

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

Chaojun Ni^{*1, 2}, Guosheng Zhao^{*2, 3, 4}, Xiaofeng Wang^{*2, 3, 4}, Zheng Zhu^{*†2},
Wenkang Qin², Guan Huang², Chen Liu⁵, Yuyin Chen⁵, Yida Wang⁵, Xueyang Zhang⁵,
Yifei Zhan⁵, Kun Zhan⁵, Peng Jia⁵, Xianpeng Lang⁵, Xingang Wang³, Wenjun Mei^{†1}

¹Peking University, China ²GigaAI, China

³Institute of Automation, Chinese Academy of Sciences, China

⁴School of Artificial Intelligence, University of Chinese Academy of Sciences, China

⁵Li Auto Inc., China

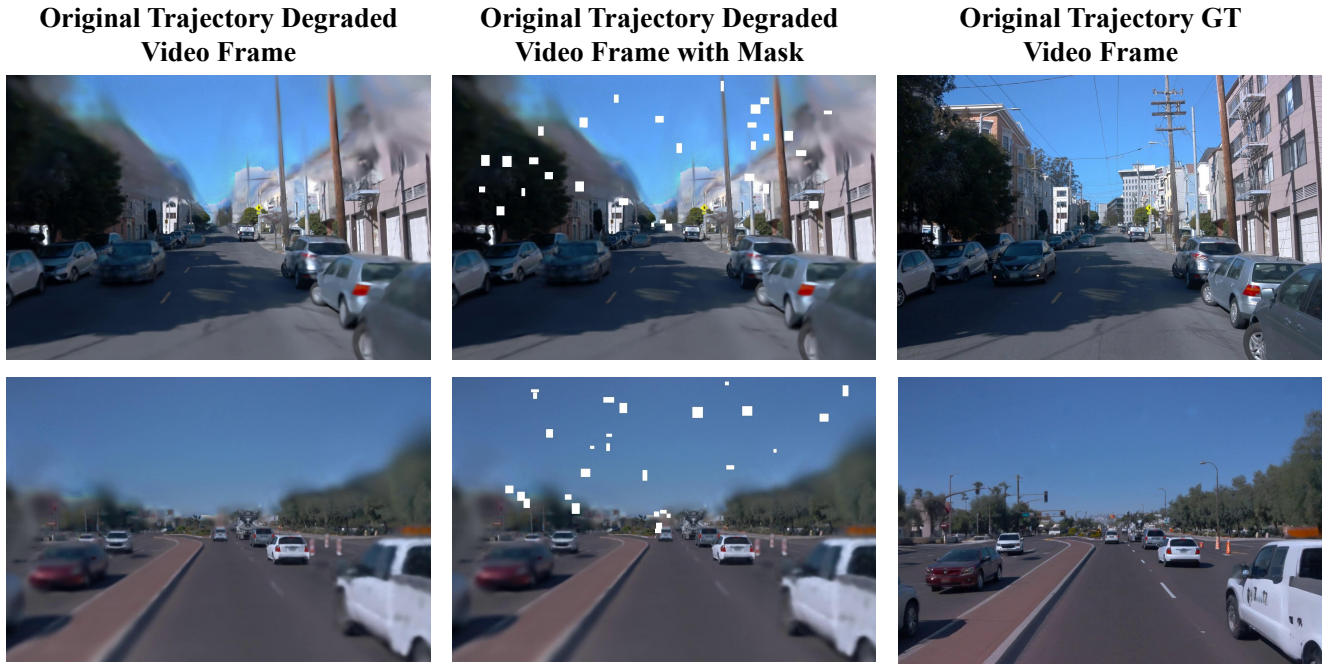


Figure 1. Examples of degraded video frames, the corresponding masked video frames, and GT video frames under the original trajectories.

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.

1. Implementation Details

Training for *DriveRestorer*. As shown in Fig. 1, 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 [10]. These metrics are specifically designed to evaluate the spatiotemporal coherence of

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 1. Eight scenes from the Waymo dataset [7] featuring high interactive activity, numerous vehicles, and complex driving trajectories.

foreground agents and background lanes, respectively.

The NTA-IoU processes images rendered under new trajectories using the YOLO11 [5] 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 [2] model to detect lanes in images rendered under new trajectories, while projecting the original trajectory lanes onto the new trajectory via geometric transformations. Finally, the mean Intersection over Union (IoU) between the projected and detected lanes is calculated.

Scene Selection. We select eight scenes from the validation set of the Waymo dataset [7]. 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. 1, 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 [10] and Street Gaussians [8]. 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.

2. Baseline

PVG [3] 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 [9] 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 subsequently adjust the Gaussians to align with the scene’s dynamic changes

S³Gaussian [4] 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 [8] 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 [10] is a method that enhances dynamic driving scene reconstruction by integrating with state-of-the-art techniques such as PVG [3], Deformable-GS [9], and S³Gaussian [4]. 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.

