CameraSquad: Achieving Content Consistency in Parallel Multi-Trajectory Camera-Controlled Video Generation

ZHUFENG XU, Institute of Computing Technology, CAS, and University of Chinese Academy of Sciences, China
XUAN GAO, Institute of Computing Technology, CAS, and School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences, China
BAILIN DENG, Cardiff University, United Kingdom
YIKANG DING, Kuaishou Technology, China
XIAOQIANG LIU, Kuaishou Technology, China
HAOXIAN ZHANG, Kuaishou Technology, China
PENGFEI WAN, Kuaishou Technology, China
HONGBO FU, The Hong Kong University of Science and Technology, Hong Kong
LIN GAO*, Institute of Computing Technology, CAS, and University of Chinese Academy of Sciences, China

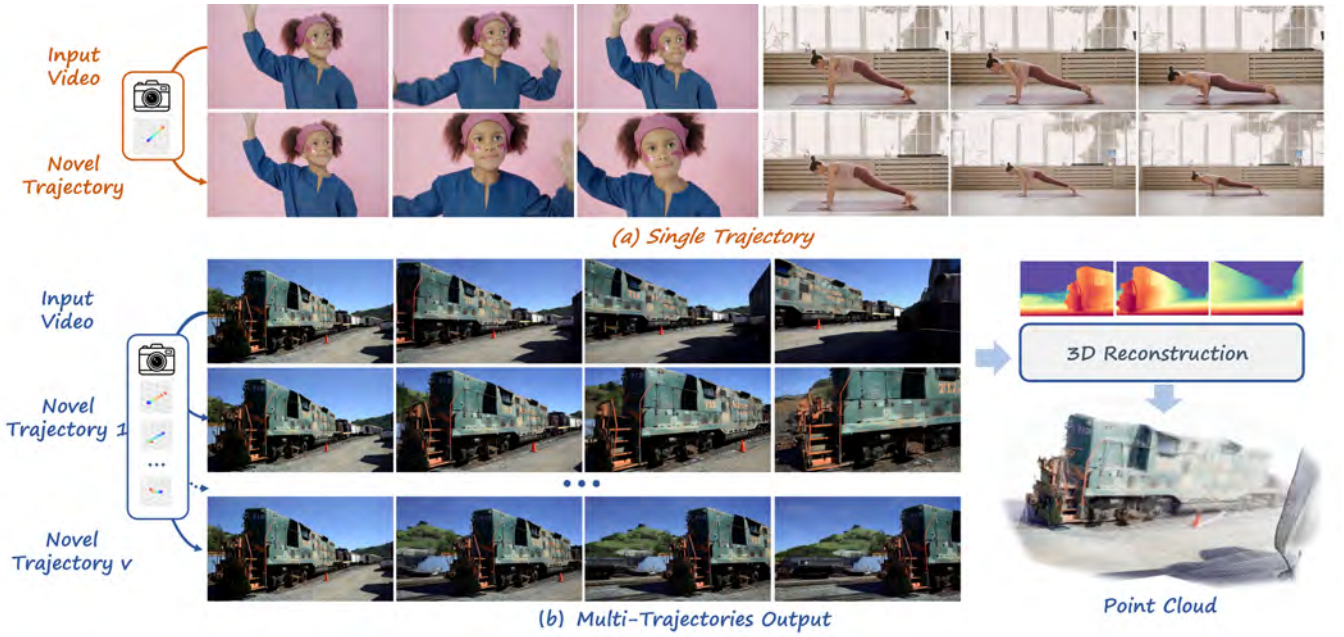

Fig. 1. Examples synthesized by CameraSquad. CameraSquad supports both single-trajectory (a) and multi-trajectory (b) generation modes. Given an input video and several sets of target camera parameters, CameraSquad can generate precise rendering results along specified trajectories, while maintaining scene consistency for objects across different trajectories. Leveraging existing powerful feed-forward models, we can back-project content-consistent dynamic point clouds based on multi-view trajectories and depth estimation results.

Camera-controlled video generation is valuable for applications ranging from visual design to providing 2D supervision for 4D generation tasks. However, existing approaches are limited to single-trajectory generation, forcing users to process multiple trajectories in separate batches. This serial inference introduces content inconsistencies across viewpoints due to the inherent randomness of diffusion models. Explicit point cloud methods can only partially address this problem, as single-viewpoint back-projection suffers from sparsity and depth estimation errors. We propose CameraSquad,

*Corresponding author: Lin Gao (gaolin@ict.ac.cn).

Authors' Contact Information: Zhufeng Xu, xuzhufeng22s@ict.ac.cn, Institute of Computing Technology, CAS, and University of Chinese Academy of Sciences, Beijing, China; Xuan Gao, gaoxuan193@mails.ucas.ac.cn, Institute of Computing Technology, CAS, and School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences, Beijing, China; Bailin Deng, DengB3@cardiff.ac.uk, Cardiff University, Cardiff, United Kingdom; Yikang Ding, dingyikang@kuaishou.com, Kuaishou Technology, Beijing, China; Xiaoqiang Liu, liuxiaoqiang@kuaishou.com, Kuaishou Technology, Beijing, China; Haoxian Zhang, zhanghaoxian@kuaishou.com, Kuaishou Technology, Beijing, China; Pengfei Wan, wanpengfei@kuaishou.com, Kuaishou Technology, Beijing, China; Hongbo Fu, fuplus@gmail.com, The Hong Kong University of Science and Technology, Hong Kong, Hong Kong; Lin Gao, gaolin@ict.ac.cn, Institute of Computing Technology, CAS, and University of Chinese Academy of Sciences, Beijing, China.

SIGGRAPH Conference Papers '26, Los Angeles, CA, USA

ACM ISBN 979-8-4007-2554-8/2026/07
https://doi.org/10.1145/3799902.3811048

a multi-trajectory camera control framework that supports both single-trajectory and parallel multi-trajectory generation. Our method achieves precise camera control while preserving input video content through decoupled content and camera control mechanisms. To ensure viewpoint consistency in multi-trajectory mode, we design a dual-mode cross-view attention mechanism that maintains consistency across parallel trajectories while guaranteeing camera control precision. Extensive experiments demonstrate that CameraSquad achieves competitive performance in camera control accuracy, consistency maintenance, and generation quality compared to existing approaches. Our project page is available at https://rabberk.github.io/CameraSquad/.

CCS Concepts: • **Computing methodologies** → **Computer vision**; *Image-based rendering*; **Computer vision tasks**.

Additional Key Words and Phrases: Camera Control, Video Generation, Multi-Trajectory Generation

ACM Reference Format:
Zhufeng Xu, Xuan Gao, Bailin Deng, Yikang Ding, Xiaoqiang Liu, Haoxian Zhang, Pengfei Wan, Hongbo Fu, and Lin Gao. 2026. CameraSquad: Achieving Content Consistency in Parallel Multi-Trajectory Camera-Controlled Video Generation. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers (SIGGRAPH Conference Papers '26), July 19–23, 2026, Los Angeles, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3799902.3811048

1 Introduction

Recent advancements [Bar-Tal et al. 2024; Blattmann et al. 2023a,b; Brooks et al. 2022; Chen et al. 2023, 2024; Guo et al. 2023; Gupta et al. 2024; Ho et al. 2022; Li et al. 2018; Lin et al. 2025b; Ning et al. 2025; Singer et al. 2022; Villegas et al. 2022; Wang et al. 2023, 2020; Zhou et al. 2022, 2024] in video model technology have showcased their robust generative capabilities, opening up promising avenues for the application of related generative techniques in film production. However, the current landscape is largely shaped by the use of Internet data, which has resulted in the predominance of foundational models primarily optimized for text-to-video generation. The reliance on text input or single images for video generation presents significant challenges in meeting user expectations for detailed and complex content. Users often seek to convey intricate narratives or fine visual elements, yet existing models struggle to effectively realize these demands. Consequently, precise control over generated video content is limited, which may hinder the creative process and affect user experience. Addressing these issues is crucial for unlocking the potential of generative video technology in the creative industry. Building upon this premise, the exploration of controllable generation in video models [Guo et al. 2024; Xiao et al. 2024; Xing et al. 2024; Yin et al. 2023] has emerged as a frontier research area, focusing on integrating new control conditions into existing backbone models.

