

GenieDrive: Towards Physics-Aware Driving World Model with 4D Occupancy Guided Video Generation

Supplementary Material

1. Implementation Details of Tri-Plane VAE

In our tri-plane VAE, we first apply a 3D convolution filter g_ϕ to downsample the occupancy $O \in \mathbb{R}^{H \times W \times D}$ into a feature volume $S \in \mathbb{R}^{h \times w \times d \times C}$, where $\frac{H}{h} = \frac{W}{w} = \frac{D}{d} = 4$ and $C = 64$. We then decompose S into three latent planes: $Z_{xy} \in \mathbb{R}^{h \times w \times C}$, $Z_{yz} \in \mathbb{R}^{w \times d \times C}$, and $Z_{xz} \in \mathbb{R}^{h \times d \times C}$. For the occupancy resolution in Occ3D-NuScenes [7], we have $H = 200$, $W = 200$, and $D = 16$, resulting in $S \in \mathbb{R}^{50 \times 50 \times 4 \times 64}$, $Z_{xy} \in \mathbb{R}^{50 \times 50 \times 64}$, $Z_{yz} \in \mathbb{R}^{50 \times 4 \times 64}$, and $Z_{xz} \in \mathbb{R}^{50 \times 4 \times 64}$. To ensure that the tri-plane latent representation can be easily processed by the subsequent attention module, we concatenate the three latent planes into a unified feature $Z \in \mathbb{R}^{50 \times 58 \times 64}$, as shown in Figure 1. In contrast, previous works such as OccWorld [9], DOME [5], and I^2 -World [6] typically compress the occupancy into a BEV feature of shape $\mathbb{R}^{50 \times 50 \times 128}$. Thus, our latent tri-plane representation occupies only

$$\frac{50 \times 58 \times 64}{50 \times 50 \times 128} \times 100\% = 58\% \quad (1)$$

of the size used in previous methods, while still achieving superior occupancy reconstruction performance (86.15 mIoU and 75.53 IoU). The superior performance and efficiency of our method primarily stem from the fact that our latent representation is more 3D-aware, rather than compressing the full 3D occupancy into a 2D BEV feature as in previous BEV-based approaches. Our compact latent representation also greatly reduces the parameter count of the subsequent occupancy prediction module, enabling our occupancy world model to achieve state-of-the-art performance while using only 3.4M parameters. We train our tri-plane VAE for 210 epochs with a dropout rate of 0.5. By epoch 140, the model already reaches 83.34 mIoU, and additional training further improves performance. The VAE can be trained very efficiently, requiring only 6 hours on $8 \times 48\text{GB}$ GPUs.

2. Details of 4D Occupancy Forecasting

We adopt an encoder-decoder design similar to that of I^2 -World [6] for occupancy forecasting. In our encoder, we propose the Mutual Control Attention (MCA), which iteratively injects information between the occupancy latent (a tri-plane) and the control latent. We then apply a shallow MLP, referred to as the transformation head, to transform the control latent into a representation used for intermediate supervision, as described in Eq. 8 of the main sub-



Figure 1. **Concatenated Tri-Plane Feature.** To make tri-plane representation more suitable for the following processing, we concatenate three planes to get a unified feature representation.

mission. We also fuse the latent tri-plane with this transformed latent using cross-attention. This fusion design allows different regions of the tri-plane latent to be influenced by the transformed control latent to varying degrees. The fused latent tri-plane is then fed into the subsequent spatial-temporal blocks. More precisely, the “spatial” component corresponds to self-attention applied within the latent tri-plane itself, while the “temporal” component concatenates the previous k latent tri-planes with the current one along the channel dimension, followed by an MLP that projects the concatenated channels from $(k + 1)C$ back to C . The output latent tri-plane is passed to the tri-plane VAE decoder to obtain the predicted occupancy for the next timestep.

