This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

DriveScape: High-Resolution Driving Video Generation by Multi-View Feature Fusion

Wei Wu[1,2], Xi Guo[2], Weixuan Tang[2], Tingxuan Huang[3],
Chiyu Wang[2], Chenjing Ding[2*]
[1]Tsinghua University [2]Sensetime Research [3]Northeastern University
{wuwei,guoxi,dingchenjing}@sensetime.com,
{tangweixuan, wangchiyu}@senseauto.com,
{huangtingxuan}@stu.neu.edu.cn

Abstract

*Recent advancements in generative models offer promising solutions for synthesizing realistic driving videos, aiding in training autonomous driving perception models. However, existing methods often struggle with high-resolution multi-view generation, mainly due to the significant memory and computational overhead caused by simultaneously inputting multi-view videos into denoising diffusion models. In this paper, we propose a driving video generation framework based on multi-view feature fusion named DriveScape for multi-view 3D condition-guided video generation. We introduce a Bi-Directional Modulated Transformer (BiMoT) module to encode, fuse and inject multi-view features along with various 3D road structures and objects, which enables high-resolution multi-view generation. Consequently, our approach allows precise control over video generation, greatly enhancing realism and providing a robust solution for creating high-quality, multi-view driving videos. Our framework achieves state-of-the-art results on the nuScenes dataset, demonstrating impressive generative quality metrics with an FID score of **8.34** and an FVD score of **76.39**, as well as superior performance across various perception tasks. This lays the foundation for more accurate environment simulation in autonomous driving. Our project homepage.*

1. Introduction

Autonomous driving has garnered significant attention, emphasizing the need for precise environmental understanding to ensure safe driving decisions [13, 38]. Bird's-eye view (BEV) maps, generated from multi-view images, serve as a crucial structured representation for various tasks, including 3D object detection, segmentation, tracking, depth estimation, and trajectory prediction [14, 15, 17, 18, 20, 34, 36, 39]. However, acquiring high-quality multi-view video data poses significant challenges due to the high labeling costs, increasing the demand for generating temporally consistent video data aligned with real-world distributions and 3D labels. Recent generative models have demonstrated the potential of synthetic data to enhance visual tasks [7, 25, 35], while diffusion models have advanced the generation of diverse and realistic driving videos, which are crucial for developing robust autonomous driving systems.

Model	Spatial Res.	Temporal Res. (Hz)	sparse condition
Drive-WM	192 × 384	2	×
WoVoGen	256 × 448	2	×
Delphi	512 × 512	2	×
GenAD	256 × 448	2	×
MagicDrive	272 × 736	-	×
DriveDreamer 1&2	448 × 256	2	×
DriveDiffusion	512 × 512	2	×
Panacea	256 × 512	2	×
Ours (DriveScape)	**576 × 1024**	**2 - 10**	✓

Table 1. Real-world multi-view driving controllable world models. DriveScape (ours) operates at a high spatiotemporal resolution, offering support for sparse layout controllability, as well as unified control for both mono-condition and multi-condition scenarios.

The most relevant approach to our work is the generation of multi-view driving videos controlled by 3D conditions. In terms of spatial resolution, general video generation tasks, such as those addressed by SVD, have achieved high-resolution video generation at 1024 × 576. However, in driving scenarios, multi-camera panoramic generation is often required. Previous methods typically need to denoise multiple views simultaneously, leading to memory constraints during both training and inference. For temporal resolution, previous methods rely on frame-by-frame 3D layout control. Due to high annotation costs, public datasets like NuScenes provide annotations at a low frame rate (2Hz), restricting models from generating videos at the

*Corresponding author

rate (2Hz), restricting models from generating videos at the same frame rate. Additionally, the cost of obtaining different types of annotations varies in practice. For instance, annotating 3D lane lines is typically more expensive than annotating bounding boxes, often resulting in lane lines being labeled at a lower frame rate. This necessitates generative models that can handle sparse conditional control and remain compatible with both sparse and non-sparse conditions simultaneously.

To address these challenges, we propose DriveScape, a solution for multi-view 3D condition-guided video generation with high spatio-temporal resolution. Unlike existing methods, DriveScape integrates adjacent views and keyframes through multi-view fusion, avoiding the computational burden of denoising multiple views simultaneously. We introduce a streamlined training and inference pipeline, free from post-processing, supporting high spatial and temporal resolution for sparse condition control. To align 3D road structural information, we propose a Bi-Directional Condition Alignment module, integrating multiple conditions into an aligned latent space for fine-grained layout controllability. By leveraging temporal and spatial information, DriveScape establishes a unified model for consistent multi-view video generation, significantly enhancing realism and applicability.

In summary, our key contributions are as follows:

- We propose a unified multi-view feature fusion framework for driving video generation without any post-processing. To the best of our knowledge, **our model is the first to achieve high-resolution, high-frame-rate, sparse condition control for multi-view driving video generation**.
- We introduce a Bi-Directional Condition Alignment module to facilitate diverse modalities of condition guidance and sparse condition control in video generation tasks. Experiments demonstrate that our model achieves decoupled, sparse, and precise control of dynamic foreground and static background.
- We achieve state-of-the-art video synthesis performance on the nuScenes dataset, excelling not only in the generation quality of multi-view images and videos but also across various perception applications.

