3D Layout-Guided Autonomous Driving Scene Generation

Yu Liu[1*]

[1]Huaxin Software College, Tianjin University of Technology, 300387, Tianjin, China

Abstract. Improving the robustness of autonomous driving perception models relies on large-scale, diverse scenario data. However, real-world road data has challenges such as high collection costs, scarcity of extreme scenarios, and complexity in multi-view labeling. Generative AI scene synthesis technology has emerged as a key solution, with diffusion models gradually replacing GAN models as the mainstream. This paper provides a systematic review of autonomous driving scene synthesis technology, outlining the evolution of the technology, clarifying the core features and logic of different generations; it focuses on analyzing the representative solution DrivingDiffusion, the first video generation framework to achieve "3D layout controllability, multi-view coordination, and temporal coherence," dissecting its architecture and core module design based on latent diffusion models (LDM). It further compares the performance of diffusion-based methods with traditional GAN-based approaches across key metrics like scene fidelity and label consistency. Moreover, it extracts the key issues and challenges in the current field; finally, it looks forward to future development directions, providing a reference for subsequent research on related virtual data generation.

1 Introduction

Training the robustness of autonomous driving perception systems requires massive amounts of scenario data, but there are three major bottlenecks in acquiring real data. The first is the high cost of collection, with individual scenario collection often exceeding ten thousand yuan; the second is the scarcity of long-tail scenarios, such as extreme weather or unexpected events, which are difficult to capture effectively; and the third is low annotation efficiency, with labeling a single scenario often taking more than 10 hours. Autonomous driving scenario synthesis technology has undergone three generations of evolution. The first generation, mainly from 2015 to 2018, focused on physical simulation methods represented by CARLA and Unity. While these technologies have strong controllability, the realism of the generated scenarios is noticeably insufficient. The second generation, from 2018 to 2021, primarily involved GAN-based generation models such as DriveGAN and SceneGAN. These models have the advantage of faster training speeds but are prone to mode collapse. The third generation has been continuously developing since 2021, centered on diffusion models like Stable Diffusion and ControlNet. These models significantly improve generation quality but

* Corresponding author's email: liuy32235@gmail.com

still have issues such as lack of 3D structure constraints, single-view limitations, and no temporal modeling capability. Current mainstream layout-guided methods also have their own shortcomings. For example, BEVGen can achieve multi-view static generation but cannot ensure temporal consistency; LayoutDiffuse lacks specialized optimization design for driving scenarios; DriveDreamer does not possess precise 3D layout control; while DrivingDiffusion fills these technological gaps, achieving full-process scenario video generation.

Existing diffusion models face three major challenges in the process of synthesizing scenes for autonomous driving. The first is a lack of cross-view consistency, manifested in issues such as multi-camera data being prone to positional misalignment and road connections breaking across different perspectives. The second is insufficient cross-frame temporal consistency, which prevents the model from generating smooth and continuous scene videos. The third is limited instance generation quality, often resulting in blurred textures and distorted shapes of objects. The core motivation of DrivingDiffusion is precisely to generate scene videos that meet the requirements of 'structural consistency, view collaboration, temporal coherence, and realistic instances' by integrating 3D layout guidance, multi-dimensional consistency mechanisms, and temporal modeling techniques, thereby specifically addressing the above problems of existing diffusion models.

DrivingDiffusion adopts a technical approach of '3D layout as the core, multi-stage cascade modeling,' and the entire scene generation process is divided into three key stages. The first stage is multi-view single-frame generation. This stage uses a 3D layout and text descriptions as generation conditions, and through the collaborative function of a built-in layout controller and cross-view attention mechanism, it ultimately generates multi-view images of the scene's first frame. The second stage focuses on single-view video generation. In this stage, the first frame image serves as a key frame, while optical flow priors and cross-frame attention mechanisms are integrated to generate subsequent frames with continuity. The third stage involves post-processing optimization. By fine-tuning model parameters and combining with a sliding window algorithm, this stage further extends the length of the generated video, ensuring the integrity and coherence of the overall generation effect.