3. Experiment Results

Qualitative Results of DriveRestorer Backbone. As shown in Fig. 2, 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

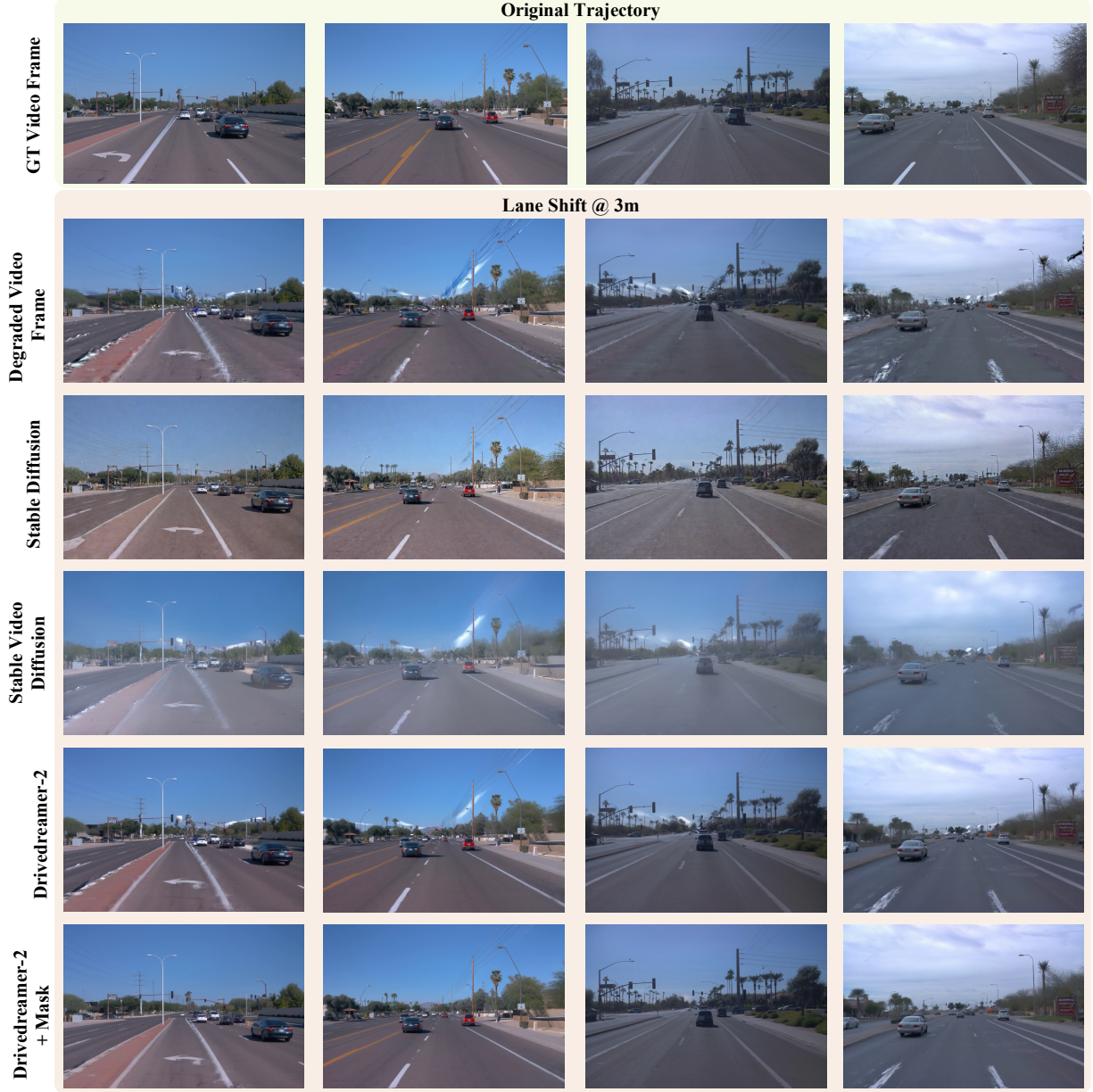


Figure 2. Qualitative comparison of the restoration effects achieved by *DriveRestorer* with different backbones. The yellow box contains the ground truth video frames of the original trajectories, while the pink boxes display the rendered video frames after the lane shift and the corresponding restored video frames by *DriveRestorer* with different backbones.

DriveRestorer based on Stable Diffusion [6] 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 [1], 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 [11] 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 [11] further en-



Figure 3. Qualitative comparison of the different stride settings in the PDUS.

hances the repair of sky defects, making the restored video frame more realistic.

Qualitative Results of Progressive Data Update Strategy. In Figure 3, 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`.

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 arXiv:2311.15127*, 2023. 3
- [2] 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. 2
- [3] 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
- [4] Nan Huang, Xiaobao Wei, Wenzhao Zheng, Pengju An, Ming Lu, Wei Zhan, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. s^3 gaussian: Self-supervised street gaussians for autonomous driving. *arXiv preprint arXiv:2405.20323*, 2024. 2
- [5] Glenn Jocher and Jing Qiu. Ultralytics yolo11, 2024. 2
- [6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3
- [7] Pei Sun, Henrik Kretschmar, 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. 2
- [8] 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. 2
- [9] 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. 2
- [10] 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
- [11] 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. 3