Driving View Synthesis on Free-form Trajectories with Generative Prior

Supplementary Material

A. Discussion on limitations

Beyond demonstrating the efficacy of the proposed method in synthesizing driving views along arbitrary trajectories, this paper also acknowledges several limitations. First, incorporating a diffusion-based generative model significantly increases reconstruction time. But such complexity might be unnecessary, as the required generative diversity is substantially constrained under our inpainting formulation. suggesting that a more lightweight model can be employed in the future. Second, our method currently employs a simple panning camera scheme to sample novel trajectories to be restored. However, this fix design may not be optimal for all scenarios, and leave some hand-designed components in the pipeline. Future work could consider to adaptively explore the optimal novel trajectories in optimization.

B. More implementation details

Metric details Previous driving scene reconstruction methods [4, 5, 14] typically utilize PSNR, SSIM [10] and LPIPS [16] for evaluate image level similarity with ground truth. However, these metrics are inapplicable for evaluating free-form trajectories due to the absence of corresponding ground-truth images for novel trajectories.

Therefore, inspired by [6, 9, 17], we propose a novel set of metrics to evaluate the recognizability of high-level traffic elements in synthesized views, based on ground truth annotations in 3D space. For vehicles, we projected their ground truth 3D bounding boxes onto the novel views to obtain corresponding 2D bounding boxes. Since IoU lacks the precision granularity needed to understand performance across various intersection over union threshold. Therefore, we calculate the Average Precision (AP[50:95]) between the projected ground truth and those detected by Faster R-CNN [7], which provides a more nuanced assessment than IoU.

For road lanes, NTL-IoU adopted in [6, 17] calculates the mIoU between the lanes on synthesized views detected by [1] and the ground truth, averaging both foreground lanes and backgrounds. However, including the background can inflate accuracy scores due to the easy detection of non-lane areas. Therefore, following standard lane segmentation evaluation [1], we focus specifically on IoU with only the foreground to better evaluate lane synthesis quality.

Warping operation ϕ in Eq. (8) of main paper Given a camera extrinsic matrix $E \in \mathbb{R}^{4\times4}$ and intrinsic matrix $K \in \mathbb{R}^{3\times3}$, the mapping from world points $p_{\text{world}} \in \mathbb{R}^3$ to pixel coordinates (x, y) and depth d can be expressed by:

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

$$(x_*, y_*, d)^\top = K p_{\text{camera}} \tag{2}$$

$$(x, y) = (\frac{x_*}{d}, \frac{y_*}{d}) \tag{3}$$

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

Validation sequences There are 18 curated driving sequences from WOD [8] used for qualitative and quantitative evaluations in our paper. The names of all segments are listed below:

- `10359308928573410754_720_000_740_000`
- `11450298750351730790_1431_750_1451_750`
- `12496433400137459534_120_000_140_000`
- `15021599536622641101_556_150_576_150`
- `16767575238225610271_5185_000_5205_000`
- `17860546506509760757_6040_000_6060_000`
- `30154365196949877l2_1300_000_1320_000`
- `66376006008140239 75_2235_000_2255_000`
- `10444454289801298640_4360_000_4380_000`
- `10588771936253546636_2300_000_2320_000`
- `10625026498155904401_200_000_220_000`
- `11017034898130016754_697_830_717_830`
- `11917887606306240 72_3880_000_3900_000`
- `11928449532664718059_1200_000_1220_000`
- `14810689888487451189_720_000_740_000`
- `44142354784453766 89_2020_000_2040_000`
- `62428225833984874 96_73_000_93_000`
- `76701030065805497 15_360_000_380_000`

Unless stated otherwise, all quantitative ablation experiments are conducted on segments `15021599536622641101_556_150_576_150` and `30154365196949877l2_1300_000_1320_000`, while the results are evaluated on novel trajectories with lateral shift $3m$.

