DriveDreamer-2: LLM-Enhanced World Models for Diverse Driving Video Generation

Guosheng Zhao[1,2,3*], Xiaofeng Wang[1,2,3*], Zheng Zhu[4*†], Xinze Chen[4], Guan Huang[4], Xiaoyi Bao[1,2,3], Xingang Wang[2,3†]

[1]School of Artificial Intelligence, University of Chinese Academy of Sciences
[2]Institute of Automation, Chinese Academy of Sciences
[3]Luoyang Institute for Robot and Intelligent Equipment
[4]GigaAI
{zhaoguosheng2021, wangxiaofeng2020, baoxiaoyi2021, xingang.wang}@ia.ac.cn,
zhengzhu@ieee.org, 501177639@qq.com, huangg22@mails.tsinghua.edu.cn

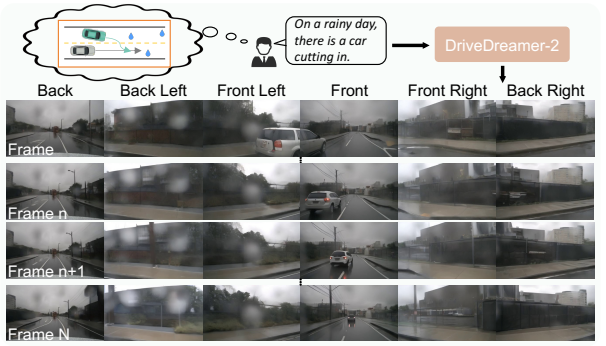
Abstract

World models have demonstrated superiority in autonomous driving, particularly in the generation of multi-view driving videos. However, significant challenges still exist in generating customized driving videos. In this paper, we propose *DriveDreamer-2*, which incorporates a Large Language Model (LLM) to facilitate the creation of user-defined driving videos. Specifically, a trajectory generation function library is developed to produce trajectories that conform to user descriptions. Subsequently, an HDMap generator is designed to learn the mapping from trajectories to road structures. Ultimately, we propose the Unified Multi-View Model (UniMVM) to enhance temporal and spatial coherence in the generated multi-view driving videos. To the best of our knowledge, *DriveDreamer-2* is the first world model to generate customized driving videos, and it can generate uncommon driving videos (e.g., vehicles abruptly cut in) in a user-friendly manner. Besides, experimental results demonstrate that the generated videos enhance the training of driving perception methods (e.g., 3D detection and tracking). Furthermore, video generation quality of *DriveDreamer-2* surpasses other state-of-the-art methods, showcasing FID and FVD scores of 11.2 and 55.7, representing relative improvements of ~30% and ~50%.

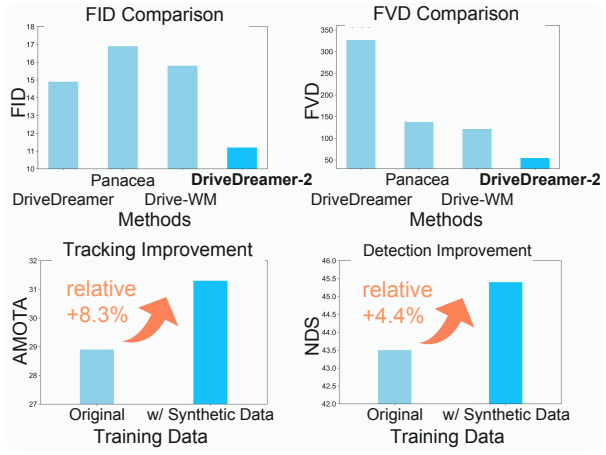
Project Page — https://drivedreamer2.github.io

Introduction

World models have seen widespread success across various applications (Hafner et al. 2019, 2020, 2023; Kim et al. 2020; Ha and Schmidhuber 2018; Seo et al. 2023; Hafner et al. 2023, 2022; Wu et al. 2023; Lin et al. 2023; Pan et al. 2022; Chen et al. 2023). In particular, world models for autonomous driving (Hu et al. 2023; Jia et al. 2023; Wang et al. 2023c,e; Yang et al. 2024; Gao et al. 2024) have recently attracted substantial interest from both industry and academia.



(a) User-customized driving video generation.



(b) Generated video quality comparison and improvement in the downstream task.

Figure 1: *DriveDreamer-2* excels in generating multi-view driving videos from user descriptions, enhancing synthetic data diversity. Its generation quality surpasses state-of-the-art methods, significantly improving downstream tasks.

Benefiting from their excellent predictive capabilities, autonomous driving world models facilitate the generation of

*These authors contributed equally.
†Corresponding author.

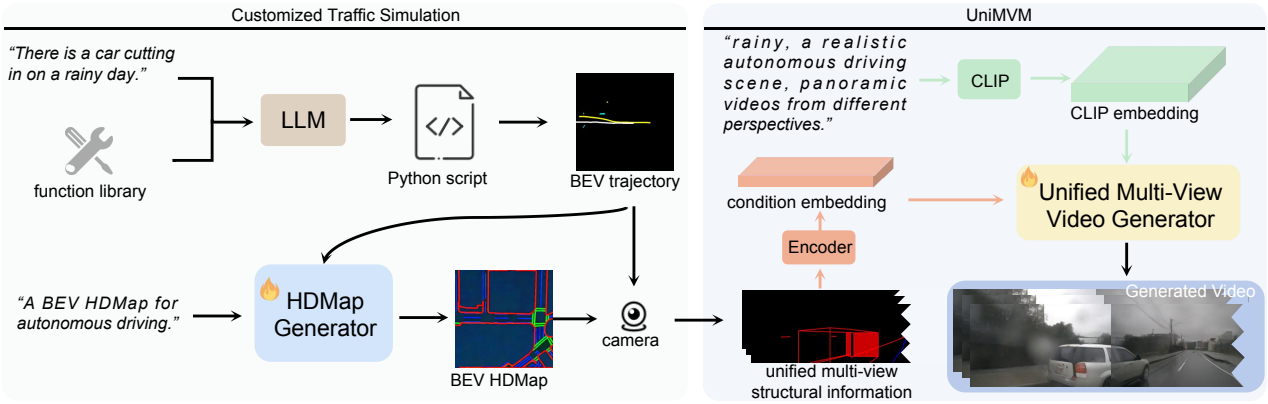


Figure 2: The overall framework of *DriveDreamer-2* involves initially generating agent trajectories according to the user query, followed by producing a realistic HDMap, and finally generating multi-view driving videos.

diverse driving videos, encompassing even long-tail scenarios. Generated videos can significantly enhance the training of driving perception systems, offering substantial benefits for practical autonomous driving applications.

