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

Supplementary Material

6. Experiment Settings

This section provides detailed implementation settings to facilitate reproducibility, including dataset descriptions, evaluation metrics, and model architecture specifications.

6.1. Datasets

Scene Generation. Our experiments are primarily conducted using the nuScenes [3] dataset. For scene generation tasks, we leverage the nuScenes-OpenOcc dataset, which provides comprehensive occupancy annotations and BEV maps, supporting the evaluation of both static and dynamic elements. Additionally, we employ GPT-4V to caption corresponding RGB images and generate textual scene descriptions as prompts for conditional generation.

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, recorded at 12 Hz, spans approximately 20 seconds, with annotations provided at 2 Hz. To train higher-frequency models, we interpolate sequences to produce 12 Hz annotations. Fine-grained control over scenes and actors is achieved by using GPT-4V to generate detailed scene and object captions. These captions provide high-level semantic descriptions and ensure precise guidance during generation. To maintain consistency, each foreground actor is assigned a unique ID, ensuring appearance coherence across frames.

Open-loop and Closed-loop Settings. We support two map environments, *singapore-onenorth* and *boston-seaport*, aligned with the DriveArena platform [51]. A total of 100 simulation sequences are defined for validation, enabling the evaluation of driving agents in both open-loop and closed-loop modes.

6.2. Evaluation Metrics

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

Mean Average Precision (mAP) and NuScenes Detection Score (NDS). We use mAP and NDS to measure detection accuracy on generated data, validating the fidelity of the simulated environment.

Progressive Driving Metric Suite (PDMS): Initially proposed by NavSim [8], PDMS evaluates trajectory performance at each timestep using metrics such as **No Collisions (NC)**, **Drivable Area Compliance (DAC)**, **Time-to-Collision (TTC)**, and other relevant indicators.

Arena Driving Score (ADS): ADS [51] combines trajectory-level metrics (e.g., PDMS) with route completion (R_c), defined as the percentage of total route distance completed ($R_c \in [0, 1]$). ADS is particularly suited for closed-loop evaluation as it accounts for safety and consistency. For instance, collisions or road deviations terminate simulations, making ADS a reliable measure of agent performance.

6.3. Model Details

OccDreamer. The scene tokenizer $\mathcal{F}_{\text{VAE}}^{\text{occ}}$ follows [41, 58] and is trained on 3D occupancy data of size $192 \times 192 \times 16$. It compresses the input $\boldsymbol{S}_k$ into a latent space $\boldsymbol{Z}^{\boldsymbol{S}_k}$ of size $48 \times 48 \times 4$ with 8 channels. A pre-trained encoder [16, 17] processes the BEV map to match the latent feature resolution.

For the denoiser ϵ_θ^s and ControlNet branch ϵ_ϕ^s, we employ 3D U-Net [16, 17] as the backbone. Training involves 60k iterations on 8 NVIDIA A800 GPUs. For scene extension, ϵ_θ^s is frozen, and ϵ_ϕ^s is fine-tuned with additional channels to condition on partial scenes. During inference, DDIM [17] with 100 sampling steps is used, and the classifier-free guidance scale is set to 7.

VideoDreamer. Based on the OpenSora codebase [60], our implementation initializes with pre-trained weights and is trained for 30k iterations on 8 NVIDIA A800 GPUs. The 4D occupancy encoder $\mathcal{F}_{\text{VAE}}^{\text{4Docc}}$ adopts the architecture from [41, 58], extracting embeddings from 4D occupancy data. The number of DiT blocks is set to 26 with $N = 13$.

During inference, rectified flow [60] is used with a classifier-free guidance scale of 7.0 and 30 sampling steps to generate videos at resolutions from 480p to 1080p. For open-loop and closed-loop evaluations, short videos of 4 frames are generated with $f = 3$ reference frames. For longer sequences, 16-frame videos are generated with $f = 4$ reference frames to ensure temporal consistency.