C. Additional 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. 1, we show the metrics for novel trajectory synthesis (IoU, AP, and FID), 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.

Noise level	IoU↑	AP ↑	FID↓	Time↓
0.4	0.2184	0.6202	96.21	**1.0**×
0.6	**0.2263**	**0.6389**	**74.34**	1.1×
0.8	0.2192	0.6259	81.55	1.2×
1.0	0.2113	0.6293	106.01	1.3×

Table 1. **Effect of the noise level.** The noise level l denotes adding the noise corresponding to the timestep $t = \lfloor lN \rfloor$, where N is the total step needed for denoising from scratch.

Gaussian model	w/ ours	IoU↑	AP ↑	FID↓
PVG [2]		0.0200	0.5401	118.41
PVG [2]	✓	**0.1872**	**0.6197**	**91.85**
StreetGaussian [12]		0.1217	0.6124	103.08
StreetGaussian [12]	✓	**0.2263**	**0.6389**	**74.34**

Table 2. **Ablation study on different Gaussian model.**

	IoU↑	AP ↑	FID↓
w/o progress.	0.2204	0.6372	75.41
w/o warp	0.2196	0.6357	105.77
Full model	**0.2263**	**0.6389**	**74.34**

Table 3. **Ablations on other design choices.** "w/o progress" denotes that the shift length is fixed at $4m$ throughout the optimization. "w/o warp" means that the denoising is started from the noisy rendered images rather than warped views.

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, refine strength of 0.6 is found to be the best balance.

Different Gaussian model Although we adopt static 3D Gaussian splatting [12] as the default scene representation in previous experiments, our method is also capable of synergy with different choices. To demonstrate the versatility of DriveX, we also evaluate our method on a representative dynamic Gaussian representation for driving view synthesis [2]. From Tab. 2, we can observe that our strategy consistently reduces the FID by a large margin by substantially reducing artifacts in novel views, and effectively recovering street lanes and cars (*i.e.* IoU, AP) that are entirely indiscernible in the baseline.

Projected LiDAR points as auxiliary oracle Although using the masked rendered image as the only condition is sufficient to achieve satisfactory performance in most scenes, incorporating projected LiDAR points as the additional oracle can greatly improve robustness in some difficult cases. In these scenes, initial poor novel view renderings would result in large masked regions, LiDAR projection can provide direct guidance for precisely restoring them. This resulted enhancement is reflected by the quantitative comparison in Tab. 4.

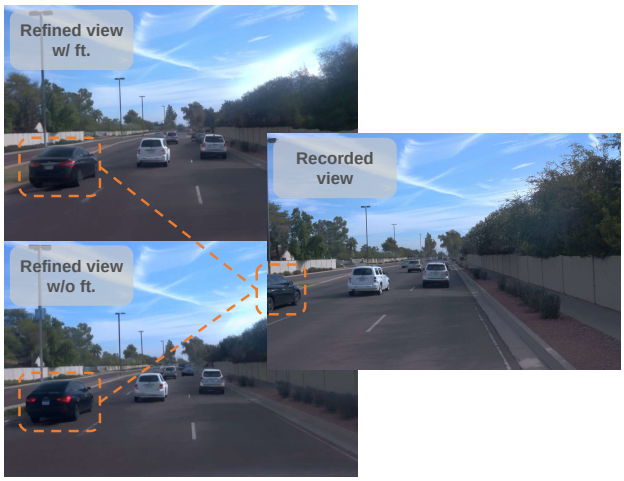

Figure 1. **Domain-specific fine-tuning enhances the style adaptation on driving videos.** Here the appearance of the black car refined by model without fine-tuning seems to lack details and has unrealistic material.

w/ render	w/ LiDAR	w/ ft.	IoU↑	AP ↑	FID↓
✓			0.2092	0.6257	90.30
✓	✓		0.2123	0.6334	86.30
	✓	✓	0.2153	0.6225	75.18
✓	✓	✓	**0.2263**	**0.6389**	**74.34**

Table 4. **Ablations on the components of generative model.** "w/ render" means using masked novel view image as condition. "w/ LiDAR" denotes adopting LiDAR projection as condition. "w/ ft." represents training diffusion model on driving video data [8].

