Beyond Inpainting: Unleash 3D Understanding for Stable Camera-Controlled Video Re-rendering

Dong-Yu Chen , Yixin Guo, Shuojin Yang*
Tai-Jiang Mu , Shi-Min Hu

Key Laboratory of Pervasive Computing, Ministry of Education
Department of Computer Science and Technology, Tsinghua University,
Beijing 100084, China
yangshuojin@tsinghua.edu.cn

Abstract. Camera control has been extensively studied in conditional video generation; however, stable altering camera trajectories while faithfully preserving video content remains a challenging task. The mainstream approach to achieving camera control is warping a 3D representation according to the target trajectory. However, such methods fail to fully leverage the 3D priors of video diffusion models (VDMs) and often fall into the Inpainting Trap, resulting in subject inconsistency and degraded generation quality. To address this problem, we propose DepthDirector, a video re-rendering framework with stable camera controllability. Our key insight is to decompose the conditional injection of video content and camera trajectories by leveraging the depth video from explicit 3D representation as camera-control guidance. Specifically, we design a View-Content Dual-Stream Condition mechanism, enabling VDMs to comprehend camera movements and leverage their 3D understanding capabilities. Additionally, we construct a large-scale multi-camera synchronized dataset named MultiCamWarp Dataset using Unreal Engine 5. Extensive experiments show that DepthDirector outperforms existing methods. Our code and dataset will be publicly available on our project website: https://eleanor6725.github.io/DepthDirector/.

1 Introduction

Recent advances in video generation, particularly video diffusion models (VDMs), have enabled high-quality, content-controllable video synthesis from diverse inputs such as text, images, and videos [16, 44, 58]. Within this rapidly evolving field, camera control is crucial for generating expressive and cinematic videos. It not only helps content creators and cinematographers in framing scene content and dictating global motion, but also craft specific atmospheres and emphasize character emotions. However, stable controllability of camera trajectories while faithfully preserving video content remains a significant challenge.

Currently, the straight-forward solution to achieving camera control is to directly use camera parameters as implicit input conditions [1, 3, 4, 17, 43, 61],

* Corresponding author.

Fig. 1: Example results synthesized by DepthDirector. DepthDirector re-shoots the source video with novel camera trajectories.

encoding position and orientation through ray embeddings [1] or relative pose encoder [3, 4]. These control mechanisms attempt to tame VDMs with camera movements, but fail to bridge the gap between the user-desired camera trajectories and the implicit 3D geometry from video latents. This results in suboptimal physical consistency and camera instability. Worse still, to learn the correspondence between camera parameters and scene geometry, such methods often rely on large-scale, multi-view synchronized rendered datasets [3,4,34,43] (e.g., **136K** videos for ReCamMaster) to implicitly learn 3D consistency. In contrast, we propose a new method that achieves superior performance with merely about **5.8%** of the training cost (using **8K** video sequences).

Another line of methods [15, 20, 36, 42, 51, 59] can achieve stable control of camera trajectories through explicit geometry-based guidance and generate final results via video **inpainting**. However, this mechanism fundamentally caps the quality of the generated content at that of the warped RGB video. As illustrated in Fig. 2, this leads to two primary issues. First, distortion: inevitable inaccuracies in 3D geometry estimated from monocular videos introduce distortions and artifacts into the warped RGB sequence, subsequently causing subject inconsistency and degraded fidelity. Second, baked-in appearance: warped RGB sequence superficially "painting" subject appearance into textures, forcing the network to ignore the view dependent effects. We attribute these artifacts to the **Inpainting Trap**, a shortcut in VDMs that bypasses the understanding of the 3D/4D world and camera transformations. This reduces a VDM with 3D priors into a naive video inpainter. To address this, we focus on designing a new geometry-based guidance mechanism that enables VDMs to internalize camera movements and faithfully preserve source content with physical consistency.

To this end, we present DepthDirector, a framework that fundamentally redefines the camera-controlled video re-rendering paradigm, shifting from pixel-

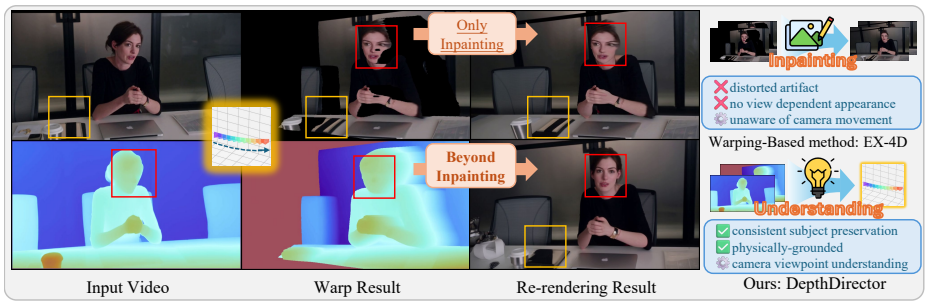

Fig. 2: Limitations of warping-based methods. Reprojected pixels exhibit noisy artifacts due to inaccurate 3D geometry. It leads to unrecoverable distortion to the identity and details of the subject, especially on human faces(See Red Box). Reflections on the table are baked as textures(Yellow Box). These **inpainting** artifacts are attributed to the unawareness of camera movements and video content.

space inpainting to structure-guided synthesis. Our key insight is to **decompose** the conditional injection of video content and camera trajectories. This decomposition forces the network to implicitly fuse both of the conditions and resolve the novel-view extrapolation from its understanding of video content and camera movements. To achieve this, we leverage temporally consistent warped depth rendered from explicit 3D representations under the target view as camera control guidance. This offers three key advantages: (1) It serves as the camera condition without appearance leakage. (2) It provides geometric information assisting VDMs in building a 3D/4D world understanding for input video. (3) It provides users with explicit and easy-to-operate camera controllability.

More specifically, DepthDirector proposes a View-Content Dual-Stream Condition mechanism. A View Stream injects warped depth sequences into the denoised tokens, ensuring precise alignment with desired camera viewpoint. Simultaneously, a Content Stream utilizes the source video injection to faithfully preserve scene structure, semantics, dynamic motions and intrinsic properties. This decomposition ensures the integration of VDMs' world-understanding prior to re-rendering, leading to stable camera controllability and consistent content preservation that goes **beyond inpainting**. As shown in Fig. 2, by introducing the decomposed conditional mechanism, face identity is effectively preserved even in challenging human-centric scenes. In contrast, warping-based methods fuse appearance and camera viewpoint information into a single warped RGB video, leading VDMs to bypass understanding and solely inpaint from this entangled condition. This explains why methods [59] that adopt source video injection still fail to maintain consistent content preservation.

