Driving Scene Synthesis on Free-form Trajectories with Generative Prior

Zeyu Yang[1*] Zijie Pan[1*] Yuankun Yang[1*] Xiatian Zhu[2] Li Zhang[1✉]

[1] Fudan University   [2] University of Surrey

https://fudan-zvg.github.io/DriveX

Abstract

*Driving scene synthesis along free-form trajectories is essential for driving simulations to enable closed-loop evaluation of end-to-end driving policies. While existing methods excel at novel view synthesis on recorded trajectories, they face challenges with novel trajectories due to limited views of driving videos and the vastness of driving environments. To tackle this challenge, we propose a novel free-form driving view synthesis approach, dubbed **DriveX**, by leveraging video generative prior to optimize a 3D model across a variety of trajectories. Concretely, we crafted an inverse problem that enables a video diffusion model to be utilized as a prior for many-trajectory optimization of a parametric 3D model (e.g., Gaussian splatting). To seamlessly use the generative prior, we iteratively conduct this process during optimization. Our resulting model can produce high-fidelity virtual driving environments outside the recorded trajectory, enabling free-form trajectory driving simulation. Beyond real driving scenes, DriveX can also be utilized to simulate virtual driving worlds from AI-generated videos.*

1. Introduction

Building virtual driving worlds engagible with driving policies plays a pivotal role in developing robust autonomous driving systems. It enables automatic synthesis of diverse driving data, including safety-critical long-tail scenarios, edge cases, and providing closed-loop evaluation for end-to-end autonomous driving systems.

Recent advancements in novel view synthesis [15, 21] have driven reconstruction-based driving simulation [4, 34, 35] due to its ability to synthesize photorealistic sensor data. These methods aim to reconstruct driving environments from individual driving sequences, typically involving single-trajectory videos from surrounding cameras with sparse views, minimal overlap, and textureless areas. De-

spite impressive results in fitting training views and interpolating along original trajectories, they still struggle to extrapolate substantially novel views far from the recorded trajectory, leading to limited flexibility and usefulness for driving simulation. To address these problems, a couple of latest works [13, 38] explored image generative prior as extra novel view regularization. However, they are largely limited in design due to **(i)** the need for fine-tuning an off-the-shelf image generation models to obtain the capability of view transformation, suffering from both additional computational cost and per-scene data scarcity, and **(ii)** absence of the desired spatiotemporal knowledge within image models. To tackle these issues, a concurrent work [40] uses a pretrained video diffusion model to generate novel trajectory training images conditioned on initial recorded frames. Without sufficient content constraint but with intrinsic hallucination risk, this approach however tends to drift away from the original scene in scene generation, making troubles for the subsequent model optimization.

To address the challenges mentioned above, we propose a novel driving scene synthesis approach for free-form trajectories, termed **DriveX**, by leveraging video generative prior with rich spatiotemporal knowledge during the model (*e.g.*, Gaussian splatting) optimization. This is achieved by designing an inverse problem that enables a video diffusion model to serve as a prior for optimizing the Gaussian over a variety of trajectories. During training, we render novel trajectory views using the in-training model. To identify potential artifacts from trajectory extrapolation, we compare each rendered view with recorded images wrapped under the same view and geometric conditions, allowing us to obtain reliable regions for each rendered image. The inverse problem is then crafted as restoring each rendered view via utilizing its corresponding reliable regions as conditioning inputs, which can be effectively executed by the video diffusion model. The output from this process is subsequently fed into the optimization as supervision to update the target model. To seamlessly exploit the generative prior, we iteratively conduct this process during Gaussian optimization. As shown in Fig. 1, our approach substantially improve the view synthesis quality in novel trajectory compared to the

---

Figure 1. Though conventional reconstruction method can well fit views on the recorded trajectory, they struggle to extrapolate onto the views along the novel trajectories (**top**). In contrast, our method significantly improves the synthesis quality on these novel views (**bottom**), thereby achieving driving scene synthesis on free-form trajectories.

traditional reconstruction methods.

The contributions of this paper are as follows: **(i)** We propose leveraging video generative prior with rich spatiotemporal knowledge for generalizable driving scene synthesis *from single-trajectory recorded videos.* **(ii)** We introduce DriveX, a novel driving scene synthesis framework for free-form trajectories that innovatively constructs an inverse problem to enable the use of a video diffusion model as a prior. **(iii)** Extensive experiments demonstrate that DriveX significantly outperforms existing state-of-the-art alternatives in driving scene synthesis using single-trajectory recorded videos. Additionally, we showcase the superiority of our method in rendering novel trajectories of virtual driving worlds from AI-generated videos, even when faced with inherent content inconsistencies. This approach enables a more economically scalable large-scale simulation without the need to collect videos for each scene.

2. Related works

Reconstruction-based driving scene synthesis Early approaches [27, 28, 32, 34, 36] for dynamic urban scenes reconstruction focus on extending neural radiance fields (NeRF) [21] in large scale unbounded scenes.These methods are restricted with low efficiency and blurry rendering performance. Recently, another line of works introduced 3D Gaussian splatting [15] into this task to produce real-time high-fidelity rendering. Some of them [35, 42] leverage its favorable explicit property to decouple the whole scene into the static background and moving foreground Gaussians, while others modeling the motion of 3D Gaussians with periodic functions [4] or deformation fields [12]. However, driving scenes featured with sparse views, little overlaps, and large textureless areas, making existing methods struggle to generalize to substantively novel views.

Generative-based driving scene synthesis Recently, pro-

pelled by the significant advancement of the diffusion model [1, 10, 22, 24, 25], many works employ video generators to construct world models [16]. Some of them introduce this paradigm into autonomous driving [6, 7, 11, 17, 20, 29, 30, 41] for simulating the plausible visual outcomes under versatile driving controls. Despite the remarkable diversity of generated results, this approach is constrained by the absence of underlying 3D model. Consequently, they cannot promise geometry and texture consistency for the generated many-trajectory videos of the same scene.

