Panacea: Panoramic and Controllable Video Generation for Autonomous Driving
Supplementary Materials

A. More Implementation Details

Decomposed 4D Attention. We present a more detailed demonstration of our decomposed 4D attention, as shown in Fig. 1, we incorporate the text context into the intra-view, inter-view, and cross-frame blocks by utilizing cross attention. This approach is consistent with the methodology employed in [5].

Construction of Camera Pose. In order to facilitate consistency across multi-view images, , we introduce a camera pose prior into control signals by encoding the camera pose into the 3D direction vector following the success in 3D perception [3, 12]. Specifically, for each point $p^c = (u, v, d, 1)$ in camera frustum space, it can be projected into 3D space by intrinsic $K \in \mathbb{R}^{4\times4}$ and extrinsic $E \in \mathbb{R}^{4\times4}$ as in Eq. 1:

$$p^{3d} = E \times K^{-1} \times p^c \tag{1}$$

Here (u, v) denotes the pixel coordinates in the image, d is the depth along the axis orthogonal to the image plane, 1 is for the convenience of calculation in homogeneous form. We pick two points and set the depth of them to $d_1 = 1, d_2 = 2$. Then the direction vector of the corresponding camera ray can be computed by:

$$\mathrm{DV}(u, v) = p^{3d}(d_2) - p^{3d}(d_1) \tag{2}$$

We normalize the direction vector by dividing the vector magnitude and multiply 255 to simply convert it into RGB pseudo-color image, as shown in Fig. 2.

Training Details of StreamPETR. StreamPETR is trained at a resolution of 512×256 instead of the 704×256 in original baseline [7]. For the single-frame baseline (StreamPETR-S), we close the query propagation and retrain the model. The batch size is 16 and the learning rate is 4e-4.

B. More Experimental Results

Lane Detection Task. To showcase our versatility across various perception tasks, we have expanded our experiments to include lane detection task on nuScenes dataset. In lane detection, as Tab. 1 (a) shows, our synthetic nuScenes validation data achieves a 59% relative mAP compared to real data using MapTR [2] for evaluation. Moreover, synthetic data enhances MapTR's training with a surprisingly 5.0 point gain (from 45.0 to 50.0, 512×288 resolution), as Tab. 1 (b) shows.

Real	Generated	mAP
✓	-	45.0
-	✓	26.4 (59%)

(a) Comparison of generated and real data on nuScenes validation set, using pre-trained MapTR.

Real	Generated	mAP
✓	-	45.0
-	✓	29.8
✓	✓	50.0 (+5.0)

(b) Comparison involving data augmentation using synthetic data for MapTR training.

Table 1. Lane detection results.

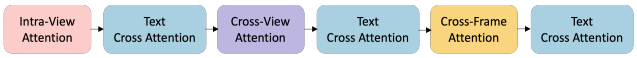

Figure 1. Illustration of the structure of decomposed 4D attention

BEV Sequence Control. Given that ControlNet [11] potentially imposes an additional cost, we explore a simpler way to introduce BEV control signals without ControlNet, merely through the concatenation of the BEV feature with the input latent codes in the channel dimension. As show in Tab. 3, the FVD and FID are 87 and 3.19 higher than using ControlNet, thus we keep the ControlNet as our default design to introduce BEV control.

Per-class Results. We provide per-class detection results in Tab. 2 on the validation set of nuScenes with StreamPETR, corresponding to the results in Tab. 3 in the main paper. As shown in Tab. 2, incorporating data generated by Panacea leads to enhanced AP for virtually all classes.

C. More Visualization Results

Multi-View Video Generation. We present more visualization results of our Panacea. As shown in Fig. 3, we exhibit multi-view videos generated from the validation set,

	mAP	car	truck	bus	trailer	constr.	pedestrian	motorcycle	bicycle	cone	barrier
Real	34.5	54.5	28.9	30.1	8.4	9.7	39.9	34.4	32.9	56.6	50.1
+Panacea	37.1+2.6	57.1+2.6	29.4+0.5	30.7+0.6	14.6+6.2	10.8+1.1	42.7+2.8	38.7+3.3	31.1-1.8	60.8+4.2	54.7+4.6

Table 2. Per-class comparison involving data augmentation using synthetic data, where 'constr' is short for construction vehicles.

