DriveGen3D: Boosting Feed-Forward Driving Scene Generation with Efficient Video Diffusion

Weijie Wang[1,2*] **Jiagang Zhu**[1,3*] **Zeyu Zhang**[1] **Xiaofeng Wang**[1,4]
Zheng Zhu[1✉] **Guosheng Zhao**[1,4] **Chaojun Ni**[1,5] **Haoxiao Wang**[2]
Guan Huang[1] **Xinze Chen**[1] **Yukun Zhou**[1] **Wenkang Qin**[1]
Duochao Shi[2] **Haoyun Li**[1,4] **Guanghong Jia**[3] **Jiwen Lu**[3]

[1]GigaAI [2]Zhejiang University [3]Tsinghua University
[4]Institute of Automation, Chinese Academy of Sciences [5]Peking University

[*]Equal contribution [✉]Corresponding author

Abstract

We present DriveGen3D, a novel framework for generating high-quality and highly controllable dynamic 3D driving scenes that addresses critical limitations in existing methodologies. Current approaches to driving scene synthesis either suffer from prohibitive computational demands for extended temporal generation, focus exclusively on prolonged video synthesis without 3D representation, or restrict themselves to static single-scene reconstruction. Our work bridges this methodological gap by integrating accelerated long-term video generation with large-scale dynamic scene reconstruction through multimodal conditional control. DriveGen3D introduces a unified pipeline consisting of two specialized components: FastDrive-DiT, an efficient video diffusion transformer for high-resolution, temporally coherent video synthesis under text and Bird's-Eye-View (BEV) layout guidance; and FastRecon3D, a feed-forward reconstruction module that rapidly builds 3D Gaussian representations across time, ensuring spatial-temporal consistency. Together, these components enable real-time generation of extended driving videos (up to 424×800 at 12 FPS) and corresponding dynamic 3D scenes, achieving SSIM of 0.811 and PSNR of 22.84 on novel view synthesis, all while maintaining parameter efficiency.

1 Introduction

The synthesis of 3D dynamic driving environments has emerged as a key research frontier in autonomous systems, driven by its wide-ranging applications in simulation, perception, and planning. While recent advances in video generation [1–7] and 3D scene reconstruction [8–16] have made substantial progress, a critical gap remains: the lack of an integrated and efficient framework that unifies long-horizon video synthesis and large-scale 3D scene reconstruction under multimodal control.

Existing methodologies typically address either temporal coherence in video generation or spatial fidelity in scene reconstruction—but not both—often requiring high computational cost or suffering from limited scalability. For example, state-of-the-art diffusion-based models like MagicDriveDiT [2] can produce high-resolution driving sequences but require up to 30 minutes to generate a single 1600×900 video, making them impractical for real-time use.

From the reconstruction perspective, optimization-based approaches [17, 8–10] are similarly time-consuming, often needing 30 minutes per scene. While recent feed-forward methods [18–20, 11–13] have reduced reconstruction time to seconds, they remain limited in scale and rarely integrate with dynamic video generation pipelines.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

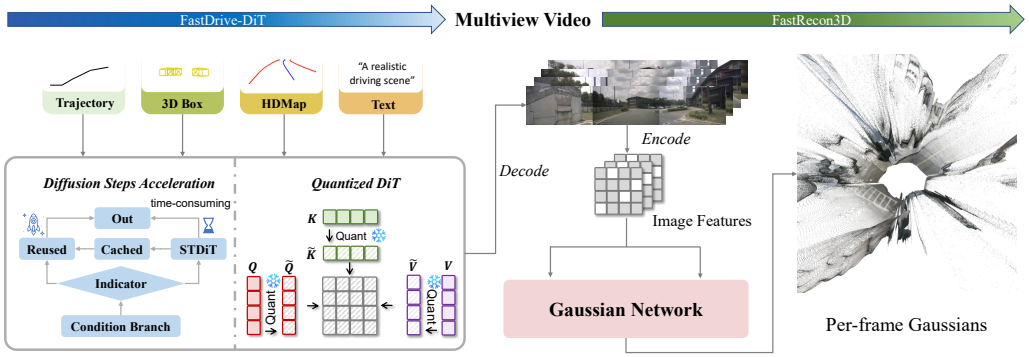

Figure 1: **Overview of DriveGen3D.** (a) Given textual and BEV layout conditions, our model first employs an accelerated Video Diffusion Transformer to synthesize a long driving video. (b) Next, a per-frame 3D Gaussian Splatting representation is utilized to construct entire scene from the generated video frames.

To bridge this gap, we propose **DriveGen3D**, an efficient and unified framework that integrates two specialized modules: **FastDrive-DiT**, an accelerated video diffusion transformer for high-resolution driving video generation, and **FastRecon3D**, a feed-forward reconstruction pipeline that builds dynamic 3D scenes from multi-view video frames in real time. FastDrive-DiT employs both diffusion step caching and quantized attention to reduce inference time by over $2\times$, while FastRecon3D leverages temporal-aware Gaussian splatting to produce high-fidelity reconstructions with minimal latency. Together, these components enable high-quality video generation and complete 3D reconstruction within 6 minutes, significantly outperforming prior methods in both efficiency and scalability, as shown in Figure 2.