Among these, camera-controllable approaches are particularly significant, aiming to generate renderings that align with user-specified camera logic, whether through descriptive text, predefined trajectories [Guo et al. 2023], or detailed intrinsic and extrinsic parameters [Bai et al. 2025; Ren et al. 2025; Yu et al. 2025a]. Precise new trajectory video generation not only provides valuable reference movements for CG production but also offers a denser 2D prior for comprehensive 4D reconstruction. However, existing methods [Bai et al. 2025; He et al. 2024; Ren et al. 2025; Yu et al. 2025a] often support generation based on a single target trajectory. In practical applications, when utilizing these techniques to provide reference for camera effects or to guide downstream 4D generation with 2D inputs, it often becomes necessary to perform serial inference across multiple trajectories. This approach, however, is susceptible to inconsistencies between different viewpoints due to the inherent randomness of video models. Such discrepancies can accumulate and adversely affect downstream 4D reconstructions, ultimately compromising the fidelity and coherence of the generated content.

To address the challenges of camera control methods in generating specific camera trajectories, particularly the inconsistency between viewpoints, a more robust framework is required. In this regard, we propose CameraSquad, a new approach that incorporates a viewpoint attention mechanism designed to ensure consistency among multiple perspectives, effectively mitigating the adverse effects of viewpoint inconsistencies. Our key innovation lies in injecting camera control signals into DiT by utilizing the RoPE (Relative Position Encoding) mechanism in the attention mechanism and introducing two cross-view attention mechanisms, which enable the parallel generation of multiple trajectories with consistent content.

Extensive experiments demonstrate that CameraSquad can achieve precise camera trajectories while producing competitive novel view rendering results compared to state-of-the-art (SOTA) methods. Moreover, CameraSquad improves geometric consistency across different trajectories, providing higher-quality priors for downstream reconstruction tasks. In summary, our contributions include:

- We propose a spatially aware camera and input video injection method that allows for precise control of camera pose while effectively maintaining the quality of content generation.
- We propose a parallel multi-trajectory camera control scheme that enables precise control over trajectory generation results in a single inference process.
- We propose a dual-mode cross-view attention mechanism that maintains content consistency of the input video while further ensuring the accuracy of camera trajectories and geometric consistency between viewpoints.

2 Related Work

Diffusion-based Novel View Generation. Leveraging 2D diffusion models for novel view generation has advanced rapidly in recent years. Recent studies advance novel view synthesis through video diffusion models, incorporating various conditioning mechanisms, including depth-warped images [Zhang et al. 2025c] and point cloud renderings [Yu et al. 2025b]. For static scenarios, generative novel view synthesis has shown remarkable progress through diffusion models. Zero-1-to-3 [Liu et al. 2023] presents diffusion models for objects with clean backgrounds, later extended to generic scenes by ZeroNVS [Sargent et al. 2024]. ReconFusion [Wu et al. 2024] refines pose estimation using PixelNeRF [Yu et al. 2021] features. Some approaches [Gao et al. 2024; Szymanowicz et al. 2025; Zhou et al. 2025] utilize diffusion models as multi-view generators conditioned on camera parameters. These methods extend to multi-view video generation for 4D synthesis, with multitasking designs supporting single or multiple views with camera poses as input. Additional

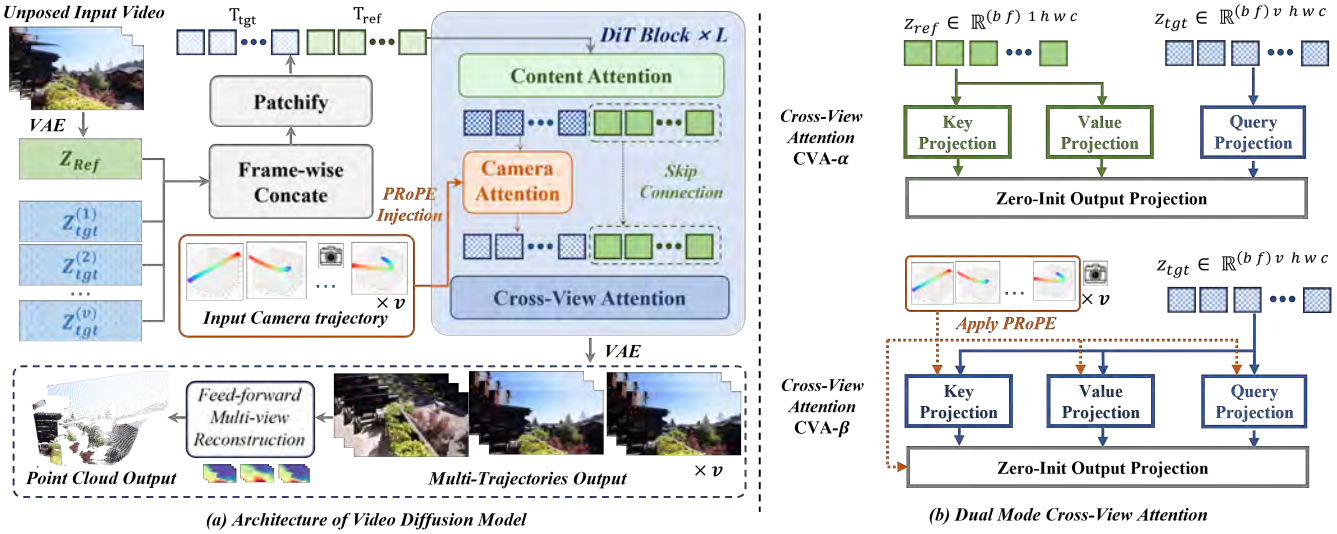

Fig. 2. **Overview of CameraSquad.** Given an input video and V target trajectories, CameraSquad concurrently generates multiple camera motion videos for the specified trajectories. As illustrated in (a), CameraSquad is built upon Wan2.2 as the base video diffusion model, where the self-attention layers are adapted into content attention, and PRoPE is employed to provide fundamental single-trajectory camera control. To ensure geometric consistency and texture invariance across perspectives during multi-trajectory generation, CameraSquad integrates a Dual-Mode Cross-View Attention mechanism (b). This module comprises two complementary forms of cross-view regulation: alpha-cross-view attention (CVA-α) for subject consistency and beta-cross-view attention (CVA-β) for relative perspective accuracy.

research [Hunyuan3D et al. 2025; Liu et al. 2024; Xiang et al. 2025] focuses on 3D object generation from images or text.

Dynamic scene synthesis induces additional complexities. NVS-Solver [You et al. 2024] leverages pretrained models for training-free depth-warped video inpainting, but encounters challenges with geometric distortions. Several methods [Dai et al. 2024; Jing et al. 2025; Liu et al. 2025; Wang et al. 2025; Wu et al. 2025; Yu et al. 2024] focus on 4D representation learning, which construct spatial-temporal models for dynamic novel view synthesis.Despite their strong rendering quality, these 4D methods are designed for scene reconstruction and view rendering, and none supports camera-controlled generation. CameraDolly [Van Hoorick et al. 2024] adapts SVD models [Blattmann et al. 2023a] using synthetic multi-view datasets from the Kubric simulator [Greff et al. 2022]. Some research efforts [Sun et al. 2025; Zhang et al. 2025a] employ LoRA to fine-tune video models for novel view synthesis; however, such methods necessitate targeted LoRA adaptation for each trajectory or video, thereby constraining their generalization capabilities and practical value.

