DiVE: DiT-based Video Generation with Enhanced Control

Junpeng Jiang[1,2]* Gangyi Hong[3,2]* Lijun Zhou[2], Enhui Ma[4,2], Hengtong Hu[2], Xia Zhou[2]
Jie Xiang[2], Fan Liu[5], Kaicheng Yu[4], Haiyang Sun[2], Kun Zhan[2], Peng Jia[2], Miao Zhang[1†]
[1]Harbin Institute of Technology (Shenzhen)   [2]Li Auto Inc.   [3]Tsinghua University
[4]Westlake University   [5]National University of Singapore
Project Page: https://liautoad.github.io/DIVE/

Abstract

Generating high-fidelity, temporally consistent videos in autonomous driving scenarios faces a significant challenge, e.g. problematic maneuvers in corner cases. Despite recent video generation works are proposed to tackcle the mentioned problem, i.e. models built on top of Diffusion Transformers (DiT), works are still missing which are targeted on exploring the potential for multi-view videos generation scenarios. Noticeably, we propose the first DiT-based framework specifically designed for generating temporally and multi-view consistent videos which precisely match the given bird's-eye view layouts control. Specifically, the proposed framework leverages a parameter-free spatial view-inflated attention mechanism to guarantee the cross-view consistency, where joint cross-attention modules and ControlNet-Transformer are integrated to further improve the precision of control. To demonstrate our advantages, we extensively investigate the qualitative comparisons on nuScenes dataset, particularly in some most challenging corner cases. In summary, the effectiveness of our proposed method in producing long, controllable, and highly consistent videos under difficult conditions is proven to be effective.

1. Introduction

Bird's-Eye-View (BEV) perception has gained significant attention for autonomous driving, highlighting its immense potential in tasks such as 3D object detection [6]. Recent approaches like StreamPETR [15] utilize multi-view videos for training, emphasizing the need for extensive, well-annotated datasets. However, gathering and annotating such data across diverse conditions is challenging and costly. To address the mentioned challenges, recent advancements in generative models show that synthetic data can effectively improve performance in various tasks like object detection and semantic segmentation.

As the involvement of temporal data in video plays a crucial role in relative perception tasks, our focus in this paper shifts to generating high-quality realistic videos. Achieving real-world fidelity requires high visual quality, cross-view and temporal consistency, and precise controllability. Notice that the potential of recent methods are limited due to disadvantages in including low resolution, fixed aspect ratios, and inconsistencies in object shape and color. Inspired by the success of Sora's performance in task of generating high-quality, temporally consistent videos, we adapt the Diffusion Transformer (DiT) for controllable multi-view video generation in our work.

Our proposed framework is among the first few works which propose to use DiT for video generation in driving scenarios, enabling precise content control by integrating bird's-eye view (BEV) layouts and scene text. Building on top of OpenSora [16] architecture, our method embeds joint cross-attention modules to manage the scene text and instance layouts from bird's-eye views. Extending the ControlNet-Transformer [3] approach for road sketches, we ensure multi-view consistency with parameter-free spatial view-inflated attention. For the aim of supporting multi-resolution generation, faster inference, and various video length, we utilize OpenSora's training strategy and introduce a novel classifier-free guidance technique to enhance control and video quality.

2. Methodology

The overall architecture of our model is illustrated in Figure 1. The parametric model proposed by OpenSora 1.1 [16] is adopted as the baseline model. To achieve precise control over foreground and background information, we incorporate layout entries and road sketches, derived from 3D geometric data through projection, into the process of layout-conditioned video generation [8]. Our proposed novel modules and training strategies will be introduced in the following sections accordingly.

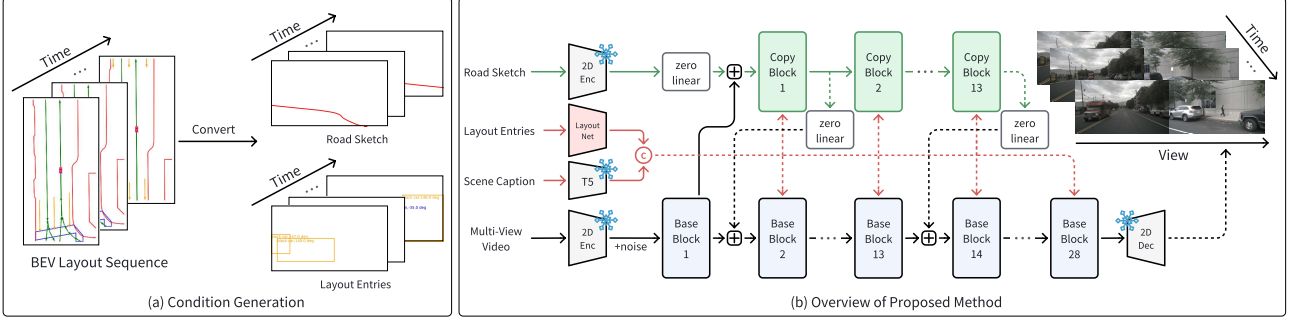*Equal contribution.
†Corresponding author.

