360Explorer: Exploring 4D Controllable World in Panoramic Videos

Xinhua Cheng[1,3], **Haiyang Zhou**[1], **Wangbo Yu**[1,2], **Tanghui Jia**[1],
Bin Lin[1], **Yunyang Ge**[1], **Weiqi Li**[1], **Li Yuan**[1,2]
[1]Peking University Shenzhen Graduate School
[2]Pengcheng Lab
[3]Rabbitpre Intelligence
chengxinhua@stu.pku.edu.cn

Abstract

We present **360Explorer**, a novel approach for generating 4D controllable panoramic videos conditioned on user-provided 3D instructions for exploring and manipulating dynamic worlds. Compared to existing perspective-based methods struggle to address spatial consistency during camera rotation in place, we introduce the panoramic view in controllable video generation models to inherently maintain the view recall consistency. By introducing dynamic point clouds as the 4D scene representations, 360Explorer unifies the modeling of camera transformations and object movements as incomplete renders to describe precise control instructions in 3D worlds. To tackle the data limitation in acquiring multi-viewpoint panoramic videos, we further propose a reverse warping strategy to construct the training dataset on easily accessible monocular panoramic videos. Extensive experiments demonstrate that 360Explorer achieves superior performance in creating 4D controllable panoramic videos with camera transformation and object movements aligned with diverse provided instructions.

1 Introduction

World Models are causal generative systems (Ha and Schmidhuber 2018; Authors 2024; Parker-Holder et al. 2024; Hu et al. 2023; Micheli, Alonso, and Fleuret 2022) to predict the changes of world states in response to actions, enabling interactive modeling and simulation of dynamic and realistic environments. Considered as an important component of world models, scalable video diffusion models (Brooks et al. 2024; Yang et al. 2024; Lin et al. 2024; Kong et al. 2024; Wan et al. 2025) with conditions have emerged as promising approaches for supporting the exploration and manipulation of dynamic worlds. By accepting spatial 3D instructions including camera trajectories (He et al. 2025a; Hou et al. 2024; Bai et al. 2025a) or motion tracking signals (Wang et al. 2024b; Gu et al. 2025) into temporal input video content, 4D controllable video generation models accomplish remarkable achievements for creating spatial and temporal coherent videos, empowering immersive experiences in dynamic 3D scenarios.

However, we observe that current 4D controllable video generation models encounter two challenges. (1) **Maintaining view recall consistency**: Since perspective-based methods struggle with preserving scene structures in viewpoints at the edge field of view (FoV) when the camera rotates large in place, their generation process leads to uncontrolled new scenes and is unable to maintain spatial consistency during view recalling due to the narrow FoV, as shown in Fig. 1. Thus, introducing a generation target with a wider FoV is beneficial for long-term spatial consistency in large-scale scenes. (2) **Unified modeling precise camera and object transformations**: While many camera-conditioned attempts (He et al. 2025a; Hou et al. 2024; Bai et al. 2025a) inject camera parameters for controlling the relative capture viewpoints, camera conditions fail to describe absolute object movements in the coordination. Other tracking-

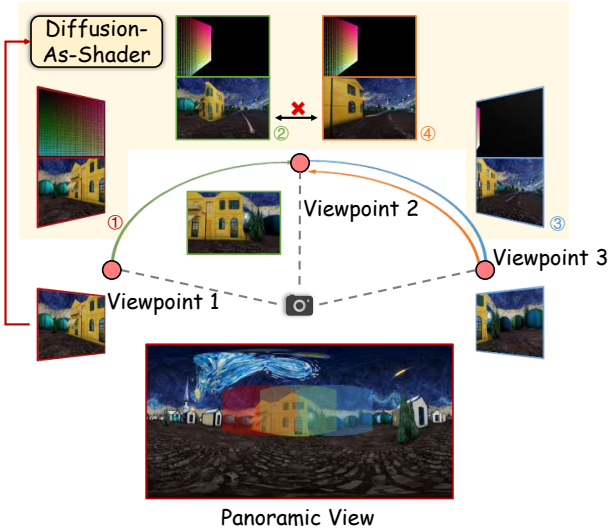

Figure 1: **Conception.** We render the perspective view ① from the panoramic view as the input for the perspective-based method Diffusion-As-Shader (DaS). We observe that DaS fails to maintain view recall consistency (②≠④) with the camera trajectory of rotating in place (①→②→③→④) while the tracking conditions (colored point maps in black background) are same. However, the panoramic view preserves the omnidirectional scene structure and achieves consistent perspective views by rendering.

conditioned attempts (Gu et al. 2025; Li et al. 2024c) leverage motion tracking as the condition for creating aligned controlled results, while tracking drags are approximate directional signals instead of accurate 3D instructions. Hence, adopting properly 4D explicit scene representations to simultaneously describe camera and object transformation as conditions for precisely driving controllable video generation models is desirable for downstream applications.

We present 360Explorer to address the two mentioned challenges: (1) We introduce panoramic views with spherical FoV as the generation targets of the controllable video generation. Compared to perspective views, panoramic views record omnidirectional scene structure, inherently achieving view recall consistency when the camera rotates in place. (2) We introduce dynamic point clouds as the 4D scene representation. By manipulating the explicit points, we obtain incomplete renders as the precise and composed 3D instructions for controlling the video generation model to produce aligned results. Combined with the panoramic views, omnidirectional point clouds capture large-scale scene structures for supporting substantial spatial movements. To overcome the lack of multi-viewpoint panoramic videos capturing the same scenario for training the controllable video generation models, we further design a reverse warping strategy for producing simulated renders describing the inverse camera trajectory and realigned with source monocular videos. To overcome the lack of panoramic video pairs with the same content but captured in different camera positions for training. We conduct experiments for comparing 360Explorer with panoramic-based 3D controllable video generation models including OmniDrag (Li et al. 2024c) and Genex (Lu et al. 2024), and state-of-the-art perspective-based 4D controllable video generation models including TrajectoryCrafter (YU et al. 2025), Diffusion-As-Shader (Gu et al. 2025), demonstrating that 360Explorer producing convincing 4D controlled panoramic video content aligned with diverse provided 3D instructions. Therefore, the main contributions of our work are summarized as:

- We present 360Explorer, a novel framework for generating 4D controllable panoramic videos to overcome the view recall inconsistency of perspective methods.
- We introduce dynamic point clouds to represent the panoramic videos for unified modeling camera transformations and object movements, and describe instructions into aligned incomplete renders for video controlling.
- To address the difficulty of acquiring multi-viewpoint panoramic videos capturing the same scenario, we propose the reverse warping strategy for simulating camera trajectory on arbitrary monocular panoramic videos.
- 360Explorer achieves superior performance than state-of-the-art methods in generating 4D controllable panoramic videos aligned with provided 3D instructions.

