DriveX: Omni Scene Modeling for Learning Generalizable World Knowledge in Autonomous Driving

Supplementary Material

Contents

6. Implementation Details

In this section, we provide more implementation details about task definitions (Section 6.1), network architectures (Section 6.2), training schemes (Section 6.3), and image semantic label generation (Section 6.4).

6.1. Task Definitions

Point Cloud Forecasting. Our DriveX model can generate future point clouds by estimating the depth of LiDAR-view rays. Following [26, 63, 70, 73], the ground truth points at future timestamps are used to initialize rays for point cloud generation. We evaluate the performance on nuScenes [3] validation set using the Chamfer Distance between predicted and ground truth point sets. In line with [63, 73], we focus on evaluating points within [-51.2m, -51.2m, -5.0m, 51.2m, 51.2m, 3.0m], a common range used in various downstream tasks.

End-to-End Driving. We conduct end-to-end driving evaluation on NAVSIM dataset, a challenging non-reactive simulation benchmark that provides reliable planning assessment. NAVSIM is built on the large-scale public driving dataset OpenScene [7] and employs a resampling strategy to reduce the prevalence of simple scenarios, resulting in a curated dataset of 103k training samples and 12k test samples. Each sample includes a trajectory with the current frame, 3 historical frames, and 8 future frames at a frequency of 2Hz. For DriveX training, we aggregate the 3 historical frames and 4 future frames from each sample, filtering overlapping frames and short sequences to construct a dataset of 250k frames. Leveraging the nonreactive simulations, NAVSIM employs closed-loop metrics centered on the Predictive Driver Model Score (PDMS) to assess performance, which is a weighted combination of five subscores: no at-fault collisions (NC), drivable area compliance (DAC), time-to-collision (TTC), comfort (Comf.), and ego progress (EP).

Occupancy Prediction and Flow estimation. We apply the DriveX model to the occupancy and occupancy flow prediction task on FlowOcc3D [30] benchmark. FlowOcc3D extends Occ3D by incorporating occupancy-level flow annotations derived from object bounding boxes, alongside the original occupancy labels. The prediction range in the ego vehicle coordinate system is [-40m, 40m] for X, Y axes and [-1m, 5.4m] for Z axis, with voxel size of 0.4m and occupancy resolution of $200\times200\times16$. For evaluation metrics, we report the mean Intersection over Union (mIoU) for each semantic class and geometric IoU (IoU_{geo}) for binary occupancy prediction, along with mean absolute velocity error (mAVE) across categories to quantify flow accuracy.

6.2. Network Architectures

As mentioned in the main paper, we validate the effectiveness of the DriveX model on two representative various tasks: end-to-end driving as well as occupancy and flow prediction. Here, we detail the network architecture and the design of the Future Spatial Attention (FSA) paradigm for both tasks.

DiffusionDrive-C. We reimplement the camera-only variant of DiffusionDrive [38], utilizing the BEVFormer [35] paradigm to construct latent 128×128 BEV features. Following the original DiffusionDrive configuration, we initialize planning queries using 20 cluster anchors for spatial cross-attention with BEV features, and employ a diffusion policy to generate 8-waypoint trajectory. Based on these predictions, DriveX generates latent BEV representations for three future timestamps. A three-layer FSA is then integrated to extract future information through spatial cross-attention. Specifically, for each planning query, we sample a total of 96 points from three future BEV representations based on its waypoints and predicted sample offsets to form the keys and values. The refined planning queries are subsequently used to predict the final trajectory.

ViewFormer. For occupancy and flow prediction tasks, we adopt the official implementation of ViewFormer [30]. Regarding trajectory prediction, we adopt the AD-MLP [37] architecture to predict two future waypoints. Given future states generated by DriveX, the FSA samples keys and values from predicted future BEV features with the number of attention heads, learnable sampling points, and attention layers set as 9, 4, and 4, respectively.

6.3. Training Schemes

While mainly following the original training protocols for downstream tasks, we specify several key training parameters and optimization settings here. In all downstream experiments, the DriveX model remains frozen.

DiffusionDrive-C is trained for 40 epochs with batch size 32 on 16 NVIDIA H20 GPUs. We use the AdamW optimizer [40] with a learning rate of 2e-4 and a weight decay of 0.001. The input images are resized to 256×704. Following DiffusionDrive [38], we focus on a spatial range of [-32.0m, -32.0m, -1.0m, 32.0m, 32.0m, 5.4m] and incorporate auxiliary supervision through object detection and map segmentation tasks.

