FreeGen: Feed-Forward Reconstruction–Generation Co-Training for Free-Viewpoint Driving Scene Synthesis

Shijie Chen[1] and Peixi Peng[1,2,3]✉

[1] Peking University Shenzhen Graduate School, China
[2] Pengcheng Laboratory, China
[3] Frontop, China
{2501111941@stu.,pxpeng}@pku.edu.cn

Abstract. Closed-loop simulation for autonomous driving requires synthesizing free-viewpoint driving scenes. However, existing datasets and generative pipelines rarely provide consistent off-trajectory observations, limiting large-scale evaluation and training. While recent generative models demonstrate strong visual realism, they struggle to simultaneously achieve interpolation consistency and extrapolation realism without per-scene optimization. To address this, we propose FreeGen, a feed-forward reconstruction-generation co-training framework for free-viewpoint driving scene synthesis. The reconstruction model provides stable geometric representations to ensure interpolation consistency, while the generation model performs geometry-aware enhancement to improve realism at unseen viewpoints. Through co-training, generative priors are distilled into the reconstruction model to improve off-trajectory rendering, and the refined geometry in turn offers stronger structural guidance for generation. Experiments demonstrate that FreeGen achieves state-of-the-art performance for free-viewpoint driving scene synthesis.

Keywords: Novel View Synthesis · Free Trajectory · Driving Scene

1 Introduction

Robust autonomous driving systems require virtual environments for scalable pretraining and closed-loop policy evaluation [22,23,25]. However, existing real [4, 15, 40] and synthetic driving data [12, 13, 43] are predominantly collected along a single trajectory, providing limited coverage of the diverse trajectories induced by different actions. The scarcity of off-trajectory observations hinders high-fidelity and consistent rendering at free viewpoints [7, 8, 47, 49]. The key challenge lies in achieving high-quality free-viewpoint driving scene synthesis across different trajectories using only single-trajectory observations.

To this end, a practical solution should satisfy three requirements: (1) **Interpolation Consistency.** Preserve geometric consistency with the observed

✉Corresponding author.

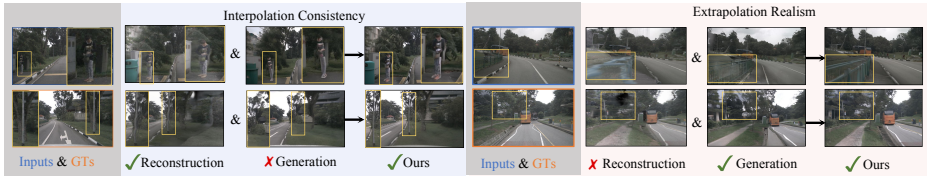

Fig. 1: Complementary strengths of reconstruction and generation methods. Reconstruction methods maintain geometry consistency but lack realistic textures, while generation methods improve realism but often distort geometry. Our method combines both, achieving consistent and realistic results.

views in their overlapping regions. (2) **Extrapolation Realism.** Generate visually plausible content that matches the underlying distribution at unseen viewpoints. (3) **Feed-forward.** Synthesize the scene in a single forward pass instead of iterative refinement, enabling scalability with data-driven advances (e.g., scaling laws) for large-scale driving scene generation.

Traditional reconstruction-based methods [7, 8, 47, 49] excel at enforcing interpolation consistency but are limited to interpolating within sparse observations, making it difficult to render free-viewpoint views. They typically require per-scene optimization, leading to prohibitive computational costs that hinder large-scale scene synthesis. Feed-forward reconstruction approaches [41, 45] enable efficient scene reconstruction in a single pass but still struggle to produce realistic details at extrapolated views. With the rapid progress of generative models, data-driven methods trained on large-scale videos [3, 14, 43] show strong potential for generating realistic views from sparse inputs, yet they generally lack fine-grained control over camera trajectories. Hybrid methods that combine generative models with reconstruction-based pipelines [10, 32, 50, 53] enhance visual detail while preserving interpolation consistency, but often require hours to days of per-scene training, undermining their practicality for feed-forward inference.

Recent methods leverage free-viewpoint rendered views to guide generative refinement for high-quality feed-forward scene synthesis [16, 42]. However, LiDAR-based conditions [42] are sparse and expensive, and often fail to provide sufficient geometric coverage for distant buildings and sky, while image-warping-based conditions [16] often introduce structural misalignment (e.g., missing utility poles and distorted pedestrians in Fig. 1). Moreover, these conditions are non-learnable, limiting their ability to benefit from large-scale image data.

We observe that reconstruction and generation methods offer complementary strengths for free-viewpoint driving scene synthesis. As shown in Fig. 1, reconstruction models excel at preserving geometric and interpolation consistency but are limited in fine details at unseen viewpoints. In contrast, generation models can produce visually realistic results from sparse observations, yet often introduce structural distortions with respect to the input views. This complementarity motivates us to unify these two paradigms in a single feed-forward framework that preserves geometric faithfulness while leveraging generative realism.

Fig. 2: Qualitative comparison of reconstruction results on and off the trajectory. The rendering quality on the trajectory (Forward 2m) is excellent. For off-trajectory (Traj.) views (Left 2m), the results without our *Co-Training* (w/o CT.) produce noticeable structural artifacts, whereas the results after our *Co-Training* (w/ CT.) reduce these artifacts and improve the rendering quality.

Recent concurrent studies [28, 54] have explored integrating feed-forward reconstruction with video diffusion models to enhance synthesis quality. However, such a unidirectional design is suboptimal. Since most driving datasets [4, 40] are collected along a single forward trajectory, the coverage of training viewpoints is restricted. Consequently, the reconstruction model inevitably introduces novel geometric biases when extrapolating to off-trajectory viewpoints. As the generative model is primarily optimized on on-trajectory data, it struggles to rectify these off-trajectory artifacts, ultimately leading to degraded refinement quality. To intuitively illustrate this limitation, we provide a qualitative comparison in Fig. 2. The reconstruction model [45] successfully renders high-fidelity views along the original trajectory. However, when extrapolating to off-trajectory viewpoints, it produces noticeable structural artifacts.