2. Related Works

2.1. Controllable generation

With the advent of diffusion models, significant progress has been made in text-to-video generation [1, 2, 5, 8, 10, 12, 24, 30, 40]. Video LDM [4] employs a latent diffusion pipeline where denoising occurs in the image latent space, greatly accelerating the process. However, text alone cannot precisely control video generation. Later methods introduced image blocks alongside text as prompts for the denoising network [37].Our goal is to generate highly realistic street-view videos, a complex scene featuring numerous elements and interactions (e.g., intricate street layouts and moving vehicles). To achieve fine-grained control, we incorporate road maps, 3D bounding boxes, and BEV keyframes, going beyond text and images to ensure detailed and precise video generation.

2.2. Multi-view video generation

Multi-view consistency and temporal consistency are two critical challenges in multi-view video generation. To address multi-view consistency, MVDiffusion [26] introduced a correspondence-aware attention module to align information across views. Tseng et al. [27] utilized epipolar geometry to regularize consistency between different views. MagicDrive [7] leveraged camera poses, bounding boxes, and road maps as priors, incorporating an additional cross-view attention block to enhance consistency. However, these methods are limited to generating multi-view images rather than videos.

2.3. Street view generation

Most street-view generation models condition on 2D layouts, such as BEV maps, 2D bounding boxes, and semantic segmentation. BEVGen [25] encapsulates all semantics within BEV for street-view generation. BEVControl [35] employs a two-stage method to generate multi-view urban scene images from BEV layouts, where a controller generates foreground and background objects, and a coordinator combines them to maintain visual consistency across views. However, projecting 3D information to 2D loses geometric details, making direct extension to video generation prone to inconsistencies across frames. To address this, we incorporate 3D bounding boxes as a condition to guide generation.DrivingDiffusion [16] adopts a multi-stage pipeline with two models and two stages of post-processing to generate frames and extend them into videos. In contrast, our method avoids the complexity of multi-stage pipelines by using an efficient, end-to-end approach.

3. Method

3.1. Overview

The overview of DriveScape is shown in Fig. 1. Based on the LDM pipeline [22], DriveScape generates street-view videos conditioned on scene annotations, such as BEV maps, 3D bounding boxes, ego states, and text for each view. The unified model, described in Section 3.2, enables high-resolution multi-view video generation without post-refinement.

The Bi-Directional Condition Alignment module, detailed in Section 3.5, ensures effective alignment between 3D road structures and temporal sparse conditions.

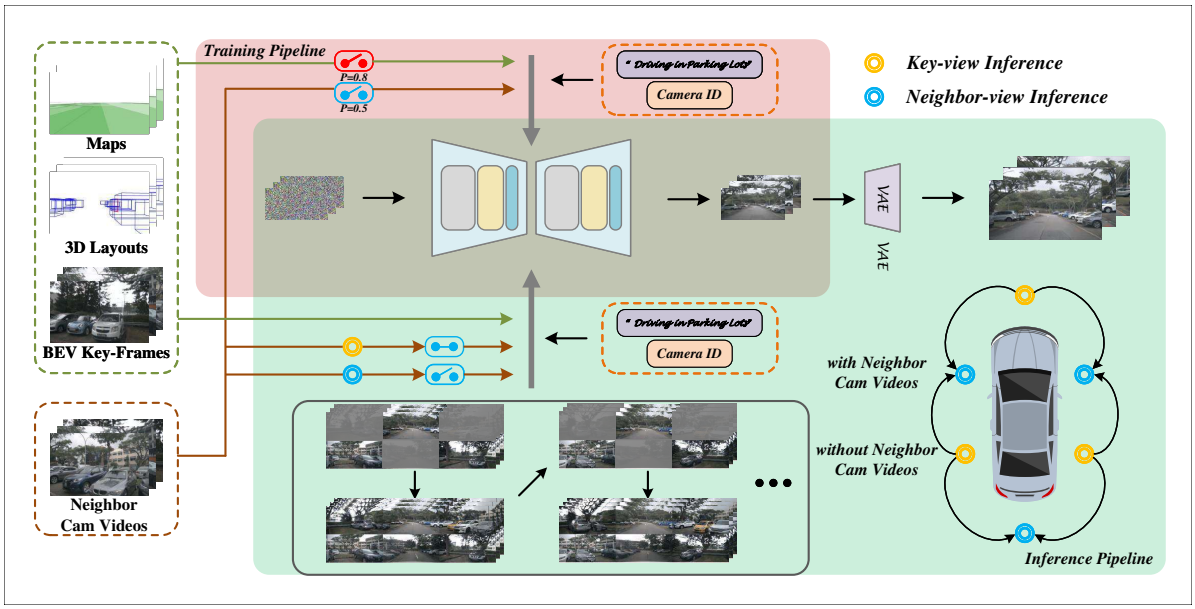


Figure 1. End-to-End Multi-view Video Generation pipeline. We use learnable embedding vectors to represent different cameras and categorize camera views as key views and neighbour views. Our training scheme guides the generation of key view videos through their neighboring frames. Additionally, we introduce the key-frames condition along with a training and inference scheme to ensure multi-view consistency simultaneously. Furthermore, our model does not require any post-refine process as DriveDiffusion ([16]). It is capable of learning multi-view and temporal consistency simultaneously, resulting in high-fidelity street-view synthesis.

