ReconDreamer: Crafting World Models for Driving Scene Reconstruction via Online Restoration

Chaojun Ni[*1,2] Guosheng Zhao[*1,4] Xiaofeng Wang[*1,4] Zheng Zhu[*1✉] Wenkang Qin[1]
Guan Huang[1] Chen Liu[3] Yuyin Chen[3] Yida Wang[3] Xueyang Zhang[3] Yifei Zhan[3]
Kun Zhan[3] Peng Jia[3] Xianpeng Lang[3] Xingang Wang[4] Wenjun Mei[2✉]
[1]GigaAI [2]Peking University [3] Li Auto Inc. [4]CASIA

Project Page: https://recondreamer.github.io

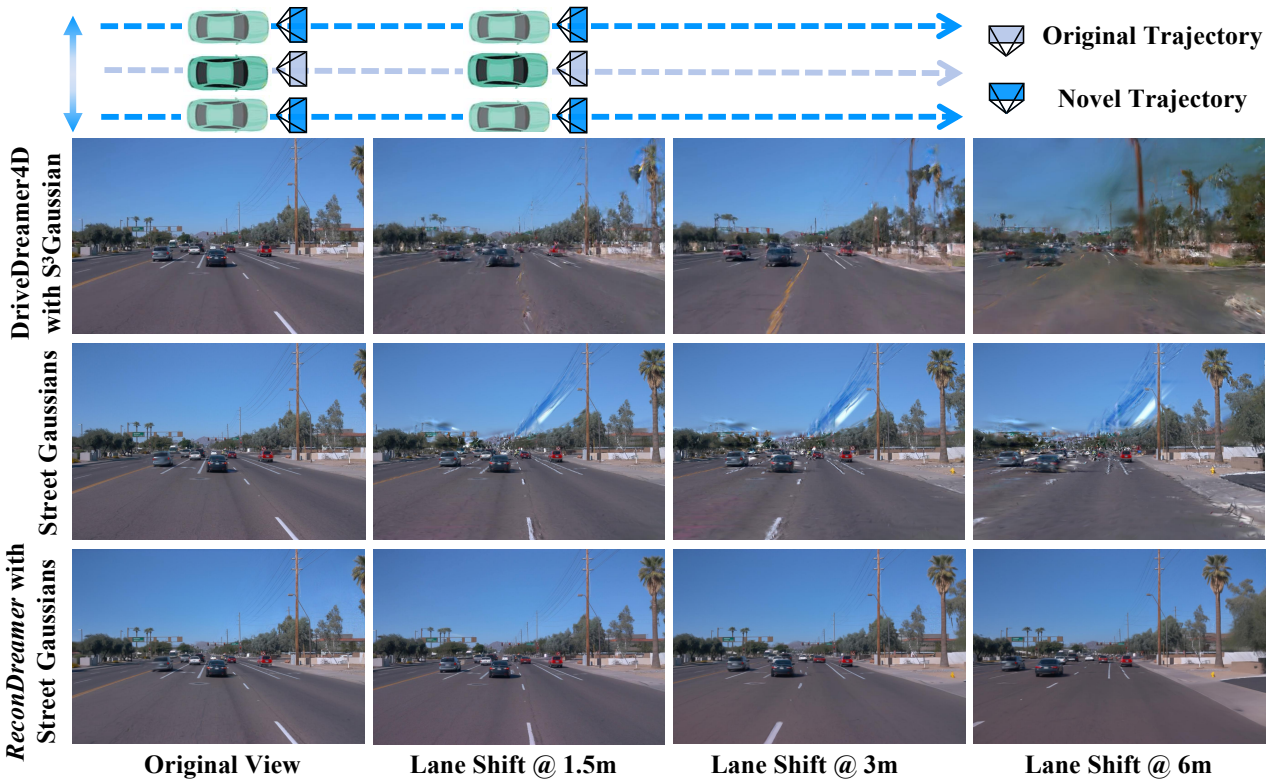

Figure 1. Dynamic driving scene reconstruction methods, such as DriveDreamer4D [68] and Street Gaussians [58], encounter significant challenges when rendering larger maneuvers (e.g., multi-lane shifts). In contrast, the proposed *ReconDreamer* significantly improves rendering quality via incrementally integrating world model knowledge.

Abstract

Closed-loop simulation is crucial for end-to-end autonomous driving. Existing sensor simulation methods (e.g., NeRF and 3DGS) reconstruct driving scenes based on conditions that closely mirror training data distributions. However, these methods struggle with rendering novel trajectories, such as lane changes. Recent works have demonstrated that integrating world model knowledge alleviates these issues. Despite their efficiency, these approaches still encounter difficulties in the accurate representation of more complex maneuvers, with multi-lane shifts being a notable example. Therefore, we introduce ReconDreamer, which enhances driving scene reconstruction through incremental integration of world model knowledge. Specifically, DriveRestorer is proposed to mitigate artifacts via online restoration. This is complemented by a progressive data update strategy designed to ensure high-quality ren-

*These authors contributed equally to this work. ✉Corresponding authors: Zheng Zhu, zhengzhu@ieee.org, Wenjun Mei, mei@pku.edu.cn.

dering for more complex maneuvers. To the best of our knowledge, ReconDreamer is the first method to effectively render in large maneuvers. Experimental results demonstrate that ReconDreamer outperforms Street Gaussians in the NTA-IoU, NTL-IoU, and FID, with relative improvements by 24.87%, 6.72%, and 29.97%. Furthermore, ReconDreamer surpasses DriveDreamer4D with PVG during large maneuver rendering, as verified by a relative improvement of 195.87% in the NTA-IoU metric and a comprehensive user study.

1. Introduction

End-to-end planning [22, 23, 26], which directly translates sensor data into control commands, is one of the most crucial tasks in autonomous driving. However, current open-loop evaluation methods fall short in providing an accurate assessment of end-to-end planning algorithms, underscoring the need for more robust evaluation frameworks [32, 34, 67]. A promising approach to address this issue involves using closed-loop evaluations conducted in real-world scenarios, which require retrieving sensor data from novel trajectory views. This demands the driving scene representation capable of reconstructing the intricate and dynamic nature of driving environments.

Closed-loop simulation predominantly hinges on scene reconstruction approaches like Neural Radiance Fields (NeRF) [15, 39, 59, 61] and 3D Gaussian Splatting (3DGS) [9, 24, 28, 58]. Despite their contributions, these techniques are fundamentally restricted by the density and diversity of training data, often limiting their rendering capabilities to scenarios that closely mimic the original training data. Consequently, they underperform in complex, high-variation driving maneuvers. Current developments in autonomous driving world models [13, 21, 49, 51, 52, 69] have introduced the capability to generate diverse videos aligned with specific driving commands, renewing the potential for more robust closed-loop simulation. The recent DriveDreamer4D [68] has further evidenced that leveraging pretrained world models as data machines can substantially improve the quality of dynamic driving scene reconstruction. However, while this training-free integration of world model knowledge is efficient, its current design still encounters challenges in executing larger maneuvers (e.g., multi-lane shifts).

In this paper, we introduce *ReconDreamer*, which enhances driving scene reconstruction via incrementally integrating knowledge from autonomous driving world models. Unlike [68] which leverages pretrained world models to directly expand novel trajectory views, *ReconDreamer* trains the world model to progressively mitigate ghosting artifacts in complex maneuver renderings. Specifically, we generate a video restoration dataset by sampling rendering outputs at various training stages. Based on the dataset, we propose the *DriveRestorer*, which is fine-tuned upon the world model to mitigate ghosting artifacts via online restoration. During the training, the masking strategy is introduced to emphasize restoration of challenging areas (e.g., sky and distant regions). Furthermore, we propose the Progressive Data Update Strategy (PDUS) to gradually restore artifacts, which ensures high-quality rendering for larger maneuvers. The proposed PDUS, by incrementally integrating world model knowledge, reduces the complexity of video restoration, making *ReconDreamer* the first approach capable of handling large viewpoint shifts in rendering (e.g., across multiple lanes, spanning up to 6 meters). As illustrated in Fig. 1, experimental results confirm that *ReconDreamer* substantially improves Street Gaussians [58] during novel trajectory rendering, achieving a relative improvement in the average NTA-IoU, NTL-IoU, and FID by 24.87%, 6.72%, and 29.97%. Additionally, *ReconDreamer* strengthens spatiotemporal coherence in larger maneuvers, outperforming DriveDreamer4D [68] with a win rate 96.88% in the user study, and a relative improvement of 195.87% in the NTA-IoU metric.