Camera-Controlled Video Generation. Camera-guided video generation focuses on using virtual camera intrinsic and extrinsic parameters as control conditions to determine the viewpoint and potential intrinsic parameters during video synthesis. This is essential for achieving specific cinematic styles, narrative perspectives, and dynamic visual effects, offering filmmakers and content creators fine-grained control over how a scene is presented. Trajectory Attention [Patrick et al. 2021] introduces trajectory attention into Video Transformers, achieving state-of-the-art results in video action recognition tasks. MotionCtrl [Wang et al. 2024b] uses camera extrinsic matrices for control, but its reliance on 1D values limits precision in complex scenarios. CameraCtrl [He et al. 2024] also employs Plücker coordinates, focusing on text-to-video generation with less emphasis on 3D consistency. CamCo [Xu et al. 2024] further enhances 3D consistency through epipolar attention. CameraCtrl II [He et al. 2025] enhances the representation capability of dynamic scenes based on its predecessor, while Direct-a-Video [Yang et al. 2024] supports camera path guidance. Cavia [Xu et al. 2025] enhances geometric consistency in camera control by optimizing noise across frame interactions based on epipolar attention.

Additionally, other works [Cheong et al. 2024; Hu et al. 2024; Yang et al. 2025] focus on addressing performance degradation issues in camera control. ReCamMaster [Bai et al. 2025] addresses camera trajectory control and viewpoint transformation for any video by training on synchronized datasets from multiple cameras, further maintaining multi-frame appearance and dynamic synchronization. Besides, some works explore complex 3D scene understanding and camera path generation [Bahmani et al. 2024; Ren et al. 2025; Yu et al. 2025a,b; Zhang et al. 2025b; Zhu et al. 2025]. However, these methods are limited to single-trajectory generation, and may result in inconsistency across multiple viewpoints when extended to multi-trajectory generation. Several autonomous driving works like DrivingDiffusion [Li et al. 2024], DriveDreamer-2 [Zhao et al. 2025], and OmniNWM [Li et al. 2025a] focus on multi-view video generation, but they do not support free-trajectory movements between shots and video-to-video generation. PlenopticDreamer [Fu et al. 2026] proposes an autoregressive, multi-in–single-out formulation to generate synchronized hallucinations across time and viewpoints. At each step, it retrieves a set of previously generated video–camera

pairs from a memory bank and conditions the next generation on these retrieved contexts.

3 Methods

Given an input video and target camera trajectories, our objective is to produce output videos that are consistent with the scene depicted in the input video, aligned with the designated trajectories, and maintain geometric coherence across different segments. CameraSquad supports both single-trajectory inference using the Camera and Content Attention mentioned in 3.1 and multi-trajectory inference modes. Unlike existing camera control methods, it can simultaneously infer camera viewpoints for multiple trajectories through the Dual Cross-view Attention described in 3.2. Furthermore, CameraSquad does not require additional explicit geometric constraints, thereby avoiding the geometric inconsistencies encountered during the serial inference process of multi-trajectory videos.

3.1 Camera and Content Attention Module

Sparse rendering results often lead to degraded texture generation in scenarios with large viewpoints. To address this issue, some explicit point cloud-based methods [Yu et al. 2025a] additionally incorporate attention control mechanisms tailored to the original video. Inspired by this, we posit that independent integration of content information and camera trajectory data would similarly benefit implicit camera control schemes. Thus, CameraSquad introduces separate modules to implement attention mechanisms for the camera and the content.

As shown in Fig. 2, to safeguard the intrinsic generative capabilities of the DiT model, we repurpose the original 3D Self-Attention components as structural proxies for Content-Attention. Methodologically, the input video is encoded through a VAE and subsequently tokenized and patchified, with the input video tokens and noised target tokens concatenated in a frame-wise manner:

$$[\mathbf{T}^{(l+1)}_{\text{content,tgt}}, \mathbf{T}^{(l+1)}_{\text{content,ref}}] = \text{Content-Attention}([\mathbf{T}^{(l)}_{\text{tgt}}, \mathbf{T}^{(l)}_{\text{ref}}]).$$

Here, $\mathbf{T}^{(l)}_{\text{tgt}} \in \mathbb{R}^{N\times D}$ and $\mathbf{T}^{(l)}_{\text{ref}} \in \mathbb{R}^{N\times D}$ represent the target noised tokens and reference tokens from input video at layer l, respectively, where N is the token count , and D is the feature dimension. The Content-Attention operation processes these sequences to generate the corresponding output sequences $\mathbf{T}^{(l+1)}_{\text{content,tgt}}$ and $\mathbf{T}^{(l+1)}_{\text{content,ref}}$ for the subsequent layer, enabling effective cross-reference learning between target and reference content.

In this study, we leverage the advantages of PRoPE (Projection-based Relative Position Encoding) [Li et al. 2025b] within the Camera-Attention mechanism. PRoPE enhances relative position encoding by fully utilizing the projection relationships between the view frustums of two cameras, incorporating both intrinsic and extrinsic parameters. The calculation of Camera-Attention is as follows,

$$\begin{aligned}\text{Camera-Attention}(T_{tgt}) &= \mathcal{T}_o\left(\text{Attention}\left(\mathcal{T}_q(Q), \mathcal{T}_{kv}(K), \mathcal{T}_{kv}(V)\right)\right)\\ &= \mathcal{T}_o\left(\text{Attention}\left(\mathcal{T}_q(P_Q(T_{tgt})), \mathcal{T}_{kv}(P_K(T_{tgt})), \mathcal{T}_{kv}(P_V(T_{tgt}))\right)\right).\end{aligned}$$

Here, $P_Q(\cdot)$, $P_K(\cdot)$, and $P_V(\cdot)$ denote learnable linear projections used to transform the input tokens into Query, Key, and Value matrices, respectively.Formally, the transformation of each component of attention is applied in the following manner,

$$\mathcal{T}_q = \text{diag}(P^T \otimes I_{d/2}, \text{RoPE}_x \otimes I_{d/4}, \text{RoPE}_y \otimes I_{d/4}) \quad (1)$$

$$\mathcal{T}_{kv} = \text{diag}(P^{-1} \otimes I_{d/2}, \text{RoPE}_x \otimes I_{d/4}, \text{RoPE}_y \otimes I_{d/4}) \quad (2)$$

$$\mathcal{T}_o = \text{diag}(P \otimes I_{d/2}, \text{RoPE}_x^{-1} \otimes I_{d/4}, \text{RoPE}_y^{-1} \otimes I_{d/4}) \quad (3)$$

In the PRoPE framework, the core transformation matrices are defined by partitioning the feature dimension d into three equal segments: the first $\frac{d}{2}$ dimensions encode 3D geometric projections, while the subsequent $\frac{d}{4}$ dimensions each handle 2D positional encoding along the x and y axes respectively. The projection matrix P typically corresponds to the camera intrinsic matrix and view matrix which represents the transformation between world and camera, and RoPE_x, RoPE_y denote rotary position embedding matrices for respective directions. I_k represents the k-dimensional identity matrix, and $\otimes$ denotes the Kronecker product. To broaden the applicability of our approach, our model is designed to work with unposed reference videos, eliminating the need for camera intrinsic and extrinsic parameters of the original footage. Thus, during Camera-Attention computation, we concentrate exclusively on $\mathbf{T}_{\text{tgt}}$ associated with virtual cameras that have specified target trajectories:

$$\begin{aligned}\mathbf{T}^{(l+1)}_{\text{camera,tgt}} &= \text{Camera-Attention}(\mathbf{T}^{(l)}_{\text{content,tgt}}),\\ \mathbf{T}^{(l+1)}_{\text{tgt}} &= \mathbf{T}^{(l+1)}_{\text{content,tgt}} + \text{Projection}(\mathbf{T}^{(l+1)}_{\text{camera,tgt}}).\end{aligned}$$

Here, Projection$(\cdot)$ serves as a linear layer. To prevent the inherent capabilities of the original video model from being compromised, Projection$(\cdot)$ injects camera control capabilities into the backbone network through zero initialization. The output from this section, after being processed with the cross attention layer, serves as the final output of this DiT Block. To prevent overfitting during the training process from undermining the original generative capabilities of the video model, we freeze this set of parameters.