3. Qualitative Comparison on Forecasting

We also provide a qualitative comparison with previous methods, including OccWorld [9], DOME [5], and I^2 -World [6], in Figure 4. We use bounding boxes to highlight the main differences between each method and the ground truth. Two scene examples are shown in Figure 4. For the first scene, OccWorld produces an unreasonable road surface, DOME diminishes vehicles in its forecasting results, and I^2 -World predicts inaccurate occupancy compared to the ground truth. In contrast, our method provides physically reasonable forecasts, particularly in capturing the dynamics of driving vehicles (blue and yellow occupancy). We further highlight that only our method consistently predicts pedestrians (red occupancy) across different timesteps, whereas the comparative methods tend to lose this detail as the timestep increases. In the second scene, the methods need to forecast the driving behavior of the truck (purple occupancy). OccWorld gradually deforms the truck, resulting in an unnatural shape in later predictions. In DOME’s

Table 1. **Reconstruction and Forecasting Performance Change in End-to-End Training.** ‘R’ denotes reconstruction and ‘F’ denotes forecasting. As the number of training epochs increases, forecasting performance gradually improves, whereas reconstruction performance decreases.

| Epoch | R. mIoU | R. IoU | F. mIoU | F. IoU |
|-------|--------------|--------------|--------------|--------------|
| 0 | 86.15 | 75.53 | 39.79 | 50.46 |
| 4 | 73.89 | 66.39 | 41.64 | 50.71 |
| 8 | 73.05 | 65.06 | 42.36 | 51.47 |
| 12 | 71.31 | 63.65 | 42.53 | 51.71 |
| 16 | 70.07 | 63.13 | 42.59 | 51.80 |
| 20 | 68.31 | 62.32 | 42.49 | 51.76 |
| 24 | 67.89 | 62.33 | 42.43 | 51.80 |

prediction, the truck eventually vanishes. I^2 -World fails to model a reasonable driving trajectory and ultimately produces an unrealistically elongated truck. In contrast, our method predicts the correct driving behavior and maintains consistent truck geometry across timesteps. It is also important to note that, in our prediction results, the right-turn behavior of the following truck is reasonable, as the previous observations do not provide sufficient guidance for predicting a straight trajectory.

4. End-to-End Training of Occupancy World Model

At first, we freeze the tri-plane VAE and train the prediction module for 48 epochs. We then unfreeze all parameters of both the VAE and the prediction module and perform end-to-end training for an additional 24 epochs. During end-to-end (E2E) training, we observe that the forecasting performance increases while the reconstruction performance decreases. The detailed performance changes during end-to-end training are reported in Table 1. As shown in the table, forecasting accuracy reaches its peak at epoch 16 and then begins to degrade, mainly due to overfitting. To further illustrate the effect of E2E training, we provide a qualitative comparison in Figure 5. As shown in the figure, after E2E training, our method can accurately forecast pedestrians, whereas the variant without E2E training often removes pedestrians in its predictions. Moreover, with E2E training, our occupancy world model better preserves the consistency of roadway features during forecasting. Additionally, E2E training helps the model predict vehicle dynamics more precisely, while the variant without E2E tends to cause vehicles to vanish in the prediction results. These results demonstrate that E2E training contributes to better detail preservation and more accurate forecasting. To illustrate the effect of end-to-end training on the comparison methods DOME [5] and I^2 -World, we present their forecasting results before and after E2E training in Figure 6. As

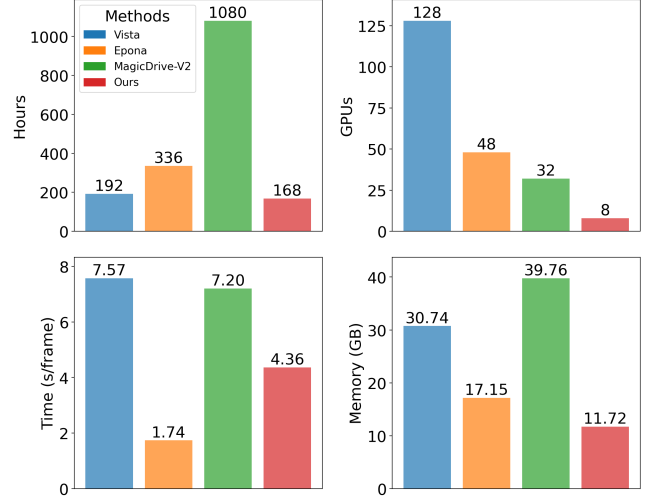


Figure 2. **Comparison of Video Generation Efficiency.** We compare our method with previous approaches in terms of training time, number of GPUs used for training, inference time, and VRAM consumption during inference.

Table 2. **Comparison of Generation Resolutions.** Vista and Epona can generate only single-view driving videos, whereas MagicDrive-V2 and our method can generate multi-view videos.

| Method | Vista [4] | Epona [8] | MagicDrive-V2 [3] | Ours |
|------------|-------------------|-------------------|----------------------------|---------------------------|
| Resolution | 576×1024 | 512×1024 | $6 \times 848 \times 1600$ | $6 \times 256 \times 512$ |

shown, E2E training significantly degrades the forecasting performance of DOME, while I^2 -World shows no improvement and also experiences a decline in prediction accuracy.

