

Supplementary Material

A Supplementary visualization

As shown in Fig. 1, our method demonstrates the capability to generate consistent multi-view video sequences under diverse lighting conditions, such as daytime and nighttime scenarios. Despite the temporal shift, the generated scenes maintain fidelity to the underlying map and object layout, demonstrating effective adaptation of scene appearance according to the time-of-day condition. This validates our approach’s robustness in handling heterogeneous environmental configurations.

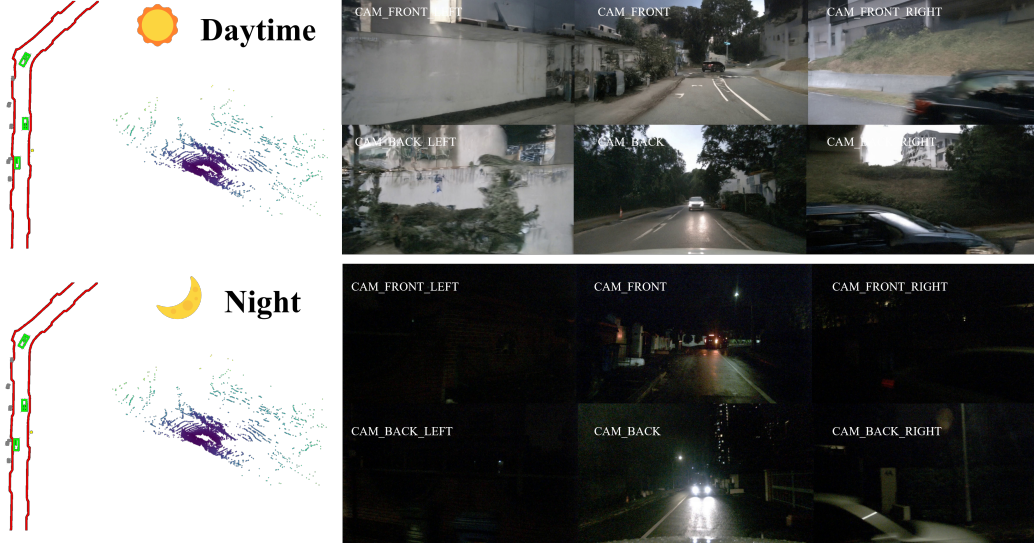


Figure 1: **Controllable generation across time-of-day.** By altering scene-level conditions, our method produces consistent multi-view videos aligned with the same underlying map and object layout, while adapting appearance to represent daytime and nighttime settings

As shown in Fig. 2, our method (second column) excels in preserving accurate layouts, object shapes, and background integrity when compared with MagicDrive (third column) and Panacea (fourth column). Notably, MagicDrive exhibits vehicle distortion and structural anomalies, while Panacea suffers from hallucinated textures and geometric misalignment. These comparisons highlight our model’s superiority in generating realistic and coherent video sequences.



Figure 2: **Qualitative comparison of video generation quality.** Our method (second column) preserves accurate layout, object shapes, and background integrity. MagicDrive (third column) shows vehicle distortion and broken structures. Panacea (fourth column) often suffers from hallucinated textures and geometric misalignment.

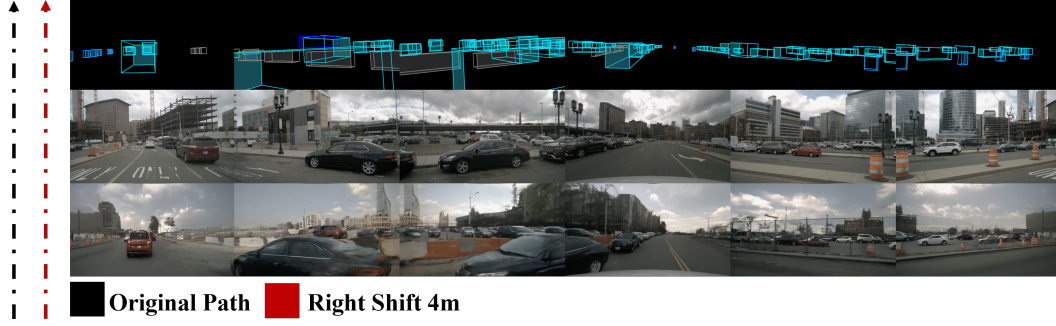


Figure 3: **Trajectory-conditioned novel view synthesis.** Given a ground-truth trajectory (middle), we modify the layout (top) by shifting the ego path 4 meters right (bottom). Our model generates plausible and consistent scenes across all views under these layout changes.