The primary contributions of this work are as follows: (1) We present *ReconDreamer*, which enhances dynamic driving scene reconstruction via incremental integration of world model knowledge. Notably, to our knowledge, *ReconDreamer* is the first method to effectively render in large maneuvers (e.g., spanning up to 6 meters). (2) The *DriveRestorer* is proposed to mitigate ghosting artifacts via online restoration. Besides, we introduce the progressive data update strategy to maintain high-quality rendering for larger maneuvers. (3) We perform comprehensive experiments to validate that *ReconDreamer* can enhance rendering quality during large maneuvers, as well as the spatiotemporal coherence of driving scene elements.

2. Related Work

2.1. Driving Scene Reconstruction Methods

NeRF and 3DGS have become prominent techniques in scene reconstruction. NeRF models [2, 3, 39, 40] use multilayer perceptron (MLP) networks to represent continuous volumetric scenes, achieving exceptional rendering quality. Recently, 3DGS [28, 64] introduced a novel approach by defining anisotropic Gaussians in 3D space and employing adaptive density control, which allows for high-quality renderings even from sparse point cloud data. Various studies have adapted NeRF [10, 15, 25, 37, 43, 47, 59, 61] and 3DGS [8, 9, 24, 35, 58, 65, 70] for driving scene reconstructions. To accommodate the dynamic nature of driving environments, some methods incorporate time as an additional parameter to capture temporal variations in dynamic scenes [1, 11, 24, 33, 36, 42, 45], while others treat the scene as

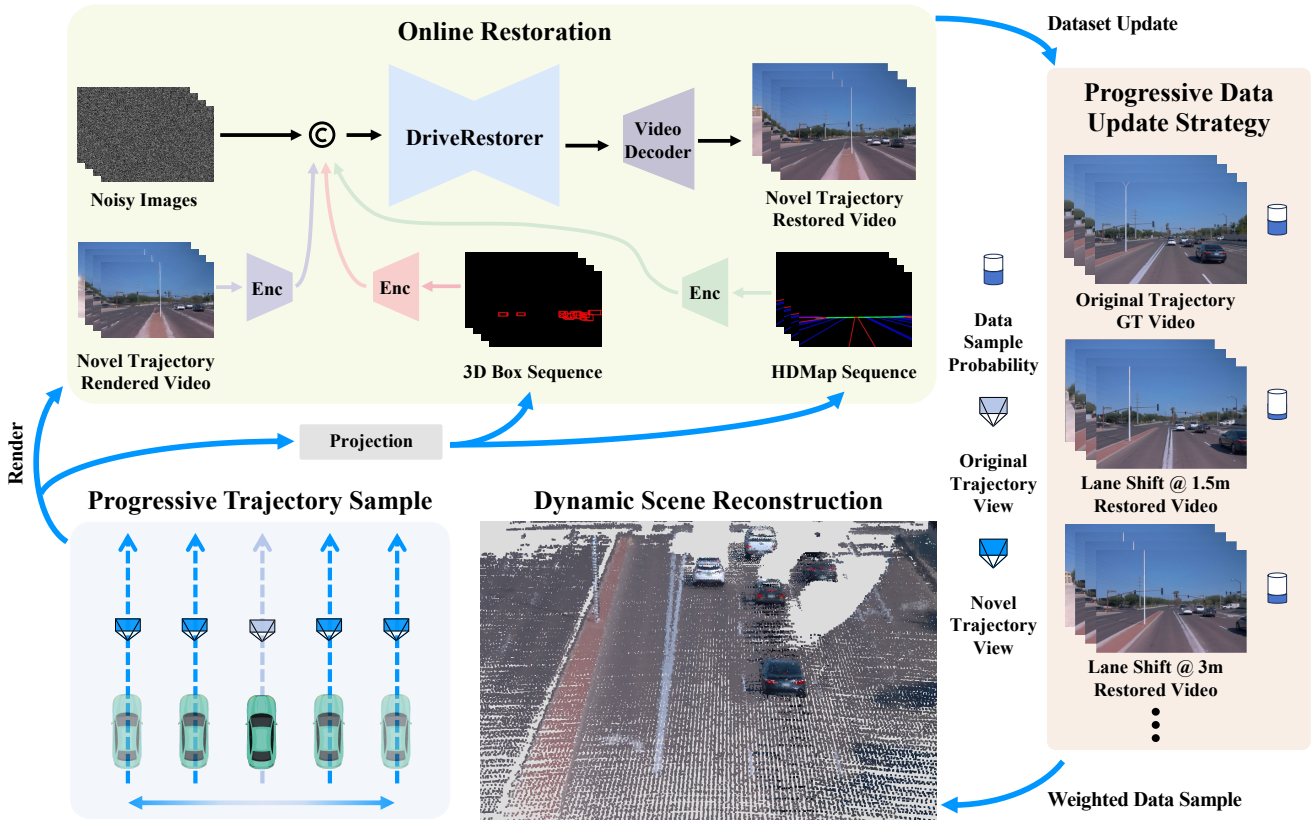

Figure 2. The overall framework of *ReconDreamer*. During the training of the dynamic driving scene reconstruction, we begin by rendering novel trajectory views. These rendered videos are subsequently processed by the *DriveRestorer* to restore their quality. Then these restored videos, together with the original video, are employed to optimize the reconstruction model. This iterative process continues until the reconstruction model converges (The training of *DriveRestorer* is omitted in this figure, and more details are in Sec. 3.2).

a combination of moving objects overlaid on a static background [30, 41, 48, 54, 56, 61]. Despite these advances, NeRF and 3DGS-based methods still encounter challenges related to data density. Their effectiveness in rendering relies heavily on sensor trajectory aligning closely with the training distribution. To tackle these challenges, methods like SGD [65], GGS [17], and MagicDrive3D [12] utilize generative models to expand the diversity of training perspectives. However, these methods are primarily focused on sparse image data or static background components, making them insufficient to fully capture the complexities of dynamic driving environments.

2.2. World Models

World models predict possible future world states as a function of imagined action sequences proposed by the actor [31, 71]. Based on world models, recent methods [4, 5, 14, 16, 18–20, 29, 38, 50, 51, 55, 57, 63, 66] have advanced the simulation of environments by generating videos that are guided by free-text actions. Leading this development is Sora [6], which employs cutting-edge generative methods to create complex visual sequences that adhere to the physical laws governing the environment. This capability not only enhances the fidelity of generated video content but also holds significant potential for applications in real-world driving scenarios. In autonomous driving, world models [13, 21, 49, 52, 60, 69] utilize predictive techniques to interpret driving environments. These methods generate realistic driving scenarios while extracting driving policies from video data, renewing the potential for more robust closed-loop simulations. The recent DriveDreamer4D [68] has further evidenced that leveraging pretrained world models as data machines improves dynamic driving scene reconstruction. Nonetheless, it still encounters challenges in executing larger maneuvers (e.g., multi-lane shifts).

3. Method

3.1. Overall Framework of ReconDreamer

Traditional scene reconstruction methods [8, 24, 28, 39, 53, 58, 62] face challenges due to the sparsity of training data. Recent approaches [17, 65, 68] alleviate this issue by leveraging generative priors to increase the data density. However, a gap remains between the generated data and the real data. In contrast, the proposed *ReconDreamer* expands the training data through an online restoration process. Notably,

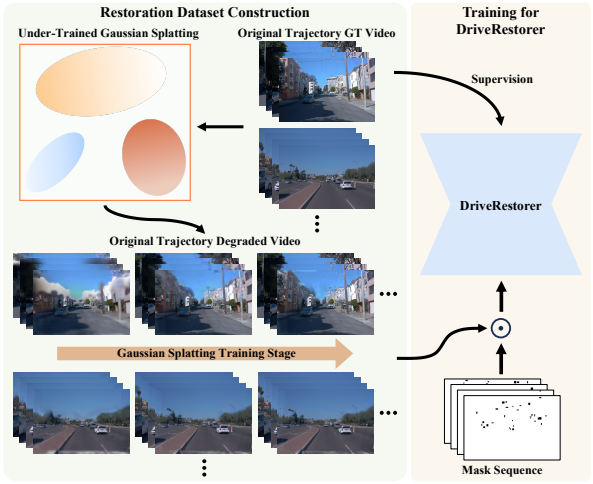

Figure 3. The restoration dataset construction for training *DriveRestorer*. Initially, we train an under-trained Gaussian Splatting model utilizing the ground truth (GT) videos from the original trajectory. During training of Gaussian Splatting model, degraded videos of the original trajectory are rendered at each stage. These degraded videos, paired with their corresponding GT videos, form the restoration dataset. A mask is then applied to the degraded videos to train DriveRestorer, supervised by the GT videos.