2 Method

2.1 Overview

DriveGen3D, as illustrated in Figure 1, is an integrated 3D driving scene generation system composed of two key components: **FastDrive-DiT** for efficient long video generation, and **FastRecon3D** for feed-forward 3D scene reconstruction. The pipeline begins with FastDrive-DiT, which synthesizes high-resolution, temporally coherent driving videos under conditional guidance. These generated videos are then passed to FastRecon3D, which reconstructs dynamic 3D scenes in a feed-forward manner using temporal-aware Gaussian splatting. Together, these modules enable rapid and scalable 3D scene generation suitable for real-time simulation and autonomous driving applications.

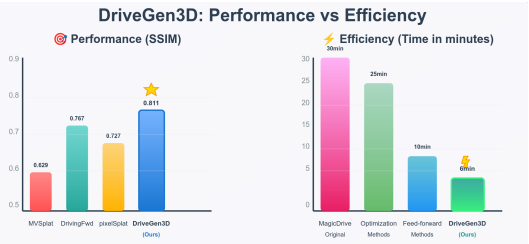

Figure 2: **Performance vs. Efficiency.** DriveGen3D achieves the highest SSIM (0.811) while significantly reducing generation time to 6 minutes—an 80% improvement over optimization-based and diffusion-based baselines—demonstrating superior video quality and real-time capability.

2.2 FastDrive-DiT

Generating 3D driving scenes, especially in the autonomous driving domain, is notoriously time-consuming due to the multi-view nature of the data. The video generation step is particularly costly because of the underlying diffusion process. For instance, MagicDriveDiT can take up to 30 minutes to produce a video of resolution $1600 \times 848 \times 6 \times 233$. To address this inefficiency, we propose **FastDrive-DiT**, an enhanced and lightweight video diffusion model built on MagicDriveDiT with targeted acceleration strategies.

Diffusion steps acceraleration. To accelerate video generation, we incorporate TeaCachee[21], a training-free caching method for diffusion models. TeaCache approximates changes in model outputs across timesteps, enabling efficient caching that achieves up to a 4.41× speedup over Open-Sora-Plan without sacrificing visual quality.

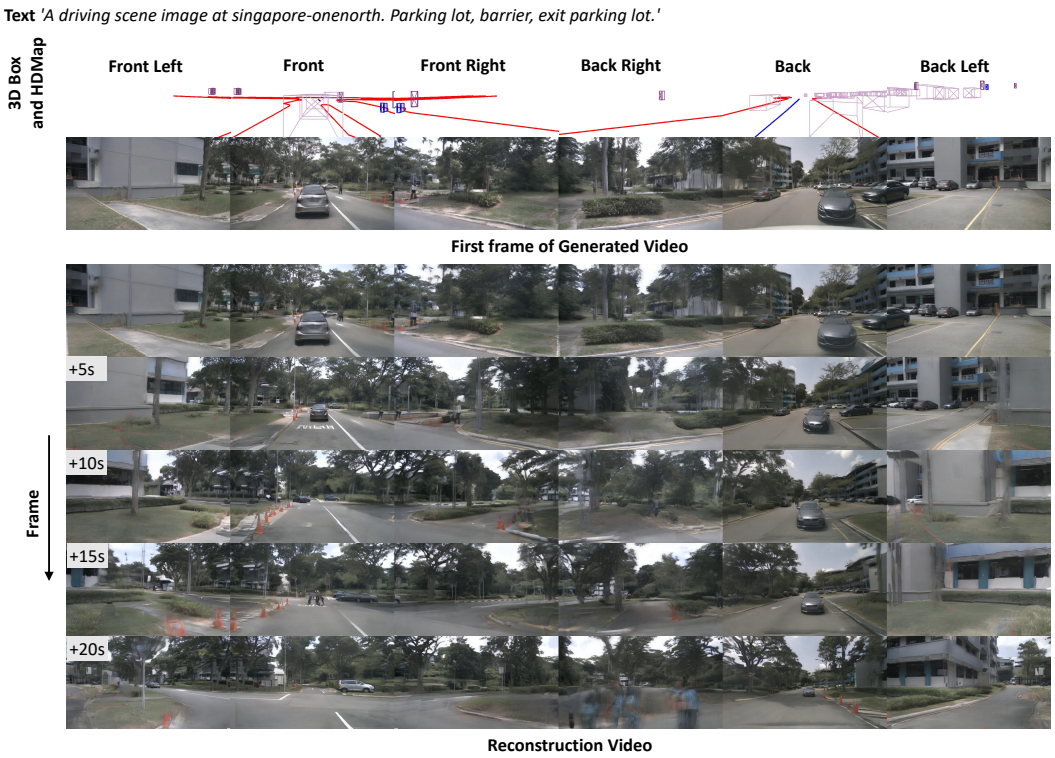

Figure 3: Visualization of multiview reconstruction video.

Our key optimization is to compute TeaCache's coefficients using only the conditional branch of the model, unlike the original implementation. This modification reduces computational overhead and further speeds up inference with no noticeable degradation in performance.