6.4. Simulation Settings.

In our implementation, the traffic flow engine [46] operates at a frequency of 10 Hz, while the control signals are set to 2 Hz, following the setup in DriveArena [51]. Every 0.5 simulation seconds, the 4D driving world updates its state and renders multi-view semantic maps as conditions for the VideoDreamer model. VideoDreamer uses the last 3 frames as reference images to generate 512×960 images, which are subsequently resized to 224×400 to serve as input for the driving agent.

Methods	Downsampling Scale	IoU$_{(\uparrow)}$	mIoU$_{(\uparrow)}$
OccWorld [58]	$H/4 \times W/4 \times T$	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 $\mathcal{F}^{\text{4Docc}}_{\text{VAE}}$ in Sec. 3.2 while *DrivingSphere*$_{3D}$ indicates $\mathcal{F}^{\text{occ}}_{\text{VAE}}$ 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 ***W***	121.4	17.34	26.21
DrivingSphere	**103.4**	**22.71**	**31.19**

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.

7. Additional Quantitative Results

7.1. Scene Reconstruction

To validate the performance of our Occupancy VAE, we conduct scene reconstruction experiments on the nuScenes validation set, evaluating both 3D and 4D scene reconstruction. These tests provide a detailed analysis of the model's ability to represent and reconstruct complex spatial and temporal elements within driving environments.

As shown in Tab. 5, for 3D scene reconstruction, our 3D Occupancy VAE significantly outperforms SemCity [22], which serves as the occupancy tokenizer baseline in Sec. 3.1. The superior performance highlights the enhanced encoding and reconstruction capabilities of our method.

For 4D scene reconstruction, we benchmark against OccWorld and OccSora, two state-of-the-art methods for large-scale 4D occupancy modeling. The 4D occupancy VAE will act as the encoder to extract the global embedding of occupancy data in Sec. 3.2. These architectural components, coupled with finely tuned experimental parameters, enable our model to capture fine-grained spatial and temporal details, ensuring accurate reconstruction of both static and dynamic elements in the scenes.

7.2. Video Generation

To validate the capabilities of VideoDreamer, we benchmark our method against state-of-the-art video generation approaches using aligned experimental settings for fairness. As shown in Tab. 6, we use BEVFusion as the detector to quantitatively assess the visual fidelity of the generated videos. The results highlight the superior performance of our method in generating high-quality, visually coherent driving scenarios.

We also perform an ablation study with a "w/o W" configuration, which conditions video generation solely on 2D sketches without incorporating occupancy data. This setup isolates the impact of the 4D driving world on the generation process. The results clearly show a significant enhancement in visual fidelity with occupancy data integration, emphasizing its crucial role in improving the realism and consistency of generated video sequences. These findings underscore the robustness of our framework in producing visually accurate driving videos and its effective use of multi-modal conditions.

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 compare occupancy scenes generated by *DrivingSphere*, SemCity [22], and real-world data. The results illustrate that our method achieves substantially higher fidelity than SemCity, closely mirroring the structural and semantic layouts of real data. Unlike the unconditional generation of SemCity, our approach leverages conditions to align with real road structures and semantic layouts, underscoring its capability for precise scene reconstruction and realistic understanding.

In Fig. 8, we illustrate the composited driving world created for a specific area. By employing the Scene Generation and Scene Extension strategies outlined in Sec. 3.1, we construct a large-scale static background with seamless spatial consistency.

8.2. Video Generation

Controllable Video Generation Fig. 9 showcases video generation results across 40 frames, highlighting *DrivingSphere*'s ability to model occlusions, depth relationships, and non-direct traffic participants such as buildings, trees, and landmarks. Our model uses occupancy data for precise control over both static and dynamic elements, ensuring consistent and realistic scene representation over time. These capabilities make our framework robust for generating complex driving environments and suitable for real-world applications requiring detailed scene understanding.

