DrivingSphere: Building a High-fidelity 4D World for Closed-loop Simulation

Supplementary Material

6. Implement details

In this section, we provide detailed implementation settings to facilitate the reproducibility of this work. Specifically, we elaborate on the datasets used, evaluation metrics, the architecture of our generation models.

6.1. Datasets

Scene Generation. Our experiments are primarily based on the nuScenes [3] dataset. For the scene generation task, we use the nuScenes-OpenOcc dataset as the data source. Each scene provides complete occupancy annotations and BEV maps, supporting the evaluation of both static and dynamic elements. We also use GPT4-V to obtain the scene description as the text prompts.

Video Generation. Following prior works [13, 28, 44], we use a standard split of 700 scenes for training and 150 scenes for validation. Each sequence is recorded at 12 Hz and lasts approximately 20 seconds, with annotations provided at 2 Hz. To train higher-frequency models, we interpolate the sequences to generate 12 Hz annotations and train models with both 2 Hz and 12 Hz versions. To achieve fine-grained control over the generated scenes and traffic actors' appearance, we utilize GPT-4v to generate detailed scene captions and object captions. These captions provide high-level semantic descriptions of the overall scene and detailed attributes for each traffic actor, enabling precise guidance during video generation. Additionally, we track each foreground actor within a single sequence to assign a unique ID, ensuring appearance consistency across frames in the generated video sequence.

Open-loop and Closed-loop Settings. We support two map environments, *singapore-onenorth* and *boston-seaport*, both aligned with the DriveArena platform. A total of 100 simulation sequences are defined as the validation set to evaluate both open-loop and closed-loop generation performance.

6.2. Evaluation Metrics

Frechet Video Distance (FVD). This metric evaluates the visual quality and temporal consistency of generated video clips, following prior methods [13, 45].

Mean Average Precision (mAP) and NuScenes Detection Score (NDS). We adopt mAP and NDS to assess the detection accuracy on generated data to validate the fidelity.

We utilize metrics from [8, 50] for comprehensive evaluation in both open-loop and closed-loop setups: Progressive Driving Metric Suite (PDMS): Initially proposed by NavSim, PDMS evaluates trajectory outputs at each timestep based on the following criteria:

No Collisions (NC): Measures whether the agent avoids collisions with road users. Drivable Area Compliance (DAC): Assesses whether the agent remains within the drivable area. Ego Progress (EP): Quantifies how effectively the agent progresses along its intended route. Time-to-Collision (TTC): Evaluates the safety of the trajectory in terms of time remaining before a collision occurs. Comfort (C): Ensures the smoothness of the driving trajectory, minimizing abrupt accelerations or turns. Arena Driving Score (ADS): ADS integrates trajectory-level performance (PDMS) with route completion to provide a holistic metric. The Route Completion (Rc) is defined as the percentage of the total route distance completed by the agent, where [0 , 1] Rc[0,1]. ADS is particularly suited for closed-loop evaluation, as it considers both safety and consistency. For instance, collisions or deviations from the road terminate the simulation, making ADS an effective differentiator of agent performance.

6.3. Model Details

OccDreamer. For the scene tokenizer $\mathcal{F}_{\text{VAE}}^{\text{occ}}$, we following [41, 57] and train a 3D occupancy VAE, which takes the occupancy data with the size of 192 × 192 × 16. $\mathcal{F}_{\text{VAE}}^{\text{occ}}$ compresses the data $\boldsymbol{S}_k$ to a latent space $\boldsymbol{Z}^{\boldsymbol{S}_k}$ with a dimension of 48 × 48 × 4 and the channel is set to 8. For the BEV map, a pre-trained encoder [16, 17] encodes the BEV representation at the same resolution as the latent feature.

As for the denoiser ϵ_θ^s and the ControlNet branch ϵ_ϕ^s, we use 3D U-Net [16, 17] as the backbone for the 3D input data. For the basic scene generation model, we train ϵ_θ^s and ϵ_ϕ^s for 60k iterations on 8 NVIDIA A800 GPUs. For the scene extension version, we freeze ϵ_θ^s and fine-tune ϵ_ϕ^s with extra channels to take the partial scene as the condition. In the inference stage, we adopt DDIM [16, 17] with 100 steps sampling. Additionally, we set the classifier-free guidance scale as 7 for the condition.