ReconDreamer progressively restores rendered data, effectively reducing the gap between generated and original data. The overall framework of *ReconDreamer* is illustrated in Fig. 2. Specifically, we first train the scene reconstruction method $\mathcal{G}$ using the original data V_{ori}. The trained methods then render novel trajectory videos $\hat{V}_{\text{novel}}$:

$$\hat{V}_{\text{novel}} = \mathcal{G}(\mathcal{T}_{\text{novel}}), \tag{1}$$

where $\mathcal{T}_{\text{novel}}$ is the novel trajectory. Notably, $\hat{V}_{\text{novel}}$ exhibits ghost artifacts due to data sparsity. Therefore, the *DriveRestorer* is introduced to restore the artifacts. The restoration resembles the diffusion denoising process [69], where structure conditions (3D boxes, and HDMaps) are employed to ensure spatiotemporal coherence of traffic elements. Note that the conditions are processed via projection transformation to align with $\mathcal{T}_{\text{novel}}$ (see Sec. 3.2 for more details). Consequently, the restored renderings V_{novel} have a smaller gap with the original video V_{ori}, making them more suitable as training samples for scene reconstruction. Additionally, to further enhance the training of $\mathcal{G}$ and enable it to render large maneuvers (e.g., multi-lane shifts), we propose the PDUS, which progressively updates the training dataset for scene reconstruction. Specifically, the novel trajectory is gradually expanded to generate large maneuver videos. These videos are then restored by *DriveRestorer* and used to update the training dataset (see Sec. 3.3 for more details). The updated dataset is then employed to optimize the reconstruction model. This iterative process continues until the reconstruction model converges.

Figure 4. Restoration data pairs for *DriveRestorer* training.

3.2. Training and Inference of DriveRestorer

Traditional scene reconstruction methods often suffer from artifacts when rendering novel trajectory views. To address this issue, we introduce *DriveRestorer* to restore these degraded renderings. In the next, we elaborate on training and inference details of the *DriveRestorer*.

Training. The major challenge of training *DriveRestorer* lies in the absence of rendering restoration datasets. Therefore, we propose a novel method for constructing restoration pairs. As illustrated in Fig. 3, we leverage under-trained reconstruction models [8, 24, 58, 62] to render videos $\hat{V}_{\text{ori}} = \mathcal{G}(\mathcal{T}_{\text{ori}})$ along the original trajectory, naturally producing ghosting artifacts due to model underfitting. These degraded frames are then paired with their corresponding ground truth video V_{ori}, forming a rendering restoration dataset. To further enhance dataset diversity, we sample rendered videos from different training stages. Consequently, the constructed rendering restoration pairs are represented as $\{\hat{V}^k_{\text{ori}}, V_{\text{ori}}\}$, where $\hat{V}^k_{\text{ori}}$ denotes the degraded video frame sampled at training stage k. Fig. 4 provides a visualization of these data pairs. Based on the constructed dataset, we train *DriveRestorer* to restore artifacts in the rendered videos. The *DriveRestorer* is fine-tuned upon the world model [69]. Specifically, we introduce degraded video frames $\hat{V}_{\text{ori}}$ as a control condition to provide appearance priors. To further emphasize the restoration of challenging regions, we apply masks to the degraded video frames $\hat{V}_{\text{ori}}$ during training. Since video quality degrades in distant areas (further from the camera center) and at the sky-scene boundary, our masks M focus primarily on these problematic regions (see the supplement for more details). During the training of *DriveRestorer*, we first feed the masked video frame $\hat{V}_{\text{mask}} = \hat{V}_{\text{ori}} \odot M$ into the encoder $\mathcal{E}$ to obtain the low-dimensional latent feature $z = \mathcal{E}(\hat{V}_{\text{mask}})$. The fine-tuning process of the world model is optimized us-

ing the diffusion loss:

$$\mathcal{L}_{\mathcal{R}} = \mathbb{E}_{\boldsymbol{z},\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon_t - \epsilon_\theta\left(\boldsymbol{z}_t, t, \boldsymbol{c}\right)\|_2^2\right], \quad (2)$$

where ϵ_t denotes the random noise at time step t, ϵ_θ is the parameterized denoising network, $\boldsymbol{z}_t$ refers to the noisy latent at time step t, and c represents the control conditions, including the degraded video $\hat{V}_{\text{mask}}$, 3D boxes, and HDMaps. The integration of video, 3D boxes, and HDMaps is similar to that of [69], with further details provided in the supplement.

Inference. After training *DriveRestorer*, we freeze its parameters to restore novel trajectory renderings:

$$V_{\text{novel}} = \mathcal{R}(\hat{V}_{\text{novel}}, \mathcal{P}(s, \mathcal{T}^k_{\text{novel}})), \quad (3)$$

where $\mathcal{R}$ is the *DriveRestorer*, s is the structure conditions (3D boxes and HDMaps), $\mathcal{P}(\cdot)$ denotes the projection transformation that aligns structure conditions with $\mathcal{T}^k_{\text{novel}}$. Note that $\mathcal{T}^k_{\text{novel}}$ is progressively expanded at different training stage k, which enhances driving scene reconstruction under large maneuver conditions (refer to Sec. 3.3 for more details). As shown in Fig. 5, the trained *DriveRestorer* can mitigate ghost artifacts in novel trajectory renderings.

Algorithm 1 Progressive Data Update Strategy

Input: Reconstruction model $\mathcal{G}$, *DriveRestorer* $\mathcal{R}$, Training step T, Update step S, Stride Δy, Original trajectory $\mathcal{T}_{\text{ori}}$, Novel trajectory video dataset D_{novel}
Output: Updated novel trajectory video dataset D_{novel}

```
k ← 0
for t in Range(0, T, S) do
    k ← k + 1
    w ← k / Σ_{j=1}^{k} j
    T^k_novel.y = T_ori.y + kΔy
    V̂_novel ← G(T^k_novel)
    V_novel ← R(V̂_novel)
    D_novel ← (1 − w) · D_novel ∪ w · V_novel
end for
```

3.3. Progressive Data Update Strategy

Based on *DriveRestorer*'s capability to restore novel trajectory videos, we propose the Progressive Data Update Strategy (PDUS) to enhance driving scene reconstruction under large maneuver conditions. The PDUS first constructs a mixed dataset $D = 0.5D_{\text{ori}} \cup 0.5D_{\text{novel}}$, where D_{ori} is the original trajectory video dataset and D_{novel} refers to the restored novel trajectory video dataset, which can be updated throughout the training process. The update strategy, detailed in Algo. 1, uses an update distance of $y = k\Delta y$ meters at k-th update step to progressively update the novel trajectory $\mathcal{T}_{\text{novel}}$. Then the reconstruction model $\mathcal{G}$ renders novel trajectory video $\hat{V}_{\text{novel}}$, which are then processed by

Figure 5. Examples of degraded video frame rendered under new trajectories and their restored frame by *DriveRestorer*.

DriveRestorer to obtain the restored novel trajectory video V_{novel}. To ensure that newly generated data provides additional priors for the reconstruction model, the updated dataset D_{novel} can be obtained as follows:

$$D_{\text{novel}} = (1 - w) \cdot D_{\text{novel}} \cup w \cdot V_{\text{novel}}, \quad (4)$$

where $w = \frac{k}{\sum_{j=1}^{k} j}$ is the sampling probability for V_{novel}.