3.2. Unified Model

Our model leverages SVD [3] as the backbone, with each denoising step processing a video from a single viewpoint. Control conditions are injected through a unified control module, including frame-wise images from adjacent viewpoints, road maps, 3D bounding boxes, and the first frame from each viewpoint. Starting with initial multi-view frames $x^0_{0:N-1}$ (N is the number of views), DriveScape predicts subsequent frames $x^{1:T}_{0:N-1}$ (T is the number of frames) under conditions such as BEV maps $m^{0:T}_{0:N-1}$, 3D bounding boxes $b^{0:T}_{0:N-1}$, BEV key frames, and neighboring camera videos. BEV maps and 3D bounding boxes are projected into the camera's First-Person View (FPV) and encoded with convolutional blocks.

As shown in Fig. 1, we use learnable embedding vectors $\{e_i\}$ to represent different cameras. Similar to the processing of canbus data (e.g., velocity and direction angle), these embeddings are passed through a multilayer perceptron (MLP) followed by a Sigmoid Linear Unit (SiLU) activation. This enables the embeddings to perform cross-attention with the image latent space, allowing the unified model to predict videos from multiple perspectives.

Embedding camera information alone cannot ensure spatial and temporal consistency, as it only captures global camera characteristics. Consistency is achieved through model design and training. For temporal consistency, we integrate temporal attention into the latent diffusion UNet [23] and control alignment module (Section 3.5). For multi-view consistency, we introduce keyframe conditions and a carefully designed training and inference scheme.

3.3. Training Stage

Neighbor frames feature fusion. DriveScape categorizes camera views into key and neighbor views (Fig. 2). Frames of the same type have minimal overlap, ensuring distinct regions. The training scheme guides key view generation using neighbor frames. During training, neighbor and key views are selected sequentially. For neighbor views, video conditions from adjacent key views are input into the network. Key views are generated sequentially, stored, and incorporated into the cross-attention mechanism with the image latent space, ensuring effective guidance for neighbor view generation.

Multi-view key-frame feature fusion. A challenge arises from the lack of specific multi-view constraints during key-view generation, leading to degraded performance. While this is generally acceptable due to minimal overlapping regions, long-term generation introduces interdependencies among views. For example, during overtaking, a vehicle initially captured by the rear camera later appears in the front camera. To address this, we propose key-frame conditions, which are the starting frames of all views, to establish correlations within the image latent space. Combined with temporal attention, the model learns the relationships across

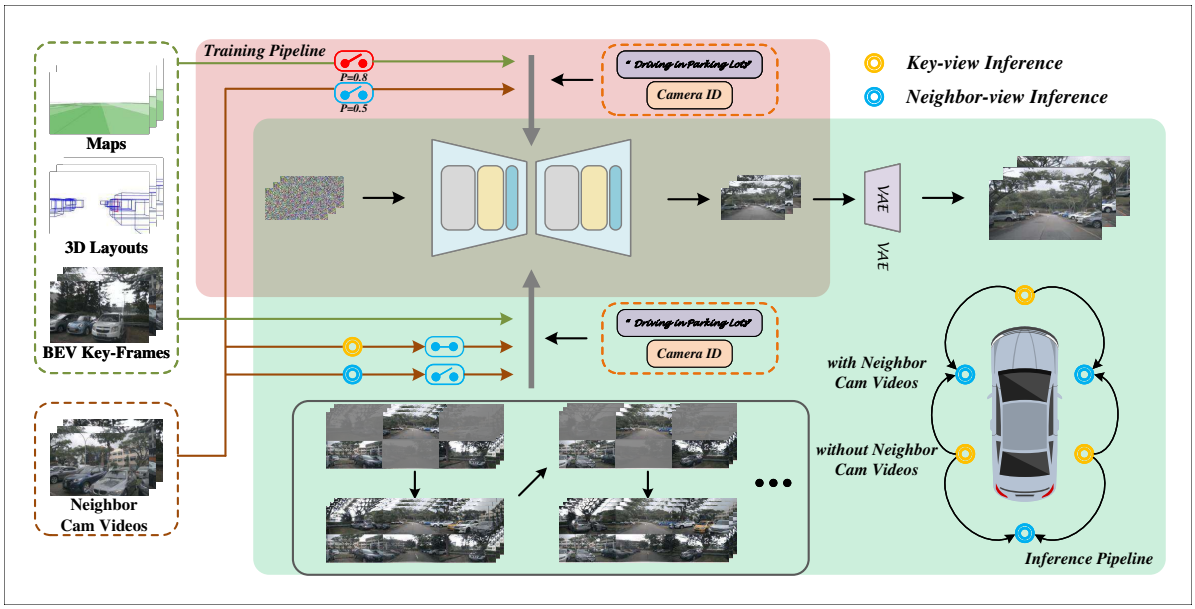


Figure 2. The DriveScape pipeline, built on the LDM framework, generates street-view videos conditioned on scene annotations, BEV maps, and 3D bounding boxes for each view. Keyframe conditions are introduced, in collaboration with temporal and spatial self-attention modules, to ensure consistency across spatial and temporal dimensions. Our unified model enables multi-view video generation without the need for complex post-processing or refinement procedures. Additionally, the Bi-Directional Modulated Transformer (BiMoT) module, consisting of two cross-attention layers with opposite directions and a temporal self-attention layer, ensures effective alignment and integration of various 3D road structural information, enabling precise control over video generation.

all views at the initial frame and the preceding frames of neighboring cameras.