Quantized DiT. SageAttention[22] is an efficient and accurate quantization method designed to accelerate attention mechanisms in transformers, which are computationally intensive with O(N²) complexity. SageAttention2[23] further enhances efficiency by introducing INT4 quantization for Q and K matrices, FP8 for P and V, and precision-enhancing techniques like outlier smoothing and FP32 Matmul buffers. It achieves 3× and 5× higher OPS than FlashAttention2 on RTX4090, with negligible accuracy loss.

We carefully analyze the inference profiling and visualize $Q \times K \times V$ of different attention blocks, pointing out potential improvement techniques. In this paper, applying sageattention to the transformer blocks saves an additional 30 seconds of inference time with nearly no performance degradation.

2.3 FastRecon3D

While the aforementioned methods enable realistic driving scene generation, applications like simulation require complete 3D scene models. To enable rapid novel scene synthesis, we introduce **FastRecon3D**, a feed-forward reconstruction module that avoids costly optimization while preserving quality.

Per-Frame 3D Representation. Following DrivingForward [11], we propose a temporal-aware Gaussian Splatting formulation that reconstructs per-frame 3D Gaussian models while maintaining 3D consistency. In our approach, each time step is represented by a set of 3D Gaussian primitives: $G_i^t = \{\mathcal{G}_i^t\}_{i=1}^{N_t} = \{(\mu_i, \mathbf{\Sigma}_i, \alpha_i, \mathbf{c}_i)\}_{i=1}^{N_t}$

Recursively reconstruction from videos. For the frame generation of the $t+1$ timestep, we extract all frames from time steps $t-1$ to $t+1$ to reconstruct the scene at time t. Formally, for each timestep t, given multi-view images $\{I_i^t\}_{i=1}^{N_t}$ and temporal neighbors $\{I_i^{t\pm\Delta}\}_{i=1}^{N_t}$, our model predicts Gaussian parameters $\mathcal{G}^t = \{\mu^t, \Sigma^t, \alpha^t, c^t\}$ and save it as a 3D model: $\mathcal{G}^t = \mathcal{F}\Big(\{I_i^{t-\Delta}, I_i^t, I_i^{t+\Delta}\}_{i=1}^{N_t}\Big)$.

By leveraging both past and future context, this recursive reconstruction method effectively captures dynamic scene elements while maintaining high spatial fidelity. As a result, our approach produces

Type	Method	Venue	PSNR ↑	SSIM ↑	LPIPS ↓
Static	MVSplat	ECCV 2024	22.83	0.629	0.317
	pixelSplat	CVPR 2024	25.00	0.727	0.298
Dynamic	UniPad	CVPR 2024	16.45	0.375	-
	SelfOcc	CVPR 2024	18.22	0.464	-
	EmerNeRF	ICLR 2024	20.95	0.585	-
	DistillNeRF	NeurIPS 2024	20.78	0.590	-
	DrivingForward (640p)	AAAI 2025	21.67	0.727	0.259
	DrivingForward (228p)	AAAI 2025	21.76	0.767	0.194
	Ours (w/ GT images)	NextVid@NeurIPS 2025	**23.71**	0.733	**0.285**
	Ours (w/ GEN images)	NextVid@NeurIPS 2025	22.84	**0.811**	0.332

Table 1: Comparison of our method against prior feed-forward and optimization-based methods. The last two rows show novel view rendering performance with either GT or generated video input. All metrics are computed at frame t given $t-1$ and $t+1$ as inputs.

Total	spatial	temporal	cross-view	cross	other
615 s	104 s	82 s	163 s	62 s	204 s

Table 2: Time cost of different attention blocks of MagicDriveDiT during inference.

	FVD ↓	mAP ↑	mIoU↑	Time cost 17f/233f ↓
MagicDriveDiT	111.58	17.10	21.92	211s/615 s
Ours (w/o Quant)	125.70	16.60	21.27	64 s/309 s
Ours	125.88	16.72	21.24	**58 s/278 s**

Table 3: Acceralerating the inference process of MagicDriveDiT. 17f and 233f denote the frames count of generated videos.

complete 3D models in a matter of seconds, meeting the demanding requirements of simulation and other real-time applications without compromising on quality.

3 Experiments

3.1 Experimental Setup

Dataset. The training dataset is obtained from the nuScenes dataset [24]. It consists of 700 training videos and 150 validation videos. For the 3D scene reconstruction model, we split the dataset into 20,000 short sequences for training.

Evaluation Metrics. For video-generation stage, we evaluate both the realism and controllability in street-view video generation. We adhere to the benchmarks from [25]. To measure video quality, we use the Frechet Video Distance (FVD). Regarding controllability, we employ mAP from 3D object detection and mIoU from BEV segmentation. For 3D scene reconstruction stage, we adopt novel view synthesis (NVS) to assess reconstruction quality, following the evaluation protocol established in DrivingForward [11]. We report PSNR, SSIM, and LPIPS [26] in Table 1.

3.2 Main Results

As shown in Table 3, equiping MagicDriveDiT with TeaCache for *condition* branch has a speedup of nearly three times and two times for 17 frames and 233 frames. The percption metrics, mAP and mIoU only shows slight decrease. Table 2 shows that cross-view attention is identified as the most computationally costly process in MagicDriveDiT.

