although SteerX incurs additional computational overhead by forwarding the scene reconstruction model multiple times, it demonstrates better inference-time scalability than BoN. Also, as the number of particles increases, SteerX achieves greater performance gains by exploring more diverse sampling trajectories, rather than relying on post-hoc selection.

4. Conclusion

We have presented SteerX, a zero-shot inference-time steering method for camera-free 3D/4D scene generation. Rather than handling geometric alignment separately in video generation or scene reconstruction, SteerX unifies both stages by tilting the data distribution toward geometrically aligned samples using a particle system based on SMC. To this end, we define two geometric reward functions specifically designed for 3D and 4D scenes. SteerX enables efficient and scalable Image/Text-to-3D/4D scene generation.

References

[1] Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Measuring multi-view consistency in generated images. *arXiv preprint arXiv:2501.06336*, 2025. 2

[2] Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. *arXiv preprint arXiv:2411.18673*, 2024. 1

[3] Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. Vd3d: Taming large video diffusion transformers for 3d camera control. *arXiv preprint arXiv:2407.12781*, 2024. 1

[4] 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. 1

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 2

[6] Guikun Chen and Wenguan Wang. A survey on 3d gaussian splatting. *arXiv preprint arXiv:2401.03890*, 2024. 1

[7] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022. 3

[8] Arnaud Doucet, Nando De Freitas, and Neil Gordon. An introduction to sequential monte carlo methods. *Sequential Monte Carlo methods in practice*, pages 3–14, 2001. 2

[9] Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. 3

[10] Stephanie Fu, Mark Hamilton, Laura Brandt, Axel Feldman, Zhoutong Zhang, and William T Freeman. Featup: A model-agnostic framework for features at any resolution. *arXiv preprint arXiv:2403.10516*, 2024. 2

[11] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024. 1

[12] Hyojun Go, Byeongjun Park, Jiho Jang, Jin-Young Kim, Soonwoo Kwon, and Changick Kim. Splatflow: Multi-view rectified flow model for 3d gaussian splatting synthesis. *arXiv preprint arXiv:2411.16443*, 2024. 1, 2, 3, 4

[13] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. 1

[14] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 1

[15] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1

[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, 2, 3, 4

[17] 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. *Advances in Neural Information Processing Systems*, 37:75125–75151, 2025. 1

[18] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022. 3

[19] 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. *Communications of the ACM*, 65(1):99–106, 2021. 1

[20] Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models. *arXiv preprint arXiv:2501.06848*, 2025. 3

[21] Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, and Yikai Wang. Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. *arXiv preprint arXiv:2411.04928*, 2024. 1, 2, 3, 4

[22] Zhenggang Tang, Yuchen Fan, Dilin Wang, Hongyu Xu, Rakesh Ranjan, Alexander Schwing, and Zhicheng Yan. Mv-dust3r+: Single-stage scene reconstruction from sparse views in 2 seconds. *arXiv preprint arXiv:2412.06974*, 2024. 1, 2

[23] Genmo Team. Mochi 1. `https://github.com/genmoai/models`, 2024. 1

[24] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, Wang Yifan, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, et al. Advances in neural rendering. In *Computer Graphics Forum*, pages 703–735. Wiley Online Library, 2022. 1

[25] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 1

[26] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20310–20320, 2024. 1

[27] Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T Barron, and Aleksander Holynski. Cat4d: Create anything in 4d with multi-view video diffusion models. *arXiv preprint arXiv:2411.18613*, 2024. 1

[28] Tong Wu, Yu-Jie Yuan, Ling-Xiao Zhang, Jie Yang, Yan-Pei Cao, Ling-Qi Yan, and Lin Gao. Recent advances in 3d gaussian splatting. *Computational Visual Media*, 10(4):613–642, 2024. 1

[29] 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, 2, 3, 4

[30] Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. *arXiv preprint arXiv:2410.24207*, 2024. 2

[31] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024. 1, 2, 3

[32] Yuyang Zhao, Chung-Ching Lin, Kevin Lin, Zhiwen Yan, Linjie Li, Zhengyuan Yang, Jianfeng Wang, Gim Hee Lee, and Lijuan Wang. Genxd: Generating any 3d and 4d scenes. *arXiv preprint arXiv:2411.02319*, 2024. 1