DreamWorld: Geometry-Grounded Video Diffusion for 3D-Consistent World Modeling

Haibo Yang[1*], Yang Chen[2], Yingwei Pan[2], Zhineng Chen[3†], Ting Yao[2], and Tao Mei[2]

[1] College of Computer Science and Artificial Intelligence, Fudan University
[2] HiDream.ai Inc.
[3] Institute of Trustworthy Embodied AI, Fudan University
yanghaibo.fdu@gmail.com, {c1enyang, pandy, tiyao, tmei}@hidream.ai, zhinchen@fudan.edu.cn

Abstract. Camera-controlled video diffusion models (VDMs) have recently emerged as powerful world models, enabling users to explore 3D scenes through flexible, user-defined camera trajectories. Nevertheless, current VDMs typically rely on implicit spatiotemporal representations without explicit 3D geometric grounding. Such geometry-agnostic modeling often leads to issues including geometrically implausible structures and cross-view spatial inconsistencies. To alleviate this, we present DreamWorld, a new recipe of world model that novelly bridges the strong spatial structure priors of 3D foundation models with the high-fidelity generative capabilities of video diffusion models for geometry-consistent 3D scene generation. Specifically, given the input image and camera trajectory, DreamWorld first learns a geometry video diffusion model to predict compact geometry features for the target novel views, functioning as explicit structure pivots to reflect the underlying 3D spatial layout. To achieve this, we introduce a distillation paradigm that transfers high-level structural knowledge from a pretrained 3D foundation model to the diffusion model, thereby enabling it to produce geometrically consistent and spatially coherent features. Conditioned on such geometry features, another appearance video diffusion model is then utilized to synthesize the final video, ensuring improved geometric plausibility and cross-view consistency while maintaining high visual fidelity. Extensive experiments demonstrate that DreamWorld outperforms existing methods in visual quality, 3D consistency, and camera controllability. Our project page is available at https://yanghb22-fdu.github.io/DreamWorld.

Keywords: World Modeling · Video Diffusion Model · 3D Geometry

[*] This work was performed when Haibo Yang was visiting HiDream.ai as a research intern.
[†] Corresponding author.

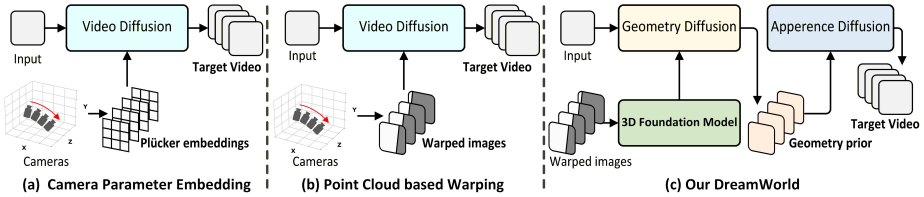

Fig. 1: Comparison between existing camera-controlled VDMs and our DreamWorld.

1 Introduction

Recent advances in video diffusion models (VDMs) and generative foundation models [4, 5, 18, 31, 59, 78, 88, 89] have significantly expanded the capability of generative models from static image synthesis to dynamic world modeling. In particular, camera-controlled VDMs [20, 27, 49, 84] have emerged as a promising paradigm for controllable 3D scene generation, where users can specify arbitrary camera trajectories to render novel views from a single input image. Acting as neural simulators of the physical world, such models aim to construct photorealistic and geometrically consistent 3D environments, unlocking the potential for a wide spectrum of applications spanning virtual reality, cinematic production, and embodied AI.

By leveraging camera parameters or point-cloud-based warped views as conditioning signals, current camera-controlled VDMs exhibit promising visual fidelity and accurate camera controllability [20, 27, 49, 84]. However, despite these successes, existing approaches predominantly rely on implicit spatiotemporal representations to imagine unobserved or occluded scene geometry under novel viewpoints, without explicit 3D geometric grounding. While video diffusion models are trained on large-scale internet video corpora and demonstrate strong capability in generating realistic frames, they are not explicitly constrained to respect 3D spatial structure [14]. As a result, these geometry-agnostic approaches often suffer from inevitable issues in novel view synthesis, such as geometrically implausible structures, distorted object shapes, and cross-view spatial inconsistencies, especially under large viewpoint changes. These issues fundamentally limit the reliability of current world models, particularly in applications requiring strong 3D consistency and structural plausibility.

In parallel to the rapid development of VDMs, 3D foundation models [34, 61, 63, 76] have demonstrated remarkable ability in learning rich structural representations from large-scale 3D data. These models encode strong spatial priors about object shapes and scene layouts, providing geometry-aware features that are robust to viewpoint changes. This observation motivates us to ask a natural question: *Can we incorporate explicit geometric structural priors from 3D foundation models into camera-controlled VDMs to trigger a geometry-grounded video diffusion model for 3D-consistent world modeling?* However, directly integrating 3D foundation models into diffusion-based video generation remains non-trivial. The former are primarily designed for multi-view reconstruction, learning to re-

cover explicit 3D geometry from sparse or dense image observations, whereas the latter operate in a high-dimensional latent video space optimized for spatiotemporal appearance synthesis. Bridging these two paradigms in a principled and effective way is still an open challenge.

In this paper, we present DreamWorld, a new recipe for geometry-consistent world modeling that tightly couples the structural priors of 3D foundation models with the high-fidelity generative power of VDMs. Our launching point is to rethink the role of geometry in current video diffusion. Instead of treating geometry as an implicit byproduct of appearance modeling, DreamWorld explicitly elevates geometry as an intermediate structural pivot within the generation pipeline, using it as a principled bridge to guide the diffusion process toward geometry-consistent 3D scene generation (Fig. 1). To materialize this idea, we first devise a distillation paradigm that transfers high-level structural knowledge from a pretrained 3D foundation model into a diffusion model. Specifically, given an input image and a user-defined camera trajectory, we train a geometry video diffusion model to predict geometry-aware latent representations of the target novel views that are aligned with features extracted from the 3D foundation model. Through this distillation process, the diffusion model inherits strong spatial structure priors, and the learnt geometry features naturally reflect the underlying 3D spatial layout of scenes. After that, we take these geometry features as intermediary structural pivots to guide another appearance video diffusion model, which synthesizes the final novel views. In this way, DreamWorld could elegantly trigger 3D-consistent world modeling in a geometry-appearance disentangled two-stage fashion: the first geometry stage focuses on structural consistency, while the second appearance stage emphasizes fine-grained visual details. Ultimately, by synergizing explicit geometric grounding with generative capacity, DreamWorld achieves a seamless harmony between structural integrity and visual realism, paving a new way for consistent 3D world modeling.

The main contribution of this work is the proposal of the geometry-grounded video diffusion paradigm that explicitly couples structural priors from 3D foundation models with video diffusion models to achieve 3D-consistent world modeling. This also leads to the elegant views of how geometry can serve as a structural pivot to bridge the gap between high-fidelity appearance synthesis and rigorous 3D spatial constraints. Extensive experiments on multiple benchmarks demonstrate that DreamWorld achieves superior performance over existing methods, in terms of visual quality, 3D consistency, and camera controllability.