3.2 Dual-Mode Cross-View Attention

Once we have implemented the single-trajectory camera control model for the aforementioned components, we achieve a multi-trajectory parallel generation scheme by inserting a new cross-view module. To ensure geometric consistency between perspectives, we incorporate an inter-view attention mechanism within specific DiT modules, enabling the model to learn patterns from various viewpoints. We use two cross-view attention mechanisms to help the model maintain consistency across multiple perspectives and preserve texture invariance: the alpha-cross-view attention (CVA-α) for maintaining subject consistency, and the beta-cross-view attention (CVA-β) for ensuring perspective accuracy (see Fig. 2).

We employ CVA-α to facilitate the preservation of texture. In this context, the reference video tokens provide the cross-attention keys and values, while the noised latent tokens from each trajectory serve as the query. We reshape all the latent tokens to ensure mutual visibility among tokens of varying perspectives at the same frame. In addition, we introduce CVA-β to further enhance camera control capabilities through multi-view supervision. Here, we similarly apply the Camera Attention mechanism mentioned previously as its structure, but adjust the single-track attention computation along the frame to a multi-track attention computation

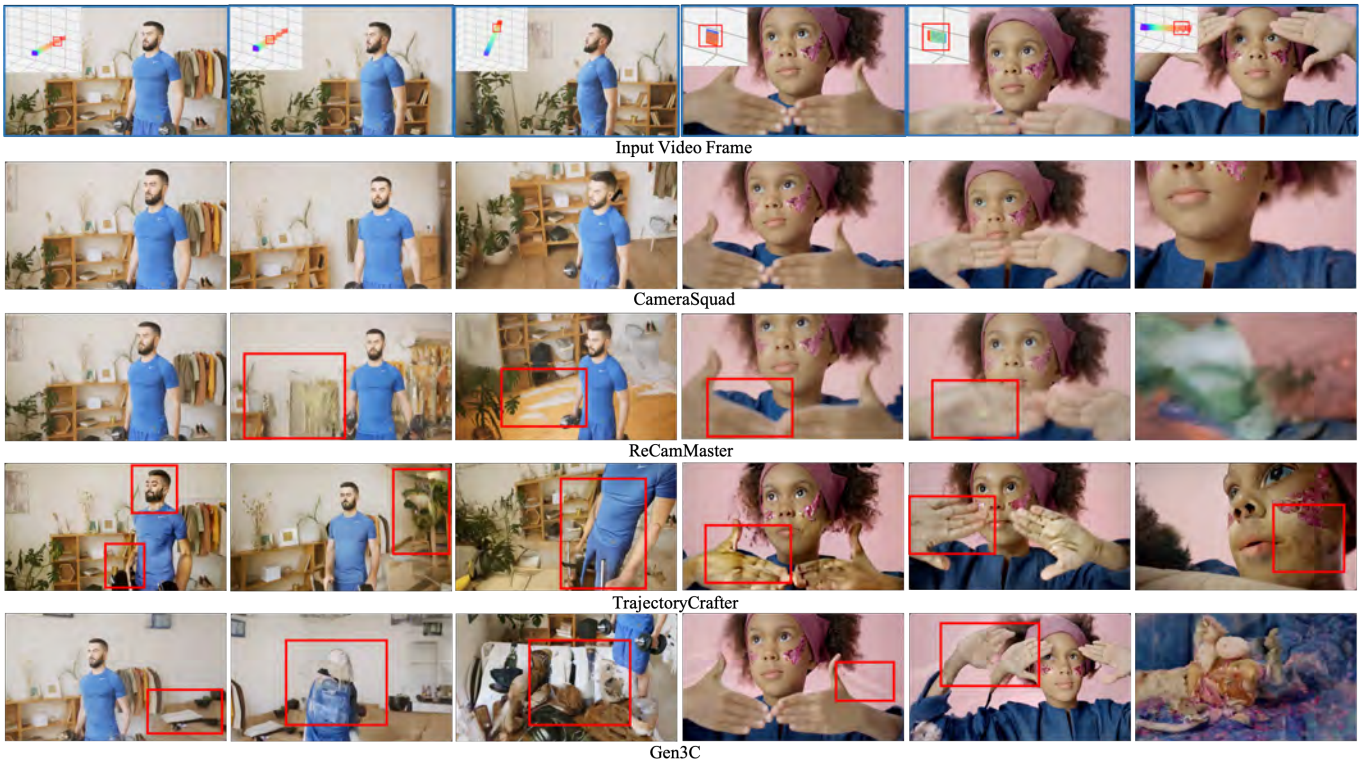

Fig. 3. **Qualitative comparison of single trajectory video synthesis.** We compare the generation quality of CameraSquad with existing state-of-the-art camera control methods on a single trajectory. Some results are also shown in the supplementary video.

Table 1. Quantitative comparison with state-of-the-art methods on WebVid and HumanVid Datasets.

Method	WebVid Dataset							HumanVid Dataset						
	Camera & Geometric				Visual Quality			Camera & Geometric				Visual Quality		
	RotErr ↓	TransErr ↓	MPI(K) ↑	MPO(K) ↑	FID ↓	FVD ↓	CLIP-V ↑	RotErr ↓	TransErr ↓	MPI(K) ↑	MPO(K) ↑	FID ↓	FVD ↓	CLIP-V ↑
ReCamMaster	1.72	3.36	931.50	834.37	30.41	365.53	89.02	2.35	3.91	933.12	820.15	37.91	306.55	90.81
TrajectoryCrafter	2.34	4.23	874.08	793.82	40.86	408.33	87.07	2.93	4.72	837.70	675.33	42.26	324.25	87.47
Gen3C	1.79	3.48	814.75	631.00	49.08	441.77	83.64	2.49	3.80	880.79	745.23	52.12	446.83	88.05
Ours	1.52	2.86	945.62	870.17	27.11	374.65	88.50	1.42	3.47	946.67	849.18	30.78	246.53	91.37

along the view dimension. Specifically, given the latent features of a reference video, $z_{ref} \in \mathbb{R}^{B\times F\times C\times H\times W}$, and the noisy target tokens, $z_{tgt} \in \mathbb{R}^{B\times F\times N\times C\times H\times W}$, we perform a reshape operation:

$$z_{\text{ref}} : (b \times 1, f, h, w, c) \rightarrow (b \times f, 1, h, w, c), \tag{4}$$

$$z_{\text{tgt}} : (b \times v, f, h, w, c) \rightarrow (b \times f, v, h, w, c). \tag{5}$$

The CVA is then formulated by substituting these projections into the scaled dot-product attention equation,

$$\text{CVA-}\alpha(\mathcal{T}_{\text{ref}}, \mathcal{T}_{\text{tgt}}) = \text{softmax}\left(\frac{P_Q(\mathcal{T}_{\text{tgt}}) \cdot P_K(\mathcal{T}_{\text{ref}})^T}{\sqrt{d_k}}\right) P_V(\mathcal{T}_{\text{ref}}),$$

$$\text{CVA-}\beta(\mathcal{T}_{\text{tgt}}) = \text{Camera-Attention}(\mathcal{T}_{\text{tgt}}).$$

Here, $\mathcal{T}_{\text{ref}}$ and $\mathcal{T}_{\text{tgt}}$ represent the set of reference tokens and target tokens, respectively. d_k is the dimension of the key vectors, and $\sqrt{d_k}$ is used as a scaling factor to stabilize the gradients. The softmax function is applied to compute the final attention weights.

3.3 Point Cloud Unprojection

By using the powerful generation and generalization capabilities of existing feed-forward models, once we obtain the specified multi-track content-consistent generation results, we can estimate the corresponding depth based on the multi-view results and reproject the point cloud representation of the scene.