To this end, we propose **FreeGen**, a feed-forward reconstruction–generation framework for free-viewpoint driving scene synthesis. We first employ a feed-forward 3D Gaussian Splatting (3DGS) model to efficiently reconstruct a scene from sparse input images without per-scene optimization. Then we render views along free-viewpoint trajectories to enforce interpolation consistency with the observed inputs. These rendered views are then used as geometric guidance for a generative refinement model, enabling realistic extrapolation at unseen viewpoints. Moreover, FreeGen does not require additional expensive annotations such as LiDAR or bounding boxes, relying solely on image data from a single trajectory, which makes it practical for large-scale driving scene synthesis.

Our method introduces a novel bidirectional refinement architecture. We build upon recent advances in feed-forward 3DGS [45] and video diffusion models [3]. To fully exploit their complementary strengths, we introduce two key designs in FreeGen. The reconstruction branch already produces structurally complete renderings with minimal visible holes, but due to limited supervision from a single trajectory, a naive generative module cannot reliably localize subtle artifacts. To address this, we **construct geometry conditions** based on

the 3DGS rendering mechanism and incorporate a **geometry-aware diffusion refinement** module to guide appearance refinement. In addition, we propose a **co-training** strategy that samples viewpoints off the input trajectory and forms a closed loop between reconstruction and generation. The generative priors distilled during refinement are fed back into the reconstruction model to improve free-viewpoint rendering quality, while the reconstruction model provides geometric guidance to construct closed-loop supervision for the generative module.

Our contributions are summarized as follows:

- We introduce FreeGen, a feed-forward reconstruction-generation co-training framework for free-viewpoint driving scene synthesis, which achieves both interpolation consistency and extrapolation realism without per-scene optimization or additional expensive annotations.
- We propose a geometry-aware diffusion refinement module to achieve high-quality refinement on reconstructed views. To further improve both reconstruction and generation performance at off-trajectory viewpoints, we introduce a closed-loop co-training strategy.
- Extensive experiments demonstrate that FreeGen outperforms existing approaches in free-viewpoint driving scene synthesis.

2 Related Work

Driving Scene Synthesis with Reconstruction. Early NeRF-based methods [1, 2, 30, 31] enable free-viewpoint synthesis but suffer from slow rendering, a bottleneck alleviated by the faster, point-based 3DGS [26]. Recent works [8, 47] adapt 3DGS for driving scenarios. MagicDrive3D [11] reconstructs scenes from synthesized images to reduce artifacts. However, these methods rely on expensive annotations (e.g., bounding boxes) [8,47], require per-scene optimization, and focus mainly on the training trajectory, limiting large-scale scalability. To improve efficiency, feed-forward 3DGS models [5, 6, 41] predict Gaussians in a single pass. For instance, Omni-Scene [45] fuses pixel and voxel features for better novel-view rendering, while InfiniCube [29] uses a voxel world to guide generation, though constrained by resolution. Despite these advances, existing feed-forward approaches still primarily target viewpoints near the original trajectory.

Driving Scene Synthesis with Diffusion Prior. Driven by diffusion models [9, 18–20, 35, 39], recent works [13, 14, 43] have advanced driving scene generation. To synthesize free-viewpoint scenes, several approaches iteratively refine rendered views and optimize 3D representations. DIFIX3D+ [46] employs a single-step diffusion model to repair 3DGS renderings. ReconDreamer [32] and DriveX [50] use video diffusion models to repair free-viewpoint renderings, while FreeSim [10] leverages progressive extrapolation. DriveDreamer4D [53] and StreetCrafter [48] distill generative repairs into 3DGS. However, these methods rely on slow per-scene optimization, limiting their scalability. ViewCrafter [51] utilizes image warping as diffusion guidance and iteratively expands the point cloud based on generated results. Similarly, FreeVS [42] and DiST-4D [16] leverage LiDAR projections or image warping as diffusion conditions, enabling effi-

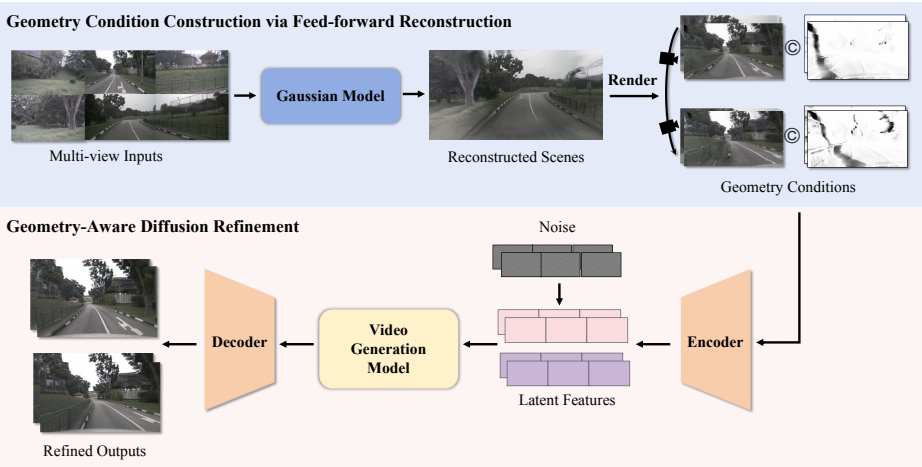

Fig. 3: Overview of the proposed FreeGen. The reconstruction model encodes multi-view inputs into Gaussian features and decodes them into 3DGS representations. The rendered views and corresponding opacity maps are then used to guide the geometry-aware diffusion model, which performs fine-grained refinement. For clarity, depth maps are omitted from the illustration.

cient training and feed-forward inference, but also suffer from sparsity or structural misalignment. Recent feed-forward methods like WorldSplat [54] and Phi-Gensis [28] integrate 4D representations with generative refinement, but their performance remains constrained by single-trajectory supervision.