2 Related Work

Novel View Synthesis. Novel View Synthesis (NVS) aims to generate unseen viewpoints of a scene from limited observations. The evolution of Novel View Synthesis (NVS) can be broadly categorized into three stages. (**i**) Optimization-based paradigms, notably represented by Neural Radiance Fields (NeRF) [1, 40, 56] and further advanced by 3D Gaussian Splatting (3DGS) [25, 29], have achieved unprecedented rendering fidelity across various scales [2, 33, 77, 85]

and dynamic environments [15, 38]. Despite their success, these methods remain fundamentally constrained by the necessity for dense viewpoint observations and computationally intensive per-scene optimization [44, 66]. (**ii**) Feed-forward models [7, 8, 12, 28, 48, 57, 62, 82] attempt to overcome these limitations by directly inferring 3D representations from sparse views. However, due to the inherent ambiguity of reconstructing 3D structures from limited data, these methods are often restricted to specific domains and suffer from structural collapse or blurred textures under extreme viewpoint variations. (**iii**) Generative diffusion models [22, 50, 79] have recently introduced a new paradigm that leverages powerful image or video priors to hallucinate disoccluded content [10, 11, 36, 37, 42, 51, 73–75, 81, 84]. Unfortunately, most existing diffusion-based works remain largely geometry-agnostic, relying on implicit latent consistency without explicit 3D grounding. This often leads to structural drift and spatial inconsistencies during camera exploration. While some efforts [27, 39, 43, 68] incorporate geometric cues via feature conditioning or depth-based warping, they remain sensitive to geometric inaccuracies and misalignment artifacts. To bridge this gap between generative quality and geometric rigor, DreamWorld explicitly integrates structured 3D embedding into video diffusion models, providing reliable geometric guidance for consistent and photorealistic synthesis.

Camera-Controlled Video Generation. With the rapid progress of video diffusion models, enhancing user-controllable camera motion has become increasingly important. Inspired by structural conditioning in text-to-image generation [32, 41, 86], various conditioning signals have been explored in the video domain, including RGB images [3, 69, 71], depth [16, 70], trajectories [45, 80], and semantic layouts [47]. Early camera-control methods incorporate numerical camera parameters or extrinsic matrices as conditioning tokens [19, 37, 53, 58, 65, 67]. AnimateDiff [19] learn motion-specific LoRA modules [24], while MotionCtrl [65] injects extrinsics directly. CamCo [72] and CameraCtrl [20] further adopt Plücker coordinates [54] for ray-aware representation. However, these approaches rely on implicit mappings from low-dimensional camera parameters to high-dimensional pixel dynamics, often limiting precision and generalization. To introduce stronger geometric constraints, recent works employ explicit 3D priors. Training-free approaches [23, 81] leverage depth-based warping to guide denoising, while MultiDiff [42] conditions video diffusion on warped views during training. More recent methods [9, 26, 39, 49, 52] incorporate point clouds or intermediate 3D representations to enhance spatial coherence. Although these strategies improve geometric consistency, they remain sensitive to warping artifacts and geometric noise.

Unlike methods relying on implicit pose embeddings or incomplete pixel-level warping, DreamWorld treats 3D geometry as an intermediate structural pivot. By distilling geometry features from 3D foundation models into the diffusion process, we bridge rigorous 3D structural priors with high-fidelity generative modeling. Our decoupled architecture first reconstructs a complete representation in a robust geometry space, then synthesizes appearance anchored to this consistent scaffold. This strategy effectively alleviates task interference, allowing

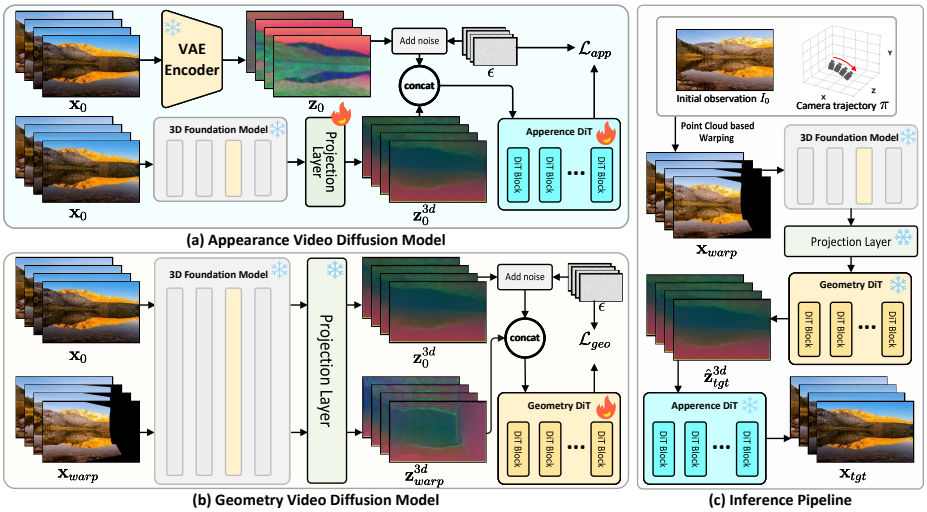

Fig. 2: An overview of our DreamWorld that briges the strong spatial structural priors from 3D foundation models with high-fidelity generative capabilities of video diffusion models for 3D-consistent world modeling. (a)-We train an appearance video diffusion model that synthesizes photorealistic RGB frames conditioned on geometry-aware representations extracted from 3D foundation models. (b)-We train a geometry video diffusion model to predict complete geometry representations for target novel views from incomplete warped observations. (c)-During inference, the geometry model first generates geometry-consistent structural features for the target camera trajectory, which are then provided to the appearance model to render high-fidelity novel-view videos.

each branch to focus on its specialized objective to ensure robust 3D consistency and superior visual quality in world generation.

3 DreamWorld

3.1 Preliminaries

Video Diffusion Models. Generative diffusion models [22, 50, 55] typically model a target data distribution by reversing a gradual noise-adding process. Given clean samples $\mathbf{x}_0 \sim p_{\text{data}}(\mathbf{x})$, the forward process perturbs the data according to $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon$, where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and $\bar{\alpha}_t$ controls the noise level at timestep t. A neural network $\epsilon_\theta(\mathbf{x}_t, t)$ is trained to predict the injected noise by minimizing $\mathbb{E}_{\mathbf{x}_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|_2^2\right]$, so that samples can be generated by iteratively denoising from Gaussian noise.

To facilitate more efficient training and straighter sampling trajectories, recent state-of-the-art video models [31,59,78] increasingly adopt the rectified flow paradigm [17]. This framework redefines the transition between noise and data as a deterministic linear interpolation. In addition, to significantly reduce computational complexity, this process is conducted in a compressed latent space.

