EFFICIENT MULTI-VIEW DRIVING SCENES GENERATION BASED ON VIDEO DIFFUSION TRANSFORMER

Junpeng Jiang[1,2] , **Gangyi Hong**[3,2] , **Hengtong Hu**[2] , **Lijun Zhou**[2] , **Tianyi Yan**[4,2] , **Yida Wang**[2] , **Kun Zhan**[2] , **Peng Jia**[2] , **Xianpeng Lang**[2] , **Miao Zhang**[1*]
[1] Harbin Institute of Technology, Shenzhen [2] Li Auto [3] Tsinghua University [4] University of Macau
jjunpeng1122@outlook.com, zhangmiao@hit.edu.cn

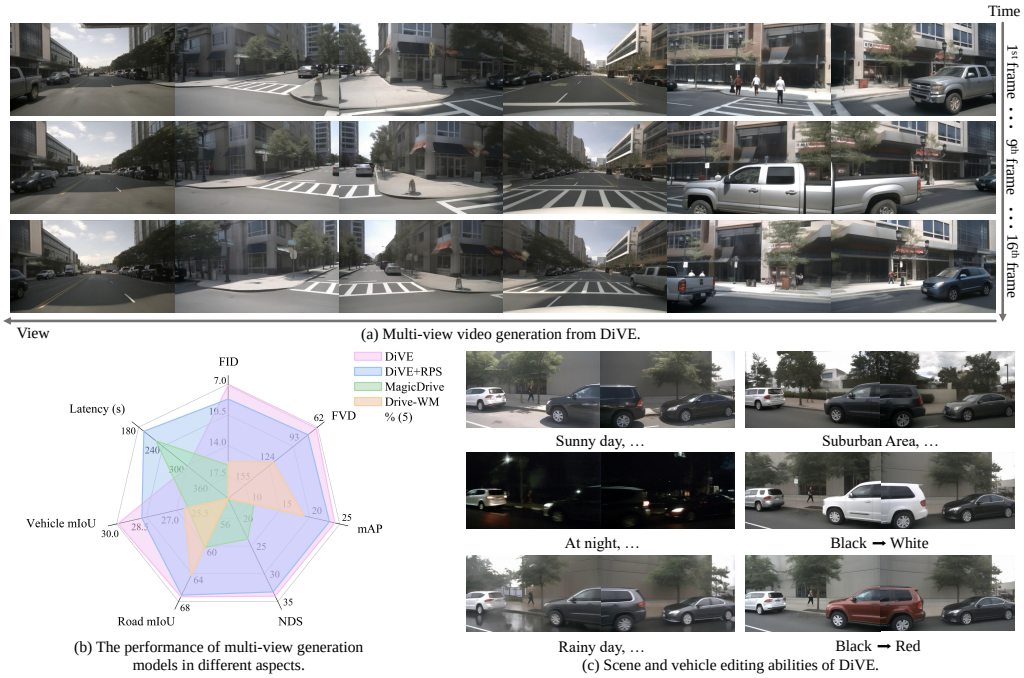


Figure 1: DIVE in qualitative and quantitative visualizations. (a) DIVE generates multi-view videos that exhibit strong realism, consistency, and controllability, making it an excellent simulator for controllably generating real-world driving scenarios under multi-modal conditions. (b) DIVE outperforms previous state-of-the-art multi-view generation models. With the proposed Resolution Progressively Sampling (RPS), we further significantly accelerate the inference of DIVE without sacrificing performance. (c) DIVE enables controllable video generation with varying attributes such as weather, time of day, location, and vehicle color.

ABSTRACT

Collecting multi-view driving scenario videos to enhance the performance of 3D visual perception tasks presents significant challenges and incurs substantial costs, making generative models for realistic data an appealing alternative. Yet, the videos generated by recent works suffer from poor quality and temporal consistency, which restricts their effectiveness in advancing perception tasks under driving scenarios. This gap highlights the need for a more robust and versatile framework capable of generating high-fidelity and temporally consistent multi-view videos, tailored to the complexities of driving scenarios. We introduce DIVE, a framework based on the Diffusion Transformer (DiT), designed to generate videos that are both temporally and cross-view consistent, aligning seamlessly with bird's-eye view (BEV) layouts and textual descriptions. Specifically, DIVE leverages cross-attention and a SketchFormer to exert precise control over multimodal data, while incorporating a view-inflated attention mechanism that adds no extra parameters, thereby guaranteeing consistency across views. To address

*Corresponding author.

the computational costs associated with high-resolution video generation, we further propose a training-free sampling strategy for acceleration called Resolution Progressively Sampling, achieving a remarkable $\times 1.62$ speedup without compensating the generation quality. In summary, DIVE delivers multi-view videos with outstanding visual quality and has demonstrated state-of-the-art performance on the nuScenes dataset. Additionally, the highly efficient and robust generation capabilities of DIVE offer promising avenues to support 3D perception models in achieving substantial performance improvements.

1 INTRODUCTION

3D visual perception tasks, such as 3D object detection and map segmentation, are crucial for autonomous driving systems. The perception methods (Huang et al., 2021; Zhou & Krähenbühl, 2022; Li et al., 2022; Yang et al., 2023a) based on bird's-eye view (BEV) holistic representations derived from multi-camera images have garnered significant attention for their ability to comprehensively understand complex environments. However, image data alone often falls short in capturing time-relevant targets such as velocity (Huang & Huang, 2022), whereas multi-view video data, with its temporal dynamics and rich information, is crucial for enhancing precise scene perception in autonomous driving systems (Wang et al., 2023; Liu et al., 2023a). Nevertheless, collecting and annotating multi-view video across various scenes is both challenging and expensive. Notably, existing studies (Swerdlow et al., 2024; Yang et al., 2023b; Gao et al., 2024; Wen et al., 2024) have validated that synthetic data can significantly improve the performance of perception models, and this paper mainly focuses on generating controllable high-quality synthetic videos to serve for the training of 3D perception models in autonomous driving.

Generally, scene text and BEV layouts are utilized to describe driving scenarios for producing diverse annotated videos in a controlled manner. Additionally, consistency and resolution are particularly important for the quality of synthetic videos, especially for tasks involving temporal modeling and perception (Wang et al., 2023; Wen et al., 2023). However, previous works (Gao et al., 2024; Wen et al., 2024) have often generated data with poor consistency, facing issues both in temporal dimensions and across viewpoints, and are limited to low-resolution videos, which can restrict the performance ceiling when using such low-quality synthetic data for model training.

