# Challenge report: Track 2 of Multimodal Perception and Comprehension of Corner Cases in Autonomous Driving

Zhiying Du, Zhen Xing
Shanghai Key Lab of Intell. Info. Processing, School of CS, Fudan University

## Abstract

*This paper presents our submission to Track 2 of the Multimodal Perception and Comprehension of Corner Cases in Autonomous Driving. While the field of autonomous driving has garnered significant interest, the collection and annotation of street scenes remain prohibitively expensive. In this work, we explore previous methods of street scene generation using diffusion models and propose our approach. By combining pre-trained image and motion layers, we achieve high-quality results with minimal training. To generate videos of arbitrary length with smooth transitions, we employ a sliding window technique to mitigate discrepancies between segments. The source code is available at* https://github.com/ZhiyingDu/ECCV-2024-Workshop-on-Multimodal-Perception-and-Comprehension-of-Corner-Cases-in-Autonomous-Driving

## 1. Introduction

In the field of autonomous driving, the demand for high-quality, annotated multi-view image and video datasets is paramount for advancing perception tasks like 3D detection [7, 13], map segmentation [6, 8], and lane detection [1, 5]. However, the significant challenges and costs associated with collecting and annotating such diverse datasets have become a bottleneck in training robust models. Recent advances in generative models[2, 9, 17], particularly diffusion models [12, 15, 16, 18], offer a promising solution by enabling the synthesis of multi-view images and videos, which can supplement or even replace real data, potentially accelerating the development of more effective autonomous driving systems.

Several methods have been proposed for generating street scenes under specific conditions, each aiming to enhance realism and consistency across views. For example, BEVGen [11] utilizes a novel cross-view conversion and spatial attention mechanism to learn relationships between camera and map views, synthesizing spatially consistent surround images that align with the BEV layout of traffic scenes. However, this approach relies on 2D bounding boxes, leading to a potential loss of height information. MagicDrive [3] addresses this limitation by incorporating a range of 3D geometric controls, such as camera poses, roadmaps, and 3D bounding boxes, alongside text descriptions via advanced encoding strategies. It also features a cross-view attention module to ensure consistency across multiple camera views, resulting in high-fidelity street scene synthesis. Similarly, Panacea [14], a more recent approach, introduces a UNet architecture with 4D attention and integrates various control signals—including images, text prompts, and the ControlNet [20] module to inject BEV sequences, enabling precise control of elements like bounding boxes, object depth, roadmaps, and camera poses for high-quality, multi-view, and panoramic video generation.

While both methods support video generation and modeling temporal information by training a new temporal module in a second stage, they encounter challenges due to mismatches in data distribution (mean and standard deviation) between the newly initialized parameters of the temporal module and the pre-trained ones. This mismatch hinders the learning process, as the model struggles to receive accurate, instance-specific gradients, negatively impacting convergence and training efficiency [10].

To address these challenges, the most straightforward approach is to introduce a pre-trained model to handle temporal information. AnimateDiff [4], for instance, facilitates the generation of animated images across various personalized T2I models, reducing the need for model-specific tuning while maintaining strong content consistency over time. Consequently, in this paper, we adopt MagicDrive [3] as our baseline and incorporate AnimateDiff to effectively model temporal information.

## 2. Method

This section outlines the method we ultimately employed in the competition.