Simulation Results As demonstrated in Fig. 11, we compare the generated simulations of DriveArena [51] and *DrivingSphere* on the same route. The results clearly show that *DrivingSphere* outperforms DriveArena in terms of temporal and spatial consistency, further establishing its su-

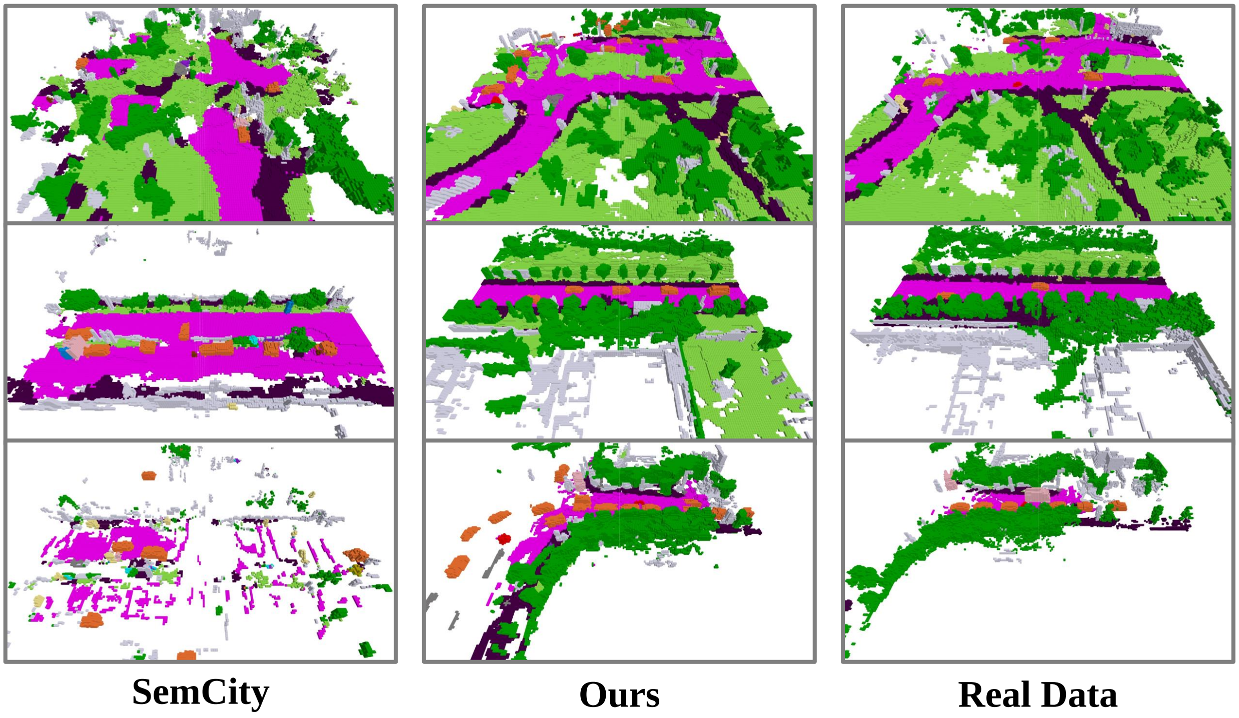

Figure 7. **Comparison between Semcity [22], *DrivingSphere* and Real Data.**

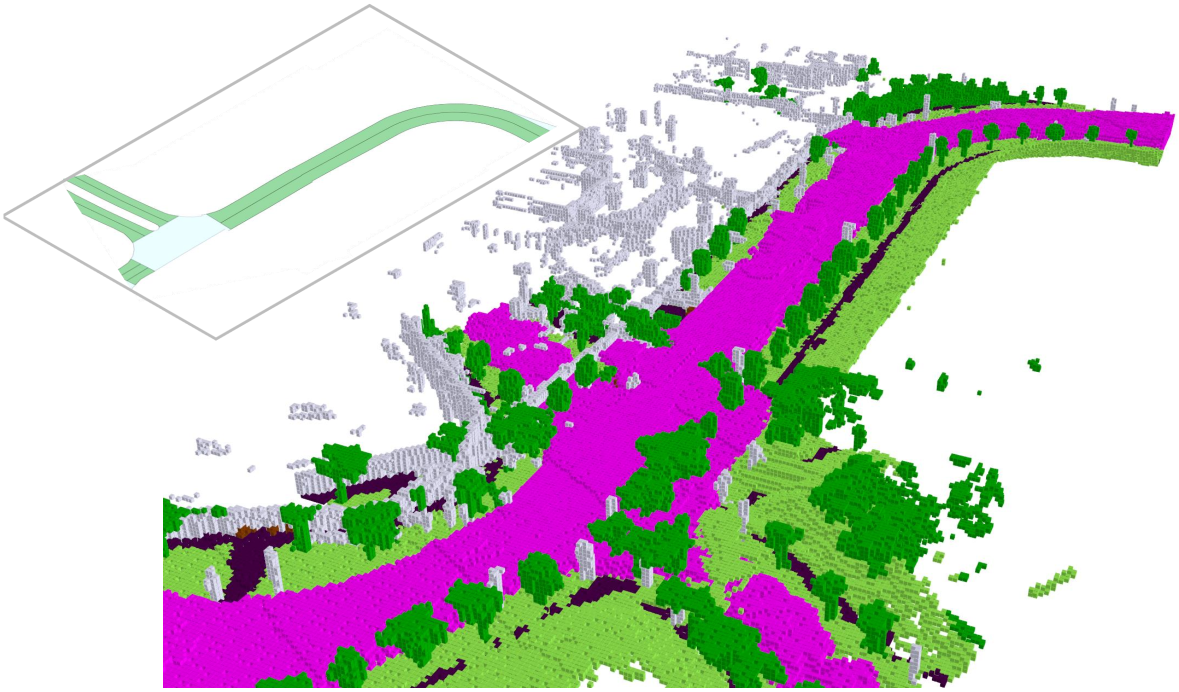

Figure 8. **Composited Driving World in a specific area.** We adpot Scene Generation and Scene Extention in Sec. 3.1 to obtain a big static background.