We utilize DA3 [Lin et al. 2025a] as the downstream depth map estimation model to effectively leverage the trajectory generation results produced by CameraSquad. This integration facilitates the estimation of depth from RGB videos, significantly enhancing the consistency of point cloud estimation, particularly when explicit camera intrinsics and extrinsics are provided. This not only produces

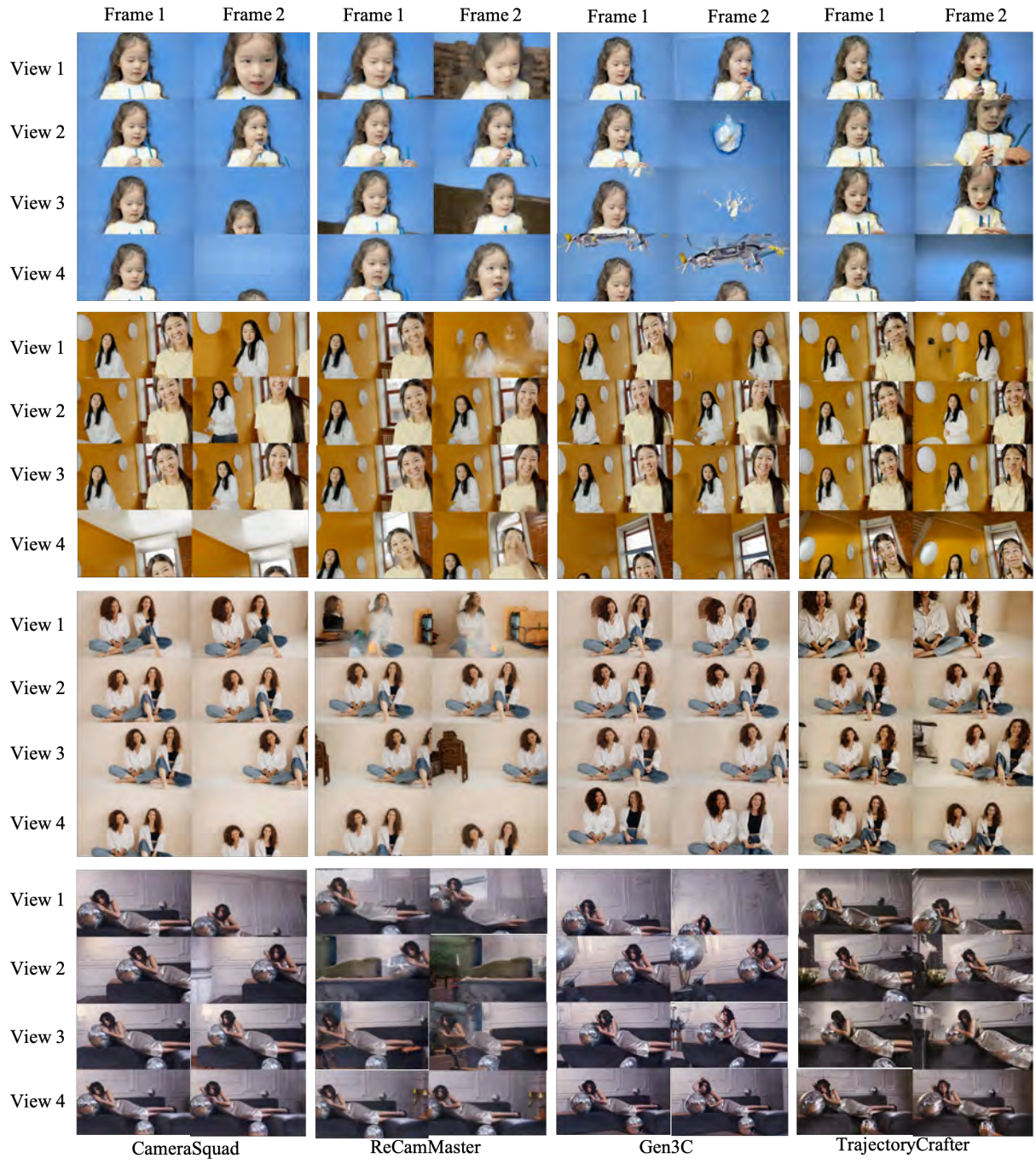

Fig. 4. **Qualitative comparison of multi-trajectory video synthesis.** We compare the generation quality of CameraSquad with existing state-of-the-art camera control methods on multi-trajectory synthesis.

a larger and more detailed point cloud compared to the single-layer point cloud projected from a single viewpoint, but also enables us to capture a dynamic point cloud representation of the relevant scene by incorporating the temporal dimension. By considering the temporal aspects, we can better represent changes in the scene over time, leading to richer and more informative 3D reconstructions.

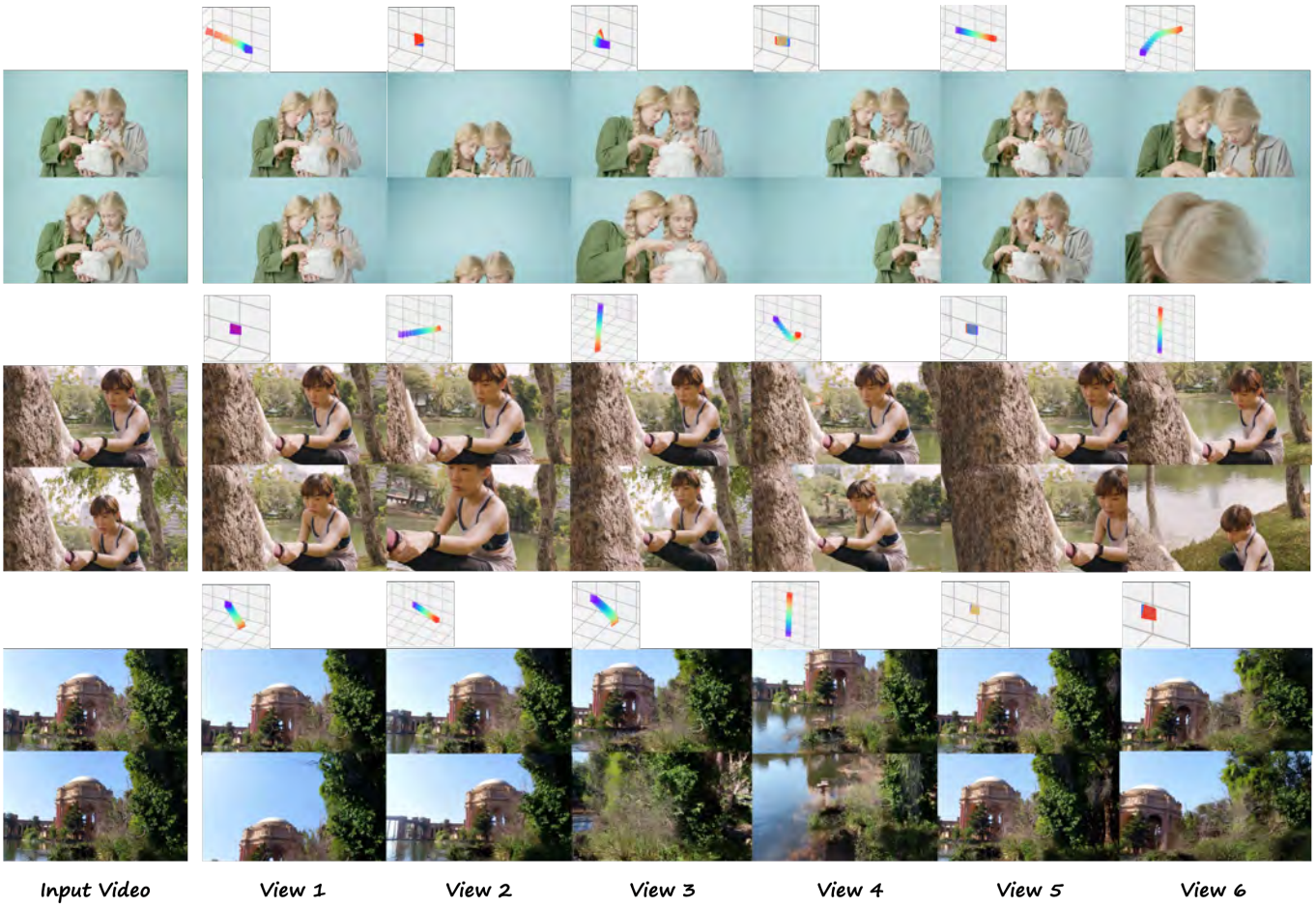

Fig. 5. **Visualization of 6-Traj. generation results.** CameraSquad enables synchronous generation of up to 6 trajectories. We present visualization results of CameraSquad on human and landscape videos, showing that our approach consistently produces stable outputs.