Specifically, a raw video $\mathbf{x}_0$ is first encoded by a pre-trained 3D Variational Autoencoder (VAE) $\mathcal{E}(\cdot)$ into a compact representation $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0)$. The forward probability path at timestep $t \in [0, 1]$ is then defined as:

$$\mathbf{z}_t = (1-t)\mathbf{z}_0 + t\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}). \tag{1}$$

The generative process is formulated as an ordinary differential equation (ODE) that transports samples from the noise distribution toward the data manifold:

$$d\mathbf{z}_t = v_\theta(\mathbf{z}_t, t, \mathbf{c})dt, \tag{2}$$

where the velocity field v_θ is parameterized by a neural network conditioned on $\mathbf{c}$, such as text prompts. During training, rectified flow based models regress the constant-velocity vector field by minimizing the following conditional flow matching objective [35]:

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{\mathbf{z}_0, \epsilon, t\sim\mathcal{U}(0,1)} \left[\| v_\theta(\mathbf{z}_t, t, \mathbf{c}) - (\epsilon - \mathbf{z}_0) \|_2^2 \right]. \tag{3}$$

The velocity model v_θ is instantiated as a Transformer-based diffusion model (DiT) [46]. For inference, we can solve the ODE in Eq. 2 via Euler discretization, starting from $t = 1$ to $t = 0$:

$$\mathbf{z}_{t-\Delta t} = \mathbf{z}_t - v_\theta(\mathbf{z}_t, t, \mathbf{c}) \cdot \Delta t, \tag{4}$$

where Δt is the discretization step.

3D Foundation Models. The emergence of 3D foundation models has established a unified paradigm for recovering spatial structures from arbitrary visual inputs. Unlike specialized models designed for isolated 3D tasks, current state-of-the-art 3D foundation models, such as Depth Anything 3 (DA3) [34] and VGGT [61], provide a generalist, feed-forward approach to joint depth and pose estimation across any number of views. Specifically, given an input video sequence $\mathbf{x} = \{I_i\}_{i=1}^L \in \mathbb{R}^{L\times 3\times H\times W}$ with L frames of resolution $H \times W$, the encoder of a 3D foundation model $\mathcal{G}(\cdot)$ processes these frames by alternating between intra-frame self-attention and inter-frame cross-attention across its Transformer layers, yielding multi-scale geometric features $\mathbf{f}_\mathbf{x}^{3d} = \mathcal{G}(\mathbf{x}) \in \mathbb{R}^{L\times K\times d\times h\times w}$ from different depths of the network, where d represents the feature embedding dimension, $h\times w$ denotes the spatial resolution of the feature maps, K is the number of selected intermediate transformer layers. Features from shallower layers typically capture low-level geometric details such as edges and local surface normals, while deeper layers encode high-level structural priors and view-invariant spatial layouts. These rich latent features can be subsequently decoded by task-specific heads to yield various 3D attributes, such as dense depth maps, camera parameters, or point clouds.

3.2 Overview

Problem Formulation. Given a single scene image $I_0 \in \mathbb{R}^{3\times H\times W}$ as the initial observation for world construction, a user-defined camera trajectory $\boldsymbol{\pi} = \{\pi_i\}_{i=1}^L$

and text prompt $\mathcal{T}$, our objective is to synthesize a high-fidelity novel view sequence $\mathbf{x}_{tgt} = \{I_i\}_{i=1}^{L} \in \mathbb{R}^{L\times 3\times H\times W}$. The synthesized sequence should satisfy three key properties: (1) trajectory adherence, such that each frame I_i should strictly follow the predefined camera pose specified by π_i; (2) text alignment, ensuring semantic consistency with the text prompt $\mathcal{T}$; and (3) scene coherence, preserving geometric structure, appearance attributes, and temporal consistency with respect to the depicted scene in the initial observation I_0. Formally, we seek to learn a conditional generative model $\mathcal{F}(\cdot)$ that approximates the conditional distribution $p(\mathbf{x}_{tgt} \mid I_0, \boldsymbol{\pi}, \mathcal{T})$:

$$\mathcal{F} : (I_0, \boldsymbol{\pi}, \mathcal{T}) \mapsto \mathbf{x}_{tgt}. \tag{5}$$

Existing Solutions Existing methods typically rely on pre-trained video diffusion models and formulate this task as a variant of camera-controllable video generation. Approaches like CameraCtrl [20] and CamCo [72] transform camera parameters $\boldsymbol{\pi}$ into pixel-wise spatial Plücker embeddings $\mathbf{P} \in \mathbb{R}^{L\times 6\times h\times w}$ [54], which are then injected into the video diffusion backbone either through an auxiliary encoder or as additional conditional tokens. While this formulation enables the model to condition the generation process on implicit camera representations, it still struggles to achieve precise camera control due to the complicated mapping from numeric camera parameters to high-dimensional video frames, particularly under complex camera motions.

Another line of work [27, 49, 84] instead explicitly reconstructs a 3D point cloud $\mathcal{P}$ from the initial observation and warps it to the target novel views, obtaining a sequence of partial RGB images $\mathbf{x}_{warp} = \{\text{Warp}(\mathcal{P}, \pi_i)\}_{i=1}^{L}$. These warped results are then fed into video diffusion models as explicit camera conditions to inpaint the disoccluded and previously unseen regions, yielding the final synthesized novel views. Although these methods achieve stronger camera controllability, they still face several challenges in maintaining high geometry consistency and visual fidelity. First, while point cloud renderings can accurately represent camera trajectories, they are inevitably plagued by substantial occlusions, unseen areas, and projection artifacts, often resulting in severely degraded visual fidelity. Conditioning on such degraded renderings that contain incomplete structures and distorted artifacts, the video diffusion model must simultaneously infer large missing regions and correct geometric & appearance inaccuracies, which substantially increases the ambiguity of the synthesis process. As a consequence, the model may over-smooth details, hallucinate inconsistent content, or deviate from the intended scene geometry, ultimately limiting the overall generation quality (Fig. 3). Second, in the absence of explicit geometric guidance of unseen regions, the video diffusion model must implicitly infer depth, occlusion relationships, and spatial layout purely from appearance and temporal correlations, which further limits the model's ability to produce geometrically plausible and structurally coherent novel views.