Settings	FVD↓	FID↓
Panacea (ControlNet)	139	16.96
Panacea (Concat)	226 (+87)	20.15 (+3.19)

Table 3. Ablation studies on how to introduce BEV control signals.

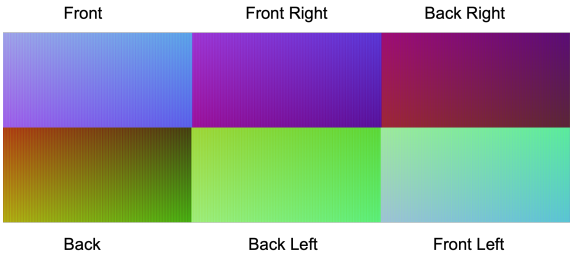

Figure 2. Visualization of pseudo-color image of camera pose.

comprising a total of 8 frames and 6 views per video. One can observe that these generated videos maintain good temporal and cross-view consistency.

Attribute Controllable Video Generation. We present visualization results with different attributes control. As illustrated in Fig. 4, we exhibit the same case with varying attribute control. Our model is capable of generating multi-view frames that align well with BEV layout and global attributes.

Generating Long Videos. Our Panacea possesses the potential to serve as a viable candidate for real-world simulation. We conduct an experiment in which we discard the BEV layout control part and use four frames as image conditions, applying iterative inference with a sliding window to generate long videos. As shown in Fig. 5, we exhibit consecutive frames of the generated videos. Panacea is proficient in generating high-quality long videos.

D. More Related Works

There are two main categories of driving scene synthesis methods, one is the generative methods based on GAN or diffusion models [6, 9], as expounded in the main paper, and the other is NeRF (Neural Radiance Fields) based methods, such as Unisim [10], Neuralsim [1] and Mars [8]. The generation quality of NeRF-based methods can be high. However, the most significant difference between this and our diffusion-based method is that NeRF-based methods can merely reconstruct pre-existing scenes that the model is trained on, consequently resulting in a lack of diversity.

E. Gen-nuScenes Dataset

We synthesis a new training dataset named Gen-nuScenes to enhance the training of Stream-PETR, which contains totally 139440 videos, each video of 8 frame length.

F. Limitations

Our task addresses a novel problem of controllable multi-view video generation for autonomous driving. Due to time and resource constraints, we have not yet undertaken more detailed designs of the model. Consequently, the quality of the videos generated by our model still leaves room for improvement. For example, the temporal and view consistency are not perfect since learning the correlation between all views and all frames poses a significant challenge, especially when the frame length is long. Our model also has some deformation issues long-termly where vehicles become very small, which is a common challenge for current video generation models. In the future, more effective designs for temporal and view consistency might need to be explored. Additionally, Panacea's computational cost of inference is relatively high. Thus we plan to enhance the efficiency of Panacea in the future. Furthermore, we still employ a relatively low spatial resolution due to time and resource limitations. Future work could involve integrating more powerful generative models, such as SD-XL [4] and more efficient manners to produce high-fidelity videos of larger spatial resolution.

References

[1] Eric Heiden, David Millard, Erwin Coumans, Yizhou Sheng, and Gaurav S Sukhatme. Neuralsim: Augmenting differentiable simulators with neural networks. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9474–9481. IEEE, 2021. 2

[2] Bencheng Liao, Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Chang Huang. Maptr: Structured modeling and learning for online vectorized HD map construction. In *ICLR*. OpenReview.net, 2023. 1

[3] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision*, pages 531–548. Springer, 2022. 1

[4] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. *CoRR*, abs/2307.01952, 2023. 2

Figure 3. Multi-view videos generated by Panacea.