Scene reconstruction with diffusion prior Leveraging generative prior to enhance the reconstruction from sparse view observations [18, 19, 37] has emerged. Extending this paradigm into driving scene reconstruction has been recently attempted [13, 38, 40]. However, they typically suffer from limited diversity and scale of driving scene data. SGD [38] and VEGS [13] contribute to define and tackle pseudo and extrapolated view synthesis in driving scenarios. However, these methods may encounter challenges with accurate pose and detail estimation under novel trajectories. GGS [8] introduces virtual lane generation module, but its dependency on modified MVSplat structure poses challenges for generalization to large-scale pretraining. A concurrent work [40] uses a world model [29] to generate novel trajectory video referenced on initial frame for trajectory augmentation. However, it tends to drift away from the original scene due to the strong built-in hallucination risk and insufficient constraints. This would inevitably hurt the subsequent model optimization. In contrast, our method dedicates to solve an inverse problem that enables a video diffusion model to be used as a prior for many-trajectory optimization of a parametric 3D model, achieving superior synthesis quality on free-form trajectories.
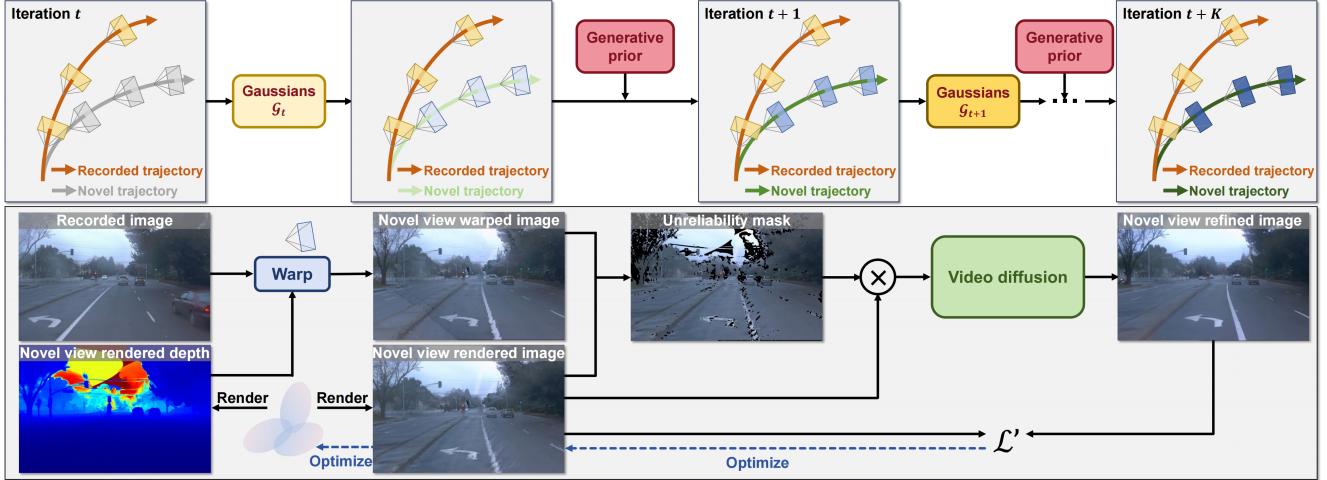
Figure 2. **The schematic illustration of DriveX**. **Top: Gaussian optimization** with both recorded trajectory and generated novel trajectory as supervision. **Bottom: Generative prior integration.** Given a novel view rendered image from Gaussians $\mathcal{G}_t$, video diffusion is applied to generate its refined version with condition of unreliability mask. We omit the loss on recorded trajectory for clarity.

3. Methodology

3.1. Preliminary: Driving scene reconstruction

Reconstruction-based driving simulation aims to recover the surrounding scenes from calibrated video $\mathcal{V}_{\text{gt}} = \{I_i\}_{i=0}^F$ recorded along a driving trajectory $\mathcal{T} = \{P_i\}_{i=0}^F$, where $P_i \in \text{SE}(3)$ is the camera pose at frame i. Then the sensor data would be rendered when the vehicle performs simulated unrecorded actions (*e.g.* lane change) in the re-constructed environment.

Recently, due to the superior fidelity and efficiency of 3D Gaussian splatting (3DGS) [15], it has been widely employed in numerous street scene reconstruction efforts [4, 12, 35, 42] as a fundamental scene representation. 3DGS represents scenes as a set of 3D Gaussians, denoted $\mathcal{G}$. The influences of each Gaussian primitive with position $\mathbf{x}_k \in \mathbb{R}^3$ and covariance $\mathbf{\Sigma} \in \mathbb{R}^{3 \times 3}$ on any position is given by an unnormalized Gaussian function $G_k(\cdot; \mathbf{x}_k, \mathbf{\Sigma})$ weighted by its opacity $o_k \in \mathbb{R}$. Given *any* camera pose P_i, it can be rendered by performing splatting and alpha blending on N sorted Gaussians visible at this view:

$$I' = R(\mathcal{G}, P_i) = \sum_{k=1}^N o_k G_k c_k(P_i) \prod_{j=1}^{k-1}(1 - o_j G_j), \quad (1)$$

where R represents the differentiable Gaussian rasterizer, I' denotes the rendered image, c_k is the view-dependent color of the k-th Gaussian. Then the Gaussians $\mathcal{G}$ can be optimized via the photometric loss between rendered images I'_i and ground truth I_i from recorded video $\mathcal{V}_{\text{gt}}$ by:

$$\mathcal{L}_{\text{img}}(I'_i, I_i) = \lambda\|I'_i - I_i\|_1 + (1-\lambda)\mathcal{L}_{\text{SSIM}}(I'_i, I_i), \quad (2)$$

where $\mathcal{L}_{\text{SSIM}}$ is the SSIM [31] loss and λ is a weight.

In the urban scenes, although the traffic participants are dynamic, they can be regarded as static in their own local coordinate frames. The transformation from each object's local frame to the global coordinate frame of the whole scene is parameterized by a series of rotations and transformations, which can be derived from its tracklet annotated by off-the-shelf 3D trackers [5, 33]. Then a complete dynamic street scene can be modeled by decoupled foreground moving Gaussians and static background Gaussians. Additionally, cube map is commonly used to model the sky with infinite distance [4, 35].

Challenges Since the camera configuration in autonomous vehicles is primarily designed for perception tasks, the captured videos often offer limited viewpoints, minimal overlap, and large textureless regions. These inherent properties bring challenges to the reconstruction of the unbounded driving scenes from single-trajectory videos. Fitting the training views at recorded trajectory $\mathcal{T}$ is insufficient to realize satisfactory extrapolation out of $\mathcal{T}$ for novel view synthesis. As a result, existing optimization-based driving scene synthesis approaches struggle to achieve high-quality view rendering along novel trajectories $\mathcal{T}' = \{P'_i | P'_i \notin \mathcal{T}\}_{i=0}^F$, often leading to obvious degeneration and artifacts.

3.2. DriveX

To address the aforementioned challenges, we propose a novel framework, DriveX, *seamlessly* integrating generative priors into the driving scene reconstruction process (*e.g.* Gaussian optimization). As shown in Fig. 2, it iteratively synthesizes novel trajectory scenes by utilizing a video diffusion as supervision to optimize the Gaussian model. To ensure the generated content aligns with the underlying scenes in detail, the generation process is appro-

3

priately conditioned on the in-training Gaussian model.