To train our new model, we construct a scalable multi-camera synchronized video dataset construction pipeline using Unreal Engine 5, named MultiCamWarp Dataset. This pipeline automatically renders realistic dynamic videos with ground truth depth in synchronized multi-view cameras. Using this system, we generate **8K** videos shot from **1K** different dynamic scenes. This dataset

enables our model to effectively learn the multi-view consistency and camera controllability with much less training cost. Experimental results show that our method outperforms state-of-the-art approaches and achieves both stable camera controllability and consistent content preservation.

2 Related Works

Video Diffusion Models. Video generation model has evolved from early approaches like Make-A-Video [40] and Gen-1 [37] to more sophisticated models. SVD [6] and VideoCrafter [7, 8] enhanced temporal coherence, while large-scale models such as Hunyuan Video [30], CogVideoX [58], and Wan [44] achieve impressive spatiotemporal consistency and 3D understanding ability.

Video Depth Estimation Recovering depth from monocular video has been widely studied. A line of works focuses on estimating temporally consistent depth without camera estimation. VDA [9] builds upon transformer model and enforce temporal consistency by additional loss. DepthCrafter [21] and ChronoDepth [39] finetune video diffusion models [6] to yield high-quality depth sequences. GeometryCrafter [54] extends this paradigm to predict per-frame point map enabling camera intrinsic estimation. Another line of works [12, 28, 32, 47, 49, 56] directly regresses per-frame depth maps and camera parameters in a feedforward manner. These methods leverage multi-task learning and large-scale datasets to achieve superior performance; however, the estimated depth maps still suffer from low resolution and inaccuracy.

Implicit Camera Conditioned Video Generation In camera-controlled video generation [2, 16, 57, 61, 62], researchers aim to directly incorporate camera parameters into video generation models to control the output viewpoint video. Camera extrinsics are injected into the diffusion model via direct parameter concatenation [50], Plücker Embedding [1, 2, 17, 31, 41], training camera encoder [3, 4, 43] and reference video [34]. Since the correspondence between camera parameters and generated views isn't straight-forward, they [3, 4, 34, 43] typically rely on large-scale multi-view rendered datasets [3,4,34,43] to implicitly learn 3D consistency and suffer from substantial camera control instabilities.

Explicit Camera Conditioned Video Generation To leverage explicit geometry reconstructed from input videos and incorporate precise camera conditioning, existing methods achieve camera control by using an anchor video [15,23, 25,36,51,59,60]. Some earlier works [5,15,53] leverage 3D point tracking [27,52] to inject camera movements, which can cause artifacts if tracking fails. In order to inject 3D information, another line of works [18, 20, 36, 51, 59, 60] uses a warp-and-inpaint strategy. They reconstruct a per-frame 3D representation and repaint on a warped video of target trajectories. However, these methods are prone to produce artifacts from noisy warps. A concurrent work WorldForge [42] designs a training-free method to mitigate the noisy artifacts, while our method introduces a new conditioning mechanism that injects camera viewpoints informed by explicit geometry, enabling higher video generation quality and achieving stable camera control.

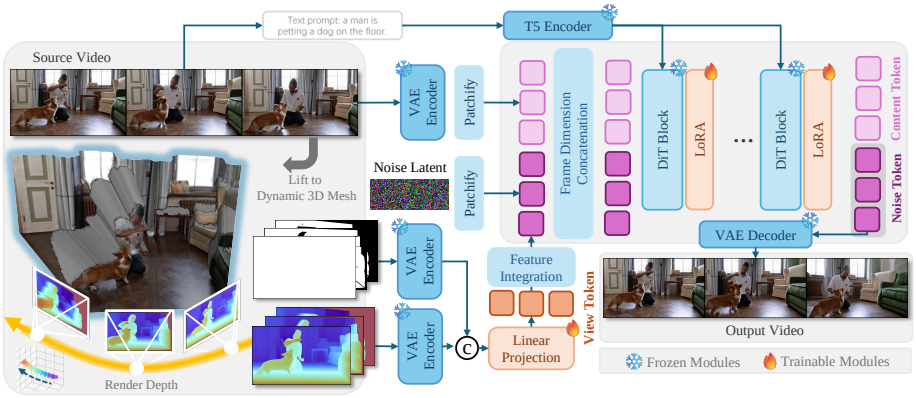

Fig. 3: Architecture overview of DepthDirector. We render depth video and occlusion mask video under target camera viewpoint from an explicit 3D mesh, injecting it into the noise latent by projection and addition. Source video is frame-wise concatenated alongside to provide content reference. Then a LoRA-based video diffusion adapter is trained to generate video following novel camera trajectories.

3 Method

As shown in Fig.3, the goal of our DepthDirector framework is to generate novel-view videos $V_T = \{O_t\}_{t=1}^N$ from an input monocular video $V_S = \{I_t\}_{t=1}^N$ and a target camera trajectory $\{P_t\}_{t=1}^N$. It consists of four key steps: (1) Constructing an explicit dynamic 3D mesh representation and rendering depth of novel views from it. (2) Injecting both input source video and warped depth video as generating conditions through a dual-stream conditional mechanism. (3) Building a multi-camera synchronized video dataset. (4) Using a lightweight video diffusion adapter to train our model.

3.1 Preliminary: Text-to-Video Base Model

Our study is conducted over Wan 2.2 pre-trained text/image-to-video foundation model [46]. This architecture comprises a 3D Variational Auto-Encoder (VAE) [29] for latent space mapping and a series of transformer blocks for sequence modeling. Each basic transformer block consists of a 3D spatial-temporal attention, a cross-attention, and a feed-forward network (FFN). The text prompt embedding c_{text} is obtained by T5 encoder ε_{T5} [35] and injected into the model through cross-attention. We adopt the Rectified Flow [33] framework to train the diffusion transformer, enabling the generation of data sample $\boldsymbol{x}$ from an initial Gaussian sample $\boldsymbol{z} \in \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Specifically, for a data point $\boldsymbol{x}$, we construct its noised version $\boldsymbol{x}_t$ at timestep t as

$$\boldsymbol{x}_t = (1-t)\boldsymbol{x} + t\boldsymbol{z}. \tag{1}$$

The training loss is a simple mean squared error (MSE):

$$\mathcal{L}_{RF}(\theta) = \mathbb{E}_{t,\boldsymbol{x},\boldsymbol{z}} \left\| \boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c}_I, \boldsymbol{c}_{\text{text}}) - (\boldsymbol{z} - \boldsymbol{x}) \right\|_2^2, \tag{2}$$

where the velocity $\boldsymbol{v}_\theta$ is parameterized by the network θ.