3 Methodology

Our key idea is to leverage the complementary strengths of reconstruction and generation to enhance both interpolation consistency and extrapolation realism in free-viewpoint scene synthesis. To achieve this, we first reconstruct the scene from multi-view inputs using a feed-forward reconstruction model, following [45]. Then we perform free-viewpoint rendering along given camera trajectories to obtain rendered views and corresponding geometry conditions (Sec. 3.1). Next, the rendered results are used to guide a diffusion refinement model (Sec. 3.2). The overall framework of our method is illustrated in Fig. 3. Furthermore, we jointly train the reconstruction and generation models to improve overall performance (Sec. 3.3), as shown in Fig. 4.

3.1 Geometry Condition Construction via Feed-forward Reconstruction

An appropriate scene representation should support both efficient reconstruction and fast rendering. Recent feed-forward 3DGS methods [5, 6, 41, 45] make

this possible by reconstructing large-scale driving scenes from multi-view images within seconds, achieving real-time rendering and learning rich scene priors from large-scale data. Motivated by this, we adopt a feed-forward 3DGS-based reconstruction network in FreeGen, following [45].

Given sparse N-view inputs $\{I_i\}_{i=1}^{N}$ with corresponding camera parameters $\{P_i\}_{i=1}^{N}$, we first predict depth maps $\{D_i\}_{i=1}^{N}$ using an off-the-shelf depth estimation model [21] to assist scene reconstruction. We then use a Gaussian encoder to extract Gaussian features. Specifically, we apply a multi-view U-Net style pixel encoder [5, 6, 41, 45] to extract pixel-aligned high-resolution feature maps $\{\mathcal{F}_i^{\text{pixel}}\}_{i=1}^{N}$. Then we fuse multi-view features into a triplane representation via a Triplane Transformer [24] with deformable attention [27]. Finally, Gaussian decoders transform both pixel and voxel features into a set of K 3D Gaussians [26] $\{G_i = (\delta_i, \alpha_i, s_i, q_i, c_i)\}_{i=1}^{K}$, where δ_i, α_i, s_i, q_i, and c_i denote the position, opacity, scale, orientation, and color of each Gaussian. We then render the color image I_{geo}, opacity map A_{geo}, and depth map D_{geo} via tile-based rasterization [26]. The color image is computed as follows:

$$I_{\text{geo}} = \sum_{i=1}^{K} c_i \, \alpha_i' \prod_{j=1}^{i-1} (1 - \alpha_j'), \tag{1}$$

where α_i' is the screen-space opacity contribution of the i-th Gaussian determined by 3DGS [26]. The opacity map A_{geo} and depth map D_{geo} are obtained similarly to the color image. Then we define the geometry conditions C_{geo} as follows:

$$C_{\text{geo}} \;=\; \big[\, I_{\text{geo}},\, D_{\text{geo}},\, A_{\text{geo}} \,\big]. \tag{2}$$

The geometry conditions provide the diffusion model with essential structural and reliability cues, enabling geometry-aware refinement. The color image provides structural and appearance information, the depth map enhances geometric consistency [16], and the opacity map reflects the reliability of the color image, where higher opacity usually corresponds to more reliable renderings.

Training Strategy. The training phase of the reconstruction branch includes two stages. In the first stage, we follow [45] and pre-train the model on the original single-trajectory sequences, forming cross-timestamp pairs to establish the reconstruction capability of the model. In the second training stage, we randomly render views at off-trajectory viewpoints and refine them with the generation branch to obtain pseudo-labels. These pseudo-labels then supervise the reconstruction branch to improve free-viewpoint rendering. Details are provided in Sec. 3.3. In the overall training phase, we supervise the reconstruction model using mean squared error loss, LPIPS loss [52], and L1 depth loss as follows:

$$\mathcal{L}_{recon} = \mathcal{L}_{mse} + \lambda_1 \mathcal{L}_{lpips} + \lambda_2 \mathcal{L}_{depth}, \tag{3}$$

where λ_1 and λ_2 are the weights for each loss component. The depth supervision uses pseudo-labels produced by the off-the-shelf depth model [21].

3.2 Geometry-Aware Diffusion Refinement

Although the reconstruction model can effectively recover scene structures from multi-view images, it still struggles to produce high-quality and realistic results at free viewpoints, especially for off-trajectory views. With the recent success of diffusion models in image generation [18, 35], we introduce a diffusion-based generation module to refine the feed-forward 3DGS renderings. Since the video diffusion model [3] has been extensively pre-trained on large-scale data, introducing additional parameters may degrade its generalization ability. Therefore, we keep our modification minimal to maintain compatibility and ensure stable generalization across diverse scenes.

To simplify the explanation, we describe the process using a single image as an example. In practice, we employ a video diffusion model to ensure temporal consistency. For the geometry conditions C_{geo} rendered by the reconstruction model, we encode the geometry conditions into a condition latent representation $\mathbf{z}_c = \mathcal{E}_c(C_{\text{geo}})$ through a learnable condition encoder $\mathcal{E}_c$ like [16]. During training, we encode the reference image I_r using a frozen image encoder $\mathcal{E}_v$ to obtain the latent representation $\mathbf{z}_v = \mathcal{E}_v(I_r)$. Gaussian noise is added to the latent to obtain a noisy version [18] as follows:

$$\mathbf{z}_\tau = \alpha_\tau \mathbf{z}_v + \sigma_\tau \boldsymbol{\epsilon}, \tag{4}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 1)$ and τ is the diffusion timestep. The geometry condition latent $\mathbf{z}_c$ is concatenated with the noise latent $\mathbf{z}_\tau$ along the channel dimension and fed into the diffusion model f_θ to predict the noise as follows:

$$\hat{\boldsymbol{\epsilon}} = f_\theta(\mathbf{z}_\tau, \mathbf{z}_c). \tag{5}$$

The training objective [18] minimizes the mean squared error between the predicted and true noise as follows:

$$\mathcal{L}_{\text{gen}} = \mathbb{E}_{\tau, \mathbf{z}_v, \boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - f_\theta(\mathbf{z}_\tau, \mathbf{z}_c)\|_2^2\right]. \tag{6}$$

This allows the diffusion model to learn geometry-aware refinement conditioned on the rendered scene structure.