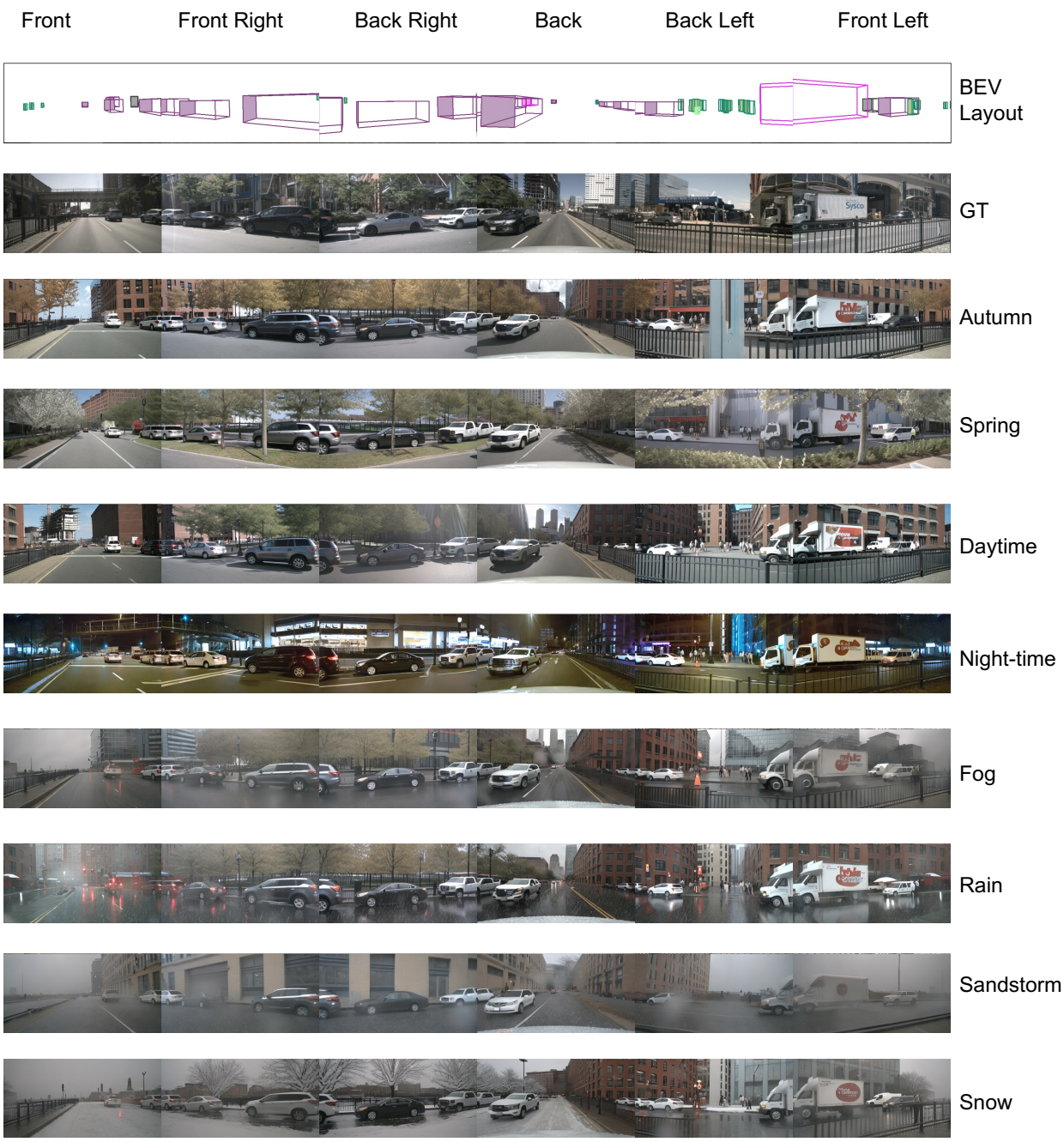

Figure 4. Multi-view frames generated by Panacea with different attribute controls.

[5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10674–10685. IEEE, 2022. 1

[6] Alexander Swerdlow, Runsheng Xu, and Bolei Zhou. Street-view image generation from a bird's-eye view layout. *CoRR*, abs/2301.04634, 2023. 2

[7] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. *CoRR*, abs/2303.11926, 2023. 1

[8] Zirui Wu, Tianyu Liu, Liyi Luo, Zhide Zhong, Jianteng Chen, Hongmin Xiao, Chao Hou, Haozhe Lou, Yuantao Chen, Runyi Yang, Yuxin Huang, Xiaoyu Ye, Zike Yan, Yongliang Shi, Yiyi Liao, and Hao Zhao. MARS: an instance-aware, modular and realistic simulator for autonomous driving. In *CICAI (1)*, pages 3–15. Springer, 2023. 2

Figure 5. Long videos generated by Panacea, from top to bottom and left to right are consecutive frames.

[9] 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. *CoRR*, abs/2308.01661, 2023. 2

[10] Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Ur-

tasun. Unisim: A neural closed-loop sensor simulator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1389–1399, 2023. 2

[11] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *CoRR*, abs/2302.05543, 2023. 1

[12] 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