2 Related Works

4D Controllable Video Generation. 4D controllable video generation models are considered important for simulating worlds. Camera-controlled video generation models are widely explored by researchers to incorporate camera parameters, including extrinsics (Wang et al. 2024b; Bai et al. 2025a) and Plücker embeddings (He et al. 2025a,b; Xu et al. 2024; Bahmani et al. 2025), into the video generation process for controlling the capture viewpoints. Despite camera-controlled video generation methods achieving success in producing high-quality video with specific camera trajectories, camera parameters cannot describe the absolute object movements in dynamic scenes, limiting their wide application in 3D worlds. Compared to directly injecting camera parameters into the generation process, many attempts introduce 3D-aware representations, including point clouds (Hou et al. 2024; Yu et al. 2024; YU et al. 2025; Wu et al. 2025) and tracking points (Feng et al. 2024; Gu et al. 2025), for simultaneously supporting camera transformations and object movements. Although their methods have proven effective for creating desired video results aligned with diverse 3D conditions, these perspective-based methods are restricted by their limited FoV, resulting in spatial inconsistency with recalled camera trajectories.

Panoramic Video Generation. Panoramic video recently attracts more interest due to its potential applications in immersive experience (Li et al. 2024b; Sun et al. 2023), embodied intelligence, and world simulation (Zhou et al. 2024a). 360DVD (Wang et al. 2024a) firstly explores generating panoramic videos based on given prompts and motion conditions with video generation models. VideoPanda (Xie et al. 2025) further enhances the video generation quality by dividing panoramic views into multiple perspective patches. Imagine360 (Tan et al. 2024) lifts standard perspective video into 360° video, enabling a dynamic scene experience in panoramic views from video anchors. HoloTime (Zhou et al. 2025) animates a static panorama with the image-to-video generation model to obtain camera-fixed 360° videos for Gaussian Splatting reconstruction. Based on the achievement of prompt-conditioned 360° video generation models, OmniDrag (Li et al. 2024c) introduces user-provided drag signals into panoramic video generation for creating videos aligned with both scene and object motion control. Genex (Lu et al. 2024) creates interactive and diverse panoramic videos from the input image to acquire imagined observations for augmenting embodied decision-making. However, both OmniDrag and Genex are panorama-to-video models that struggled to handle dynamic inputs, and their methods only receive approximate directions as conditions instead of precise provided instructions.

3 Methods

In this section, we present the details of 360Explorer pipeline for producing accurate and high-quality 4D world exploration and manipulation in panoramic views. We firstly introduce how we represent the dynamic world in point clouds and describe provided 3D instructions in panoramic worlds, including camera transformation and object movements in Sec. 3.1. We then present how we overcome the lack of multi-viewpoint panoramic videos from accessible monocular panoramic videos in Sec. 3.2. We finally give the details of the model architecture of 360Explorer, and the overview is shown in Fig. 3 in Sec. 3.3.

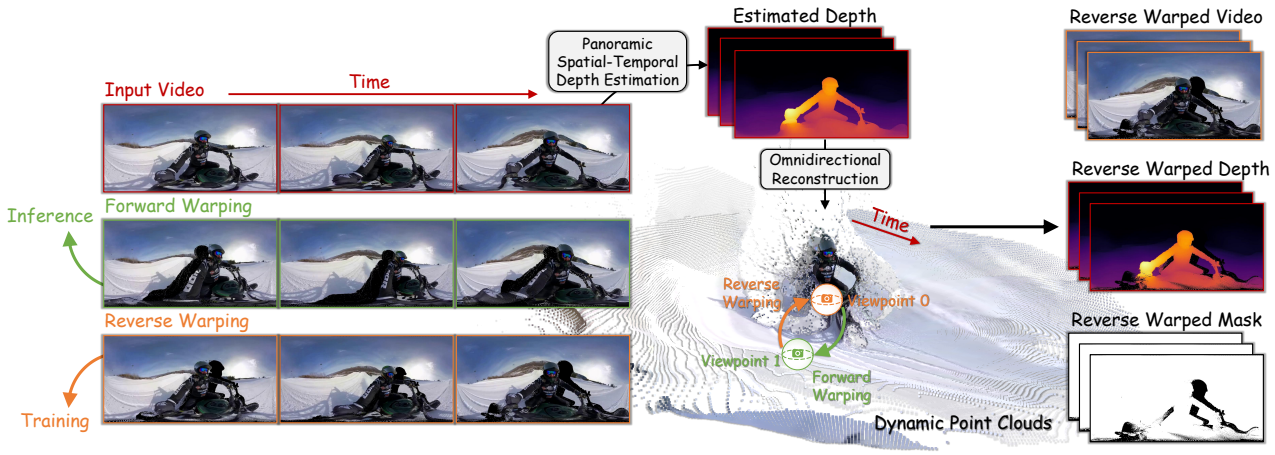

Figure 2: 360Explorer reconstructs dynamic point clouds from input video and corresponding estimated depth. To construct the training pairs under the lack of multi-viewpoint panoramic videos, we propose the reverse warping strategy to estimate the corresponding incomplete renders with the input video at the same viewpoint. In inference stage, we adjust the camera positions or manipulate the object clouds to convert 3D instructions into forward warping renders for producing exploration results.

3.1 Describing 3D Instructions in Renders

Given a panoramic video $\boldsymbol{I}_s = \{\boldsymbol{I}_s^i \in \mathbb{R}^{H\times W\times 3}\}_{i=1}^F$ in equirectangular projection with $H = W/2$, we omnidirectional reconstruct corresponding dynamic point clouds as the explicit representation for world exploration and manipulation, where F, H, W denotes the frame number, height and width of the video respectively. We first estimate the panoramic depth maps $\boldsymbol{D}_s = \{\boldsymbol{D}_s^i \in \mathbb{R}^{H\times W}\}_{i=1}^F$ from the source video. We then reconstruct the source video into the dynamic point clouds $\boldsymbol{P}_s = \{\boldsymbol{P}_s^i \in \mathbb{R}^{H\times W\times 3}\}_{i=1}^F$ according to $\boldsymbol{D}_s$ in the panoramic camera coordinate system:

$$\phi = u/H * \pi, \quad \theta = (1 - v/W) * 2\pi, \tag{1}$$

$$\boldsymbol{d} = [\sin(\phi) * \cos(\theta), \sin(\phi) * \sin(\theta), \cos(\phi)], \tag{2}$$

$$\boldsymbol{P}_s^i[u, v] = \boldsymbol{D}_s^i[u, v] * \boldsymbol{d}, \tag{3}$$

where $\boldsymbol{d} \in \mathbb{R}^3$ is the ray direction of the specific pixel position $[u, v]$, with $u \in \{1, ..., H\}, v \in \{1, ..., W\}$.