Table 1 provides a comprehensive comparison of DriveGen3D against both optimization-based dynamic methods and feed-forward reconstruction methods. Notably, when using generated images instead of ground truth, DriveGen3D maintains competitive reconstruction quality with a PSNR of 22.84 and achieves the highest SSIM of 0.811, demonstrating strong temporal coherence and structure preservation in generated scenes. This suggests that despite operating on synthetic inputs, DriveGen3D produces reliable and structurally consistent 3D reconstructions, validating the effectiveness of its end-to-end pipeline

4 Conclusion

We present DriveGen3D, an efficient framework for synthesizing high-resolution, long-duration driving videos from textual descriptions and BEV layouts, and generating high-quality large-scale dynamic scenes. Our approach marks a significant advancement in world modeling by bypassing conventional voxel-based generation paradigms [27] through a novel integration of longitudinal video generation and scene reconstruction modules. This architecture enables faithful reproduction of real-world driving scenarios and thus paves the way for novel applications in autonomous vehicle simulation and dynamic world modeling.

References

[1] Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. arXiv preprint arXiv:2310.02601, 2023.

[2] Ruiyuan Gao, Kai Chen, Bo Xiao, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrivedit: High-resolution long video generation for autonomous driving with adaptive control. arXiv preprint arXiv:2411.13807, 2024.

[3] Xiaofan Li, Yifu Zhang, and Xiaoqing Ye. Drivingdiffusion: Layout-guided multi-view driving scenarios video generation with latent diffusion model. In European Conference on Computer Vision, pages 469–485. Springer, 2024.

[4] 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.

[5] 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.

[6] 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 Advances in Neural Information Processing Systems, 2024.

[7] Zhuoran Yang, Xi Guo, Chenjing Ding, Chiyu Wang, and Wei Wu. Physical informed driving world model, 2024.

[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.

[9] Hongyu Zhou, Jiahao Shao, Lu Xu, Dongfeng Bai, Weichao Qiu, Bingbing Liu, Yue Wang, Andreas Geiger, and Yiyi Liao. Hugs: Holistic urban 3d scene understanding via gaussian splatting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 21336–21345, June 2024.

[10] 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 IEEE Conf. Comput. Vis. Pattern Recog., pages 21634–21643, 2024.

[11] Qijian Tian, Xin Tan, Yuan Xie, and Lizhuang Ma. Drivingforward: Feed-forward 3d gaussian splatting for driving scene reconstruction from flexible surround-view input. arXiv preprint arXiv:2409.12753, 2024.

[12] Hao Lu, Tianshuo Xu, Wenzhao Zheng, Yunpeng Zhang, Wei Zhan, Dalong Du, Masayoshi Tomizuka, Kurt Keutzer, and Yingcong Chen. Drivingrecon: Large 4d gaussian reconstruction model for autonomous driving. arXiv preprint arXiv:2412.09043, 2024.

[13] Xin Fei, Wenzhao Zheng, Yueqi Duan, Wei Zhan, Masayoshi Tomizuka, Kurt Keutzer, and Jiwen Lu. Driv3r: Learning dense 4d reconstruction for autonomous driving. arXiv preprint arXiv:2412.06777, 2024.

[14] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting, 2024.

[15] Junge Zhang, Qihang Zhang, Li Zhang, Ramana Rao Kompella, Gaowen Liu, and Bolei Zhou. Urban scene diffusion through semantic occupancy map, 2024.

[16] Chaojun Ni, Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Wenkang Qin, Guan Huang, Chen Liu, Yuyin Chen, Yida Wang, Xueyang Zhang, et al. Recondreamer: Crafting world models for driving scene reconstruction via online restoration. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 1559–1569, 2025.

[17] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph., 2023.

[18] David Charatan, Sizhe Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In IEEE Conf. Comput. Vis. Pattern Recog., 2024.

[19] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. arXiv preprint arXiv:2403.14627, 2024.

[20] Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth. arXiv preprint arXiv:2410.13862, 2024.

[21] Feng Liu, Shiwei Zhang, Xiaofeng Wang, Yujie Wei, Haonan Qiu, Yuzhong Zhao, Yingya Zhang, Qixiang Ye, and Fang Wan. Timestep embedding tells: It's time to cache for video diffusion model. arXiv preprint arXiv:2411.19108, 2024.

[22] Jintao Zhang, Jia Wei, Pengle Zhang, Jun Zhu, and Jianfei Chen. Sageattention: Accurate 8-bit attention for plug-and-play inference acceleration. In Int. Conf. Learn. Represent., 2025.

[23] Jintao Zhang, Haofeng Huang, Pengle Zhang, Jia Wei, Jun Zhu, and Jianfei Chen. Sageattention2: Efficient attention with thorough outlier smoothing and per-thread int4 quantization, 2024.

[24] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In IEEE Conf. Comput. Vis. Pattern Recog., June 2020.