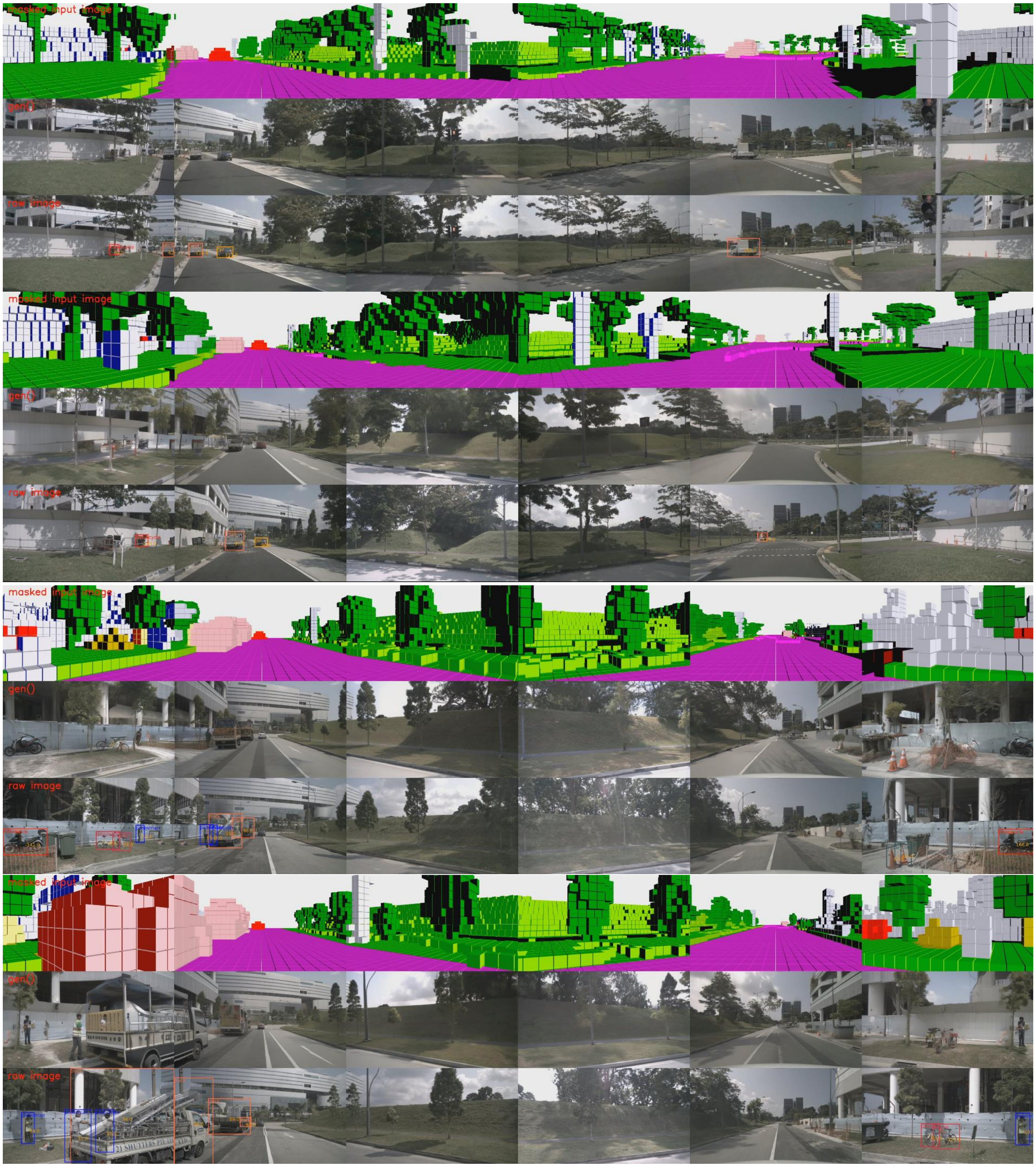

Figure 9. **Generated Video sequnences in nuScene.** Top: Occupancy condition, Middle: Our generated video, Bottom: Ground truth video sequence.

periority in generating coherent and realistic driving simulations.

Long-term Video Generation In an attached video demo, we present