To achieve this, we delicately design an inverse problem [14, 23], which can be formulated as recovering clean novel view images $\mathcal{V}$ from the artifact-heavy novel view rendered images $\mathcal{V}'$ on novel trajectories $\mathcal{T}'$. Concretely, $\mathcal{V}'$ can be regarded as a measurement of unknown ground truth $\mathcal{V}$, as:

$$\mathcal{V}' = f(\mathcal{V}) + \epsilon, \tag{3}$$

where ϵ is a random noise, and the measurement function f distorts $\mathcal{V}$ due to poor generalization ability of Gaussian model in conventional methods. With this design, a video generative model [37] can be readily employed to address this problem effectively. The refined video can then serve as additional supervision within Gaussian optimization.

Solving inverse problem with generative prior The key to solve the inverse problem lies in accurately identifying the artifacts in the novel view rendered video. Inspired from the sparse view reconstruction methods [19, 34], we construct an unreliability mask representing the artifacts via comparing the rendered images with recorded images, which is used to directly guide the diffusion model focus on refining these unreliable regions while preserving reliable areas.

Specifically, given a sequence of viewpoints $\mathcal{T}' = \{P'_i\}_{i=0}^{F}$ along a novel trajectory, a video diffusion model $\mathcal{D}$ is employed to estimate $\mathcal{V}_i$ in Eq. (3):

$$\mathcal{V}'_t = \{R(\mathcal{G}_t, \ P'_i)\}_{i=0}^{F}, \tag{4}$$

$$\mathcal{V}'_{t,\mathrm{refine}} = \mathcal{D}\left(\mathcal{V}'_t, \ \mathcal{M}\right), \tag{5}$$

$$\mathcal{L}' \triangleq \mathcal{L}_{\mathrm{img}}(\mathcal{V}'_t, \ \mathcal{V}'_{t,\mathrm{refine}}), \tag{6}$$

where $\mathcal{V}'_t$ is rendered from the Gaussian model $\mathcal{G}_t$ at current iteration t through Eq. (1), the refined video $\mathcal{V}'_{t,\mathrm{refine}}$ can be regard the estimation of $\mathcal{V}$ in Eq. (3), and $\mathcal{M}$ is the mask indicating unreliable regions. Typically, $\mathcal{D}$ takes noise-perturbed images as input, conditioned on the masked video, and outputs the refined video $\mathcal{V}'_{t,\mathrm{refine}}$. Then $\mathcal{V}'_{t,\mathrm{refine}}$ is used as the additional supervision for the novel trajectory loss $\mathcal{L}'$ (Eq. (6)) to improve the novel view rendering. Note we omit the loss on recorded trajectory [35] in the whole context for clarity.

To assess geometry accuracy in novel view rendered images, we derive $\mathcal{M}$ by comparing the rendered images with novel view warped images from the recorded trajectory. Let $I_{\mathrm{ren}} \in \mathcal{V}'_t$ denote a rendered image, and let I_{rec} be the recorded image that are closest to I_{ren}. We define a warping operation ψ in Eq. (7) to project a 3D point p onto the image plane given camera pose P:

$$(x, y, d) = \psi(p|P), \tag{7}$$

where x, y represent the pixel coordinates, and d is the depth (Please refer to Appendix for details). Notably, ψ is invertible, allowing us to unproject an image with its corresponding depth map back to 3D points given the camera pose.

Algorithm 1 Iterative refinement

Require: Initial Gaussian model $\mathcal{G}_0$, novel trajectories $\mathcal{T}'$, video diffusion $\mathcal{D}$, total steps T, warm up steps T_0
1: **for** $t = 0, \cdots, T - 1$ **do**
2:    $\mathcal{V}_t \leftarrow \{R(\mathcal{G}_t, P_i)\}_{i=0}^{F}$
3:    Computing loss $\mathcal{L}_{\mathrm{img}}(\mathcal{V}_t, \mathcal{V}_{\mathrm{gt}})$
4:    **if** $t \geq T_0$ **then**
5:       $\mathcal{V}'_t \leftarrow \{R(\mathcal{G}_t, P'_i)\}_{i=0}^{F}$
6:       **if** $(t - T_0) \mod K = 0$ **then**
7:          $k \leftarrow t$
8:          $\mathcal{V}'_{k,\mathrm{refine}} \leftarrow \mathcal{D}\left(\mathcal{V}'_t, \mathcal{M}\right)$
9:       **end if**
10:      Computing loss $\mathcal{L}' \leftarrow \mathcal{L}_{\mathrm{img}}(\mathcal{V}'_t, \mathcal{V}'_{k,\mathrm{refine}})$
11:    **end if**
12:    Backwarding loss and updating $\mathcal{G}_{t+1}$
13: **end for**
14: **return** $\mathcal{G}_T$

Consequently, we can obtain a pseudo image $\hat{I}_{\mathrm{ren}}$ under the pose P_{ren} by sampling the color in I_{rec} using the warped pixel coordinates $(\mathbf{x}, \mathbf{y}, \mathbf{d})$ from unwarped I_{ren} to I_{rec}, as expressed by Eq. (8).

$$(\mathbf{x}, \mathbf{y}, \mathbf{d}) = \psi(\psi^{-1}(I_{\mathrm{ren}}, D_{\mathrm{ren}}|P_{\mathrm{ren}})|P_{\mathrm{rec}}), \tag{8}$$

where D_{ren} is the rendered depth corresponding to $\hat{I}_{\mathrm{ren}}$. Since it is hard to guarantee the pixel-wise correspondence between I_{ren} and $\hat{I}_{\mathrm{ren}}$, we choose SSIM to evaluate similarity at the patch-structure level. The unreliability mask $\mathcal{M}$ is then obtained by applying a threshold τ to SSIM score:

$$\mathcal{M} = \mathbb{1}\left(\mathrm{SSIM}(I_{\mathrm{ren}}, \hat{I}_{\mathrm{ren}}) < \tau\right). \tag{9}$$

Because the warping process involves both rendered depth and rendered image, the discrepancy can serve as an indicator of geometric or appearance unreliability.

Since the warped image provides a rough overview about the novel trajectory, making it unnecessary to start denoising from scratch when we refine the videos using Eq. (5). Hence, we perturb $\hat{I}_t$ by adding Gaussian noise at some noise level. This consideration not only reduces unnecessary time costs but also avoids color shifts by preserving accurate low-frequency components in original video.