[25] Zhiying Du and Zhen Xing. Challenge report: Track 2 of multimodal perception and comprehension of corner cases in autonomous driving. In ECCV 2024 Workshop on Multimodal Perception and Comprehension of Corner Cases in Autonomous Driving, 2024.

[26] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 586–595, 2018.

[27] Yifan Lu, Xuanchi Ren, Jiawei Yang, Tianchang Shen, Zhangjie Wu, Jun Gao, Yue Wang, Siheng Chen, Mike Chen, Sanja Fidler, et al. Infinicube: Unbounded and controllable dynamic 3d driving scene generation with world-guided video models. arXiv preprint arXiv:2412.03934, 2024.

[28] Bohan Li, Jiazhe Guo, Hongsi Liu, Yingshuang Zou, Yikang Ding, Xiwu Chen, Hu Zhu, Feiyang Tan, Chi Zhang, Tiancai Wang, et al. Uniscene: Unified occupancy-centric driving scene generation. arXiv preprint arXiv:2412.05435, 2024.

[29] Yuqing Wen, Yucheng Zhao, Yingfei Liu, Fan Jia, Yanhui Wang, Chong Luo, Chi Zhang, Tiancai Wang, Xiaoyan Sun, and Xiangyu Zhang. Panacea: Panoramic and controllable video generation for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6902–6912, 2024.

[30] 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 IEEE Conf. Comput. Vis. Pattern Recog., 2024.

[31] 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 IEEE Conf. Comput. Vis. Pattern Recog., 2022.

[32] Ziyang Xie, Junge Zhang, Wenye Li, Feihu Zhang, and Li Zhang. S-nerf: Neural radiance fields for street views, 2023.

[33] 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 CAAI International Conference on Artificial Intelligence, pages 3–15. Springer, 2023.

[34] Qiang Zhang, Seung-Hwan Baek, Szymon Rusinkiewicz, and Felix Heide. Differentiable point-based radiance fields for efficient view synthesis, 2023.

[35] Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting, 2024.

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

[37] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. arXiv preprint arXiv:2408.16061, 2024.

[38] Weijie Wang, Donny Y Chen, Zeyu Zhang, Duochao Shi, Akide Liu, and Bohan Zhuang. Zpressor: Bottleneck-aware compression for scalable feed-forward 3dgs. arXiv preprint arXiv:2505.23734, 2025.

[39] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In IEEE Conf. Comput. Vis. Pattern Recog., pages 5294–5306, 2025.

[40] Chen Ziwen, Hao Tan, Kai Zhang, Sai Bi, Fujun Luan, Yicong Hong, Li Fuxin, and Zexiang Xu. Long-lrm: Long-sequence large reconstruction model for wide-coverage gaussian splats. arXiv preprint arXiv:2410.12781, 2024.

[41] Chuanrui Zhang, Yingshuang Zou, Zhuoling Li, Minmin Yi, and Haoqian Wang. Transplat: Generalizable 3d gaussian splatting from sparse multi-view images with transformers. In AAAI Conf. Artif. Intell., volume 39, pages 9869–9877, 2025.

[42] Yunsong Wang, Tianxin Huang, Hanlin Chen, and Gim Hee Lee. Freesplat: Generalizable 3d gaussian splatting towards free view synthesis of indoor scenes. Adv. Neural Inform. Process. Syst., 37:107326–107349, 2024.

[43] Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. arXiv preprint arXiv:2410.24207, 2024.

[44] Duochao Shi, Weijie Wang, Donny Y. Chen, Zeyu Zhang, Jiawang Bian, Bohan Zhuang, and Chunhua Shen. Revisiting depth representations for feed-forward 3d gaussian splatting. arXiv preprint arXiv:2506.05327, 2025.

[45] Weijie Wang, Yeqing Chen, Zeyu Zhang, Hengyu Liu, Haoxiao Wang, Zhiyuan Feng, Wenkang Qin, Zheng Zhu, Donny Y. Chen, and Bohan Zhuang. Volsplat: Rethinking feed-forward 3d gaussian splatting with voxel-aligned prediction. arXiv preprint arXiv:2509.19297, 2025.

[46] Chaojun Ni, Xiaofeng Wang, Zheng Zhu, Weijie Wang, Haoyun Li, Guosheng Zhao, Jie Li, Wenkang Qin, Guan Huang, and Wenjun Mei. Wonderturbo: Generating interactive 3d world in 0.72 seconds. arXiv preprint arXiv:2504.02261, 2025.

[47] Xuanchi Ren, Yifan Lu, Hanxue Liang, Jay Zhangjie Wu, Huan Ling, Mike Chen, Francis Fidler, Sanja annd Williams, and Jiahui Huang. Scube: Instant large-scale scene reconstruction using voxsplats. In Adv. Neural Inform. Process. Syst., 2024.

A Implementation Details

We train the model with a resolution of 424×800. We inference the model on NVIDIA H20 GPUs. When assessing the time cost of our proposed method, the baselines for the video generation model are MagicDriveDiT (17f) and MagicDriveDiT (233f). The 3D scene reconstruction stage is trained on NVIDIA H20 GPUs for 2 days.