For Gaussian Splatting-based driving scene reconstruction methods, we use original trajectory dataset D_{ori} to train the Gaussian parameters ϕ:

$$\mathcal{L}_{\text{ori}}(\phi) = \lambda_1\mathcal{L}^{\text{RGB}}_{\text{ori}} + \lambda_2\mathcal{L}^{\text{Depth}}_{\text{ori}} + \lambda_3\mathcal{L}^{\text{SSIM}}_{\text{ori}}, \quad (5)$$

where $\mathcal{L}^{\text{RGB}}_{\text{ori}}, \mathcal{L}^{\text{Depth}}_{\text{ori}}, \mathcal{L}^{\text{SSIM}}_{\text{ori}}$ are reconstruction losses typically used in the Gaussian Splatting optimization [58], and $\lambda_1, \lambda_2, \lambda_3$ are loss weights. Additionally, for the generated novel trajectory data D_{novel}, we adopt the depth-free training strategy from [68] to train *ReconDreamer*:

$$\mathcal{L}_{\text{novel}}(\phi) = \lambda_1\mathcal{L}^{\text{RGB}}_{\text{novel}} + \lambda_3\mathcal{L}^{\text{SSIM}}_{\text{novel}}. \quad (6)$$

The overall loss function for mixed training is:

$$\mathcal{L}(\phi) = \mathcal{L}_{\text{ori}} + \mathcal{L}_{\text{novel}}. \quad (7)$$

4. Experiments

In this section, we present our experimental setup, which encompasses the datasets, implementation details, and evaluation metrics. Subsequently, both quantitative and qualitative results are provided to demonstrate that the proposed *ReconDreamer* can effectively render large maneuvers while also enhancing spatiotemporal coherence. Finally, we perform experiments to explore the different stride settings in progressive data update strategy and evaluate different *DriveRestorer* backbone choices.

Method	Lane Shift @ 3m			Lane Shift @ 6m			Lane Change			Average		
	NTA-IoU ↑	NTL-IoU ↑	FID↓	NTA-IoU ↑	NTL-IoU ↑	FID↓	NTA-IoU ↑	NTL-IoU ↑	FID↓	NTA-IoU ↑	NTL-IoU ↑	FID↓
PVG [8]	0.242	49.99	134.74	0.118	48.32	212.19	0.256	50.70	105.29	0.205	49.67	150.74
S^3Gaussian [24]	0.059	48.24	234.54	0.014	47.69	311.59	0.175	49.05	124.90	0.083	48.33	223.68
Deformable-GS [62]	0.142	50.63	161.20	0.069	48.86	260.83	0.240	51.62	92.34	0.150	50.37	171.46
Street Gaussians [58]	0.498	53.19	130.75	0.374	49.27	213.04	0.496	56.03	86.46	0.456	52.83	143.42
ReconDreamer with Street Gaussians	**0.539**	**54.58**	**93.56**	**0.467**	**52.58**	**149.19**	**0.554**	**56.63**	**73.91**	**0.517**	**54.60**	**105.55**

Table 1. Comparison of NTA-IoU, NTL-IoU, and FID scores for different novel trajectory views with various methods [8, 24, 58, 62].

Method	Lane Shift @ 3m			Lane Shift @ 6m			Lane Change			Average		
	NTA-IoU ↑	NTL-IoU ↑	FID↓	NTA-IoU ↑	NTL-IoU ↑	FID↓	NTA-IoU ↑	NTL-IoU ↑	FID↓	NTA-IoU ↑	NTL-IoU ↑	FID↓
PVG [8]	0.242	49.99	134.74	0.118	48.32	212.19	0.256	50.70	105.29	0.205	49.67	150.74
DriveDreamer4D with PVG [68]	0.340	51.32	129.05	0.121	49.28	210.37	0.438	53.06	**71.52**	0.300	51.22	136.98
ReconDreamer with PVG	**0.474**	**52.06**	**124.12**	**0.358**	**50.14**	**177.58**	**0.464**	**53.21**	74.32	**0.432**	**51.80**	**125.34**
S^3Gaussian[24]	0.059	48.24	234.54	0.014	47.69	311.59	0.175	49.05	124.90	0.083	48.33	223.68
DriveDreamer4D with S^3Gaussian [68]	0.263	**51.07**	173.49	0.031	48.04	303.95	**0.495**	**53.42**	**66.93**	0.263	**50.84**	181.46
ReconDreamer with S^3Gaussian	**0.267**	50.26	**110.41**	**0.177**	**50.54**	**152.89**	0.413	51.62	123.61	**0.286**	50.80	**128.97**
Deformable-GS [62]	0.142	50.63	161.20	0.069	48.86	260.83	0.240	51.62	92.34	0.150	50.37	171.46
DriveDreamer4D with Deformable-GS [68]	0.181	50.72	169.79	0.078	48.90	259.19	0.335	52.93	77.32	0.198	50.85	168.77
ReconDreamer with Deformable-GS	**0.416**	**52.26**	**106.99**	**0.294**	**50.54**	**143.88**	**0.443**	**53.78**	**76.24**	**0.384**	**52.19**	**109.04**

Table 2. Comparison of NTA-IoU, NTL-IoU, and FID scores for different novel trajectory views with DriveDreamer4D [68].

	ReconDreamer Win Rate	
	Street Gaussians[58]	DriveDreamer4D with PVG[68]
Lane Shift @ 3m	97.92%	96.88%
Lane Shift @ 6m	100.00%	100.00%
Lane Change	93.75%	93.75%
Average	**97.22%**	**96.88%**

Table 3. Comparing the win rates of *ReconDreamer* in rendering large maneuvers.

4.1. Experiment Setup

Dataset. We conduct experiments in eight highly interactive scenes from the Waymo dataset [46]. These scenes are characterized by numerous vehicles in various positions, following complex driving trajectories, with multiple lanes that increase the complexity of foreground and background reconstruction (see supplementary materials for specific scene IDs and frame numbers).

Implementation Details. To showcase the capability of *ReconDreamer* in rendering large maneuvers, we conduct comprehensive comparisons against several state-of-the-art methods for dynamic driving scene reconstruction. These methods include Deformable-GS [62], S^3Gaussian [24], PVG [8], Street Gaussians [58], and DreamerDriver4D [68]. For the training phase, we divide the scenes into multiple segments, each containing 40 frames, and exclusively use data from the front-facing camera. Furthermore, we set up the training strategies and hyperparameters for each baseline method to match their original configurations, ensuring consistent training for 50,000 iterations. After the training, we assess the model's performance across three distinct novel trajectories: 3 meters lateral shift from the original trajectory, 6 meters lateral shift from the original trajectory, and lane change involving a lateral move to a parallel lane.

Metrics. Following DriveDreamer4D [68], we use Novel Trajectory Agent IoU (NTA-IoU), Novel Trajectory Lane IoU (NTL-IoU), FID as evaluation metrics. Additionally, we conduct a user study to evaluate the quality of the rendered video. More details are in supplementary materials.

4.2. Main Results

Comparison with Scene Recosntruction Baselines. In Tab. 1, we compare *ReconDreamer* with different dynamic driving scene reconstruction methods [8, 24, 58, 62]. The experimental results demonstrate that the current state-of-the-art dynamic driving scene reconstruction method, Street Gaussians [58], outperforms other traditional approaches (*i.e.*, PVG [8], S^3Gaussian and Deformable-GS [62]) across various novel trajectory renderings. Therefore, we investigate how much *ReconDreamer* can improve performance based on [58]. Specifically, *ReconDreamer* with Street Gaussians outperforms Street Gaussians across all metrics for different novel trajectory renderings. The average scores (NTA-IoU, NTL-IoU, FID) are relatively improved by 13.38%, 3.35%, and 26.40%, respectively. Notably, *ReconDreamer* with Street Gaussians effectively renders large maneuvers (e.g., Lane Shift @ 6m), relatively surpassing Street Gaussians by 24.87%, 6.72%, and 29.97% on the NTA-IoU, NTL-IoU, and FID metrics, respectively.