World modeling in autonomous driving is challenging due to its complexity and vast sampling space. Early approaches (Hu et al. 2022; Gao et al. 2022) mitigate these problems by incorporating world modeling within the Bird's Eye View (BEV) semantic segmentation space. However, these methods are mainly focused on simulated environments. In the recent evolution of autonomous driving technologies, there has been a substantial leap forward in the development of world models, driven by significant advances in the field of video generation (Ranzato et al. 2014; Srivastava, Mansimov, and Salakhudinov 2015; Kalchbrenner et al. 2017; Weissenborn, Täckström, and Uszkoreit 2019; Wang et al. 2024b; Hong et al. 2022; Villegas et al. 2023; Kondratyuk et al. 2023; Mathieu, Couprie, and LeCun 2015; Vondrick, Pirsiavash, and Torralba 2016; Denton and Fergus 2018; Hsieh et al. 2018; Kumar et al. 2019). This progress has been propelled by the utilization of cutting-edge diffusion models (Dhariwal and Nichol 2021; Ho, Jain, and Abbeel 2020; Ho et al. 2022b; Nichol et al. 2021; Nichol and Dhariwal 2021; Rombach et al. 2022; Harvey et al. 2022; Ho et al. 2022a; Yang, Srivastava, and Mandt 2022; Singer et al. 2022; Khachatryan et al. 2023; Wang et al. 2023b), exemplified by notable contributions such as DriveDreamer (Wang et al. 2023c), Drive-WM (Wang et al. 2023e), MagicDrive (Gao et al. 2023), Panacea (Wen et al. 2023), GAIA-1 (Hu et al. 2023), and the integration of Large Language Models (LLM) like ADriver-I (Jia et al. 2023). These sophisticated models have played a pivotal role in pushing the boundaries of world modeling capabilities, allowing researchers and engineers to explore more complex and realistic driving scenarios. However, it is important to note that a majority of these methods rely heavily on structured information (e.g. 3D boxes, HDMaps, and optical flow) or real-world image frames. This dependence not only constrains interactivity, but also limits the diversity of generated videos.

To address these challenges, we introduce *DriveDreamer-2*, the first world model to generate diverse driving videos in a user-friendly manner. Specifically, the traffic simulation task has been disentangled into the generation of foreground conditions (trajectories of the ego-car and other agents) and background conditions (HDMaps of lane boundary, lane divider, and pedestrian crossing). For foreground generation, a functional library is constructed to generate agent trajectories based on user text input, employing an LLM (Cho et al. 2024; Wang et al. 2024a; Ma et al. 2023; Mao et al. 2023b) as the interface. For background conditions, we propose the HDMap generator that employs a diffusion model to simulate road structures. In this process, the previously generated agent trajectories are involved as conditional inputs, which allows the HDMap generator to learn the associations between foreground and background conditions in driving scenes. Building upon generated traffic conditions, the Unified Multi-view Video Model (UniMVM) is proposed to generate multi-view driving videos, which is designed to unify both intra- and cross-view spatial consistency, enhancing the overall temporal and spatial coherence.

Extensive experiment results show that *DriveDreamer-2* is capable of producing diverse user-customized videos, including uncommon scenarios where vehicles abruptly cut in (depicted in Fig. 1). Besides, *DriveDreamer-2* can generate high-quality driving videos with an FID of 11.2 and FVD of 55.7, relatively improving previous best-performing methods by ∼30% and ∼50%. Furthermore, experiments verify that *DriveDreamer-2*-generated driving videos can enhance the training of autonomous driving perception methods, improving detection by 4% and tracking by 8%. The main contributions of this paper are summarized as follows:

- We present *DriveDreamer-2*, the first world model to generate diverse driving videos user-friendly.
- We propose a traffic simulation pipeline employing only text prompts as input, which can be utilized to generate diverse traffic conditions for driving video generation.
- UniMVM is presented to seamlessly integrate intra- and cross-view spatial consistency, elevating the overall temporal and spatial coherence in generated driving videos.

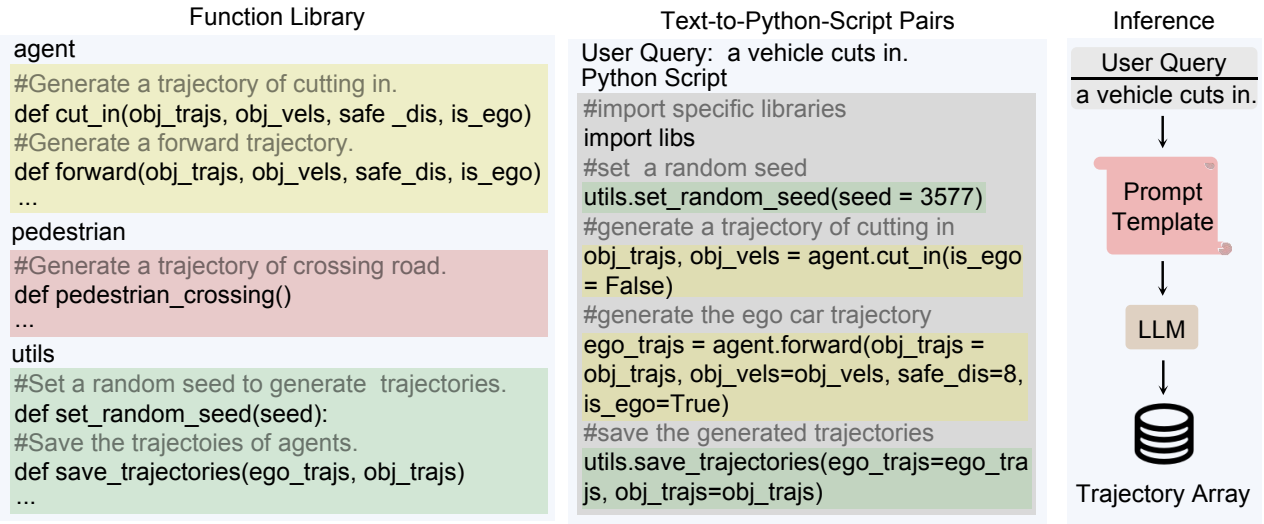


Figure 3: The overview of customized trajectory generation, which involves using a function library to create Text-to-Python-Script pairs, fine-tuning an LLM with the resulting dataset, and generating trajectories in response to user queries.