ViewFormer is trained for 24 epochs with batch size 16 on 16 NVIDIA H20 GPUs. The model is optimized using AdamW with an initial learning rate of 2e-4 and a weight decay of 0.01, where the learning rate is decayed by a factor of 0.1 at epochs 19 and 23. All experiments are conducted using a default image size of 256×704.

DriveX. The main training settings are outlined in the main paper. We provide additional implementation details here. To obtain comprehensive geometric information of the environment, we aggregate LiDAR points from the past four frames and future five frames, utilizing an off-the-shelf tracker to specifically track and aggregate points from moving objects. Subsequently, we sample points using a voxel size of 0.4 to construct LiDAR-view rays. During the dynamic-aware sampling phase, rays are sampled based on moving bounding boxes. The sampling procedure prioritizes high-confidence trackers within a range of [-30m, 30m] along both the X and Y axes. To ensure learning efficiency, we implement the Farthest Point Sampling (FPS) strategy, limiting the total number of rays to 2000 per scene. Regarding the image-view ray sampling, we adopt a straightforward strategy where the sampling probability for each pixel is determined by the inverse of its semantic category frequency.

6.4. Image Semantic Label Generation

We primarily follow the approach of OccNeRF [68], utilizing Grounded SAM [46] for open-vocabulary semantic segmentation. However, we observe that this method struggles to segment large objects completely. To address this limitation and improve segmentation quality, we integrate OpenSeeD [69], a state-of-the-art segmentation framework, specifically focusing on improving the segmentation of "manmade" and "sky" categories. The segmentation results from both methods are merged to form a semantic label set comprising 15 categories (excluding 'other' and 'other flat' categories in [30]), aligned with the classification scheme defined in [68].

Method	Day	Night	Sunny	Rainy
ViewFormer	43.42	28.57	42.28	44.14
+ DriveX-S	**44.26**	**29.61**	**43.25**	**44.93**
Improvement	***+0.84***	***+1.04***	***+0.97***	***+0.79***

Table 9. Occupancy prediction performance (mIoU) under different lighting and weather conditions on nuScenes validation set.

Dataset fraction	Chamfer Distance (m^2) ↓				Time
	1.0s	2.0s	3.0s	Avg.	
10%	2.08	2.30	2.51	2.30	1.2h
50%	1.36	1.53	1.70	1.53	5.7h
100%	**0.80**	**0.99**	**1.28**	**1.02**	11.2h

Table 10. Ablation study on data scale. All experiments are conducted with the same training epochs.

7. Efficiency and Robustness Analysis

DriveX employs a shared world encoder across tasks with minimal computational overhead, adding only 48ms latency and 12.8M parameters for occupancy and flow prediction task over ViewFormer [30] (150ms, 103.8M). Similarly, for end-to-end driving task, it introduces merely 42ms latency and 11.9M parameters over DriffusionDrive-C [38] (90ms, 56.0M). Both experiments are conducted on an H20 GPU. Additionally, we evaluate DriveX's robustness under different weather and lighting conditions in Table 9. DriveX achieves consistent performance gains across varying environments, confirming its practical reliability.

8. Additional Experiments

Data Scaling. A significant promise of self-supervised world models lies in their ability to achieve consistent performance improvements with increasing data scale. We conduct this study on varying fractions of NuScenes training set. Table 10 demonstrates consistent performance gains across all metrics as training data increases. When scaling from 50% to 100% of the dataset, the model still achieves substantial gains, reducing Chamfer Distance by 0.51m^2.

9. Qualitative Results

We visualize some qualitative results for point cloud forecasting on nuScenes validation set in Figure 5. Given historical latent representations and current visual observations, our DriveX model can generate future point clouds at 0.5s intervals over a 3-second horizon. The upper part presents a scenario where the ego-vehicle moves straight and passes other objects. DriveX accurately models the spatial relationships in the dynamic scene, producing precise point cloud predictions. Notably, for two distant objects with sparse point representation in the ground truth (highlighted in orange), DriveX successfully predicts their locations as the ego vehicle approaches them. The lower part demonstrates a challenging scenario where the ego-vehicle navigates through an intersection. The spatial lay-