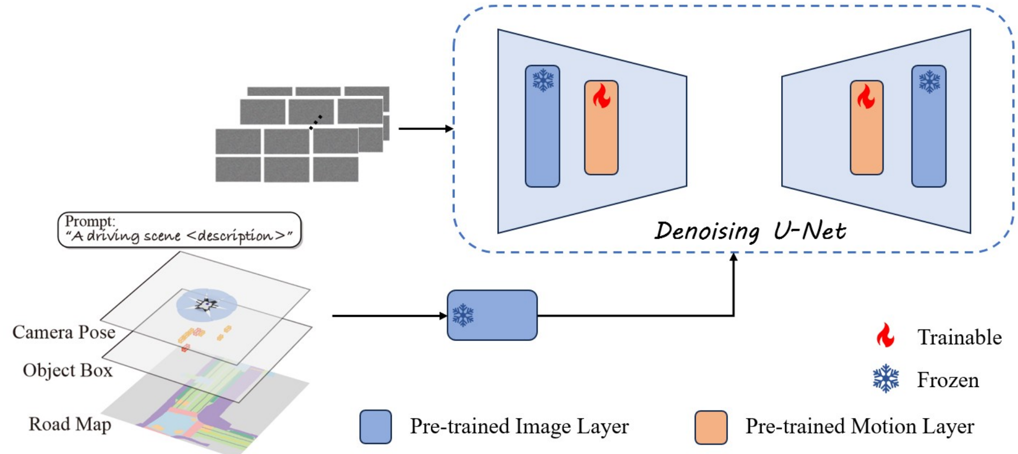


Figure 1. Overview of our method for street-view video generation. This method leverages pre-trained motion layer to model the frame consistency.

### 2.1. Temporal Consistency Modeling

To ensure temporal consistency across video frames, we extend the image diffusion model to the video domain by integrating pre-trained motion layer. As illustrated in

Figure 1, we integrated the motion module of AnimateDiff into the image model of MagicDrive to effectively model temporal information. We input $\boldsymbol{z}_{\boldsymbol{t}}^{1:K}$ to this video diffusion model by reshaping the input features from $\mathbb{R}^{B\times F\times N\times C\times H\times W}$ into $\mathbb{R}^{(BFN)\times C\times H\times W}$. Within temporal modules, we reshape the features into $\mathbb{R}^{(BNHW)\times F\times C}$ to compute cross-frame information along the temporal dimension.

### 2.2. Long video generation

By leveraging temporal consistency modeling and pre-trained image model, we can generate temporally consistent videos of arbitrary length through segment-by-segment processing. However, due to the limitations of the temporal attention block in capturing long-range consistency between segments, unnatural transitions and inconsistent details may still occur. To address this challenge, we applied the sliding window technique from MagicAnimate [19] during the inference stage. We divide the long condition sequence into multiple segments with temporal overlap, where each segment has a length of $K$. First, we sample noise $\boldsymbol{z}^{1:F}$ for the entire video with $F$ frames and partition it into overlapping noise segments $\{\boldsymbol{z}^{1:K}, \boldsymbol{z}^{K-s+1:2K-s}, ..., \boldsymbol{z}^{n(K-s)+1:n(K-s)+K}\}$, where $n = \lceil (F-K)/(K-s) \rceil$ and $s$ is the overlap stride, with $s < K$. If $(F-K) \mod (K-s) \neq 0$, meaning the last segment is shorter than $K$, we pad it with the first few frames to form a full $K$ frame segment. Additionally, sharing the same initial noise $\boldsymbol{z}^{1:K}$ across all segments improves video quality. At each denoising timestep $\boldsymbol{t}$, we predict noise and obtain $\epsilon_\theta^{1:K}$ for each segment, then merge them into $\epsilon_\theta^{1:F}$ by averaging the overlapping frames. When $\boldsymbol{t} = 0$, we generate the final video $I^{1:F}$.

## 3. Experiments

In this section, we mainly introduce the implementation details and experimental results.

### 3.1. Implementation Details

Apart from adjusting the learning rate, we used the default configuration in MagicDrive. We observed that a lower learning rate facilitated easier model convergence. We finalized the learning rate at 1e-5.

### 3.2. Experimental Results

Preliminary experiments were made to explore the validity of the method. A subset of the results are shown by figure 2 and table 1. As you can see, table 1 reveals an interesting

Table 1. Quantitative comparison of our method and MagicDrive. The top results are highlighted in **black**

| Metrics | FVD ↓ | mAP ↑ | mIoU ↑ |
|---|---|---|---|
| MagicDrive | **218.1200** | 11.8617 | 18.3429 |
| Ours | 232.5072 | **12.7845** | **19.4639** |

observation: our pre-trained motion layer performs worse on FVD compared to fine-tuning a new temporal module. Upon closer investigation, we found that our model is still in the process of converging, with the FVD values of the latest checkpoints gradually decreasing. We believe that with sufficient training steps, our method will eventually outperform the fine-tuned a new temporal module.