Iterative refinement We have detailed how to use generative prior to solve a video inverse problem to improve the reconstruction quality. In turn, the generation can also benefit from these improvements, as it is conditioned on $\mathcal{G}_t$ whose enhancements could provide more accurate information for generation. This reciprocal enhancement motivates an *iterative refinement*. As illustrated in Algorithm 1, we first warm up the optimization for T_0 steps using conventional reconstruction method [35] on recorded trajectory, and then
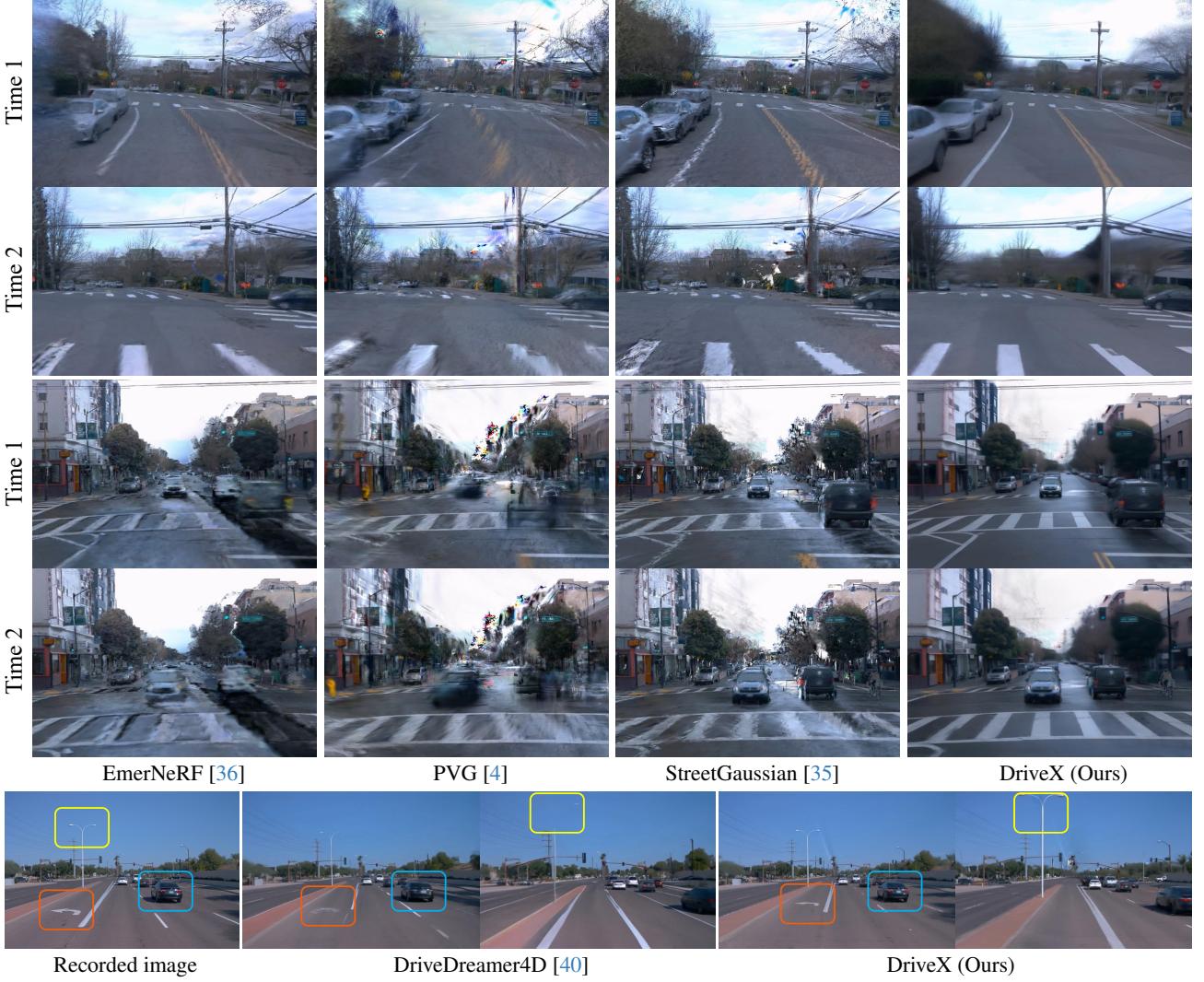
4

Figure 3. **Qualitative comparisons of novel trajectories on Waymo**. **Top:** For each method, images are rendered to the left of the recorded trajectory by $3m$. **Bottom:** Our results are consistent with the recorded image compared to DriveDreamer4D [40].

store the refined video $\mathcal{V}'_{t,\text{refine}}$ as a buffer, which will be updated every K steps. Due to the additional time demands of running video diffusion model, it is inefficient to set $K = 1$, *i.e.*, the buffer is updated every step. Please refer to Sec. 4.3 for the trade-off between different K. Since the generated contents are conditioned on Gaussian model $\mathcal{G}_t$, they are constrained with the recorded scene, thus ensuring the consistency between reconstruction and generation.

Novel trajectory sampling Another important design space is how to sample the novel trajectory $\mathcal{T}' = \{P'_i\}_{i=0}^F$. The ideal trajectory should balance the following two requirements: first, maximizing the utility of the generative model, *i.e.*, the generated novel views should contribute as much as possible to enhance reconstruction quality; second, minimizing inconsistency between the recorded and gener-

ated images. Therefore, we adopt a panning camera trajectory that starts from a recorded front view $P'_0 = [R_0|T_0]$ and gradually shifts laterally:

$$P'_i = [R_0|\frac{i}{F}s\mathbf{v} + T_0], \tag{10}$$

where $\mathbf{v} \in \mathbb{R}^3$ is the shifting direction, and $s \in \mathbb{R}$ controls the maximum shifting length. This design not only provides novel views beyond the recorded trajectory, but also allows video generation model to extract detailed references from the initial frame. At the early stage of optimization, rendering quality may deteriorate as the camera diverges from the recorded trajectory, which offers limited guidance for video generation. To address this issue, we initially use a small shifting length s and progressively extend it during optimization, ensuring our method's stability and robust-

ness across diverse scenes.

3.3. Application on generated videos

In addition to real-world scenes, the driving videos used for reconstruction can also be created by a video generator [7], which explores new possibilities for the automatic creation of diverse virtual driving environments. However, the AI-generated videos present more severe challenges because of the lack of corresponding LiDAR depth and accurate pixel-to-pixel matching between different frames. These challenges further necessitate the novel view supervision provided by the generative priors.

Therefore, we devise an additional recipe for the robust optimization of the AI-generated videos. Specifically, since the generated videos cannot be precisely aligned with the conditioned camera parameters, we employ a feed-forward approach [39] to estimate the camera parameters and corresponding depths for each frame. For scenes containing moving vehicles, we manually annotate their tracklets as the initialization for the local coordinate frames. During optimization, we fine-tune the camera pose to mitigate potential prediction errors. In the computation of unreliability, as the generated video is monocular, we use the nearest previous frames that can include the current shifted view content as the auxiliary source images in Eq. (8) to ensure well-defined unreliability scores for all areas of the novel view.