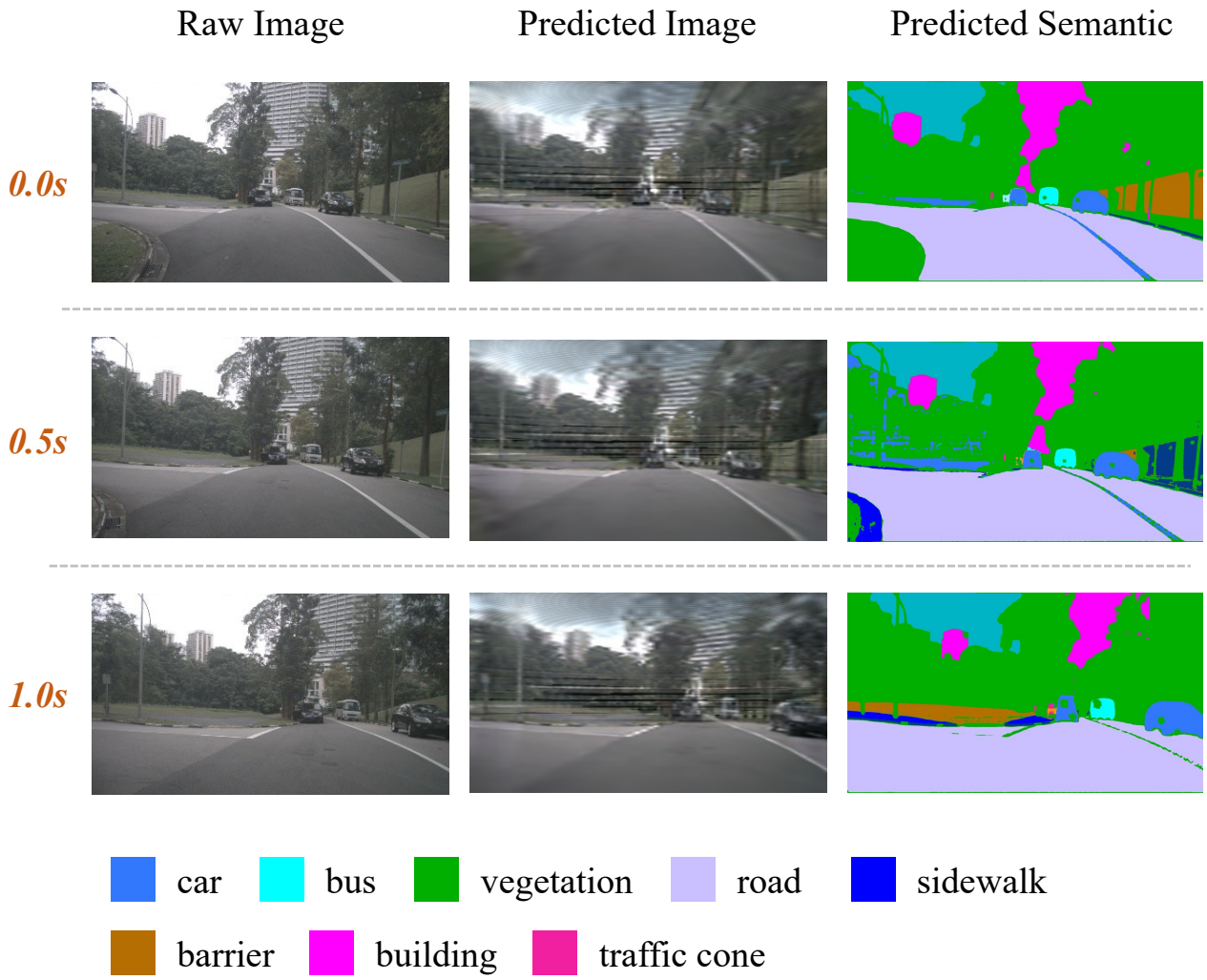

Figure 4. Qualitative results of image generation and semantic prediction on nuScenes validation set. For simplicity, we only visualize the front-view camera images. **Left:** Raw images. **Middle:** Images reconstructed from BEV space latent world features. **Right:** Semantic predicted from BEV space latent world features. Our model can reconstruct sharp boundaries of buildings and cars, while accurately capturing small objects such as traffic cones.

out undergoes complex changes during the crossing process, such as the curved roadside structure. The generated point clouds demonstrate DriveX's ability to understand the related position and orientation between the ego-vehicle and its surrounding environment. Furthermore, Figure 4 presents qualitative results for image generation and semantic segmentation over a 1-second future horizon. DriveX exhibits remarkable performance in scene reconstruction, accurately preserving lighting conditions and producing sharp semantic maps. While minor artifacts appear in the generated images and slight noise exists in the predicted semantics, we attribute these limitations to DriveX's lightweight architecture, as it avoids the computational overhead of intensive video diffusion models [11, 55, 56].

