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

Supplementary Material

1. Supplementary Visualization Results

As introduced in the main text, our method is the first to achieve high-resolution, controllable multi-view driving video generation, capable of handling sparse conditions and dynamic frame rates. Due to space limitations in the main text and the challenge of visually presenting high-fidelity videos in print, we provide a web page in the supplementary materials for additional results.

2. Discussions

To provide a clearer understanding of our work, we address some potential issues in this section.

Q1. Potential applications in autonomous driving.

- Our method can edit real videos and, with the support of additional training data, simulate corner cases such as traffic accidents, overtaking, lane changes, and animals obstructing the road. On the project page, we provide several videos in which the bounding boxes of the original videos have been edited to simulate collisions between other vehicles and the ego vehicle.
- Our method can be integrated with traffic flow simulation methods [13, 24], where the traffic simulation generates layout sequences for various driving scenarios, and DriveScape transforms these layout sequences into highly realistic videos.
- Based on DriveScape and HD maps, we can construct a closed-loop simulation environment for autonomous driving algorithms, such as UniAD [12]. Using our model, the environment can dynamically respond to the decisions made by the driving algorithm and generate a realistic environment for the next moment. This enables automated testing of driving algorithms, significantly reducing the testing costs associated with autonomous driving systems.

Q2. Limitation and future works.

- The current generation efficiency remains relatively slow; DDIM inference with 25 steps takes approximately 20 minutes, which restricts the model's applicability. In future work, we aim to improve efficiency by incorporating methods such as LCM [15] and Rectified Flow [6].
- Similar to other video generation algorithms [14, 22], the generation quality for non-rigid objects, such as humans, remains suboptimal. The primary reason for this limitation is the relatively small size of the training dataset, Nuscenes. To address this issue, we plan to collect additional data to enhance performance.
- Our model successfully achieves multi-task training for image-to-video (I2V) and text-to-video (T2V) generation. However, due to the limited scale of the dataset, the low quality of text data, and the fact that the pre-trained model was not originally designed for T2V tasks, the performance of I2V is significantly superior to T2V. In future work, we aim to further improve the T2V capabilities of the model.

3. Related Works

3.1. Controllable generation

With the advent of diffusion models, significant progress has been made in the field of text-to-video generation [1, 2, 5, 8, 10, 11, 16, 21, 27]. Video LDM [4] utilizes a latent diffusion pipeline, where the diffusion denoising process operates on image latents, significantly accelerating the denoising process. However, text alone cannot provide precise control over video generation. To address this limitation, subsequent methods incorporated image inputs alongside text as prompts for the denoising network to achieve better control [26]. Our goal is to generate highly realistic street videos, which represent a particularly challenging scenario due to their complexity. These scenes involve numerous elements and interactions (e.g., intricate street layouts, moving vehicles, etc.), necessitating additional information for fine-grained control. In our approach, we integrate road maps, 3D bounding boxes, and BEV keyframes to achieve precise control over video generation.

3.2. Multi-view video generation

Multi-view consistency and temporal consistency are two critical challenges in multi-view video generation. To maintain multi-view consistency, MVDiffusion [18] proposed a correspondence-aware attention module to align information across multiple views. Additionally, [19] employed epipolar geometry to regularize the consistency between different views. MagicDrive [7] utilized camera pose, bounding boxes, and road maps as priors, integrating an additional cross-view attention block to enhance consistency. However, these methods are limited to generating multi-view images rather than videos.

3.3. Street view generation

Most street view generation models rely on 2D layouts as input conditions, such as BEV maps, 2D bounding boxes, and semantic segmentation. BEVGen [17] encapsulates all