Motivation. By carefully revisiting the limitations of existing methods, we derive two key insights: First, we identify that a reliable conditional signal for video diffusion model to generate novel views must satisfy two critical requirements: (**i**) it should possess strong 3D priors to accurately reflect the underlying

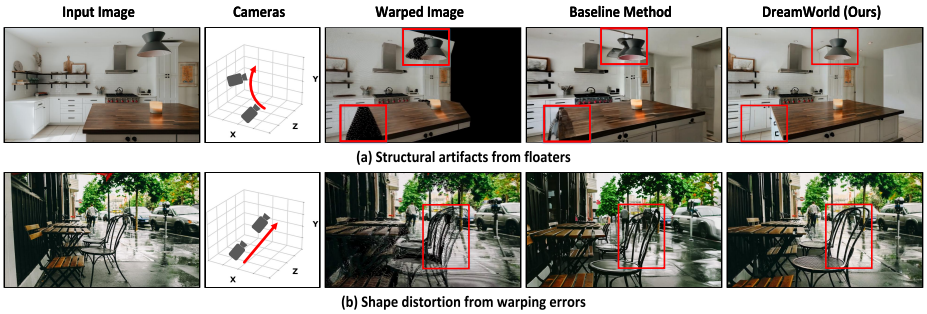

Fig. 3: Video generation directly conditioned on warped partial images often suffers from severe visual artifacts and structural distortions. In contrast, our DreamWorld leverages geometry features from 3D foundation models to first infer complete structural representations and then performs appearance synthesis conditioned on them, ensuring both strong geometric consistency and high visual fidelity.

scene layout, and (**ii**) it must be spatially robust to maintain consistency under extreme camera motion and without inheriting the artifacts of point cloud renderings. We find that the spatial features extracted from 3D foundation models (e.g., DA3 [34]) serve as an ideal structural scaffold. Unlike implicit embeddings that rely on model hallucination or explicit warping that suffers from rendering distortions, these features provide a view-consistent geometric representation that is inherently stable and prior-rich. Second, we emphasize the necessity of a geometry-appearance decoupled generative paradigm. That means the generation process should begin in a compact geometry space, followed by the addition of appearance details, rather than directly operating in the high-dimensional RGB space. Building upon two insights, we decompose the camera-controlled video generation into two stages. In the first stage, we train a geometry video diffusion model to predict DA3 features of the target novel views to serve as a geometry structure pivot for subsequent appearance generation. This model primarily focuses on enforcing cross-view structural consistency and accurate camera adherence, thereby significantly reducing the complexity of the learning objective. In the second stage, we train an appearance video diffusion model to synthesize the final novel views conditioned on the first stage's DA3 features, which focuses exclusively on synthesizing high-fidelity textures and photorealistic details. By decoupling geometry and appearance in this manner, we can pursue stronger cross-view consistency and improved visual fidelity.

3.3 Geometry-Grounded Video Generation

We first introduce our devised appearance video diffusion model $\mathcal{F}_{app}(\cdot)$, which synthesizes photorealistic RGB frames conditioned on geometry representations derived from 3D foundation models. Given a raw video $\mathbf{x}_0$, the pre-trained VAE encoder $\mathcal{E}(\cdot)$ first maps it into a compact VAE latent $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0)$. In addition,

as described in Sec. 3.1, we extract its multi-scale geometric features $\mathbf{f}^{3d}_{\mathbf{x}_0}$ from a pre-trained 3D foundatation model, which captures both fine-grained geometry details and global scene layouts. We then fine-tune a pre-trained video diffusion model to predict denoised VAE latents conditioned on geometric features.

Note that it is not trivial to directly inject these geometry features into the video diffusion backbone. The representations extracted from 3D foundation models are typically high-dimensional, and their spatial structures are not naturally aligned with the latent space of the VAE. To address this issue, we introduce a lightweight MLP-based compression module $\mathcal{M}(\cdot)$ to project geometry features $\mathbf{f}^{3d}_{\mathbf{x}_0}$ into a diffusion-friendly representation $\mathbf{z}^{3d}_0 = \mathcal{M}(\mathbf{f}^{3d}_{\mathbf{x}_0})$. To incorporate this geometry guidance, $\mathbf{z}^{3d}_0$ is concatenated with the noisy video latents $\mathbf{z}_t$ along the channel dimension to fed into the denosing network of the video diffusion model. This allows our appearance video diffusion model $\mathcal{F}_{app}(\cdot)$ to offload the burden of spatial reasoning and focus exclusively on appearance fidelity. During training, we optimize $\mathcal{F}_{app}(\cdot)$ with parameter ϕ by minimizing a flow-matching objective coupled with a KL-regularization:

$$\mathcal{L}_{app} = \mathbb{E}_{\mathbf{z}_0,\epsilon,t}\left[\left\|v_\phi(\mathbf{z}_t, t, \mathbf{z}^{3d}_0, I_0, \mathcal{T}) - (\epsilon - \mathbf{z}_0)\right\|^2_2\right] + \lambda\mathcal{L}_{\text{KL}}, \tag{6}$$

where $\mathcal{L}_{\text{KL}}$ is the KL divergence objective to enforce the projected 3d feature $\mathbf{z}^{3d}_0$ to follow a Gaussian distribution. After training, our devised appearance video diffusion model can denoise VAE latents conditioned on 3D-aware representations, and the resulting latents are then decoded into RGB frames via VAE decoder. In this way, we can faithfully render high-fidelity pixels by leveraging geometry-aware structural features as conditioning signals, ensuring that the synthesized frames remain consistent with the underlying scene geometry.

3.4 Geometry Feature Generation

While the aforementioned appearance video diffusion model $\mathcal{F}_{app}(\cdot)$ ensures high-fidelity pixel synthesis under a given geometry structure prior, a critical challenge then arises: during inference, how to predict such geometry structural guidance $\hat{\mathbf{z}}^{3d}_0$ for target novel views, only given the initial observations I_0 and user-defined camera trajectory $\boldsymbol{\pi}$. We achieve this by introducing a geometry video diffusion model $\mathcal{F}_{geo}(\cdot)$. Specifically, we first project the initial observation I_0 to target viewpoints using point cloud based warping, yielding a sequence of incomplete warped frames $\mathbf{x}_{warp}$ corresponding to the camera trajectory $\boldsymbol{\pi}$. We then feed these warped frames into the encoder of the pre-trained 3D foundation model $\mathcal{G}(\cdot)$ and the learnt projection module $\mathcal{M}(\cdot)$ in the apperance video diffusion model to extract incomplete geometry features $\mathbf{z}^{3d}_{warp} = \mathcal{M}(\mathbf{f}^{3d}_{\mathbf{x}_{warp}})$, where $\mathbf{f}^{3d}_{\mathbf{x}_{warp}} = \mathcal{G}(\mathbf{x}_{warp})$. For the ground-truth video $\mathbf{x}_0$ corresponding to the warped frames $\mathbf{x}_{warp}$, we perform the same process as above to obtain ground-truth geometry features $\mathbf{z}^{3d}_0$. Then, we train our geometry video diffusion model $\mathcal{F}_{geo}(\cdot)$ in the geometry feature space, to predict complete geometry features for the target novel views by conditioning on the incomplete geometry features $\mathbf{z}^{3d}_{warp}$.

Formally, the denoising neural network of our geometry video diffusion model is trained through feature-level flow-matching objective:

$$\mathcal{L}_{geo} = \mathbb{E}_{\mathbf{z}_0^{3d},\epsilon,t}\left[\left\|v_\theta(\mathbf{z}_t^{3d}, t, \mathbf{z}_{warp}^{3d}, I_0, \mathcal{T}) - (\epsilon - \mathbf{z}_0^{3d})\right\|_2^2\right], \tag{7}$$

where $\mathbf{z}_t^{3d}$ is the noised geometry feature at timestep t and θ is the parameter of the denoising neural network of $\mathcal{F}_{geo}(\cdot)$. Once trained, the geometry video diffusion model can inherit the geometry structure knowledge from 3D foundation models to synthesize the complete geometry features through a conditional denoising process.