Comparison with DriveDreamer4D. We compare the pro-

Preatrained Model	Lane Shift @ 3m			Lane Shift @ 6m			Lane Change			Average		
	NTA-IoU ↑	NTL-IoU ↑	FID↓	NTA-IoU ↑	NTL-IoU ↑	FID↓	NTA-IoU ↑	NTL-IoU ↑	FID↓	NTA-IoU ↑	NTL-IoU ↑	FID↓
Stable Diffusion [44]	0.504	54.51	112.25	0.389	52.13	178.16	0.509	56.42	80.90	0.467	54.35	123.77
Stable Video Diffusion[4]	0.500	53.33	120.46	0.435	49.67	199.40	0.527	56.05	89.89	0.487	53.02	136.58
DriveDreamer-2[69]	0.525	54.42	100.32	0.456	52.17	152.23	**0.551**	**56.67**	92.30	0.511	54.42	114.95
DriveDreamer-2+Mask	**0.539**	**54.52**	**93.56**	**0.467**	**52.56**	**149.19**	0.544	**56.67**	**73.91**	**0.517**	**54.58**	**105.55**

Table 4. Comparison of NTA-IoU, NTL-IoU, and FID scores for *DriveRestorer* with different backbones.

Stride	Lane Shift @ 3m			Lane Shift @ 6m			Lane Change			Average		
	NTA-IoU ↑	NTL-IoU ↑	FID↓	NTA-IoU ↑	NTL-IoU ↑	FID↓	NTA-IoU ↑	NTL-IoU ↑	FID↓	NTA-IoU ↑	NTL-IoU ↑	FID↓
0.5m	0.533	54.10	120.23	0.452	51.82	162.34	0.521	56.78	77.34	0.502	54.23	119.97
1.0m	0.536	54.13	114.23	0.459	51.93	162.46	0.523	56.89	79.27	0.506	54.32	118.65
1.5m	**0.539**	**54.52**	**93.56**	**0.467**	**52.56**	**149.19**	**0.554**	56.67	73.91	**0.517**	**54.58**	**105.55**
3.0m	0.521	54.26	97.54	0.451	52.16	150.43	0.523	**57.10**	**72.09**	0.498	54.51	106.69
6.0m	0.518	54.55	104.72	0.447	52.24	168.22	0.513	56.40	84.34	0.493	54.40	119.09

Table 5. Comparison of NTA-IoU, NTL-IoU, and FID scores for different stride settings in progressive data update strategy. Note that no progressive data update strategy is adopted when stride is set to 6m.

posed *ReconDreamer* with DriveDreamer4D[68], both of which leverage world model priors to enhance driving scene reconstruction. Specifically, we implement both approaches on PVG[8], Deformable-GS[62], and S^3Gaussian[24]. The experimental results shown in Table 2 indicate that these three original methods face challenges when rendering foreground vehicles and lane lines across new trajectories. While DriveDreamer4D improves over baseline algorithms in all metrics, its performance is limited in large maneuvers (6m lane shift), with NTA-IoU values of just 0.121, 0.031, and 0.078. In contrast, *ReconDreamer*, which uses world model priors via an online restoration process instead of a training-free approach, shows superior performance compared to DriveDreamer4D. When compared to the versions of DriveDreamer4D implemented with PVG[8], S^3Gaussian[24], and Deformable-GS[62], *ReconDreamer* significantly boosts the average NTA-IoU scores by 44.00%, 8.75%, and 93.94%, respectively. Furthermore, it achieves relative improvements in FID of 8.50%, 28.93%, and 35.39%, respectively. Notably, in the large maneuver rendering (6m lane shift), ReconDreamer outperforms DriveDreamer4D, with relative improvements in NTA-IoU reaching 195.87%, 470.97%, and 276.92%, respectively.

User Study. Additionally, a user study is conducted to evaluate the rendering quality of various methods on novel trajectories, with a particular focus on the presence of ghost artifacts in the rendered video. We select DriveDreamer4D with PVG [68] and Street Gaussians [58] as the comparison methods. As shown in Tab. 3, the results indicate that the *ReconDreamer* approach significantly outperformed the above two methods in terms of user preference, which have an average of 97.22% and 96.88% win rates.

Qualitative Results. As illustrated in Fig. 6, we present view renderings under new trajectories involving large viewpoint shifts, alongside their corresponding ground truth video from the original trajectories. The renderings from the baseline algorithms show significant ghosting and speckling in the background, with jagged lane markings and distorted or missing background elements such as trees. As the camera moves, numerous irregular white artifacts become visible, particularly in Street Gaussians [58]. Furthermore, foreground vehicles are severely distorted and affected by speckles. However, our method significantly improves the rendering quality. In the background, *ReconDreamer* maintains excellent spatiotemporal consistency between the novel trajectory and the original trajectory. Additionally, the foreground vehicles remain clear and accurately reflect the viewpoint changes.

4.3. Ablation Study

***DriveRestorer* Backbone.** As shown in Tab. 4, we compare the performance of *DriveRestorer* using different backbones and fine-tuning strategies, including Stable Diffusion [44], Stable Video Diffusion [4], DriveDreamer-2 [69], and DriveDreamer-2 with mask. All these methods utilize Street Gaussians [58] for reconstruction. The *DriveRestorer* based on the Stable Diffusion [44] demonstrates strong performance, achieving relative improvements over the baseline by 2.41%, 2.88%, and 13.71% in NTA-IoU, NTL-IoU, and FID metrics, respectively. However, image-based restoration models perform moderately on metrics such as NTA-IoU, primarily due to their lack of spatiotemporal consistency, which leads to positional deviation of vehicles after restoration. Meanwhile, the video-based method Stable Video Diffusion [4] provides better spatial continuity but faces challenges with high fine-tuning difficulty and lack of controllability, leading to color discrepancies and difficulties in restoring details such as lane lines. DriveDreamer-2 [69], built upon Stable Video Diffusion [4], introduces additional control conditions, such as 3D boxes and HDMaps, significantly enhancing the NTA-IoU and NTL-IoU. Compared to Stable Diffusion [44], DriveDreamer-2 improves the NTA-IoU by 9.42% and the NTL-IoU by 0.13%. Compared to Stable Video Diffusion [4], the improvements are