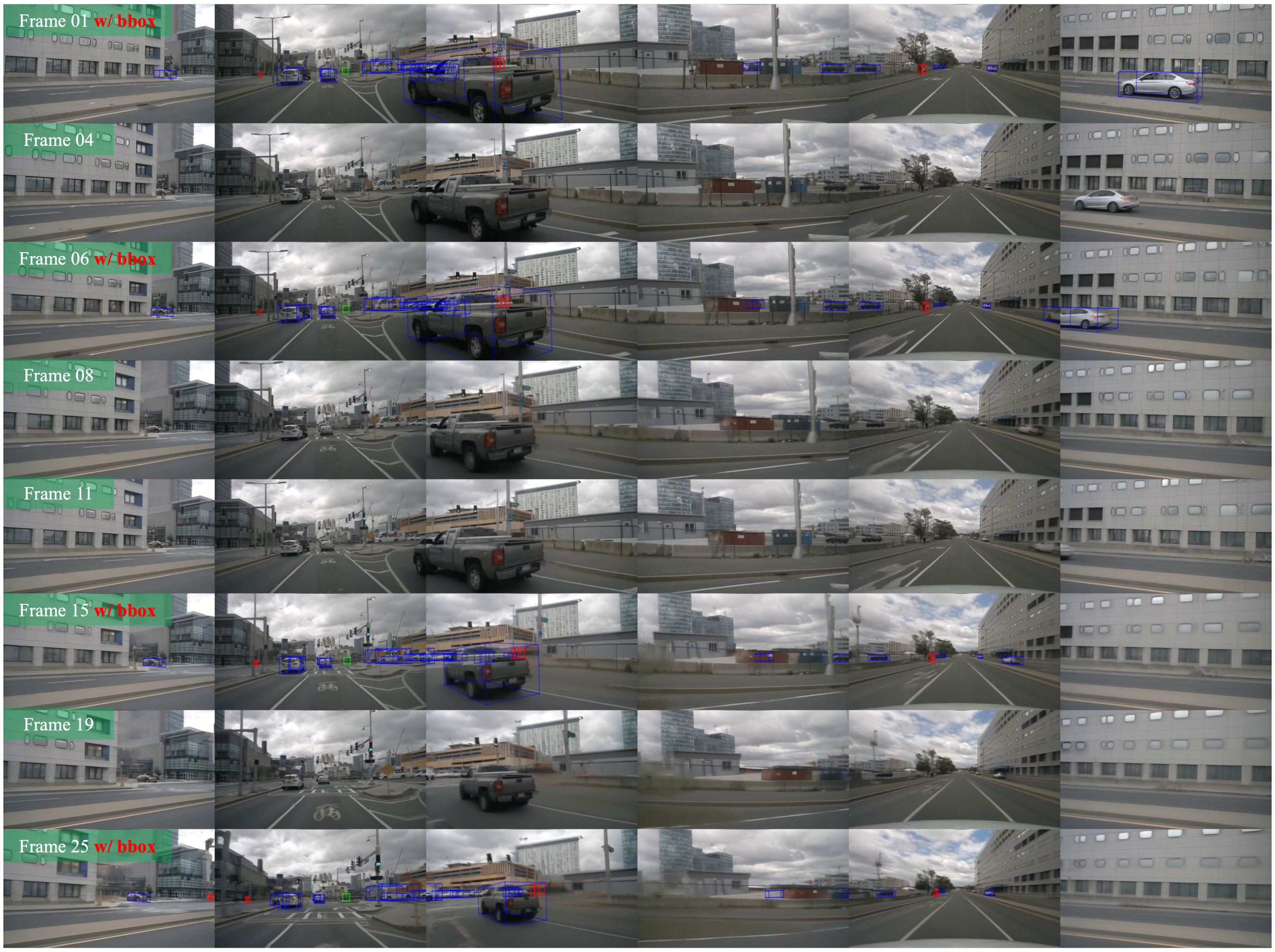

Figure 1. **More examples of sparse control.** We projected the 3D bounding box onto the video frames. Frames without the w/ bbox label are unannotated frames. As shown, the generated videos align well with the 3D bounding boxes, and the transitions during the unannotated frames are remarkably smooth. For additional results, please refer to the project_page/index.html file included in the supplementary materials.

semantic information within the BEV representation to approach street view generation. BEVControl [25] proposes a two-stage method for generating multi-view urban scene images from a BEV layout. In this approach, a controller generates foreground and background objects, and a coordinator combines them to preserve visual consistency across different views. However, projecting 3D information into 2D inherently loses 3D geometric details, which can lead to inconsistencies when extending these methods directly to video generation, particularly across multiple frames. To address this issue, we introduce 3D bounding boxes as additional conditions to guide the generation process. DrivingDiffusion [14] employs a multi-stage scheme with two models and two stages of post-processing to first generate frames and then extend them into videos. While effective, these methods rely on a complex multi-stage pipeline. In contrast, our method adopts an end-to-end pipeline that is both efficient and effective.