3.5 Inference Pipeline

During inference, we generate the target novel views in a two-stage manner. Given the initial observation I_0 and user-defined camera trajectory $\boldsymbol{\pi}$, we first employ our geometry video diffusion model $\mathcal{F}_{geo}(\cdot)$ to predict a sequence of geometry representations $\hat{\mathbf{z}}_{tgt}^{3d}$ corresponding to the target viewpoints. Then, the predicted geometry features $\hat{\mathbf{z}}_{tgt}^{3d}$ are provided as geometry structural conditioning signals to the appearance video diffusion model $\mathcal{F}_{app}(\cdot)$, which performs conditional denoising in the VAE latent space and subsequently decodes the denoised latents into the final RGB frames $\mathbf{x}_{tgt}$.

4 Experiments

4.1 Experimental Settings

Datasets. We train DreamWorld on a mixed dataset of RealEstate10K [90] and SpatialVID [60]. RealEstate10K provides indoor sequences with diverse camera trajectories and accurate pose annotations. To enhance scene diversity and real-world generalization, we incorporate SpatialVID [60], a large-scale in-the-wild dataset enriched with explicit geometric annotations. To ensure training stability, we curate a high-quality subset of $223K$ clips from both datasets, filtered based on motion magnitude, pose reliability, and visual consistency. This combined training set covers a balanced distribution of indoor and outdoor environments, establishing a robust foundation for spatially grounded world modeling.

Implementation Details. The appearance video diffusion model $\mathcal{F}_{app}(\cdot)$ and the geometry video diffusion model $\mathcal{F}_{geo}(\cdot)$ are initialized from pre-trained Wan 2.1 model [59]. The training process is decoupled into two progressive stages. In the first stage, we jointly optimize the appearance video diffusion model and the projection module $\mathcal{M}$ to establish a reliable mapping from the geometric manifold to the visual space. To bridge the gap between training and inference, we introduce small perturbations to the geometry features during this stage, which improves the model's robustness to the generated structural scaffolds. In the second stage, we freeze the projection module $\mathcal{M}$ and train the geometry video diffusion model to perform feature-level distillation and structural completion.

Table 1: Quantitative comparison of single-image novel view synthesis on RealEstate10K [90] and Tanks-and-Temples dataset [30].

Method	RealEstate10K Dataset											
	Easy Set						Hard Set					
	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	R_{dist} ↓	T_{dist} ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	R_{dist} ↓	T_{dist} ↓
MotionCtrl [65]	14.12	0.538	0.413	88.13	4.372	8.778	13.05	0.495	0.453	97.66	7.116	9.484
CameraCtrl [20]	19.09	0.706	0.325	42.18	1.254	2.796	17.15	0.554	0.395	75.76	1.875	3.104
ViewCrafter [84]	19.21	0.719	0.315	39.75	0.713	2.086	17.76	0.677	0.376	52.86	1.161	2.682
Gen3C [49]	20.73	0.726	0.279	33.97	0.556	1.582	18.04	0.681	0.334	46.67	0.838	2.347
DreamWorld (Ours)	**23.15**	**0.781**	**0.249**	**25.18**	**0.301**	**1.112**	**20.94**	**0.739**	**0.267**	**31.16**	**0.521**	**1.318**
Method	**Tanks-and-Temples Dataset**											
	Easy Set						Hard Set					
	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	R_{dist} ↓	T_{dist} ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	R_{dist} ↓	T_{dist} ↓
MotionCtrl [65]	14.88	0.554	0.497	80.73	7.447	8.603	13.45	0.524	0.514	89.55	8.731	9.402
CameraCtrl [20]	15.79	0.587	0.396	71.31	1.656	2.913	14.01	0.532	0.493	82.01	1.975	3.124
ViewCrafter [84]	18.11	0.671	0.327	45.67	0.966	2.175	17.41	0.645	0.353	57.76	1.437	2.891
Gen3C [49]	18.95	0.683	0.308	41.95	0.884	1.677	17.95	0.659	0.331	48.42	1.266	2.510
DreamWorld (Ours)	**21.66**	**0.721**	**0.256**	**29.85**	**0.377**	**1.214**	**19.59**	**0.704**	**0.283**	**36.95**	**0.584**	**1.337**

Table 2: Quantitative comparison on the WorldScore benchmark [14]. The performance of all baseline methods in this table are directly taken from [13].

Method	3D Consist.	Photo Consist.	Style Consist.	Cam. Ctrl.	Obj. Ctrl.	Cont. Align.	Subj. Qual.	Avg. ↑
WonderWorld [83]	82.85	67.86	55.79	**92.32**	47.63	**79.09**	**69.03**	<u>70.65</u>
AETHER [91]	79.84	58.68	72.09	57.44	52.26	28.06	41.11	55.64
Uni3C [6]	78.59	85.48	88.32	62.94	45.83	47.40	57.00	66.51
Voyager [27]	56.00	80.68	72.89	45.92	<u>57.69</u>	48.36	44.74	58.04
FantasyWorld [13]	<u>83.31</u>	<u>86.11</u>	**94.22**	57.05	34.46	38.45	57.40	64.43
DreamWorld (Ours)	**84.96**	**92.70**	<u>92.98</u>	<u>81.97</u>	**63.11**	<u>50.37</u>	<u>59.17</u>	**75.04**

For both stages, we employ the AdamW optimizer with a constant learning rate of 5×10^{-5} and a global batch size of 32. Each stage is trained for $15,000$ iterations to ensure stable convergence. All training sequences consist of 81-frame clips processed at a resolution of 832×480.

Baselines and Metrics. Following [84], we first use RealEstate10K [90] and Tanks and Temples [30] as test set for zero-shot novel view synthesis. We compare our method against two primary technical paradigms: implicit embedding-based approaches, such as MotionCtrl [65], CameraCtrl [20], and explicit warping-based frameworks, including ViewCrafter [84] and Gen3C [49]. Evaluation is conducted on both easy and hard subsets using PSNR, SSIM [64], LPIPS [87], FID [21], and pose accuracy metrics (R_{dist}, T_{dist}). To further evaluate holistic world modeling, we follow [13] and benchmark our performance on the photorealistic subset of the WorldScore [14] dataset. We compare DreamWorld against state-of-the-art 3D world models, including WonderWorld [83], AETHER [91], Uni3C [6], Voyager [27], and FantasyWorld [13]. This multi-axis assessment provides a rigorous test for the robustness of our decoupled paradigm by evaluating camera control, 3D consistency, and content alignment in complex scenarios.

Fig. 4: Qualitative comparison results. Unlike baseline methods prone to warping artifacts (green boxes) or structural reasoning failures (red boxes), DreamWorld ensures stable, photorealistic views via a reliable geometric scaffold.