In Figure 6 and Figure 7, we present qualitative results for two downstream tasks: occupancy prediction with flow estimation and end-to-end driving. As demonstrated in Figure 6, ViewFormer accurately predicts both occupancy and motion flow for static environment elements and dynamic agents (pedestrians, vehicles, etc.). Figure 7 illustrates the end-to-end driving performance, where our DriveX exhibits robust behavior across diverse driving scenarios. The model successfully executes various driving maneuvers, including breaking, lane changes, intersection negotiation, and turns.

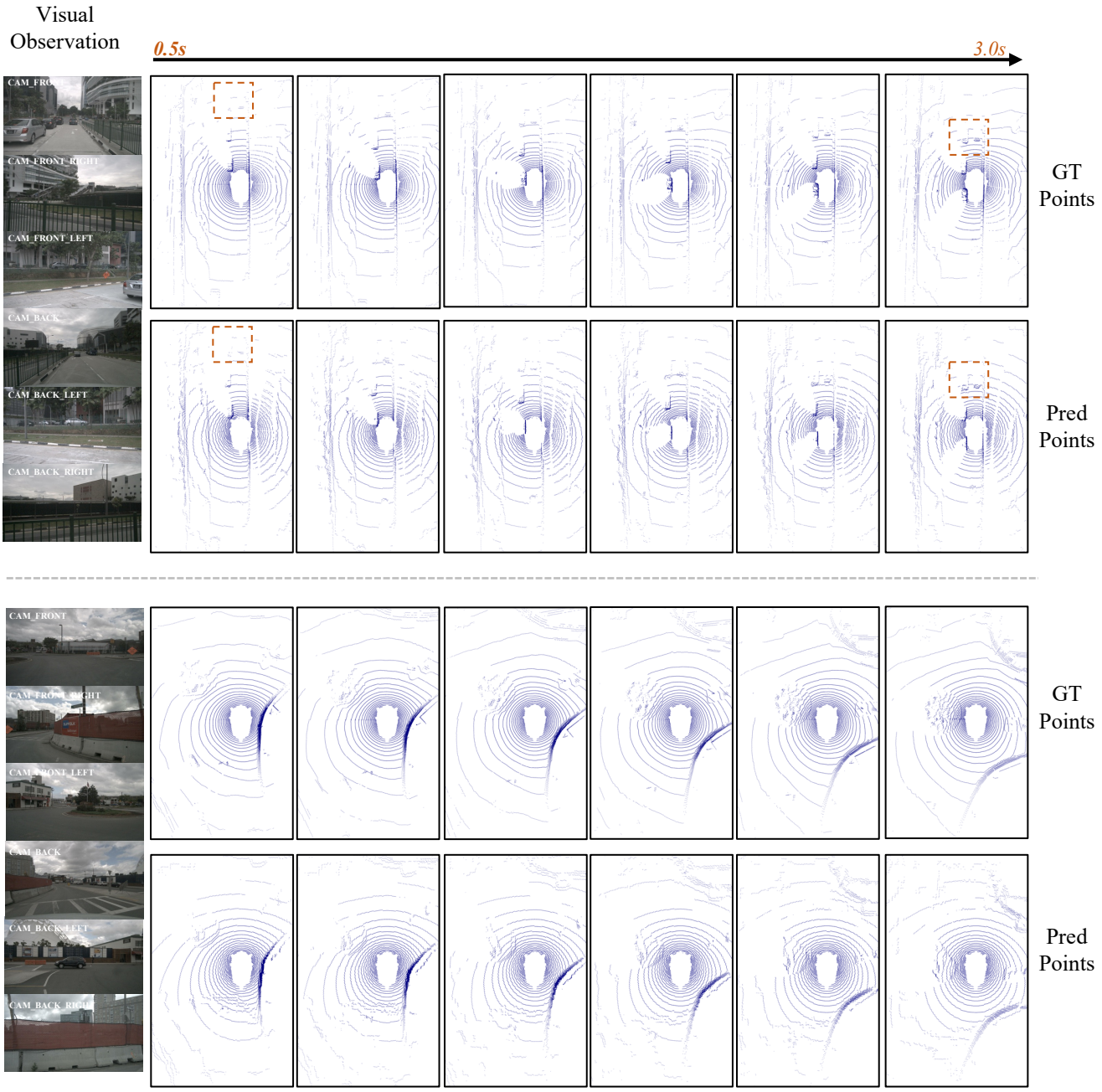

Figure 5. Qualitative results of point clouds forecasting on nuScenes validation set. **Left:** Current six-view camera images. **Right:** 3-second future point cloud predictions, where ground truth points (top) and predicted points (bottom) are visualized. The first row shows a straight driving scenario, while the second row presents a complex intersection crossing scenario. Orange rectangles highlight two challenging distant objects from the ego-vehicle.

