MagicDrive-V2: High-Resolution Long Video Generation for Autonomous Driving with Adaptive Control

Ruiyuan Gao[1], Kai Chen[2], Bo Xiao[3], Lanqing Hong[4], Zhenguo Li[4], Qiang Xu[1]

[1]The Chinese University of Hong Kong [2]Hong Kong University of Science and Technology [3]Huawei Cloud [4]Huawei Noah's Ark Lab

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Abstract

The rapid advancement of diffusion models has greatly improved video synthesis, especially in controllable video generation, which is vital for applications like autonomous driving. Although DiT with 3D VAE has become a standard framework for video generation, it introduces challenges in controllable driving video generation, especially for frame-wise geometric control, rendering existing methods ineffective. To address these issues, we propose *MagicDrive-V2*, a novel approach that integrates the MVDiT block and spatial-temporal conditional encoding to enable multi-view video generation and precise geometric control. Additionally, we introduce an efficient method for obtaining contextual descriptions for videos to support diverse textual control, along with a progressive training strategy using mixed video data to enhance training efficiency and generalizability. Consequently, *MagicDrive-V2* enables multi-view driving video synthesis with 3.3× resolution and 4× frame count (compared to current SOTA), rich contextual control, and geometric controls. Extensive experiments demonstrate *MagicDrive-V2*'s ability, unlocking broader applications in autonomous driving.

Motivation

❗ *We want to further push the limit of driving video generation for practical use cases!*

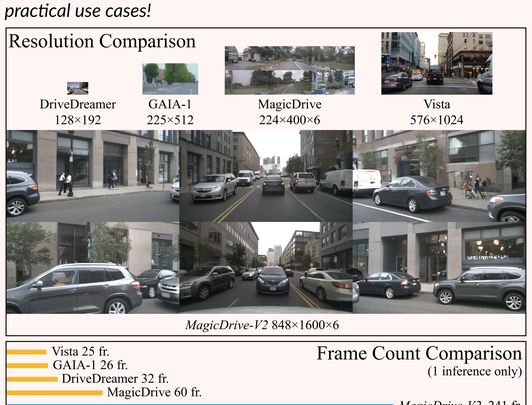


Figure 1. Resolution and frame count comparison for driving video generation models. Most of the previous work falls short of high resolution or long video generation.

Method

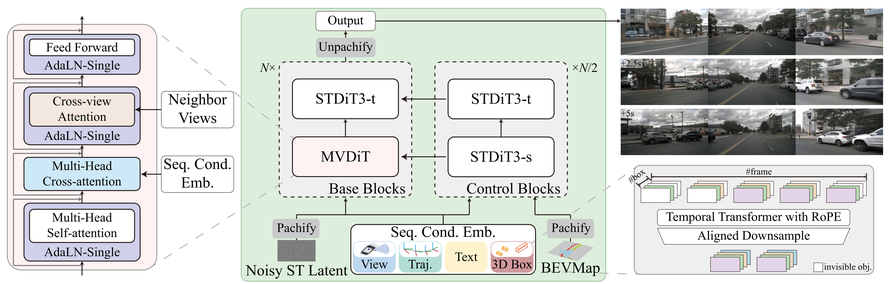
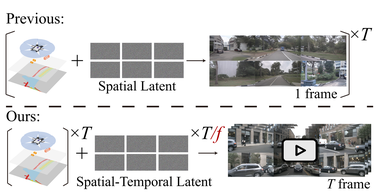
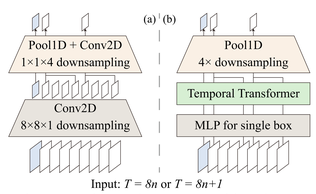


Figure 2. **Architecture Overview of *MagicDrive-V2***. To incorporate different conditions for video generation, *MagicDrive-V2* adopts a two-branch architecture with basic STDiT3 blocks. We propose the MVDiT block for multi-view consistency and Spatial-Temporal (Box/Traj.) Encoder to inject condition into the Spatial-Temporal (ST) Latent.

Spatial-Temporal Latents Need Spatial-Temporal Control



(a) Different from the spatial-only latent from 2D VAE, the spatial-temporal latents from 3D VAE require spatial-temporal condition injection (ours) for frame-wise geometry controls.



(b) **Spatial-Temporal Encoder for Maps** (a) **and Boxes** (b). Only temporally aligned embedding works for spatial-temporal latent control.

Diverse Text Control

💡 **Image caption** is good enough to describe the environmental information.

Figure 4. *MagicDrive-V2* supports diverse text control by re-captioning the video center frame.

Mixed Training and Progressive Scale-up

💡 Mixed video training enable our model to generate 8× more frames during inference.

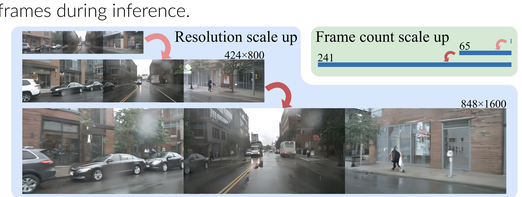


Figure 5. Progressive bootstrap training in *MagicDrive-V2*. For high-resolution long video generation, we train the model to progressively scale up from both the resolution and the frame count dimensions.

Results

Qualitative Comparison

Controllability

Figure 6. *MagicDrive-V2* generates high-resolution (e.g., 424×800 here) street-view videos for 241 frames (i.e., the full length of nuScenes videos, approximately 20 seconds at 12 FPS) with multiple controls (i.e., road map, object boxes, ego trajectory, and text). Notably, the 241-frame length at 424×800 is unseen during training, demonstrating our method's generalization capability to video length. We annotate the ego-vehicle trajectory and selected objects to aid localization, with same-color boxes denoting the same object.