In terms of core contributions, this work first outlines the technological evolution in the field of autonomous driving scene synthesis, thereby clarifying the core development trends of 'layout guidance, multi-view collaboration, and temporal modeling.' Next, it provides an in-depth analysis of the working principles of each core module in the DrivingDiffusion framework, offering detailed support for understanding this technical solution. At the same time, a four-dimensional evaluation system is constructed, and systematic experiments are conducted to verify the superiority of the methods under consideration in scene synthesis. Finally, the study also analyzes existing problems in the current field and looks ahead to future research directions.

2 Related Work

2.1 Classification and Evolution of Autonomous Driving Scenario Synthesis Technology

Autonomous driving scenario synthesis technology can be divided into static image generation and dynamic video generation, with representative methods, core principles, and advantages and disadvantages varying across different technical approaches. In the field of static image generation, the mainstream approach is physics-based simulation methods, represented by CARLA. Its core principle is to construct virtual environments through a

physics engine, with the advantages of strong controllability and support for automatic annotation, but it has drawbacks such as low realism and coarse details.

Dynamic video generation encompasses multiple technical directions. Among them, GAN-based methods, represented by DriveGAN, rely on adversarial training between a generator and discriminator. Their prominent advantages are fast training speed and low inference latency, but they are prone to mode collapse and semantic distortion, which fundamentally shares limitations in representational capacity with early generative flow models like Glow [1]. General diffusion methods, such as Stable Diffusion [2], use text or edge maps to guide latent diffusion, producing high-quality, detail-rich content, though they lack 3D constraints and are limited to a single viewpoint. Layout-guided methods, such as BEVGen [3], generate multi-view images through BEV layout mapping, ensuring structural consistency and correct physical logic, yet suffer from low instance quality and lack temporal coherence. Text-guided methods like Make-A-Video extend image diffusion models with spatiotemporal U-Net, supporting diverse scene types without layout constraints, but they are less targeted and have insufficient temporal performance. Driving-specific methods, such as DriveDreamer [4], learn road dynamics based on a world model, achieving temporally coherent and road-rule-compliant results, but lack 3D control and multi-view capabilities. In contrast, the layout-guided DrivingDiffusion, centered on '3D layout, keyframes, sliding window,' has the advantages of controllable layouts, multi-dimensional consistency, and high instance quality, with the only drawbacks being high computational cost and limitations in long video generation.

2.2 Work Related to Core Supporting Technologies

2.2.1 Latent Diffusion Model (LDM)

The core innovation of LDM is transferring the diffusion process to a low-dimensional latent space, which ensures generation quality while reducing computational cost. The theoretical foundation of traditional diffusion models comes from Denoising Diffusion Probabilistic Models (DDPM) [5]. Training such models in a 256×256 pixel space (with a dimensionality of 196,608) requires hundreds of GPU days. LDM compresses this into a low-dimensional latent representation using an autoencoder; for example, with a downsampling factor of 8, the dimensionality can be reduced to 3,072, lowering computational complexity by tens of times. Its workflow is divided into two steps: in the perceptual compression stage, the model uses an autoencoder that incorporates perceptual loss, adversarial loss, and regularization to convert images into latent representations to ensure information preservation; in the latent space diffusion stage, the forward process gradually adds noise to the latent representation, while the reverse process predicts the noise through a conditional U-Net network, completing training with an MSE loss. At the same time, it leverages cross-attention mechanisms to integrate conditions such as text and layout, supporting layout-driven scene synthesis.