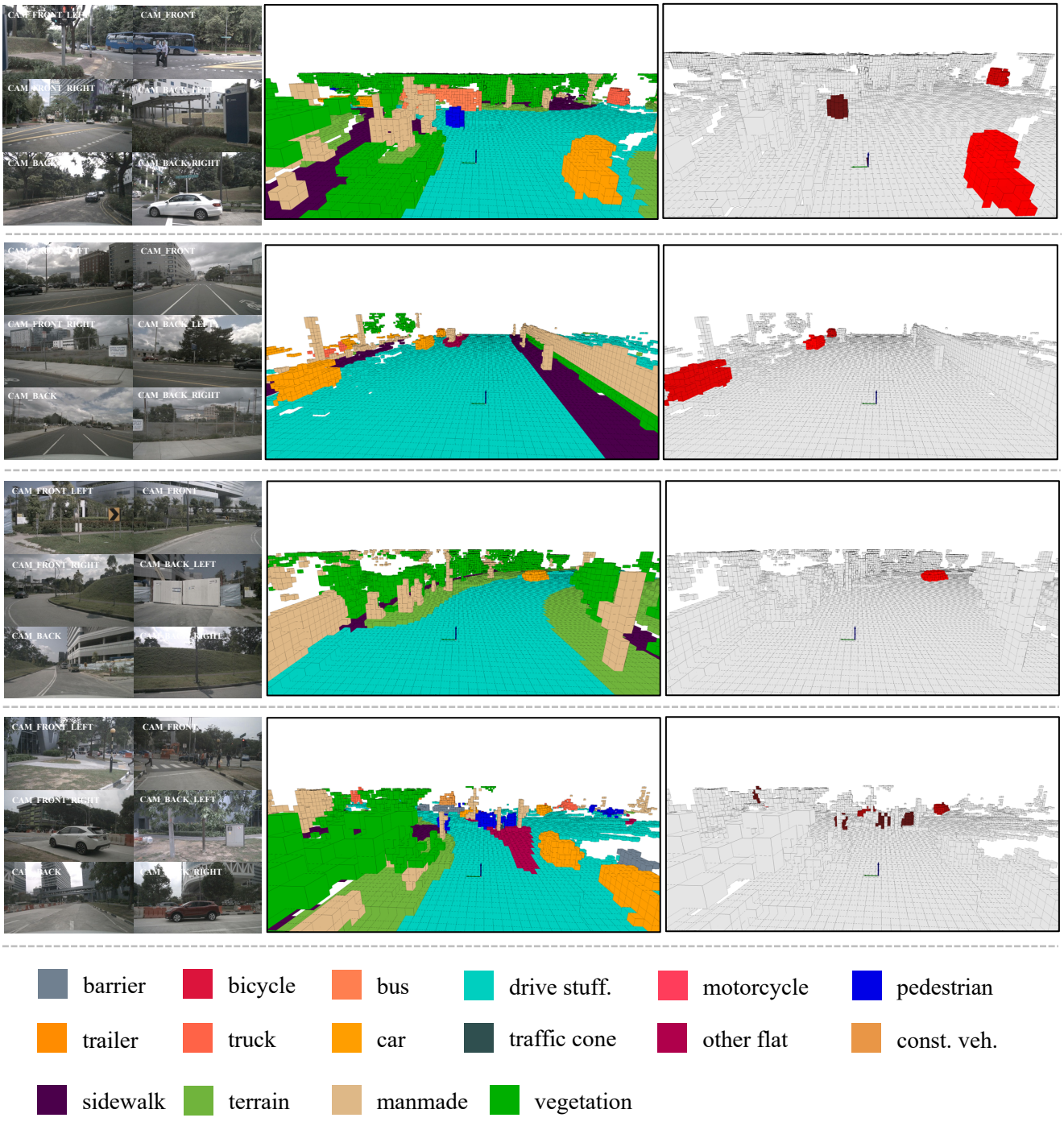

Figure 6. Qualitative results of occupancy and occupancy flow prediction on the NuScenes validation set. **Left:** Input visual images. **Middle:** Occupancy prediction results. **Right:** Occupancy flow prediction results, where dynamic occupancy is visualized using a color gradient from dark red to red based on flow magnitude.

Figure 7. Qualitative results of end-to-end driving on NAVSIM test set. DriveX successfully handles challenging scenarios including breaking, lane changing, intersection negotiation, and turning maneuvers.