Training Strategy. The diffusion model is also trained in two stages. In the first stage, following DiST-4D [16], we formulate cross-timestamp pairs to pre-train the diffusion model. We utilize these cross-temporal renderings as geometric conditions alongside their corresponding ground-truth frames, ensuring the model learns to refine appearances under explicit geometric constraints. In the second stage, we introduce a closed-loop supervision scheme to improve generation quality at off-trajectory viewpoints. Details are provided in Sec. 3.3. The depth supervision follows the same protocol in Sec. 3.1.

3.3 Co-Training

Considering that both the reconstruction and generation models are trained with single-trajectory data, coverage of viewpoint variation is limited for free-viewpoint synthesis. Inspired by closed-loop training [16, 34] and noting that

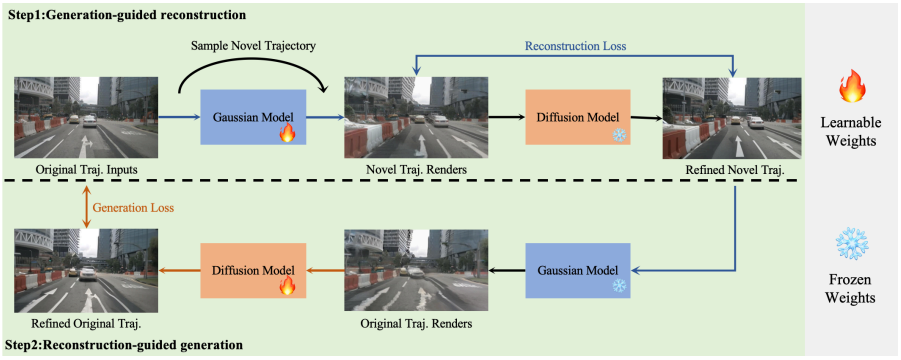

Fig. 4: Illustration of the Co-Training strategy. FreeGen adopts a closed-loop co-training strategy between the pre-trained Gaussian reconstruction model and the pre-trained diffusion refinement model. Novel trajectories (Traj.) are sampled and rendered by the Gaussian model, then refined by the diffusion model to form pseudo-supervision. The refined results are fed back to the Gaussian model for reconstruction loss, while the diffusion model learns from generation loss on the original trajectory.

both models are learnable, we design a co-training scheme that lets the two models promote each other, as shown in Fig. 4. The procedure alternates two steps and freezes one model at a time to keep optimization stable. Many existing methods, such as DIFIX3D [46] and StreetCrafter [48], distill diffusion priors into 3DGS by iteratively refining renderings at inference time. However, this process is computationally expensive and slow. Inspired by their success, we leverage our learnable reconstruction branch to integrate these priors through a co-training loop, eliminating the need for test-time optimization.

Step 1: Generation-guided reconstruction. To enhance the performance at off-trajectory viewpoints, we sample novel camera poses and render them using the pre-trained Gaussian model. The frozen pre-trained diffusion model then refines renderings to create pseudo-labels. Then we update the Gaussian model using these labels and the original captured data. This step effectively transfers generative priors to the Gaussian model. To prevent error accumulation, the inputs to the Gaussian model remain the original data, and generated outputs are used only as auxiliary supervision. Following the practice in StreetCrafter [48], we set the weight of this auxiliary loss to 0.1 to ensure training stability. Considering the strong baseline capabilities of our pre-trained models, we sample poses within a shift range of $[-3m, 3m]$ and record offsets for use in the next step.

Step 2: Reconstruction-guided generation. Using the updated Gaussian model, we render the off-trajectory refined results from Step 1 back to the original trajectory viewpoints. The diffusion model refines these re-rendered views under geometric conditions and the outputs are supervised by the original trajectory ground-truth frames with the generation losses in Sec. 3.2. This step strengthens geometry-aware refinement in the generation branch. Notably, although utilizing generated data for reconstruction typically risks error propagation, our design

naturally alleviates this risk. Since the supervision relies strictly on the original trajectory data to prevent training drift, the variations introduced by rendering from generated data effectively act as a form of robust data augmentation.

Through alternating Step 1 and Step 2, the reconstruction branch learns stronger off-trajectory geometry and the generation branch learns geometry-consistent appearance. The closed loop progressively aligns structural consistency and visual realism across free-viewpoints.

4 Experiments

4.1 Setup

Dataset. We conduct experiments on nuScenes [4], following [16]. Following prior works [16,45], the 2 Hz keyframe annotations are interpolated to 12 Hz. We use the official split with 700 scenes for training and 150 scenes for validation.
Evaluation. For *off-trajectory evaluation*, we follow the protocols in [16, 42]. Specifically, we sample frames every two frames along the original trajectory, and apply lateral camera shifts $\tau \in \{\pm 1\text{m}, \pm 2\text{m}, \pm 4\text{m}\}$ to each sampled frame. We synthesize RGB videos under the shifted viewpoints and compute Fréchet Inception Distance (FID) [37] and Fréchet Video Distance (FVD) [38] between the synthesized videos and the captured videos from the original trajectory. These metrics reflect appearance realism and temporal coherence under viewpoint extrapolation. For *on-trajectory evaluation*, we assess the performance across three difficulty levels. Specifically, we select target viewpoints at temporal intervals of 1, 2, and 4 frames, designating them as *Easy*, *Medium*, and *Hard* settings, respectively. We report peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) [44], and perceptual similarity (LPIPS) [52] to assess reconstruction accuracy and perceptual fidelity along the original trajectory.
Baselines. We compare FreeGen with several baselines. PVG [7], EmerNeRF [49], StreetGaussian [47], and OmniRe [8] represent reconstruction approaches without a generative base model, while Image warping [16] and Omni-Scene [45] are compared as feed-forward reconstruction baselines. For generation-based refinement, we evaluate the SD-Turbo-based DIFIX3D+ [46] (integrated with PVG [7]), alongside SVD-based FreeVS [42] and DiST-4D [16], all within a 6-frame evaluation setting. Finally, we list results for DiT-based concurrent works PhiGensis [28] (OpenSora V2.0 [17]) and WorldSplat [54] (OpenSora V1.2 [17]) for reference.
Implementation Details. To ensure a fair comparison with prior works [16, 42], we set the video length to $T = 6$ and adopt a base resolution of 768×432 for overall evaluation, preserving the original 1600×900 aspect ratio. For optimization, the reconstruction model undergoes 50K pre-training and 20K co-training steps, using loss weights $\lambda_1 = 0.05$ and $\lambda_2 = 0.01$ [45]. During the co-training phase, the auxiliary loss weight for the reconstruction branch is set to 0.1 to ensure training stability. Conversely, the generation model is trained for 20K pre-training and 20K co-training steps with a constant learning rate of 1×10^{-5},