4. Experiments

4.1. Experimental setup

Datasets For evaluation of the improvement for real-world reconstruction, we conduct experiments on 12 selected sequences comprising surrounding videos and synchronized LiDAR point clouds from Waymo Open Dataset [26]. Official tracklets are used to crop the point cloud of dynamic traffic participants and initialize their per-frame pose. For fair comparison, all images are downsampled into 640×920 in both training and testing. To test the applicability of DriveX in generated driving videos, we assess our method on 12 driving scene videos produced by video diffusion models [7]. Each video is generated given the initial frame from nuScenes validation set as the condition.

Metrics To reasonably benchmark the improvement on the existing reconstruction-based street view synthesis pipeline, we quantitatively compared DriveX with several representative baselines on views from the novel trajectory. Since the ground truth sensor data on these trajectories is unavailable, it is infeasible to directly assess the *pixel-level* fidelity. Therefore, we devised a novel benchmark to comprehensively evaluate the synthesis results at these views. Specifically, the Fréchet Inception Distance (FID) [9] between synthesized views in novel trajectories and the captured images from the original trajectories is employed to assess the re-

alism of synthesized images at the *distribution level*. Moreover, we report Lane IoU and vehicle AP to evaluate the *distinguishability and fidelity of two safety-critical high-level traffic components*, *i.e.*, road lanes and vehicles, in novel trajectory viewpoints. Refer to the appendix for detailed calculations of these metrics.

Implementation details For reconstruction of real video where calibrated data and LiDAR are available, we first warm up the optimization with Gaussian model [35] for 50,000 steps and then integrate generative priors for following 30,000 steps. For novel trajectory videos, the initial view P_0' of a novel trajectory $\mathcal{T}'$ is selected from every 3 recorded forward cameras, with shifting direction $\mathbf{v}$ oriented to the left or right, and shifting length s ranging from $2m$ to $4m$. During the iterative refinement, the buffer for storing refined videos is updated every $K = 2000$ steps. We use ViewCrafter [37] as our generative prior. The refine strength (*i.e.* the level of noise adding to images) is set to 0.6. The threshold τ in Eq. (9) is set to 0.65. To alleviate the conflicts between recorded data and generated results on moving cars, their supervision strength is reduced to 0.2. For generated videos, we first estimate the camera pose [39] and predict the depth and then perform similar process.

4.2. Comparison with state of the art

Results on real-world videos For comparisons on real-world reconstruction, we primarily focused on improvements in substantial novel views beyond the recorded trajectory. To this end, we synthesized novel driving videos along trajectories shifting from the recorded trajectories by $\pm 1m$, $\pm 2m$ and $\pm 3m$. Table 1 demonstrate our significant improvement on FID, AP and IoU, surpassing the leading StreetGaussian [35] with a 9.3% decrease in FID, 1.6% increase in AP and 21.1% increase in IoU under the largest $3m$ shifting length. These enhancements are visually corroborated in Fig. 3, where safety-critical scene elements can be well synthesized in novel views by our methods. It can be observed that by integrating generative priors, our approach demonstrates robust rendering quality under large trajectory deviations, resulting in more realistic driving environments. Notably, once trained, **our model solely relies on the Gaussians for inference, enabling efficient rendering speed.**

For fair comparison with DriveDreamer4D [40] which is not open-source, we adopt the same experiment setting (sequences, metric and its best reported Gaussian counterpart [4]). The results are presented in Tab. 2 and bottom of Fig. 3. Our model markedly outperforms DriveDreamer4D with variety of Gaussian models on all metrics, visual quality, and the consistency of traffic components.

Results on generated videos Another intriguing application of the proposed method is transforming AI-generated

	±1m			±2m			±3m			Inference speed
	IoU↑	AP↑	FID↓	IoU↑	AP↑	FID↓	IoU↑	AP↑	FID↓	(FPS)
EmerNeRF [36]	0.1676	0.5149	91.40	0.1372	0.4671	111.67	0.1369	0.4198	132.48	0.12
S^3Gaussian [12]	0.1423	0.5301	70.71	0.0865	0.4781	100.29	0.0430	0.4341	126.22	8.4
PVG [4]	0.1335	0.5406	71.63	0.0632	0.4686	102.92	0.0365	0.4267	137.12	**48**
StreetGaussian [35]	0.2002	0.5945	49.44	0.1668	0.5793	80.05	0.1385	0.5342	105.37	34
DriveX (Ours)	**0.2065**	**0.5949**	**48.65**	**0.1892**	**0.5827**	**76.44**	**0.1755**	**0.5427**	**95.60**	34

Table 1. **Novel trajectory evaluation on Waymo.** We compare images rendered in novel trajectories with different shift lengths, where the IoU is computed between the ground truth and those detected by TwinLiteNet [3]. **Once trained, our model solely relies on the Gaussians for inference.** The inference speeds of all methods are measured on the NVIDIA RTX A6000 GPU.

	Gaussian model	Lane change		Acceleration		Deceleration		Average	
		NTA-IoU↑	NTL-IoU↑	NTA-IoU↑	NTL-IoU↑	NTA-IoU↑	NTL-IoU↑	NTA-IoU↑	NTL-IoU↑
DriveDreamer4D [40]	PVG [4]	0.428	53.00	0.411	53.10	0.421	53.78	0.420	53.29
DriveX (Ours)	PVG [4]	0.474	58.53	0.563	65.36	0.611	66.97	0.549	63.62
DriveX (Ours)	StreetGauss. [35]	0.516	62.77	0.569	65.14	0.599	66.23	0.561	64.71

Table 2. **Comparision with DriveDreamer4D [40]** under its evaluation setting. For fair comparison, we report the same **metric** proposed in [40] measured on the same **sequences**, and same **Gaussian model** [4] with the best reported results in [40].

Method	FID ↓		
	±0m	±1m	±2m
Recon-only [35]	45.45	96.76	146.61
DriveX (Ours)	**45.44**	**92.77**	**142.73**

Table 3. **Quantitative comparison on generated video.** We compared the FID score on videos generated by Vista [7].

Gaussian model	w/ ours	IoU↑	AP ↑	FID↓
PVG [4]		0.0200	0.5401	118.41
PVG [4]	✓	**0.0819**	**0.6188**	**101.04**
StreetGaussian [35]		0.1217	0.6124	103.08
StreetGaussian [35]	✓	**0.2092**	**0.6257**	**90.30**

Table 4. **Ablation study on different Gaussian model.** Unless stated otherwise, all ablation experiments are measured on views at the novel trajectory shifted $3m$ from the recorded trajectory.