Videodreamer. Our implementation is based on the Open-Sora codebase [59], initialized with pre-trained weights. The training process is carried out on 8 NVIDIA A800 GPUs, comprising 30k iterations. As for the 4D occupancy encoder $\mathcal{F}_{\text{VAE}}^{\text{4Docc}}$, we also borrow the network architecture from [41, 57]. $\mathcal{F}_{\text{VAE}}^{\text{4Docc}}$ takes 4D occupancy data as the input and extract the embedding. The number of DiT blocks is set to 26 and $N = 13$. For inference, we utilize rectified flow [59] with a classifier-free guidance scale of 7.0, performing 30 sampling steps to generate videos at various resolutions from 480p to 1080p. For the open-loop and closed-loop evaluation, we generate the video of 4 frames

Methods	Downsampling Scale	IoU$_{(\uparrow)}$	mIoU$_{(\uparrow)}$
OccWorld [57]	$H/4 \times W/4 \times T$	62.29	66.38
OccSora [41]	$H/8 \times W/8 \times T/8$	27.4	37
DrivingSphere$_{4D}$	$H/4 \times W/4 \times T$	**93.1**	**73.89**
Semcity [22]	-	95.8	76.9
DrivingSphere$_{3D}$	$H/4 \times W/4$	**97.2**	**86.81**

Table 5. **Quantitative results of Occupancy Tokenizer for Occupancy Reconstruction.** *DrivingSphere*$_{4D}$ indicates $\mathcal{F}_{\text{VAE}}^{\text{4Docc}}$ in Sec. 3.2 while *DrivingSphere*$_{4D}$ indicates $\mathcal{F}_{\text{VAE}}^{\text{occ}}$ in Sec. 3.1.

Methods	FVD	mAP$_{(\uparrow)}$	NDS$_{(\uparrow)}$
RealData [3]	-	62.29	66.38
MagicDrive [13]	-	12.30	23.32
DriveDreamer [43]	340.8	-	-
Panacea [45]	139	11.58	22.31
Drive-WM [44]	122.7	20.66	-
DrivingSphere w/o $\boldsymbol{W}$	121.4	17.34	26.21
DrivingSphere	**103.4**	**22.71**	**31.79**

Table 6. **Comparison of SOTA video generation methods on nuScenes validation set.** We use BEVFusion as the 3D detector. *'w/o W'* indicates that the model uses no occupancy but uses the 2D sketch as the condition.

with $f = 3$ frames as the condition while generating long videos, we generate the 16 frames video sequence and generally use $f = 4$ frames as the condition.

7. Additional Quantitative Results

7.1. Scene Reconstruction

To validate the effectiveness of our Occupancy VAE, we conduct scene reconstruction experiments on the nuScenes validation set. These experiments evaluate both 3D and 4D scene reconstruction, providing a comprehensive analysis of the model's capability. As shown in Tab. 5, for 3D scene reconstruction, we compare our trained 3D Occupancy VAE with SemCity, which is utilized as the occupancy tokenizer in Section 3.1. The results demonstrate that our Occupancy VAE achieves superior performance, highlighting its ability to encode and reconstruct 3D occupancy data effectively. For 4D scene reconstruction, we benchmark against OccWorld and OccSora, widely regarded as state-of-the-art methods for large-scale 4D occupancy generation. The results clearly show that our Occupancy VAE outperforms these approaches across all evaluation metrics, establishing new benchmarks for 4D scene reconstruction quality. These significant improvements are primarily attributed to the carefully designed network architecture, including the Projection Module and Expansion & Squeeze Strategy, as well as meticulously tuned experimental parameters. Such architectural innovations enable the model to capture fine-grained spatial and temporal information, ensuring accurate and efficient reconstruction of both static and dynamic elements within the scenes. This experiment underscores the robustness and effectiveness of the proposed framework in handling complex 3D and 4D scene representations.

7.2. Video Generation

To further validate the capabilities of VideoDreamer, we align our experimental settings with state-of-the-art video generation methods, ensuring a fair comparison. As presented in Tab. 6, we employ BEVFusion as the detector to quantitatively evaluate the visual fidelity of the generated videos. The results demonstrate that our method achieves superior performance, highlighting its ability to generate high-quality and visually coherent driving scenarios. Additionally, we conduct an ablation study by introducing the configuration "w/o W", which uses only 2D sketches as conditions without incorporating occupancy data. This ablation effectively isolates the contribution of the 4D driving world to the video generation process. The results clearly illustrate the significant improvement in visual fidelity when occupancy data is integrated, confirming the critical role of the occupancy condition in enhancing the realism and consistency of generated video sequences. This experiment underscores the robustness of our framework in producing visually accurate driving videos and its ability to leverage multi-modal conditions effectively.