3.2 Warped Depth for Camera Control

To provide stable and easy-to-operate camera controllability, we first construct an explicit 3D representation for each frame. Given the input video $V_S = \{I_t\}_{t=1}^N$, we use the 3D foundation model [49] to estimate the camera extrinsics $P_S = \{P_t\}_{t=1}^N$, and per-frame point map $X_S = \{X_t\}_{t=1}^N$. In order that the rendered depth includes less noisy artifacts, we transform the point cloud into a 3D mesh by connecting adjacent pixel points [20]. Then we render the depth $D^r = \{d_t^r\}_{t=1}^N$ and occlusion mask $M^r = \{M_t^r\}_{t=1}^N$ of the 3D mesh under target camera trajectory $P_T = \{P_t\}_{t=1}^N$. The depth of the 3D mesh at a novel view contains occluded parts and out of frame regions, where the occlusion mask M^r will be labeled zero. Therefore, D^r and M^r provide sufficient informations to represent the camera movements and target view layout.

To inject this geometric information into the noise latent $\boldsymbol{x}_t$, the depth should be encoded into the same domain as the video. Therefore, we fit the rendered depth value D^r into RGB domain, and the pretrained video VAE can be leveraged to encode depth video. We first normalize the raw depth values into $[0, 1]$ in log space and then map them into RGB values by a predefined colormap.

3.3 View-Content Dual-Stream Condition

Having the colored depth video D^r and masks M^r from dynamic 3D mesh, as well as the source video V_S, we learn a conditional distribution using a conditional video diffusion model. To decompose the injection of both the viewpoint information from 3D mesh renders and the source video, we propose a dual-stream conditioning strategy, forcing the network to **implicitly fuse** both of the conditions and resolve the novel-view extrapolation from its understanding of video content and camera movements.

The View Stream injects the warped depth sequence into the noise tokens, ensuring precise alignment with the target camera viewpoint. First, we encode D^r and M^r using the VAE encoder $\mathcal{E}$, concatenating them in the channel dimension, and project them into view tokens $\hat{\boldsymbol{x}_v}$ in noise latent space through a linear projector F_c:

$$\hat{\boldsymbol{x}_v} = F_c([\mathcal{E}(D^r), \mathcal{E}(M^r)]_{\text{channel-dim}}). \tag{3}$$

Here, $[\cdot, \cdot]$ denotes the concatenation operation. Then, given noise latent tokens $\boldsymbol{x}_t$ in VAE compressed space, we integrate the view tokens $\hat{\boldsymbol{x}_v}$ with patchified noise latent tokens $\hat{\boldsymbol{x}_t}$ by addition, seamlessly incorporating geometric priors into the video synthesis pipeline:

$$\hat{\boldsymbol{x}_t} = \texttt{patchify}(\boldsymbol{x}_t) + \hat{\boldsymbol{x}_v}. \tag{4}$$

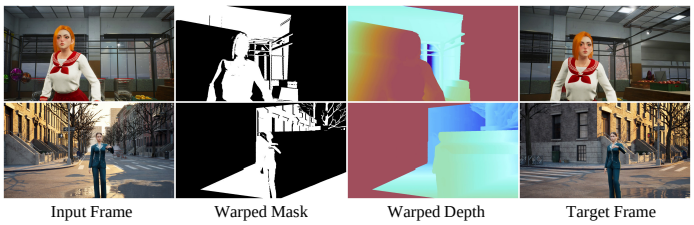

Fig. 4: Illustration of our dataset: MultiCamWarp Dataset.

While view stream resolves target camera viewpoint, it lacks semantic context for reconstructing scene appearance. We inject the source video to provide rich appearance details and dynamic motions. To achieve better synchronization and content consistency with the source video, we concatenate the source video tokens with the target video tokens along the frame dimension:

$$\hat{\boldsymbol{x}}_i = [\texttt{patchify}(\boldsymbol{x}_s), \hat{\boldsymbol{x}}_t]_{\text{frame-dim}}. \tag{5}$$

where $\boldsymbol{x}_s \in \mathbb{R}^{b\times f\times s\times d}$ is the VAE compressed latent of source video, $\hat{\boldsymbol{x}}_i \in \mathbb{R}^{b\times 2f\times s\times d}$ is the input of diffusion transformer. In other words, the input token number is doubled compared to the vanilla video generation process. Then self attention layers switch information among all tokens, ensuring the implicit condition fusion of both stream. This dual-stream design is critical for disentanglement: it relieves the View Stream from infering appearance, allowing it to focus purely on viewpoint control, while the Content Stream strictly preserves original content and temporal fidelity.

3.4 MultiCamWarp Video Dataset

To finetune the model for our task to re-render the source video under a novel camera trajectory, we need a new paired multi-view video dataset. Specifically, the training data should include multiple shots captured in the same scene simultaneously, alongside with the ground-truth depth map to generate view condition. Acquiring such data in real-world scenarios is extremely costly, and publicly available multi-view synchronized datasets [3, 4, 14, 26, 34, 38, 55] are also limited by constrained camera movements and lack of geometry information, making them unsuitable for our task. Therefore we chose to leverage the rendering engine to generate the training data.