Video generation models based on the Diffusion Transformer (DiT) (Peebles & Xie, 2023) architecture have recently demonstrated their ability to produce impressive high-quality videos, such as Sora (OpenAI, 2024). However, these methods (Zheng et al., 2024; Yang et al., 2024) generate videos only rely on text conditioning but not multi-modal conditions, and effective models for multi-view video generation in challenging driving scenarios remain lacking. Considering DiT's potential for temporal consistent and high-quality video generation, we aim to implement controllable generation of multi-view driving scene videos based on this architecture.

In this paper, we introduce DIVE, the first DiT-based video generation model with enhanced multi-modal control, tailored for generating controllable multi-view driving scene videos. Building upon OpenSora 1.1 (Zheng et al., 2024) as base model, we exploit its strengths in temporal priors and long-range dependencies to establishing a strong foundation for temporal consistency and high-resolution generation. We employ a cross-attention mechanism to simultaneously process driving scene descriptions, decomposed 3D objects, and camera information, to align complex scenes and precisely control foreground elements and consistent motion trajectories. For BEV road map alignment, inspired by (Chen et al., 2024), we employ SketchFormer to effectively handle road sketches. To ensure multi-view consistency, we incorporate a view-inflated attention that adds no extra parameters. Following (Zheng et al., 2024), we adopt a multi-resolution and multi-phase training strategy. This not only aids the model in learning features across different scales, but also enhances its ability to generate multi-resolution videos. Additionally, we introduce a first-k frame masking strategy to enable the generation of controllable, infinitely long multi-view videos in a rollout manner.

Nonetheless, generating high-resolution videos cannot overlook the computational burdens involved in the process. High-resolution generation typically requires more computations and longer inference times. Therefore, the ability to sample high-resolution videos with faster inference speeds is a crucial factor in evaluating the effectiveness of a generating model. To address this issue, we present an efficient and training-free inference strategy called Resolution Progressively Sampling,

which is built on multi-resolution generative ability. Specially, we perform inference at lower resolutions during the initial and mid-stages, gradually transitioning to desired high resolution in the later stages.

Our key contributions can be summarized as:

- We introduce DIVE, the first framework that attempts to apply the DiT architecture to generate multi-camera perspectives videos in driving scenarios. This framework enables effective simultaneous control over multiple conditions while ensuring strong consistency across views and in temporal aspects.
- We propose Resolution Progressively Sampling (RPS), a training-free acceleration strategy, where the inference mode with progressively increasing resolution alleviates the computational burden required for the early-stage inference. Equipped with RPS, DIVE achieves a $\times 1.62$ speedup with minimal performance degradation.
- We achieve state-of-the-art generation performance on the nuScenes (Caesar et al., 2020) dataset, with a notably significant reduction in FVD score by 36.7 compared to current SOTA methods. We also demonstrate that our consistent and high-resolution generated data can further improve the performance of current perception models.

2 Related Work

Diffusion Transformer for Generation. Diffusion Transformer (DiT)(Peebles & Xie, 2023) replaces U-Net with Vision Transformers (Dosovitskiy, 2020) to enhance scalability in Latent Diffusion Models (LDMs)(Rombach et al., 2022). While SD3 (Esser et al., 2024) advances DiT for text-to-image synthesis, Sora(OpenAI, 2024), OpenSora (Zheng et al., 2024), and CogVideoX (Yang et al., 2024) extend it to video generation using temporal or spatiotemporal attention. However, applying DiT to multi-view driving scenarios remains unexplored, as it demands cross-view consistency and precise control over foreground and background elements.

Multi-View Generation for Driving Scenes. Multi-view driving scene generation relies on BEV layouts for synthesis. BEVGen (Swerdlow et al., 2024) generates street-view images autoregressively using spatial embeddings and camera bias. BEVControl (Yang et al., 2023b) employs attribute-specific controllers and cross-view-cross-element attention to ensure consistency, while MagicDrive (Gao et al., 2024) uses 3D encoding and cross-view attention. Panacea (Wen et al., 2024) adopts a two-stage pipeline with decomposed 4D attention and BEV-guided ControlNet for panoramic videos. Drive-WM (Wang et al., 2024b) integrates end-to-end planning with video generation. Despite progress, challenges in resolution, fidelity, and cross-view coherence remain, motivating our approach.

High-Resolution Generation. High-resolution diffusion models aim to transcend fixed resolution limits. DemoFusion (Du et al., 2024) employs progressive upscaling with noise inversion, but requires full inference steps per scale. CheapScaling (Guo et al., 2024) introduces tuning-free pivot replacement and U-Net-specific time-aware upsampling for multi-scale synthesis. Megafusion (Wu et al., 2024) adopts a tuning-free truncate-and-relay strategy but lacks rectified flow optimization. Unlike these image-centric methods, our approach leverages DiVE's multi-resolution generation with resolution-aware timestep shift (Esser et al., 2024), enabling efficient high-quality video synthesis through low-to-high resolution sampling while reducing computational costs.

3 Preliminary

Rectified Flow (Liu et al., 2023b) bridges DDPM (Ho et al., 2020) (SDE-based, high-quality but slow) and DDIM (Song et al., 2020) (ODE-based, fast but lower fidelity) by optimizing straight ODE trajectories between distributions. Given an initial distribution π_1 and a target data distribution π_0, it trains a velocity field υ_θ to minimize path curvature via:

$$\ell(\theta) := \mathbb{E}_{x_1, x_0}\left[\int_0^1 \|\upsilon_\theta(x_t, t, c) - (x_1 - x_0)\|_2^2 \, \mathrm{d}t\right] , \tag{1}$$

where $x_0 \sim \pi_0$, $x_1 \sim \pi_1$. $x_t := (1-t)x_0 + tx_1$ is a linear interpolation between x_0 and x_1. Rectified flow guarantees well-defined and unique ODE solutions by preventing paths from crossing,