Fine-tuning generative model on driving videos The comparison between the 2^{nd} and 4^{th} row of Tab. 4 demonstrates that the diffusion model fine-tuned on in-domain driving videos outperforms that trained solely on general-purpose 3D datasets [13], especially in FID. This improvement comes from the generative model's adaptation to the style of real driving scenes. As shown in Fig. 1, the counterpart without domain-specific fine-tuning tends to generate cartooned foreground vehicles without detailed texture. Therefore, beyond the improvement in quantitative metrics, it also simplifies our pipeline by removing the need for some hand-designed processing, such as lowering the refine weight for foreground vehicles.

Other designs We also ablate other design choices in Tab. 3, including progressive increasing shifted length and warped condition. The results show that the final qual-

Figure 2. **Reconstruction results on generated videos.** All ground truth videos used for reconstruction are generated by Vista [3]. The first and the second rows show the novel trajectory synthesis results by the baseline [12] and our method, respectively.

Method	FID ↓		
	$\pm 0m$	$\pm 1m$	$\pm 2m$
Recon-only [12]	45.45	96.76	146.61
DriveX (ours)	**45.44**	**92.77**	**142.73**

Table 5. **Quantitative comparison on generated video.** We compared the FID score on videos generated by Vista [3].

ity degrades without any element, which indicates their helpfulness on the rendering quality and the stability of optimization.

D. Application on generated videos

Another intriguing application of the proposed method is transforming AI-generated video into a re-enactable driving world. 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.

Details Specifically, since the generated videos cannot be precisely aligned with the conditioned camera parameters, we employ a feed-forward approach [15] 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. (9) of main paper to ensure well-defined unreliability scores for all areas of the novel view. Besides, we include an entropy loss for the opacity of each Gaussian to prevent potential overfitting.

Figure 3. **Results on the PandaSet dataset [11].** The results are on the novel trajectory shifted $2.0m$ from the recoded trajectory.

Results To quantitatively evaluate this capability, we compared DriveX and the reconstruction-only baseline [12] on 12 driving videos generated by Vista [3]. Due to the evident stylistic gap between the generated videos and those in WOD, we only use the diffusion model trained on general-purpose 3D datasets as generative prior without LiDAR projection as condition. The FID of videos rendered across different novel trajectories is reported in Tab. 5. The results indicate that the realism of rendered videos is effectively 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 $\pm 1m$ and $\pm 2m$ 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. 2. As can be seen, the baseline method exhibits significant degradation in views that are far from the recorded trajectory.

E. Results on the PandaSet dataset

To further validate the versatility of the proposed method on a broader range of scene types and sensor configurations, we apply our approach to two sequences indexed 001 and 003 from the PandaSet dataset [11], each consisting of 80

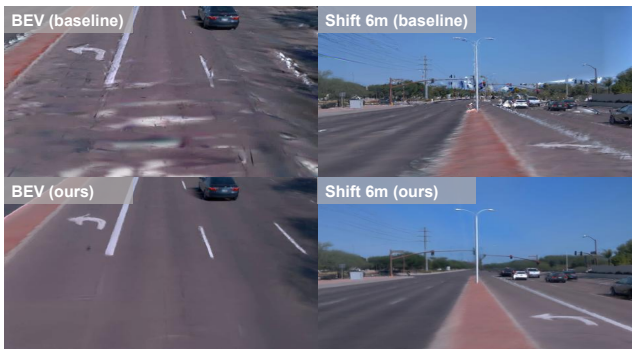

Figure 4. **Results on more aggressive view change.**

Figure 5. **Failure cases.** Our method cannot *reconstruct* a feasible appearance from scratch for the side of a vehicle that is invisible from all recorded views.

frames. The image resolution is downsampled to 450×800 for both training and evaluation. In Fig. 3, we compared our method with the state-of-the-art reconstruction baseline StreetGaussian [12] on the trajectory with shift length $2.0m$. The result suggests that our method can significantly reduce artifacts in synthesized images and improve the discernibility of safety-critical traffic elements, which is also corroborated by our better FID.