4.2 Quantitative Results

Table 1 summarizes the quantitative results for single-image novel view synthesis, where DreamWorld achieves state-of-the-art performance across all evaluation metrics and benchmarks. Remarkably, compared to the best competitor, Gen3C [49], DreamWorld significantly boosts the Hard Set PSNR from 18.04 to 20.94 on RealEstate10K and from 17.95 to 19.59 on Tanks-and-Temples. The consistently lower FID values and reduced pose errors further confirm that DreamWorld excels in generating photorealistic sequences that also strictly adhere to intended camera trajectories. Unlike implicit parameter embedding methods [20,65] that lack explicit 3D grounding, or warping-based models [49,84] that are inherently prone to projection artifacts and disocclusion holes, DreamWorld

leverages a geometry video diffusion model to synthesize a holistic structural scaffold. By bypassing the inherent brittleness of raw pixel-level warping, our proposed framework achieves a seamless harmony between rigorous 3D consistency and high-fidelity appearance synthesis.

Table 2 presents the quantitative comparison on the WorldScore benchmark [14], where DreamWorld achieves the highest Average score, establishing a new state-of-the-art among current world models. A key observation is the significant gap in consistency metrics between DreamWorld and its competitors. For instance, while WonderWorld [83] achieves strong camera controllability, its Photo Consistency (67.86) lags far behind DreamWorld's 92.70, illustrating that existing models often sacrifice visual stability for motion accuracy. In contrast, by treating geometry as a decoupled structural pivot, DreamWorld resolves the "task entanglement" inherent in joint architectures like Voyager [27] and FantasyWorld [13]. Specifically, Voyager's reliance on static priors leads to substantial structural drift (56.00 in 3D Consistency), whereas FantasyWorld's unified optimization forces a compromise between rigorous 3D grounding and visual realism. Our decoupled paradigm ensures the appearance stage focuses on high-fidelity synthesis anchored in a reliable 3D layout, resulting in superior consistency. These results validate that disentangling geometry from appearance effectively alleviates the structural-appearance conflict, providing a more stable and 3D-consistent world generation than existing frameworks.

4.3 Qualitative Results

Fig. 4 presents the qualitative comparison results. WonderWorld [83] exhibits the most severe structural degradation. Due to its reliance on a sparse surfel-based initialization optimized primarily for low-latency single-view rendering, it suffers from significant geometry missing when moving to novel viewpoints. Meanwhile, Gen3C [49] relies on explicit point cloud based warping, but its quality is strictly bounded by the initial 3D prior. As highlighted in the green boxes, stretched areas and holes from the warping stage lead to blurry results and distorted shapes in the final frames. Notably, FantasyWorld [13] attempts to address these issues by jointly modeling video and 3D tasks. While this unified approach improves view alignment, it leads to "task entanglement", where the two objectives interfere with each other. As shown in the red boxes, this conflict makes it difficult for the model to reason about complex structures like thin tree branches, resulting in broken or floating artifacts. In contrast, DreamWorld consistently produces more stable and coherent novel views with plausible geometry and realistic textures. By decoupling structural completion from appearance generation, each branch can focus its specialized capacity on a single objective. Specifically, the geometry branch is dedicated to robust 3D reasoning without the burden of texture synthesis, while the appearance branch focuses entirely on rendering photorealistic details. Our approach achieves both superior structural integrity and higher appearance quality, outperforming existing frameworks even in scenes with intricate details.

Table 3: Ablation study on the core design choices in DreamWorld. Experiments are conducted on the easy set of RealEstate10K [90].

Method	Condition Space		Geometry & Appearance		PSNR ↑	SSIM ↑	LPIPS ↓
	VAE	3D	Joint	Decoupled			
Baseline	✓	✗	✗	✗	19.47	0.721	0.287
DreamWorld w/o Geometry Diffusion	✗	✓	✗	✗	20.05	0.724	0.283
DreamWorld w/o Geo-App Decoupled	✗	✓	✓	✗	21.87	0.752	0.256
DreamWorld (Ours)	✗	✓	✗	✓	**23.15**	**0.781**	**0.249**

4.4 Ablation Study

In an effort to study the effectiveness of each development in our DreamWorld, we conducted a systematic ablation study on the RealEstate10K dataset. Table 3 demonstrates the quantitative results of several ablated runs. The baseline of our DreamWorld directly encode the warped partial images into the VAE latent space as conditions to synthesize novel views, without explicit geometric modeling, which leads to the lowest performance.

Effect of Geometry Diffusion. *DreamWorld w/o Geometry Diffusion* replaces the generative geometry prediction with raw incomplete geometry features extracted from the 3D reconstruction pipeline. While using 3D features provides a better geometric conditions than the VAE-based Baseline (20.05 vs. 19.47 PSNR), the results remain sub-optimal. These limitations arise from the inherent sparsity of warped features, where disocclusion artifacts inevitably lead to structural gaps. Without a geometry completion stage to fill these missing regions, the appearance model must simultaneously infer both the missing structure and textures, which introduces significant synthesis ambiguity. In contrast, our full model employs geometry diffusion to reconstruct a complete structural pivot from sparse warped features, achieving the best performance (23.15 PSNR). This geometry diffusion stage enables the model to generatively complete missing regions, ensuring that the subsequent appearance synthesis is grounded in a stable and geometrically consistent foundation.

Effect of Decoupled Geometry-Appearance Training. *DreamWorld w/o Geo-App Decoupled* jointly predicts geometry and appearance in a single objective, similar to the design of existing world models [13]. Compared to the baseline, this approach achieves a further performance gain (21.87 PSNR) because it introduces explicit 3D features that provide much-needed spatial awareness for scene synthesis. However, the performance gains remain limited as the model must simultaneously optimize for two fundamentally different goals: 3D structural reasoning and high-fidelity rendering. This coupling leads to task entanglement, which increases optimization difficulty and results in unstable structural predictions for complex scenes. In contrast, DreamWorld separates these two objectives into a "geometry-then-appearance" pipeline. By first generating a consistent geometric scaffold and then synthesizing appearance conditioned on it, the model can focus on structural completion and visual rendering in a more

specialized manner. This decoupled design alleviates task interference and leads to more stable and geometry-consistent world generation.

Overall, these ablations validate the key design principle of our proposed DreamWorld: explicitly generating geometry as an intermediate structural pivot and decoupling it from appearance modeling is essential for achieving reliable 3D-consistent world modeling.

5 Conclusions

In this paper, we presented DreamWorld, a geometry-grounded framework for controllable world modeling with video diffusion models. We identify the lack of explicit structural grounding as a key reason of 3D inconsistency in existing camera-controlled VDMs. To address this, we propose a decoupled geometry–appearance paradigm, where a dedicated geometry branch performs structural completion in a compact geometry-aware latent space, and an appearance branch synthesizes high-fidelity visuals conditioned on the completed structural anchors. This "geometry-then-appearance" design avoids task entanglement and ensures that the model achieves both stable geometric consistency and high-quality rendering. Extensive experiments on RealEstate10K, Tanks and Temples, and WorldScore demonstrate that DreamWorld consistently improves visual quality, cross-view consistency, and camera controllability over prior methods, validating the importance of explicit 3D grounding in generative world models.