Table 1: Quantitative comparison of free-viewpoint synthesis quality. We compare FreeGen with baselines under different trajectory shift settings. Results for DIFIX3D+ [46] are reproduced by adapting its official implementation, and results for PhiGensis [28] and WorldSplat [54] are directly reported from their original papers, while others are from DiST-4D [16]. **Base Model** specifies the pre-trained generative backbone used for video or image synthesis. Image warping and Omni-Scene [45] serve as the geometric conditions for DiST-4D [16] and ours, respectively.

Method	Base Model	Shift ±1m		Shift ±2m		Shift ±4m	
		FID↓	FVD↓	FID↓	FVD↓	FID↓	FVD↓
PVG [7]	-	48.15	246.74	60.44	356.23	84.50	501.16
EmerNeRF [49]	-	37.57	171.47	52.03	294.55	76.11	497.85
StreetGaussian [47]	-	32.12	153.45	43.24	256.91	67.44	429.98
OmniRe [8]	-	31.48	152.01	43.31	254.52	67.36	428.20
Image warping [16]	-	32.27	216.69	60.64	344.68	93.27	398.36
Omni-Scene [45]	-	12.92	117.49	20.96	205.99	39.47	370.79
DIFIX3D+ [46]	SD-Turbo [36]	17.36	143.04	21.73	154.51	30.43	263.69
FreeVS [42]	SVD [3]	51.26	431.99	62.04	497.37	77.14	556.14
DiST-4D [16]	SVD [3]	10.12	45.14	12.97	68.80	17.57	105.29
Ours	SVD [3]	**9.47**	**34.11**	**11.25**	**43.28**	**14.30**	**66.19**
PhiGensis [28]	OpenSora V2.0 [33]	9.80	43.80	11.71	67.54	15.51	103.13
WorldSplat [54]	OpenSora V1.2 [17]	8.25	40.17	11.26	47.41	13.38	64.07

as original ground-truth data is available for supervision. All experiments run on two NVIDIA A100 GPUs. We initialize the reconstruction branch from Omni-Scene [45] and the generation branch from SVD [3].

4.2 Main Results

Off-trajectory Quantitative Evaluation. Table 1 reports results under three lateral shift settings. Classical reconstruction methods [7, 8, 47, 49] suffer from blur at off-trajectory viewpoints because no observations exist outside the input views. FreeVS [42] introduces extra geometry conditions by projecting LiDAR, but the point cloud is sparse, which limits performance. DIFIX3D+ [46] employs 3DGS renderings to condition the diffusion model, but its per-scene optimized 3DGS lacks learned priors and inevitably overfits to the observed regions. DiST-4D [16] uses image warping to provide geometry conditions and obtains noticeable gains, yet warping cannot faithfully represent full scene structure and the pipeline still depends on LiDAR and boxes with heavy preprocessing to recover depth. Benefiting from the holistic 3D representation provided by Omni-Scene [45], FreeGen achieves the best results at all shift levels. The improvements are especially large on FVD, indicating higher temporal coherence and realism under free-viewpoint changes. FreeGen uses only images and an off-the-shelf depth estimator to obtain depth in seconds, which is fast and easy to deploy at large scale. Furthermore, we compare FreeGen with state-of-the-art DiT-based

Table 2: Quantitative comparison of interpolation quality on the recorded trajectory. We evaluate the performance under three temporal sparsity settings.

Method	@1 (Dense)				@2 (Medium)				@4 (Sparse)			
	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
PVG [7]	23.33	**0.73**	0.41	39.47	22.08	**0.71**	0.41	47.25	20.47	**0.67**	0.45	70.94
DIFIX3D+ [46]	21.73	0.65	0.37	17.59	20.78	0.62	0.39	19.08	19.68	0.59	0.41	23.42
DiST-4D [16]	19.26	0.66	0.36	13.42	18.47	0.62	0.39	15.68	17.55	0.58	0.43	20.64
Ours	**23.34**	0.72	**0.30**	**9.61**	**22.29**	0.69	**0.32**	**11.25**	**20.77**	0.65	**0.36**	**14.81**

methods. We outperform PhiGensis [28] across all shift settings. Even when compared to WorldSplat [54], which leverages spatial annotations (HDMaps and 3D bounding boxes), FreeGen achieves competitive FID scores and superior FVD. This comparison highlights the robustness of FreeGen in generating high-quality free-viewpoint results without relying on costly annotations.

On-trajectory Quantitative Evaluation. Table 2 presents the interpolation results on the original trajectory across three temporal sparsity levels. The advanced reconstruction method PVG [7] achieves remarkable reconstruction quality under dense observations. However, its performance degrades rapidly as the input becomes sparse, indicating a strong tendency to overfit to the observed views. In contrast, generative approaches like DIFIX3D+ [46] and DiST-4D [16] use 3DGS and point cloud as geometric conditions to guide diffusion-based refinement. Since DIFIX3D+ relies on 3DGS renderings, it maintains better reconstruction metrics than DiST-4D. Nevertheless, since these generative models tend to hallucinate and over-enhance scene details, their pixel-aligned reconstruction metrics fall behind PVG, even though their visual realism is superior. Our FreeGen overcomes these limitations by learning robust data priors through a feed-forward reconstruction branch, which faithfully recovers scene structures even given highly sparse inputs. Furthermore, our co-training strategy significantly boosts the generative branch's ability to maintain structural consistency and high visual quality. Consequently, FreeGen achieves the best PSNR and LPIPS, yields an SSIM on par with PVG, and dramatically lowers the FID score.