- Extensive experiments show that *DriveDreamer-2* crafts diverse, customized driving videos, and improves FID by ∼30% and FVD by ∼50% over SOTA methods and enhances the training of various driving perception models.

DriveDreamer-2

Fig. 2 illustrates the overall framework of *DriveDreamer-2*. It starts with a customized traffic simulation to generate foreground agent trajectories and background HDMaps. A finetuned LLM translates user prompts into agent trajectories, which condition the HDMap generator to simulate road structures. Building on these traffic conditions, UniMVM unifies intra- and cross-view spatial consistency, enhancing the temporal and spatial coherence of the generated videos. In the subsequent sections, we delve into the details of the customized traffic simulation and the UniMVM framework.

Customized Traffic Simulation

Driving simulators stand as a cornerstone in self-driving development, aiming to offer a controlled environment to mimic real-world conditions. Previous traffic simulation methods (Tan et al. 2023; Feng et al. 2023; Zhong et al. 2023b,a) have focused on how vehicles move within a given HDMap. Unlike these methods, we propose a customized traffic simulation pipeline that generates corresponding trajectories and HDMaps based on user input. Specifically a trajectory-generation function library is designed to finetune the LLM, which facilitates transferring user prompts into diverse agent trajectories, encompassing maneuvers such as cut-in and U-turn. Additionally, the pipeline incorporates the HDMap generator to simulate the background road structures. During this phase, the previously generated agent trajectories serve as conditional inputs, ensuring that the resulting HDMap adheres to traffic constraints. In the following, we elaborate on the finetuning process of the LLM and the framework of the HDMap generator.

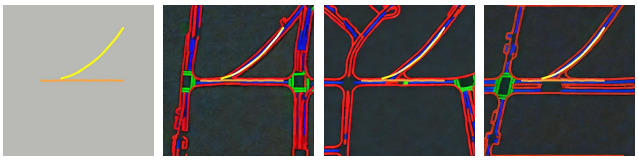


Figure 4: The proposed HDMap generator can generate diverse BEV HDMaps based on the same BEV trajectory input. The orange and yellow colors represent the motion trajectories of the ego car and other vehicles, respectively. Road boundaries are indicated in red, lane dividers are shown in blue, and pedestrian crossings are highlighted in green.

Finetuning LLM for Trajectory Generation Previous traffic simulation methods (Mao et al. 2023a; Zhong et al. 2023a,b) necessitate the intricate specification of parameters, involving details such as the agent's speed, position, acceleration, and mission goal. To simplify this intricate process, we design a function library and incorporate an LLM to enable the efficient transformation of user-friendly language inputs into comprehensive traffic simulation scenarios. As depicted in Fig. 3, the constructed function library encompasses 18 functions, including agent functions (steering, constant speed, acceleration, and braking), pedestrian functions (walking direction and speed), and other utility functions such as saving trajectories. Building upon these functions, Text-to-Python-Scripts pairs are manually curated for finetuning LLM (GPT-3.5). The scripts include a range of fundamental scenarios such as lane-changing, overtaking, following other vehicles, and executing U-turns. Additionally, we encompass more uncommon scenarios like pedestrians abruptly crossing, and vehicles cutting into the lane. Taking the user input *a vehicle cuts in* as an example, the corresponding script involves the following steps: initially

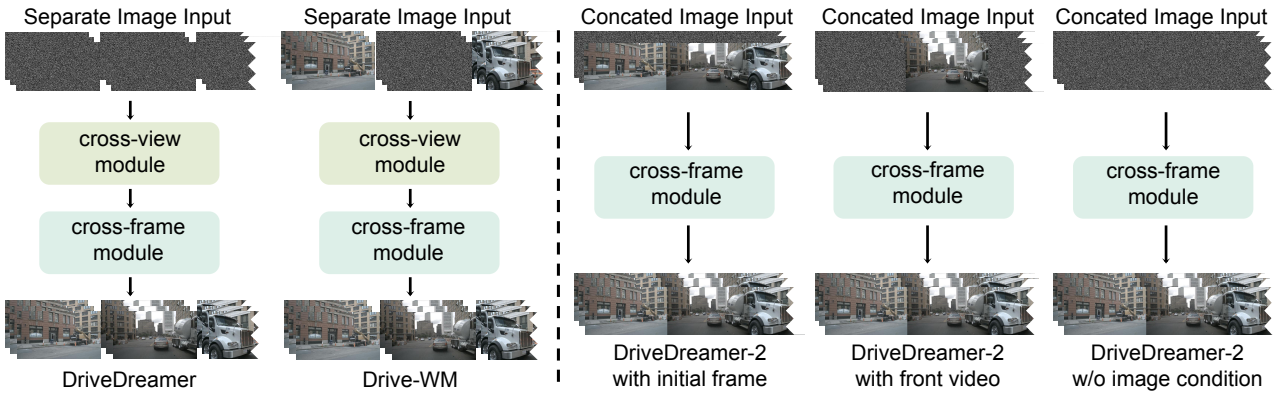


Figure 5: The comparison of multi-view video generation paradigms. All structural conditions and text prompts are omitted here to emphasize the distinctions between our UniMVM and previous methods. By adjusting the mask, UniMVM can generate videos conditioned on the initial frame, front view video, and without image input.

generating a trajectory of cutting in (`agent.cut_in()`), followed by generating the corresponding ego car trajectory (`agent.forward()`), and ultimately utilizing the saving function from utilities to directly output the trajectory of the ego-car and other agents in array format. In the inference phase, we follow (Mao et al. 2023a) to expand prompt inputs to a pre-defined template, and the finetuned LLM can directly output the trajectory array.