B More Experiments

Diffusion steps acceraleration. Firstly, we visualize the input differences and output differences in consecutive timesteps in Figure B. It is observed that the default configurations of TeaCache, i.e. *ALL*, exhibits a U-shape for model output, with downward tend initially, nearly constant for the middle and upward until the end. The same is for the *uncondition* branch. The *condition* branch exhibits a slightly different phenomenon, with the start of model output much smaller. We also apply simple polynomial fitting to fit a relationship between model input and output and the use these cofficients to predict model output according to the input. As shown in the black lines in Figure B, the quality of fitting is best for the *condition* branch, while *ALL* and *uncondition* do not fit well for the start. We attribute this to the different input and output relationships in the start of diffusion process. So we finally only apply TeaCache to the *condition* branch of MagicDriveDiT.

Quantized DiT. Secondly, we profile the time cost of MagicDriveDiT. As shown in Table 2, cross-view attention is identified as the most computationally costly process in MagicDriveDiT. We further analyze different attention components in MagicDriveDiT and plot the distribution of Q, K, V, as depicted in Figure A. Notably, Q and K exhibit a trend similar to that in Figure 4 of SageAttention2. An interesting phenomenon is observed: the numeric range of V follows the order: spatial $>$ temporal $>$ cross-view. Specifically, the cross-view range is 10 times smaller than the spatial range and 5 times smaller than the temporal range. Given the principle of quantization, a smaller range is more conducive to quantization. Therefore, a promising solution is not only to apply SageAttention to cross-view attention but also substitute the high-precision quantization method for V with a lower-resolution one. Concretely, for V, the FP8, E4M3 data type can be replaced with FP8, E5M2 data type and INT8. We leave the latter for future attempt. In this paper, applying sageattention to the transformer blocks saves an additional 30 seconds (233 frames) of inference time after diffusion steps acceraleration while maintaining performance.

Visualization. In Figure C, we compare the generated videos of baseline model, baseline model with diffusion steps acceraleration and further quantization with SageAttention. With the proposed techniques, no obvious change is observed. It can still generate video longer than 15 s with a much faster speed. This highlights the efficiency of *DriveGen3D*.

Full pipeline results. In Figure D, we show a typical output result of *DriveGen3D*. *DriveGen3D* can generate and reconstruct videos for more than 20S, 12FPS.

C Related Work

C.1 Video generation for driving scene

Recent advancements in street view generation and autonomous driving scene synthesis have significantly improved the fidelity and controllability of synthetic data. MagicDrive [1] introduces a framework for street view generation with diverse 3D geometry controls, such as camera poses and 3D bounding boxes, enhancing tasks like BEV segmentation and 3D object detection through cross-view consistency. MagicDriveDiT [2] extends this by addressing high-resolution, long video generation for autonomous driving, leveraging flow matching and spatial-temporal conditional encoding to achieve superior scalability and control. InfiniCube [27] focuses on unbounded dynamic 3D driving scene generation, combining scalable 3D representations with video models to ensure geometric and appearance consistency across large-scale scenes. DrivingDiffusion [3] synthesizes multi-view driving videos with precise layout control, ensuring cross-view and cross-frame consistency through a spatial-temporal diffusion framework. UniScene [28] unifies the generation of semantic occupancy, video, and LiDAR data, employing a progressive generation process to reduce complexity and improve downstream task performance. DriveDreamer [4] pioneers real-world-driven world models,

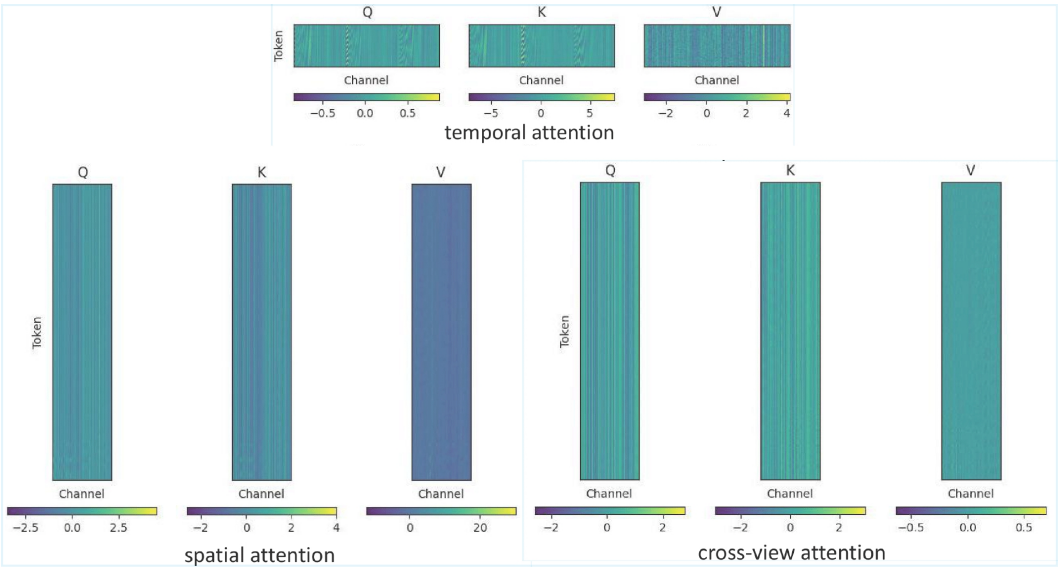

Figure A: Typical examples of tensors' data distribution in different attention blocks of MagicDriveDiT.

