SparseWorld-TC: Trajectory-Conditioned Sparse Occupancy World Model

Supplementary Material

A. Additional Quantitative Experiments

A.1. Ray-level mIoU

SparseOcc proposes RayIoU (Ray-level mIoU) to solve the inconsistency penalty along the depth axis raised in traditional voxel-level mIoU criteria. We evaluate our SparseWorld-TC-Large* model with this ray-level metric and report the results in Table 5.

Table 5. RayIoU scores [%] of 4D occupancy forecasting performance on the Occ3D-nuScenes [37] benchmark.

	RayIoU_{1m}	RayIoU_{2m}	RayIoU_{4m}	RayIoU
Recon.	35.4	43.0	47.8	42.1
1s	29.5	36.5	41.4	35.8
2s	25.8	32.1	36.8	31.6
3s	23.5	29.4	33.8	28.9

A.2. Ablation Study on Trajectory Embedding

We further explore the influence of embedding modules for the trajectory condition. The modules named "TE", "PE" and "STE" represent time embedding, position embedding and spatiotemporal embedding, respectively. The performance of our SparseWorld-TC-Small model in different settings is summarized in Table 6.

Table 6. Ablation study on embedding modules. Avg. denotes average performance of mIoU or IoU in 1s, 2s, and 3s.

Modules			mIoU (%) ↑	IoU (%) ↑
TE	PE	STE	Avg.	Avg.
			15.44	32.19
✓			17.45	35.06
✓	✓		23.07	47.53
✓	✓	✓	**25.60**	**49.02**

A.3. Ablations of More Model Settings

In Table 7, "w/o Deformable Attn" replaces deformable attention with a simpler multi-view feature aggregation, while "w/o Temporal Attn" replaces temporal attention with frame-wise processing. We further ablate the effect of increasing number of anchors and points per anchor in Table 8. Roughly, more anchors and points per anchor both improve the performance.

Table 7. Ablation on Removing Temporal and Deformable Attention. Avg. denotes average of mIoU or IoU in 1s, 2s, and 3s.

Modules	Avg.mIoU (%) ↑	Avg.IoU (%) ↑
Ours-Small	**25.60**	**49.02**
w/o Temporal Attn.	25.07	47.53
w/o Deformable Attn.	23.89	46.55

A.4. Hardware Cost versus Model Size and Performance

As shown in Table 8, the + denotes increasing the number of anchors (N) or points per anchor (M). We compare on backbones with both ResNet and DINOv3(*).

Table 8. GPU Memory Usage and Performance of Evaluation.

Model	N	M	Memory (GB)	mIoU (%)
Ours-Small	600	128	5.4	25.60
Ours-Large	4800	16	7.7	26.42
Ours-Large+	4800	64	9.2	27.25
Ours-Small*	600	128	7.1	27.71
Ours-Large*	4800	16	14.4	29.89
Ours-Large*+	9600	16	24.6	30.88

A.5. Per-class Performance

As shown in Table 10 and Table 11, we report the performance per-class of our SparseWorld-TC-Large* (Ours-Large* for short). Our approach not only maintains the geometric consistency of static scenes, but also predicts the dynamic objects relatively accurately.

A.6. Improvement on Small Targets

The performance on smaller objects, including motorcycles and pedestrians, improves with a larger number of anchor queries. Emphasizing small objects through loss reweighting also enhances the model capability, as evidenced in Table 9.

Table 9. Performance Improvement for Small Objects

Model	Avg. mIoU (%) ↑	
	motorcycle	pedestrian
Ours-Large*	13.64	6.86
Q=4800 → Q=9600	14.72 (+1.08)	8.29 (+1.43)
Class Reweighting	14.64 (+1.00)	8.63 (+1.83)

B. Additional Qualitative Experiments

B.1. Feedforward Gaussian

Inspired by advances in feedforward Gaussian methods [31, 32, 36], we extend the original model with additional MLPs to decode Gaussian parameters from latent features. Then we utilize the differential Gaussian rasterization proposed in 3DGS [13] to render the predicted front-view image and calculate L1 loss with the GT image.