	IoU↑	AP ↑	FID↓
mask all	0.0899	0.5781	107.58
w/o mask	0.1106	0.5965	90.49
w/ mask	**0.2092**	**0.6257**	**90.30**

Table 5. **Effects of unreliability mask.**

Buffer interval	IoU↑	AP ↑	FID↓	Time↓
500	0.1070	0.5954	**86.56**	4.8×
2,000	**0.2092**	**0.6257**	90.30	1.7×
5,000	0.1104	0.5992	105.25	1.4×
Inf	0.0945	0.5845	95.08	1.0×

Table 6. **Ablations on buffer intervals.** *Inf* denotes the novel trajectory supervision only generated once time at iteration T_0.

video into a re-enactable driving world. To quantitatively evaluate this capability, we deploy DriveX and the reconstruction-only baseline [35] on 12 driving videos generated by [7]. The FID of videos rendered across different novel trajectories is reported in Tab. 9. The results indicate that the realism of rendered videos is well preserved within a certain range of shifted length. Compared to the reconstruction-only baseline, our DriveX achieves 4.1% and 2.6% FID reduction in shifting length of ±1m and ±2m respectively. To obtain real world distance, we align the predicted depth and LiDAR depth of the first frame by the least square estimation. These advancements are also clearly illustrated in Fig. 4. As can be seen, the baseline method exhibits significant degradation in views that are far from the recorded trajectory.

4.3. Ablation study

Different Gaussian model Although we adopt StreetGaussian [35] as the default Gaussian counterpart in previous experiments, our method is also capable of synergy with different choices. To demonstrate the versatility of DriveX, we

evaluate on a dynamic counterpart PVG [4]. From Tab. 4,

Figure 4. **Reconstruction results on generated videos.** All ground truth videos used for reconstruction are generated by Vista [7]. The first and the second rows show the novel trajectory synthesis results by the baseline [35] and our method, respectively.
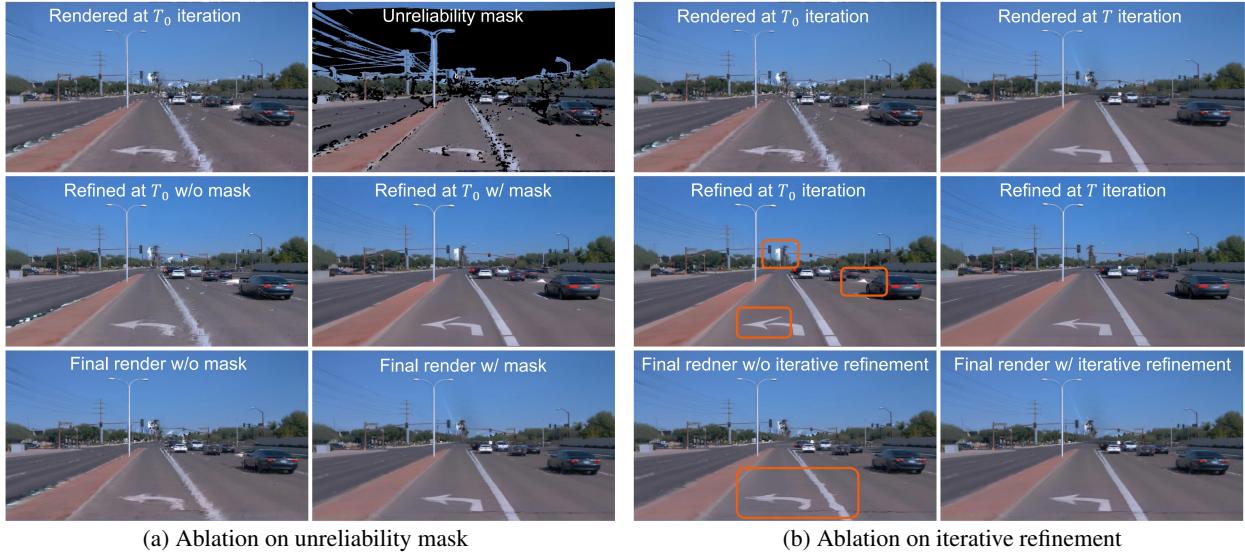

(a) Ablation on unreliability mask

(b) Ablation on iterative refinement

Figure 5. **Ablations on (a) unreliability mask and (b) iterative refinement.** Fig.(a) shows the refined results without (middle left) and with (middle right) unreliability mask (top right). The final rendered results exhibit significant artifacts without mask (bottom left). Fig.(b) shows the rendered (top) and restored results at iteration T_0 (middle left) and T (middle right). It can be seen that single-time refinement can not completely eliminate artifacts, resulting in inferior final rendered results compared to iterative refinement (bottom). T_0 and T indicate the first and last iteration after warm up, respectively.

we can observe that our strategy consistently reduces the FID by a large margin, and effectively recovering street lanes and cars (*i.e.* IoU, AP) that are entirely indiscernible in the baseline methods. Unless stated otherwise, all ablation experiments are conducted on two Waymo sequences.

Unreliability mask The unreliability mask is designed to guide the generation of video diffusion. From Tab. 5 and Fig. 5(a) we can observe: (i) When all regions are masked, where video diffusion model does not know any information from the in-training Gaussian, leading to inferior results; (ii) When the mask is disabled, the generative model can still refine the rendered results toward the video distribution, but with minor refinement; (iii) Equipped with proposed unreliability mask, our full model achieves superior performance in both quantitative and qualitative ablation.

Iterative refinement The interval of iterative refinement offers the trade-off between quality and efficiency. When the novel trajectory supervision only generated at iteration T_0 (see Fig. 5(b)), the quality of the novel view is constrained to the in-training Gaussian model which is not fully optimized. To the other extreme, we are unable to afford the time cost of running the diffusion at every step. From Tab. 6, a small interval 500 leads to $4.8\times$ training time. Hence we finally chose the interval 2,000 to strike the balance.

5. Conclusion

In this paper, we present a novel framework that integrates generative priors into the reconstruction of the driving scene from single-trajectory recorded videos, addressing limitations in extrapolating novel views beyond a recorded trajectory. To reasonably leverage the video generative model

to provide consistent novel view supervision, we construct a tailored inverse problem to refine the novel view rendered images and progressively enhance the street scene model. The extensive experiments demonstrate that the proposed method significantly improves the synthesis quality for novel trajectories, both in real and AI-generated scenes. These results pave the new way for driving scene synthesis on free-form trajectories.