ultra-long video generation on private data. This example generates 600 continuous frames at 10 Hz over a 1-minute duration, showcasing *DrivingSphere*'s

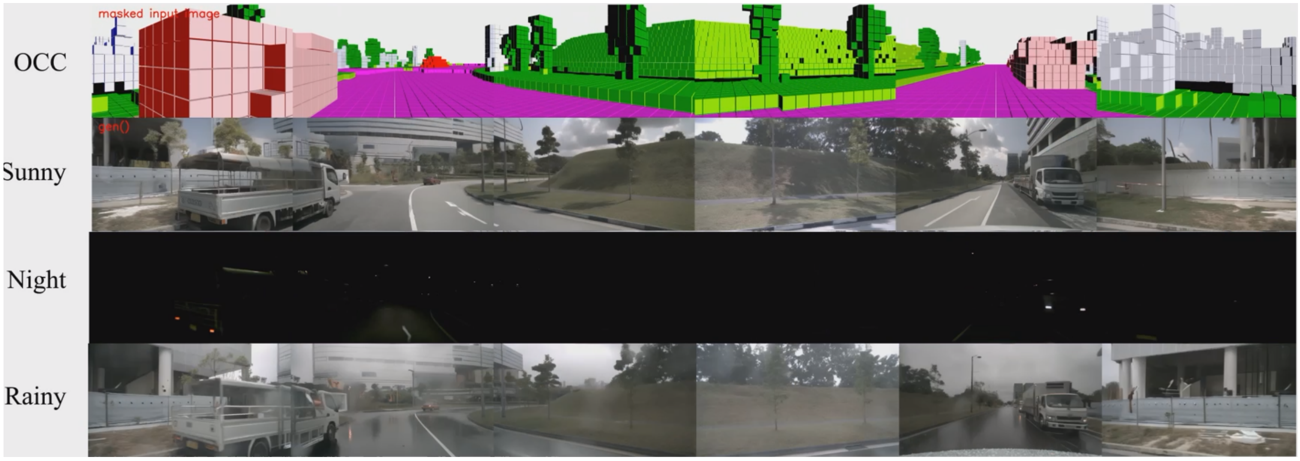

Figure 10. **Controllable Generation with scene captions.** The visual reuslts vary with the give scene description.