4. Implementation Details

4.1. Unified Model

In this section, we present a detailed description of our unified model architecture, which comprises two main components: the Unet-Spatial-Temporal model, as defined by the publicly available Stable Video Diffusion [3], and an encoding model designed to handle multiple conditions through the use of BiMoT injectors.

The Unet-Spatial-Temporal Model can be divided into two components: the spatial part and the temporal part. In our work, we trained only the temporal part for the following reasons. The complete Unet model was pretrained on

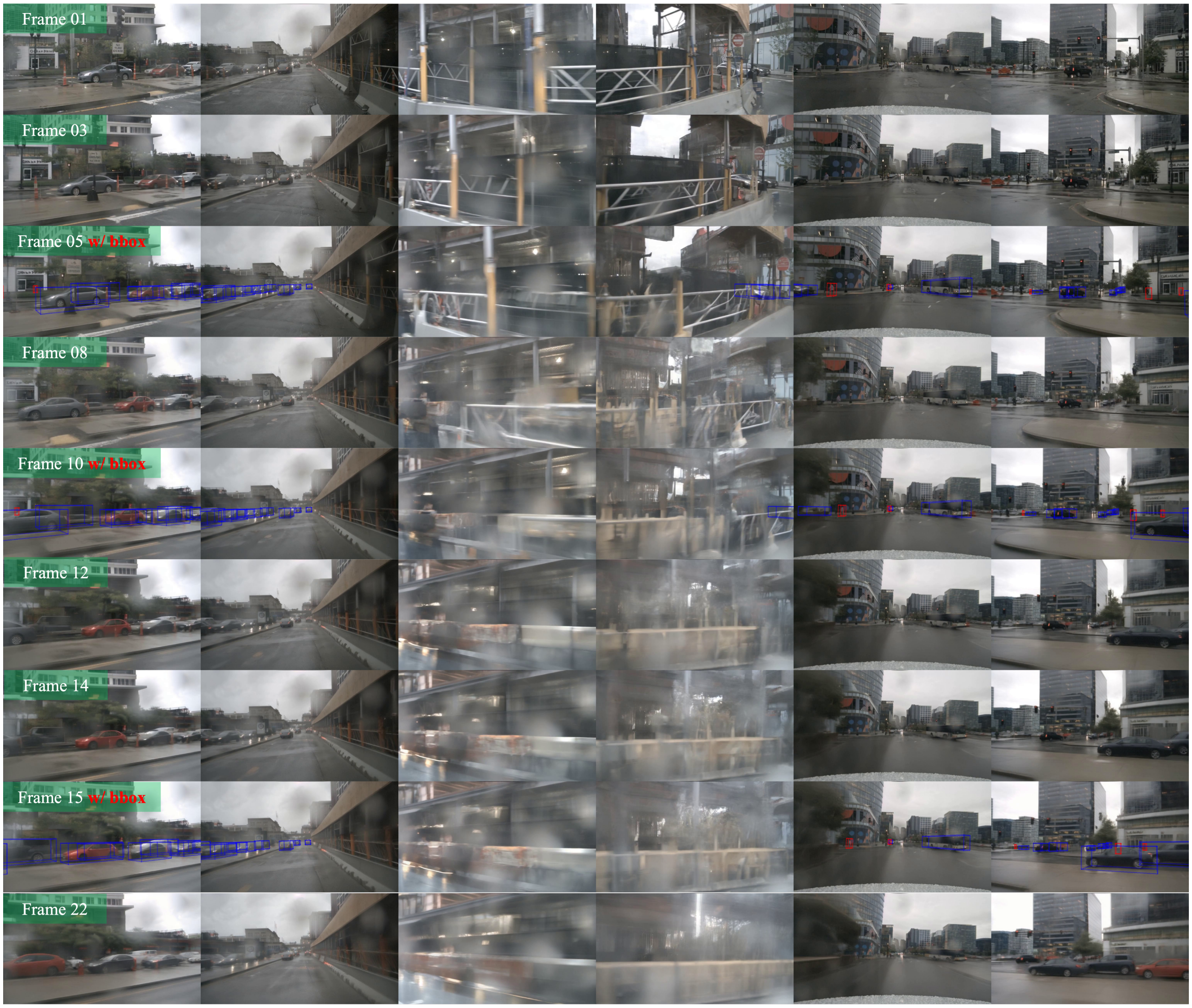

Figure 2. **More examples of sparse control.**

sequences of 14 frames at a resolution of 576×1024, using a comprehensive and diverse dataset that likely includes a wide range of objects and contexts. Consequently, the spatial part of the model has already developed a robust understanding of spatial feature representations across various environments and scenarios. This pretraining has endowed the spatial component with a broad generalization capability, allowing it to be directly applied to driving scene datasets without the need for additional training or fine-tuning. In contrast, the temporal part of the model requires further training because temporal dynamics can vary significantly across different types of video content. Driving scenes, in particular, exhibit unique temporal patterns, such as the motion of vehicles, changes in lighting conditions over time, or camera movement, which may not have been fully captured during the initial pretraining phase. By training the temporal components on driving scene datasets, the model is better equipped to learn the specific temporal dependencies and dynamics of these scenarios. This additional training is crucial for generating coherent and realistic video sequences over time, ensuring fidelity to the unique characteristics of driving scenes.

The Encoding Architecture for handling multiple conditional inputs is structured with separate encoding heads and BiMoT injectors. Each encoding head consists of a MaxPooling2D layer with a stride of 2, followed by two convolutional layers of size 3×3 with a stride of 1, integrated with residual connections. The number of channels within these encoding heads is designed to match that of the Unet-Spatial-Temporal model, with a notable exception for the

Method	Real	Generated	mAP↑	mAOE↓	mAVE↓	NDS↑
Panacea	✓	-	34.5	59.4	29.1	46.9
Panacea	-	✓	22.5	72.7	46.9	36.1
Panacea	✓	✓	37.1	54.2	27.3	49.2
DriveScape	-	✓	41.5	49.1	27.4	51.8
DriveScape	✓	✓	**42.8**	**45.5**	**27.1**	**53.2**

Table 1. Comparison involving data augmentation using synthetic data with Panacea. We attempt training exclusively using synthetic data and also explore integrating it with real data.