Figure 1. The structural implementation with each individual components in our proposed method.

2.1. Mutli-Conditioned Spatial-Temporal DiT

Following OpenSora 1.1 [16], we utilize a pre-trained and frozen variational autoencoder from LDM [14] to extract latent features $z \in \mathbb{R}^{V \times T \times 4 \times h \times w}$ from the input multi-view video clip, where V represents the number of views, T denotes the length of frames, h and w denote the height and width of the latent features, respectively. These features are then modeled for spatiotemporal information using a 3D patch embedded. The textual input is encoded into 200 tokens using the T5 [13] language model.

Spatial View-Inflated Attention. To guarantee the multi-view consistency during generation, we replace the commonly used cross-view attention modules [4, 8] with a parameter-free view-inflated attention mechanism. Specifically, we extend 2D spatial self-attention to enable cross-view interactions by reshaping the input from $B \times V \times T \times H' \times W' \times C$ to $B \times T \times (VH'W') \times C$ and treating $VH'W'$ as the sequence length. Consequently, our proposed approach improves the multi-view coherence without compensating with additional parameters.

Caption-Layout Joint Cross-Atttention. Following MagicDrive [4], we use a cross-attention mechanism to integrate scene captions and layout entries. Whereas the layout entries, *i.e.* instance details such as 2D coordinates, heading and ID, are Fourier-encoded and combined into a unified embedding. Instance captions are encoded using a pre-trained CLIP [12] model. These embeddings are concatenated and processed through an MLP, producing the final layout embedding, which, along with the scene caption embedding, conditions the cross-attention mechanism.

ControlNet-Transformer. Delving into details, we introduce ControlNet-Transformer to ensure the precision towards the road sketch control inspired by PixArt-δ [3]. Practically, a pre-trained VAE extracts the latent features from road sketches, which are then processed by a 3D patch embedder for the sake of consistency issue with our main network. To parameterize our mentioned design, 13 duplicated blocks are integrated with the first 13 base blocks with the DiT [11] architecture. Each duplicated block combines the road sketch features and base block outputs, using spatial self-attention to reduce the computational overhead.

2.2. Training

Variable Resolution and Frame Length. Following OpenSora [16], we adopt the Bucket strategy, which ensures that videos within each batch have a consistent resolution and frame length.

Rectified Flow. Inspired by OpenSora 1.2 [16], we replace IDDPM [9] with rectified flow [7] during the later training stages for increased stability and reduced inference steps. Rectified flow, an ODE-based generative model, defines the forward process between data and normal distributions as

$$x_t = (1 - t)x_0 + tx_1 , \qquad (1)$$

where x_1 is a data sample, and x_0 is a sample from the normal distribution. The loss function is constructed as

$$\ell(\theta) := \mathbb{E}_{x_1, x_0}[\|v_\theta(x_t, t, c) - (x_1 - x_0)\|_2^2] , \qquad (2)$$

with c encompassing the three conditions. Sampling is performed from $t = 1$ to $t = 0$ in N steps via

$$x_{t-\frac{1}{N}} = x_t - \frac{1}{N}v_\theta(x_t, t, c), \forall t \in \{1, 2, ..., N\} / N . \quad (3)$$

First-k Frame Masking. To enable arbitrary-length video generation, we propose a first-k frame masking strategy, allowing the model to seamlessly predict future frames from the preceding ones. Formally, given a binary mask m indicating the frames to be masked—where the unmasked frames serve as the condition for future frame generation—we update x_t as

$$x_t \leftarrow x_t \odot (1 - m) + x_1 \odot m , \qquad (4)$$

with losses calculated only on unmasked frames. During inference, video is generated autoregressively, with the last-k frames of the previous clip conditioning the next.