We build the entire data generating system in Unreal Engine 5 [13]. Specifically, we first collect multiple 3D environments as "backgrounds". Then, we place animated characters within these environments as the "main subjects" of the videos. We then position multiple cameras facing the subjects and moving along predefined trajectories to simulate the process of simultaneous shooting. This allows us to render datasets with synchronized cameras that include dynamic objects. To scale up the data volume, we construct a set of camera movement

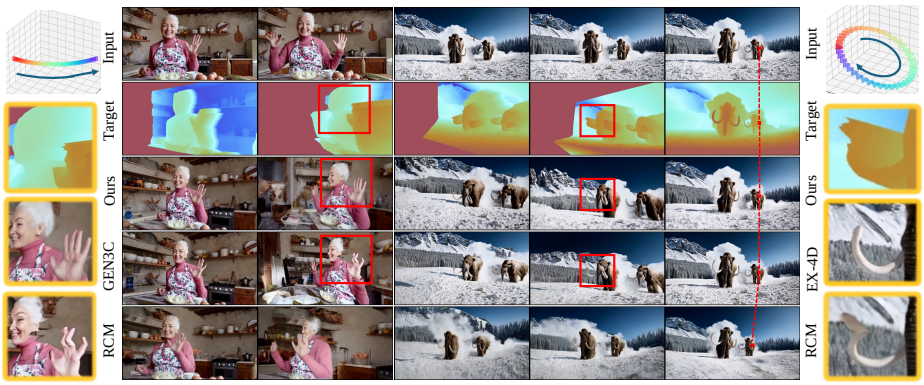

Fig. 5: Qualitative comparison with state-of-the-art methods. Ours DepthDirector achieves both stable camera controllability and content preservation. GEN3C [36] and EX-4D [20] generate distorted artifact due to inaccuracy of depth estimation (see zoom-out details within the yellow border). RCM(ReCamMaster) [3] not only loses consistency at large rotation angle, but also fails to achieve stable camera control. It fails to return back to the original viewpoint in the "orbit" camera trajectory (see red dot on the right).

rules for automatically batch generation of natural and diverse camera trajectories. As shown in Fig. 4, for each dynamic scene, we render RGB value and depth value under 8 random camera trajectories, and build training data-pairs consists of warped depth video, source view video and target view video.

Specifically, during inference, the depth scales of different input videos vary significantly. To mitigate this issue, we introduce random scaling and shifting operations before mapping the raw depth values to the RGB domain, enhancing the model's robustness to depth-scale variations. Moreover, thanks to the dual-stream injection mechanism, the model can adaptively balance geometric and appearance information, effectively alleviating the depth distribution gap between training and inference.

Note that this pipeline can automatically build training videos on a headless Linux server with NVIDIA RTX GPU with a speed of 10 fps, making it easy to scale up. In total, we obtain 8K visually-realistic videos shot from 1K different dynamic scenes in 40 high-quality 3D environments with 8K different camera trajectories. Each video has a resolution of 720×1280 pixels and 81 frames.

3.5 Lightweight Adapter for Video Diffusion

Instead of finetuning the entire VDM, we utilize a Low-Rank Adaptation (LoRA) [19] approach with a small rank. This allows the pre-trained VDM to remain frozen while only updating the adapter parameters.

The training process adheres to the original flow matching objective [11], ensuring rapid convergence and stable performance without requiring extensive

Fig. 6: Qualitative comparison with warping-based methods. Our method preserves facial identity under camera view changes, while other methods produce noticeable distortions.

computational resources. The model is trained to predict flow matching velocity, and the training objective is defined as:

$$\mathcal{L} = \mathbb{E}_{t,\boldsymbol{z}}[\omega(t)||\boldsymbol{v}_\theta(\boldsymbol{x}_t, V_S, D^r, M^r, t; \theta) - (\boldsymbol{z} - \boldsymbol{x})||_2^2]. \tag{6}$$

Here $\boldsymbol{z}$, $\boldsymbol{x}$, $\boldsymbol{x}_t$, $\omega(t)$ denote the ground-truth noise, ground-truth data, noisy latents, and training weight at diffusion timestep t, respectively; V_S is the input video; $\boldsymbol{v}_\theta$ denotes our denoising model with parameters θ.

4 Experiments

4.1 Implementation Details

Our approach builds upon the Wan2.2-TI2V-5B model [44, 46]. The video diffusion backbone remains frozen during training, while our lightweight adapter employs a LoRA rank of 32. The input videos are resized to a resolution of 704×1280 with 81 frames per sequence. The adapter is optimized using the AdamW optimizer with a learning rate of 1×10^{-4}. Our method is implemented on 8 NVIDIA A100 GPUs. Training is completed in 4 days, and inference generates each video in approximately 4 minutes using 50 denoising steps. Please refer to the supplementary material for additional implementation details.

4.2 Evaluation

Metrics and Dataset For evaluation, we randomly sampled 150 in-the-wild web videos from Koala dataset [48], which include diverse dynamic scenes and also challenging human-centric scenes. To evaluate both camera controllability and novel-view synthesis capabilities, we assess our model using 10 diverse

Table 1: Quantitative comparison with state-of-the-art methods. The best numbers are bolded and the second ones are unlerlined.

Method	Camera Accuracy			Identity Preservation		VBench			
	RE ↓	TE ↓	CamMC ↓	RS ↑	IFS ↑	Cs.S.↑	Cs.B.↑	Img.Q↑	M.S.↑
TC [59]	1.464	0.133	2.084	0.567	0.916	93.85	<u>94.75</u>	71.59	98.37
GEN3C [36]	1.397	**0.076**	1.982	0.618	<u>0.943</u>	<u>94.50</u>	93.80	<u>72.04</u>	<u>99.21</u>
EX4D [20]	**1.186**	0.101	**1.687**	<u>0.653</u>	0.925	94.18	94.58	70.65	98.12
Ours(DC)	<u>1.382</u>	<u>0.096</u>	<u>1.962</u>	**0.669**	**0.959**	**94.62**	**94.76**	**72.88**	**99.28**
RCM [3]	3.576	0.180	5.063	0.576	0.932	92.78	91.15	67.27	<u>99.17</u>
Ours(1.3B)	<u>1.349</u>	<u>0.073</u>	<u>1.913</u>	<u>0.653</u>	<u>0.942</u>	<u>93.22</u>	<u>92.32</u>	<u>70.00</u>	98.95
Ours(Full)	**1.010**	**0.053**	**1.434**	**0.663**	**0.954**	**94.27**	**94.46**	**72.90**	**99.21**

trajectories for each video. Our testing suite combines basic rotations and translations (pan, rotate, arc, zoom) with complex paths (orbit, helix, S-curve, and parabola). During these maneuvers, the camera rotates around the main subject at maximum horizontal angles between $\pm 30^\circ$ and $\pm 60^\circ$.

To evaluate camera trajectory accuracy, we employ the state-of-the-art camera parameters estimation model ViPE [22] to extract camera rotation R_i and translation T_i for the i-th frame in the video. We then compute the Rotation Distance (RE), Translation Error (TE), and Camera Motion Consistency (CamMC), following CamI2V [61]. For content preservation evaluation, we assess human identity preservation metrics for all evaluation videos with human faces. We employ ArcFace [10] embedding similarity to assess two key aspects. First, Reference Similarity (RS) calculates the similarity between the first frame of source video and generated frames to evaluate identity preservation. Second, InterFrame Similarity (IFS) quantifies the similarity between consecutive video frames to evaluate the stability of identity features during camera movements. We also evaluate our method on VBench [24] metrics for comprehensive perceptual quality measurement, including: Subject Consistency(Cs.S.), Background Consistency(Cs.B.), Imaging Quality(Img.Q) and Motion Smoothness(M.S.).