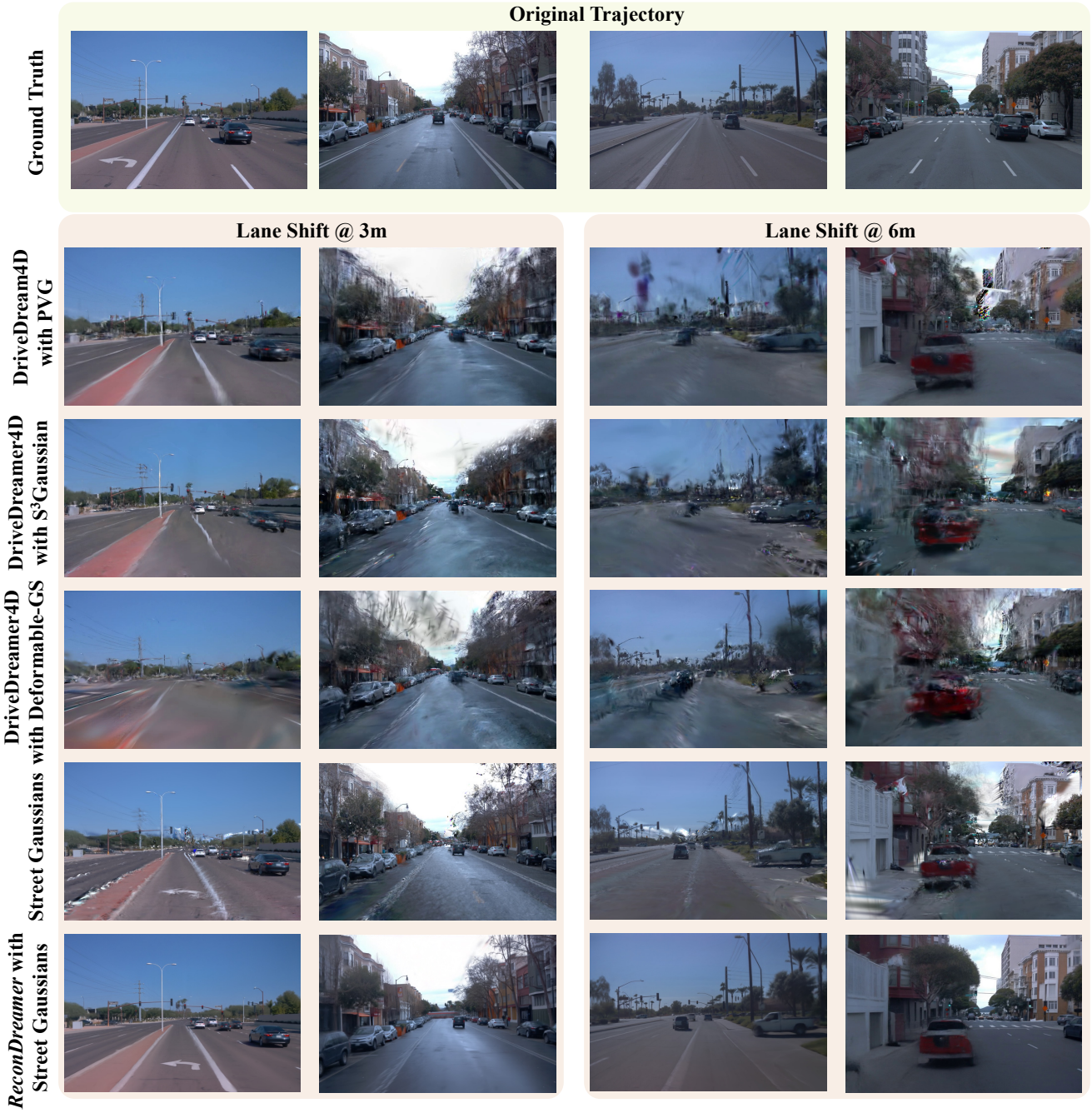

Figure 6. Qualitative comparisons of novel trajectory renderings for lane shift @ 3m and lane shift @ 6m. The yellow box contains the ground truth video frames of the original trajectories, while the pink boxes display the rendered video frames after the lane shift.

4.93% for NTA-IoU and 2.64% for NTL-IoU. Additionally, the FID score improves by 7.13% and 15.84% compared to the respective baselines. Finally, incorporating masks during the fine-tuning process of DriveDreamer-2 [69] further boosts the overall performance of the model. The supplement provides additional visual comparisons.

Progressive Data Update Strategy. To analyze the influence of different stride settings ($\Delta y = 0.5$, 1, 1.5, 3, and 6) in the PDUS, we conduct experiments on *ReconDreamer* with Street Gaussians. As shown in Tab. 5, the model achieves optimal performance when the stride is set to 1.5. Smaller stride values (0.5 and 1.0) lead to a 2.90% and 2.13% decrease in NTA-IoU, and a 0.64% and 0.48% reduction in NTL-IoU, respectively. Additionally, the FID score shows a significant drop of 13.66% and 12.41% respectively. Larger stride values cause excessive noise and ghosting during the

rendering of novel trajectories, exceeding the restoration capacity of *DriveRestorer*. Notably, a stride of 6 is equivalent to disabling the PDUS, resulting in the worst average NTA-IoU and NTL-IoU, which confirms the effectiveness of the proposed PDUS.

5. Discussion and Conclusion

Closed-loop simulation is crucial for autonomous driving, requiring accurate scene reconstruction for realistic simulations. Methods like NeRF and 3DGS struggle with novel trajectory renderings. DriveDreamer4D alleviates the challenge by using a pre-trained world model but still struggles with larger maneuvers. To overcome these limitations, we propose *ReconDreamer*, which crafts world models for driving scene reconstruction via online restoration. With the introduction of *DriveRestorer* to reduce ghosting artifacts and a progressive data update strategy, *ReconDreamer* is the first method capable of rendering large maneuvers, spanning up to 6 meters. Experiments show *ReconDreamer* surpasses Street Gaussians in NTA-IoU, NTL-IoU, and FID with improvements of 24.87%, 6.72%, and 29.97%, respectively. It also outperforms DriveDreamer4D with PVG in large maneuver rendering, with a comprehensive user study and 195.87% improvement in NTA-IoU.

References

[1] Benjamin Attal, Jia-Bin Huang, Christian Richardt, Michael Zollhoefer, Johannes Kopf, Matthew O'Toole, and Changil Kim. Hyperreel: High-fidelity 6-dof video with ray-conditioned sampling. In *CVPR*, 2023. 2

[2] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, 2022. 2

[3] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *ICCV*, 2023. 2

[4] 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 arXiv:2311.15127*, 2023. 3, 7, 15

[5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023. 3

[6] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 3

[7] 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. 12

[8] Yurui Chen, Chun Gu, Junzhe Jiang, Xiatian Zhu, and Li Zhang. Periodic vibration gaussian: Dynamic urban scene reconstruction and real-time rendering. *arXiv preprint arXiv:2311.18561*, 2023. 2, 3, 4, 6, 7, 13, 14

[9] Ziyu Chen, Jiawei Yang, Jiahui Huang, Riccardo de Lutio, Janick Martinez Esturo, Boris Ivanovic, Or Litany, Zan Gojcic, Sanja Fidler, Marco Pavone, Li Song, and Yue Wang. Omnire: Omni urban scene reconstruction. *arXiv preprint arXiv:2408.16760*, 2024. 2

[10] Kai Cheng, Xiaoxiao Long, Wei Yin, Jin Wang, Zhiqiang Wu, Yuexin Ma, Kaixuan Wang, Xiaozhi Chen, and Xuejin Chen. Uc-nerf: Neural radiance field for under-calibrated multi-view cameras in autonomous driving. *arXiv preprint arXiv:2311.16945*, 2023. 2

[11] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *CVPR*, 2023. 2

[12] Ruiyuan Gao, Kai Chen, Zhihao Li, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrive3d: Controllable 3d generation for any-view rendering in street scenes. *arXiv preprint arXiv:2405.14475*, 2024. 3

[13] 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. *arXiv preprint arXiv:2405.17398*, 2024. 2, 3

[14] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023. 3

[15] Jianfei Guo, Nianchen Deng, Xinyang Li, Yeqi Bai, Botian Shi, Chiyu Wang, Chenjing Ding, Dongliang Wang, and Yikang Li. Streetsurf: Extending multi-view implicit surface reconstruction to street views. *arXiv preprint arXiv:2306.04988*, 2023. 2

[16] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Li Fei-Fei, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. *arXiv preprint arXiv:2312.06662*, 2023. 3

[17] Huasong Han, Kaixuan Zhou, Xiaoxiao Long, Yusen Wang, and Chunxia Xiao. Ggs: Generalizable gaussian splatting for lane switching in autonomous driving. *arXiv preprint arXiv:2409.02382*, 2024. 3

[18] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 3

[19] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *NeurIPS*, 2022.

[20] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 3

[21] 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, 3

[22] Shengchao Hu, Li Chen, Penghao Wu, Hongyang Li, Junchi Yan, and Dacheng Tao. St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In *ECCV*, 2022. 2

[23] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *CVPR*, 2023. 2

[24] Nan Huang, Xiaobao Wei, Wenzhao Zheng, Pengju An, Ming Lu, Wei Zhan, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. s^3gaussian: Self-supervised street gaussians for autonomous driving. *arXiv preprint arXiv:2405.20323*, 2024. 2, 3, 4, 6, 7, 14

[25] Muhammad Zubair Irshad, Sergey Zakharov, Katherine Liu, Vitor Guizilini, Thomas Kollar, Adrien Gaidon, Zsolt Kira, and Rares Ambrus. Neo 360: Neural fields for sparse view synthesis of outdoor scenes. In *ICCV*, 2023. 2

[26] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *ICCV*, 2023. 2

[27] Glenn Jocher and Jing Qiu. Ultralytics yolo11, 2024. 12

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

[29] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vighnesh Birodkar, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023. 3

[30] Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas J Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas Funkhouser. Panoptic neural fields: A semantic object-aware neural scene representation. In *CVPR*, pages 12871–12881, 2022. 3

[31] Yann LeCun and Courant. A path towards autonomous machine intelligence version 0.9.2, 2022-06-27. 2022. 3

[32] Hao Li, Ming Yuan, Yan Zhang, Chenming Wu, Chen Zhao, Chunyu Song, Haocheng Feng, Errui Ding, Dingwen Zhang, and Jingdong Wang. Xld: A cross-lane dataset for benchmarking novel driving view synthesis. *arXiv preprint arXiv:2406.18360*, 2024. 2