Currently our feedforward 3DGS implementation only supports the front-view camera, which supports both 256x704 and 128x352 resolutions. To alleviate the voids caused by query sparsity, we boost the number of anchors and points per anchor to 7200 and 32.

Figure 7. Gaussian splatting reconstruction during training.

The reconstruction and future forecasting of sensor observation, as shown Figure 7 and Figure 8, demonstrate the potential of our model in leveraging the Gaussian representation and achieving self-supervised training in the future.

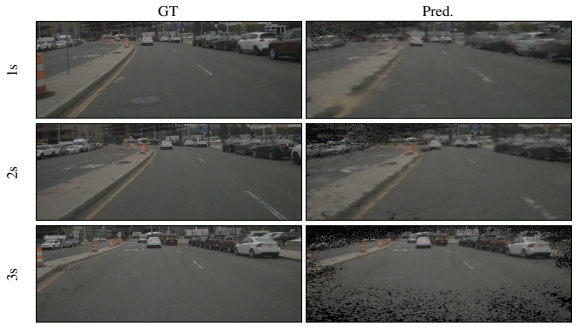

Figure 8. Future observation forecasting on validation set.

B.2. Convergence Visualization

Figure 9 shows how the model converge from noise to the detailed occupancy.

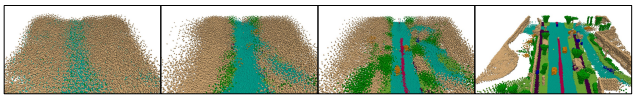

Figure 9. Model Convergence Visualization.

B.3. Failure Case Analysis

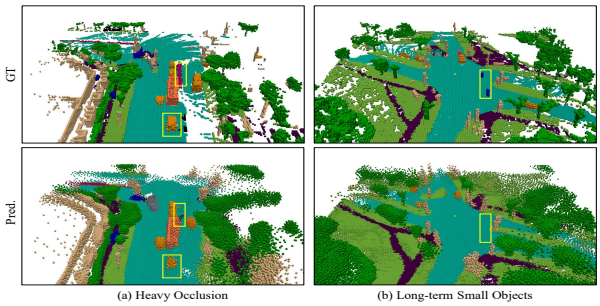

Figure 10. Failure cases: (left) the car and barrier with heavy occlusion, (right) small objects appear in the future 6 seconds.

As shown in Fig. 10, we visualize typical failure cases.

B.4. Additional Trajectory-conditioned Prediction

We additionally visualize the occupancy prediction results under different trajectory conditions, as shown in Fig. 11. More visualizations are shown in Fig. 12 and Fig. 13.

Table 10. Per-class mIoU [%] performance of 4D occupancy forecasting on Occ3D-nuScenes [37].

Ours-Large*	mIoU	IoU	others	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	manmade	vegetation
Recon.	37.92	57.65	13.15	45.06	26.72	40.26	46.44	24.62	25.38	21.47	28.99	34.84	35.92	76.87	44.76	51.14	49.71	40.98	38.38
1s	32.76	55.28	12.25	42.04	19.22	29.07	32.03	22.42	15.70	10.28	24.24	30.55	27.37	74.98	43.12	49.53	48.43	39.03	36.68
2s	29.62	53.56	11.40	38.75	15.61	20.72	25.49	19.93	13.46	6.01	19.55	26.27	22.35	74.46	42.20	48.34	47.42	36.82	34.76
3s	27.28	51.71	10.36	35.19	12.38	16.90	22.22	17.73	11.75	4.29	15.14	23.78	20.04	73.56	40.62	46.71	45.89	34.49	32.66

Table 11. Per-class RayIoU [%] performance of 4D occupancy forecasting on Occ3D-nuScenes [37].