8. Additional Visualtion Results

In this section, we provide more quality visualization results and a video is also attached in the materials for better visualization of temporal results.

8.1. Scene Generation

In Fig. 7, we present a comparison between the occupancy scenes generated by our method, SemCity, and real-world data. The visual results clearly demonstrate that our method achieves significantly higher fidelity compared to SemCity, closely approximating the structural and semantic layout of real-world data. It is important to note that SemCity is an unconditional generation method, and as such, its outputs are unpaired with the real data used for comparison. In contrast, our method leverages conditions, ensuring consistency with the road structures and semantic layouts of the real data. This alignment highlights the strength of our approach in generating occupancy scenes that are not only visually realistic but also semantically coherent, demonstrating its suitability for tasks requiring precise scene understanding and reconstruction.

8.2. Video Generation

Controllable Video Generation In Fig. 11, we showcase the results of video generation spanning 40 frames. The

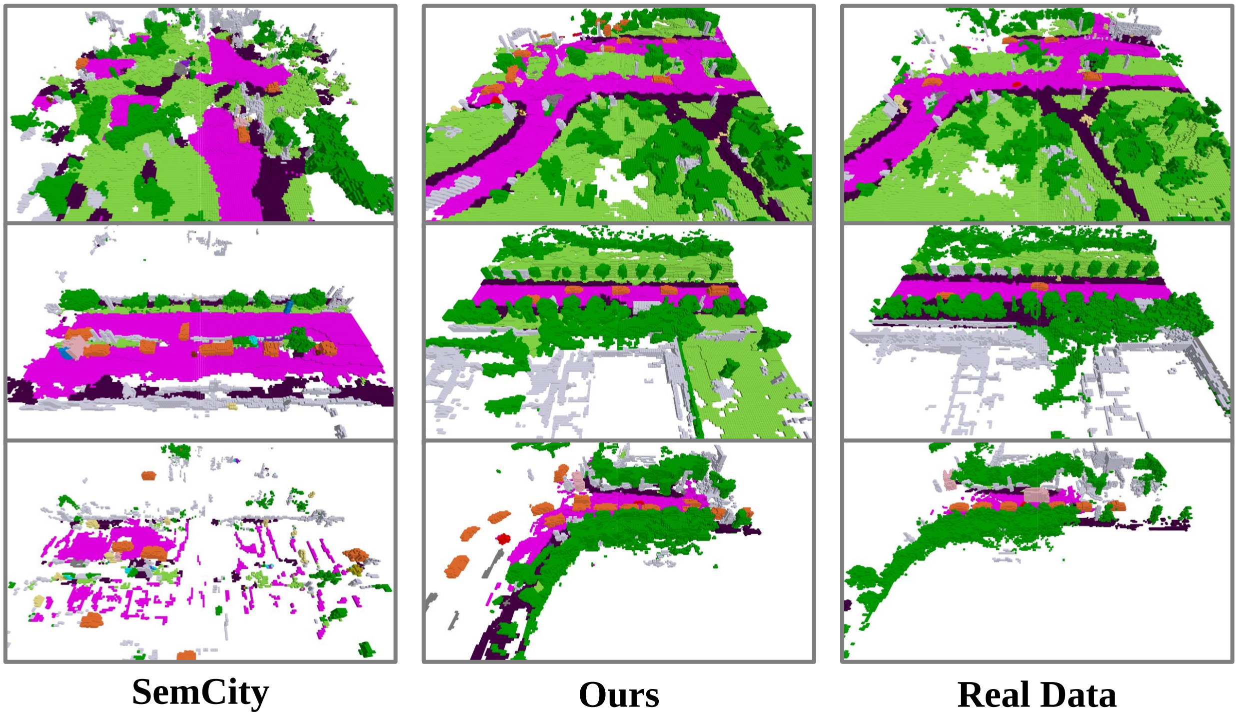

Figure 7. **Comparison between Semcity [22], *DrivingSphere* and Real Data.**

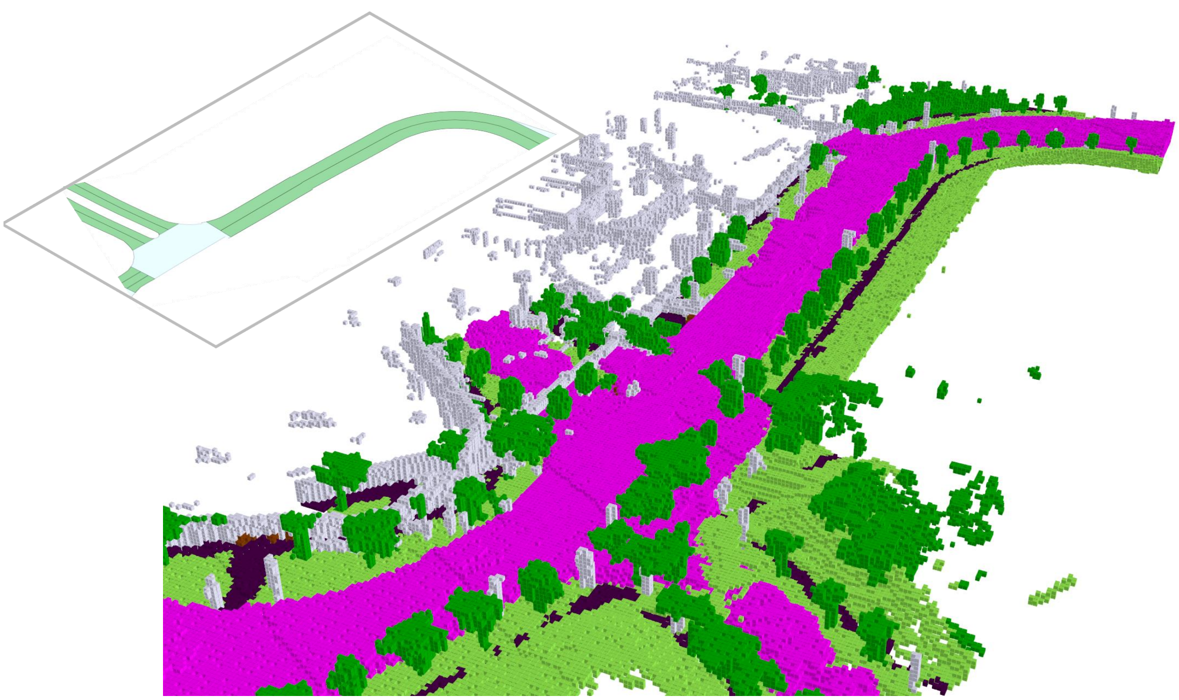

Figure 8. **Composited Driving World in a specific area.** We adpot Scene Generation and Scene Extention in Sec. 3.1 to obtain a big static background.