5. Efficiency of Video Generation

Previous video-based driving world models or driving video generation models are typically trained from scratch, which is computationally expensive. We summarize the training and inference efficiency of different methods in Figure 2. As shown in the figure, previous works such as Vista [4], Epona [8], and MagicDrive-V2 [3] require a large number of GPUs (32–128) and long training periods (192–1080 h). In contrast, we leverage pretrained video generation models to reduce training cost. With our two-stage generation framework and fine-tuning strategy, the total training time is reduced to one week (3 days for training the first-stage occupancy world model and 4 days for fine-tuning the video model) using only 8 GPUs. Moreover, our method also achieves superior inference efficiency: the average generation speed reaches 4.36 s per frame, and the VRAM consumption is only 11.72 GB. MagicDrive-V2 requires 39.76 GB of VRAM by performing sequence parallelism across 8 GPUs, whereas other methods, including ours, can perform inference on a single GPU. We also list the genera-

tion resolution of each method in Table 2. Vista and Epona generate only single-view videos, while MagicDrive-V2 and our method support six-view video generation.

6. Generality of the Same-Height-Across-Views MVA Design

To efficiently model multi-view consistency, we perform attention across different views at the same height. However, our method is not limited to yaw changes, as multi-view consistency does not rely solely on MVA. Instead, the generated occupancy provides strong guidance for consistent scene appearance across views and for temporal evolution. Therefore, when the vehicle experiences changes in roll or pitch, the occupancy captures these variations and further reflects them in the generated videos. As shown in Fig. 3, our method generalizes well to scenarios with varying roll and pitch angles.

7. More Video Generation Results

7.1. Long Video Generation

Our *GenieDrive-L* produces 81-frame multi-view driving videos, and by applying the rollout operation, it can further generate 241-frame (~ 20 s) sequences—the longest video length in the NuScenes [1] dataset. We provide two representative samples in Figure 7. As shown, even after two rollouts, our method consistently maintains high generation quality and strong multi-view coherence in both daytime and nighttime scenarios.

7.2. Driving Scenario Editing

By editing the occupancy and then generating driving videos guided by the edited occupancy, our method can easily remove or insert objects within driving scenes. To illustrate how our method performs scene editing, we compare the original video with the corresponding edited version in Figure 8. As shown, our method gradually removes the car in the first scene and naturally inserts a truck onto the roadway in the second scene. The edited results appear natural and reasonable, maintaining both spatial and temporal consistency in the generated driving videos. These results further demonstrate that our method enables effective and realistic editing of driving scenarios. This convenient and controllable scene editing capability can greatly enhance out-of-distribution driving data generation.

7.3. Sim-to-Real Driving Scenario Generation

The sim-to-real gap is largely caused by the unrealistic rendering quality of the simulator. However, there is no obvious discrepancy between synthetic occupancy and real-world occupancy. Therefore, we leverage occupancy from the CARLA simulator [2] and use our method to transfer the synthetic occupancy into realistic multi-view driving



Figure 3. Generality on varying roll and pitch angles.

videos. As shown in Figure 9, we visualize the original simulated driving scenes alongside the corresponding sim-to-real results produced by our method. The results show that *GenieDrive* can accurately capture the driving behavior in simulation and generate corresponding realistic outcomes in real-world scenarios. Moreover, our method preserves fine details from the synthetic scenes, such as surrounding vegetation and vehicles. This capability can substantially enhance the realism of synthetic data, thereby further mitigating the sim-to-real gap.