Describing 3D Instructions in Incomplete Renders. Since we obtain the dynamic point clouds $\boldsymbol{P}_s$, we introduce the forward warping process for the inference stage to produce desirable exploration and manipulation panoramic videos, as shown in Fig. 2. We precisely describe the 3D instructions as the aligned displacements of the dynamic point clouds $\boldsymbol{P}_s$, and render transformed point clouds $\boldsymbol{P}_f$ for achieving forward warped video $\boldsymbol{I}_f$, depth $\boldsymbol{D}_f$:

$$\boldsymbol{d} = \boldsymbol{P}_f^i[u, v]/||\boldsymbol{P}_f^i[u, v]||_2, \tag{4}$$

$$a = \lfloor \arccos(\boldsymbol{d}[2]/\pi) * H \rfloor, \tag{5}$$

$$b = \lfloor (1 - \arctan(\boldsymbol{d}[0], \boldsymbol{d}[1])/2\pi) * W \rfloor, \tag{6}$$

$$\boldsymbol{I}_f^i[a, b] = \boldsymbol{I}_s^i[u, v], \quad \boldsymbol{D}_f^i[a, b] = ||\boldsymbol{P}_f^i[u, v]||_2. \tag{7}$$

We denote that mask $\boldsymbol{M}_f = \{\boldsymbol{M}_f^i \in \{0, 1\}^{H\times W}\}_{i=1}^F$ represents the incomplete regions in renders caused by the point occlusions are also produced by $\boldsymbol{M}_f^i[a, b] = 1$, where 1 denotes there has corresponding content at specific pixel position $[a, b]$ and 0 denotes voids.

Compared to a camera transformation composed of translation $\boldsymbol{T} \in \mathbb{R}^{F\times 3}$ and rotation $\boldsymbol{R} \in \mathbb{R}^{F\times 3\times 3}$ in perspective view, the panoramic camera transformation is simplified to translation due to the panoramic view inherently supporting perspective sampling at omnidirectional directions. For describing camera translation $\boldsymbol{T}_c$ as incomplete renders, we fix the camera position at the origin of the coordinate system and convert the camera translation into a inverse translation $\boldsymbol{T}_c^{-1} = -\boldsymbol{T}_c$ of the dynamic point clouds for convenience:

$$\boldsymbol{P}_f = \boldsymbol{P}_s - \boldsymbol{T}_c, \tag{8}$$

For describing object movements $\boldsymbol{T}_o$, we firstly leverage video segmentation models SAM2 (Ravi et al. 2024) to determine points belonging to the object, and then apply $\boldsymbol{T}_o$ on belonging points and produce corresponding incomplete renders for world manipulation:

$$\boldsymbol{S} = \text{SAM2}(\boldsymbol{I}_s, [x, y]), \tag{9}$$

$$\boldsymbol{P}_f = [\boldsymbol{P}_s[\boldsymbol{S}] + \boldsymbol{T}_o, \boldsymbol{P}_s[1 - \boldsymbol{S}]], \tag{10}$$

where segmentation mask $\boldsymbol{S} = \{\boldsymbol{S}_i \in \{0, 1\}^{H\times W}\}_{i=1}^F$ and $[x, y]$ is the query pixel position at the first frame.

Panoramic Spatial-Temporal Depth Estimation. Existing panoramic depth estimation approaches (Rey-Area, Yuan, and Richardt 2022; Wang and Liu 2024) are developed for panoramic images, thus failing to maintain temporal coherence across video frames. While perspective video depth estimation methods (Hu et al. 2025; Chen et al. 2025) achieve superior performance in temporal consistency, they struggle to address the geometric properties of equirectangular panoramic videos, as shown in Fig. 4. The geometric properties of panoramas with equirectangular projection ($360° \times 180°$) can be summarized as: (1) Horizontal field of view is cyclic, *i.e.*, the left and right borders should coincide seamlessly; (2) Distortion increases monotonically with latitude away from the equator, becoming extreme near the zenith and nadir. To acquire temporally consistent video

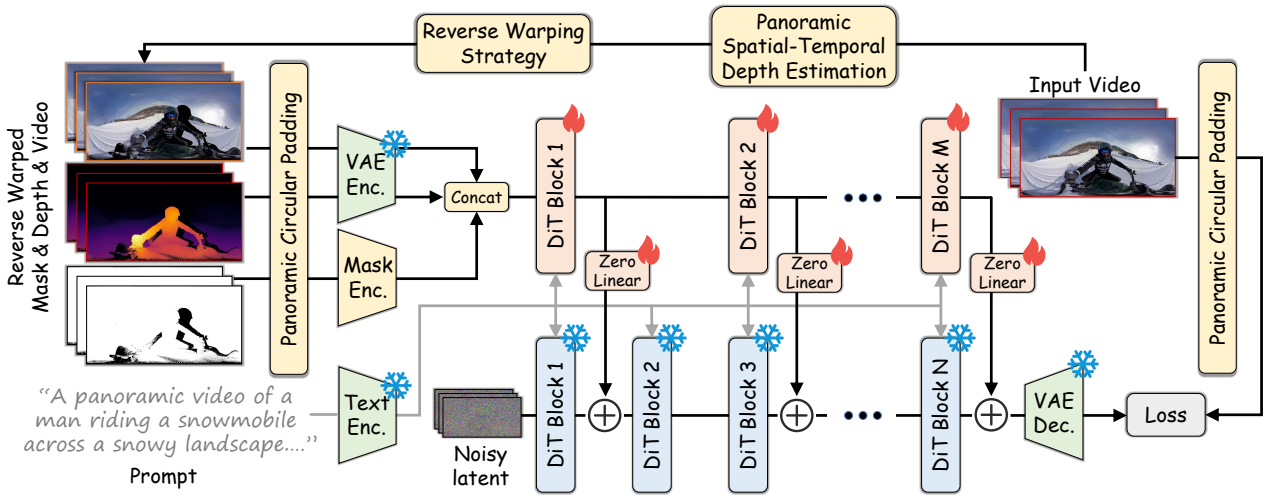

Figure 3: **Architecture of 360Explorer**. 360Explorer is a transformer-based video diffusion model with distributed blocks as the trainable control branch to receive conditions. We encode reverse warped frames and colored depth maps with the frozen VAE and process masks with the reshape operation. Conditions are concatenated in channel dimension for injecting conditions from instruction-aligned renders into the generation model.

