MagicDrive: Street View Generation with Diverse 3D Geometry Control

Ruiyuan Gao[1]*, Kai Chen[2]*, Enze Xie[3]†, Lanqing Hong[3], Zhenguo Li[3], Dit-Yan Yeung[2], Qiang Xu[1]†

[1]The Chinese University of Hong Kong [2]Hong Kong University of Science and Technology [3]Huawei Noah's Ark Lab * Equal contribution † Corresponding authors

ICLR

Abstract

Recent advancements in diffusion models have significantly enhanced the data synthesis with 2D control. Yet, precise 3D control in street view generation, crucial for 3D perception tasks, remains elusive. Specifically, utilizing Bird's-Eye View (BEV) as the primary condition often leads to challenges in geometry control (*e.g.*, height), affecting the representation of object shapes, occlusion patterns, and road surface elevations, all of which are essential to perception data synthesis, especially for 3D object detection tasks. In this paper, we introduce MagicDrive, a novel street view generation framework, offering diverse 3D geometry controls including camera poses, road maps, and 3D bounding boxes, together with textual descriptions, achieved through tailored encoding strategies. Besides, our design incorporates a cross-view attention module, ensuring consistency across multiple camera views. With MagicDrive, we achieve high-fidelity street-view image & video synthesis that captures nuanced 3D geometry and various scene descriptions, enhancing tasks like BEV segmentation and 3D object detection.

Motivation and Task

Figure 1. Multi-camera street view generation from MagicDrive.

Task: Given a set of bounding boxes, a BEV map for road semantics, several camera poses, and text descriptions of the scene, MagicDrive generates realistic street images/videos which support perception task training.

Motivation: Annotated data is expensive, so we generate.

- **Controllable generation** and **diffusion models** make it possible for annotated data generation.
- **3D scene control** has not been investigated.
- **Cross-view & temporal consistency** is crucial for training perception tasks.

Method

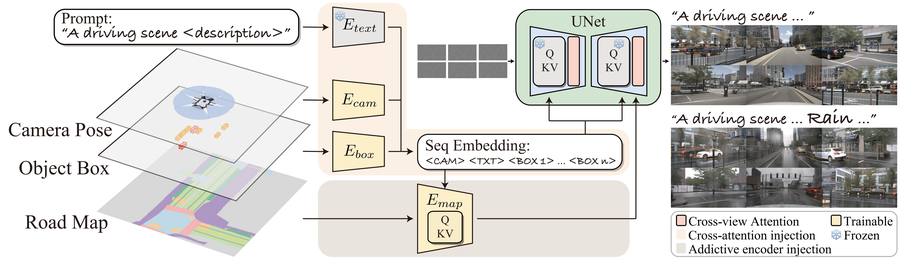


Figure 2. Overview of MagicDrive for street-view image generation. MagicDrive generates highly realistic images, exploiting geometric information from 3D annotations by independently encoding road maps, object boxes, and camera parameters for precise, geometry-guided synthesis. Additionally, MagicDrive accommodates guidance from descriptive conditions (*e.g.*, weather).

💡 Addictive branch (*e.g.*, ControlNet) can be used for conditions from different views.

3D Bounding Box Encoding

❓ Why not project all geometric conditions to the image view and apply 2D conditions?

(a) 3D bounding boxes show position relationships. (b) Road surface elevation guided by 3D bounding boxes.

Figure 3. 3D bounding boxes are crucial for street view synthesis. 2D boxes or BEV maps-only lost distance, height, and elevation.

Multi-view Consistency

💡 Cross-view attention (w/o prior) is good enough and extendable to video generation.

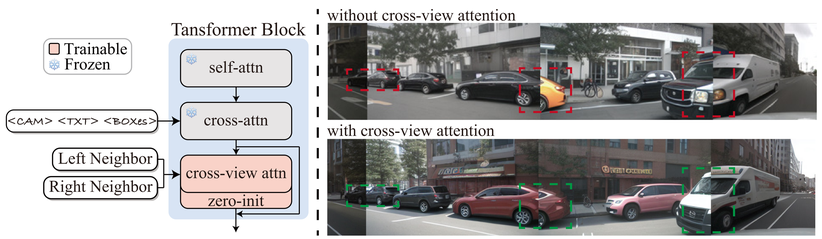
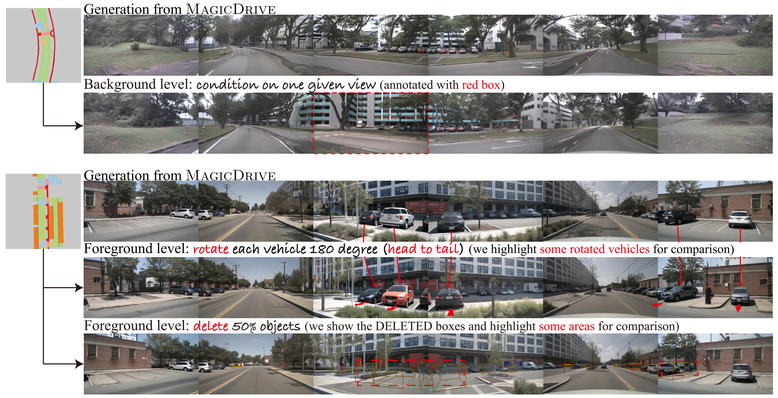


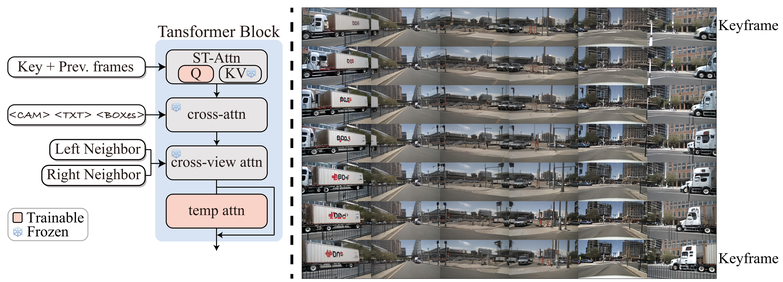
Figure 4. Cross-view Attention. *left*: we introduce cross-view attention to the pre-trained UNet after the cross-attention module. *right*: cross-view attention guarantees consistency across multiple views.

Results

Multi-level Control



Extend to Video Generation



Training Support for Perception Tasks

Table 1. Comparison about support for 3D object detection model (*i.e.*, BEVFusion).

Modality	Data	mAP ↑	NDS ↑
C	w/o synthetic data	32.88	37.81
	w/ MagicDrive	35.40 +2.52	39.76 +1.95
C+L	w/o synthetic data	65.40	69.59
	w/ MagicDrive	67.86 +2.46	70.72 +1.13

Table 2. Comparison about support for BEV segmentation model (*i.e.*, CVT).

Data	Vehicle mIoU ↑	Road mIoU ↑
w/o synthetic data	36.00	74.30
w/ BEVGen	36.60 +0.60	71.90 -2.40
w/ MagicDrive	40.34 +4.34	79.56 +5.26