Sparse multi-modal feature fusion. Our model supports training with videos at various frame rates, ranging from 2Hz to 10Hz, while maintaining sample conditions at 2Hz. A lightweight encoder is used to encode the 2Hz map and bounding boxes (BBoxes), with condition-missing frames padded with zeros. Additionally, an extra mask channel is introduced to indicate the presence of conditions at each frame, where mask values are 0 or 1. These inputs are processed through the BiMoT module. Embeddings for unconditioned frames are learned and integrated into the conditions to align with the video frame rate. This diverse frame-rate training scheme, combined with the BiMoT module, enhances sparse condition control capability.

Conditions dropout. To enhance robustness, key views are trained in random order. Neighbor camera videos and conditions are dropped with probabilities of 50% and 20%, respectively. Unlike [16], our model requires no post-refinement, simultaneously learning multi-view and temporal consistency to improve quality and realism.

3.4. Inference Stage

During inference, our pipeline enables efficient and consistent generation of multi-view videos. Similar to the training stage, the process consists of two patterns: key-view inference and neighbor-view inference, as shown in Fig. 2. Key-view camera videos are generated simultaneously without cross-view condition control, as the key-view cameras are intentionally placed non-adjacently. For example, one camera is positioned at the front, while two are at the rear—one on the left and the other on the right. This strategic arrangement ensures comprehensive coverage from multiple perspectives with minimal overlap.

In this configuration, neighbor-view cameras—two at the front and one at the rear—naturally receive cross-view conditional information from their adjacent key-view cameras. This design allows neighbor-view inference to maintain consistency with conditions from neighboring cameras, ensuring cross-view coherence.

Leveraging this setup, our inference pipeline excels in parallel processing, requiring only two consecutive iterations to generate multi-view videos. This approach not only accelerates video production but also ensures cross-view consistency and realism throughout.

3.5. Bi-Directional Modulated Transformer

To integrate multi-conditions from different modalities into the UNet model and ensure effective articulation of each condition, we propose the Bi-Directional Modulated Transformer (BiMoT) module. This module aligns features from various modalities within the latent space, reinforces tem-

poral coherence among the merged conditions, and injects the aligned conditions into the network. As illustrated in Fig. 1, the module consists of two cross-attention layers with opposite directions and a temporal self-attention layer.

Latent to Condition Attention To address diverse conditions with different modalities, we employ a latent-to-condition cross-attention mechanism to standardize feature representations. Pixel-level features are used as signals to merge various condition features. Specifically, let F_{cond} represent condition features and F_{latent} represent video latent features. The output is computed as:

$$F_{cond}^{'} = \text{crossattn}(F_{cond}, F_{latent}), \tag{1}$$

where the query is F_{cond}, and the key and value are F_{latent}. This cross-attention operation facilitates the flow of video latent information into the condition space, providing a unified representation for integrating all modalities.

Temporal Attention. We apply temporal self-attention to ensure that the merged features propagate effectively along the temporal dimension. Sparse condition information is also incorporated into the image latent space, naturally unifying all modalities into a single representation. This enables temporal information from latent features to interact with sparse conditions over time, guiding the generation process with sparse control. This mechanism ensures consistency and precise control over the temporal dimensions.

Condition to Video Attention. Similar to the latent-to-condition cross-attention, a condition-to-video cross-attention mechanism is applied:

$$F_{latent}^{'} = \text{crossattn}(F_{latent}, F_{cond}). \tag{2}$$

This operation facilitates the flow of merged condition feature information into the image latents, ultimately enabling unified control over video generation.

Coarse to Fine Condition Mixture. BiMoT is flexible in accommodating various types of conditions, enabling easy extension to additional inputs. We begin with maps and 3D layouts, then introduce BEV key-frames, and finally integrate neighbor camera videos. This progression allows generation control to transition from a coarse level, driven by maps and 3D layouts, to a finer level, guided by pixel-level details. This ensures both structural coherence and detailed fidelity. The flexibility of BiMoT enables fine-grained layout controllability with unified mono-condition and multi-condition control.

4. Experiments

In this section, we evaluate our model through key metrics, outlining the experimental setup, dataset, and evaluation criteria. We then analyze qualitative results and conduct ablation studies to validate our approach.

4.1. Experiment Details

Dataset. The training data is sourced from the real-world driving dataset nuScenes [6], a widely used benchmark for BEV segmentation and 3D object detection in driving scenarios. Following the official configuration, we use 700 street-view scenes for training and 150 for validation, targeting 8 object classes and 8 road classes.