Comparison Baselines For comparative evaluation, we involve the state-of-the-art methods capable of camera-controllable video synthesis, including warping-based methods [20,36,59], and implicit camera conditioned methods [3,34]. Other earlier methods [5, 15, 25, 53, 60] are omitted for their relatively inferior performance. All methods are evaluated using identical camera trajectories. Note that the camera trajectories of warping-based methods are defined relative to the camera pose of each frame, a.k.a. relative camera, while those of implicit camera conditioned methods are defined relative to the first frame, a.k.a. absolute camera. To ensure fair comparisons, we conducted our experiments under both settings. For relative camera, we follow warping-based methods to use DepthCrafter [21] and a fixed focal length to extract per-frame depth map, named as **Ours(DC)**. And all warping-based methods receive the same depth estimation results as inputs, eliminating any advantage from differing geometric priors. For absolute

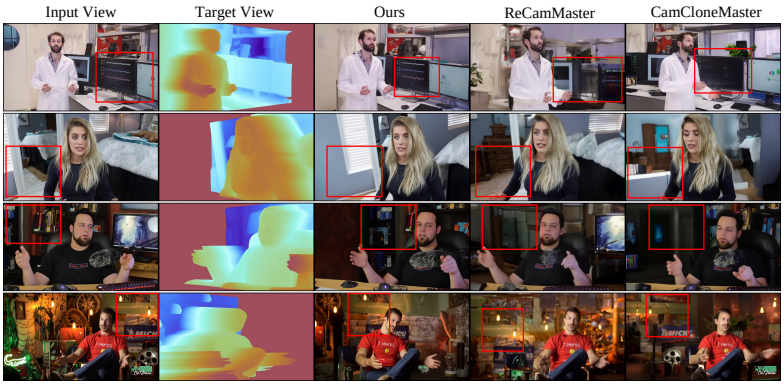

Fig. 7: Qualitative comparison with implicit-controlled methods. Our method generates consistent background, while baseline methods fail to preserve the environment details from input views.

camera, we use $\pi 3$ [49] to estimate input camera parameters and per-frame point maps as in Sec.3.2, named as **Ours(Full)**. CamCloneMaster requires reference videos as guidance, limiting the diversity of input camera trajectories. Therefore, we only evaluate arc trajectories by rendering reference videos with UE5.

Quantitative Comparison Quantitative results are presented in Table.1. Regarding Camera Accuracy, DepthDirector reaches achieves near-zero error, comparable to warping-based methods [20, 36, 59]. Note that the target camera trajectory is defined on the monocular point map which is inheritly inaccurate. The marginal gap in camera accuracy is introduced when DepthDirector refines this imperfect monocular geometry to prioritize visual quality. We contend that this negligible inaccuarcy on metrics is already imperceptible to users. Furthermore, under absolute camera settings, our method significantly outperforms implicit camera conditioned methods [3], demonstrating that DepthDirector provides stable camera controllability. In terms of identity preservation, DepthDirector demonstrates superior performance against all baseline methods, highlighting our method's ability to leverage the 3D understanding of VDMs to generate results with high visual quality. Finally, for comprehensive perceptual quality, DepthDirector achieves the best performance on all VBench metrics.

Qualitative Comparison As shown in Fig. 5, although warping-based methods can strictly follow the target trajectories, they produce distortion artifacts from inheritly inaccurate depth. Implicit methods fail to accurately align with the desired camera views and also lose consistency at large viewpoint change. Compared with them, DepthDirector not only stably aligns with target views but also generates 3D consistent content with high visual quality. This is attributed to our decomposed condition mechanism. Fig. 6 demonstrates that warping-based methods greatly suffer from inaccurate warped video, especially on human faces. However, our method can go **beyond inpainting**, leveraging its 3D understanding ability to generate 3D consistent human subject. As shown in Fig. 7,

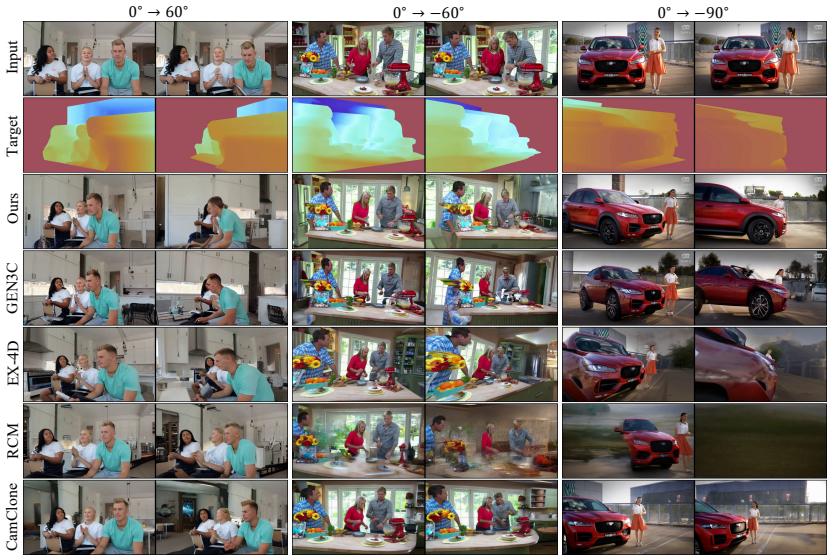

Fig. 8: Qualitative comparison on Extreme Novel View. DepthDirector can achieve high-quality video generation results with stable camera controllability even with camera rotations of up to 90 degrees.

Table 2: Quantitative comparison on Extreme Novel View.

