

Supplementary Material for DriveDreamer4D: World Models Are Effective Data Machines for 4D Driving Scene Representation

Guosheng Zhao^{1, 2, 3*}, Chaojun Ni^{4*}, Xiaofeng Wang^{1, 2, 3*}, Zheng Zhu^{5*†},
Xueyang Zhang⁶, Yida Wang⁶, Guan Huang⁵, Xinze Chen⁵, Boyuan Wang^{1, 2, 3},
Youyi Zhang⁷, Wenjun Mei⁴, Xingang Wang^{2, 3†}

¹Institute of Automation, Chinese Academy of Sciences, China

²School of Artificial Intelligence, University of Chinese Academy of Sciences, China

³Luoyang Institute for Robot and Intelligent Equipment, China ⁴Peking University, China

⁵GigaAI, China ⁶Li Auto Inc., China ⁷Technical University of Munich, Germany

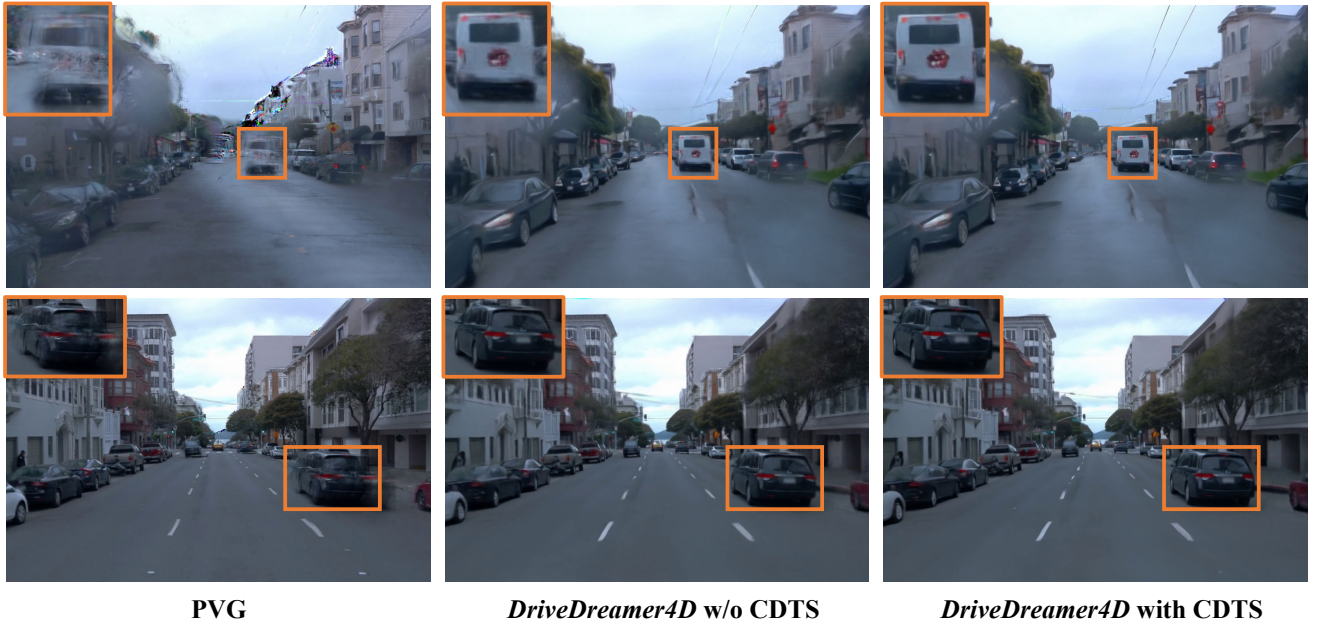


Figure 1. Visual comparisons in the novel trajectories for the Cousin Data Training Strategy (CDTS) ablation study. The orange boxes emphasize the superior performance of *DriveDreamer4D* and the further improvements in detail rendering brought by CDTS.

In the supplementary material, we begin by introducing the three baseline methods employed in our work. Next, we elaborate on the implementation of *DriveDreamer4D*, covering the training for novel trajectory video generation, the selection of scenes, and the setup of the user study. Finally, additional visualizations are presented to illustrate the improved rendering quality achieved through Cousin Data Training Strategy (CDTS) and showcase the performance of *DriveDreamer4D* in speed change scenarios.

*These authors contributed equally to this work.

†Corresponding authors. zhengzhu@ieee.org, xingang.wang@ia.ac.cn.