depth with panoramic geometry, we propose a panoramic spatial-temporal depth estimation strategy, which refines the depth estimation results predicted by the perspective video method with the panoramic frame depth estimation. For overcoming the horizontal continuous, we expand panoramas $\boldsymbol{I}_s$ in the right border with duplicated regions at the opposite side for panoramic depth prediction. We then linearly blend the depth estimation in repeated parts to guarantee the horizontal continuity and crop the extension regions to obtain the depth $\boldsymbol{D}_{pers}$. For overcoming the latitude distortion, we select the first panoramic frame $\boldsymbol{I}_s^0$ as the reference and achieve the corresponding panoramic metric depth $\boldsymbol{D}_{pano}^0$ predicted by a panoramic depth estimation model as the geometrical reference. To refine the perspective depth, we optimize a patched scaling factor $\{\alpha_j \in \mathbb{R}\}_{j=1}^{H/K}$ to represent the distortion degree in different latitudes, with a panoramic patch composed of adjacent K rows sharing the same factor:

$$\hat{\boldsymbol{D}}_{pers}^0 = \boldsymbol{D}_{pers}^0[Kj : K(j+1)], \tag{11}$$

$$\hat{\boldsymbol{D}}_{pano}^0 = \boldsymbol{D}_{pano}^0[Kj : K(j+1)], \tag{12}$$

$$\mathcal{L}_{depth} = \sum_{j=1}^{H/K} ||\alpha_j \hat{\boldsymbol{D}}_{pers}^0 - \hat{\boldsymbol{D}}_{pano}^0||_2 + \lambda \mathcal{L}_{\text{TV}}(\alpha_j), \tag{13}$$

where $\mathcal{L}_{\text{TV}}$ denotes the total variation regularization (Zhou et al. 2024b; Li et al. 2024a) to impose α_j smoother and λ is the balanced weight. After optimizing the scale factor $\{\alpha_j\}_{j=1}^{H}$, we rescale the $\boldsymbol{D}_{pers}$ per rows along frames to gain $\boldsymbol{D}_s$ with temporal consistency and reasonable panoramic geometry.

3.2 Simulated Renders on Monocular Videos

For training 4D controllable panoramic video generation models, multi-viewpoint panoramic videos capturing the same dynamic scene are expected for constructing video training pairs that include both the content before and after camera translation. However, existing multi-viewpoint datasets (Grauman et al. 2022; Greff et al. 2022; Zheng et al. 2023; Bai et al. 2025a) are collected from perspective views, restricting their usage in panoramic scenarios. Thus we propose a **reverse warping strategy** for creating aligned warped renders on arbitrary monocular panoramic videos with a simulated camera translation trajectory, producing reverse warped video $\boldsymbol{I}_r$, depth $\boldsymbol{D}_r$ and mask $\boldsymbol{M}_r$ for model training, as shown in Fig. 2.

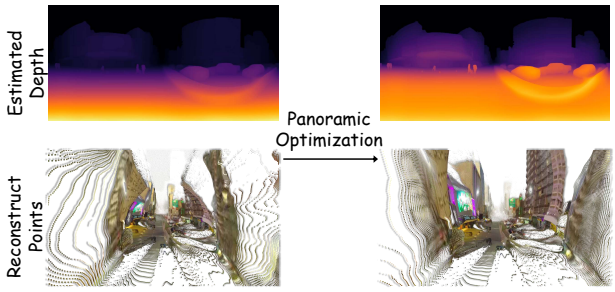

Figure 4: Perspective video estimator fails to tackle panoramas, leading to significant distortion increasing with latitude. By introducing the optimized scaling factor, rescaled estimated depth demonstrates more reasonable geometry.

Concretely, we introduce a forward-reverse process for a camera translation trajectory $\boldsymbol{T}_c$, resulting in the reverse warping simulating the inverse translation $-\boldsymbol{T}_c$ and producing incomplete renders realigned with the viewpoints of the source video $\boldsymbol{I}_s$. Since we obtain the forward warped video $\boldsymbol{I}_f$, depth $\boldsymbol{D}_f$ and mask $\boldsymbol{M}_f$, the dynamic point clouds $\boldsymbol{P}_r$ can be reconstructed from $\boldsymbol{D}_f$ by Eq. 3 with the modification that $[u, v] \in \{(j, k)|\boldsymbol{M}_f^i[j, k] = 1\}$. We then translate

point clouds with $\boldsymbol{P}_r = \boldsymbol{P}_f - (-\boldsymbol{T}_c)$ and achieve reverse renders by Eq. 7 for training. Therefore, we construct the reverse warped video $\boldsymbol{I}_r$ realigned with the source video $\boldsymbol{I}_s$ describing the inverse trajectory $-\boldsymbol{T}_c$.

Training and Inference. In the training stage, we construct training pairs including reverse warped video $\boldsymbol{I}_r$, depth $\boldsymbol{D}_r$, and mask $\boldsymbol{M}_r$ as the condition, and take the source video $\boldsymbol{I}_s$ as the supervision. In the inference stage, we generate forward warped video $\boldsymbol{I}_f$, depth $\boldsymbol{D}_f$, and mask $\boldsymbol{M}_f$ by moving reconstructed dynamic point clouds as model input, and produce high-quality panoramic world exploration and manipulation aligned with provided 3D instructions.