Limitations and Broader Impact. DreamWorld focuses on static scenes and may falter with complex dynamics or severe occlusions. Extending geometry grounding to spatiotemporal representations is a vital future direction. While consistent world generation empowers simulation and embodied AI, its high realism necessitates ethical vigilance to prevent misuse in deceptive content creation.

Acknowledgement: This work was supported by the Key Science & Technology Project of Anhui Province No. 202523o09050002 and the Science and Technology Commission of Shanghai Municipality under Grant 25511104000.

References

1. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-NeRF: a multiscale representation for anti-aliasing neural radiance fields. In: ICCV (2021)
2. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-NeRF 360: unbounded anti-aliased neural radiance fields. In: CVPR (2022)
3. 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)
4. Cai, Q., Chen, J., Chen, Y., Li, Y., Long, F., Pan, Y., Qiu, Z., Zhang, Y., Gao, F., Xu, P., et al.: Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. arXiv preprint arXiv:2505.22705 (2025)

5. Cai, Q., Chen, J., Gao, C., Gong, Z., Li, Y., Pan, Y., Peng, Y., Qiu, Z., Yu, K., Zhang, Y., et al.: Hidream-o1-image: A natively unified image generative foundation model with pixel-level unified transformer. arXiv preprint arXiv:2605.11061 (2026)
6. Cao, C., Zhou, J., Li, S., Liang, J., Yu, C., Wang, F., Xue, X., Fu, Y.: Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. In: SIGGRAPH Asia (2025)
7. Charatan, D., Li, S.L., Tagliasacchi, A., Sitzmann, V.: PixelSplat: 3D Gaussian splats from image pairs for scalable generalizable 3D reconstruction. In: CVPR (2024)
8. Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., Su, H.: MVSNeRF: fast generalizable radiance field reconstruction from multi-view stereo. In: ICCV (2021)
9. Chen, L., Zhou, Z., Zhao, M., Wang, Y., Zhang, G., Huang, W., Sun, H., Wen, J.R., Li, C.: Flexworld: Progressively expanding 3d scenes for flexiable-view synthesis. arXiv e-prints (2025)
10. Chen, Y., Pan, Y., Li, Y., Yao, T., Mei, T.: Control3d: Towards controllable text-to-3d generation. In: ACM MM (2023)
11. Chen, Y., Pan, Y., Yang, H., Yao, T., Mei, T.: Vp3d: Unleashing 2d visual prompt for text-to-3d generation. In: CVPR (2024)
12. Chen, Y., Xu, H., Zheng, C., Zhuang, B., Pollefeys, M., Geiger, A., Cham, T.J., Cai, J.: MVSplat: efficient 3D Gaussian splatting from sparse multi-view images. In: ECCV (2024)
13. Dai, Y., Jiang, F., Wang, C., Xu, M., Qi, Y.: Fantasyworld: Geometry-consistent world modeling via unified video and 3d prediction. In: ICLR (2026)
14. Duan, H., Yu, H.X., Chen, S., Fei-Fei, L., Wu, J.: Worldscore: A unified evaluation benchmark for world generation. In: ICCV (2025)
15. Duan, Y., Wei, F., Dai, Q., He, Y., Chen, W., Chen, B.: 4d-rotor gaussian splatting: towards efficient novel view synthesis for dynamic scenes. In: SIGGRAPH (2024)
16. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models. In: ICCV (2023)
17. 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: ICML (2024)
18. Gao, C., Ding, L., Cai, X., Huang, Z., Wang, Z., Xue, T.: Controllable first-frame-guided video editing via mask-aware loRA fine-tuning. In: ICLR (2026)
19. Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In: ICLR (2024)
20. He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., Yang, C.: Cameractrl: Enabling camera control for text-to-video generation. In: ICLR (2024)
21. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017)
22. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
23. Hou, C., Wei, G., Zeng, Y., Chen, Z.: Training-free camera control for video generation. In: ICLR (2024)
24. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: ICLR (2022)
25. Huang, B., Yu, Z., Chen, A., Geiger, A., Gao, S.: 2D Gaussian splatting for geometrically accurate radiance fields. In: SIGGRAPH (2024)

26. Huang, J., Yang, Y., Yang, B., Ma, L., Ma, Y., Liao, Y.: Gen3r: 3d scene generation meets feed-forward reconstruction. In: CVPR (2026)
27. 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. TOG (2025)
28. Jin, H., Jiang, H., Tan, H., Zhang, K., Bi, S., Zhang, T., Luan, F., Snavely, N., Xu, Z.: Lvsm: A large view synthesis model with minimal 3d inductive bias. In: ICLR (2025)
29. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. TOG (2023)
30. Knapitsch, A., Park, J., Zhou, Q.Y., Koltun, V.: Tanks and temples: Benchmarking large-scale scene reconstruction. TOG (2017)
31. 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)
32. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: CVPR (2023)
33. Li, Z., Müller, T., Evans, A., Taylor, R.H., Unberath, M., Liu, M.Y., Lin, C.H.: Neuralangelo: High-fidelity neural surface reconstruction. In: CVPR (2023)
34. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. In: ICLR (2026)
35. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: ICLR (2023)
36. Liu, F., Sun, W., Wang, H., Wang, Y., Sun, H., Ye, J., Zhang, J., Duan, Y.: ReconX: reconstruct any scene from sparse views with video diffusion model. arXiv preprint arXiv:2408.16767 (2024)
37. Liu, R., Wu, R., Hoorick, B.V., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: CVPR (2024)
38. Luiten, J., Kopanas, G., Leibe, B., Ramanan, D.: Dynamic 3D Gaussians: tracking by persistent dynamic view synthesis. In: 3DV (2024)
39. Ma, B., Gao, H., Deng, H., Luo, Z., Huang, T., Tang, L., Wang, X.: You see it, you got it: Learning 3d creation on pose-free videos at scale. In: CVPR (2025)
40. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)
41. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: AAAI (2024)
42. Müller, N., Schwarz, K., Rössle, B., Porzi, L., Bulò, S.R., Nießner, M., Kontschieder, P.: Multidiff: Consistent novel view synthesis from a single image. In: CVPR (2024)
43. Müller, N., Schwarz, K., Rössle, B., Porzi, L., Bulo, S.R., Nießner, M., Kontschieder, P.: Multidiff: Consistent novel view synthesis from a single image. In: CVPR (2024)
44. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. TOG (2022)
45. Niu, M., Cun, X., Wang, X., Zhang, Y., Shan, Y., Zheng, Y.: Mofa-video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model. In: ECCV (2024)
46. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV (2023)