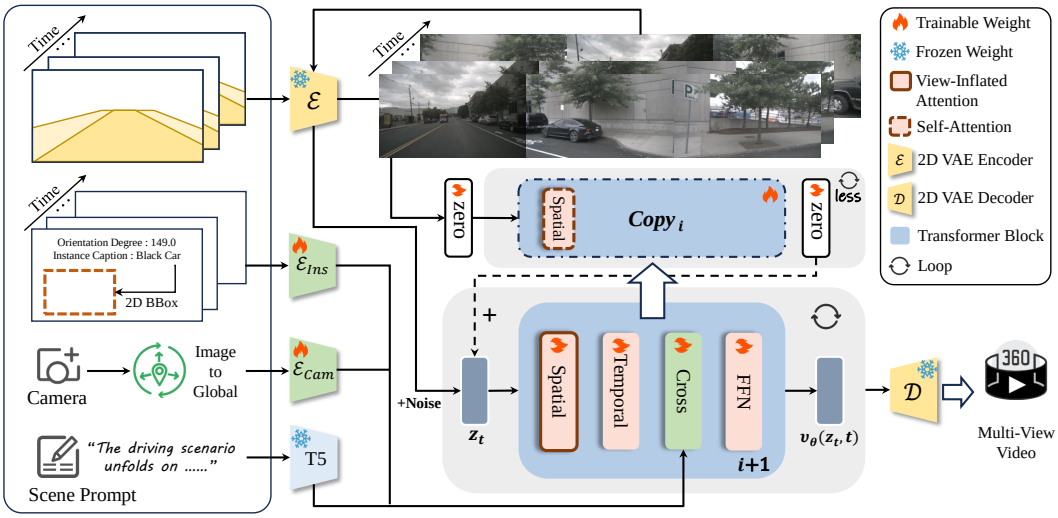


Figure 2: Overview of DIVE for multi-view video generation. Our model encodes four inputs for controllable generation: scene description words for global context, camera information for motion control, bounding boxes that locate 3D objects placement, and road sketches for road conditions. Each block in DIVE consists of spatial attention, temporal attention, cross attention, and an MLP. Notably, the view-inflated attention, which enhances view consistency, is integrated into the spatial attention mechanism within the backbone network.

which also leads to a theoretical reduction in convex transport costs by reconfiguring flows along straight paths between distributions. Practically, this approach results in nearly straight trajectories that require only a few reflow steps, enabling fewer Euler discretization steps and thus improving computational efficiency while minimizing discretization error.

Diffusion Transformer (DiT) (Peebles & Xie, 2023) replaces UNet backbones with scalable ViT-like (Dosovitskiy, 2020) blocks, where each layer combines multi-head self-attention and MLP modules. By integrating class and timestep-conditioned AdaLN layers, DiT enables efficient conditional generation while maintaining superior scalability over traditional diffusion architectures.

4 METHODOLOGY

In this section, we first elucidate the mechanisms by which DIVE attains cross-view consistency and multi-view controllability in Sec. 4.1. Then the training strategies are demonstrated in Sec. 4.2 accordingly. Finally, Sec. 4.3 addresses our strategy to foster a training-free inference acceleration.

4.1 DIVE

Figure 2 outlines our model's architecture, adopting OpenSora 1.1 (Zheng et al., 2024) as the baseline. We extract latent features z of multi-view video x using a frozen LDM (Rombach et al., 2022) pre-trained VAE, then encode them with a 3D patch embedder to capture spatiotemporal dynamics.

Unified Cross-Attention for Multimodal Conditioning. As shown in Figure 2, DIVE integrates multimodal conditions through a unified cross-attention with three coordinated encoding pathways:

- **Linguistic Guidance.** We establish a dual-scale text conditioning system where scene-level context and instance-specific captions are encoded through T5 (Raffel et al., 2020) and CLIP text encoder (Radford et al., 2021) respectively, producing semantic tokens $\mathcal{L} \in \mathbb{R}^{200\times d}$ and $\mathcal{T} \in \mathbb{R}^{n_{\text{ins}}\times d}$, where n_{ins} is the number of instances.
- **Geometric Grounding.** Building on spatial disentanglement approach (Yang et al., 2023b), 3D instances are projected into 2D space where bounding boxes $\mathcal{B}$ and orientation angles θ undergo Fourier encoding $\mathcal{F}$ (Mildenhall et al., 2020). These geometric features are fused with $\mathcal{T}$ via parameterized blending:

$$\mathcal{I} = \Phi\left(\mathcal{F}(\mathcal{B}), \mathcal{F}(\theta), \mathcal{T}\right) \ , \tag{2}$$

where Φ denotes our geometric fusion MLP.

- **Ego-motion Awareness.** To ensure cross-view consistency and motion direction, we derive camera-aware embeddings $\mathcal{P}$ through coordinate transformation:

$$\mathcal{P} = \Phi\left(\mathcal{F}\left(\begin{pmatrix} R & t \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} K^{-1} & 0 \\ 0 & 1 \end{pmatrix}\right)\right) , \tag{3}$$

  where rotation matrix $R \in \mathbb{R}^{3\times3}$ and translation vector $t \in \mathbb{R}^{3\times1}$ define the camera's pose in the global coordinate system. $K \in \mathbb{R}^{3\times3}$ is the camera intrinsic matrix.

The final condition $\mathcal{C} = [\mathcal{L}; \mathcal{I}; \mathcal{P}] \in \mathbb{R}^{(200+n_{\text{ins}}+1)\times d}$ establishes cross-modal correlations through cross-attention, enabling joint conditioning on semantic, geometric, and dynamic constraints.

SketchFormer for Road Guidance. Drawing from (Chen et al., 2024), we introduce SketchFormer, a novel framework for geometry-aware road generation. Our method injects latent sketch guidance through a three-stage process: (1) A pre-trained VAE compresses road sketches into disentangled embeddings; (2) A shared 3D patch embedder—identical to the one in the primary network—aligns these embeddings with the feature space; and (3) A cascade of 13 mirrored fusion cells (synchronized with the early layers of primary network) progressively blend sketch semantics via zero-initialized linear layers. This hierarchical conditioning ensures spatial alignment while mitigating feature collision.

View-Inflated Attention. Departing from parameter-heavy cross-view attention (Gao et al., 2024; Wen et al., 2024), we propose view-inflated attention: a simple yet effective approach that reshapes input features from $B \times V \times T \times H' \times W' \times C$ to $B \times T \times (VH'W') \times C$ prior to attention computation. This parameter-free operation implicitly enables cross-view interaction by treating $VH'W'$ as token length, achieving comparable consistency without introducing additional learnable weights. For SketchFormer, we deliberately omit this reshaping step to maintain training stability and computational efficiency, as sketches provide inherent spatial constraints.