Figure 2. Visual results of out method.

## 4. Conclusions

Our primary contribution is the introduction of a pre-trained motion layer into the model, enabling the generation of videos with strong temporal consistency by training only this layer. Additionally, our experiments revealed that using a smaller learning rate can improve convergence in street scene generation tasks. However, our method faces some challenges, such as the separation of perspective consistency and frame consistency due to the two-stage training process. Future work could focus on integrating these two aspects to achieve more cohesive video generation.

## References


[1] Li Chen, Chonghao Sima, Yang Li, Zehan Zheng, Jiajie Xu, Xiangwei Geng, Hongyang Li, Conghui He, Jianping Shi, Yu Qiao, et al. Persformer: 3d lane detection via perspective transformer and the openlane benchmark. In *European Conference on Computer Vision*, pages 550–567. Springer, 2022. 1

[2] Qijun Feng, Zhen Xing, Zuxuan Wu, and Yu-Gang Jiang. Fdgaussian: Fast gaussian splatting from single image via geometric-aware diffusion model. *arXiv preprint arXiv:2403.10242*, 2024. 1

[3] 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. 1

[4] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 1

[5] Shaofei Huang, Zhenwei Shen, Zehao Huang, Zi-han Ding, Jiao Dai, Jizhong Han, Naiyan Wang, and Si Liu. Anchor3dlane: Learning to regress 3d anchors for monocular 3d lane detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17451–17460, 2023. 1

[6] Yanqin Jiang, Li Zhang, Zhenwei Miao, Xiatian Zhu, Jin Gao, Weiming Hu, and Yu-Gang Jiang. Polarformer: Multi-camera 3d object detection with polar transformer. In *Proceedings of the AAAI conference on Artificial Intelligence*, pages 1042–1050, 2023. 1

[7] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022. 1

[8] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, and Xiangyu Zhang. Petrv2: A unified framework for 3d perception from multi-camera images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3262–3272, 2023. 1

[9] Tianyi Lu, Xing Zhang, Jiaxi Gu, Hang Xu, Renjing Pei, Songcen Xu, and Zuxuan Wu. Fuse your latents: Video editing with multi-source latent diffusion models. *arXiv preprint arXiv:2310.16400*, 2023. 1

[10] Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*, 2024. 1

[11] Alexander Swerdlow, Runsheng Xu, and Bolei Zhou. Street-view image generation from a bird's-eye view layout. *IEEE Robotics and Automation Letters*, 2024. 1

[12] Shuyuan Tu, Qi Dai, Zihao Zhang, Sicheng Xie, Zhi-Qi Cheng, Chong Luo, Xintong Han, Zuxuan Wu, and Yu-Gang Jiang. Motionfollower: Editing video motion via lightweight score-guided diffusion. *arXiv preprint arXiv:2405.20325*, 2024. 1

[13] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3621–3631, 2023. 1

[14] 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. 1

[15] Zejia Weng, Xitong Yang, Zhen Xing, Zuxuan Wu, and Yu-Gang Jiang. Genrec: Unifying video generation and recognition with diffusion models. *arXiv preprint arXiv:2408.15241*, 2024. 1

[16] Zhen Xing, Qi Dai, Zihao Zhang, Hui Zhang, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. Vidiff: Translating videos via multi-modal instructions with diffusion models. *arXiv preprint arXiv:2311.18837*, 2023. 1

[17] Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *arXiv preprint arXiv:2310.10647*, 2023. 1

[18] Zhen Xing, Qi Dai, Zejia Weng, Zuxuan Wu, and Yu-Gang Jiang. Aid: Adapting image2video diffusion models for instruction-guided video prediction. *arXiv preprint arXiv:2406.06465*, 2024. 1

[19] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1481–1490, 2024. 2

[20] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1