References

- [1] 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 *CVPR*, 2020. 3
- [2] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *CoRL*, 2017. 3
- [3] Ruiyuan Gao, Kai Chen, Bo Xiao, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrive-v2: High-resolution long video generation for autonomous driving with adaptive control. In *ICCV*, 2025. 2
- [4] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. In *NeurIPS*, 2024. 2
- [5] Songen Gu, Wei Yin, Bu Jin, Xiaoyang Guo, Junming Wang, Haodong Li, Qian Zhang, and Xiaoxiao Long. Dome: Taming diffusion model into high-fidelity controllable occupancy world model. *arXiv:2410.10429*, 2024. 1, 2, 6
- [6] Zhimin Liao, Ping Wei, Ruijie Zhang, Shuaijia Chen, Haoxuan Wang, and Ziyang Ren. I2-world: Intra-inter tokenization for efficient dynamic 4d scene forecasting. In *ICCV*, 2025. 1, 6
- [7] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. In *NeurIPS*, 2023. 1
- [8] Kaiwen Zhang, Zhenyu Tang, Xiaotao Hu, Xingang Pan, Xiaoyang Guo, Yuan Liu, Jingwei Huang, Li Yuan, Qian Zhang, Xiao-Xiao Long, et al. Epona: Autoregressive diffusion world model for autonomous driving. In *ICCV*, 2025. 2
- [9] Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. Occworld: Learning a 3d occupancy world model for autonomous driving. In *ECCV*, 2024. 1

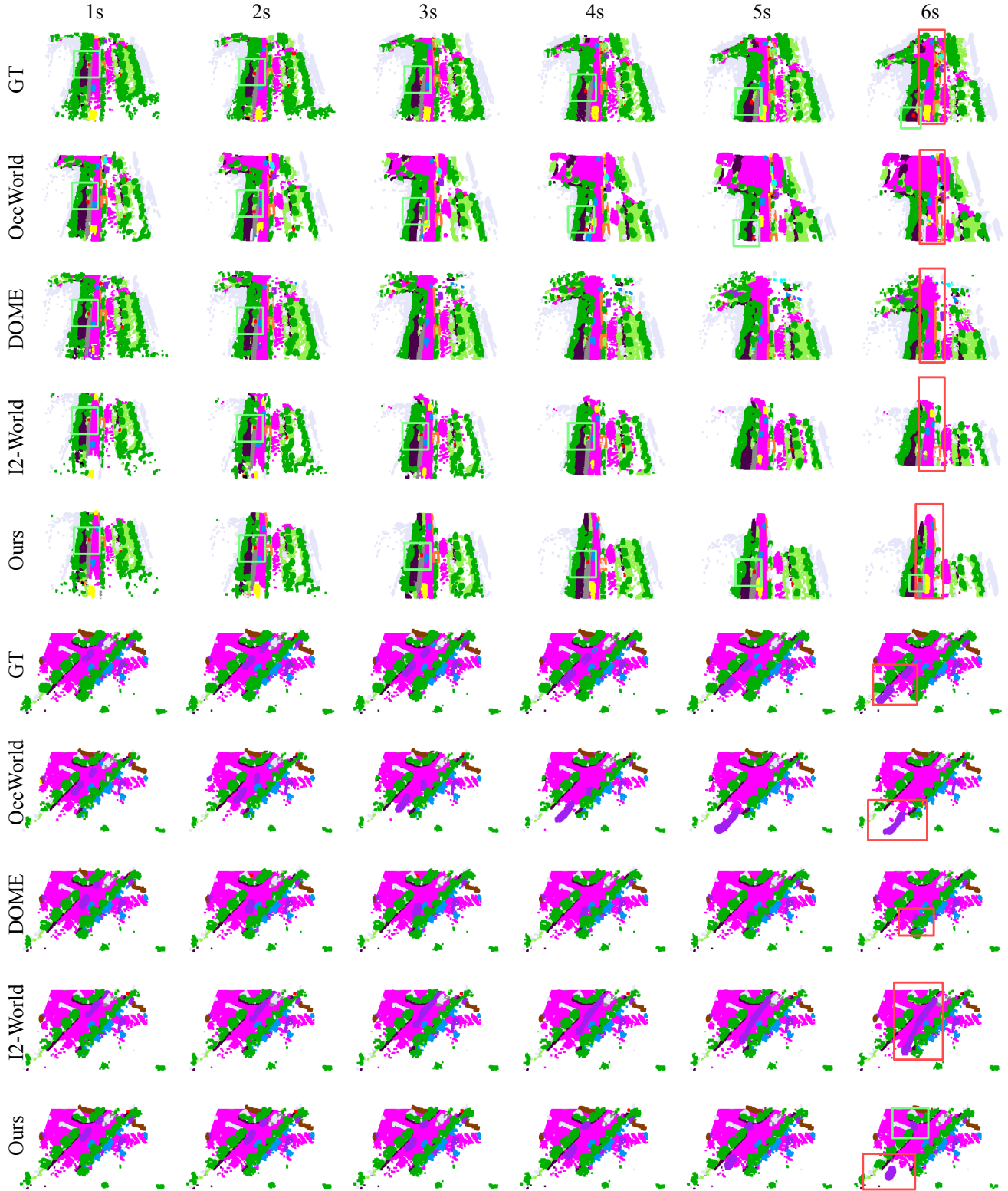


Figure 4. **Qualitative Comparison of 4D Occupancy Forecasting.** We highlight the differences using bounding boxes. Previous methods tend to produce unreasonable predictions or miss important details such as pedestrians. In contrast, our method generates physically reasonable results while preserving the detailed structures of the driving scene.