‡Project Page: <https://drivedreamer4d.github.io>

1. Baselines

To demonstrate the effectiveness and generalizability of our method, three different 4D Gaussian Splatting (4DGS) baselines are selected for the experiments. In this section, we briefly introduce the three baselines employed in this paper: PVG [3], S³Gaussian [5], and Deformable-GS [10].

PVG [3] introduces a unified representation model known as Periodic Vibration Gaussians (PVGs), which vibrate over time with optimizable parameters, including vibration directions, lifespan, and life peak (the moment of highest opacity), to effectively represent dynamic scenes. The

Scene	Start Frame	End Frame
segment-10359308928573410754_720_000_740_000_with_camera_labels.tfrecord	120	159
segment-12820461091157089924_5202_916_5222_916_with_camera_labels.tfrecord	0	39
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-17152649515605309595_3440_000_3460_000_with_camera_labels.tfrecord	60	99
segment-17860546506509760757_6040_000_6060_000_with_camera_labels.tfrecord	90	129
segment-2506799708748258165_6455_000_6475_000_with_camera_labels.tfrecord	80	119
segment-3015436519694987712_1300_000_1320_000_with_camera_labels.tfrecord	40	79

Table 1. Selected scenes from the validation set of the Waymo dataset [8].

Method	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
PVG	15.73	0.7093	0.2568
<i>DriveDreamer4D</i> with PVG	17.81	0.7601	0.2260

Table 2. PSNR, SSIM and LPIPS scores on EUVS test set.

model employs a self-supervised approach to optimize these Gaussians and achieves static-dynamic decomposition by classifying them based on their lifespans. This method allows PVG to effectively represent the characteristics of various objects and elements in dynamic urban scenes.

S³Gaussian [5] proposes a self-supervised street Gaussian method to model complex 4D dynamic scenes. Each scene is represented using 3D Gaussians to preserve explicitness, and a spatial-temporal field network is employed to compactly model the 4D dynamics. To facilitate efficient scene reconstruction without costly annotations, it utilizes a self-supervised approach to decompose dynamic and static 3D Gaussians.

Deformable-GS [10] represents scenes using a canonical space defined by Gaussian distributions. It models scene dynamic by employing a deformation network to predict offsets for the Gaussian parameters. These offsets adjust the Gaussians to align with the dynamic elements of the scene. Additionally, Deformable-GS has demonstrated strong performance in both synthetic and indoor datasets.

2. Implementation Details

In Sec. 2, we primarily introduce the training for novel trajectory video generation, the selection of scenes, and the details of the user study.

Training for Novel Trajectory Video Generation. As depicted in the upper part of Fig. 2 (in the main text), a controllable driving video generation model is crucial for producing novel trajectory videos. Specifically, we follow the approach outlined in [11] to train such a model on the Waymo dataset [8]. Unlike [11], which focuses on multi-view video generation using the nuScenes dataset [2], our

Method	NTA-IoU \uparrow	NTL-IoU \uparrow	FID \downarrow
PVG	0.256	50.70	105.29
VEGS	0.417	51.95	109.31
FreeVS	0.426	52.08	128.63
<i>DriveDreamer4D</i> with PVG	0.438	53.06	71.52

Table 3. Comparison of NTA-IoU, NTL-IoU and FID scores with more SOTA methods across lane change on Waymo.

Method	Frames	Views	NTA-IoU \uparrow	NTL-IoU \uparrow	FID \downarrow
PVG	40	3	0.334	52.94	106.91
<i>DriveDreamer4D</i> with PVG	40	3	0.497	55.11	69.48
PVG	100	1	0.320	52.08	86.66
<i>DriveDreamer4D</i> with PVG	100	1	0.508	56.27	65.39

Table 4. Experiments of different frames and views on Waymo.

work concentrates solely on front-view video generation. This focus allows us to increase the number of frames to 40 and the resolution to 960×640 , a significant improvement compared to the previous 8 frames at a resolution of 448×256 . The increase in both frame count and resolution contributes to an enhanced performance of the reconstruction model, particularly for novel trajectory generation. As for the training data, it comprises the entire Waymo training split, consisting of 798 videos. To enhance the dataset, we further divide these videos into 40-frame clips, resulting in approximately 64K clips. Additionally, the training process is initialized with parameters from SVD [1], with 3D bounding boxes, HDMaps, and text incorporated as control conditions. And, the AdamW optimizer [7] is employed for parameter optimization, with a learning rate of 5×10^{-5} , a batch size of 8, and a total of 50K iterations. All experiments are conducted on an NVIDIA H20 (96GB) GPUs.