3.3 Renders Conditional Video Generation

By describing provided 3D instructions into aligned incomplete renders, we expect a conditional video generation model that receives warped renders and generates target exploration videos. Thus, our 360Explorer is finetuned from the video editing model VACE (Jiang et al. 2025), which accepts source videos and corresponding masks for outputting inpainted videos. VACE follows a similar design to the ControlNet that modifies the Wan2.1 (Wan et al. 2025) model by duplicating multiple distributed blocks as the conditional branch and introducing zero-initialized linear layers for adding the output feature to corresponding denoising DiT blocks. Before encoding input videos into latents by VAE encoder, VACE employs mask multiplication to decouple videos into reactive frames $\boldsymbol{I}_{re} = \boldsymbol{I} \cdot (1 - \boldsymbol{M})$ containing pixels to be changed and inactive pixels $\boldsymbol{I}_{in} = \boldsymbol{I} \cdot \boldsymbol{M}$. Entire embeddings is finally obtained by concatenating inactivate embedding $\boldsymbol{e}_{in} = \text{VAE}(\boldsymbol{I}_{re})$, reactivate embedding $\boldsymbol{e}_{re} = \text{VAE}(\boldsymbol{I}_{in})$ and mask embedding $\boldsymbol{e}_M = \text{Reshape}(\boldsymbol{M})$ in channel dimension and fed into conditional branch for generating controllable videos.

However, the decoupling process in VACE is not entirely suitable for 360Explorer since warped frames $\boldsymbol{I}_r[u, v] = \boldsymbol{0}$ when $\boldsymbol{M}_r[u, v] = 0$, resulting in reactivate pixels $\boldsymbol{I}_r \cdot (1 - \boldsymbol{M}_r)$ are unable to provide information. Contrary, introducing depth cues for controlling the generation process is crucial for creating geometrically reasonable videos. Thus, we incorporate depth as the condition by normalization and color conversion, leading to colored depth maps $\boldsymbol{C}_r$. Therefore, conditional embedding of 360Explorer is acquired by concatenating frame embedding $\boldsymbol{e}_I = \text{VAE}(\boldsymbol{I}_r \cdot \boldsymbol{M})$, depth embedding $\boldsymbol{e}_D = \text{VAE}(\boldsymbol{C}_r \cdot \boldsymbol{M})$ and mask embedding $\boldsymbol{e}_M = \text{Reshape}(\boldsymbol{M})$ in channel dimension. We finetune the condition branch, including distributed blocks and zero-initialized linear layers, while freezing the encoders and main branch blocks.

Panoramic Circular Padding. We note that horizontal left-right end continuity is an important property for panoramic view, which is crucial for immersive exploration. Thus, we propose a panoramic circular padding strategy that duplicates the border on both sides of the panoramic view to maintain the horizontal continuity. Specifically, we first resize the conditions and supervisions from $H \times W$ to defined panoramic size $H_p \times W_p$, and we then repeat the left and right $(W - W_p)/2$ regions and pad them on the opposite sides, leading to results with $H_p \times W$ size and the aspect ratio larger than $2 : 1$ due to $H_p = W_p/2 < W$ in panoramic view. While existing video generation models support mixed-resolution inference, their pre-trained weights perform best on the recommended aspect ratio, *e.g.*, $4 : 3$ or $16 : 9$, thus we scale the resize the frames along the height to satisfy the original aspect ratio of pre-trained video models. In the inference stage, we execute the inverse operation, including rescaling the frames and cropping the padding pixels, which creates continuous panoramic videos.

4 Experiments

4.1 Implementation Details

Dataset. We adopt the high-quality Filtered-24k subset of the 360-1M (Wallingford et al. 2024) panoramic dataset. By employing the data curation pipeline proposed in Open-Sora Plan (Lin et al. 2024), including jump cutting, motion calculation, OCR cropping, aesthetic and Image quality filtration, we filter and construct 80K paired source video clips and corresponding incomplete renders for training. We further leverage Qwen2.5-VL-7B (Bai et al. 2025b) as the video captioner to generate detailed text prompts.

Training. We finetune 360Explorer based on the pretrained Wan2.1-VACE-14B (Jiang et al. 2025) weights. During training, we set the training video shape as $F \times H \times W = 49 \times 480 \times 832$, and the panoramic size in panoramic circular padding $H_p \times W_p = 400 \times 800$. In our panoramic spatial-temporal depth estimation, we adopt Video-Dpeth-Anything (Chen et al. 2025) and DreamScene360 (Zhou et al. 2024b) as the perspective and panoramic depth predictor, respectively. The learning rate of optimizing scaling factor $\{\alpha_j\}_{j=1}^{H/K}$ is 1×10^{-2} for 1500 iterations, with patch size $K = 8$, balanced weight $\lambda = 0.1$, and the optimization process takes about 3 seconds. The training stage of 360Explorer is conducted on 16 A100 GPUs with the batch size of 1 for each GPU and the learning rate of 5×10^{-5} for 10000 iterations, which takes about 3 days for training.

4.2 Comparison with Panoramic Methods

To verify the effectiveness of 360Explorer for immersive world exploration, we compare 360Explorer with existing controllable panoramic videos generating methods, including Genex (Lu et al. 2024) and OmniDrag (Li et al. 2024c). Since both Genex and Omnidrag only support driving static images as input, we collect 50 realistically captured panoramic images for testing by duplicating the images to obtain still videos as our input for fair comparison. We present the qualitative comparisons in Fig. 5 and quantitative comparisons in Tab. 1, respectively. For evaluation metrics, we use the VBench (Huang et al. 2024) protocol, which includes Subject Consistency and Background Consistency, Motion Smoothness, and Imaging Quality. As shown in Tab. 1, 360Explorer significantly outperforms the compared methods across all video quality metrics, demonstrating state-of-the-art performance in novel panoramic view generation. Moreover, we underline that both Genex and Omnidrag are unable to receive the precise camera movement instructions: Genex can only generate videos in moving forward with unpredictable distances, resulting

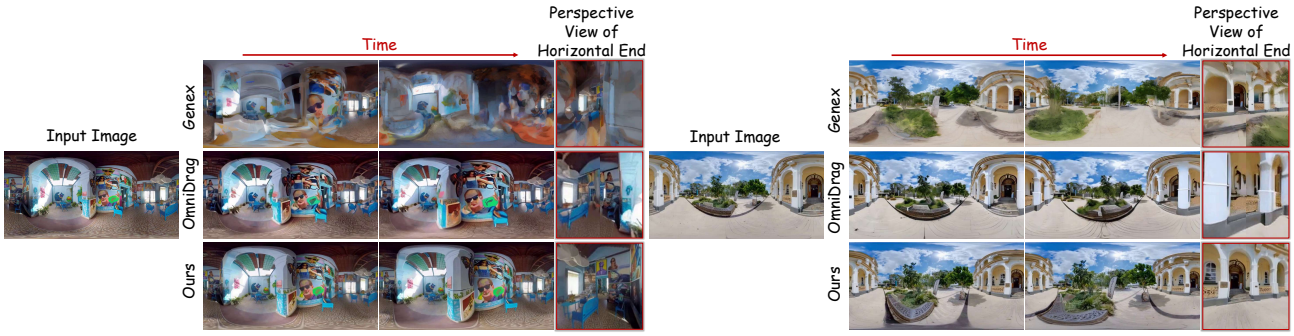

Figure 5: Qualitative comparisons with controllable panoramic video generation methods. The direction of camera translation is *forward*. We underline that both Genex and OmniDrag can only accept static images and generate inaccurate results with approximate directions of camera movement and discontinuous horizontal ends. However, our 360Explorer achieves high-quality exploration videos aligned with precise camera instructions.