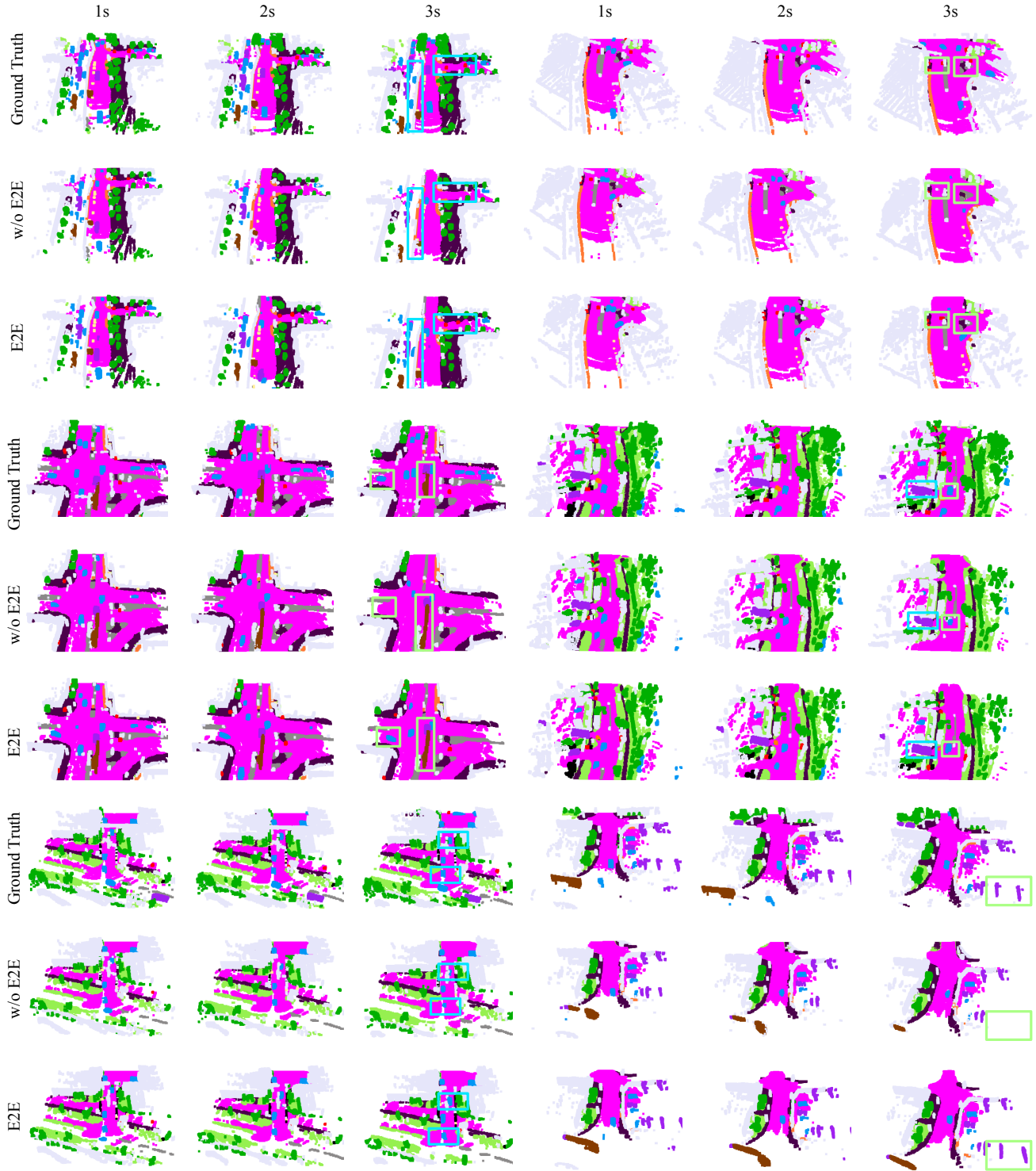


Figure 5. **Qualitative Comparison Before and After End-to-End Training.** We highlight the differences using bounding boxes. The visualization results show that end-to-end training helps our occupancy world model forecast **pedestrians**, **cars**, **trucks**, **trailers**, and **roadway** features more accurately. In contrast, the variant without end-to-end training tends to lose these details in its predictions.