HDMap Generation A comprehensive traffic simulation not only entails the trajectories of foreground agents but also necessitates the generation of background HDMap elements such as lanes and pedestrian crosswalks. Therefore, the HDMap generator is proposed to ensure the background elements do not conflict with the foreground trajectories. In the HDMap generator, we formulate the background elements generation as a conditional image generation problem, where the conditional input is the BEV trajectory map $\mathcal{T}_b \in \mathcal{R}^{3\times H_b\times W_b}$, and the target is the BEV HDMap $\mathcal{H}_b \in \mathcal{R}^{3\times H_b\times W_b}$. Different from previous conditional image generation approaches (Zhang, Rao, and Agrawala 2023; Li et al. 2023) that predominantly rely on outline conditions (edges, depths, boxes, segmentation maps), the proposed HDMap generator explores the correlations between the foreground and background traffic elements. Specifically, the HDMap generator is constructed upon an image-generation diffusion model. To train the generator, we curate a trajectory-to-HDMap dataset $\mathcal{D} = \{\mathcal{T}_b, \mathcal{H}_b\}$. In the trajectory map, distinct colors are assigned to represent different agent categories. Meanwhile, the target HDMap comprises three channels, representing lane boundaries, lane dividers, and pedestrian crossings, respectively. Within the HDMap generator, we employ stacks of 2D convolution layers to incorporate the trajectory map condition. The resulting feature maps $C_{\mathcal{T}}$ are then seamlessly integrated into the diffusion model using (Zhang, Rao, and Agrawala 2023). In the training stage, the diffusion forward process gradually adds noise ϵ to the latent feature $\mathcal{Z}_0$, resulting in the noisy latent feature $\mathcal{Z}_{T_b}$. Then we train ϵ_θ to predict the noise we added, and the HDMap generator ϕ is optimized via

$$\min_{\phi} \mathcal{L} = \mathbb{E}_{\mathcal{Z}_0,\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t,c}\left[\|\epsilon - \epsilon_\theta(\mathcal{Z}_t, t, c)\|_2^2\right], \quad (1)$$

where time step t is uniformly sampled from $[1, T_b]$. As shown in Fig. 4, leveraging the proposed HDMap generator allows us to generate diverse HDMaps based on the same trajectory conditions. It is noteworthy that the generated HDMaps not only adhere to traffic constraints (lane boundaries positioned on either side of lane dividers, and pedestrian crossings at intersections) but also seamlessly integrate with trajectories.

UniMVM

Utilizing structured information generated by the customized traffic simulation, multi-view driving videos can be generated via the framework of DriveDreamer (Wang et al. 2023c). However, the view-wise attention introduced in previous methods (Wang et al. 2023c; Yang et al. 2023) can not guarantee multi-view consistency. To address this, (Wang et al. 2023e; Wen et al. 2023; Li, Zhang, and Ye 2023) employ image or video conditions to improve consistency across different views, but at the cost of reduced generation efficiency and diversity. In *DriveDreamer-2*, we introduce UniMVM, a novel framework that unifies multi-view video generation both with and without adjacent view conditions. This ensures temporal and spatial coherence while maintaining generation speed and diversity.

Formulation In multi-view video dataset p_{data}, $\mathrm{x} \in \mathcal{R}^{K\times T\times 3\times H\times W}$ is a sequence of T images with K views, with height H and width W. Let x_i denote the sample of i-th view, then the multi-view video joint distribution $p(\mathrm{x}_1, ..., \mathrm{x}_\mathrm{K})$ can be obtained by (Wang et al. 2023e):

$$p(\mathrm{x}_1, ..., \mathrm{x}_\mathrm{K}) = p(\mathrm{x}_1)p(\mathrm{x}_2|\mathrm{x}_1)...p\left(\mathrm{x}_K|\,\mathrm{x}_1..., \mathrm{x}_{K-1}\right). \quad (2)$$

Eq. 2 indicates that adjacent view videos can be expanded with multiple generation steps, which is inefficient. In the proposed UniMVM, we draw inspiration from the Eq. 2 to expand the view. However, unlike Drive-WM (Wang et al. 2023e) which requires the independent generation of

Method	Conditions	FID↓	FVD↓
DriveDreamer (Wang et al. 2023c)	-	26.8	353.2
DriveDreamer-2	-	**25.0**	**105.1**
Drive-WM (Wang et al. 2023e)	3-view videos†	**15.8**	122.7
DriveDreamer-2	1-view video	18.4	**74.9**
DriveDreamer (Wang et al. 2023c)	1st-frame multi-view image	14.9	340.8
DrivingDiffusion (Li, Zhang, and Ye 2023)	1st-frame multi-view image†	15.8	332.0
Panacea (Wen et al. 2023)	1st-frame multi-view image†	16.9	139.0
DriveDreamer-2	1st-frame multi-view image	**11.2**	**55.7**

Table 1: Comparison of the generation quality on nuScenes validation set. † denotes that the conditions are generated.

views, UniMVM unifies multiple views as a complete patch. Specifically, we concatenate the multi-view video in the order of {FL, F, FR, BR, B, BL}[1] to obtain the spatially unified image $\mathrm{x}' \in \mathcal{R}^{T\times 3\times H\times KW}$. Then we can obtain the multi-view driving video distribution $p(\mathrm{x}')$:

$$\begin{aligned} p(\mathrm{x}') &= p(\mathrm{x}'\cdot(1-m), \mathrm{x}'\cdot m) \\ &= p(\mathrm{x}'\cdot m)p(\mathrm{x}'\cdot(1-m)|\mathrm{x}'\cdot m), \end{aligned} \tag{3}$$

where m represents the mask of one of the all views. As shown in Fig. 5, we compare the paradigm of UniMVM with that of DriveDreamer and Drive-WM. In contrast to these counterparts, UniMVM unifies multiple views into a complete patch for video generation without introducing cross-view parameters. Furthermore, various driving video generation tasks can be accomplished via adjusting the mask m. Specifically, when m is set to mask future $T-1$ frames, UniMVM enables future video prediction based on the input of the first frame. Configuring m to mask {FL, FR, BR, B, BL} views empowers UniMVM to achieve multi-view video outpainting, leveraging a front-view video input. Furthermore, UniMVM can generate multi-view videos when m is set to mask all video frames, and experiments confirm its ability to produce temporally and spatially coherent videos with enhanced efficiency and diversity.