4.2 Training DiVE

Multi-Scale Training. We propose a hierarchical training paradigm that progressively learns scale-aware representations through three distinct phases: (1) **Conditional Image Generation**, which establishes cross-modal grounding by mapping 3D constraints to multi-view images; (2) **Low-Resolution Video Training**, focusing on global spatiotemporal patterns and low-frequency features; and (3) **High-Resolution Video Refinement**, dedicated to capturing fine-grained visual details. This multi-scale framework provides dual benefits: (1) implicit data augmentation through scale variation, and (2) native support for Resolution Progressively Sampling (Section 4.3).

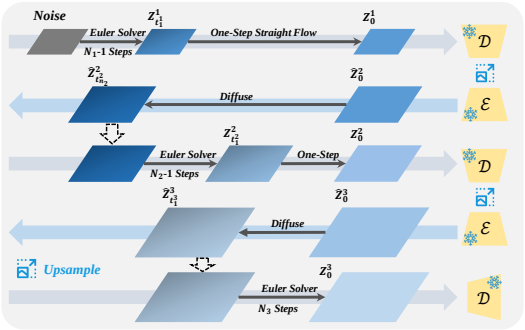


Figure 3: The overall process of Resolution Progressively Sampling. Larger quadrilaterals represent higher resolutions, and deeper colors indicate more noisy regions.

First-k Frame Masking. To enable arbitrary-length video generation, we introduce a first-k frame masking strategy, allowing the model to seamlessly predict future frames from the preceding ones. Formally, given a binary mask $m \in \mathbb{R}^T$, where $m_i = 1$ for $i \leq k$ and 0 otherwise and the masked frames serve as the condition for future frame generation, we update x_t as

$$x_t \leftarrow x_t \odot (1 - m) + x_1 \odot m , \tag{4}$$

with losses calculated only on unmasked frames. At inference, generation proceeds autoregressively, using the last-k frames as context for seamless extension.

4.3 Resolution Progressively Sampling

To alleviate the computational pressure in multi-view high-resolution video generation, we propose a training-free acceleration strategy: Resolution Progressively Sampling (RPS). It leverages the inherent multi-resolution capability of DiVE through two key innovations:

- **Progressive Resolution Scaling.** Inspired by (Guo et al., 2024; Du et al., 2024; Wu et al., 2024), we employ a multi-stage sampling framework. It iteratively refines the video from low-resolution (*e.g.*, 240p) latent z_1 to high-resolution (*e.g.*, 480p) x, with s stages and n_k steps per stage ($N = \sum_{k=1}^{s} n_k$). The sampling timesteps of each stage S_k decrease from $t_{n_k}^k$ to t_1^k. Between stages, after completing the first $n_k - 1$ steps, a one-step straight flow at t_1^k estimates z_0^k, which is decoded by the VAE into a clear video at the current resolution. This is then upsampled and encoded into the initial state of the next stage via diffusion, creating a cyclic progression from low to high resolution until the target resolution $h_s \times w_s$ is achieved.
- **Resolution-Aware Timestep Shift.** Inspired by (Esser et al., 2024; Hoogeboom et al., 2023), higher resolutions require greater noise to effectively disrupt the signal. The relationship between $t_{n_{k+1}}^{k+1}$ and t_1^k follows a shift function:

$$t_{n_{k+1}}^{k+1} = \frac{t_1^k \sqrt{(h_{k+1} w_{k+1})/(h_k w_k)}}{1 + t_1^k \left(\sqrt{(h_{k+1} w_{k+1})/(h_k w_k)} - 1 \right)}, \tag{5}$$

  ensuring increased noise at the $k+1$ stage.

As shown in Figure 1 and Table 4, RPS achievse 1.62× faster inference versus baseline, while enhancing visual fidelity through log-SNR consistency $\log \frac{h_k w_k}{h_{k+1} w_{k+1}}$ (Hoogeboom et al., 2023).

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUPS

Dataset. We conduct experiments on the nuScenes (Caesar et al., 2020) dataset, a publicly available 3D perception dataset for driving scenarios. The nuScenes dataset comprises 700 video sequences for training and 150 for validation. Each sequence is recorded at 12 Hz and lasts approximately 20 seconds, with annotations provided at 2 Hz. To achieve high-frame-rate generation, we utilized the 12 Hz interpolated annotations provided by W-CODA* for both training and evaluation.

Quality Metrics. To evaluate the quality of the generated video, we utilize two primary metrics: the frame-wise Fréchet Inception Distance (FID) (Heusel et al., 2017) and the Fréchet Video Distance (FVD) (Unterthiner et al., 2018). FID assesses the quality of individual frames, whereas FVD evaluates both the quality and temporal consistency of the video.

Controllability Metrics. To assess controllability, we employ CVT (Zhou & Krähenbühl, 2022) and BEVFusion (Liu et al., 2023c) to conduct a quantitative analysis of two perception tasks—BEV segmentation and 3D object detection. We generate the corresponding data based on the annotations from the validation set and evaluate performance using a model pretrained on real data. Additionally, we generate data on the training set and utilize the video-based perception method StreamPETR (Wang et al., 2023) to evaluate the effectiveness of DIVE in augmenting data.

5.2 IMPLEMENTATION DETAILS.

Our implementation is based on the OpenSora 1.1 (Zheng et al., 2024) codebase, initialized with pretrained weights. The multi-scale training process is carried out on 8 NVIDIA A800 GPUs, with each of the three phases comprising $20k$, $30k$ and $80k$ iterations, respectively. For inference, we utilize rectified flow (Liu et al., 2023b) with a classifier-free guidance (Ho & Salimans, 2021) scale of 2.0, performing 30 sampling steps to generate videos at a resolution of 480p (480×854) and 16 frames. Our Resolution Progressively Sampling acceleration strategy conducts 10 sampling steps at resolutions of 240p (240×426), 360p (360×640), and 480p (480×854) sequentially. When generating long videos, we generally set k to 4.