Method	VBench ↑				Human Evaluation ↑			
	Subject Consistency	Background Consistency	Motion Smoothness	Imaging Quality	Frame Quality	Temporal Consistency	Horizontal Continuity	Camera Alignment
Genex	0.8508	0.9038	0.9749	0.6838	2.67	2.71	2.58	2.25
OmniDrag	0.9585	0.9670	0.9897	0.7105	3.03	3.62	2.35	2.74
Ours	**0.9664**	**0.9681**	**0.9927**	**0.7407**	**3.82**	**4.07**	**3.78**	**4.12**

Table 1: Quantitative comparison and user study with controllable panoramic video generation methods.

in severe interpenetration issues. OmniDrag accepts drag signals to control camera movement, while the movement directions of generated videos conflict with the provided drags. Furthermore, we specifically sample the perspective view of the horizontal ends from the videos generated by different methods, and the results demonstrate that Genex and OmniDrag suffer from severe horizontal left-right end inconsistency, while our 360Explorer produces panoramic videos with continuous horizontal ends for immersive explorations. To further comprehensively evaluate the quality of controlled panoramic videos, we conducted a user study and collected 20 feedback from humans to investigate the performance of different methods. The human evaluation includes 4 criteria, including panoramic frame quality, cross-frame temporal consistency, horizontal left-right end continuity, and alignment with camera instructions. For each metric, participants are required to select a score from 1 to 5 (5 is the best), and we calculate the average score as the criterion, as shown in Tab. 1. The evaluation results reveal that 360Explorer achieves superior performance, especially on the left-right end continuity and the camera instructions alignment, which is essential for panoramic world exploration.

4.3 Comparison with 4D Controllable Methods

Due to the absence of available 4D controllable panoramic video generation methods currently, we construct compared baselines based on state-of-the-art perspective methods, including TrajectoryCrafter (Traj.Crafter) (YU et al. 2025), Diffusion-As-Shader (DaS) (Gu et al. 2025), and video inpaint model VACE. We firstly obtained incomplete renders

Method	Subject Consistency	Background Consistency	Motion Smoothness	Imaging Quality
VACE†	0.9326	0.9451	0.9881	0.6971
Traj.Crafter†	0.9393	0.9438	0.9899	0.7406
DaS†	0.9410	0.9441	**0.9917**	0.7055
Ours	**0.9440**	**0.9479**	0.9905	**0.7447**

Table 2: Quantitative comparison with 4D controllable video generation methods. † denotes that compared methods are *modified for tackling panoramic views* by receiving depths or renders produced by our proposed panoramic depth estimation and point clouds reconstruction.

for source panoramic videos under different 3D instructions by our proposed panoramic spatial-temporal depth estimation, dynamic point clouds reconstruction, and forward warping process. We then input our panoramic incomplete renders (Videos and masks for Traj.Crafter and VACE, first frame and depths for DaS) into mentioned methods to construct corresponding panoramic variants for overcoming the difficulty that compared methods are not proposed for panoramic views. We collect 50 real-world panoramic videos, and we generate 4 different camera trajectories for each video. We present the qualitative comparisons in Fig. 6 and quantitative comparisons in Tab. 2, which demonstrate that compared variants struggle to tackle novel panoramic views with camera translation and object movements, while 360Explorer produces high-quality panoramic exploration

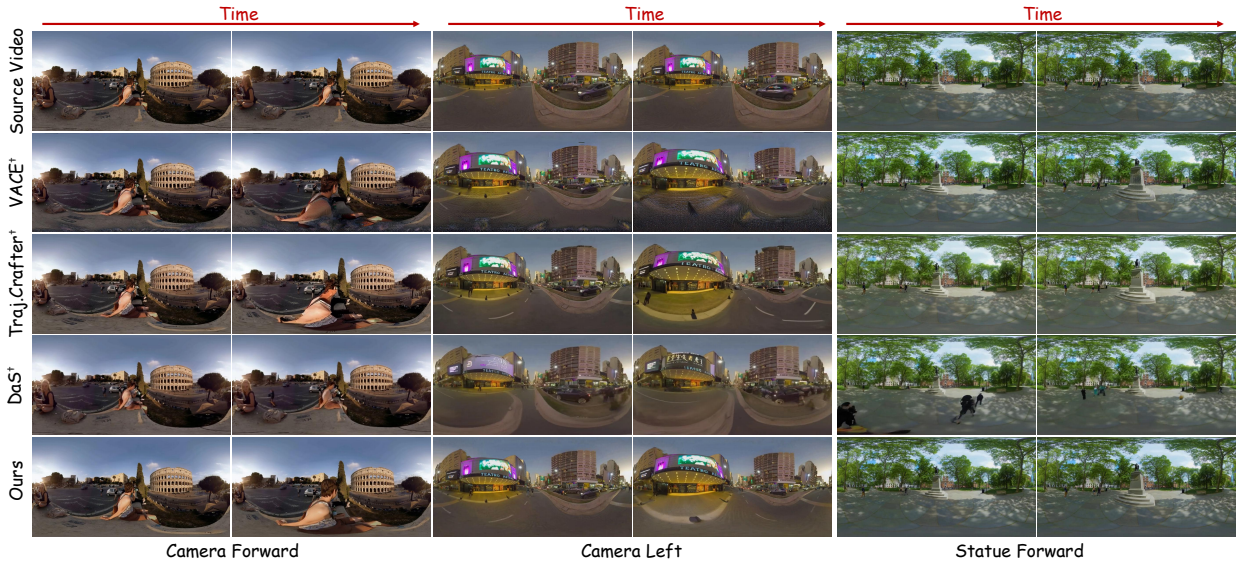

Figure 6: Qualitative comparisons with 4D controllable panoramic video generation methods. † denotes that compared methods are *modified for tackling panoramic views*.