encoders processing Maps and 3D layouts. For these encoders, the channel count is deliberately halved. This design ensures that when features from the Maps and 3D layout encoders are concatenated, the resulting channel count aligns with other encoders, thereby enabling uniform injection of conditional features through the BiMoT injectors. The rationale for concatenating Maps and 3D layout features and injecting them simultaneously lies in their shared purpose: both provide structural information about the scene, forming a foundation for spatial composition in the generative process. In contrast, other conditional inputs typically convey detailed, dynamic content. These inputs are injected separately into the Unet to ensure they maintain their distinct influence on the generated output, preserving the integrity of their unique guidance.

4.2. Training Details

In training our comprehensive end-to-end model, which integrates a variety of conditions such as the reference camera position (camera_id), neighboring camera videos, and more, we have adopted several strategies to enhance training efficiency and boost overall model performance.

Firstly, we employ randomized conditioning throughout the training process. Specifically, there is a 50% probability of dropping the neighboring camera videos and a 20% probability of dropping the maps, 3D layouts, and BEV keyframes. This approach trains the network to develop multi-task generative capabilities, mitigating over-reliance on any specific set of conditional inputs.

Secondly, we implemented a specialized training pattern to handle samples captured from different camera perspectives. Each training batch is constructed with six consecutive samples, captured from different camera positions at the same time. These samples are arranged in a localized random order during the continuous training process. This strategy ensures diverse representations of visual angles and enhances the model's ability to learn from multi-view data. By doing so, the model gains a deeper understanding of temporal consistency and the 3D structure of scenes, leading to improved performance in multi-view video generation.

4.3. SparseCtrl Implement Details

To compare with existing works on sparse conditional control, we adapted SparseCtrl [9] in our setting to support multi-condition control. The original implementation of SparseCtrl is based on AnimateDiff, a 2D text-to-image (T2I) model with motion modules. To fairly compare the control ability of SparseCtrl with our control module, we re-implement SparseCtrl using the SVD framework. Figure 3 illustrates the detailed pipeline. Our implementation incorporates two main principles of SparseCtrl: the Spatial-Temporal Sparse Condition Encoder and the concept of ”zero-initialized layers.” We adopt the SVD encoder as the Condition Encoder. Following the original SparseCtrl implementation, different conditions are channel-wise concatenated after being processed through the VAE encoder and are then fed into the Condition Encoder. Each hidden state is added to the corresponding feature maps of the backbone encoder via a zero-initialized layer. In our training setup, we freeze the spatial blocks of the backbone and train all other blocks to optimize performance.