2.2.2 Layout-to-Image Generation Technology

The evolution of layout-to-image generation technology has gone through three generations. The first generation is represented by GAN-based methods such as SPADE [6]. These methods use layouts as conditions to achieve basic controllability, but they are prone to mode collapse and perform poorly in restoring complex layouts. The second generation is represented by diffusion-based methods like LayoutDiffuse. This method, developed further from LayoutDiffusion [7], relies on diffusion models to enhance image realism and mitigate mode collapse, but due to the lack of 3D optimization design, it struggles to meet the physical

realism requirements of autonomous driving scenarios. The third generation is represented by driving-specific methods such as BEVGen and DrivingDiffusion [3]. These methods introduce 3D representations like BEV layouts to improve the structural consistency of generated results. Among them, DrivingDiffusion strengthens the alignment of multi-view images through the combination of a 3D layout controller and cross-view attention mechanisms, and its performance surpasses similar methods like BEVGen.

2.2.3 Temporal Consistency Modeling Techniques

The mainstream methods for temporal consistency modeling can be divided into three categories. The first category is the spatiotemporal module extension method, represented by VDM. It adds spatiotemporal attention modules and 3D convolution layers to the traditional U-Net architecture of image diffusion models, enabling the generation of videos with rich details. However, it requires massive annotated video data, which makes it difficult to implement in autonomous driving scenarios. The second category is the keyframe control method, represented by DrivingDiffusion. It uses the first frame as the keyframe and relies on a cross-frame attention mechanism to establish the correlation between the current frame, the keyframe, and previous frames, reducing dependence on data. Additionally, a pre-trained pose regression network is introduced to further enhance temporal stability. The third category is the optical flow prior integration method, represented by DriveDreamer [4]. It uses optical flow prediction to guide smooth transitions of object motion, but errors in optical flow prediction are prone to accumulate and propagate, leading to scene drift issues when generating long videos.

3 Comparison of Autonomous Driving Scenario Synthesis Technologies

Currently, autonomous driving scenario synthesis technology has developed multiple technical paths, generally categorized into single-view single-frame, single-view multi-frame, and multi-view multi-frame methods. Each type focuses on different core capabilities but also has obvious shortcomings. Single-view single-frame methods, represented by ControlNet [8] and Stable Diffusion [2], can generate high-quality static images but do not support multi-view or temporal generation, making it difficult to meet the training needs of multi-camera perception. ControlNet improves structural controllability through additional control signals, offering better structural restoration accuracy than Stable Diffusion. However, both lack 3D constraints, resulting in insufficient structural consistency in complex scenes. Single-view multi-frame methods, such as DriveDreamer [4], have the capability to generate continuous video but only support a single view and may produce rough image quality and poor temporal consistency. They rely on a flow prior to guide the sequence, lack 3D layout constraints, and are prone to unreasonable phenomena such as target position jumps. Multi-view single-frame methods, exemplified by BEVGen [3], can achieve collaborative multi-view generation and have basic scene structure restoration capabilities. However, they cannot generate temporal videos and have low target instance restoration accuracy, making it difficult to meet downstream requirements for instance detail. DrivingDiffusion uses a multi-stage cascading framework, balancing multi-view output with multi-frame generation capabilities. Through designs such as 3D layout constraints, cross-view attention, and temporal optimization, it achieves coordinated improvements in structural consistency, temporal coherence, and instance quality, addressing the core shortcomings of existing methods.

4 Experimental Results and Analysis

The experiments were conducted based on the nuScenes and CARLA datasets, where nuScenes serves as a multimodal autonomous driving benchmark dataset, providing standardized test samples and evaluation metrics for the quantitative assessment of scene synthesis methods [1, 9]. The evaluation dimensions cover image quality, structural consistency, temporal coherence, instance quality, and downstream data augmentation effects. Core metrics include FID, SSIM, road and vehicle mIoU, FVD, and Object NDS, with a focus on comparing the overall performance differences among various synthesis techniques.