Scene Selection. All selected scenes are sourced from the validation set of the Waymo dataset [8] and are carefully chosen based on their distinctive characteristics. Specifically, the selection prioritizes scenes that exhibit significant motion dynamics, such as large-scale maneuvers, as these scenarios pose greater challenges for both video re-



Figure 2. Qualitative comparisons of novel trajectory renderings during lane change scenarios. The orange boxes highlight that *DriveDreamer4D* significantly enhances the rendering quality across various baselines (PVG [3], S^3 Gaussian [5], Deformable-GS [10]).

construction and trajectory generation tasks. Tab. 1 shows all 8 scenes selected for our experiments. The official file names of these scenes, as provided in [8], are listed along with their respective starting and ending frames.

User Study. For the eight different scenes mentioned above, we create 72 comparison videos for the user study, covering three novel trajectories (acceleration, deceleration,

and lane change) under three different baselines. To ensure fairness, the baseline and our method were randomly assigned to the left or right side of each comparison video. For each comparison, the participants are asked to choose the result they deem the most accurate or realistic (either the left or right side).

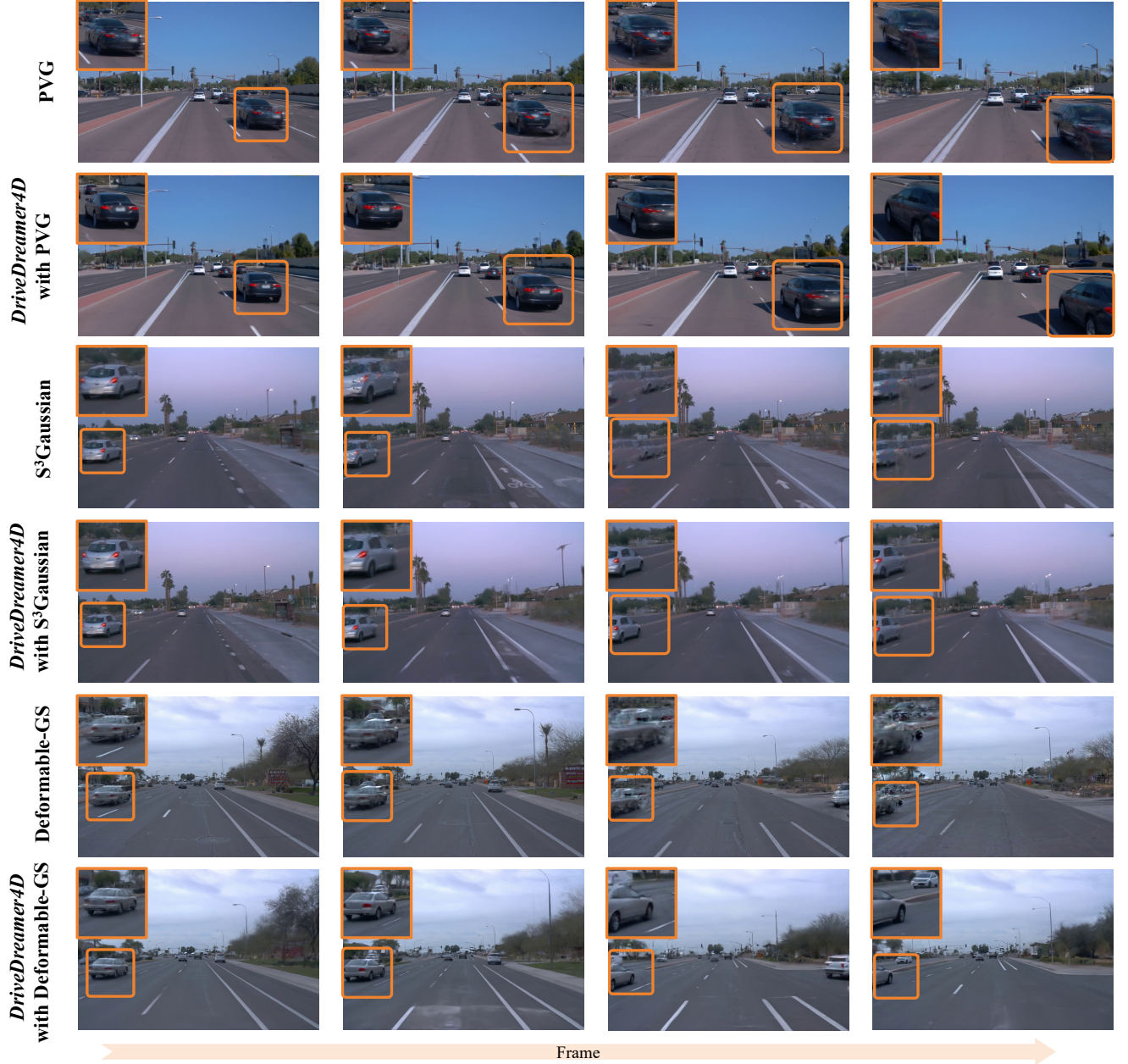


Figure 3. Qualitative comparisons of novel trajectory renderings during speed change scenarios. The orange boxes highlight that *DriveDreamer4D* significantly enhances the rendering quality across various baseline methods (PVG [3], S^3 Gaussian [5], Deformable-GS [10]).

3. More Experiments

In this section, our experiments primarily focus on *DriveDreamer4D* with PVG [3], highlighting key metrics for lane change trajectories. For the EUVS benchmark [4], we test on the Level 1 (translation) dataset, as it aligns closely with our objectives. Tab. 2 demonstrates the impressive results on extrapolated views. Furthermore, we compare *DriveDreamer4D* with VEGS [6] and FreeVS [9]. As shown in

Tab. 3, *DriveDreamer4D* outperforms these SOTA methods. For experiments of more frames and views, please see Tab. 4.

4. Visualization

In this part, we present additional visualization results, including qualitative analyses from the Cousin Data Training Strategy (CDTS) ablation study and visual comparisons for

speed change scenarios.

As mentioned in Sec. 4.3 of the main text, we perform an ablation study on the CDTS using PVG [3]. For clarity, *DriveDreamer4D* in this ablation study refers to *DriveDreamer4D* with PVG. As shown in Fig. 1, *DriveDreamer4D* demonstrates significant improvement over the baseline methods, regardless of whether CDTS is applied. Notably, the baseline methods struggle to accurately reconstruct the positions of vehicles in novel trajectories, resulting in severe ghosting artifacts. In contrast, *DriveDreamer4D* excels at rendering the vehicle positions with high precision, significantly enhancing rendering performance. Moreover, with the introduction of CDTS, *DriveDreamer4D* further enhances the reconstruction quality of dynamic vehicles, particularly at the edges, providing more detailed and accurate representations.