Evaluation Metrics. To evaluate the effectiveness of our approach in delivering realism, continuity, and precise control over multiple conditions, we employ four key metrics for comparison against existing multi-view image and video generation methods. For realism, we use the Fréchet Inception Distance (FID) [11] to measure image quality. Temporal coherence is assessed using the Fréchet Video Distance (FVD) [28], ensuring consistent and fluid motion in the generated videos.Controllability is evaluated through two perception-based methods: CVT [39] for BEV segmentation and BEVFusion [19] for 3D object detection. Both methods are trained on the nuScenes dataset, providing a reliable benchmark for performance evaluation. Additionally, since our model generates videos, we follow Panacea [33] and use the video-based perception model StreamPETR [29] to evaluate the 3D object detection task.

4.2. Experiments Results

Qualitative Results. We showcase the quality and consistency of our video generation results in Fig. 3. Our model achieves notable consistency across both frames and viewpoints. Specifically, vehicles are accurately positioned according to the provided 3D layouts, and street delineations closely match the supplied projected maps. Remarkably, even when bounding boxes extend beyond the visible area, our model adheres to the specified conditions and generates the remaining objects with high fidelity.

Sparse Condition Control As shown in Fig. 5, our model generates consistent videos in both time and space under sparse conditions, achieving a resolution of 576 $\times$ 1024 and a frame rate of 10 fps. To compare with related works on sparse control [9], we implemented a sparse control model with multi-condition fusion, as existing methods lack support for multi-condition control. A visualization comparing the reproduced SparseCtrl [9] and our method is presented in Fig. 4. Our method clearly demonstrates superior spatial consistency. Moreover, the FID and FVD scores for the reproduced SparseCtrl are 11.15 and 116.00, respectively, significantly worse than those achieved by our model (8.34 and 76.39). For additional visual comparisons and details on the multi-condition fusion SparseCtrl, please refer to the supplementary materials and the accompanying web page.

Quantitative Results. We report quantitative metrics on the nuScenes validation set in Table 2. Using FID and FVD, we evaluate image and video quality, respectively. Following MagicDrive [7] and DrivingDiffusion [16], we

Method	FID↓	FVD↓	Road mIoU ↑	Vehicle mIoU ↑	Object NDS ↑
Oracle	-	-	73.68	34.84	41.21
BEVGen[25]	25.54	-	50.20 (-23.48)	5.89 (-28.95)	-
BEVControl [35]	24.85	-	60.80 (-12.88)	26.80 (-8.04)	-
MagicDrive [7]	16.20	-	61.05 (-12.63)	27.01 (-7.83)	23.32 (-17.89)
DriveDreamer [31]	52.60	452.0	-	-	-
DrivingDiffusion [16]	15.83	332.0	63.20 (-10.48)	31.60 (**-3.24**)	33.1 (-8.11)
Drive-WM [32]	15.80	122.70	65.07 (-8.61)	27.19 (-7.65)	-
Delphi [21]	15.08	113.5	-	-	36.58 (-4.63)
SparseCtrl[9]	11.15	116.00	-	-	-
Ours	**8.34**	**76.39**	**68.21 (-5.47)**	30.27 (-4.57)	36.5 (**-4.71**)

Table 2. The comparative analysis of our method against others on the nuScenes validation dataset is detailed for each task. Our end-to-end approach outperformed most baseline models without requiring any post-refinement process.

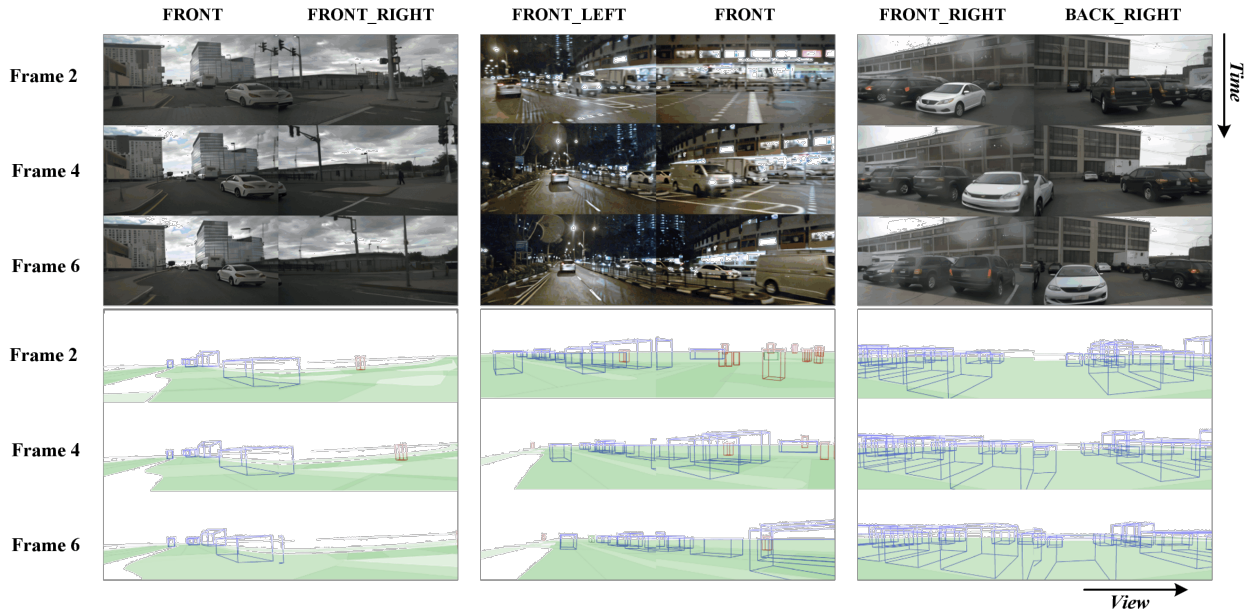


Figure 3. Showcases for multi-view video generation by DriveScape demonstrate frame-to-frame and viewpoint consistency. Vehicles generated by our model are accurately positioned according to the provided 3D layouts, while the street depictions closely align with the projected maps.