Ours-Large*	RayIoU	others	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	manmade	vegetation
Recon.	42.1	11.7	47.3	31.7	63.6	56.3	28.7	30.4	35.2	32.3	37.1	52.4	70.4	40.7	38.1	40.2	53.1	46.2
1s	35.8	10.5	44.8	19.8	47.2	43.0	26.7	17.6	21.6	29.1	31.0	44.5	66.0	37.9	35.5	38.3	50.8	44.2
2s	31.6	9.7	41.9	15.9	34.4	34.1	25.1	15.1	13.7	25.8	21.9	39.7	62.8	36.1	33.6	36.7	48.3	41.8
3s	28.9	9.0	39.3	12.9	28.5	29.9	23.9	13.5	10.1	22.5	19.1	36.7	60.2	34.1	31.8	34.8	45.8	39.3

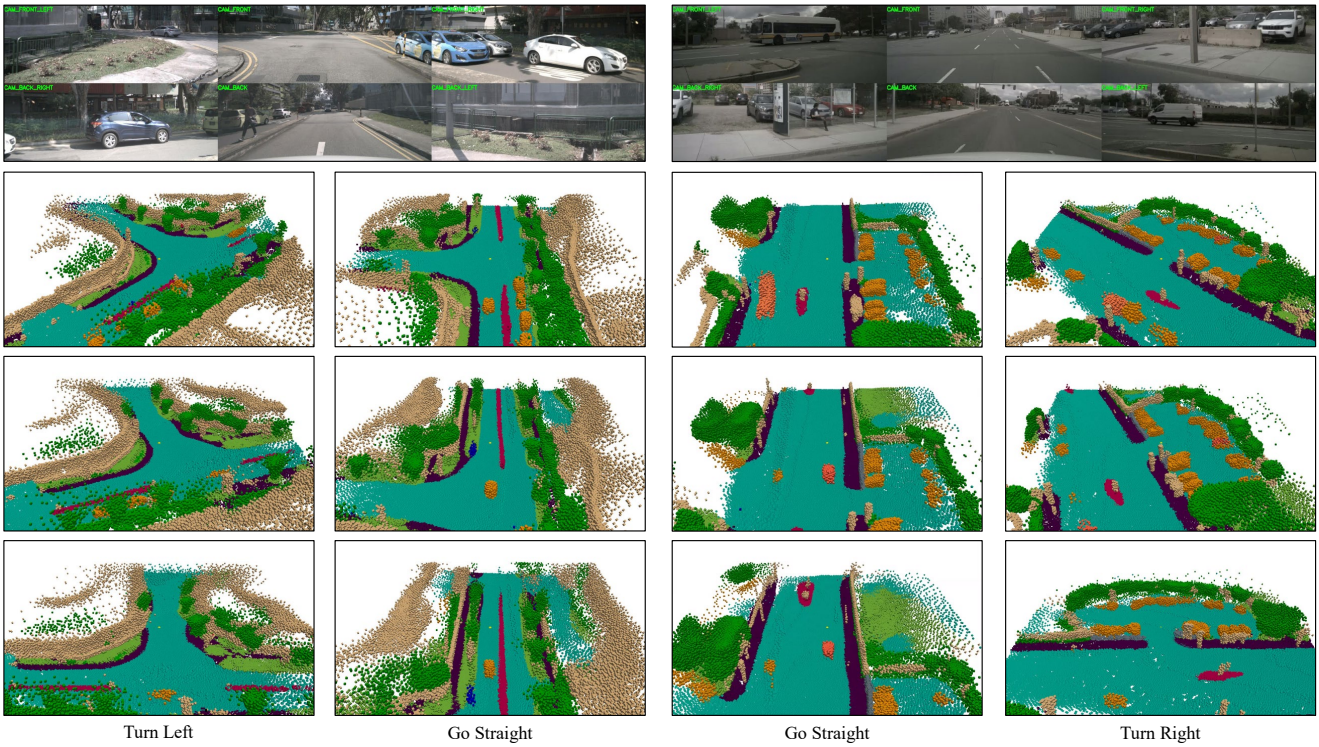

Figure 11. Additional qualitative results in the different trajectory conditions.