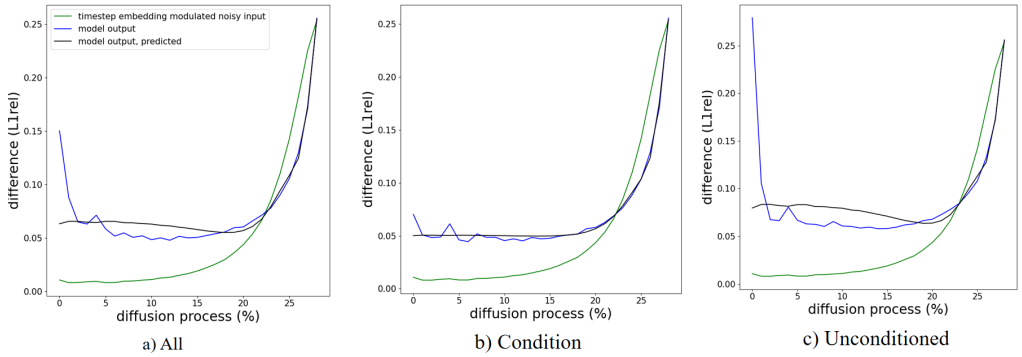

Figure B: Visualization of input and output differences across consecutive timesteps. We separately plot the *all*, *conditioned*, and *unconditioned* components.

using diffusion models to capture complex driving environments and enhance driving video generation and action prediction. Panacea [29] integrates a novel 3D attention and a two-stage generation pipeline to maintain coherence, supplemented by the ControlNet framework for meticulous control by the Bird's-Eye-View (BEV) layouts. DriveDreamer-2 [5] integrates LLMs to generate user-defined driving videos, improving temporal and spatial coherence while surpassing state-of-the-art methods in video quality metrics like FID and FVD. Together, these works advance the field of autonomous driving by providing scalable, controllable, and high-fidelity synthetic data generation frameworks.

C.2 Reconstruction for driving scene

The reconstruction of dynamic driving scenes [8, 17, 30, 31, 3, 32–36] has emerged as a critical task in autonomous systems and immersive environment modeling. Contemporary approaches predominantly leverage Gaussian splatting-based representations due to their inherent balance between rendering efficiency and geometric expressiveness. Early methodologies in this domain adopted optimization-based paradigms, exemplified by works such as StreetGaussian [8], DrivingGaussian [10], and HUGS [9]. These frameworks optimize scene representations per instance for specific street segments (typically under 100 meters in scale) through iterative refinement processes spanning approximately 30 minutes.