Video Generation Based on the UniMVM formulation, our approach first unifies the traffic conditions, which results in sequences of HDMaps $\{\mathcal{H}_i\}_{i=0}^{N-1} \in \mathcal{R}^{N\times 3\times H\times KW}$ and 3D boxes $\{\mathcal{B}_i\}_{i=0}^{N-1} \in \mathcal{R}^{N\times C\times H\times KW}$ (N is the frame number of video clip, and C is the category number). Note that sequences of 3D boxes can be derived from agent trajectories, and the sizes of 3D boxes are determined based on the respective agent category. Unlike the original DriveDreamer, *DriveDreamer-2* projects 3D boxes directly onto the image plane as control conditions, eliminating the need for position and category embeddings. We adopt three encoders to embed HDMaps, 3D boxes, and image frames into latent space features $y_{\mathcal{H}}$, $y_{\mathcal{B}}$, and $y_{\mathcal{I}}$. Then we concatenate the spatially aligned conditions $y_{\mathcal{H}}$, $y_{\mathcal{B}}$ with $\mathcal{Z}_t$ to obtain the input $\mathcal{Z}_{in}$, where $\mathcal{Z}_t$ is the noisy latent feature generated from $y_{\mathcal{I}}$ by the diffusion process. For video generator training, all parameters are optimized via denoising score matching (Karras et al. 2022).

[1]F: Front, L: Left, R: Right, B: Back.

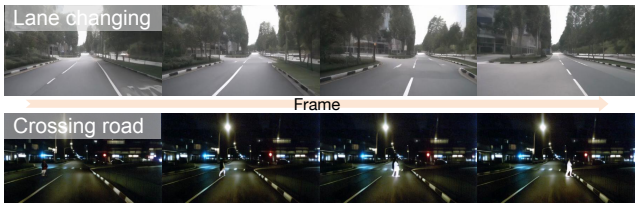


Figure 6: User-customized driving videos generated by *DriveDreamer-2*. The top row depicts a scene where the ego car changes lanes, while the bottom row shows an unexpected pedestrian crossing the road at night.

Initial frame	Real	Generated	mAP↑	mAOE↓	mAVE↓	NDS↑
-	✓	-	31.7	67.9	33.0	43.5
✓	✓	✓	32.6	61.7	**29.7**	45.2
-	✓	✓	**32.9**	**61.5**	30.4	**45.4**

Table 2: Comparison involving data augmentation using synthetic data on 3D object detection.

Experiment

Experiment Details

Dataset. The training dataset is derived from the nuScenes dataset (Caesar et al. 2019), comprising 700 training videos and 150 validation videos. Each video contains approximately 20 seconds of footage captured by six surround-view cameras at a frame rate of 12Hz, resulting in around 1 million frames for training. Following (Wang et al. 2023c,d), we preprocess the nuScenes dataset to calculate 12Hz annotations. HDMaps and 3D boxes are transformed to BEV perspective to train HDMap generation, and these annotations are projected to pixel coordinate to train video generation.

Training. For agent trajectory generation, we employ GPT-3.5 as the LLM and finetune it utilizing a text-to-script dataset to specialize in trajectory generation knowledge. The proposed HDMap generator is built upon SD2.1 (Rombach et al. 2022) with the ControlNet parameters (Zhang, Rao, and Agrawala 2023) being trainable. It is trained for 55K iterations with a batch size of 24 at a resolution of 512×512. The video generator leverages SVD (Blattmann et al. 2023) for its robust video generation capabilities, with all parameters finetuned. During UniMVM training, the model under-

Initial frame	Real	Generated	AMOTA↑	AMOTP↓	IDS↓
-	✓	-	28.9	1.419	687
✓	✓	✓	31.2	1.396	**542**
-	✓	✓	**31.3**	**1.387**	593

Table 3: Comparison involving data augmentation using synthetic data on multi-object tracking.

Method	Cross-view module	UniMVM	FID↓	FVD↓
DriveDreamer (Wang et al. 2023c)	✓	-	14.9	340.8
DriveDreamer-2	✓	-	17.2	94.6
	-	✓	**11.2**	**55.7**

Table 4: The ablation study on the backbone and UniMVM. Cross-view module denotes the cross-view attention used in previous methods.

went 200K iterations with a batch size of 1, an $N = 8$ frame length, $K = 6$ views, and a spatial size of 256×448. All the experiments are conducted on NVIDIA A800 (80GB) GPUs, and we use the AdamW optimizer (Kingma and Ba 2014) with a learning rate 5×10^{-5}.

Evaluation. Extensive experiments are conducted to assess *DriveDreamer-2*. For qualitative experiments, we visualize customized driving video generation to validate that *DriveDreamer-2* can produce diverse driving videos in a user-friendly manner. Additionally, visualization comparisons are conducted between UniMVM and other generative paradigms to demonstrate that *DriveDreamer-2* excels in generating temporally and spatially coherent videos. For quantitative experiments, the frame-wise Fréchet Inception Distance (FID) (Parmar, Zhang, and Zhu 2022) and Fréchet Video Distance (FVD) (Unterthiner et al. 2018) are utilized as metrics. Besides, StreamPETR (Wang et al. 2023a), building upon a ResNet-50 (He et al. 2016) backbone, is trained at the same resolution of 256×448 to evaluate the improvements of 3D object detection and multi-object tracking achieved by our generated results.

User-Customized Driving Video Generation

DriveDreamer-2 offers a user-friendly interface for generating driving videos. As depicted in Fig. 1a, users are only required to input a text prompt (e.g., *on a rainy day, there is a car cut in*). Then *DriveDreamer-2* produces multi-view driving videos aligned with the text input. Fig. 6 illustrates another two customized driving videos. The upper one depicts the process of the ego car changing lanes to the left during the daytime. The lower one showcases an unexpected pedestrian crossing the road at night, prompting the ego car to brake to avoid the collision. Notably, the generated videos demonstrate an exceptional level of realism, where we can even observe the reflection of high beams on the pedestrian.

Figure 7: Visualization comparison between *DriveDreamer-2* generation with and without UniMVM. The upper part depicts generation without UniMVM, while the lower part illustrates generation with UniMVM.

Quality Evaluation of Generated Videos

To verify the video generation quality, we compared *DriveDreamer-2* with various driving video generation approaches on the nuScenes validation set. Evaluations are conducted under three settings: no image condition, video condition, and first-frame multi-view image condition. Additionally, Drive-WM, DrivingDiffusion and Panacea adopt a two-stage pipeline, first generating visual conditions and then generating videos. Results in Tab. 1 show that *DriveDreamer-2* achieves high-quality outcomes across all settings. Without an image condition, it achieves an FID of 25.0 and an FVD of 105.1, significantly improving over its predecessor. Under a single-view video condition, *DriveDreamer-2* shows a 39% relative improvement in FVD compared to Drive-WM. With the first-frame multi-view image condition, *DriveDreamer-2* achieves an FID of **11.2** and an FVD of **55.7**, surpassing previous methods by a considerable margin.