Classifier-free Guidance for Multi-Conditions. We observe that extending classifier-free guidance from the text
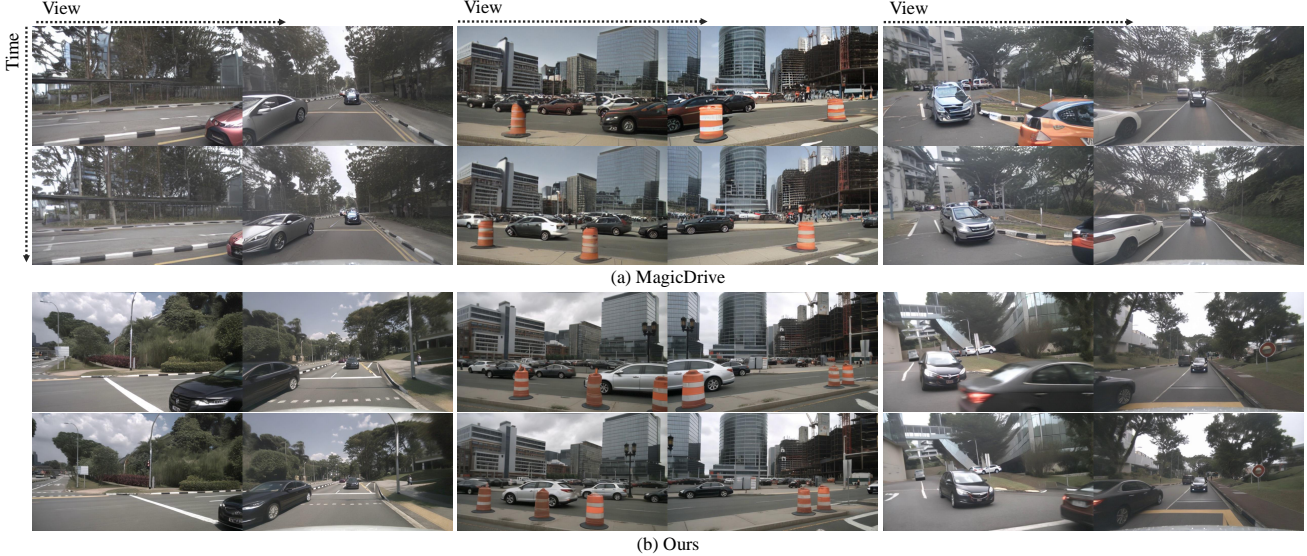
(a) MagicDrive

(b) Ours

Figure 2. Qualitative comparison of multi-view videos generated by our model and MagicDrive.

(a) "Sunny day, …"
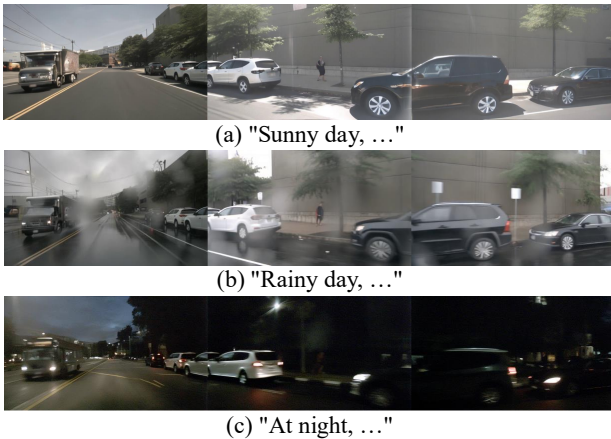

(b) "Rainy day, …"

(c) "At night, …"

Figure 3. The use-case of scene editing.

condition to layout entries and road sketches enhances conditional control precision and visual quality. During training, we set the text condition c_T, layout condition c_L, and sketch condition c_R to ϕ with a 5% probability each, and also enforce a 5% probability where all three conditions are simultaneously set to ϕ. The guidance scales λ_T, λ_L, λ_R correspond to the scene caption, layout entries, and road sketch, respectively, and measure the alignment between the sampling results and conditions. Inspired by [1] , the modified velocity estimate is as follows:

$$v'_\theta = v_\theta(x_t, \phi, \phi, \phi) \tag{5}$$
$$+ \lambda_T \cdot (v_\theta(x_t, c_T, c_L, c_R) - v_\theta(x_t, \phi, c_L, c_R)) \tag{6}$$
$$+ \lambda_L \cdot (v_\theta(x_t, \phi, c_L, c_R) - v_\theta(x_t, \phi, \phi, c_R)) \tag{7}$$
$$+ \lambda_R \cdot (v_\theta(x_t, \phi, \phi, c_R) - v_\theta(x_t, \phi, \phi, \phi)) . \tag{8}$$

3. Experiments

3.1. Setups

Dataset and Evaluation Metrics. We train and evaluate our model using the nuScenes [2] dataset and the interpolated 12Hz annotations provided by challenge. The generated multi-view videos are assessed based on distribution similarity (FVD), temporal consistency (DTC [10] and CTC [12]), visual quality (MUSIQ [5]) and controllability. Controllability is evaluated through two perception tasks: 3D object detection and BEV segmentation, with BEVFormer [6] serving as the perception model.