As shown in Fig. 3, given an original trajectory, we altered the ego path by shifting it 4 meters to the right. The results show that our method can generate plausible and consistent scenes across all views under layout changes, indicating its strong generalization ability to adapt to dynamic environments without compromising the realism or consistency of the generated content.

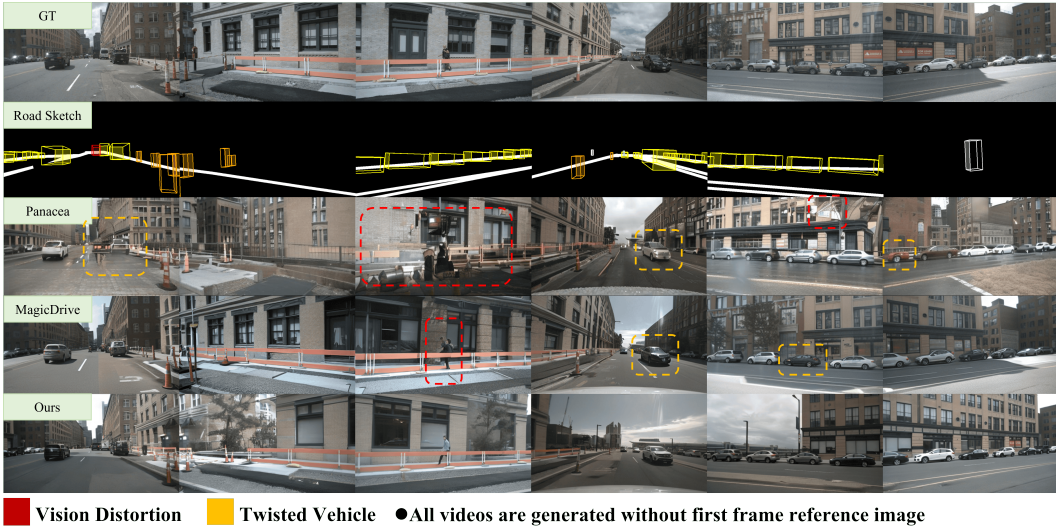


Figure 4: **Qualitative comparison of video generation.** From top to bottom: (1) Ground-truth images, (2) Road sketch input, (3) Panacea [3], (4) MagicDrive [1], (5) Ours. Panacea suffers from hallucinated textures and geometric misalignment. MagicDrive shows vehicle distortion and broken structures. In contrast, **ours** preserves accurate layout, object shapes, and background integrity.

Our method produces high-quality video samples that exhibit strong geometric fidelity and visual coherence, even in complex and dynamic scenes. As illustrated in Fig. 4, our model accurately preserves the shapes of vehicles, the structure of lanes, and the textures of the surrounding environment. In comparison, MagicDrive introduces noticeable object deformations and layout inconsistencies, while Panacea generates hallucinated content and suffers from distorted backgrounds.



Figure 5: **Joint generation of LiDAR and multi-view video.** Our method generates spatially aligned LiDAR and camera views conditioned on a shared straight street layout.



Figure 6: **Joint generation of LiDAR and multi-view video.** Our method generates spatially aligned LiDAR and camera views conditioned on a Busy junction layout.



Figure 7: **Joint generation of LiDAR and multi-view video.** Our method generates spatially aligned LiDAR and camera views conditioned on a turning layout.

22 Figs. 5, 6 and 7 demonstrate our method’s capability to generate spatially aligned LiDAR and
 23 camera views under different urban layouts. As shown in Fig. 5, focusing on a straight street
 24 layout, our method demonstrates precise alignment between LiDAR point clouds and camera views,
 25 ensuring accurate representation of scene geometry and appearance. This confirms its effectiveness
 26 in straightforward driving environments. As shown in Fig. 6, our method is tested on a busy junction
 27 layout with increased complexity, including multiple vehicles and dynamic interactions. Notably,
 28 it maintains exceptional spatial alignment and contextual fidelity, underscoring its robustness in
 29 navigating intricate, real-world urban environments. Fig. 7 demonstrates our method’s capability to
 30 generate spatially aligned LiDAR and camera views under a turning scenario. The results show that
 31 our method maintains exceptional spatial alignment and contextual fidelity, even in more complex
 32 driving situations involving turns.