5. Additional Experiments

Following Panacea [23], we also utilize generated data to assist in training the perception model StreamPETR [20]. Specifically, we use the layouts from the training set to generate 10,000 short videos, each consisting of 8 frames. We pre-train StreamPETR (R50) with these generated videos and subsequently fine-tune it using real data. As shown in Tab. 1, the results demonstrate that our model achieves superior performance in both the pre-training and fine-tuning stages.

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. 1

[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. 1

[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. 1

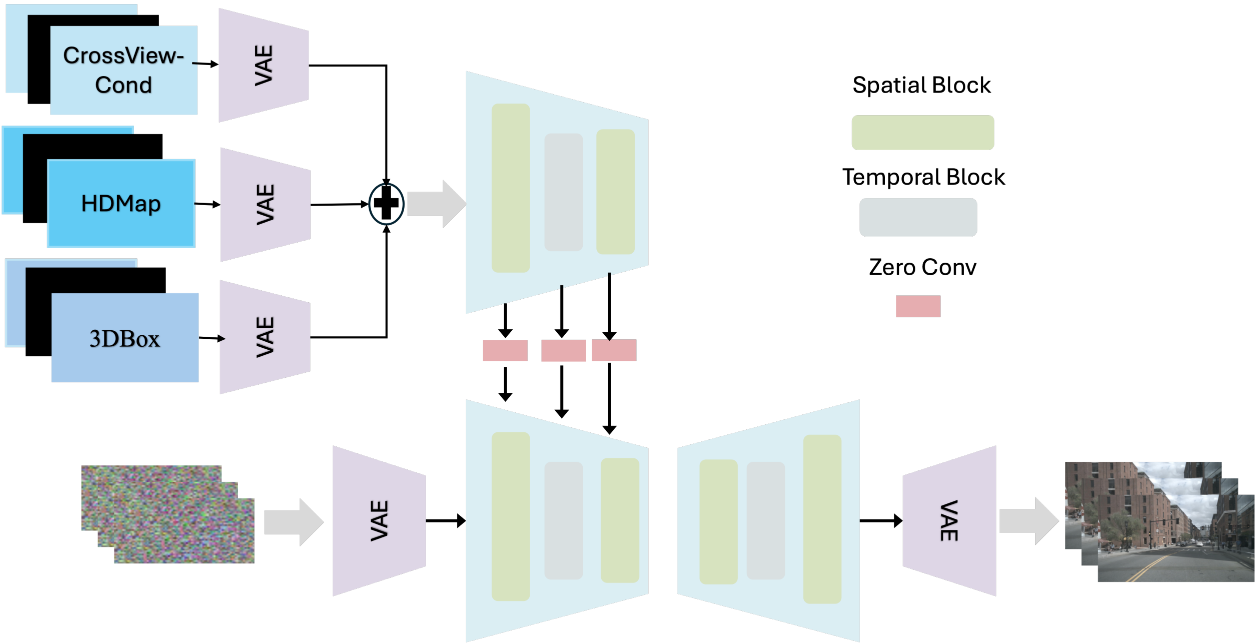

Figure 3. Pipeline of our implementation of SparseCtrl

[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. 1

[6] 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. 1

[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

[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. 1

[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. 4

[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. 1

[11] 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. 1

[12] 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

[13] Daniel Krajzewicz, Jakob Erdmann, Michael Behrisch, and Laura Bieker. Recent development and applications of sumo-simulation of urban mobility. *International journal on advances in systems and measurements*, 5(3&4), 2012. 1

[14] 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. 1, 2

[15] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 1

[16] 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. 1

[17] 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

[18] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion, 2023. 1

[19] 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. 1

[20] 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. 4

[21] 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. 1

[22] 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. 1

[23] 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. 4

[24] Licheng Wenl, Daocheng Fu, Song Mao, Pinlong Cai, Min Dou, Yikang Li, and Yu Qiao. Limsim: A long-term interactive multi-scenario traffic simulator. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1255–1262. IEEE, 2023. 1

[25] 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. 2

[26] 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. 1

[27] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models, 2023. 1