5.3 MAIN RESULTS

Quantitative Results. Table 1 shows the performance of DIVE and previous methods on the nuScenes validation. Our model achieves state-of-the-art FID (11.62) and FVD (68.4) scores, surpassing existing multi-view image generation models, including BEVGen (Swerdlow et al., 2024),

*W-CODA's homepage: https://coda-dataset.github.io/w-coda2024/

Table 1: Quantitative comparison of driving scenario generation methods evaluated on the nuScenes validation set. BEV segmentation and 3D object detection tasks use models pre-trained on nuScenes data. '+RPS' denotes the Resolution Progressively Sampling acceleration technique. DIVE (Ours) achieves top performance in all metrics, with and without RPS. The best (**bold**) and second-best (underlined) results are highlighted. Higher/lower metric values are preferred as indicated by ↑ / ↓.

Method	Avenue	Resolution	FID↓	FVD↓	BEV segmentation		3D object detection	
					Road mIoU↑	Vehicle mIoU↑	mAP↑	NDS↑
Real Data	-	-	-	-	73.67	34.81	35.54	41.21
BEVGen	RA-L'24	224×400	25.54	-	50.20	5.89	-	-
BEVControl	arXiv'23	-	24.85	-	60.80	26.80	19.64	-
MagicDrive	ICLR'24	224×400	16.20	-	61.05	27.01	12.30	23.32
DriveDreamer	ECCV'24	256×448	26.80	353.2	-	-	-	-
DriveDreamer-2	arXiv'24	256×448	25.00	105.1	-	-	-	-
Panacea	CVPR'24	256×512	16.96	139.0	55.78	22.74	11.58	22.31
Drive-WM	CVPR'24	192×384	15.80	122.7	65.07	27.19	20.66	-
Ours	-	480×854	**7.14**	**68.4**	**68.16**	**30.50**	**25.75**	**33.61**
Ours+RPS	-	480×854	8.70	78.8	67.92	29.31	24.87	32.84

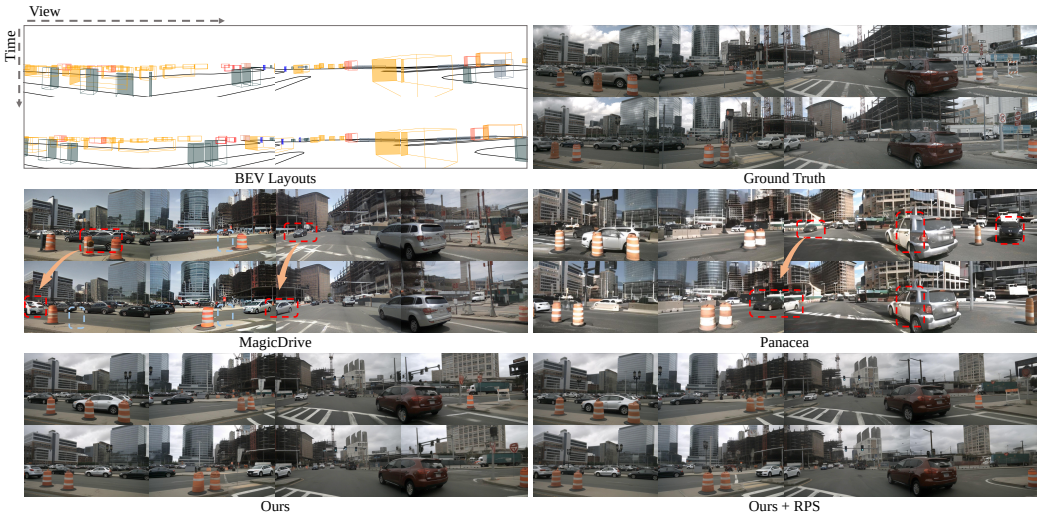


Figure 4: Qualitative comparison of DIVE with MagicDrive and Panacea. We use dashed boxes to highlight some of the noticeable issues in MagicDrive and Panacea, and arrows to indicate the changes in vehicle positions over time. In contrast, DIVE demonstrates superior realism, temporal and cross-view consistency, and controllability, both before and after applying RPS.

BEVControl Yang et al. (2023b), and MagicDrive (Gao et al., 2024), as well as video generation models such as DriveDreamer (Wang et al., 2024a), DriveDreamer-2 (Zhao et al., 2024), Panacea (Wen et al., 2024), Drive-WM (Wang et al., 2024b). This demonstrates DIVE's unique advantage in maintaining temporal consistency while preserving frame-level quality. The control precision analysis reveals deeper strengths: BEV segmentation performance (Road and Vehicle mIoU) aligns closely with real data distributions, while 3D detection performance (mAP and NDS) validates its precise spatial alignment capabilities. Despite minor performance dips under RPS, DIVE consistently outperforms existing methods. These quantitative findings confirm DIVE's dual capability to simultaneously optimize photorealism and geometric fidelity.

Qualitative Results. In Figure 4, DIVE demonstrates superior generation quality compared to MagicDrive (Gao et al., 2024) and Panacea (Wen et al., 2024), which suffer from issues like vehicle fragmentation and unrealistic artifacts (*e.g.*, rearview mirrors on vehicle backs). DIVE excels in producing highly realistic vehicles with precise fidelity, maintaining both temporal and cross-view consistency, unlike the significant color and shape variations in MagicDrive and Panacea. Additionally, DIVE ensures accurate scene controllability, generating correct object counts, road layouts, and realistic zebra crossings, while offering flexible customization of weather, time, archi-

Table 2: Comparison about support for StreamPETR.

Real	DiVE	mAP↑	mAOE↓	mAVE↓	NDS↑
✓	-	35.5	52.7	32.1	47.3
-	✓	27.3	56.3	50.3	40.4
✓	✓	36.7	42.1	32.4	49.2

Table 3: Ablation on the different spatial attention types.

Attention Type	FVD↓	Object mAP↑	Map mIoU↑
Self	93.14	24.57	35.87
Left-Self-Right	86.15	25.88	36.97
View-Inflated	**86.13**	**26.30**	**37.41**

tectural styles, and even vehicle colors (Figure 1 (c) and Figure 7) — a feature often lacking in prior methods. Notably, even when using the RPS inference acceleration strategy, the visual quality of DiVE's generated results remains almost indistinguishable from those produced without acceleration, demonstrating the effectiveness of RPS.