Training Details. We train our method in four stages using eight NVIDIA A800 GPUs. In the first stage, we fine-tune on OpenSora 1.1 [16] checkpoints with fixed-resolution images of 512×512 for $30k$ steps to control layout and sketch, training the ControlNet-Transformer, spatial attention, and layout net with spatial self-attention in base blocks. In the second stage, we train the model for $26k$ steps with variable resolutions (144p, 240p, 360p) and frame lengths to adapt to the nuScenes dataset, continuing to use spatial self-attention. The final two stages replace IDDPM with rectified flow, training for $20k$ steps at 144p to 360p, then $80k$ steps at higher resolutions (480p to full).

Inference Details. We perform sampling inference using rectified flow with 30 steps, choosing 480p resolution for a balance between inference time and visual quality. Each inference round uses a frame length of 16. We set λ_L and λ_R to 2.0, adjusting λ_T to 1.0 for night scenes and 7.0 for other scenes to achieve the best results.

Method	FVD$_\downarrow$	Object mAP$_\uparrow$	Map mIoU$_\uparrow$	DTC$_\uparrow$	CTC$_\uparrow$	IQ$_\uparrow$
MagicDrive	221.90	11.73	18.44	0.8755	0.9251	48.85
Ours	**94.60**	**24.55**	**35.96**	**0.9132**	**0.9446**	**51.82**

Table 1. Quantitative comparison with MagicDrive. DTC, CTC and IQ represent DINO Temporal Consistency, CLIP Temporal Consistency and Imaging Quality, respectively. The best performances are presented in **bold**.

3.2. Quality of Controllable Generation

To assess the quality of generated videos, we compare our method with the challenge baseline, MagicDrive [4], using evaluations on 16-frame sequences. As shown in Table 1, our model outperforms MagicDrive in terms of data distribution similarity, temporal consistency, visual quality, and controllability. Additionally, Figure 2 illustrates that the videos produced by our model exhibit both higher visual quality and better spatial consistency. Figure 3 demonstrates the scene editing capability of our method, where the weather in the generated video changes according to the caption, while other objects remain unchanged.

Method	FVD$_\downarrow$	Object mAP$_\uparrow$	Map mIoU$_\uparrow$	Score$_\uparrow$
CFG$_{T,L,R}$	94.60	24.55	**35.96**	**2.5962**
CFG$_{T,L}$	89.12	24.70	34.40	2.5487
CFG$_T$	**83.63**	20.05	34.26	2.1749
CFG$_{MagicDrive}$	164.48	**26.18**	35.02	2.3618

Table 2. Ablation on the classifier-free guidance.

3.3. Ablation Study

Effect of Proposed Classifier-free Guidance. We compared different classifier-free guidance methods, both with and without unconditional layout and sketch considerations, as detailed in Table 2. The "Score" is calculated as in the 1st round of the challenge, with CFG$_{T,L,R}$ being our proposed method. Excluding unconditional sketch (CFG$_{T,L}$) or both (CFG$_T$) yielded slightly better FVD but showed more pronounced differences in BEV segmentation and 3D object detection. We also evaluated CFG$_{MagicDrive}$ from Magic-Drive [4], which performed well in controllability but had only satisfactory FVD. Ultimately, CFG$_{T,L,R}$ achieved the best overall score.

4. Conclusion

In this paper, we present the first DiT-based controllable multi-view video generation model tailored for driving scenarios. The integration of ControlNet-Transformer and joint cross-attention facilitates precise control over BEV layouts. Spatial view-inflated attention, combined with a comprehensive set of training and inference strategies, ensures high-quality and consistent video generation. Comparisons with MagicDrive and various visualizations further demonstrate the model's superior control and consistency in generated videos.

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, pages 18392–18402, 2023. 3

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, and Others. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, pages 11621–11631, 2020. 3

[3] Junsong Chen, Yue Wu, Simian Luo, Enze Xie, Sayak Paul, Ping Luo, Hang Zhao, and Zhenguo Li. Pixart-δ: Fast and controllable image generation with latent consistency models. *arXiv preprint arXiv:2401.05252*, 2024. 1, 2

[4] Ruiyuan Gao, Kai Chen, Enze Xie, Hong Lanqing, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. In *ICLR*, 2024. 2, 4

[5] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *ICCV*, pages 5148–5157, 2021. 3

[6] 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 *ECCV*, pages 1–18. Springer, 2022. 1, 3

[7] Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023. 2

[8] Enhui Ma, Lijun Zhou, Tao Tang, Zhan Zhang, Dong Han, Junpeng Jiang, Kun Zhan, Peng Jia, Xianpeng Lang, Haiyang Sun, et al. Unleashing generalization of end-to-end autonomous driving with controllable long video generation. *arXiv preprint arXiv:2406.01349*, 2024. 1, 2

[9] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, pages 8162–8171. PMLR, 2021. 2

[10] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 3

[11] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023. 2

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 2, 3

[13] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 2

[14] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2

[15] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *CVPR*, pages 3621–3631, 2023. 1

[16] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024. 1, 2, 3