47. Peruzzo, E., Goel, V., Xu, D., Xu, X., Jiang, Y., Wang, Z., Shi, H., Sebe, N.: Vase: Object-centric appearance and shape manipulation of real videos. arXiv preprint arXiv:2401.02473 (2024)
48. Ren, X., Lu, Y., Liang, H., Wu, Z., Ling, H., Chen, M., Fidler, S., Williams, F., Huang, J.: Scube: Instant large-scale scene reconstruction using voxsplats. In: NeurIPS (2024)
49. 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: CVPR (2025)
50. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
51. Sargent, K., Li, Z., Shah, T., Herrmann, C., Yu, H.X., Zhang, Y., Chan, E.R., Lagun, D., Fei-Fei, L., Sun, D., et al.: Zeronvs: Zero-shot 360-degree view synthesis from a single image. In: CVPR (2024)
52. Seo, J., Fukuda, K., Shibuya, T., Narihira, T., Murata, N., Hu, S., Lai, C.H., Kim, S., Mitsufuji, Y.: Genwarp: Single image to novel views with semantic-preserving generative warping. In: NeurIPS (2024)
53. Shi, R., Chen, H., Zhang, Z., Liu, M., et al.: Zero123++: a single image to consistent multi-view diffusion base model. arXiv preprint arXiv:2310.15110 (2023)
54. Sitzmann, V., Rezchikov, S., Freeman, B., Tenenbaum, J., Durand, F.: Light field networks: Neural scene representations with single-evaluation rendering. In: NeurIPS (2021)
55. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
56. Tancik, M., Weber, E., Ng, E., Li, R., Yi, B., Kerr, J., Wang, T., Kristoffersen, A., Austin, J., Salahi, K., Ahuja, A., McAllister, D., Kanazawa, A.: Nerfstudio: A modular framework for neural radiance field development. In: SIGGRAPH (2023)
57. Tang, J., Chen, Z., Chen, X., Wang, T., Zeng, G., Liu, Z.: Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In: ECCV (2024)
58. 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: ECCV (2024)
59. Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
60. Wang, J., Yuan, Y., Zheng, R., Lin, Y., Gao, J., Chen, L.Z., Bao, Y., Zhang, Y., Zeng, C., Zhou, Y., et al.: Spatialvid: A large-scale video dataset with spatial annotations. arXiv preprint arXiv:2509.09676 (2025)
61. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: CVPR (2025)
62. Wang, Q., Wang, Z., Genova, K., Srinivasan, P., Zhou, H., Barron, J.T., Martin-Brualla, R., Snavely, N., Funkhouser, T.: Ibrnet: Learning multi-view image-based rendering. In: CVPR (2021)
63. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: CVPR (2024)
64. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. TIP (2004)
65. Wang, Z., Yuan, Z., Wang, X., Chen, T., Xia, M., Luo, P., Shan, Y.: Motionctrl: A unified and flexible motion controller for video generation. In: SIGGRAPH (2024)
66. Wang, Z., Shen, T., Nimier-David, M., Sharp, N., Gao, J., Keller, A., Fidler, S., Müller, T., Gojcic, Z.: Adaptive shells for efficient neural radiance field rendering. TOG (2023)

67. Watson, D., Saxena, S., Li, L., Tagliasacchi, A., Fleet, D.J.: Controlling space and time with diffusion models. In: ICLR (2025)
68. Wu, R., Mildenhall, B., Henzler, P., Park, K., Gao, R., Watson, D., Srinivasan, P.P., Verbin, D., Barron, J.T., Poole, B., et al.: Reconfusion: 3d reconstruction with diffusion priors. In: CVPR (2024)
69. Xing, J., Liu, H., Xia, M., Zhang, Y., Wang, X., Shan, Y., Wong, T.T.: Tooncrafter: Generative cartoon interpolation. TOG (2024)
70. Xing, J., Xia, M., Liu, Y., Zhang, Y., Zhang, Y., He, Y., Liu, H., Chen, H., Cun, X., Wang, X., et al.: Make-your-video: Customized video generation using textual and structural guidance. TVCG (2024)
71. Xing, J., Xia, M., Zhang, Y., Chen, H., Wang, X., Wong, T.T., Shan, Y.: Dynamicrafter: Animating open-domain images with video diffusion priors. In: ECCV (2024)
72. Xu, D., Nie, W., Liu, C., Liu, S., Kautz, J., Wang, Z., Vahdat, A.: Camco: Camera-controllable 3d-consistent image-to-video generation. arXiv preprint arXiv:2406.02509 (2024)
73. Yang, H., Chen, Y., Pan, Y., Yao, T., Chen, Z., Mei, T.: 3dstyle-diffusion: Pursuing fine-grained text-driven 3d stylization with 2d diffusion models. In: ACM MM (2023)
74. Yang, H., Chen, Y., Pan, Y., Yao, T., Chen, Z., Ngo, C.W., Mei, T.: Hi3d: Pursuing high-resolution image-to-3d generation with video diffusion models. In: ACM MM (2024)
75. Yang, H., Chen, Y., Pan, Y., Yao, T., Chen, Z., Wu, Z., Jiang, Y.g., Mei, T.: Dreammesh: Jointly manipulating and texturing triangle meshes for text-to-3d generation. In: ECCV (2024)
76. Yang, J., Sax, A., Liang, K.J., Henaff, M., Tang, H., Cao, A., Chai, J., Meier, F., Feiszli, M.: Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In: CVPR (2025)
77. Yang, J., Ivanovic, B., Litany, O., Weng, X., Kim, S.W., Li, B., Che, T., Xu, D., Fidler, S., Pavone, M., Wang, Y.: Emernerf: Emergent spatial-temporal scene decomposition via self-supervision. arXiv preprint arXiv:2311.02077 (2023)
78. 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. arXiv preprint arXiv:2408.06072 (2024)
79. Ye, X., Du, Y., Tao, Y., Chen, Z.: Textssr: diffusion-based data synthesis for scene text recognition. In: ICCV (2025)
80. Yin, S., Wu, C., Liang, J., Shi, J., Li, H., Ming, G., Duan, N.: Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. arXiv preprint arXiv:2308.08089 (2023)
81. You, M., Zhu, Z., Liu, H., Hou, J.: Nvs-solver: Video diffusion model as zero-shot novel view synthesizer. In: ICLR (2025)
82. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelNeRF: Neural radiance fields from one or few images. In: CVPR (2021)
83. Yu, H.X., Duan, H., Herrmann, C., Freeman, W.T., Wu, J.: Wonderworld: Interactive 3d scene generation from a single image. In: CVPR (2025)
84. Yu, W., Xing, J., Yuan, L., Hu, W., Li, X., Huang, Z., Gao, X., Wong, T.T., Shan, Y., Tian, Y.: Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. TPAMI (2025)
85. Yu, Z., Sattler, T., Geiger, A.: Gaussian opacity fields: Efficient adaptive surface reconstruction in unbounded scenes. TOG (2024)

86. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023)
87. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
88. Zhang, Z., Long, F., Pan, Y., Qiu, Z., Yao, T., Cao, Y., Mei, T.: Trip: Temporal residual learning with image noise prior for image-to-video diffusion models. In: CVPR (2024)
89. Zhang, Z., Long, F., Qiu, Z., Pan, Y., Liu, W., Yao, T., Mei, T.: Motionpro: A precise motion controller for image-to-video generation. In: CVPR (2025)
90. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. TOG (2018)
91. Zhu, H., Wang, Y., Zhou, J., Chang, W., Zhou, Y., Li, Z., Chen, J., Shen, C., Pang, J., He, T.: Aether: Geometric-aware unified world modeling. In: ICCV (2025)