Method	Mild ($0° \rightarrow \pm30°$)			Large ($0° \rightarrow \pm60°$)			Extreme ($0° \rightarrow \pm90°$)		
	Cs.S.↑	Cs.B.↑	Img.Q.↑	Cs.S.↑	Cs.B.↑	Img.Q.↑	Cs.S.↑	Cs.B.↑	Img.Q.↑
EX-4D [20]	94.37	<u>94.58</u>	70.77	90.09	91.85	69.43	86.03	89.02	67.30
GEN3C [36]	94.43	94.05	**72.98**	90.97	<u>92.12</u>	<u>70.75</u>	87.73	90.07	67.64
RCM [3]	93.81	91.64	68.83	89.79	89.54	64.79	83.54	86.31	58.42
CamClone [34]	**95.35**	93.63	70.46	**92.63**	91.83	69.99	**90.04**	<u>90.46</u>	<u>69.32</u>
Ours(Full)	<u>94.70</u>	**95.22**	<u>72.92</u>	<u>91.91</u>	**92.93**	**72.66**	<u>89.92</u>	**90.91**	**71.83**

DepthDirector outperforms the implicit methods on background consistency by a large margin. This is because the warped depth video provides richer geometric cues, helping the model preserve the scene geometry and appearance details.

Comparison on Extreme Novel View We also performed a stress test to demonstrate the robustness of our method under extreme viewpoint change. Using VBench [24] metrics, we assessed performance across three increasingly challenging angular ranges: Mild ($0° \rightarrow \pm30°$), Large ($0° \rightarrow \pm60°$), and Extreme ($0° \rightarrow \pm90°$). We selected validation videos with fixed camera from our full evaluation set to eliminate the difference of target trajectory definition between relative and absolute camera settings. For all compared methods, the target camera originates from the input view and rotates around the main subject at specified horizontal angles. We compare against EX-4D [20], GEN3C [36],

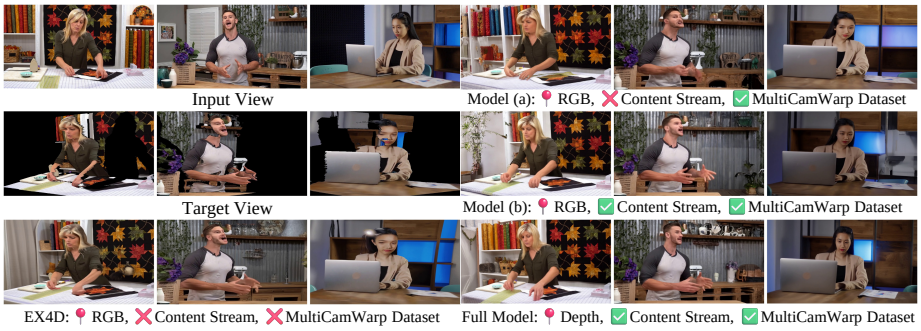

Fig. 9: Ablation Study on Model Design. This demonstrates the importance of decomposed condition injection by depth condition and View-Content Dual-Stream mechanism.

ReCamMaster [3], CamCloneMaster [34]. We use $\pi 3$ [49] to extract depth map for GEN3C and our method, to demonstrate our superior performance given same geometry priors. EX-4D still utilized DepthCrafter due to its strict dependence on high-resolution depth maps. As shown in Fig.8 and Tab.2, our method generates temporally consistent results with high fidelity and visual quality in extreme novel views, demonstrating its robust understanding of target camera movements. Notably, while CamCloneMaster achieves a higher subject consistency score, this is primarily because it fails to reach the desired camera rotation. Excluding this outlier, our method achieves highest score among all metrics. Further visualization and analysis can be found in the Supplementary Materials.

4.3 Ablation Study

Ablation on View-Content Dual-Stream Condition

We extend more experiments to demonstrate how each component contributes to the goal of going beyond inpainting. We focus on metrics demonstrating content consistency, including identity preservation(RS, IFS) and VBench metrics(Cs.S, Cs.B). We train Model (a), which replaces the warped depth video with the warped RGB video and removes the source video concatenation in our View-Content Dual-Stream Condition mechanism, thus maintaining the same as EX-4D [20]. The differences between Model (a) and EX-4D are: (1) different base models, and (2) EX-4D only uses monocular video datasets to train the inpainting ability,

Table 3: Ablation study on Model Design.

Method	Identity		VBench	
	RS ↑	IFS ↑	Cs.S.↑	Cs.B.↑
Model (a)	0.5954	0.9606	94.80	94.57
Model (b)	<u>0.6463</u>	<u>0.9621</u>	<u>95.20</u>	94.43
Model (c)	0.5804	0.9628	95.17	<u>94.66</u>
Full Model	**0.6887**	**0.9661**	**95.29**	**94.66**

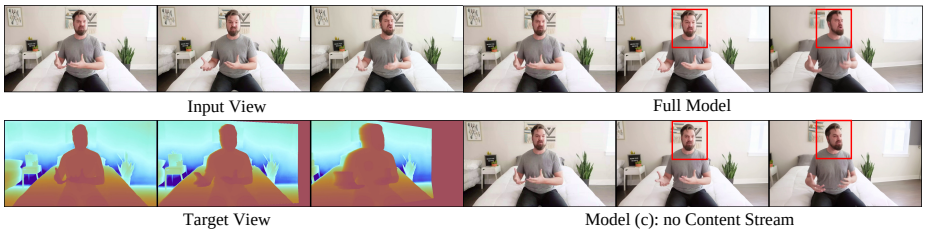

Fig. 10: Ablation Study on Model Design. Model without injecting source video cannot recover the dynamic expression change of source video.

whereas Model (a) uses a multi-camera dataset(our MultiCamWarp Dataset) for training. And Model (b) adds the source video concatenation based on Model (a) and still uses warped RGB video as condition. Model (c) adopt the same view condition stream with full model but without the source video concatenation.

As shown in Fig.9 and Table.3, due to the effectiveness of our View-Content Dual-Stream Condition mechanism and the high-quality multi-camera information of our MultiCamWarp Dataset, the Model (a) and Model (b) both try to learn repainting and correction based on the inaccurate warped video. Training with multi-camera dataset helps Model (a) to learn repainting and correction instead of solely inpainting. The content stream of our Dual-Stream Condition (source video concatenation) further helps Model (b) to maintain content preservation, achieving better repainting results. However, the correction is still hard to learn due to the Inpainting Trap, and these models still generate unnatural artifacts and cannot fully recover the details. Compared with them, our Full Model can generate consistent content with high fidelity, demonstrating the necessity of depth condition. As shown in Fig.10, model(c) without injecting source video infers video content from input image of first frame, and therefore fails to preserve detailed motions, such as facial expressions.