Table 2. Quantitative comparison with SOTA methods on VBench metrics.

Dataset	Method	AQ ↑	IQ ↑	MS ↑	BC ↑	SC ↑
WebVid	ReCamMaster	0.4327	0.5412	0.9816	0.9209	0.9352
	TrajectoryCrafter	0.4443	0.5713	0.9772	0.9133	0.9342
	Gen3C	0.3893	0.5209	0.9701	0.8923	0.9037
	CameraSquad (ours)	0.4390	0.5663	0.9821	0.9218	0.9402
HumanVid	ReCamMaster	0.5277	0.6223	0.9836	0.9163	0.8954
	TrajectoryCrafter	0.5314	0.6790	0.9821	0.9296	0.9190
	Gen3C	0.4912	0.6322	0.9797	0.8911	0.8754
	CameraSquad (ours)	0.5377	0.6703	0.9891	0.9313	0.9260

Table 3. User study results comparing visual quality across methods.

Method	**VQ ↑**	**SC ↑**	**MS ↑**	**TA ↑**	**CC ↑**
ReCamMaster	2.95	2.79	3.01	3.13	3.06
TrajectoryCrafter	3.01	3.01	3.11	3.30	3.10
GEN3C	2.77	2.99	3.14	3.38	3.05
CameraSquad (Ours)	**3.97**	**4.01**	**4.03**	**3.97**	**3.76**

4 Experiment

4.1 Experiment Settings

Implementation Details. We employ Wan2.2 5B [Wan et al. 2025] as foundational model and conduct training on ReCamMaster [Bai et al. 2025] and SynCamMaster [Bai et al. 2024] datasets. CVA modules are inserted into even-numbered DiT blocks with alternating modes. Our training supports simultaneous generation of up to 6 trajectories. We train a single-trajectory model at a low resolution of 256x416 with a learning rate of 1e-5 to enable the video model to perceive camera control conditions. Subsequently, we incorporate cross-view attention to facilitate multi-track generation, supporting parallel generation of six views. This phase utilizes a learning rate of 5e-6, with the two stages of the learning process comprising 8k steps and 6k steps with a batch size of 16.

Evaluation Details. To validate the consistency of perspectives, we randomly sampled from the WebVid [Bain et al. 2021] and HumanVid [Wang et al. 2024a] datasets, selecting four trajectories for

Fig. 6. **More results of qualitative comparison of single trajectory video synthesis.**

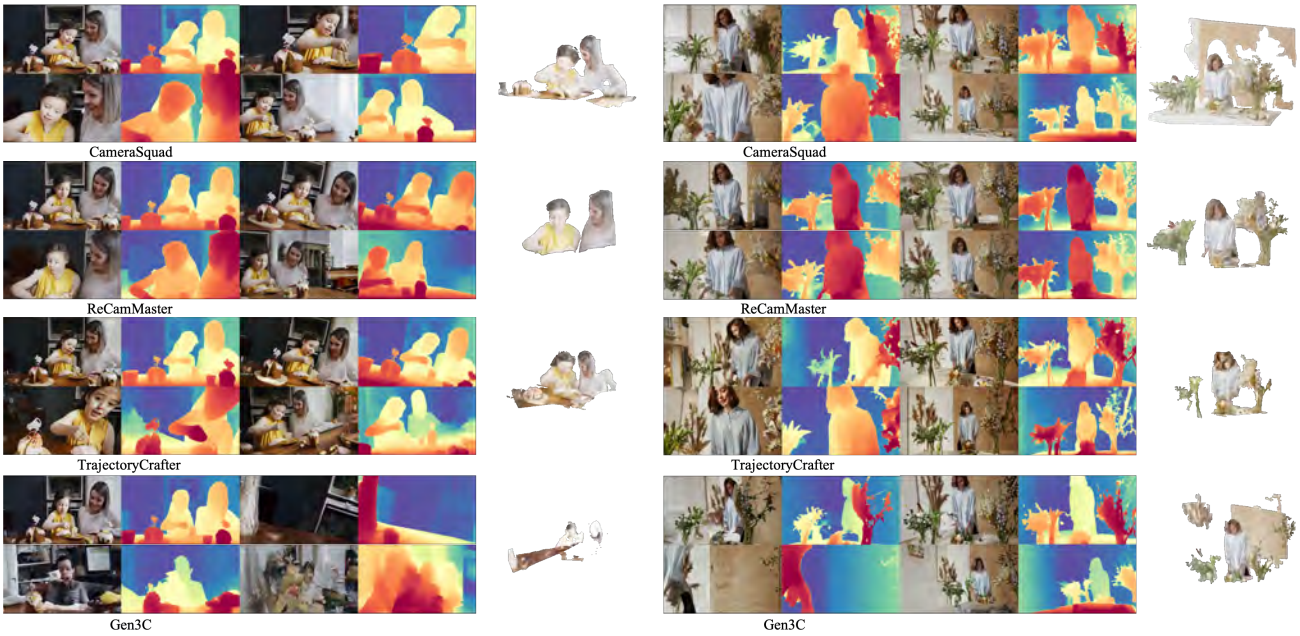

Fig. 7. **Qualitative comparison of multi-view point cloud reprojection.** We compare the point cloud quality of CameraSquad with existing state-of-the-art camera control methods by reprojecting multi-view frames using DA3 estimated depth into point clouds.

Fig. 8. **Ablation study results.** We validate the qualitative results under different structural designs and training strategy designs.

Table 4. Ablation study on structural design and training strategy.

Config.	RotErr ↓	TransErr ↓	MPI(K) ↑	MPO(K) ↑	FID ↓	FVD ↓	CLIP-V ↑
Structural Design Ablations							
Full model	1.42	3.47	946.67	849.18	30.78	246.53	91.37
w/ Raymap	2.33	5.65	912.34	829.91	31.21	301.99	90.67
w/ Parallel ins.	2.17	3.98	877.78	834.59	46.34	396.71	83.47
w/o CVA	1.83	3.62	881.20	811.28	38.32	247.13	89.09
w/o CVA-β	1.77	3.68	905.26	821.20	32.09	277.60	86.12
w/o CVA-α	2.12	4.34	923.98	822.43	39.10	343.48	86.33
Training Strategy Ablations							
Full training	1.42	3.47	946.67	849.18	30.78	246.53	91.37
w/o Noise inj.	1.92	4.01	879.19	805.32	51.39	330.29	83.98
w/o two strat.	1.58	3.60	937.12	845.55	32.70	249.39	90.21

quality generation for each input video, with a total of 800 videos aggregated from each dataset. We utilize RotErr and TransErr [He et al. 2024] to assess the precision of camera control, while employing metrics such as FVD [Unterthiner et al. 2019], FID [Heusel et al. 2017], CLIP-V [Radford et al. 2021] to evaluate the quality of generation. Additionally, we use Mat. Pix. [Shen et al. 2024] to verify the pixel matching degree between input-output pairs (MPI) and among the output perspectives (MPO), which can most intuitively reflect the content consistency between viewpoints and the matching degree of the downstream reprojected point clouds.

We evaluate our method using five key VBench [Huang et al. 2024] metrics: Aesthetic Quality (AQ) measures the visual appeal and artistic quality of generated videos, Imaging Quality (IQ) assesses the technical quality including resolution, clarity, and visual fidelity, Motion Smoothness (MS) evaluates the temporal consistency and smoothness of motion between frames, Background Consistency (BC) measures the stability and coherence of background elements throughout the video, and Subject Consistency (SC) evaluates the consistency of main subjects or objects.