[33] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *CVPR*, 2021. 2

[34] Zhiqi Li, Zhiding Yu, Shiyi Lan, Jiahan Li, Jan Kautz, Tong Lu, and Jose M Alvarez. Is ego status all you need for open-loop end-to-end autonomous driving? In *CVPR*, 2024. 2

[35] Zhuopeng Li, Yilin Zhang, Chenming Wu, Jianke Zhu, and Liangjun Zhang. Ho-gaussian: Hybrid optimization of 3d gaussian splatting for urban scenes. *arXiv preprint arXiv:2403.20032*, 2024. 2

[36] Haotong Lin, Sida Peng, Zhen Xu, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Efficient neural radiance fields for interactive free-viewpoint video. In *SIGGRAPH Asia*, 2022. 2

[37] Fan Lu, Yan Xu, Guang Chen, Hongsheng Li, Kwan-Yee Lin, and Changjun Jiang. Urban radiance field representation with deformable neural mesh primitives. In *ICCV*, 2023. 2

[38] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024. 3

[39] 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. *Communications of the ACM*, 2021. 2, 3

[40] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM ToG*, 2022. 2

[41] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In *CVPR*, pages 2856–2865, 2021. 3

[42] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021. 2

[43] Konstantinos Rematas, Andrew Liu, Pratul P Srinivasan, Jonathan T Barron, Andrea Tagliasacchi, Thomas Funkhouser, and Vittorio Ferrari. Urban radiance fields. In *CVPR*, 2022. 2

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 7, 14

[45] Liangchen Song, Anpei Chen, Zhong Li, Zhang Chen, Lele Chen, Junsong Yuan, Yi Xu, and Andreas Geiger. Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 2023. 2

[46] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 6, 12, 14

[47] 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

[48] Adam Tonderski, Carl Lindström, Georg Hess, William Ljungbergh, Lennart Svensson, and Christoffer Petersson. Neurad: Neural rendering for autonomous driving. In *CVPR*, 2024. 3

[49] 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 arXiv:2309.09777*, 2023. 2, 3

[50] Xiaofeng Wang, Kang Zhao, Feng Liu, Jiayu Wang, Guosheng Zhao, Xiaoyi Bao, Zheng Zhu, Yingya Zhang, and Xingang Wang. Egovid-5m: A large-scale video-action dataset for egocentric video generation. *arXiv preprint arXiv:2411.08380*, 2024. 3

[51] Xiaofeng Wang, Zheng Zhu, Guan Huang, Boyuan Wang, Xinze Chen, and Jiwen Lu. Worlddreamer: Towards general world models for video generation via predicting masked tokens. *arXiv preprint arXiv:2401.09985*, 2024. 2, 3

[52] 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, 3

[53] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *CVPR*, 2024. 3

[54] Zirui Wu, Tianyu Liu, Liyi Luo, Zhide Zhong, Jianteng Chen, Hongmin Xiao, Chao Hou, Haozhe Lou, Yuantao Chen, Runyi Yang, et al. Mars: An instance-aware, modular and realistic simulator for autonomous driving. In *ICAI*, 2023. 3

[55] Jiannan Xiang, Guangyi Liu, Yi Gu, Qiyue Gao, Yuting Ning, Yuheng Zha, Zeyu Feng, Tianhua Tao, Shibo Hao, Yemin Shi, et al. Pandora: Towards general world model with natural language actions and video states. *arXiv preprint arXiv:2406.09455*, 2024. 3

[56] Ziyang Xie, Junge Zhang, Wenye Li, Feihu Zhang, and Li Zhang. S-nerf: Neural radiance fields for street views. *arXiv preprint arXiv:2303.00749*, 2023. 3

[57] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 3

[58] Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng. Street gaussians for modeling dynamic urban scenes. *arXiv preprint arXiv:2401.01339*, 2024. 1, 2, 3, 4, 5, 6, 7, 12, 14

[59] 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. *arXiv preprint arXiv:2311.02077*, 2023. 2

[60] Xuemeng Yang, Licheng Wen, Yukai Ma, Jianbiao Mei, Xin Li, Tiantian Wei, Wenjie Lei, Daocheng Fu, Pinlong Cai, Min Dou, Botian Shi, Liang He, Yong Liu, and Yu Qiao. Drivearena: A closed-loop generative simulation platform for autonomous driving. *arXiv preprint arXiv:2408.00415*, 2024. 3

[61] Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. Unisim: A neural closed-loop sensor simulator. In *CVPR*, 2023. 2, 3

[62] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *CVPR*, 2024. 3, 4, 6, 7, 13, 14

[63] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 3

[64] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *CVPR*, 2024. 2

[65] 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. *arXiv preprint arXiv:2403.20079*, 2024. 2, 3

[66] Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. Make pixels dance: High-dynamic video generation. *arXiv preprint arXiv:2311.10982*, 2023. 3

[67] Jiang-Tian Zhai, Ze Feng, Jinhao Du, Yongqiang Mao, Jiang-Jiang Liu, Zichang Tan, Yifu Zhang, Xiaoqing Ye, and Jingdong Wang. Rethinking the open-loop evaluation of end-to-end autonomous driving in nuscenes. *arXiv preprint arXiv:2305.10430*, 2023. 2

[68] 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 arXiv:2410.13571*, 2024. 1, 2, 3, 5, 6, 7, 12, 14

[69] 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 arXiv:2403.06845*, 2024. 2, 3, 4, 5, 7, 8, 15

[70] 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*, pages 21634–21643, 2024. 2

[71] Zheng Zhu, Xiaofeng Wang, Wangbo Zhao, Chen Min, Nianchen Deng, Min Dou, Yuqi Wang, Botian Shi, Kai Wang, Chi Zhang, et al. Is sora a world simulator? a comprehensive survey on general world models and beyond. *arXiv preprint arXiv:2405.03520*, 2024. 3

Figure 7. Examples of degraded video frames, the corresponding masked video frames, and GT video frames under the original trajectories.

In the supplementary material, we provide detailed information on the training for *DriveRestorer* , the selected scenes, the metrics, and the user study. Additionally, we present more qualitative results to compare the restoration effects achieved by *DriveRestorer* with different backbones and to evaluate the impact of varying stride settings in the PDUS.

6. Implementation Details

Training for *DriveRestorer*. As shown in Fig. 7,the frames rendered by the reconstruction model often exhibit significant degradation at the boundary between the sky and the background and the areas far from the camera in the image center. To address these issues, we introduce a masking strategy, applying random masks to these degraded regions to guide the model in repairing them.

Metrics. As mentioned in the main text, we utilize Novel Trajectory Agent Intersection over Union (NTA-IoU) and Novel Trajectory Lane Intersection over Union (NTL-IoU) to assess the quality of the rendered video, both metrics proposed in DriveDreamer4D [68]. These metrics are specifically designed to evaluate the spatiotemporal coherence of foreground agents and background lanes, respectively.

The NTA-IoU processes images rendered under new trajectories using the YOLO11 [27] detector to extract 2D bounding boxes of vehicles. Meanwhile, by applying geometric transformations to the 3D bounding boxes from the original trajectories, they can be accurately projected onto the new trajectory perspective, thus obtaining the ground truth 2D bounding boxes in the new trajectory view. Each projected 2D bounding box will find the nearest 2D bounding box generated by the detector and compute their Intersection over Union (IoU).

Similarly, the NTL-IoU employs the TwinLiteNet [7] model to detect lane in the images rendered under the new trajectories, and the lane from the original trajectories will also be projected onto the new trajectory through corresponding geometric transformations. Finally, the mean Intersection over Union (IoU) between the projected and detected lane lines is calculated.

Scene Selection. We select eight scenes from the validation set of the Waymo dataset [46]. These scenes feature high levels of interactive activity, with numerous vehicles in varied positions and exhibiting complex driving trajectories. Additionally, these scenes include multiple lanes, which increases the complexity of foreground and background reconstruction. As shown in Table. 6, we provide a detailed list of the segment IDs.

User Study. In the user study, we compare our results with two baseline models: DriveDreamer4D with PVG [68] and Street Gaussians [58]. This comparison is conducted across the eight scenarios we selected, with an emphasis on the overall quality of the videos, including the consistency and clarity of the background, as well as the positional accuracy of foreground objects. In each comparison, our method and the baseline methods are randomly assigned to the top or bottom of the video, and participants are asked to choose the option they find most satisfactory.