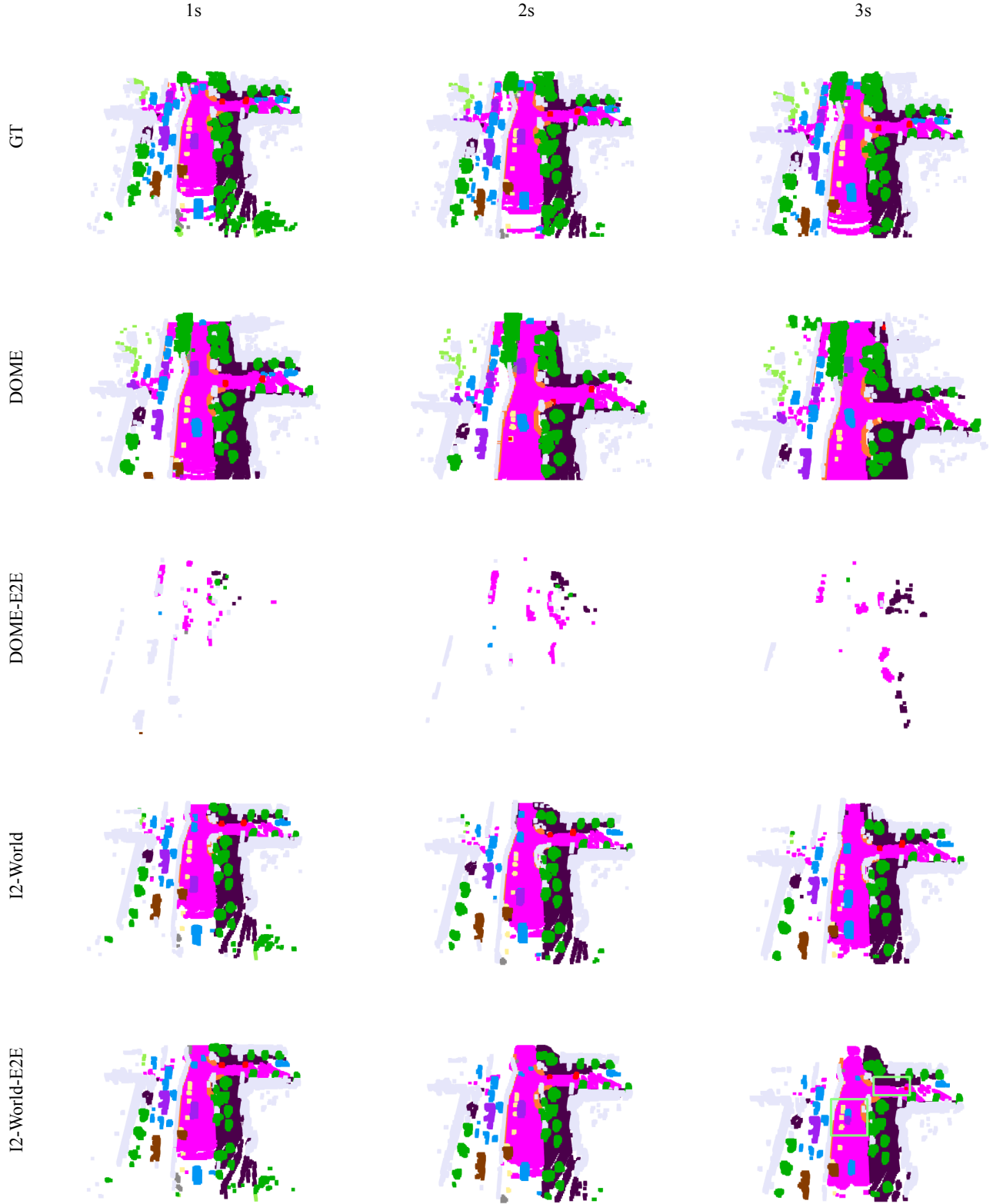


Figure 6. **Effect of End-to-End Training on the Comparison Methods.** We visualize the impact of end-to-end (E2E) training on the comparison methods DOME [5] and I^2 -World [6] by presenting the ground truth along with their predictions before and after E2E training. For DOME, the forecasting capability completely breaks down after E2E training. For I^2 -World, E2E training fails to produce more accurate forecasts and further leads to noticeable loss of scene details.