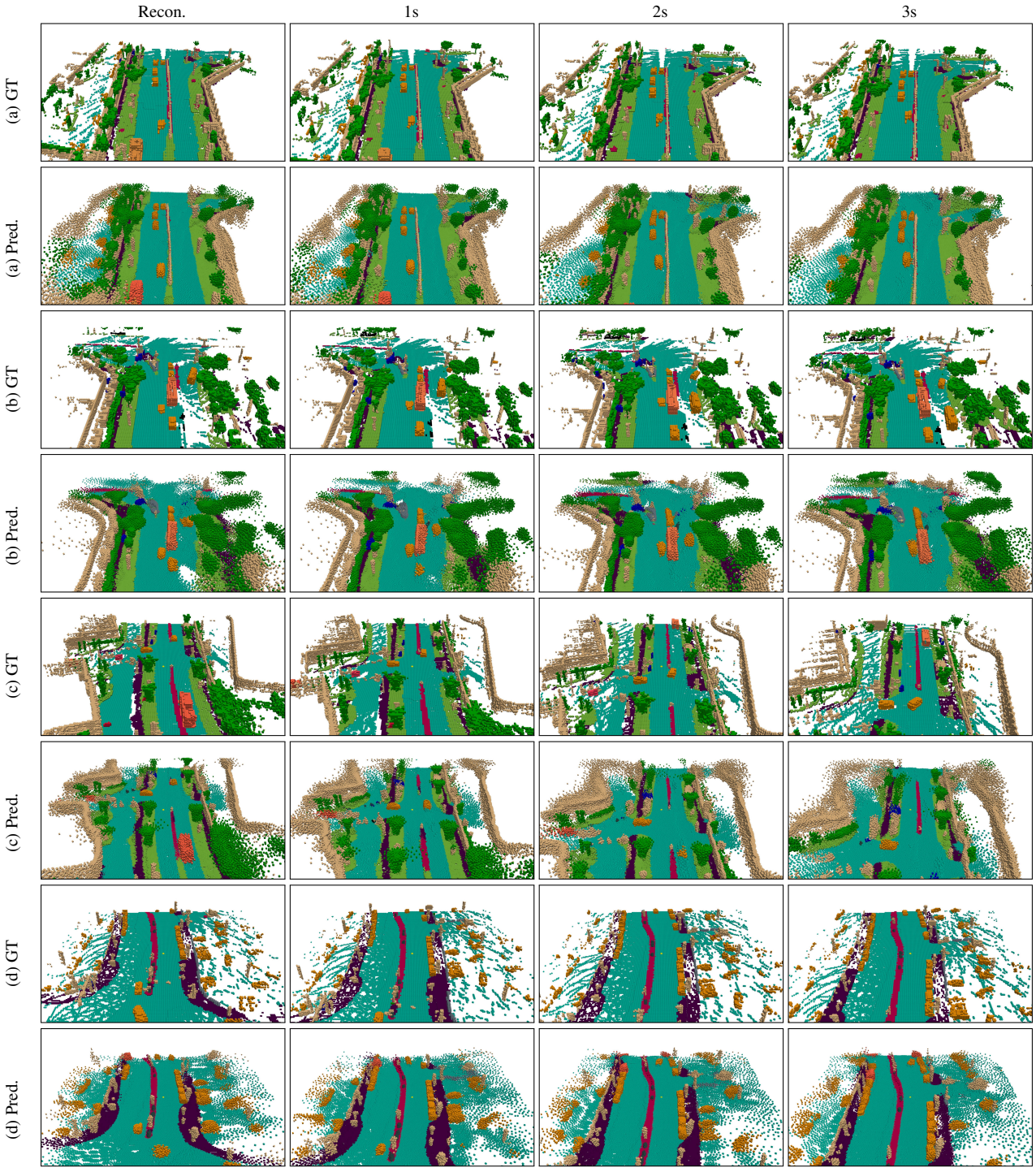

Figure 12. Additional qualitative results of our proposed SparseWorld-TC are presented here.