To validate the quality of the generated data, we augmented the training of StreamPETR for 3D object detection and multi-object tracking using both real and generated videos from the nuScenes training set. The results, summarized in Tab. 2 and Tab. 3, show significant performance gains. Using the initial frame as a condition, the generated videos improve 3D detection metrics (mAP and NDS) by 2.8% and 3.9%, respectively, and tracking metrics (AMOTA and AMOTP) by 8.0% and 1.6%. Notably, *DriveDreamer-2* can generate high-quality videos without image conditions, further enhancing content diversity and performance. Without image conditions, improvements are 3.8% for mAP, 4.4% for NDS, 8.3% for AMOTA, and 2.3% for AMOTP, compared to the baseline.

Ablation Study

We conduct an ablation study to investigate the effect of diffusion backbone and the proposed UniMVM, and the results are in Tab. 4. Compared to SD1.4 (Rombach et al. 2022) used in DriveDreamer, SVD (Blattmann et al. 2023) provides richer prior knowledge of videos, resulting in 17.2 FID and 94.6 FVD. The introduction of SVD results in an almost 70% improvement in FVD. Additionally, we also note that a slight decrease in FID, which we hypothesize is attributed to the introduction of the cross-view module, disrupting the SVD's ability to learn spatial features. To fully unleash the potential of SVD in multi-view video generation, we propose UniMVM, which unifies constraints

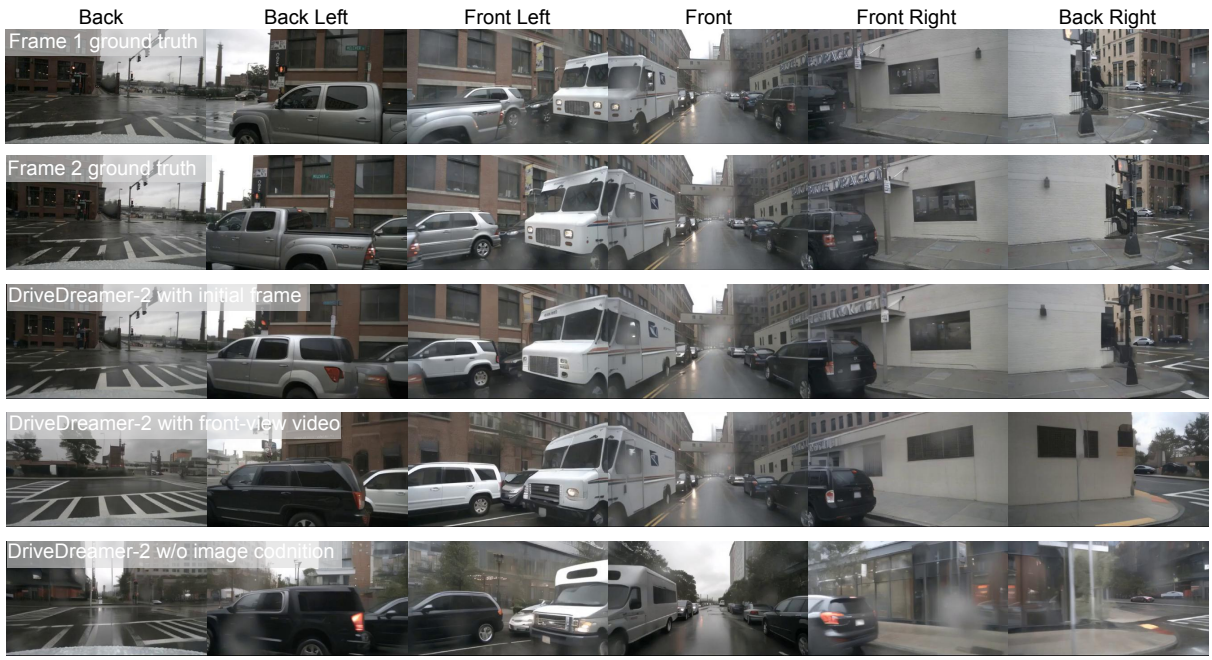


Figure 8: Visualization comparison with different conditions. Under different conditions, *DriveDreamer-2* aintains high multi-view consistency. When conditioned on the 1st frame (row 1), the generated second frame (row 3) closely matches the ground truth (row 2). Using front-view video increases diversity in generation (row 4), with only front views aligning well with the ground truth. Notably, without any conditioning image, *DriveDreamer-2* generates the highest diversity (row 5), showing significant color variations in cars and backgrounds compared to the ground truth.

Method	Conditions	FID↓	FVD↓
DriveDreamer (Wang et al. 2023c)	-	26.8	353.2
	1st-frame multi-view image	14.9	340.8
DriveDreamer-2	-	25.0	105.1
	1-view video	18.4	74.9
	1st-frame multi-view image	**11.2**	**55.7**

Table 5: The ablation study on different conditions.

on intra- and cross-view, achieving remarkable FID and FVD scores of 11.2 and 55.7, respectively. These represent relative improvements of ∼30% and ∼80% compared to DriveDreamer. As depicted in Fig. 7, the introduction of UniMVM leads to significant improvements in generating multi-view videos, both in foreground and background aspects. These qualitative results highlight UniMVM's significant capability to ensure multi-view consistency.

Moreover, we explore the influence of various conditions on driving video generation, as shown in Tab. 5 and Fig. 8. The first row in Fig. 8 illustrates the ground truth (GT) of the first frame, representing the style of the GT video. Meanwhile, the second row displays the GT of the second frame, representing the GT of the generated multi-view frame. *DriveDreamer-2* with the initial frame can generate results that are highly similar to the GT video, achieving optimal results in terms of FID and FVD, with scores of 11.2 and 55.7, respectively. The video generated by *DriveDreamer-2* with the front-view video retains some aspects of the GT scene while also introducing some diversity, resulting in 17.2 FID and 94.6 FVD. *DriveDreamer-2* can also generate extremely competitive results even without any image conditioning, achieving FID and FVD scores of 25.0 and 105.1, respectively. Notably, *DriveDreamer-2* exhibits the highest diversity in this setting, where the generated appearance of cars and the street backgrounds differ significantly from GT.