Qualitative Evaluation. In Fig. 5, we present a qualitative comparison between FreeGen and established baselines. Traditional per-scene reconstruction methods, such as PVG [7] and StreetGaussian [47], suffer from pronounced blurring and ghosting artifacts in fine details. On the generative side, DIFIX3D+ [46] tends to over-enhance the scene and often produces spatial misalignments due to its over-reliance on reference frames. Although DiST-4D [16] generates highly realistic textures, it frequently introduces severe structural distortions. In contrast, FreeGen maintains structural consistency while delivering high-quality extrapolation results. Notably, our feed-forward reconstruction branch outperforms per-scene optimized models at novel viewpoints, demonstrating that robust priors learned from large-scale data play a crucial role in free-viewpoint synthesis. Besides, Fig. 6 further illustrates the spatial and temporal capabilities of FreeGen. As shown in Fig. 6(a), under challenging lateral shifts of 2m to the left

Fig. 5: Qualitative comparison on free-viewpoint synthesis. Compared to optimization-based [7, 47] and generative baselines [16, 46], our FreeGen achieves superior visual quality and preserves structural consistency.

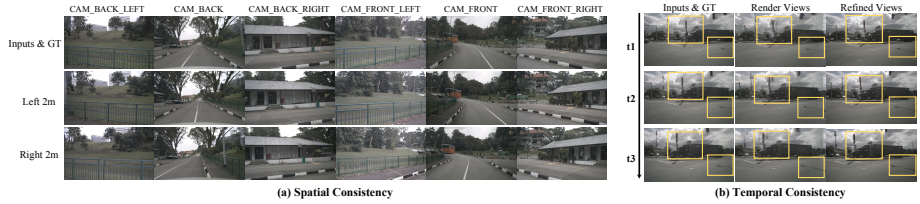

Fig. 6: Qualitative results of FreeGen. (a) Spatial consistency: Synthesized novel views under ±2m lateral shifts. **(b) Temporal consistency:** Consecutive frames generated under a −1m shift. Our FreeGen maintains sharp details and stable geometry across both spatial and temporal dimensions.

and right, our method synthesizes detailed appearances while preserving scene geometry. Moreover, Fig. 6(b) validates our temporal consistency. Across consecutive frames, the refined views remain highly stable, exhibiting no noticeable flickering or geometric drift over time.

4.3 Ablation Study

Main Ablation Studies. We conduct an ablation under a leftward shift of −2m to validate each design choice. Table 3 reports the results for settings (a) to (d). *(a) Reconstruction only (Rec.).* Novel views are rendered by the reconstruction branch without generative refinement. This yields competitive FID due to faithful geometry but limited off-trajectory information and single-frame input lead to worse FVD. *(b) Generation (Gen.).* Using the rendered color and depth to guide generation branch improves both metrics, but the model lacks an explicit notion of reliable regions to refine. *(c) Opacity guidance (Opacity).* Adding the opacity map provides reliability cues, which further improves performance. *(d) Co-training on both models.* Jointly updating reconstruction and generation branches achieves the best results. The closed loop transfers appearance priors

Table 3: Ablation study on the contribution of each module. All metrics are evaluated under a left 2m shift using six camera views. **Rec.** and **Gen.** indicate enabling the reconstruction and generation branches. **Opacity** denotes using the opacity maps as an additional geometric condition. **Co-training** indicates enabling our complete alternating co-training pipeline (details in Table 4).

Settings	Rec.	Gen.	Opacity	Co-training	FID-2m↓	FVD-2m↓
(a)	✓				23.25	246.73
(b)	✓	✓			15.49	60.53
(c)	✓	✓	✓		13.86	57.21
(d)	✓	✓	✓	✓	**12.87**	**48.39**

Table 4: Ablation study of co-training dynamics under a left 2m shift on six cameras. **Gen. Only** directly fine-tunes the generation branch with a frozen reconstruction branch. **Rec. Only** updates the reconstruction branch via generative priors but freezes the generation branch. **Full Co-training** represents our complete alternating optimization pipeline. **Rec. CT** and **Gen. CT** indicate co-training the reconstruction branch and the generation branch respectively.

Setting	Rec.CT	Gen.CT	Rec.		Rec. + Gen.	
			FID↓	FVD↓	FID↓	FVD↓
Pre-trained	-	-	23.25	246.73	13.86	57.21
Rec. Only	✓	-	**22.23**	**223.39**	14.39	54.93
Gen. Only	-	✓	23.25	246.73	13.16	51.78
Full Co-training	✓	✓	**22.23**	**223.39**	**12.87**	**48.39**

to the reconstruction branch and supplies geometry-consistent supervision to the generation branch, leading to the highest quality free-viewpoint results.

Ablation of Co-training Dynamics. Table 4 shows the ablation study of different co-training strategies. When we only train the reconstruction branch, the final results are not optimal because the data distribution of geometric conditions changes, making it difficult for the generation branch to adapt to them. When we only train the generation branch, it is limited by the rendering errors of the original reconstruction branch in off-trajectory views. Our full co-training strategy achieves the best results.

Analysis of Geometry Conditions. Fig. 7 compares three types of conditions. FreeVS [42] projects LiDAR to the image plane, but the condition remains sparse and is almost unusable for distant structures. DiST-4D [16] warps images to the target view and shows clearer layouts, yet it cannot represent full scene geometry and exhibits noticeable structural drift at larger viewpoint shifts. Our condition captures scene structure and fine details and aligns well with the target view.

Analysis of Opacity Guidance. In Fig. 8, we visualize the effect of opacity guidance. The opacity map clearly highlights reliable and unreliable regions. With this guidance, the generative model can focus refinement on uncertain areas, which improves the overall quality of the results.