4.1 Quantitative Performance Comparison

Table 1. Quantitative comparison of various methods on the nuScenes validation set

Method	Multiple perspectives	Multi-frame	FID↓	FVD↓	Road mIoU↑	Vehicle mIoU↑	Object NDS↑
BEVGen [4]	√	×	25.54	-	50.2	5.9	-
DriveDreamer [5]	×	√	52.6	452	-	-	-
ControlNet [3]	×	×	25.1	-	61.0	28.5	27.3
Stable Diffusion [2]	×	×	28.3	-	48.6	22.3	25.1
DrivingDiffusion	√	√	15.83	332	63.2	31.6	33.1

As shown in Table 1, each type of method has obvious shortcomings. Single-view single-frame methods cannot meet the requirements of multi-view temporal sequences; single-view multi-frame methods lack image quality and temporal coherence; multi-view single-frame methods have low instance reconstruction accuracy. DrivingDiffusion performs best across all key metrics, balancing the ability to generate multi-view multi-frame outputs, with a significant overall performance advantage.

4.2 Comparison of Downstream Data Augmentation Effects

Table 2. Comparison of Performance Improvements of BEVFusion Perception Model by Various Synthetic Data and Downstream Data Augmentation Effects

Training Data Configuration	Increase in synthetic data volume	Object NDS↑	mAOE↓	mATE↓
Only real data	0	0.412	0.5613	0.2845
Real data 2000 frames of synthetic data	2000	0.418 (+1.5%)	0.5295 (-5.7%)	0.2732 (-4.0%)
Real data 6000 frames of synthetic data	6000	0.434 (+2.2%)	0.5130 (-8.6%)	0.2618 (-7.9%)
Real data 10,000 frames of synthetic data	10000	0.438 (+2.4%)	0.5087 (-9.4%)	0.2583 (-9.2%)

As shown in Table 2, all types of synthetic data can improve the performance of the BEVFusion perception model [10]. BEVFusion, as a representative framework for LiDAR and camera fusion, shows performance improvements that directly reflect the supportive value of synthetic data for real perception tasks. Among them, the data generated by

DrivingDiffusion shows increasingly better performance as the quantity increases, outperforming data generated by ControlNet and DriveDreamer in terms of overall object detection performance and pose estimation error optimization, highlighting its greater downstream practical value [4, 7].

5 Discussion

5.1 Core Advantages and Technological Innovation

DrivingDiffusion has multiple core advantages. It addresses structural consistency issues through precise control with 3D layout, ensures synchronized optimization of perspective and timing through multi-dimensional consistency, and effectively balances instance quality with overall structure. At the same time, it offers strong downstream utility and holds engineering application value.

In terms of technological innovation, DrivingDiffusion achieves precise feature mapping via a 3D layout controller, enhances consistency using cross-view and cross-frame attention mechanisms, strengthens target semantics with a local prompting module, and balances long video generation with efficiency through a sliding window algorithm.

5.2 Limitations

There are currently four technical gaps in the field of autonomous driving scene synthesis. The first is the poor mapping between 3D layout and images. Existing methods mostly rely on 2D or simplified 3D representations for guidance, making it difficult to capture the 3D topology of roads and easily leading to physically inconsistent scenes such as floating vehicles or broken roads. The second is the weak multi-view collaboration capability. Existing methods lack effective cross-view constraints, causing misalignment in object positions and proportions or discontinuities in road edges across multi-camera images, which can interfere with the training of multi-view fusion modules in perception models. The third is the lack of balance between temporal coherence and instance quality. Most methods either simplify instance details to ensure temporal smoothness or focus on improving instance generation quality while neglecting temporal consistency, making it difficult to simultaneously meet perception models' training needs for accurate object recognition and motion prediction. The fourth is the insufficient verification of the practical usefulness of generated data. Existing studies often focus on optimizing image quality metrics but rarely quantify how generated data improves performance on perception tasks, resulting in a disconnect between generation technology and downstream applications. DrivingDiffusion, through designs such as a 3D layout controller and cross-view and cross-frame consistency mechanisms, specifically addresses the first three issues and also validates the practical usefulness of the generated data.