Discussion and Conclusion

This paper introduces *DriveDreamer-2*, which pioneers the generation of user-customized driving videos. Leveraging an LLM, *DriveDreamer-2* translates user queries into foreground agent trajectories. It then uses the proposed HDMap generator to create background traffic conditions based on these trajectories. The structured conditions facilitate video generation, enhanced by UniMVM for temporal and spatial coherence. Extensive experiments verify that *DriveDreamer-2* can generate uncommon driving scenarios, such as abrupt vehicle maneuvers. The generated videos significantly enhance the training of driving perception methods. Importantly, *DriveDreamer-2* achieves FID and FVD scores of 11.2 and 55.7, respectively, representing relative improvements of approximately 30% and 50% over state-of-the-art methods. These results affirm its efficacy in multi-view driving video generation.

References

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2019. nuScenes: A multimodal dataset for autonomous driving. *CVPR*.

Chen, L.; Wu, P.; Chitta, K.; Jaeger, B.; Geiger, A.; and Li, H. 2023. End-to-end Autonomous Driving: Challenges and Frontiers. *arXiv preprint arXiv:2306.16927*.

Cho, J. H.; Ivanovic, B.; Cao, Y.; Schmerling, E.; Wang, Y.; Weng, X.; Li, B.; You, Y.; Krähenbühl, P.; Wang, Y.; et al. 2024. Language-Image Models with 3D Understanding. *arXiv preprint arXiv:2405.03685*.

Denton, E.; and Fergus, R. 2018. Stochastic video generation with a learned prior. In *ICML*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *NeurIPS*.

Feng, L.; Li, Q.; Peng, Z.; Tan, S.; and Zhou, B. 2023. Trafficgen: Learning to generate diverse and realistic traffic scenarios. In *ICRA*.

Gao, R.; Chen, K.; Xie, E.; Hong, L.; Li, Z.; Yeung, D.-Y.; and Xu, Q. 2023. Magicdrive: Street view generation with diverse 3d geometry control. *arXiv preprint arXiv:2310.02601*.

Gao, S.; Yang, J.; Chen, L.; Chitta, K.; Qiu, Y.; Geiger, A.; Zhang, J.; and Li, H. 2024. Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability. *arXiv preprint arXiv:2405.17398*.

Gao, Z.; Mu, Y.; Shen, R.; Chen, C.; Ren, Y.; Chen, J.; Li, S. E.; Luo, P.; and Lu, Y. 2022. Enhance Sample Efficiency and Robustness of End-to-end Urban Autonomous Driving via Semantic Masked World Model. *arXiv preprint arXiv:2210.04017*.

Ha, D.; and Schmidhuber, J. 2018. Recurrent world models facilitate policy evolution. *NeurIPS*.

Hafner, D.; Lee, K.-H.; Fischer, I.; and Abbeel, P. 2022. Deep hierarchical planning from pixels. *NeurIPS*.

Hafner, D.; Lillicrap, T.; Ba, J.; and Norouzi, M. 2019. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*.

Hafner, D.; Lillicrap, T.; Norouzi, M.; and Ba, J. 2020. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*.

Hafner, D.; Pasukonis, J.; Ba, J.; and Lillicrap, T. 2023. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*.

Harvey, W.; Naderiparizi, S.; Masrani, V.; Weilbach, C.; and Wood, F. 2022. Flexible diffusion modeling of long videos. *NeurIPS*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; et al. 2022a. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *NeurIPS*.

Ho, J.; Saharia, C.; Chan, W.; Fleet, D. J.; Norouzi, M.; and Salimans, T. 2022b. Cascaded diffusion models for high fidelity image generation. *JMLR*.

Hong, W.; Ding, M.; Zheng, W.; Liu, X.; and Tang, J. 2022. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*.

Hsieh, J.-T.; Liu, B.; Huang, D.-A.; Fei-Fei, L. F.; and Niebles, J. C. 2018. Learning to decompose and disentangle representations for video prediction. *NeurIPS*.

Hu, A.; Corrado, G.; Griffiths, N.; Murez, Z.; Gurau, C.; Yeo, H.; Kendall, A.; Cipolla, R.; and Shotton, J. 2022. Model-based imitation learning for urban driving. *NeurIPS*.

Hu, A.; Russell, L.; Yeo, H.; Murez, Z.; Fedoseev, G.; Kendall, A.; Shotton, J.; and Corrado, G. 2023. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*.

Jia, F.; Mao, W.; Liu, Y.; Zhao, Y.; Wen, Y.; Zhang, C.; Zhang, X.; and Wang, T. 2023. Adriver-i: A general world model for autonomous driving. *arXiv preprint arXiv:2311.13549*.

Kalchbrenner, N.; Oord, A.; Simonyan, K.; Danihelka, I.; Vinyals, O.; Graves, A.; and Kavukcuoglu, K. 2017. Video pixel networks. In *ICML*.

Karras, T.; Aittala, M.; Aila, T.; and Laine, S. 2022. Elucidating the design space of diffusion-based generative models. *NeurIPS*.

Khachatryan, L.; Movsisyan, A.; Tadevosyan, V.; Henschel, R.; Wang, Z.; Navasardyan, S.; and Shi, H. 2023. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*.

Kim, S. W.; Zhou, Y.; Philion, J.; Torralba, A.; and Fidler, S. 2020. Learning to simulate dynamic environments with gamegan. In *CVPR*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kondratyuk, D.; Yu, L.; Gu, X.; Lezama, J.; Huang, J.; Hornung, R.; Adam, H.; Akbari, H.; Alon, Y.; Birodkar, V.; et al. 2023. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*.

Kumar, M.; Babaeizadeh, M.; Erhan, D.; Finn, C.; Levine, S.; Dinh, L.; and Kingma, D. 2019. Videoflow: A flow-based generative model for video. *arXiv preprint arXiv:1903.01434*.

Li, X.; Zhang, Y.; and Ye, X. 2023. DrivingDiffusion: Layout-Guided multi-view driving scene video generation with latent diffusion model. *arXiv preprint arXiv:2310.07771*.