Generated Videos for Data Augmentation. To assess whether DiVE-generated data can enhance perception tasks, we adopt a similar approach to Panacea by synthesizing a new training dataset with DiVE and evaluating its impact using StreamPETR (Wang et al., 2023). When trained separately on the original nuScenes set and the DiVE-generated dataset, we achieve mAP and NDS scores of 27.3 and 40.4, respectively — 85.4% and 76.9% of those obtained using only real data (Table 2). This highlights the significant potential of synthetic data as a valuable training resource. Moreover, augmenting the original dataset with synthetic data further boosts StreamPETR's performance across nearly all metrics, underscoring DiVE's practical effectiveness in advancing perception tasks.

5.4 Ablation Study

Given the high resolution and frame count of DiVE-generated videos, full validation inference demands substantial resources. To streamline evaluation, we adopt W-CODA's approach, generating only the first 16 frames per scene across four runs, assessing quality via FVD and controllability using BEVFormer (Li et al., 2022) for 3D object detection and BEV segmentation.

View-Inflated Attention. We compare view-inflated attention with self-attention (which lacks view interaction) and left-self-right attention (focusing on neighboring views) to validate its effectiveness. As shown in Table 3, view-inflated attention excels in both generation quality and controllability. While left-self-right attention improves local consistency, it struggles with global semantic coherence, highlighting the superiority of view-inflated attention for achieving cross-view consistency and maintaining long-term generation quality.

Resolution Progressively Sampling (RPS). Table 4 demonstrates the validity of the resolution-aware timestep shift in RPS and the reasonableness of timestep distribution across resolutions. The i-j-k notation denotes the number of steps at 240p, 360p, and 480p resolutions, respectively. Without timestep shift (*e.g.*, 10-10-10), performance consistently drops, underscoring its importance. While increasing steps at higher resolutions improves quality, it also raises inference time; thus, 10-10-10 strikes the optimal balance between performance and efficiency.

Table 4: Ablation of Resolution Progressively Sampling.

Steps Choices	Timestep Shift	FVD↓	Object mAP↑	Map mIoU↑	Latency (s)↓
10-10-10	-	103.92	25.29	37.60	189.7
0-0-30	-	86.13	26.30	37.41	306.6
20-5-5	✓	119.62	24.54	36.72	134.4
5-5-20	✓	89.64	26.26	37.92	252.7
10-10-10	✓	97.74	26.26	37.68	189.7

6 Conclusion

We present DiVE, a pioneering DiT-based framework for generating multi-view driving scene videos. Through various improvements to the existing DiT architecture, DiVE is capable of generating videos that precisely align with 3D annotations and maintain temporal and multi-view consistency. We also introduce Resolution Progressively Sampling, an inference acceleration strategy that significantly improves efficiency while maintaining generation quality. DiVE achieves SOTA performance both before and after acceleration. By generating higher-quality videos efficiently, DiVE stands out as the preferred choice for enhancing data used in training perception tasks of autonomous driving.

REFERENCES

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

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.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Ruoyi Du, Dongliang Chang, Timothy Hospedales, Yi-Zhe Song, and Zhanyu Ma. Demofusion: Democratising high-resolution image generation with no $$$. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6159–6168, 2024.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

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 *The Twelfth International Conference on Learning Representations*, 2024.

Lanqing Guo, Yingqing He, Haoxin Chen, Menghan Xia, Xiaodong Cun, Yufei Wang, Siyu Huang, Yong Zhang, Xintao Wang, Qifeng Chen, et al. Make a cheap scaling: A self-cascade diffusion model for higher-resolution adaptation. *arXiv preprint arXiv:2402.10491*, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning*, pp. 13213–13232. PMLR, 2023.

Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022.

Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021.

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*, pp. 1–18. Springer, 2022.

Haisong Liu, Yao Teng, Tao Lu, Haiguang Wang, and Limin Wang. Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18580–18590, 2023a.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023b.

Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, pp. 2774–2781. IEEE, 2023c.

B Mildenhall, PP Srinivasan, M Tancik, JT Barron, R Ramamoorthi, and R Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, 2020.

OpenAI. Sora. 2024. URL `https://openai.com/index/sora/`.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

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.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

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

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

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*, pp. 3621–3631, 2023.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. In *European conference on computer vision*, 2024a.

Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the Future: Multiview Visual Forecasting and Planning with World Model for Autonomous Driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

Yuqing Wen, Yucheng Zhao, Yingfei Liu, Binyuan Huang, Fan Jia, Yanhui Wang, Chi Zhang, Tiancai Wang, Xiaoyan Sun, and Xiangyu Zhang. Panacea+: Panoramic and controllable video generation for autonomous driving. *arXiv preprint arXiv:2408.07605*, 2023.

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*, pp. 6902–6912, 2024.

Haoning Wu, Shaocheng Shen, Qiang Hu, Xiaoyun Zhang, Ya Zhang, and Yanfeng Wang. Megafusion: Extend diffusion models towards higher-resolution image generation without further tuning. *arXiv preprint arXiv:2408.11001*, 2024.

Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17830–17839, 2023a.

Kairui Yang, Enhui Ma, Jibin Peng, Qing Guo, Di Lin, and Kaicheng Yu. Bevcontrol: Accurately controlling street-view elements with multi-perspective consistency via bev sketch layout. *arXiv preprint arXiv:2308.01661*, 2023b.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. *arXiv preprint arXiv:2403.06845*, 2024.

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. *arXiv preprint arXiv:2412.20404*, 2024.

Brady Zhou and Philipp Krähenbühl. Cross-view transformers for real-time map-view semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13760–13769, 2022.

APPENDIX

A MORE IMPLEMENTATION DETAILS

A.1 CLASSIFIER-FREE GUIDANCE

We find that simply setting the text condition to *null* in the unconditional sampling of Classifier-free Guidance (CFG) (Ho & Salimans, 2021) is insufficient. While this often improves the fidelity of the generated videos, it leads to blurriness in 3D objects, weakening the model's control over them. Inspired by Gao et al. (2024), we extend the unconditional sampling of CFG. In the unconditional sampling, we set text condition c_T, 3D objects c_O and road sketches c_R to *null* ϕ. This approach maximizes the model's generation fidelity and controllability. The modified velocity estimate is as follows:

$$v'_\theta = v_\theta(x_t, \phi, \phi, \phi) \tag{6}$$
$$+ \lambda \cdot (v_\theta(x_t, c_T, c_O, c_R) - v_\theta(x_t, \phi, \phi, \phi)) \, . \tag{7}$$

Notably, we observe that when generating night scenes, CFG causes the scenes to be overly dark, as shown in Figure 5 (a). This is because, without CFG, DIVE already generates night scenes with reasonable brightness. However, CFG further emphasizes the darkness, leading to the disappearance of many foreground objects. To address this issue, we do not set the text condition to *null* in the unconditional sampling, while still setting the objects and sketches to *null*. This approach allows us to generate more realistic night scenes, as shown in Figure 5 (b). The corresponding modified velocity estimate is as follows:

(a) Set the text condition to *null* in the unconditional sampling

(b) Set the text condition to c_T in the unconditional sampling

Figure 5: Night scene generation in different CFG methods.