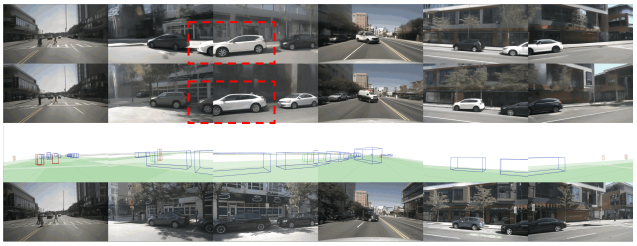


Figure 4. Results in each row are Ours, SparseCtrl, input condition and GT respectively, which illustrate that our method has better spatial consistency than SparseCtrl.

generate data from validation set labels and assess it with a pre-trained perception model. The decline in Road mIoU and Vehicle mIoU, compared to original data, measures the effectiveness of our framework in producing perception-friendly synthetic data.

To align with previous studies, we resized generated images to 224×400. We compared our FID results with three image generation models: BEVGen [25], BEVControl [35], and MagicDrive [7]. Our method significantly outperformed these models. Notably, video generation models typically yield higher FID scores than image models. For instance, Drive-WM, using only an image model, achieved better FID than the full model, highlighting the challenges of video generation.

For video quality, assessed via the FVD metric, our approach showed substantial improvements over recent works

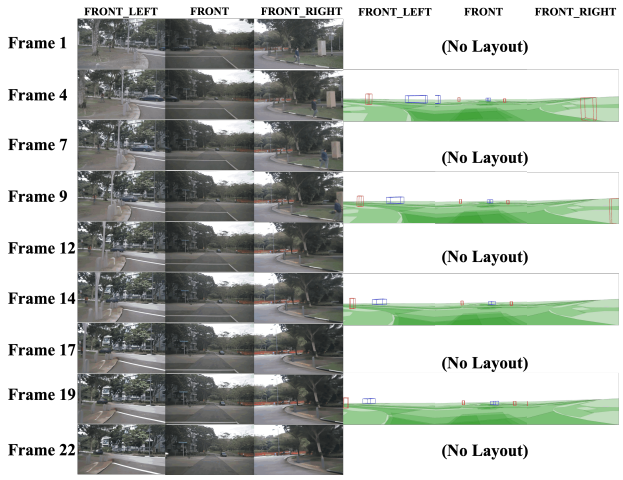


Figure 5. Showcase for sparse Condition Control.

such as DriveDreamer [31], DrivingDiffusion [16], and Drive-WM [32]. In BEV segmentation metrics, our method achieved superior perceptual performance, except for a slightly lower Road mIoU compared to DrivingDiffusion, which utilized a post-processing network for refinement. To maintain simplicity, our end-to-end system applies no post-processing beyond image resizing for perceptual evaluation. The perception models we used operate on single-image inference, not leveraging the temporal consistency of our generated videos. Given our significant lead in FVD, our method holds strong potential for training video BEV perception models.

data	NDS ↑	FID ↓	FVD ↓
Oracle	46.9	-	-
Panacea	32.1 (-14.8)	16.96	139
Oracle	53.7	-	-
Ours	**44.5** (-9.2)	**8.34**	**76.39**

Table 3. Compared with Panacea, we evaluate results on the nuScenes validation set using the pre-trained StreamPETR model.

4.3. Ablation Studies

To refine and evaluate the efficacy of our proposed framework, we conducted ablation studies, as detailed in Table 4. Various strategies for implementing conditional guidance were tested. Removing the entire BiMoT module resulted in the FID increasing from 8.34 to 20.98 and the FVD rising from 76.39 to 142.81, underscoring the module's importance for multi-condition encoding. Similarly, removing the temporal attention layer within BiMoT caused a significant performance drop, with FID and FVD increasing to 20.67 and 123.03, respectively, highlighting the role of temporal modelling in BiMoT.Additionally, during testing, we separately removed the key-frame condition and the neighbor condition. Both led to a performance decline, indicating the importance of these conditions for facilitating information exchange between views.

TeAttn	BiMOT	KF	Ne	FID	FVD
	✓	✓	✓	20.98	142.81
		✓	✓	20.67	123.03
✓	✓		✓	11.60	108.80
✓	✓	✓		12.18	111.36
✓	✓	✓	✓	**8.34**	**76.39**

Table 4. Ablation Study. TeAttn is temporal attention module in BiMOT, KF is key-frame condition and Ne is neighbor condition

4.4. Visualization of Different Settings

In our visual showcases, we demonstrate the model's proficiency in manipulating video predictions for a single scene under various control conditions. The control conditions, sourced from frames of identical perspectives but randomly selected from other scenes, test the decoupling of conditions and the robustness of our method across diverse scenarios.