Li, Y.; Liu, H.; Wu, Q.; Mu, F.; Yang, J.; Gao, J.; Li, C.; and Lee, Y. J. 2023. Gligen: Open-set grounded text-to-image generation. In *CVPR*.

Lin, J.; Du, Y.; Watkins, O.; Hafner, D.; Abbeel, P.; Klein, D.; and Dragan, A. 2023. Learning to Model the World with Language. *arXiv preprint arXiv:2308.01399*.

Ma, Y.; Cao, Y.; Sun, J.; Pavone, M.; and Xiao, C. 2023. Dolphins: Multimodal language model for driving. *arXiv preprint arXiv:2312.00438*.

Mao, J.; Qian, Y.; Zhao, H.; and Wang, Y. 2023a. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*.

Mao, J.; Ye, J.; Qian, Y.; Pavone, M.; and Wang, Y. 2023b. A language agent for autonomous driving. *arXiv preprint arXiv:2311.10813*.

Mathieu, M.; Couprie, C.; and LeCun, Y. 2015. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *ICML*.

Pan, M.; Zhu, X.; Wang, Y.; and Yang, X. 2022. Iso-dream: Isolating and leveraging noncontrollable visual dynamics in world models. *NeurIPS*.

Parmar, G.; Zhang, R.; and Zhu, J.-Y. 2022. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*.

Ranzato, M.; Szlam, A.; Bruna, J.; Mathieu, M.; Collobert, R.; and Chopra, S. 2014. Video (language) modeling: a baseline for generative models of natural videos. *arXiv preprint arXiv:1412.6604*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Seo, Y.; Hafner, D.; Liu, H.; Liu, F.; James, S.; Lee, K.; and Abbeel, P. 2023. Masked world models for visual control. In *CoRL*.

Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*.

Srivastava, N.; Mansimov, E.; and Salakhudinov, R. 2015. Unsupervised learning of video representations using lstms. In *ICML*.

Tan, S.; Ivanovic, B.; Weng, X.; Pavone, M.; and Kraehenbuehl, P. 2023. Language conditioned traffic generation. *arXiv preprint arXiv:2307.07947*.

Unterthiner, T.; Van Steenkiste, S.; Kurach, K.; Marinier, R.; Michalski, M.; and Gelly, S. 2018. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*.

Villegas, R.; Babaeizadeh, M.; Kindermans, P.-J.; Moraldo, H.; Zhang, H.; Saffar, M. T.; Castro, S.; Kunze, J.; and Erhan, D. 2023. Phenaki: Variable length video generation from open domain textual description. *ICLR*.

Vondrick, C.; Pirsiavash, H.; and Torralba, A. 2016. Generating videos with scene dynamics. *NeurIPS*.

Wang, S.; Liu, Y.; Wang, T.; Li, Y.; and Zhang, X. 2023a. Exploring Object-Centric Temporal Modeling for Efficient Multi-View 3D Object Detection. *arXiv preprint arXiv:2303.11926*.

Wang, S.; Yu, Z.; Jiang, X.; Lan, S.; Shi, M.; Chang, N.; Kautz, J.; Li, Y.; and Alvarez, J. M. 2024a. OmniDrive: A Holistic LLM-Agent Framework for Autonomous Driving with 3D Perception, Reasoning and Planning. *arXiv preprint arXiv:2405.01533*.

Wang, X.; Yuan, H.; Zhang, S.; Chen, D.; Wang, J.; Zhang, Y.; Shen, Y.; Zhao, D.; and Zhou, J. 2023b. VideoComposer: Compositional Video Synthesis with Motion Controllability. *arXiv preprint arXiv:2306.02018*.

Wang, X.; Zhu, Z.; Huang, G.; Chen, X.; and Lu, J. 2023c. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*.

Wang, X.; Zhu, Z.; Huang, G.; Wang, B.; Chen, X.; and Lu, J. 2024b. WorldDreamer: Towards General World Models for Video Generation via Predicting Masked Tokens. *arXiv preprint arXiv:2401.09985*.

Wang, X.; Zhu, Z.; Zhang, Y.; Huang, G.; Ye, Y.; Xu, W.; Chen, Z.; and Wang, X. 2023d. Are We Ready for Vision-Centric Driving Streaming Perception? The ASAP Benchmark. In *CVPR*.

Wang, Y.; He, J.; Fan, L.; Li, H.; Chen, Y.; and Zhang, Z. 2023e. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. *arXiv preprint arXiv:2311.17918*.

Weissenborn, D.; Täckström, O.; and Uszkoreit, J. 2019. Scaling autoregressive video models. *arXiv preprint arXiv:1906.02634*.

Wen, Y.; Zhao, Y.; Liu, Y.; Jia, F.; Wang, Y.; Luo, C.; Zhang, C.; Wang, T.; Sun, X.; and Zhang, X. 2023. Panacea: Panoramic and Controllable Video Generation for Autonomous Driving. *arXiv preprint arXiv:2311.16813*.

Wu, P.; Escontrela, A.; Hafner, D.; Abbeel, P.; and Goldberg, K. 2023. Daydreamer: World models for physical robot learning. In *CoRL*.

Yang, J.; Gao, S.; Qiu, Y.; Chen, L.; Li, T.; Dai, B.; Chitta, K.; Wu, P.; Zeng, J.; Luo, P.; et al. 2024. Generalized predictive model for autonomous driving. In *CVPR*.

Yang, K.; Ma, E.; Peng, J.; Guo, Q.; Lin, D.; and Yu, K. 2023. BEVControl: Accurately Controlling Street-view Elements with Multi-perspective Consistency via BEV Sketch Layout. *arXiv preprint arXiv:2308.01661*.

Yang, R.; Srivastava, P.; and Mandt, S. 2022. Diffusion probabilistic modeling for video generation. *arXiv preprint arXiv:2203.09481*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *ICCV*.

Zhong, Z.; Rempe, D.; Chen, Y.; Ivanovic, B.; Cao, Y.; Xu, D.; Pavone, M.; and Ray, B. 2023a. Language-Guided Traffic Simulation via Scene-Level Diffusion. *arXiv preprint arXiv:2306.06344*.

Zhong, Z.; Rempe, D.; Xu, D.; Chen, Y.; Veer, S.; Che, T.; Ray, B.; and Pavone, M. 2023b. Guided conditional diffusion for controllable traffic simulation. In *ICRA*.