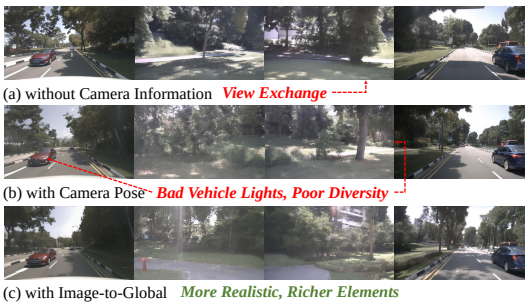
$$v'_\theta = v_\theta(x_t, c_T, \phi, \phi) \tag{8}$$
$$+ \lambda \cdot (v_\theta(x_t, c_T, c_O, c_R) - v_\theta(x_t, c_T, \phi, \phi)) \, . \tag{9}$$

A.2 MORE TRAINING DETAILS

All three stages of DIVE are trained with the Brain Floating Point (BF16) precision, and the AdamW optimizer with a learning rate of 1e-4 is adopted. We employ the Bucket mechanism for multi-resolution training. Specifically, for the main resolutions of 240p, 360p, and 480p, the batch sizes are 4, 2, and 1, respectively. The resolution for StreamPETR (Wang et al., 2023) training is 480p, which is different from the 256×704 of the baseline.

B ABLATION OF CAMERA INFORMATION

We find that DIVE can generate multi-view videos with reasonable motion directions even without camera information, but this is not always reliable. When adjacent views lack both the road sketch and 3D objects, the generated video is more likely to exhibit viewpoint exchange issues, as shown in Figure 6 (a). To address this issue, we encode camera information into the model. Unlike previous methods (Gao et al., 2024) that rely on camera poses, we use an image-to-global coordinate transformation matrix. This choice stems from our observation that using the image-to-global transformation matrix results in more realistic and



(a) without Camera Information

(b) with Camera Pose

(c) with Image-to-Global

Figure 6: Effect of camera information.

element-rich scenes. For example, as shown in Figure 6 (c), the middle two views include additional elements such as fire hydrants and buildings, which are absent when using camera poses, as illustrated in Figure 6 (b).

C MORE VISUALIZATION RESULTS

Flexible Controllable Generation. Figure 7 illustrates the generation results under the variation of objects conditions. After rotating all the objects by 180 degree, DIVE remains capable of generating reasonable and high-fidelity results, and the orientation after rotation is highly accurate. We also present the generation result after removing all the objects. It is clearly observed that the architecture of the scene has changed, which demonstrates the diverse generation ability of DIVE.

(a) Generation from the original nuScenes validation set

(b) Rotate all objects 180 degree (head to tail)

(c) Remove all objects

Figure 7: Flexible controllable generation from DIVE.

More Qualitative Comparison. We present more qualitative comparisons with MagicDrive (Gao et al., 2024) and Panacea (Wen et al., 2024) in Figure 8, 9. DIVE consistently exhibits a visual quality that is closer to the real world, as well as more natural and reasonable cross-view and temporal consistency.

Long Video Generation. Figure 10 presents the results of long video generation by DIVE, where each row depicts the outcomes of the keyframes. Surprisingly, DIVE maintains excellent temporal consistency even during the continuous generation of 240 frames, avoiding repetitive patterns. Moreover, the color of the vehicle remains unchanged regardless of the temporal sequence or the presence of other vehicles in the scene. This level of consistent generation is rare in previous works.

D LIMITATIONS

Although DIVE can generate driving scenes that are remarkably close to the real world at present, its controllability remains less satisfactory and lags behind the perception results based on real data. Alternative conditional embedding methods or data types warrant further exploration. For instance, AdaLN (Peebles & Xie, 2023) could potentially be employed to aggregate a series of control conditions.Additionally, the specific utilization strategy for generated data as augmentation samples is underdeveloped. Efficient and effective use of generated data for augmentation may yield greater benefits than developing a superior generative model. Future work could benefit from leveraging techniques related to dataset distillation.