As shown in Fig. 7, the experiments explore the influence of different 3D layout controls. Using a single reference frame from a highway segment, we first depict vehicles driving normally, then transition to vehicles traversing grassy terrain. Remarkably, our model maintains temporal consistency even in this unconventional context. Additional experiments involve removing or adding large vehicles, with results aligning closely with the specified conditions, showcasing precise control.

Furthermore, as illustrated in Fig. 8, we investigate the impact of different maps on video generation. The generated scenes align impressively with the shapes of the provided map regions, maintaining layout feature consistency despite map changes. These experiments provide strong evidence of our model's capability to render specified control conditions into precise and coherent visual narratives within the generated video content.

5. Conclusions

The paper introduces DriveScape, the first end-to-end multi-view driving video generative model. It enables consistent multi-view video generation without complex post-processing or refinement. This unified framework streamlines multi-view condition-guided video generation. We also propose a Bi-Directional Modulated Transformer module for diverse condition guidance. Our approach achieves decoupled and precise control of dynamic foreground and static background, delivering state-of-the-art video synthesis on the nuScenes dataset. It excels in multi-view image and video generation quality and perception applications. In summary, our method outperforms others in generation quality and ease of application.

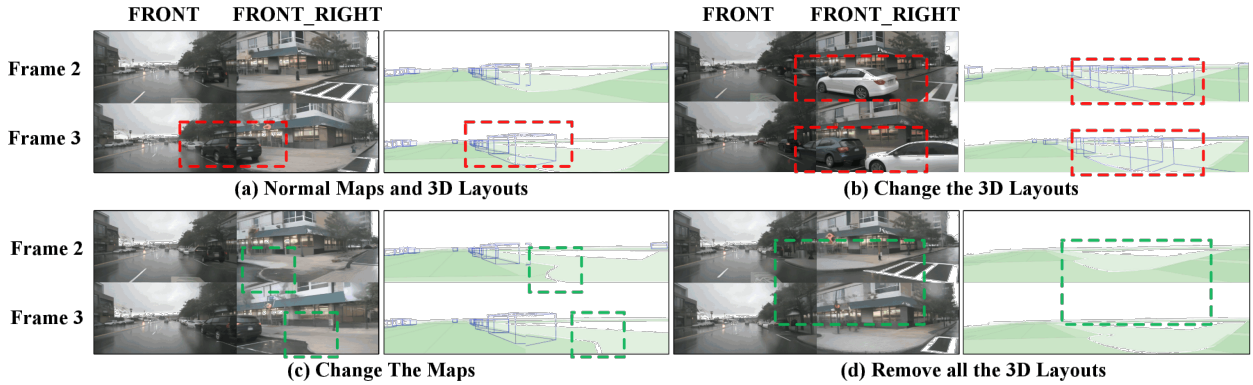


Figure 6. Showcase for maps and 3D layouts control under different conditions. (a): control by the reference maps and 3D Layouts; (b) change 3D Layouts with more vehicles; (c) change the maps with curved pedestrian crosswalk; (d) remove all the 3D Layouts.

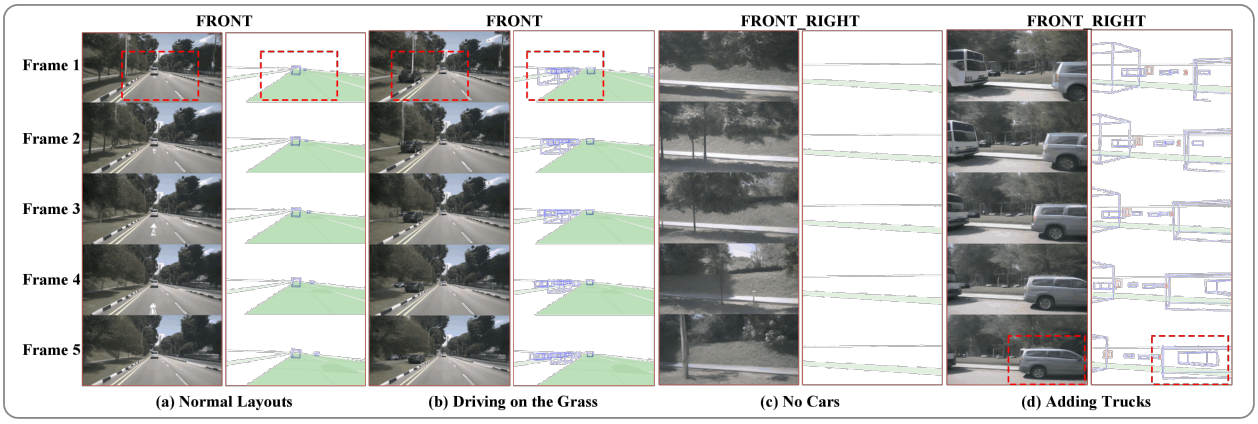


Figure 7. Showcases for 3D layouts control under different conditions. DriveScape can precisely control dynamic objects. (a) reference layouts; (b) move car to the grass; (c) remove all cars; (d) add trucks to the scene.