As shown in Fig. 2, we present the novel trajectory view synthesis during lane change. Images rendered by the baseline methods exhibit issues where foreground vehicles incorrectly change lanes in sync with the camera’s motion, and some vehicles are incompletely rendered. Additionally, the background is filled with speckles and ghosting. Especially shown in the rightmost column of Fig. 2, baseline methods often produce blurred, ghosted foreground vehicles and background speckles in the sky, alongside blurred lane markings. Our method, however, significantly improves rendering quality, as highlighted by the orange boxes, with sharper vehicle contours and reduced background artifacts like speckles and ghosting. For more details, please refer to the file `videos/lane_change_comparison.mp4`. More qualitative analysis of novel trajectory view renderings are shown in Fig. 3, focusing on speed change scenarios. Our method significantly enhances the positional accuracy of foreground vehicles and background elements under speed change scenarios. Specifically, baseline results (PVG [3], S^3 Gaussian [5], Deformable-GS [10]) are displayed in rows 1, 3, and 5. It is evident that the baseline methods face challenges with perspective synthesis in speed-change scenarios, resulting in inaccurate positional shifts (such as blurring or disappearance of foreground vehicles). In contrast, the integration of *DriveDreamer4D* enables the 4DGS algorithms to achieve superior spatial consistency and significantly improved rendering quality, as illustrated by the orange boxes in the Fig. 3. More details can be found in the file `videos/speed_change_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. 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 multi-modal dataset for autonomous driving. In *CVPR*, 2020. 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. 1, 3, 4, 5
- [4] Xiangyu Han, Zhen Jia, Boyi Li, Yan Wang, Boris Ivanovic, Yurong You, Lingjie Liu, Yue Wang, Marco Pavone, Chen Feng, and Yiming Li. Extrapolated urban view synthesis benchmark. *arXiv preprint arXiv:2412.05256*, 2024. 4
- [5] 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. 1, 2, 3, 4, 5
- [6] 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. 4
- [7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2
- [8] 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, 3
- [9] Qitai Wang, Lue Fan, Yuqi Wang, Yuntao Chen, and Zhaoxiang Zhang. Freevs: Generative view synthesis on free driving trajectory. *arXiv preprint arXiv:2410.18079*, 2024. 4
- [10] 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. 1, 2, 3, 4, 5
- [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. 2