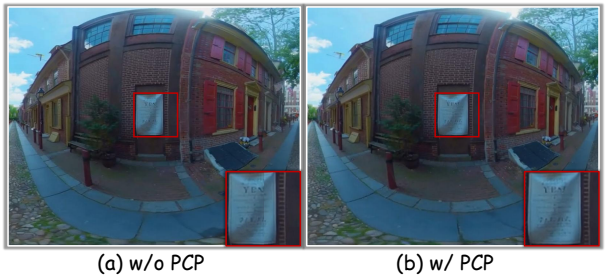

Figure 7: Qualitative ablation study of panoramic circular padding (PCP) strategy. We duplicate the generated frame and concatenate the left and right ends to check continuity.

results aligned with provided 3D instructions. Although DaS outperforms slightly higher at the Motion Smoothness score, we observe from qualitative comparisons that their higher performance is achieved at the expense of alignment with 3D instructions.

4.4 Ablation Study

Architecture. To demonstrate the effectiveness of our design for the 360Explorer Architecture, we conduct quantitative and qualitative ablation studies in Tab. 3 We create two variants to verify that the structure information inherited from both modalities successfully contributes to the conditioned generation results. For "w/o Mask" variant, we set the warped mask $\boldsymbol{M} = \mathbf{1}$, and we reuse reactivate embedding $\boldsymbol{e} = \text{VAE}(\boldsymbol{I} \cdot (1 - \boldsymbol{M}))$ instead of $\boldsymbol{e} = \text{VAE}(\boldsymbol{C} \cdot \boldsymbol{M})$ for "w/o Depth" variant. Qualitative results show that both depth and mask signals promote the creation ability of 360Explorer.

Panoramic Circular Padding. We present the qualitative comparison in Fig. 7 to verify that the proposed panoramic continuous strategy benefits horizontal left-right end continuity for panoramic views. Although our incomplete renders naturally maintain the spatial consistency through the 3D-aware warping conditioned by the reconstructed dynamic point clouds, directly generating exploration videos still leads to significant end inconsistency. For introducing copied padding at the left and right ends during both training and inference stages, 360Explorer performs superior horizontal continuity, which is essential for immersive panoramic explorations.

Method	Subject Consistency	Background Consistency	Motion Smoothness	Imaging Quality
w/o Mask	0.9412	0.9456	0.9896	0.7398
w/o Depth	0.9435	0.9470	0.9903	0.7429
Full	**0.9440**	**0.9492**	**0.9905**	**0.7447**

Table 3: Quantitative ablation study of utilizing depth branch and mask branch in architecture settings.

5 Conclusions

In this work, we propose a novel approach named 360Explorer for creating 4D panorama exploration and manipulation results aligned with provided videos and 3D instructions. By introducing the dynamic point clouds as the scene representation and panoramic views as the generation targets, 360Explorer unified modeling the camera transformations and object movements and inherently maintains the consistency during camera rotation in place. We further design the reverse warping strategy for constructing training pairs to tackle the limitation in acquiring multi-viewpoint panoramic videos. Conducted experiments demonstrate that 360Explorer accomplishes superior performance in generating controlled panoramic videos.

References

Authors, G. 2024. Genesis: A Universal and Generative Physics Engine for Robotics and Beyond.

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

Bai, J.; Xia, M.; Fu, X.; Wang, X.; Mu, L.; Cao, J.; Liu, Z.; Hu, H.; Bai, X.; Wan, P.; et al. 2025a. Recammaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025b. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Brooks, T.; Peebles, B.; Homes, C.; DePue, W.; Guo, Y.; Jing, L.; Schnurr, D.; Taylor, J.; Luhman, T.; Luhman, E.; Ng, C. W. Y.; Wang, R.; and Ramesh, A. 2024. Video generation models as world simulators.

Chen, S.; Guo, H.; Zhu, S.; Zhang, F.; Huang, Z.; Feng, J.; and Kang, B. 2025. Video Depth Anything: Consistent Depth Estimation for Super-Long Videos.

Feng, W.; Liu, J.; Tu, P.; Qi, T.; Sun, M.; Ma, T.; Zhao, S.; Zhou, S.; and He, Q. 2024. I2VControl-Camera: Precise Video Camera Control with Adjustable Motion Strength. *arXiv preprint arXiv:2411.06525*.

Grauman, K.; Westbury, A.; Byrne, E.; Chavis, Z.; Furnari, A.; Girdhar, R.; Hamburger, J.; Jiang, H.; Liu, M.; Liu, X.; et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18995–19012.

Greff, K.; Belletti, F.; Beyer, L.; Doersch, C.; Du, Y.; Duckworth, D.; Fleet, D. J.; Gnanapragasam, D.; Golemo, F.; Herrmann, C.; et al. 2022. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3749–3761.

Gu, Z.; Yan, R.; Lu, J.; Li, P.; Dou, Z.; Si, C.; Dong, Z.; Liu, Q.; Lin, C.; Liu, Z.; et al. 2025. Diffusion as Shader: 3D-aware Video Diffusion for Versatile Video Generation Control. *arXiv preprint arXiv:2501.03847*.

Ha, D.; and Schmidhuber, J. 2018. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31.

He, H.; Xu, Y.; Guo, Y.; Wetzstein, G.; Dai, B.; Li, H.; and Yang, C. 2025a. Cameractrl: Enabling camera control for video diffusion models. In *The Thirteenth International Conference on Learning Representations*.

He, H.; Yang, C.; Lin, S.; Xu, Y.; Wei, M.; Gui, L.; Zhao, Q.; Wetzstein, G.; Jiang, L.; and Li, H. 2025b. Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. *arXiv preprint arXiv:2503.10592*.

Hou, C.; Wei, G.; Zeng, Y.; and Chen, Z. 2024. Training-free camera control for video generation. *arXiv preprint arXiv:2406.10126*.

Hu, A.; Russell, L.; Yeo, H.; Murez, Z.; Fedoseev, G.; Kendall, A.; Shotton, J.; and Corrado, G. 2023. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*.

Hu, W.; Gao, X.; Li, X.; Zhao, S.; Cun, X.; Zhang, Y.; Quan, L.; and Shan, Y. 2025. DepthCrafter: Generating Consistent Long Depth Sequences for Open-world Videos. In *CVPR*.

Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; Chanpaisit, N.; 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.

Jiang, Z.; Han, Z.; Mao, C.; Zhang, J.; Pan, Y.; and Liu, Y. 2025. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*.