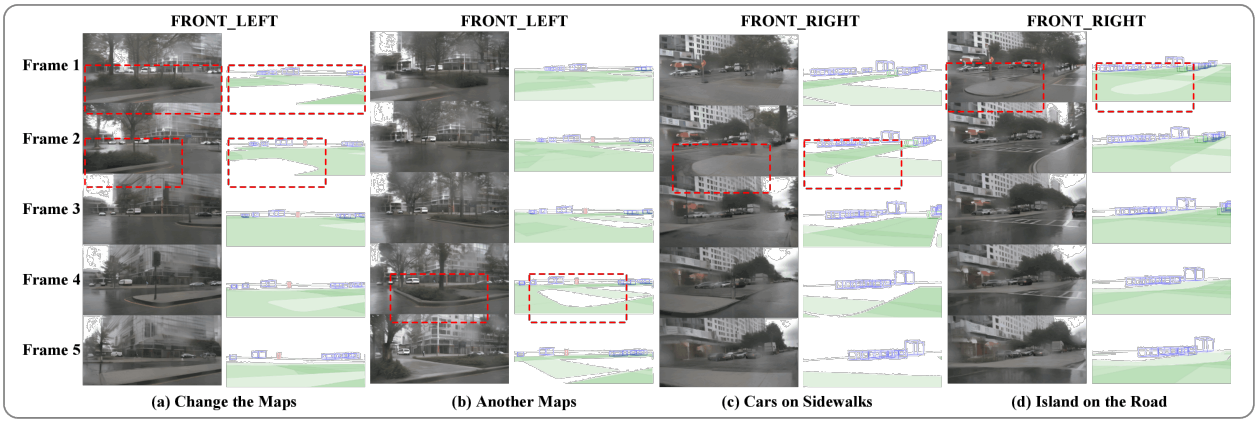


Figure 8. Showcases for maps control under. DriveScape can precisely control static background. (a) replace the driable area with green belts; (b) replace the driable area with green belts; (c) replace the driable area with sidewalks; (d) add island on the road.

References

[1] Jie An, Songyang Zhang, Harry Yang, Sonal Gupta, Jia-Bin Huang, Jiebo Luo, and Xi Yin. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. *arXiv preprint arXiv:2304.08477*, 2023. 2

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. 2

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2

[4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 2

[5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 2

[6] 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*, pages 11621–11631, 2020. 5

[7] Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. *arXiv preprint arXiv:2310.02601*, 2023. 1, 2, 5, 6

[8] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2

[9] Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *European Conference on Computer Vision*, pages 330–348. Springer, 2025. 5, 6

[10] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022. 2

[11] 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. 5

[12] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2

[13] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Goal-oriented autonomous driving. *arXiv preprint arXiv:2212.10156*, 2022. 1

[14] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022. 1

[15] 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. 1

[16] Xiaofan Li, Yifu Zhang, and Xiaoqing Ye. Drivingdiffusion: Layout-guided multi-view driving scene video generation with latent diffusion model. *arXiv preprint arXiv:2310.07771*, 2023. 2, 3, 4, 5, 6

[17] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1477–1485, 2023. 1

[18] 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*, pages 1–18. Springer, 2022. 1

[19] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. *Advances in Neural Information Processing Systems*, 35:10421–10434, 2022. 5

[20] Haicheng Liao, Yongkang Li, Zhenning Li, Chengyue Wang, Zhiyong Cui, Shengbo Eben Li, and Chengzhong Xu. A cognitive-based trajectory prediction approach for autonomous driving. *arXiv preprint arXiv:2402.19251*, 2024. 1

[21] 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. 6

[22] 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*, pages 10684–10695, 2022. 2

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 3

[24] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data, 2022. 2

[25] Alexander Swerdlow, Runsheng Xu, and Bolei Zhou. Street-view image generation from a bird's-eye view layout. *IEEE Robotics and Automation Letters*, 2024. 1, 2, 6

[26] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion, 2023. 2

[27] Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16773–16783, 2023. 2

[28] 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. 5

[29] 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. 5

[30] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation, 2023. 2

[31] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*, 2023. 6

[32] 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. *arXiv preprint arXiv:2311.17918*, 2023. 6

[33] 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*, pages 6902–6912, 2024. 5

[34] 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*, pages 17830–17839, 2023. 1

[35] 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*, 2023. 1, 2, 6

[36] Yuxiang Yang, Yingqi Deng, Jing Zhang, Jiahao Nie, and Zheng-Jun Zha. Bevtrack: A simple and strong baseline for 3d single object tracking in bird's-eye view. *arXiv e-prints*, pages arXiv–2309, 2023. 1

[37] David Junhao Zhang, Dongxu Li, Hung Le, Mike Zheng Shou, Caiming Xiong, and Doyen Sahoo. Moonshot: Towards controllable video generation and editing with multimodal conditions, 2024. 2

[38] Wenzhao Zheng, Ruiqi Song, Xianda Guo, and Long Chen. Genad: Generative end-to-end autonomous driving. *arXiv preprint arXiv:2402.11502*, 2024. 1

[39] 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*, pages 13760–13769, 2022. 1, 5

[40] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models, 2023. 2