F. More quanlitative results

More aggressive view change. We test the proposed method on some challenging cases to assess its robustness. In Fig. 4 (left), the view is rendered from the camera elevated by $3m$ above the recorded trajectory, with a pitch angle of $30°$. In Fig. 4 (right), we increase the lateral shift length to $6m$. To accommodate these aggressively changed views, we include them in the refined trajectory during optimization. As shown in the results, the baseline method almost totally fails to synthesize meaningful images under these viewpoints, while our solution substantially improves the fidelity.

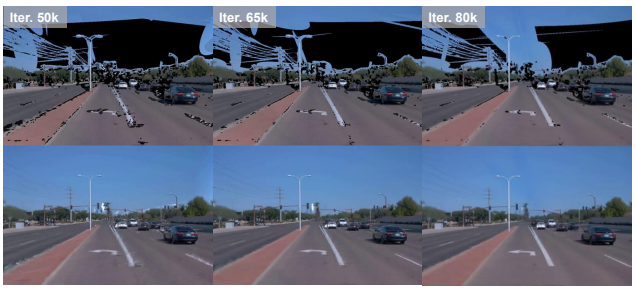

Figure 6. **The evolution of unreliability mask during optimization.**

Failure case. Despite the robustness and generalization ability on diverse scenes as demonstrated above, it should be noted that the proposed framework only aims to provide regularization for unconstrained reconstruction. As a reconstruction pipeline, it does not work well on totally unseen regions of the scene. For example, in Fig. 5, our method cannot create a feasible appearance from scratch for the side of a vehicle that is invisible from all recorded views and thus lacks proper geometry initialization before the refinement stage.

The evolution of unreliability mask during optimization. In Fig. 6, we present novel view renderings and the corresponding unreliability masks at different training iterations. The results clearly show that the reliable regions gradually expand during iterative refinement, demonstrating the effectiveness of the proposed unreliability mask mechanism–it can serve as a reliable cue for the rendered images rather than introducing compounding errors. Additionally, the distribution of the mask indicates that unreliable regions mainly appear near edge areas of the visual elements, which is the rationale behind the adoption of edge-aware inpainting.

More comparisons with sota methods The video comparison with the state-of-the-art alternates can be found in our project page.

References

[1] 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. 1

[2] 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. 2

[3] 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. 3

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

[5] 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

[6] Chaojun Ni, Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Wenkang Qin, Guan Huang, Chen Liu, Yuyin Chen, Yida Wang, Xueyang Zhang, Yifei Zhan, Kun Zhan, Peng Jia, Xianpeng Lang, Xingang Wang, and Wenjun Mei. Recondreamer: Crafting world models for driving scene reconstruction via online restoration. *arXiv preprint*, 2024. 1

[7] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 1

[8] 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. 1, 2

[9] Qitai Wang, Lue Fan, Yuqi Wang, Yuntao Chen, and Zhaoxiang Zhang. Freevs: Generative view synthesis on free driving trajectory. In *ICLR*, 2025. 1

[10] 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. 1

[11] Pengchuan Xiao, Zhenlei Shao, Steven Hao, Zishuo Zhang, Xiaolin Chai, Judy Jiao, Zesong Li, Jian Wu, Kai Sun, Kun Jiang, et al. Pandaset: Advanced sensor suite dataset for autonomous driving. In *ITSC*, 2021. 3

[12] 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. 2, 3, 4

[13] 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

[14] 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

[15] 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. 3

[16] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 1

[17] 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