Kong, W.; Tian, Q.; Zhang, Z.; Min, R.; Dai, Z.; Zhou, J.; Xiong, J.; Li, X.; Wu, B.; Zhang, J.; et al. 2024. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*.

Li, R.; Pan, P.; Yang, B.; Xu, D.; Zhou, S.; Zhang, X.; Li, Z.; Kadambi, A.; Wang, Z.; Tu, Z.; et al. 2024a. 4k4dgen: Panoramic 4d generation at 4k resolution. *arXiv preprint arXiv:2406.13527*.

Li, W.; Zhao, S.; Chen, B.; Cheng, X.; Li, J.; Zhang, L.; and Zhang, J. 2024b. Resvr: Joint rescaling and viewport rendering of omnidirectional images. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 78–87.

Li, W.; Zhao, S.; Mou, C.; Sheng, X.; Zhang, Z.; Wang, Q.; Li, J.; Zhang, L.; and Zhang, J. 2024c. OmniDrag: Enabling Motion Control for Omnidirectional Image-to-Video Generation. *arXiv preprint arXiv:2412.09623*.

Lin, B.; Ge, Y.; Cheng, X.; Li, Z.; Zhu, B.; Wang, S.; He, X.; Ye, Y.; Yuan, S.; Chen, L.; et al. 2024. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*.

Lu, T.; Shu, T.; Xiao, J.; Ye, L.; Wang, J.; Peng, C.; Wei, C.; Khashabi, D.; Chellappa, R.; Yuille, A.; et al. 2024. Genex: Generating an explorable world. *arXiv preprint arXiv:2412.09624*.

Micheli, V.; Alonso, E.; and Fleuret, F. 2022. Transformers are sample-efficient world models. *arXiv preprint arXiv:2209.00588*.

Parker-Holder, J.; Ball, P.; Bruce, J.; Dasagi, V.; Holsheimer, K.; Kaplanis, C.; Moufarek, A.; Scully, G.; Shar, J.; Shi, J.; Spencer, S.; Yung, J.; Dennis, M.; Kenjeyev, S.; Long, S.; Mnih, V.; Chan, H.; Gazeau, M.; Li, B.; Pardo, F.; Wang, L.; Zhang, L.; Besse, F.; Harley, T.; Mitenkova, A.; Wang, J.; Clune, J.; Hassabis, D.; Hadsell, R.; Bolton, A.; Singh, S.; and Rocktäschel, T. 2024. Genie 2: A Large-Scale Foundation World Model.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Rey-Area, M.; Yuan, M.; and Richardt, C. 2022. 360MonoDepth: High-Resolution 360° Monocular Depth Estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 3752–3762. IEEE.

Sun, X.; Li, W.; Zhang, Z.; Ma, Q.; Sheng, X.; Cheng, M.; Ma, H.; Zhao, S.; Zhang, J.; Li, J.; et al. 2023. OPDN: Omnidirectional position-aware deformable network for omnidirectional image super-resolution. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1293–1301.

Tan, J.; Yang, S.; Wu, T.; He, J.; Guo, Y.; Liu, Z.; and Lin, D. 2024. Imagine360: Immersive 360 Video Generation from Perspective Anchor. *arXiv preprint arXiv:2412.03552*.

Wallingford, M.; Bhattad, A.; Kusupati, A.; Ramanujan, V.; Deitke, M.; Kembhavi, A.; Mottaghi, R.; Ma, W.-C.; and Farhadi, A. 2024. From an image to a scene: Learning to imagine the world from a million 360 videos. *Advances in Neural Information Processing Systems*, 37: 17743–17760.

Wan, T.; Wang, A.; Ai, B.; Wen, B.; Mao, C.; Xie, C.-W.; Chen, D.; Yu, F.; Zhao, H.; Yang, J.; et al. 2025. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*.

Wang, N.; and Liu, Y. 2024. Depth Anywhere: Enhancing 360 Monocular Depth Estimation via Perspective Distillation and Unlabeled Data Augmentation. In Globersons, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J. M.; and Zhang, C., eds., *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Wang, Q.; Li, W.; Mou, C.; Cheng, X.; and Zhang, J. 2024a. 360dvd: Controllable panorama video generation with 360-degree video diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6913–6923.

Wang, Z.; Yuan, Z.; Wang, X.; Li, Y.; Chen, T.; Xia, M.; Luo, P.; and Shan, Y. 2024b. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, 1–11.

Wu, T.; Yang, S.; Po, R.; Xu, Y.; Liu, Z.; Lin, D.; and Wetzstein, G. 2025. Video World Models with Long-term Spatial Memory. *arXiv preprint arXiv:2506.05284*.

Xie, K.; Sabour, A.; Huang, J.; Paschalidou, D.; Klar, G.; Iqbal, U.; Fidler, S.; and Zeng, X. 2025. VideoPanda: Video panoramic diffusion with multi-view attention. *arXiv preprint arXiv:2504.11389*.

Xu, D.; Nie, W.; Liu, C.; Liu, S.; Kautz, J.; Wang, Z.; and Vahdat, A. 2024. Camco: Camera-controllable 3d-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*.

Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. 2024. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*.

YU, M.; Hu, W.; Xing, J.; and Shan, Y. 2025. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. *arXiv preprint arXiv:2503.05638*.

Yu, W.; Xing, J.; Yuan, L.; Hu, W.; Li, X.; Huang, Z.; Gao, X.; Wong, T.-T.; Shan, Y.; and Tian, Y. 2024. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*.

Zheng, Y.; Harley, A. W.; Shen, B.; Wetzstein, G.; and Guibas, L. J. 2023. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19855–19865.

Zhou, H.; Cheng, X.; Yu, W.; Tian, Y.; and Yuan, L. 2024a. Holodreamer: Holistic 3d panoramic world generation from text descriptions. *arXiv preprint arXiv:2407.15187*.

Zhou, H.; Yu, W.; Guan, J.; Cheng, X.; Tian, Y.; and Yuan, L. 2025. HoloTime: Taming Video Diffusion Models for Panoramic 4D Scene Generation. *arXiv preprint arXiv:2504.21650*.

Zhou, S.; Fan, Z.; Xu, D.; Chang, H.; Chari, P.; Bharadwaj, T.; You, S.; Wang, Z.; and Kadambi, A. 2024b. Dreamscene360: Unconstrained text-to-3d scene generation with panoramic gaussian splatting. In *European Conference on Computer Vision*, 324–342. Springer.