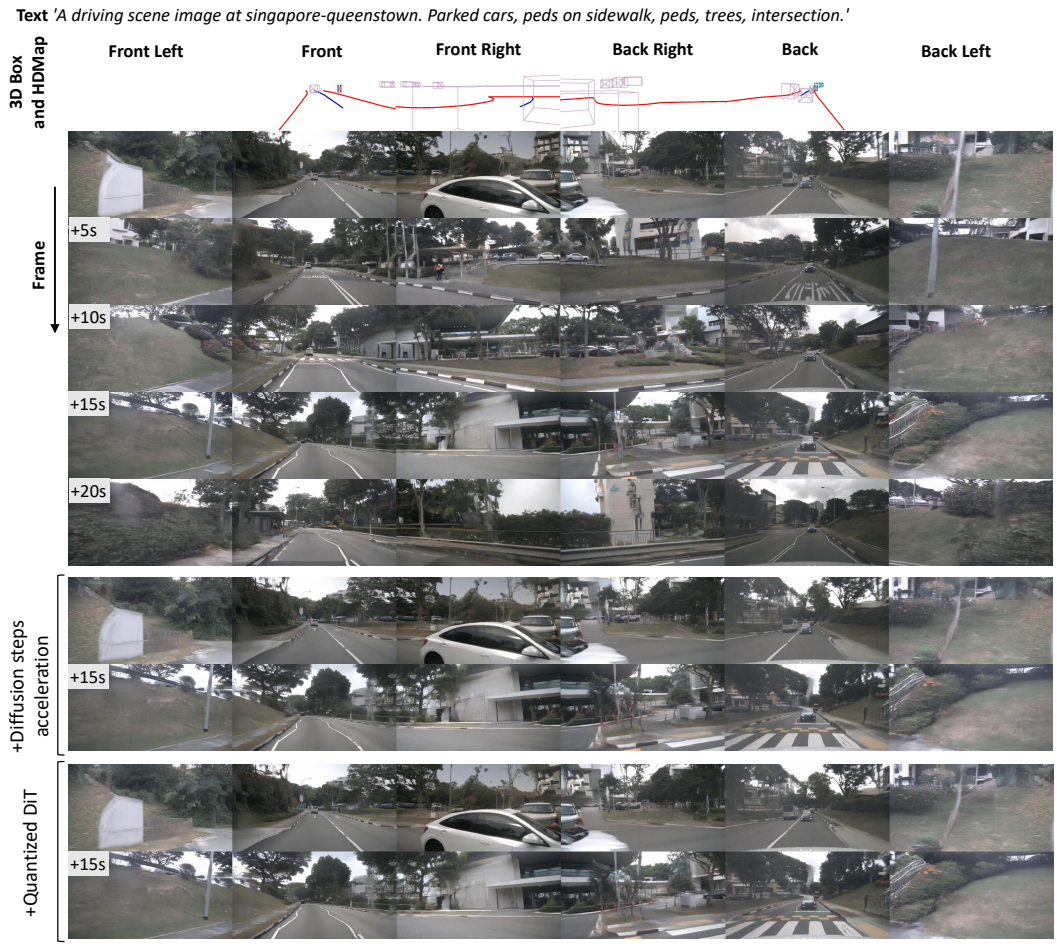

Figure C: Comparison of video generation for MagicDriveDiT, Diffusion steps acceleration and Quantized DiT.

Recent advancements have shifted toward feed-forward architectures [18–20, 37–46] to enable rapid 3D reconstruction. Methods like PixelSplat [18], MVSplat [19], and DepthSplat [20] employ large pretrained networks to directly infer Gaussian parameters from multi-view inputs, reducing reconstruction time from minutes to seconds. Though these approaches demonstrate generalizability across scenes, they often sacrifice reconstruction fidelity in geometrically complex regions or under sparse observational constraints. Parallel innovations address the temporal dimension of driving scenes: InfiniCube [27] extends the Scube [47] framework to 3D street generation, disentangling dynamic vehicles from static backgrounds via hybrid voxel-video control mechanisms. Drive3R [13] adapts the Spann3R [37] architecture for per-frame 3D scene reconstruction through temporal consistency priors. DrivingRecon [12] mimics StreetGaussians' pipeline but replaces iterative optimization with feed-forward prediction, achieving real-time capability at moderate resolutions. DrivingForward [11] enhances sparse-view reconstruction robustness by jointly learning pose estimation and depth prediction modules within its network architecture. These advancements collectively highlight two persistent limitations: 1) Existing feed-forward 3D reconstruction methods operate at constrained spatial resolutions (typically $\leq 512 \times 512$). 2) Prior works on generative models for conditional scene synthesis required substantial computational resources and complex framework integration, which hindered their widespread adoption.

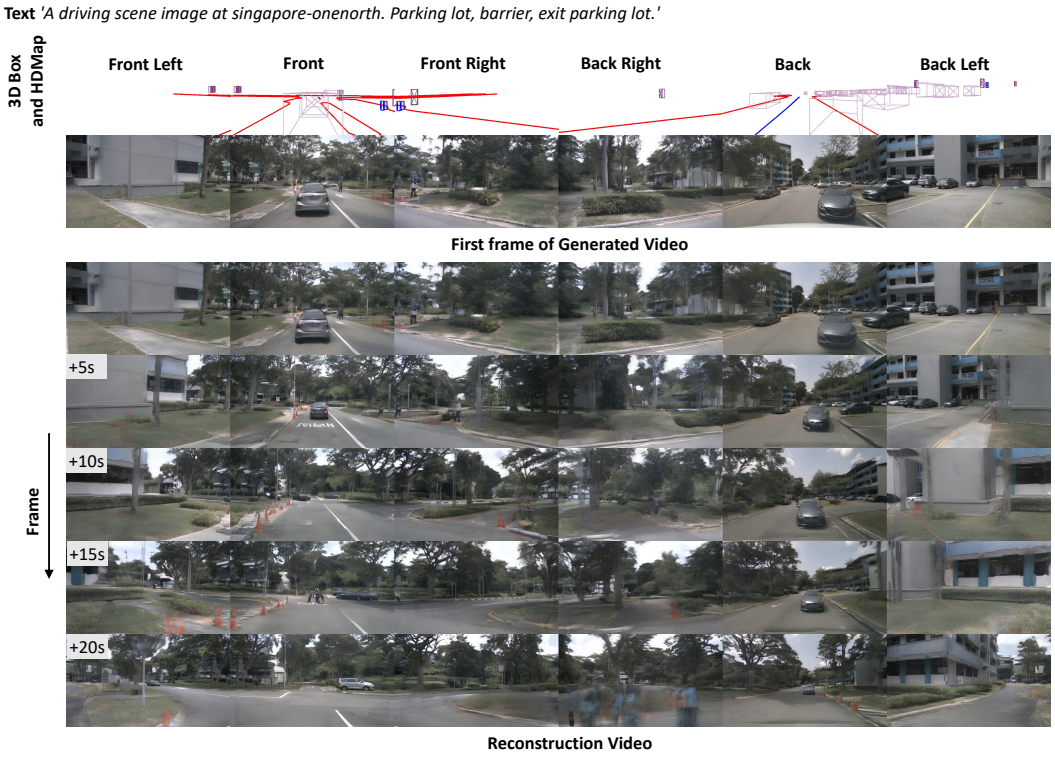

Figure D: Visualization of multiview reconstruction video.