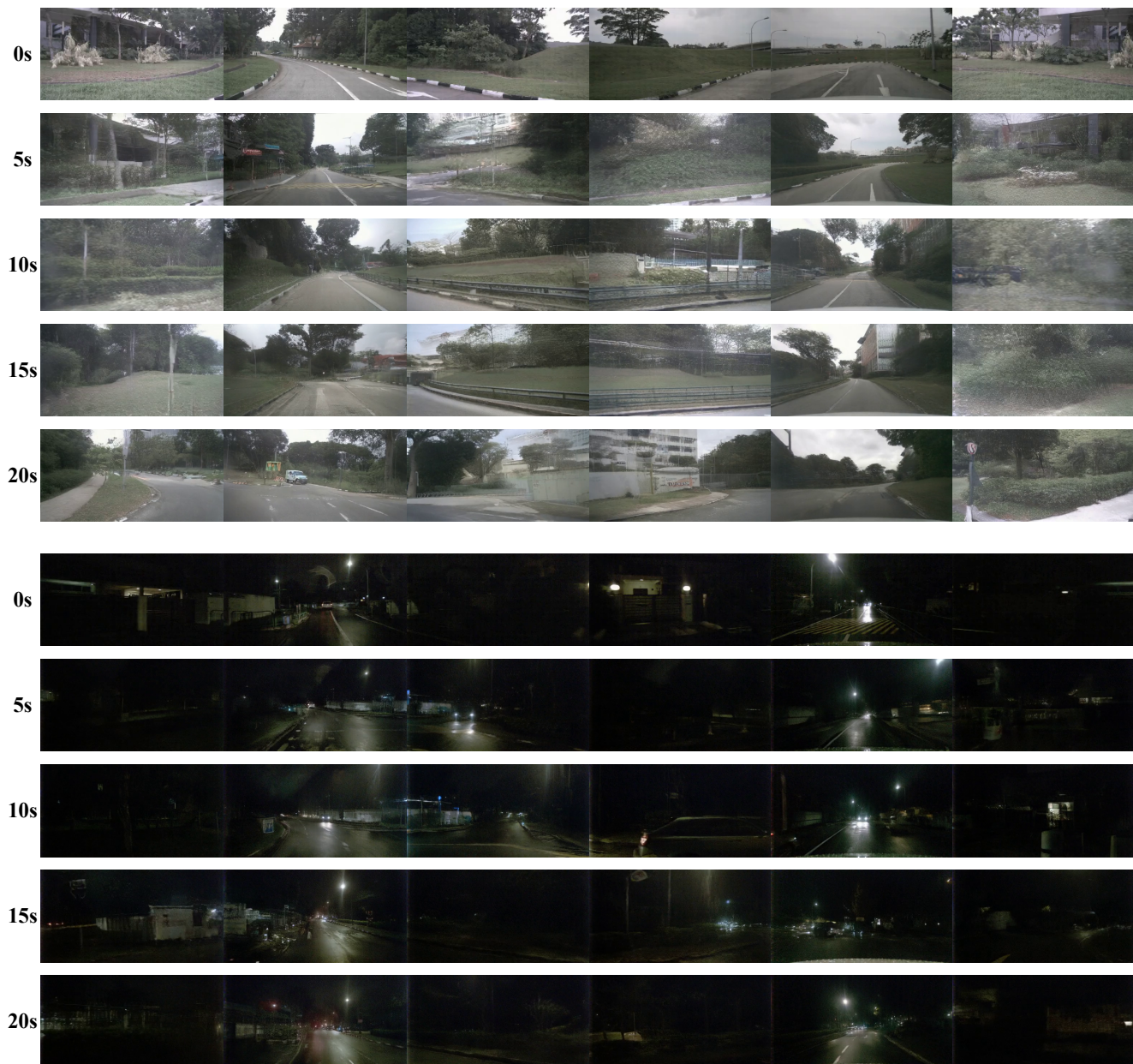
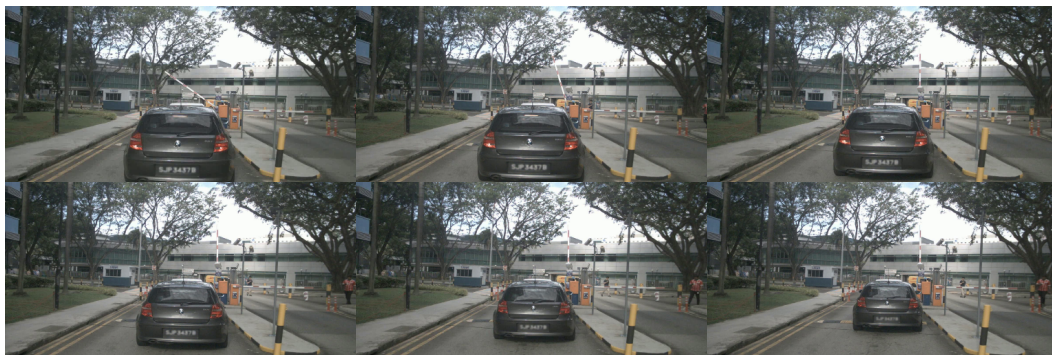
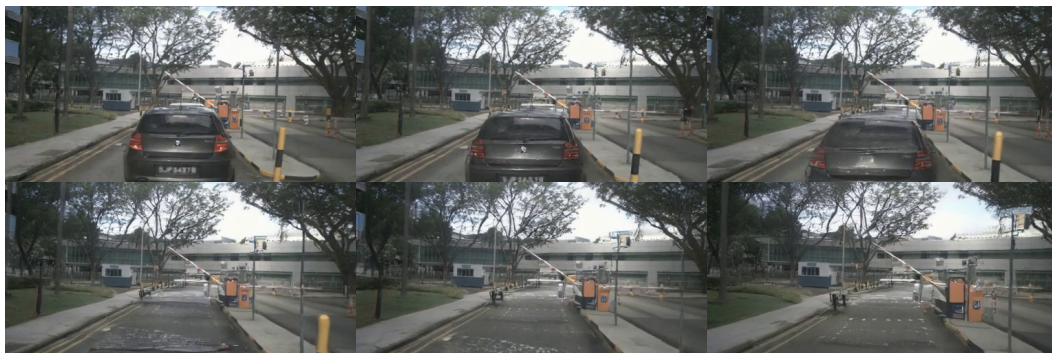