5.3 Future Research Directions

Future work can advance the optimization and expansion of DrivingDiffusion from multiple dimensions: achieving model lightweighting through techniques such as knowledge distillation and quantization-pruning to meet real-time deployment requirements on in-vehicle platforms; for long video scenarios, introducing Transformer's temporal modeling capabilities and dynamically adjusting the sliding window step to improve the coherence of long video generation; for complex scenarios, integrating trajectory prediction technology with traffic flow rule databases to enhance the simulation capability of complex dynamic

scenes; at the same time, extending to a multimodal direction by incorporating LiDAR point clouds, voice commands, and other conditions to enrich the input dimensions of generation tasks; it is also possible to explore end-to-end perception-generation integrated solutions to enable adaptive generation for specific scenarios. Overall, the development of this technology could follow an evolution path from static images to dynamic videos, from single conditions to multimodal conditions, and from offline generation to online deployment.

6 Conclusion

This article reviews the evolution of autonomous driving scene synthesis technology and analyzes the DrivingDiffusion framework. Based on the latent diffusion model, this method, through multi-module collaborative design, addresses the core challenges of general diffusion models in driving scene synthesis. Experiments show that DrivingDiffusion outperforms baseline methods in overall performance, and the generated data effectively improves the performance of the BEVFusion perception model. Although there are limitations such as high computational costs and insufficient long-video coherence, its technical approach lays the foundation for the field. In the future, with the development of technologies like lightweight models and complex scene modeling, autonomous driving scene synthesis will achieve a leap from offline data augmentation to online real-time deployment, promoting the commercialization of autonomous driving.

References

1. D. L. Sohl-Dickstein J. Bengio. Density estimation using real NVP. IEEE Trans. Pattern Anal. Mach. Intell., **40**(11): 2698-2711, (2017), DOI:10.1109/TPAMI.2017.2768399
2. R. R. Blattmann, D. Lorenz, et al. High-resolution image synthesis with latent diffusion models. IEEE Trans. Pattern Anal. Mach. Intell.,**45**(11): 13024-13035, (2022), DOI: 10.1109/TPAMI.2022.3221685
3. P. S. Chen Y. Jiang, et al. BEVGen: Multi-view image generation from bird's-eye view layouts for autonomous driving. Comput. Vis. Image Underst., **231**: 103586, (2023), DOI: 10.1016/j.cviu.2023.103586
4. Y. W. Zhang M. Li, et al. DriveDreamer: World model-driven autonomous driving scenario generation. IEEE Robot. Autom. Lett., **8**(7): 4120-4127, (2023), DOI: 10.1109/LRA.2023.3289781
5. J. H. Jain P. Abbeel. Denoising diffusion probabilistic models. Proc. Adv. Neural Inf. Process. Syst., Vancouver, Canada, 6840-6851 (2020)
6. X. H. Wang, G. Ding, et al. LayoutDiffusion: Layout-guided image generation with self-attention guidance. IEEE Trans. Multim., **25**: 5620-5632, (2023), DOI: 10.1109/TMM.2022.3209499
7. L. Z. Agrawala. ControlNet: Adding conditional control to text-to-image diffusion models. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., New Orleans, USA, 3788-3797 (2023)
8. H. C. Bankiti A. H. Lang, et al. nuScenes: A multimodal dataset for autonomous driving. IEEE Trans. Pattern Anal. Mach. Intell., **44**(10): 6365-6379, (2022), DOI: 10.1109/TPAMI.2021.3119520

9. Z. L. Chen, K. Chen, et al. BEVFusion: A simple and robust LiDAR-camera fusion framework. IEEE Trans. Intell. Transp. Syst., **25**(3): 2566-2576, (2024), DOI: 10.1109/TITS.2023.3290568
10. J. Z. Park, P. Isola, et al. Semantic image synthesis with spatially adaptive normalization. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Long Beach, USA, 2337-2346 (2019)