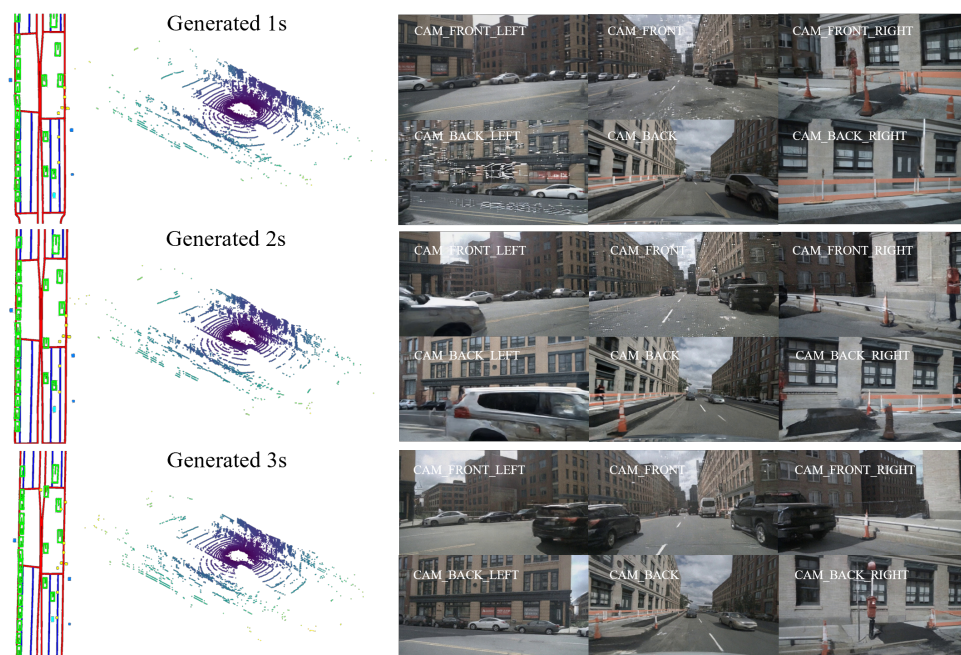


Figure 8: Long-term multi-view video generation over 3 seconds in an urban driving scene, conditioned on the straight street layout.

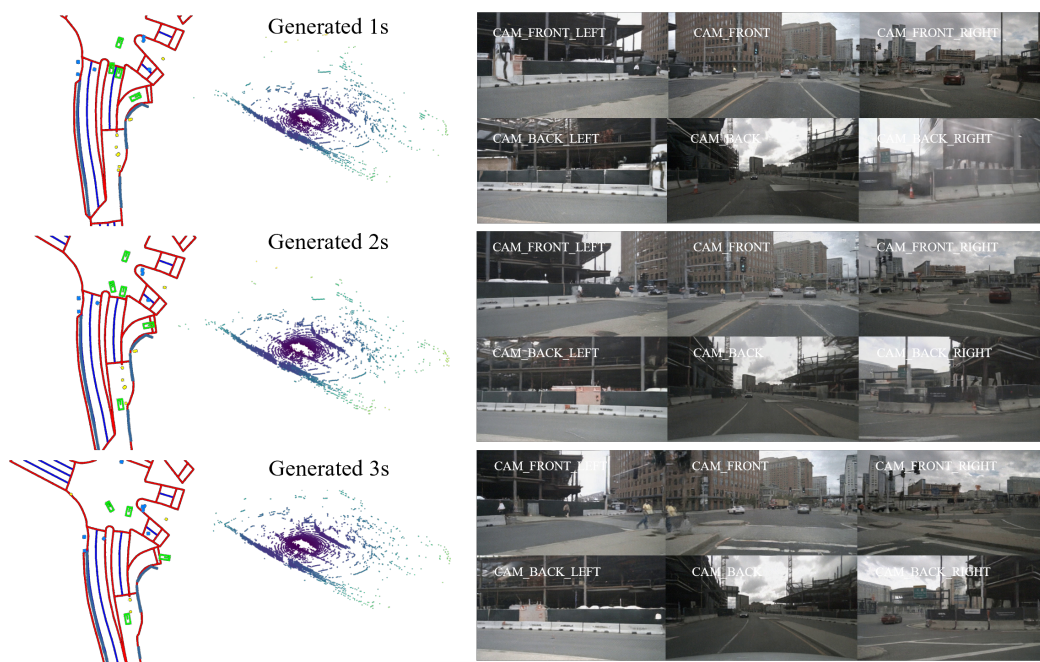


Figure 9: Long-term multi-view video generation over 3 seconds in an urban driving scene, conditioned on the busy junction layout.