References

[1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint*, 2023. 2

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 1, 2

[3] Quang-Huy Che, Dinh-Phuc Nguyen, Minh-Quan Pham, and Duc-Khai Lam. Twinlitenet: An efficient and lightweight model for driveable area and lane segmentation in self-driving cars. In *MAPR*, 2023. 7, 1

[4] Yurui Chen, Chun Gu, Junzhe Jiang, Xiatian Zhu, and Li Zhang. Periodic vibration gaussian: Dynamic urban scene reconstruction and real-time rendering. *arXiv preprint*, 2023. 1, 2, 3, 5, 6, 7

[5] Lue Fan, Feng Wang, Naiyan Wang, and Zhao-Xiang Zhang. Fully sparse 3d object detection. In *NeurIPS*, 2022. 3

[6] Ruiyuan Gao, Kai Chen, Enze Xie, HONG Lanqing, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. In *ICLR*, 2023. 2

[7] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. In *NeurIPS*, 2024. 2, 6, 7, 8

[8] Huasong Han, Kaixuan Zhou, Xiaoxiao Long, Yusen Wang, and Chunxia Xiao. Ggs: Generalizable gaussian splatting for lane switching in autonomous driving. *arXiv preprint*, 2024. 2, 1

[9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017. 6

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 2

[11] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023. 2

[12] Nan Huang, Xiaobao Wei, Wenzhao Zheng, Pengju An, Ming Lu, Wei Zhan, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. s^3gaussian: Self-supervised street

[13] Sungwon Hwang, Min-Jung Kim, Taewoong Kang, Jayeon Kang, and Jaegul Choo. Vegs: View extrapolation of urban scenes in 3d gaussian splatting using learned priors. In *ECCV*, 2024. 1, 2

[14] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *NeurIPS*, 2022. 4

[15] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 2023. 1, 2, 3

[16] Yann LeCun. A path towards autonomous machine intelligence. *Open Review*, 2022. 2

[17] Xiaofan Li, Yifu Zhang, and Xiaoqing Ye. Drivingdiffusion: Layout-guided multi-view driving scene video generation with latent diffusion model. In *ECCV*, 2023. 2

[18] Fangfu Liu, Wenqiang Sun, Hanyang Wang, Yikai Wang, Haowen Sun, Junliang Ye, Jun Zhang, and Yueqi Duan. Reconx: Reconstruct any scene from sparse views with video diffusion model. *arXiv preprint*, 2024. 2

[19] Xi Liu, Chaoyi Zhou, and Siyu Huang. 3dgs-enhancer: Enhancing unbounded 3d gaussian splatting with view-consistent 2d diffusion priors. In *NeurIPS*, 2024. 2, 4

[20] Jiachen Lu, Ze Huang, Zeyu Yang, Jiahui Zhang, and Li Zhang. Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation. In *ECCV*, 2025. 2

[21] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1, 2

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

[23] Litu Rout, Negin Raoof, Giannis Daras, Constantine Caramanis, Alex Dimakis, and Sanjay Shakkottai. Solving linear inverse problems provably via posterior sampling with latent diffusion models. In *NeurIPS*, 2024. 4

[24] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2020. 2

[25] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint*, 2023. 2

[26] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 6, 1

[27] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *CVPR*, 2022. 2

[28] Haithem Turki, Jason Y Zhang, Francesco Ferroni, and Deva Ramanan. Suds: Scalable urban dynamic scenes. In *CVPR*, 2023. 2

gaussians for autonomous driving. *arXiv preprint*, 2024. 2, 3, 7

[29] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint*, 2023. 2

[30] Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *CVPR*, 2024. 2

[31] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. In *IEEE TIP*, 2004. 3, 1

[32] Zian Wang, Tianchang Shen, Jun Gao, Shengyu Huang, Jacob Munkberg, Jon Hasselgren, Zan Gojcic, Wenzheng Chen, and Sanja Fidler. Neural fields meet explicit geometric representations for inverse rendering of urban scenes. In *CVPR*, 2023. 2

[33] Hai Wu, Wenkai Han, Chenglu Wen, Xin Li, and Cheng Wang. 3d multi-object tracking in point clouds based on prediction confidence-guided data association. In *IEEE TITS*, 2021. 3

[34] Ziyang Xie, Junge Zhang, Wenye Li, Feihu Zhang, and Li Zhang. S-nerf: Neural radiance fields for street views. In *ICLR*, 2023. 1, 2, 4

[35] Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng. Street gaussians: Modeling dynamic urban scenes with gaussian splatting. In *ECCV*, 2024. 1, 2, 3, 4, 5, 6, 7, 8

[36] Jiawei Yang, Boris Ivanovic, Or Litany, Xinshuo Weng, Seung Wook Kim, Boyi Li, Tong Che, Danfei Xu, Sanja Fidler, Marco Pavone, et al. Emernerf: Emergent spatial-temporal scene decomposition via self-supervision. In *ICLR*, 2024. 2, 5, 7

[37] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint*, 2024. 2, 4, 6

[38] Zhongrui Yu, Haoran Wang, Jinze Yang, Hanzhang Wang, Zeke Xie, Yunfeng Cai, Jiale Cao, Zhong Ji, and Mingming Sun. Sgd: Street view synthesis with gaussian splatting and diffusion prior. In *WACV*, 2024. 1, 2

[39] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint*, 2024. 6

[40] Guosheng Zhao, Chaojun Ni, Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Boyuan Wang, Youyi Zhang, Wenjun Mei, and Xingang Wang. Drivedreamer4d: World models are effective data machines for 4d driving scene representation. *arxiv preprint*, 2024. 1, 2, 5, 6, 7

[41] Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. *arXiv preprint*, 2024. 2

[42] Xiaoyu Zhou, Zhiwei Lin, Xiaojun Shan, Yongtao Wang, Deqing Sun, and Ming-Hsuan Yang. Drivinggaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes. In *CVPR*, 2024. 2, 3

Driving Scene Synthesis on Free-form Trajectories with Generative Prior

Supplementary Material

6. More implementation details

Used sequences The results in Table 1 of the main text are computed from 12 sequences curated from WOD [26]. The names of all sequences are listed below:

- `segment-10444454289801298640`
- `segment-10625026498155904401`
- `segment-6242822583398487496`
- `segment-8822503619482926605`
- `segment-10359308928573410754`
- `segment-11450298750351730790`
- `segment-12496433400137459534`
- `segment-15021599536622641101`
- `segment-16767575238225610271`
- `segment-17860546506509760757`
- `segment-3015436519694987712`
- `segment-6637600600814023975`

For the reconstruction from generated videos, the videos are generated given the following images from nuScenes [2] validation set:

- `n015-2018-09-25-13-17-43+0800__CAM_FRONT__1537853021662460`
- `n015-2018-10-08-15-36-50+0800__CAM_FRONT__1538984433512460`
- `n015-2018-09-25-13-17-43+0800__CAM_FRONT__1537852899162460`
- `n015-2018-10-08-15-44-23+0800__CAM_FRONT__1538984912912460`
- `n015-2018-10-08-15-44-23+0800__CAM_FRONT__1538984834412460`
- `n015-2018-10-08-15-44-23+0800__CAM_FRONT__1538984833912460`
- `n015-2018-07-18-11-41-49+0800__CAM_FRONT__1531885347362460`
- `n015-2018-10-08-15-36-50+0800__CAM_FRONT__1538984560512460`
- `n015-2018-10-08-15-36-50+0800__CAM_FRONT__1538984563012460`
- `n008-2018-08-30-15-31-50-0400__CAM_FRONT__1535657697012404`
- `n008-2018-09-18-14-35-12-0400__CAM_FRONT__1537295844362404`
- `n008-2018-08-30-15-31-50-0400__CAM_FRONT__1535657694012404`

Metric details Previous driving scene reconstruction methods [8, 13, 38] typically utilize PSNR, LIPIS and SSIM [31] for quality evaluation. However, these metrics are inapplicable for evaluating free-form trajectories due to the absence of corresponding ground-truth images for novel trajectories. Latest research [40] proposed Novel Trajectory Agent IoU (NTA-IoU) and Novel Trajectory Lane IoU (NTL-IoU) for traffic components evaluation. Accordingly, we compared NTA-IoU and NTL-IoU, as shown in Table 2.

Furthermore, we propose improvements to the evaluation metrics of NTA-IoU and NTL-IoU. For vehicle quality, similar to NTA-IoU, we projected the original 3D vehicle bounding boxes onto the new viewpoints to generate corresponding 2D bounding boxes. IoU lacks the precision granularity needed to understand performance across various intersection over union threshold. Therefore, we calculate the Average Precision (AP[50:95]) which provides a more nuanced assessment of detection accuracy.

For lane quality, NTL-IoU calculates the mIoU between the rendered and the ground truth lanes, averaging both foreground lanes and backgrounds. However, including the background can inflate accuracy scores due to the easy de-

Noise level	IoU↑	AP ↑	FID↓	Time↓
0.4	0.1122	0.5904	98.31	1.0×
0.6	**0.2092**	**0.6257**	**90.30**	1.1×
0.8	0.0996	0.5874	92.26	1.2×
1.0	0.0773	0.5979	117.91	1.3×

Table 7. **Effect of the noise level.** In this table, we evaluate the performance under different refine strengths. The metrics are measured on views at the novel trajectory shifted $3m$ from the recorded trajectory.

tection of non-lane areas. Therefore, following official lane detection benchmarks [3] we focus specifically on IoU with only the foreground to better evaluate lane recognition performance.

For the evaluation of the capability as driving the scene generator, we mainly assess the realism of synthesized images across different trajectories by measuring their FID with the images at recorded trajectories.

More details For reconstruction from real data, we adopt regularization loss as introduced in [35], including sky loss, LiDAR depth loss and entropy loss for objects. In our framework, the LiDAR depth loss is used for both recorded trajectory and novel trajectory. In addition, we balance the recorded and generated data with 50% of the training data being generated. For reconstruction from generated videos, we additionally include an entropy loss for the opacity of each Gaussian to prevent potential overfitting.

Warping operation ϕ in Eq. (7) Given a camera extrinsic matrix $E \in \mathbb{R}^{4 \times 4}$ and intrinsic matrix $K \in \mathbb{R}^{3 \times 3}$, the mapping from world points $p_{\text{world}} \in \mathbb{R}^3$ to pixel coordinates (x, y) and depth d can be expressed by:

$$\begin{bmatrix} p_{\text{camera}} \\ 1 \end{bmatrix} = (*, *, d, 1)^\top = E \begin{bmatrix} p_{\text{world}} \\ 1 \end{bmatrix} \quad (11)$$

$$(x_*, y_*, d)^\top = K p_{\text{camera}} \quad (12)$$

$$(x, y) = (\frac{x_*}{d}, \frac{y_*}{d}) \quad (13)$$

Note that all the calculation above is invertible, allowing the mapping from (x, y, d) to world points p_{world}.

7. More ablations

Refine strength When we refine the rendered videos using video diffusion, we also study the effect the refine strength (level of noise added to the images). In Tab. 7, we show the metrics for novel trajectory synthesis (IoU, AP, and FID),

	IoU↑	AP↑	FID↓
w/o pose refinement	N/A	N/A	116.86
w/o depth prior	N/A	N/A	98.62
Full model	N/A	N/A	95.73
w/o progress.	0.1075	0.5959	91.22
w/o warp	0.0652	0.5962	119.24
Full model	0.2092	0.6257	90.30

Table 8. **Ablations on other design choices.** The metric in the first three lines is measured on the generated video, the others are tested on real scene.

	FID ↓		
	±0m	±1m	±2m
MagicDrive-t [6]	125.07	130.23	138.31
Vista [7]	97.48	99.19	148.07
Recon-only [35]	41.55	95.32	144.01
Ours	**36.54**	**71.47**	**119.11**

Table 9. **Comparison with the video-based driving world model.** The FID score is calculated between the recorded videos from the nuScenes dataset [2] and the novel trajectory videos synthesized by each method.

along with the total optimization time. With a low refine strength, the refined image adheres to the artifact-heavy rendered image, resulting in minimal improvement. On the other hand, a high refine strength yields better novel views but may not keep fidelity to the original scene. Additionally, higher refine strength requires more denoising steps, increasing the training time. Consequently, a refine strength of 0.6 is found to be the best balance.

Other designs We also ablate other design choices in Tab. 8, including the pose refinement and depth prior in the reconstruction from generative video, as well as progressive increasing shifted length and warped condition. The results show that the final quality degrades without any element. The results indicate that the exclusion of any element leads to a decline in the rendering quality and the stability of optimization.

8. Comparison with video generative model

In this section, we compare the novel trajectory video synthesis with representative open-sourced generative-based driving world models [6, 7]. Since these works are primarily trained on nuScenes dataset [2], we still conduct the experiments on 12 sequences from this dataset mentioned in Sec. 6. Different from the experiment in Tab. 3, in this experiment, our method takes ground truth recorded trajectory videos as input. The FID between the recorded videos and the synthesized novel trajectory videos from the differ-

ent methods are reported in Tab. 9.

Note that it is difficult to precisely control the trajectory for generative model. Though we adjust the action control, it still tend to generate videos in the original trajectory. Although all methods do not have significant artifacts, DriveX still demonstrates a clear advantage in FID metrics. This is because it can effectively avoid hallucinations that may be produced by generative models, making its result better aligned with the ground truth. Another advantage difficult to demonstrate by this metric is that DriveX can achieve precise camera pose control. In contrast, other methods are constrained by the diversity and imbalance of actions existing in public driving datasets, thus making it difficult to achieve accurate free-form trajectory synthesis.

9. More results

Fig. 6 shows more comparison results. We recommend readers to watch the video in the supplementary material.

Figure 6. **More comparison results**.