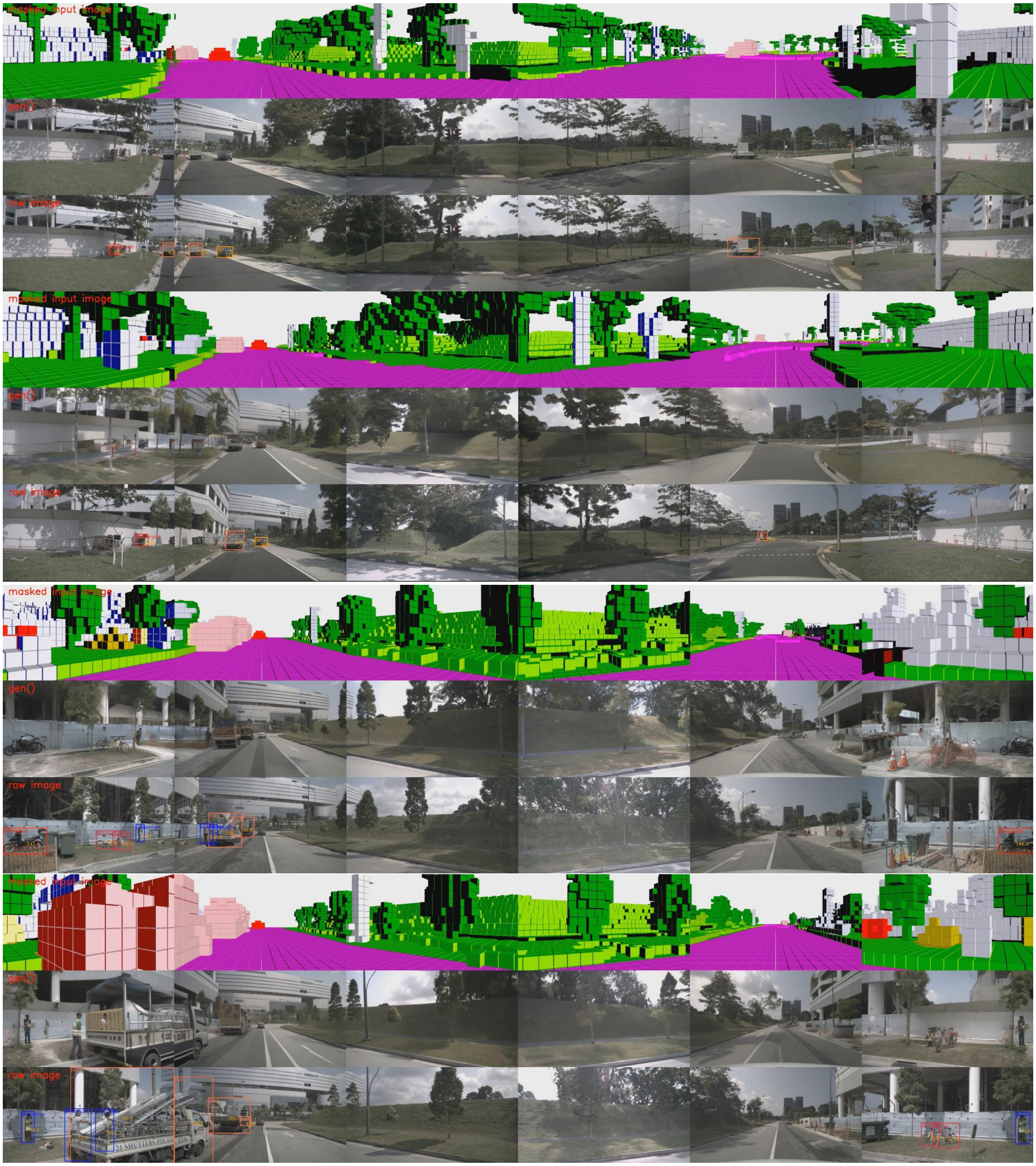

Figure 9. **Generated Video sequnences in nuScene.** Top: Occupancy condition, Middle: Our generated video, Bottom: Ground truth video sequence.

generated sequences demonstrate the effectiveness of our method in accurately modeling not only the occlusions and depth relationships of foreground objects but also in precisely controlling the generation of non-direct traffic participants, such as trees, buildings, and man-made landmarks.

Our approach leverages the accurate control provided by

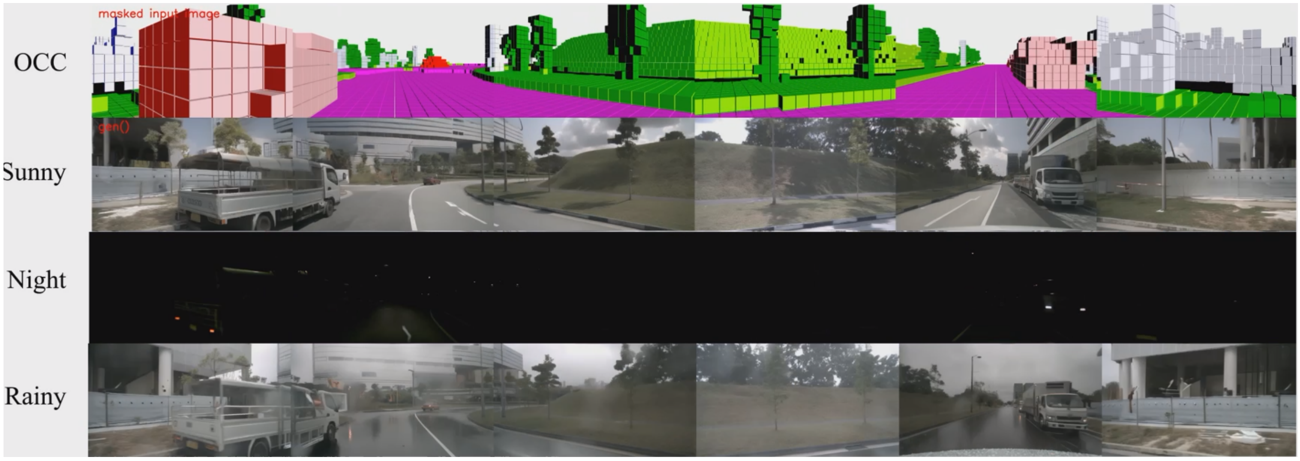

Figure 10. **Controllable Generation with scene captions.** The visual reuslts vary with the give scene description.

occupancy data, enabling consistent and realistic representation of both dynamic and static elements within the scene. This capability highlights the robustness of our model in generating complex and coherent driving environments over extended temporal horizons, making it well-suited for real-world applications that require high fidelity and detailed scene understanding.