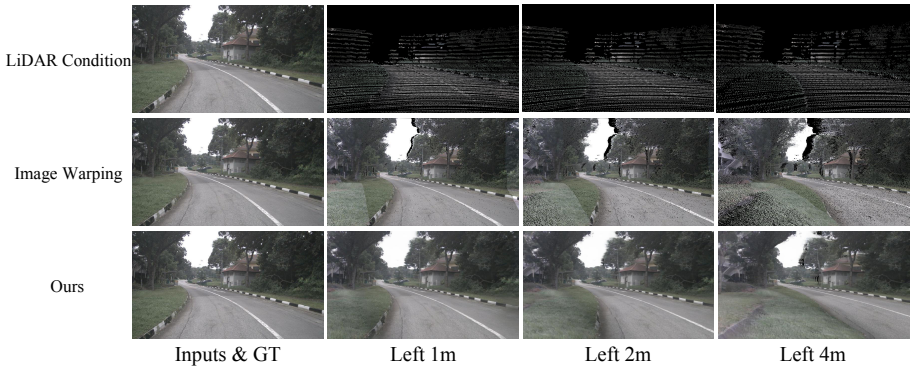

Fig. 7: Qualitative comparison of geometric conditions. We construct target views with lateral shifts of 1 m, 2 m, and 4 m. For LiDAR-based conditions, we aggregate 10 consecutive frames and project the point clouds to the target view. For image warping, we use depth to warp the input images to the target camera. Our condition is obtained by rendering from the feed-forward reconstruction model.

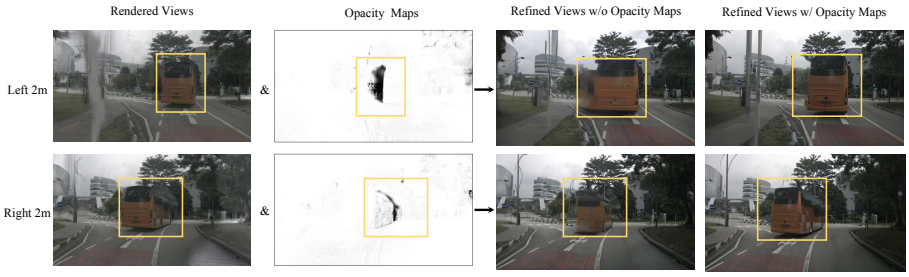

Fig. 8: Ablation on the opacity map. We compare generation results with and without opacity guidance to assess its effect.

5 Conclusion

We have introduced FreeGen, a novel feed-forward framework for free-viewpoint driving scene synthesis from the single trajectory observations. Our approach effectively addresses the critical challenge of balancing interpolation consistency and extrapolation realism. By co-training a reconstruction model with a generation model, FreeGen leverages the stable geometric guidance from reconstruction to ensure structural consistency, while harnessing the power of generative priors to enhance visual realism at unseen viewpoints. Extensive experiments demonstrate that FreeGen surpasses existing methods, enabling the efficient synthesis of high-fidelity and consistent driving scenes.

Acknowledgements

The study was funded by the Shenzhen Science and Technology Program (KQTD 20240729102051063), the National Natural Science Foundation of China under contracts No. 62422602, No. 62372010, No. 62425101, No. 62332002, No. 62206281, Key Laboratory Grants 241-HF-D05-01, and the major key project of the Peng Cheng Laboratory (PCL2021A13 and PCL2025A02). Computing support was provided by Pengcheng Cloudbrain.

References

1. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5470–5479 (2022)
2. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Zip-nerf: Anti-aliased grid-based neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19697–19705 (2023)
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. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
5. Charatan, D., Li, S.L., Tagliasacchi, A., Sitzmann, V.: pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19457–19467 (2024)
6. 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: European Conference on Computer Vision. pp. 370–386. Springer (2024)
7. Chen, Y., Gu, C., Jiang, J., Zhu, X., Zhang, L.: Periodic vibration gaussian: Dynamic urban scene reconstruction and real-time rendering. International Journal of Computer Vision **134**(3), 83 (2026)
8. Chen, Z., Yang, J., Huang, J., De Lutio, R., Martinez Esturo, J., Ivanovic, B., Litany, O., Gojcic, Z., Fidler, S., Pavone, M., et al.: Omnire: Omni urban scene reconstruction. In: International Conference on Learning Representations. vol. 2025, pp. 85508–85527 (2025)
9. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
10. Fan, L., Zhang, H., Wang, Q., Li, H., Zhang, Z.: Freesim: Toward free-viewpoint camera simulation in driving scenes. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12004–12014 (2025)
11. Gao, R., Chen, K., Li, Z., Hong, L., Li, Z., Xu, Q.: Magicdrive3d: Controllable 3d generation for any-view rendering in street scenes. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5724–5733 (2026)

12. Gao, R., Chen, K., Xiao, B., Hong, L., Li, Z., Xu, Q.: Magicdrive-v2: High-resolution long video generation for autonomous driving with adaptive control. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 28135–28144 (2025)
13. Gao, R., Chen, K., Xie, E., Hong, L., Li, Z., Yeung, D.Y., Xu, Q.: Magicdrive: Street view generation with diverse 3d geometry control. In: International Conference on Learning Representations. vol. 2024, pp. 22841–22860 (2024)
14. Gao, S., Yang, J., Chen, L., Chitta, K., Qiu, Y., Geiger, A., Zhang, J., Li, H.: Vista: A generalizable driving world model with high fidelity and versatile controllability. Advances in Neural Information Processing Systems **37**, 91560–91596 (2024)
15. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012)
16. Guo, J., Ding, Y., Chen, X., Chen, S., Li, B., Zou, Y., Lyu, X., Tan, F., Qi, X., Li, Z., et al.: Dist-4d: Disentangled spatiotemporal diffusion with metric depth for 4d driving scene generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27231–27241 (2025)
17. Guo, X., Ding, C., Dou, H., Zhang, X., Tang, W., Wu, W.: Infinitydrive: Breaking time limits in driving world models. arXiv preprint arXiv:2412.01522 (2024)
18. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
19. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
20. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. Advances in neural information processing systems **35**, 8633–8646 (2022)
21. Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., Wang, K., Yu, G., Shen, C., Shen, S.: Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
22. Hu, S., Chen, L., Wu, P., Li, H., Yan, J., Tao, D.: St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In: European Conference on Computer Vision. pp. 533–549. Springer (2022)
23. Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., et al.: Planning-oriented autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17853–17862 (2023)
24. Huang, Y., Zheng, W., Zhang, Y., Zhou, J., Lu, J.: Tri-perspective view for vision-based 3d semantic occupancy prediction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9223–9232 (2023)
25. Jiang, B., Chen, S., Xu, Q., Liao, B., Chen, J., Zhou, H., Zhang, Q., Liu, W., Huang, C., Wang, X.: Vad: Vectorized scene representation for efficient autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8340–8350 (2023)
26. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
27. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Yu, Q., Dai, J.: Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)