4.2 Comparsion Results

As shown in Tab. 1, Fig. 3, Fig. 4, Fig. 5 and Fig. 6, the quantitative evaluation results on WebVid and HumanVid datasets demonstrate that our method achieves significant performance improvements in camera control accuracy, geometric consistency, and visual quality. On the WebVid dataset, our method outperforms ReCamMaster [Bai et al. 2025], TrajectoryCrafter [Yu et al. 2025a] and Gen3C [Ren et al. 2025] with rotation error of 1.52 and translation error of 2.86, while achieving the competitive FID and FVD scores. On HumanVid, our method achieves superior FID and FVD scores, substantially outperforming all baselines while attaining the highest CLIP-V score. VBench results in Tab. 2 show our method leads across key metrics, comprehensively validating our approach's advantages.

In Fig. 7, we perform depth estimation and point cloud reprojection using multi-view generation results with DA3 as the depth estimator. All methods use the same adaptive confidence threshold lower percentile of 40 (DA3's default), filtering out points below this confidence. This evaluates perspective consistency through reprojected point cloud quality, discarding inconsistent pixels. Our method demonstrates superior pixel preservation across viewpoints, while ReCamMaster shows background region losses and TrajectoryCrafter/Gen3C exhibit subject region losses. This highlights varying geometric consistency across existing methods in different scenarios, whereas our approach achieves superior scene consistency compared to contemporaries.

4.3 User Study

We conduct a user study to evaluate the visual quality of the generated videos. We employ five distinct scoring metrics: Video Quality (VQ), Subject Consistency (SC), and Motion Smoothness (MS), which align with the established metrics utilized in VBench. In addition to these, we introduce two supplementary metrics, Trajectory Accuracy (TA) and Cross-view Consistency (CC) to assess camera control precision and the consistency of visual representations across different viewpoints. For this evaluation, we recruited 29 participants who rated 10 sample groups, with each group containing 1 input

video and 4 paired output videos from different methods. Participants received clear metric definitions and provided ratings on a 5-point Likert scale (1-5). The evaluation outcomes are systematically presented in Table 3. The results demonstrate that our method significantly outperforms all baseline approaches across all five evaluation dimensions.

4.4 Ablation Study

We perform a series of ablation studies to validate the structural design of the model and the training strategy. All ablation studies are conducted using the HumanVid dataset. Structurally, we compared the PRoPE-based camera control with that based on Raymap. Furthermore, we explored the effects of parallel insertion of input video tokens into each self-attention layer against sequentially inputting them as the output of the preceding layer. Specifically, in our full model, the reference tokens from the previous DiT block are concatenated sequentially with the target tokens. The parallel insertion involves compressing the input video, and then concatenating it to the input of each DiT block. After processing through the block, only the portions corresponding to the noised tokens are retained. We evaluated the performance implications of integrating cross-view attention mechanisms, comparing simultaneous deployment with individual usage. Regarding the training strategy, we implement noise injection to mitigate the sim-to-real gap and examine the impact of initial low-resolution camera control training before transitioning to full-resolution on convergence speed.

As shown in Fig. 8 and Tab. 4, our ablation studies validate key design choices. PRoPE outperforms Raymap with lower camera control errors. Sequential token insertion surpasses parallel insertion in quality maintenance. Complete cross-view attention achieves optimal performance over individual components. Noise injection significantly improves sim-to-real transfer, with removal causing quality degradation.

5 Conclusion

We present CameraSquad, a novel approach for multi-trajectory parallel generation with implicit camera control. CameraSquad achieves high-precision camera-controlled video generation and, through our cross-view attention mechanism, produces geometrically consistent results across viewpoints compared to existing camera control methods. By leveraging feed-forward depth estimators to back-project point cloud representations, CameraSquad generates more consistent and scene-preserving results, yielding higher confidence depth estimates and superior point cloud quality.

However, our method has limitations due to the lack of explicit geometric priors. When encountering abstract scenes or out-of-distribution cases, our model may fail to generate satisfactory results. This is partly attributed to our reliance on synthetic training data. Furthermore, the camera movements in existing open-source multi-trajectory datasets are mostly distributed around relatively the same starting point and within a relatively small range, which limits our current work's ability to further extrapolate to larger fields of view. We believe that future availability of larger real-world multi-view video datasets with precise camera annotations will significantly improve our model's generation performance.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62561160115, 62322210), the Beijing Major Science and Technology Project (No. Z251100008125009), the Innovation Funding of ICT, CAS (No. E561100) and sponsored by CAAI-MindSpore Open Fund,developed on OpenICommunity.

References

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

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. 2025. Recammaster: Camera-controlled generative rendering from a single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14834–14844.

Jianhong Bai, Menghan Xia, Xintao Wang, Ziyang Yuan, Xiao Fu, Zuozhu Liu, Haoji Hu, Pengfei Wan, and Di Zhang. 2024. Syncammaster: Synchronizing multi-camera video generation from diverse viewpoints. *arXiv preprint arXiv:2412.07760* (2024).

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1728–1738.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. 2024. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*. 1–11.

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

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. 2023b. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 22563–22575.

Tim Brooks, Janne Hellsten, Miika Aittala, Ting-Chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei Efros, and Tero Karras. 2022. Generating long videos of dynamic scenes. *Advances in Neural Information Processing Systems* 35 (2022), 31769–31781.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. 2023. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512* (2023).

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. 2024. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7310–7320.

Soon Yau Cheong, Duygu Ceylan, Armin Mustafa, Andrew Gilbert, and Chun-Hao Paul Huang. 2024. Boosting camera motion control for video diffusion transformers. *arXiv preprint arXiv:2410.10802* (2024).

Yanran Dai, Jing Li, Yuqi Jiang, Haidong Qin, Bang Liang, Shikuan Hong, Haozhe Pan, and Tao Yang. 2024. Real-time distance field acceleration based free-viewpoint video synthesis for large sports fields. *Computational Visual Media* 10, 2 (2024), 331–353.

Xiao Fu, Shitao Tang, Min Shi, Xian Liu, Jinwei Gu, Ming-Yu Liu, Dahua Lin, and Chen-Hsuan Lin. 2026. Plenoptic Video Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin Brualla, Pratul Srinivasan, Jonathan Barron, and Ben Poole. 2024. CAT3D: Create Anything in 3D with Multi-View Diffusion Models. *Advances in Neural Information Processing Systems* 37 (2024), 75468–75494.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. 2022. Kubric: a scalable dataset generator. (2022).

Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. 2024. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *European Conference on Computer Vision*. Springer, 330–348.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. 2023. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*