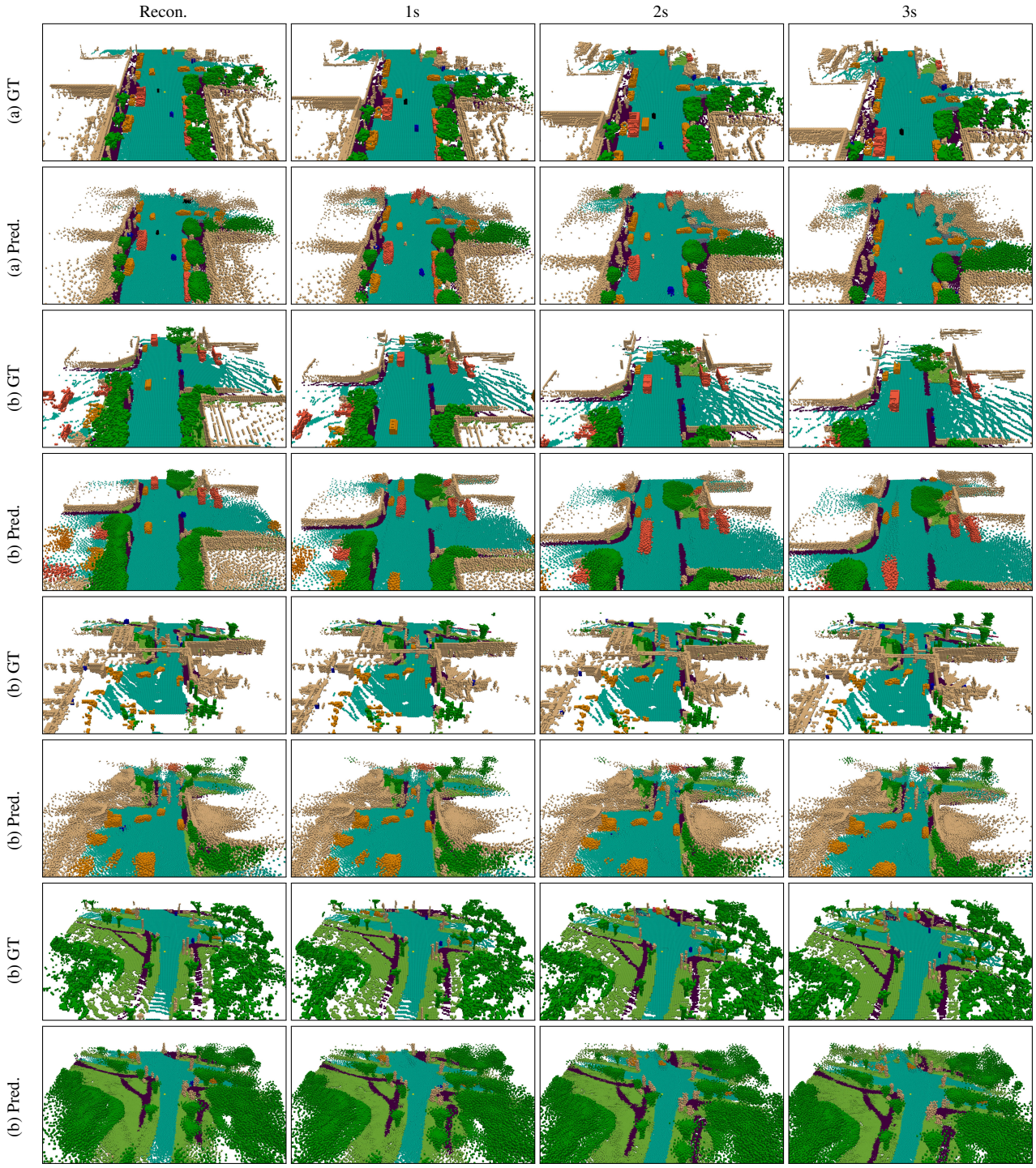

Figure 13. Additional qualitative results of our proposed SparseWorld-TC are presented here.