Mechanism Explanation Our core insight is to decompose the injection of video content and camera trajectories, forcing the network to infer novel views through geometric reasoning rather than inpainting. We visualize the attention weights between a noise patch(Query) and source video patches(Keys) in Fig.11. DepthDirector precisely retrieves the corresponding patch from source video by strong attention, demonstrating the geometric reasoning ability. In contrast, Model (b) only slightly attends to a rough region. This means that the source video produces a weak contribution to the noise patch, and this weak correspondence fails to guide the model to learn consistency.

Ablation on Base Model We change our base model from Wan2.2-TI2V-5B [46] to Wan2.1-T2V-1.3B [45], and train at a resolution of 843×480. **Ours(1.3B)** in Table 1 shows that changing to a lighter base model only introduces a slight performance drop. Note that, compared with ReCamMaster which is also trained on 1.3B base model, **Ours(1.3B)** still outperforms in all metrics.

Fig. 11: Attention visualization. Green box marks the query noise patch; the heatmap shows how source video patches attended by it. We normalize weight by a factor(0.04), red means higher. DepthDirector maintains strong and precise attentions to corresponding patch from source video, producing consistent outputs, while the RGB variant produces weak and rough attentions, leading to distorted structures.

5 Conclusion

In this paper, we propose DepthDirector, a framework for reproducing dynamic scenes from input videos under novel camera trajectories. We achieve both stable camera controlability and consistent content preservation. Our key idea is to leverage warped depth video as the camera-view condition, achieving decomposition of conditional injection of video content and viewpoint control. We design a View-Content Dual-Stream Condition mechanism that injects warped depth video to guide viewpoint changes and frame-wise concatenates source video for content reference. Our framework enables video diffusion models to go beyond inpainting and unleash their 3D understanding capabilities, faithfully generating videos of unseen views that are accurately aligned with the target camera trajectory.

Acknowledgements

This work was supported by Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM101), National Natural Science Foundation of China (No. 62220106003, 62572267) and Tsinghua University Initiative Scientific Research Program. We also thank Xiao-Lei Li, Dr. Hangtao Feng and Dr. Wensen Feng for valuable discussions.

References

1. Bahmani, S., Skorokhodov, I., Qian, G., Siarohin, A., Menapace, W., Tagliasacchi, A., Lindell, D.B., Tulyakov, S.: Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22875–22889 (2025)

2. Bahmani, S., Skorokhodov, I., Siarohin, A., Menapace, W., Qian, G., Vasilkovsky, M., Lee, H.Y., Wang, C., Zou, J., Tagliasacchi, A., et al.: Vd3d: Taming large video diffusion transformers for 3d camera control. In: The Thirteenth International Conference on Learning Representations (2025)
3. Bai, J., Xia, M., Fu, X., Wang, X., Mu, L., Cao, J., Liu, Z., Hu, H., Bai, X., Wan, P., et al.: Recammaster: Camera-controlled generative rendering from a single video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14834–14844 (2025)
4. Bai, J., Xia, M., Wang, X., Yuan, Z., Liu, Z., Hu, H., Wan, P., Zhang, D.: Syncammaster: Synchronizing multi-camera video generation from diverse viewpoints. In: The Thirteenth International Conference on Learning Representations (2025)
5. Bian, W., Huang, Z., Shi, X., Li, Y., Wang, F.Y., Li, H.: Gs-dit: Advancing video generation with pseudo 4d gaussian fields through efficient dense 3d point tracking. arXiv preprint arXiv:2501.02690 (2025)
6. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
7. Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., et al.: Videocrafter1: Open diffusion models for high-quality video generation. arXiv preprint arXiv:2310.19512 (2023)
8. Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7310–7320 (2024)
9. Chen, S., Guo, H., Zhu, S., Zhang, F., Huang, Z., Feng, J., Kang, B.: Video depth anything: Consistent depth estimation for super-long videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22831–22840 (2025)
10. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2019)
11. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Proceedings of Machine Learning Research. vol. 235, pp. 12606–12633 (2024)
12. Fang, X., Gao, J., Wang, Z., Chen, Z., Ren, X., Lyu, J., Ren, Q., Yang, Z., Yang, X., Yan, Y., Lyu, C.: Dens3r: A foundation model for 3d geometry prediction. arXiv preprint arXiv:2507.16290 (2025)
13. Games, E.: Unreal engine 5. `https://www.unrealengine.com/en-US/unreal-engine-5` (2022), accessed Feb 7, 2026
14. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., et al.: Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19383–19400 (2024)
15. Gu, Z., Yan, R., Lu, J., Li, P., Dou, Z., Si, C., Dong, Z., Liu, Q., Lin, C., Liu, Z., et al.: Diffusion as shader: 3d-aware video diffusion for versatile video generation control. In: ACM SIGGRAPH 2025 Conference Papers. pp. 62:1–62:12 (2025)

16. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In: The Twelfth International Conference on Learning Representations (2024)
17. He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., Yang, C.: Cameractrl: Enabling camera control for text-to-video generation. arXiv preprint arXiv:2404.02101 (2024)
18. Hou, C., Chen, Z.: Training-free camera control for video generation. In: The Thirteenth International Conference on Learning Representations (2025)
19. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. In: The Tenth International Conference on Learning Representations (2022)
20. Hu, T., Peng, H., Liu, X., Ma, Y.: Ex-4d: Extreme viewpoint 4d video synthesis via depth watertight mesh. arXiv preprint arXiv:2506.05554 (2025)
21. Hu, W., Gao, X., Li, X., Zhao, S., Cun, X., Zhang, Y., Quan, L., Shan, Y.: Depthcrafter: Generating consistent long depth sequences for open-world videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2005–2015 (2025)
22. Huang, J., Zhou, Q., Rabeti, H., Korovko, A., Ling, H., Ren, X., Shen, T., Gao, J., Slepichev, D., Lin, C.H., et al.: Vipe: Video pose engine for 3d geometric perception. arXiv preprint arXiv:2508.10934 (2025)
23. Huang, T., Zheng, W., Wang, T., Liu, Y., Wang, Z., Wu, J., Jiang, J., Li, H., Lau, R., Zuo, W., et al.: Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. ACM Transactions on Graphics **44**(6), 245:1–245:15 (2025)
24. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)
25. Jin, W., Dai, Q., Luo, C., Baek, S.H., Cho, S.: Flovd: Optical flow meets video diffusion model for enhanced camera-controlled video synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2040–2049 (2025)
26. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3334–3342 (2015)
27. Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., Rupprecht, C.: Cotracker: It is better to track together. In: European Conference on Computer Vision. pp. 18–35. Springer (2024)
28. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., et al.: Mapanything: Universal feed-forward metric 3d reconstruction. In: 2026 International Conference on 3D Vision (3DV). pp. 499–509. IEEE (2026)
29. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: The 2nd International Conference on Learning Representations (2014)
30. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)