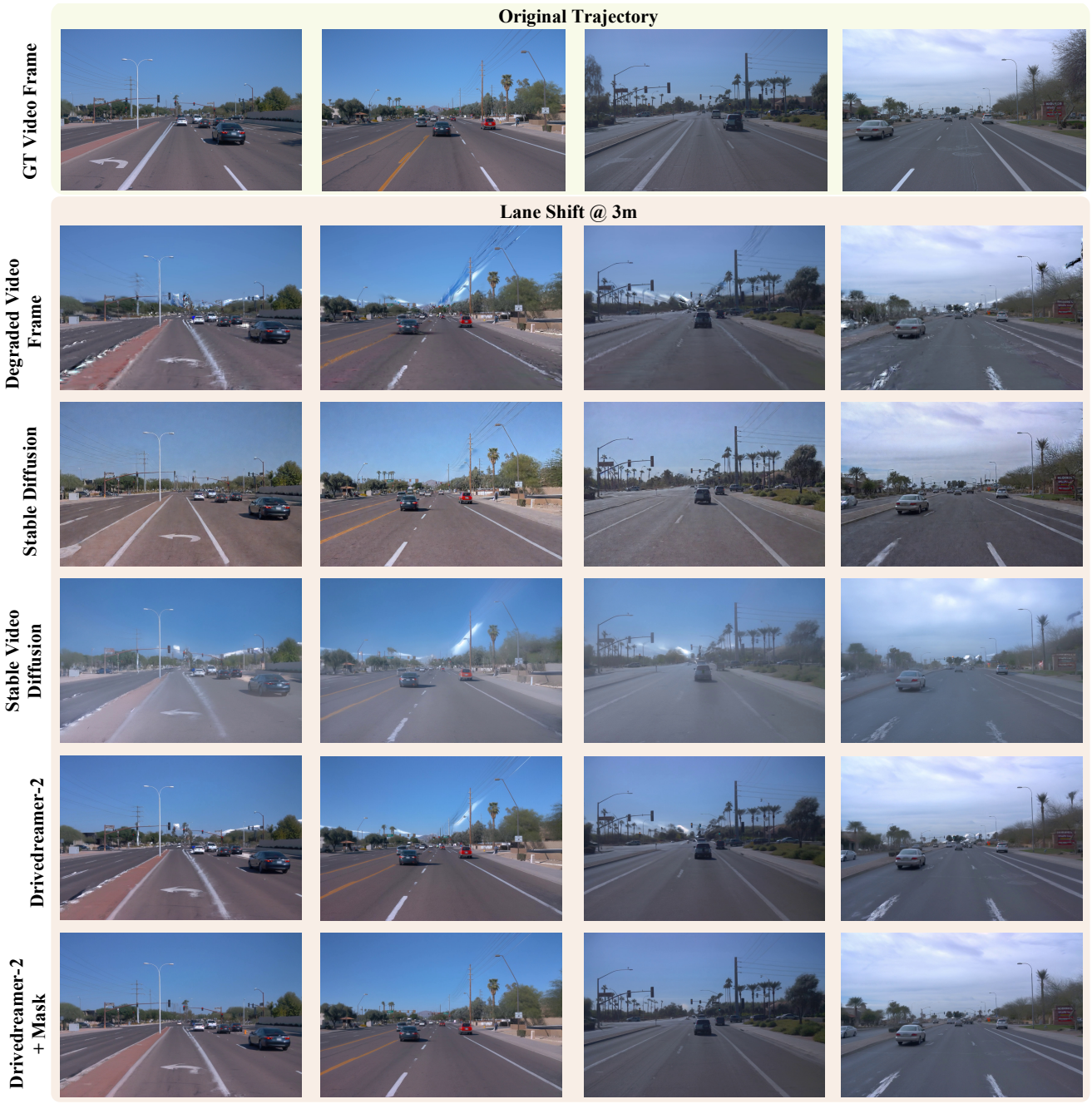

Figure 8. Qualitative comparison of the restoration effects achieved by *DriveRestorer* with different backbones. The yellow box contains the ground truth video frames of the original trajectories, while the pink boxes display the rendered video frames after the lane shift and the corresponding restored video frames by *DriveRestorer* with different backbones.

7. Baseline

PVG [8] introduces a novel unified representation model designed to capture dynamic scenes through the use of time-varying Gaussian distributions. These Gaussians are characterized by adjustable properties such as vibration direction, duration, and peak intensity. The approach distinguishes between static and dynamic elements by sorting the Gaussians according to their durations.

Deformable-GS [62] establishes a canonical space where scenes are represented using Gaussian distributions. For capturing dynamics, it employs a deformation network to forecast the offsets of Gaussian attributes, which subse-

Scene	Start Frame	End Frame
segment-10359308928573410754_720_000_740_000_with_camera_labels.tfrecord	120	159
segment-11450298750351730790_1431_750_1451_750_with_camera_labels.tfrecord	0	39
segment-12496433400137459534_120_000_140_000_with_camera_labels.tfrecord	110	149
segment-15021599536622641101_556_150_576_150_with_camera_labels.tfrecord	0	39
segment-16767575238225610271_5185_000_5205_000_with_camera_labels.tfrecord	0	39
segment-17860546506509760757_6040_000_6060_000_with_camera_labels.tfrecord	90	129
segment-3015436519694987712_1300_000_1320_000_with_camera_labels.tfrecord	40	79
segment-6637600600814023975_2235_000_2255_000_with_camera_labels.tfrecord	70	109

Table 6. Eight scenes from the Waymo dataset [46] featuring high interactive activity, numerous vehicles, and complex driving trajectories.

Figure 9. Qualitative comparison of the different stride settings in the PDUS.

quently adjust the Gaussians to align with the scene's dynamic changes

S^{3}Gaussian [24] is a method designed for efficient 3D scene reconstruction that operates without the need for expensive annotations. It achieves this by using 4D consistency to divide scenes into dynamic and static components, representing each with 3D Gaussians for detailed precision and employing a spatial-temporal field network to model the 4D dynamics compactly.

Street Gaussians [58] is a dynamic scene modeling method based on Gaussian Splatting for driving scenes. It separately models the static background and foreground vehicles. By utilizing boxes predicted by a pre-trained model, Street Gaussians warps the Gaussians of foreground vehicles and refines them during training.

DriveDreamer4D [68] is a method that enhances dynamic driving scene reconstruction by integrating with state-of-the-art techniques such as PVG [8], Deformable-GS [62], and S^{3}Gaussian [24]. It leverages world model priors to synthesize novel trajectory videos, where structured conditions are explicitly utilized to control the spatial-temporal consistency of traffic elements.

8. Experiment Results

Qualitative Results of *DriveRestorer* Backbone. As shown in Fig. 8, we compare restoration effects achieved by *DriveRestorer* with different backbones. The images rendered under the new trajectories exhibit several defects, including distorted and blurred distant trees, flying points in the sky, and partially obscured foreground vehicles. The *DriveRestorer* based on Stable Diffusion [44] demonstrates

promising performance, repairing the background and effectively correcting the distortion of foreground vehicles. However, image restoration methods lack spatial continuity, causing the repaired foreground vehicles to appear in incorrect positions or even exhibit color changes. For instance, in the second column, some distant vehicles that are originally red turned into grey. The video-based method, Stable Video Diffusion [4], offers improved spatial continuity but encounters challenges due to the great difficulty of fine-tuning. Although it restores many distorted vehicles, the video frames show significant color differences from the original and sky defects remain unrepaired. Then, DriveDreamer-2 [69] introduces control conditions, such as 3D boxes and HDMaps, which resolve the issue of color discrepancies and improve the restoration of background elements like lane lines. Finally, incorporating masks during the fine-tuning process of DriveDreamer-2 [69] further enhances the repair of sky defects, making the restored video frame more realistic.

Qualitative Results of Progressive Data Update Strategy. In Figure 9, we compare different stride settings in the PDUS, including $\Delta y = 1.5$ and 3. Although *ReconDreamer* is effective in enhancing image quality and reducing artifacts for both stride values, an excessively large stride can lead to poorer reconstruction of lane markings and distant scenes.

We provide a video that includes more comparisons with the baseline. For further details, please refer to the file videos/comparison.mp4.