(2023).
Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Fei-Fei Li, Irfan Essa, Lu Jiang, and José Lezama. 2024. Photorealistic video generation with diffusion models. In *European Conference on Computer Vision*. Springer, 393–411.
Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. 2024. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101* (2024).
Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. 2025. Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13416–13426.
Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).
Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. 2022. Video diffusion models. *Advances in neural information processing systems* 35 (2022), 8633–8646.
Teng Hu, Jiangning Zhang, Ran Yi, Yating Wang, Hongrui Huang, Jieyu Weng, Yabiao Wang, and Lizhuang Ma. 2024. Motionmaster: Training-free camera motion transfer for video generation. *arXiv preprint arXiv:2404.15789* (2024).
Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. 2024. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21807–21818.
Team Hunyuan3D, Shuhui Yang, Mingxin Yang, Yifei Feng, Xin Huang, Sheng Zhang, Zebin He, Di Luo, Haolin Liu, Yunfei Zhao, et al. 2025. Hunyuan3D 2.1: From Images to High-Fidelity 3D Assets with Production-Ready PBR Material. *arXiv preprint arXiv:2506.15442* (2025).
Xinyi Jing, Tao Yu, Renyuan He, Yu-Kun Lai, and Kun Li. 2025. Frnerf: Fusion and regularization fields for dynamic view synthesis. *Computational Visual Media* (2025).
Bohan Li, Zhuang Ma, Dalong Du, Baorui Peng, Zhujin Liang, Zhenqiang Liu, Chao Ma, Yueming Jin, Hao Zhao, Wenjun Zeng, et al. 2025a. OmniNWM: Omniscient Driving Navigation World Models. *arXiv preprint arXiv:2510.18313* (2025).
Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. 2025b. Cameras as Relative Positional Encoding. *Advances in Neural Information Processing Systems* (2025).
Xiaofan Li, Yifu Zhang, and Xiaoqing Ye. 2024. DrivingDiffusion: Layout-guided multi-view driving scenarios video generation with latent diffusion model. In *European Conference on Computer Vision*. Springer, 469–485.
Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. 2018. Video generation from text. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.
Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. 2025a. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647* (2025).
Shanchuan Lin, Xin Xia, Yuxi Ren, Ceyuan Yang, Xuefeng Xiao, and Lu Jiang. 2025b. Diffusion Adversarial Post-Training for One-Step Video Generation. In *International Conference on Machine Learning*. PMLR, 37959–37974.
Anran Liu, Xiaoxiao Long, Yuan Liu, Ping Luo, and Wenping Wang. 2025. Sem-iNeRF: Camera pose refinement by inverting neural radiance fields with semantic feature consistency. *Computational Visual Media* (2025).
Qihao Liu, Yi Zhang, Song Bai, Adam Kortylewski, and Alan Yuille. 2024. Direct-3d: Learning direct text-to-3d generation on massive noisy 3d data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6881–6891.
Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 2023. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9298–9309.
Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. 2025. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *Computational Visual Media* (2025).
Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and Joao F Henriques. 2021. Keeping your eye on the ball: Trajectory attention in video transformers. *Advances in neural information processing systems* 34 (2021), 12493–12506.
Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.
Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. 2025. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 6121–6132.
Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. 2024. Zeronvs: Zero-shot 360-degree view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9420–9429.
Xuelun Shen, Zhipeng Cai, Wei Yin, Matthias Müller, Zijun Li, Kaixuan Wang, Xiaozhi Chen, and Cheng Wang. 2024. Gim: Learning generalizable image matcher from internet videos. *arXiv preprint arXiv:2402.11095* (2024).
Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792* (2022).
Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhu, Jun Zhang, and Yikai Wang. 2025. Dimensionx: Create any 3d and 4d scenes from a single image with decoupled video diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13695–13706.
Stanislaw Szymanowicz, Jason Y Zhang, Pratul Srinivasan, Ruiqi Gao, Arthur Brussee, Aleksander Holynski, Ricardo Martin-Brualla, Jonathan T Barron, and Philipp Henzler. 2025. Bolt3d: Generating 3d scenes in seconds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 24846–24857.
Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. 2019. FVD: A new metric for video generation. (2019).
Basile Van Hoorick, Rundi Wu, Ege Ozguroglu, Kyle Sargent, Ruoshi Liu, Pavel Tokmakov, Achal Dave, Changxi Zheng, and Carl Vondrick. 2024. Generative camera dolly: Extreme monocular dynamic novel view synthesis. In *European Conference on Computer Vision*. Springer, 313–331.
Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. 2022. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399* (2022).
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. 2025. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314* (2025).
Chaoyang Wang, Ashkan Mirzaei, Vidit Goel, Willi Menapace, Aliaksandr Siarohin, Avalon Vinella, Michael Vasilkovsky, Ivan Skorokhodov, Vladislav Shakhrai, Sergey Korolev, et al. 2025. 4Real-Video-V2: Fused View-Time Attention and Feedforward Reconstruction for 4D Scene Generation. *arXiv preprint arXiv:2506.18839* (2025).
Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. 2023. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571* (2023).
Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. 2020. Imaginator: Conditional spatio-temporal gan for video generation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 1160–1169.
Zhenzhi Wang, Yixuan Li, Yanhong Zeng, Youqing Fang, Yuwei Guo, Wenran Liu, Jing Tan, Kai Chen, Tianfan Xue, Bo Dai, et al. 2024a. Humanvid: Demystifying training data for camera-controllable human image animation. *Advances in Neural Information Processing Systems* 37 (2024), 20111–20131.
Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. 2024b. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*. 1–11.
Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T Barron, and Aleksander Holynski. 2025. Cat4d: Create anything in 4d with multi-view video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 26057–26068.
Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. 2024. Reconfusion: 3d reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 21551–21561.
Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. 2025. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 21469–21480.
FU Xiao, Xian Liu, Xintao Wang, Sida Peng, Menghan Xia, Xiaoyu Shi, Ziyang Yuan, Pengfei Wan, Di Zhang, and Dahua Lin. 2024. 3dtrajmaster: Mastering 3d trajectory for multi-entity motion in video generation. In *The Thirteenth International Conference on Learning Representations*.
Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, et al. 2024. Make-your-video: Customized video generation using textual and structural guidance. *IEEE Transactions on Visualization and Computer Graphics* 31, 2 (2024), 1526–1541.
Dejia Xu, Yifan Jiang, Chen Huang, Liangchen Song, Thorsten Gernoth, Liangliang Cao, Zhangyang Wang, and Hao Tang. 2025. Cavia: Camera-controllable Multi-view Video Diffusion with View-Integrated Attention. In *International Conference on Machine Learning*. PMLR, 69293–69317.

Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. 2024. Camco: Camera-controllable 3d-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509* (2024).

Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. 2024. Direct-a-video: Customized video generation with user-directed camera movement and object motion. In *ACM SIGGRAPH 2024 Conference Papers*. 1–12.

Xiaoda Yang, Jiayang Xu, Kaixuan Luan, Xinyu Zhan, Hongshun Qiu, Shijun Shi, Hao Li, Shuai Yang, Li Zhang, Checheng Yu, et al. 2025. Omnicam: Unified multimodal video generation via camera control. *arXiv preprint arXiv:2504.02312* (2025).

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

Meng You, Zhiyu Zhu, Hui Liu, and Junhui Hou. 2024. Nvs-solver: Video diffusion model as zero-shot novel view synthesizer. *arXiv preprint arXiv:2405.15364* (2024).

Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. 2021. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4578–4587.

Heng Yu, Chaoyang Wang, Peiye Zhuang, Willi Menapace, Aliaksandr Siarohin, Junli Cao, Laszlo A Jeni, Sergey Tulyakov, and Hsin-Ying Lee. 2024. 4real: Towards photorealistic 4d scene generation via video diffusion models. *Advances in Neural Information Processing Systems* 37 (2024), 45256–45280.

Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. 2025a. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*. 100–111.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. 2025b. ViewCrafter: Taming Video Diffusion Models for High-fidelity Novel View Synthesis. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 01 (2025), 1–18.

David Junhao Zhang, Roni Paiss, Shiran Zada, Nikhil Karnad, David E Jacobs, Yael Pritch, Inbar Mosseri, Mike Zheng Shou, Neal Wadhwa, and Nataniel Ruiz. 2025a. Recapture: Generative video camera controls for user-provided videos using masked video fine-tuning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2050–2062.

Mengchen Zhang, Tong Wu, Jing Tan, Ziwei Liu, Gordon Wetzstein, and Dahua Lin. 2025b. Gendop: Auto-regressive camera trajectory generation as a director of photography. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 18229–18239.

Songchun Zhang, Huiyao Xu, Sitong Guo, Zhongwei Xie, Hujun Bao, Weiwei Xu, and Changqing Zou. 2025c. Spatialcrafter: Unleashing the imagination of video diffusion models for scene reconstruction from limited observations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 27794–27805.

Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. 2025. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 10412–10420.

Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. 2022. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018* (2022).

Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. 2025. Stable virtual camera: Generative view synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12405–12414.

Yuan Zhou, Qiuyue Wang, Yuxuan Cai, and Huan Yang. 2024. Allegro: Open the black box of commercial-level video generation model. *arXiv preprint arXiv:2410.15458* (2024).

Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, and Tong He. 2025. Aether: Geometric-aware unified world modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8535–8546.