31. Kuang, Z., Cai, S., He, H., Xu, Y., Li, H., Guibas, L.J., Wetzstein, G.: Collaborative video diffusion: Consistent multi-video generation with camera control. In: Advances in Neural Information Processing Systems. pp. 16240–16271 (2024)
32. Lin, C., Lin, Y., Pan, P., Yu, Y., Hu, T., Yan, H., Fragkiadaki, K., Mu, Y.: Movies: Motion-aware 4d dynamic view synthesis in one second. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 295–306 (2026)
33. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations (2023)
34. Luo, Y., Shi, X., Bai, J., Xia, M., Xue, T., Wang, X., Wan, P., Zhang, D., Gai, K.: Camclonemaster: Enabling reference-based camera control for video generation. In: SIGGRAPH Asia 2025 Conference Papers. pp. 20:1–20:10 (2025)
35. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research **21**(140), 1–67 (2020)
36. Ren, X., Shen, T., Huang, J., Ling, H., Lu, Y., Nimier-David, M., Müller, T., Keller, A., Fidler, S., Gao, J.: Gen3c: 3d-informed world-consistent video generation with precise camera control. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6121–6132 (2025)
37. RunwayML: Gen-1: The next step forward for generative ai (2023), `https://runwayml.com/research/gen-1`, accessed May 7, 2025
38. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1010–1019 (2016)
39. Shao, J., Yang, Y., Zhou, H., Zhang, Y., Shen, Y., Guizilini, V., Wang, Y., Poggi, M., Liao, Y.: Learning temporally consistent video depth from video diffusion priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22841–22852 (2025)
40. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. In: The Eleventh International Conference on Learning Representations (2023)
41. Sitzmann, V., Rezchikov, S., Freeman, B., Tenenbaum, J., Durand, F.: Light field networks: Neural scene representations with single-evaluation rendering. In: Advances in Neural Information Processing Systems. pp. 19313–19325 (2021)
42. Song, C., Yang, Y., Zhao, T., Li, R., Zhang, C.: Taming video models for 3d and 4d generation via zero-shot camera control. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 40352–40363 (2026)
43. Van Hoorick, B., Wu, R., Ozguroglu, E., Sargent, K., Liu, R., Tokmakov, P., Dave, A., Zheng, C., Vondrick, C.: Generative camera dolly: Extreme monocular dynamic novel view synthesis. In: European Conference on Computer Vision. pp. 313–331. Springer (2024)
44. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
45. Wan-AI: Wan2.1-t2v-1.3b: Text-to-video generation model. `https://huggingface.co/Wan-AI/Wan2.1-T2V-1.3B` (2025), accessed Feb 7, 2026
46. Wan-AI: Wan2.2-ti2v-5b: Text-to-video generation model. `https://huggingface.co/Wan-AI/Wan2.2-TI2V-5B` (2025), accessed Feb 7, 2026

47. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5294–5306 (2025)
48. Wang, Q., Shi, Y., Ou, J., Chen, R., Lin, K., Wang, J., Jiang, B., Yang, H., Zheng, M., Tao, X., et al.: Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8428–8437 (2025)
49. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: *pi*3: Permutation-equivariant visual geometry learning. arXiv preprint arXiv:2507.13347 (2025)
50. Wang, Z., Yuan, Z., Wang, X., Li, Y., Chen, T., Xia, M., Luo, P., Shan, Y.: Motionctrl: A unified and flexible motion controller for video generation. In: ACM SIGGRAPH 2024 Conference Papers. pp. 114:0–114:11 (2024)
51. Wang, Z., Cho, J., Li, J., Lin, H., Yoon, J., Zhang, Y., Bansal, M.: Epic: Efficient video camera control learning with precise anchor-video guidance. arXiv preprint arXiv:2505.21876 (2025)
52. Xiao, Y., Wang, Q., Zhang, S., Xue, N., Peng, S., Shen, Y., Zhou, X.: Spatialtracker: Tracking any 2d pixels in 3d space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20406–20417 (2024)
53. Xiao, Z., Ouyang, W., Zhou, Y., Yang, S., Yang, L., Si, J., Pan, X.: Trajectory attention for fine-grained video motion control. In: The Thirteenth International Conference on Learning Representations (2025)
54. Xu, T.X., Gao, X., Hu, W., Li, X., Zhang, S.H., Shan, Y.: Geometrycrafter: Consistent geometry estimation for open-world videos with diffusion priors. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6632–6644 (2025)
55. Xu, Z., Xu, Y., Yu, Z., Peng, S., Sun, J., Bao, H., Zhou, X.: Representing long volumetric video with temporal gaussian hierarchy. ACM Transactions on Graphics **43**(6), 171:1–171:18 (November 2024)
56. Xu, Z., Zhou, H., Peng, S., Lin, H., Guo, H., Shao, J., Yang, P., Yang, Q., Miao, S., He, X., et al.: Towards depth foundation models: Recent trends in vision-based depth estimation. Computational Visual Media **12**(2), 243–271 (2026)
57. Yang, S., Hou, L., Huang, H., Ma, C., Wan, P., Zhang, D., Chen, X., Liao, J.: Direct-a-video: Customized video generation with user-directed camera movement and object motion. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–12 (2024)
58. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. In: The Thirteenth International Conference on Learning Representations (2025)
59. Yu, M., Hu, W., Xing, J., Shan, Y.: Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 100–111 (2025)
60. Zhang, D.J., Paiss, R., Zada, S., Karnad, N., Jacobs, D.E., Pritch, Y., Mosseri, I., Shou, M.Z., Wadhwa, N., Ruiz, N.: Recapture: Generative video camera controls for user-provided videos using masked video fine-tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2050–2062 (2025)
61. Zheng, G., Li, T., Jiang, R., Lu, Y., Wu, T., Li, X.: Cami2v: Camera-controlled image-to-video diffusion model. arXiv preprint arXiv:2410.15957 (2024)

62. Zheng, S., Peng, Z., Zhou, Y., Zhu, Y., Xu, H., Huang, X., Fu, Y.: Vidcraft3: Camera, object, and lighting control for image-to-video generation. arXiv preprint arXiv:2502.07531 (2025)