28. Lu, H., Ma, Z., Jiang, G., Ge, W., Li, B., Cai, Y., Zheng, W., Zhang, Y., Chen, Y.: 4d driving scene generation with stereo forcing. arXiv preprint arXiv:2509.20251 (2025)
29. Lu, Y., Ren, X., Yang, J., Shen, T., Wu, Z., Gao, J., Wang, Y., Chen, S., Chen, M., Fidler, S., et al.: Infinicube: Unbounded and controllable dynamic 3d driving scene generation with world-guided video models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27272–27283 (2025)
30. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
31. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM transactions on graphics (TOG) **41**(4), 1–15 (2022)
32. Ni, C., Zhao, G., Wang, X., Zhu, Z., Qin, W., Huang, G., Liu, C., Chen, Y., Wang, Y., Zhang, X., et al.: Recondreamer: Crafting world models for driving scene reconstruction via online restoration. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1559–1569 (2025)
33. Peng, X., Zheng, Z., Shen, C., Young, T., Guo, X., Wang, B., Xu, H., Liu, H., Jiang, M., Li, W., Wang, Y., Ye, A., Ren, G., Ma, Q., Liang, W., Lian, X., Wu, X., Zhong, Y., Li, Z., Gong, C., Lei, G., Cheng, L., Zhang, L., Li, M., Zhang, R., Hu, S., Huang, S., Wang, X., Zhao, Y., Wang, Y., Wei, Z., You, Y.: Open-sora 2.0: Training a commercial-level video generation model in 200k. arXiv preprint arXiv:2503.09642 (2025)
34. Ren, X., Shen, T., Huang, J., Ling, H., Lu, Y., Nimier-David, M., Müller, T., Keller, A., Fidler, S., Gao, J.: Gen3c: 3d-informed world-consistent video generation with precise camera control. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6121–6132 (2025)
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
36. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation. In: European Conference on Computer Vision. pp. 87–103. Springer (2024)
37. Seitzer, M.: pytorch-fid: FID Score for PyTorch. `https://github.com/mseitzer/pytorch-fid` (August 2020), version 0.3.0
38. Skorokhodov, I., Tulyakov, S., Elhoseiny, M.: Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2 (2021)
39. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
40. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2446–2454 (2020)
41. Tian, Q., Tan, X., Xie, Y., Ma, L.: Drivingforward: Feed-forward 3d gaussian splatting for driving scene reconstruction from flexible surround-view input. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 7374–7382 (2025)
42. Wang, Q., Fan, L., Wang, Y., Chen, Y., Zhang, Z.: Freevs: Generative view synthesis on free driving trajectory. In: International Conference on Learning Representations. vol. 2025, pp. 58640–58662 (2025)

43. Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., Lu, J.: Drivedreamer: Towards real-world-drive world models for autonomous driving. In: European conference on computer vision. pp. 55–72. Springer (2024)
44. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
45. Wei, D., Li, Z., Liu, P.: Omni-scene: Omni-gaussian representation for ego-centric sparse-view scene reconstruction. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22317–22327 (2025)
46. Wu, J.Z., Zhang, Y., Turki, H., Ren, X., Gao, J., Shou, M.Z., Fidler, S., Gojcic, Z., Ling, H.: Difix3d+: Improving 3d reconstructions with single-step diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26024–26035 (2025)
47. Yan, Y., Lin, H., Zhou, C., Wang, W., Sun, H., Zhan, K., Lang, X., Zhou, X., Peng, S.: Street gaussians: Modeling dynamic urban scenes with gaussian splatting. In: European Conference on Computer Vision. pp. 156–173. Springer (2024)
48. Yan, Y., Xu, Z., Lin, H., Jin, H., Guo, H., Wang, Y., Zhan, K., Lang, X., Bao, H., Zhou, X., et al.: Streetcrafter: Street view synthesis with controllable video diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 822–832 (2025)
49. Yang, J., Ivanovic, B., Litany, O., Weng, X., Kim, S.W., Li, B., Che, T., Xu, D., Fidler, S., Pavone, M., et al.: Emernerf: Emergent spatial-temporal scene decomposition via self-supervision. In: International Conference on Learning Representations. vol. 2024, pp. 16739–16766 (2024)
50. Yang, Z., Pan, Z., Yang, Y., Zhu, X., Zhang, L.: Driving view synthesis on free-form trajectories with generative prior. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 28083–28092 (2025)
51. 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. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
52. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
53. Zhao, G., Ni, C., Wang, X., Zhu, Z., Zhang, X., Wang, Y., Huang, G., Chen, X., Wang, B., Zhang, Y., et al.: Drivedreamer4d: World models are effective data machines for 4d driving scene representation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12015–12026 (2025)
54. Zhu, Z., Wu, Z., Zhu, Z., Zhou, L., Sun, H., Wan, B., Ma, K., Chen, G., Ye, H., Xie, J., et al.: Worldsplat: Gaussian-centric feed-forward 4d scene generation for autonomous driving. arXiv preprint arXiv:2509.23402 (2025)