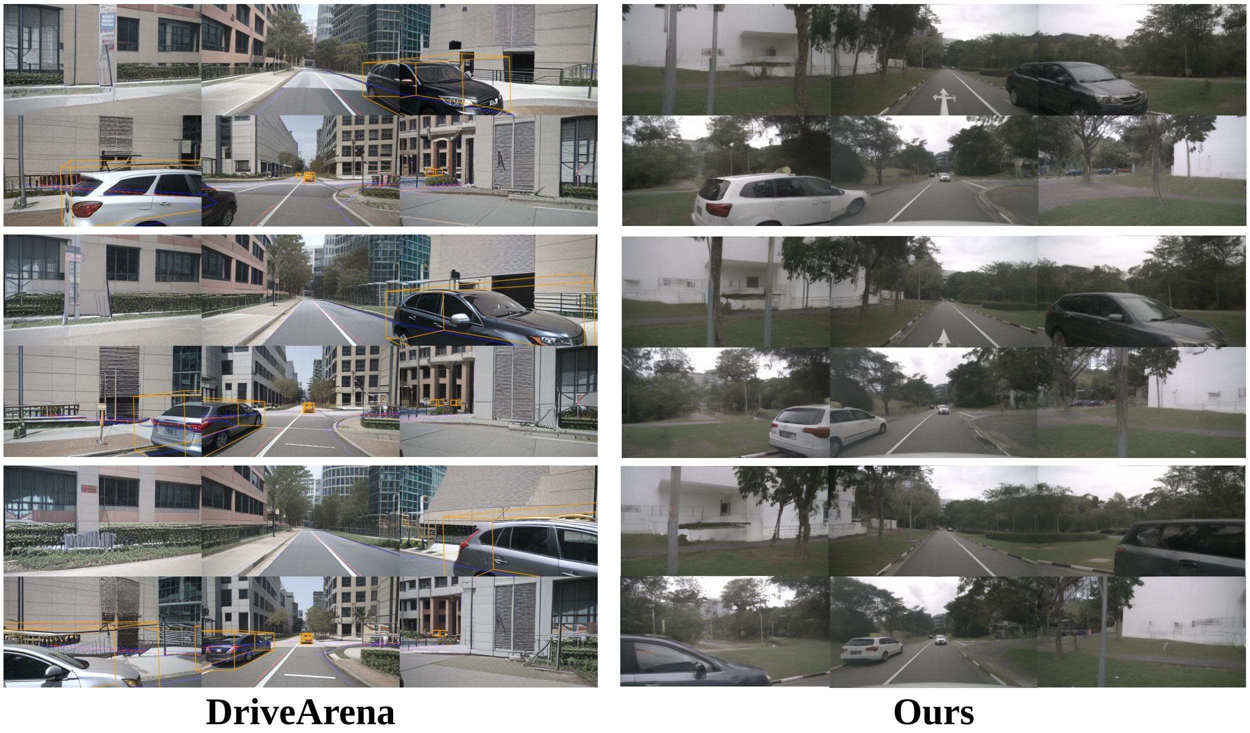

Figure 11. **Comparison with DriveArena [51].** The visual output of DriveArena and *DrivingSphere* on the same route demonstrates superior temporal and spatial consistency in generated simulations.

capacity for maintaining high fidelity and consistency over extended temporal horizons.

9. Limitations and Future Work

Efforts to optimize the computational efficiency of 4D occupancy and video generation pipelines will be central to future work. Techniques like model pruning, quantization, and adaptive sampling will be explored to minimize computational costs while maintaining high fidelity. Real-time rendering capabilities will also be prioritized to facilitate online validation.

Expanding environmental diversity in simulations will be another focus area. Future enhancements will include modeling extreme weather conditions (e.g., heavy rain, snow, fog), varying road geometries, and rare traffic scenarios. These improvements aim to enable more comprehensive robustness testing for autonomous driving systems under challenging and diverse conditions.