Figure 7. **Long Video Generation Examples.** We provide two examples of our generated 20-second multi-view driving videos: one captured under daytime conditions and the other at night. Our method maintains both generation quality and multi-view consistency even over such long 20-second sequences.

Prior to
Removal



Post
Removal



Prior to
Insertion



Post
Insertion

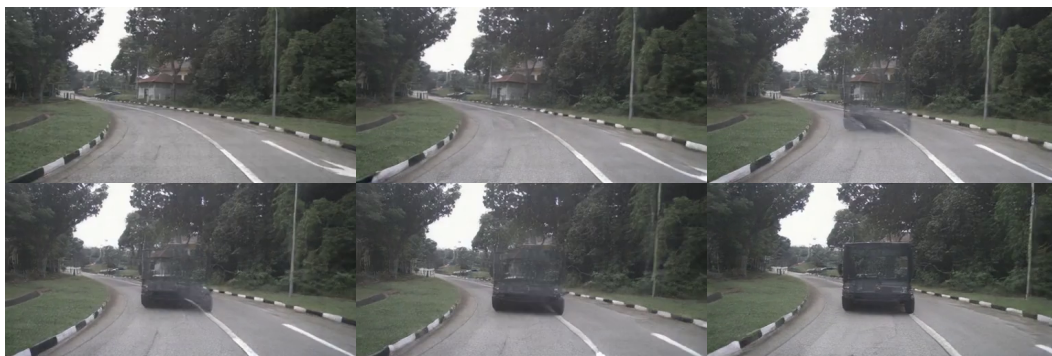


Figure 8. **Driving Scenes Editing.** We visualize the editing process applied to driving videos. Both *Removal* and *Insertion* operations take effect progressively over time. Compared with the original video, the edited results demonstrate that our method can effectively remove and insert objects within driving scenes.



Figure 9. **Sim-to-Real Generation.** The left side shows the BEV map of simulated driving scenes in the CARLA simulator. Our method possesses the ability to transform these simulated scenarios into realistic multi-view driving videos. The visualization results demonstrate that our method not only generates accurate ego-vehicle behaviors, such as *left turns* and *overtaking*, but also preserves important scene details, including surrounding vehicles highlighted with red boxes.