Simulation Results

Long-term Video Generation We also provide a demo of ultra-long video generation on private data (refer to the attached video). The demo showcases a video generated at 10 Hz over a duration of 1 minute, resulting in an impressive 600 frames of continuous generation.

9. Limitations and Future Work

Optimizing the computational pipeline for generating and rendering 4D occupancy and video data will be a key focus. Techniques such as model pruning, quantization, and adaptive sampling will be explored to reduce computational overhead without compromising fidelity. Additionally, enabling real-time rendering capabilities will make the system more practical for online validation.

Expanding the diversity of simulated environments is critical for robustness testing. Future work will aim to model a broader range of conditions, including extreme weather (e.g., heavy rain, snow, and fog), varying road geometries, and rare traffic scenarios. This enhancement will allow the simulation to evaluate autonomous driving systems under more comprehensive and challenging conditions.

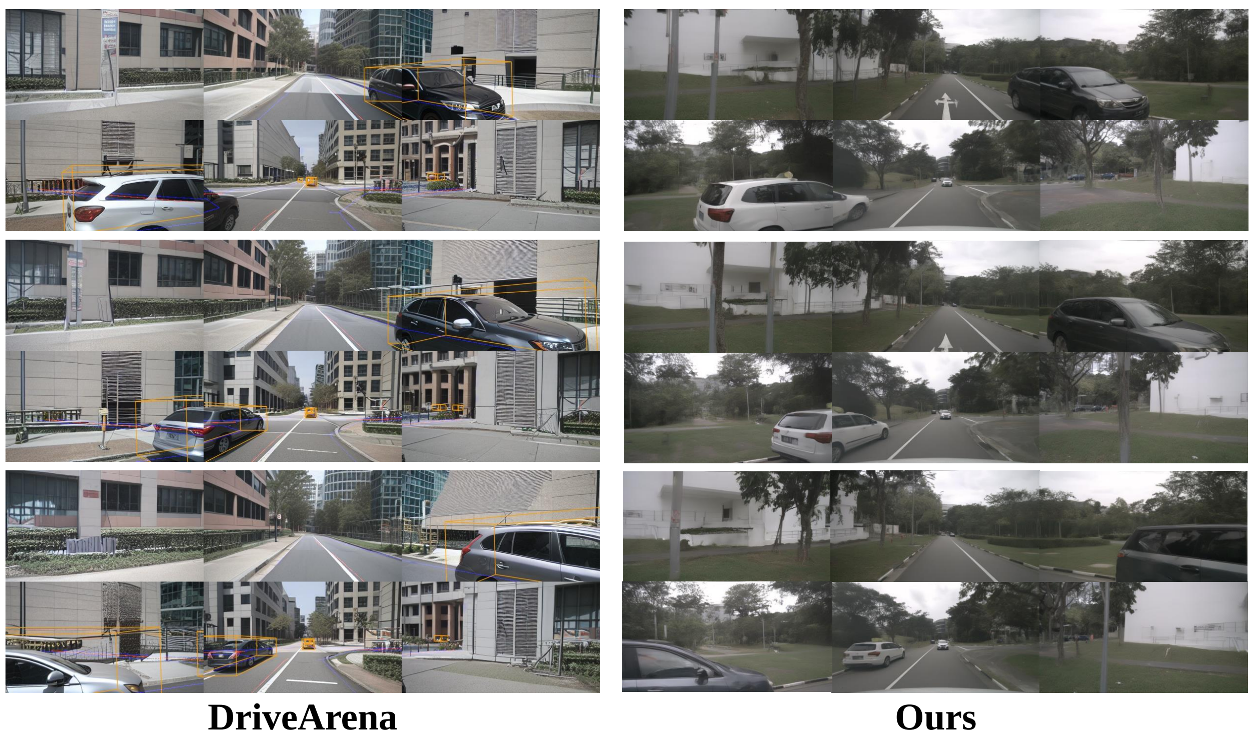

Figure 11. **Comparison with DriveArena [50].** The visual output of DriveArena and *DrivingSphere* on the same route demonstrates superior temporal and spatial consistency in generated simulations.