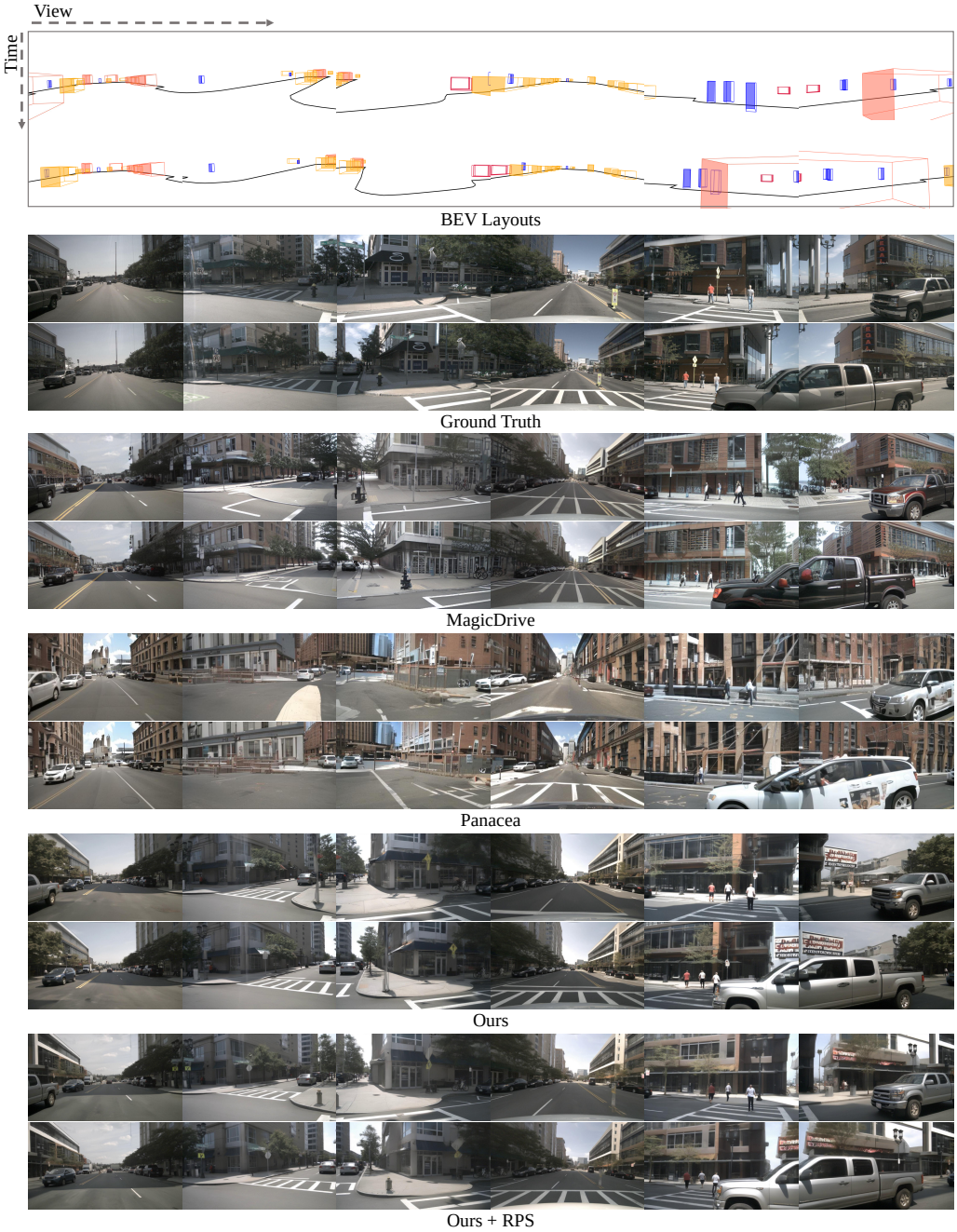


Figure 8: Qualitative comparison with MagicDrive and Panacea on driving scene from nuScenes validation set.

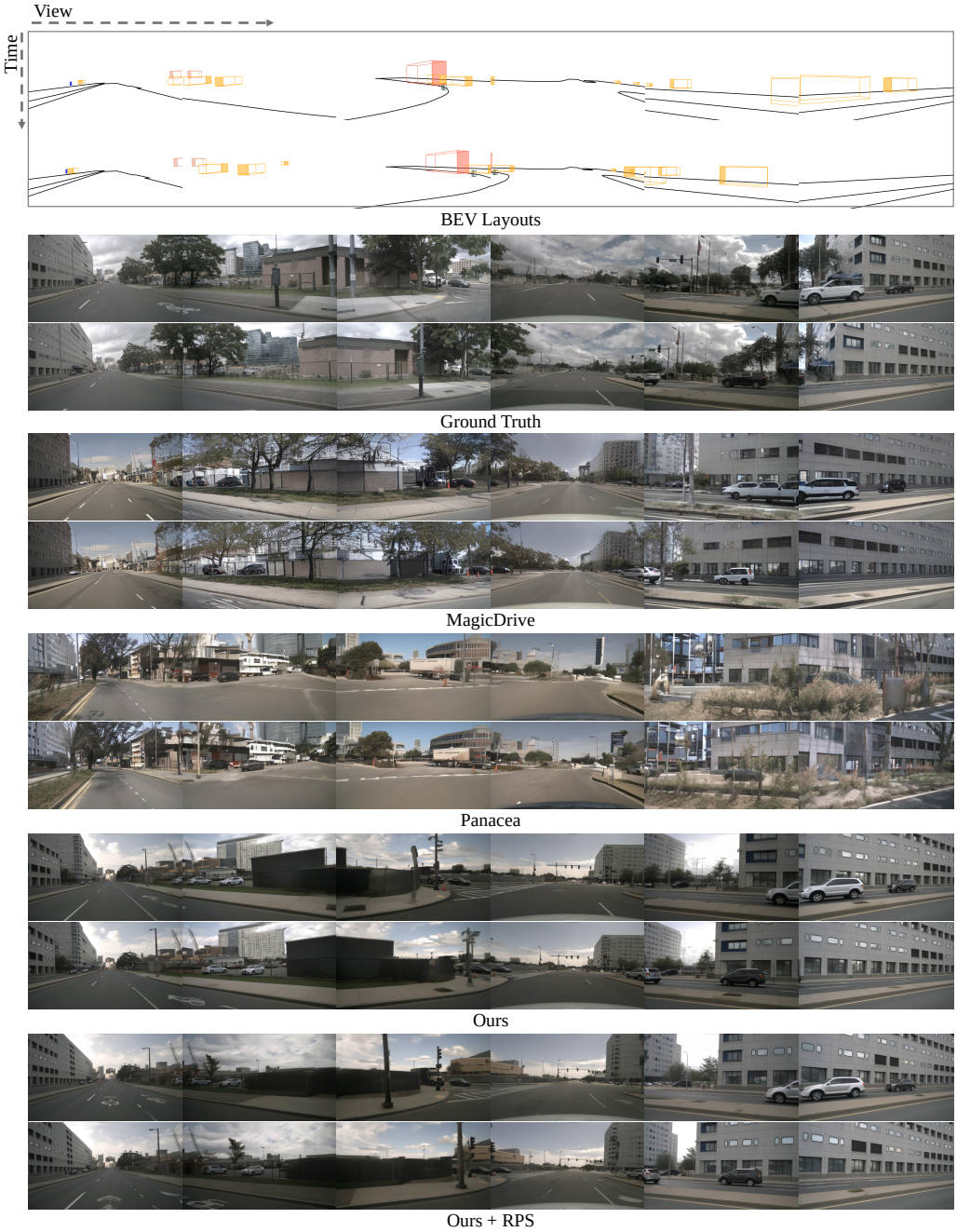


Figure 9: Qualitative comparison with MagicDrive and Panacea on driving scene from nuScenes validation set.

Figure 10: Long video generated by DiVE.