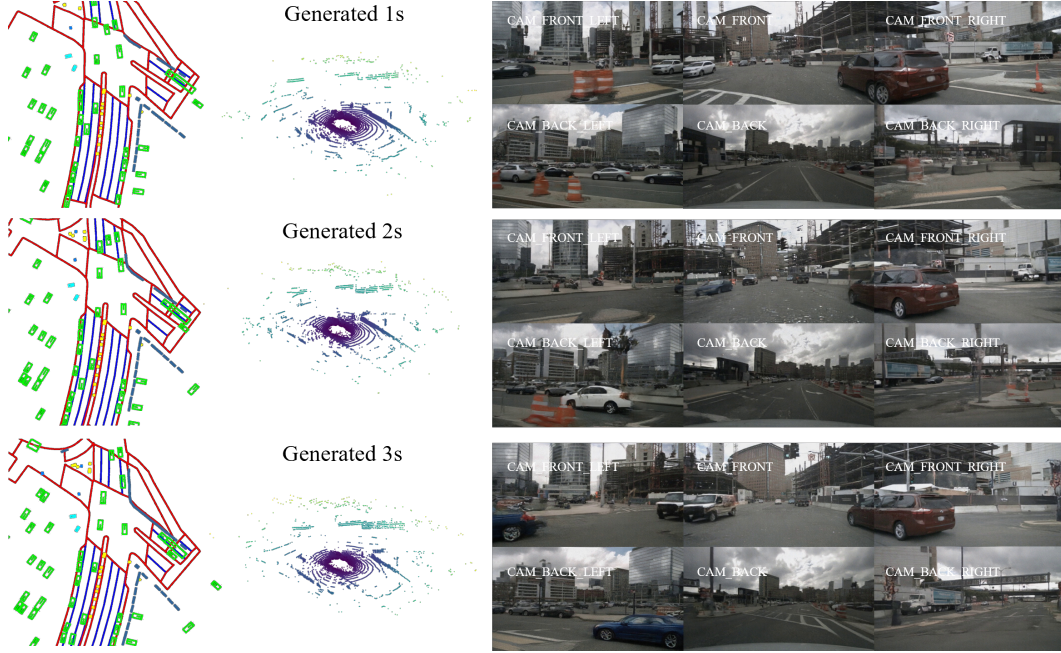


Figure 10: Long-term multi-view video generation over 3 seconds in an urban driving scene, conditioned on the complex crossroad layout.

Figs. 8, 9, and 10 demonstrate our method’s capability for long-term consistent multi-view video generation over three seconds in diverse urban driving environments. In Fig. 8, sequences are generated under a straight street layout, with synthesized frames exhibiting strong spatial-temporal coherence and high visual fidelity, confirming our method’s reliability in stable scenarios. Fig. 9 evaluates performance in a busy junction setting with dynamic traffic interactions. Despite the increased complexity, the generated sequences maintain both temporal consistency and semantic accuracy, capturing realistic vehicle behaviors and scene dynamics—demonstrating the model’s generalization to complex urban environments. Finally, Fig. 10 presents results from a challenging crossroad scenario. The generated multi-view sequences preserve spatial-temporal coherence and realism, accurately reflecting dynamic interactions among multiple vehicles.

B Additional Downstream Evalutaion

Table 1: Effect of Multimodal Data Generation on 3D Object Detection (BEVFusion)

Method	Input	mAP↑	NDS↑
Baseline [2]	C&L	66.87	69.65
Ours(+cam_gen)	C&L	67.09 (+0.22)	70.12 (+0.47)
Ours(+lidar_gen)	C&L	67.69(+0.82)	70.58 (+0.93)
Ours(+cam_gen&lidar_gen)	C&L	67.78(+0.91)	71.13(+1.48)

As shown in Tab.1, we evaluate the effectiveness of our generative data on the BEVFusion[2] framework for 3D object detection. Our approach yields consistent improvements across all settings, increasing the mAP from 66.87 to 67.78 and NDS from 69.65 to 71.13. Notably, joint generation of both camera and LiDAR modalities achieves the highest gains (+0.91 mAP / +1.48 NDS), demonstrating the complementary benefits of multimodal generation. These results validate the utility of high-quality synthetic data in enhancing downstream perception tasks, especially in data-scarce or long-tail scenarios.

51 **References**

- 52 [1] Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang
53 Xu. Magicdrive: Street view generation with diverse 3d geometry control. *arXiv preprint*
54 *arXiv:2310.02601*, 2023.
- 55 [2] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao
56 Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework.
57 *Advances in Neural Information Processing Systems*, 35:10421–10434, 2022.
- 58 [3] Yuqing Wen, Yucheng Zhao, Yingfei Liu, Fan Jia, Yanhui Wang, Chong Luo, Chi Zhang, Tiancai
59 Wang, Xiaoyan Sun, and Xiangyu Zhang. Panacea: Panoramic and controllable video generation
60 for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and*
61 *Pattern Recognition*, pages 6902–6912, 2024.