Appendix

A. Prediction Model Comparisons

Table 1 shows the parameter numbers of different prediction models and their inference times. We see that i) Although MotionNet has the fastest speed, its performance is significantly surpassed by other methods. FIERY and BEVerse both use RNN-based prediction models. Our method is close to FIERY's, but approximately two-time as fast as BEVerse's. ii) Compared to other methods, though our prediction model has the fewest parameters, we achieve better performance. In summary, the proposed STPT module is parameter-efficient with prominent performance and leading speed.

Methods	IoU	VPQ	# param.	Inference Time
MotionNet†	35.4	30.6	11.26M	22ms
FIERY†	38.3	32.1	12.37M	134ms
BEVerse†	40.2	34.0	13.17M	322ms
TBP-Former	41.9	36.9	9.42M	165ms

Table 1. The number of parameters and inference time of prediction models. †: We use MotionNet, FIERY, and BEVerse's prediction models to replace our proposed STPT.

B. Metrics

In ablation experiments for the prediction model, we use Video Recognition Quality (VRQ) and Video Segmentation Quality (VSQ) for additional prediction metrics following FIERY. The formulas of VRQ and VSQ are shown below.

$$\mathrm{VRQ} = \sum_{t=0}^{H} \frac{|TP_t|}{|TP_t| + \frac{1}{2}|FP_t| + \frac{1}{2}|FN_t|}$$

$$\mathrm{VSQ} = \sum_{t=0}^{H} \frac{\sum_{(p_t,q_t)\in TP_t} \mathrm{IoU}(p_t,q_t)}{|TP_t|}$$

where H is the sequence length, TP_t represents the set of true positives, FP_t represents the set of false positives and FN_t represents the set of false negtives at timestamp t.

C. Additional Visualization

In Fig. 1, we show more qualitative comparisons with other vision-centric PnP methods, including FIERY and BEVerse. From left to right, we show the prediction results of FIERY, BEVerse and TBP-Former, and the ground truth. We see that TBP-Former is superior to other methods in the integrity of segmentation and the accuracy of prediction.

D. Results on BEV Detection task

Though we mainly focus on the joint perception and prediction (PnP) task in the paper, our proposed framework is also capable of other BEV perception tasks. We implement CenterPoint's detection head after the generated BEV representations to execute BEV detection task. We calculate vehicles' Average Precision under BEV (APBEV) at various thresholds to compare TBP-Former and previous methods. Table 2 demonstrates that our approach achieves comparable performance.

Methods	Input RGB Resolution	AP@0.3	AP@0.5	AP@0.7
BEVFormer-small	1280 × 720	55.42	35.76	13.13
BEVFormer-base	1600 × 900	56.94	40.40	16.38
BEVerse	1408 × 512	54.30	37.86	16.12
Ours	480 × 224	55.43	38.49	17.98

Table 2. BEV Detection results on nuScenes validation dataset.

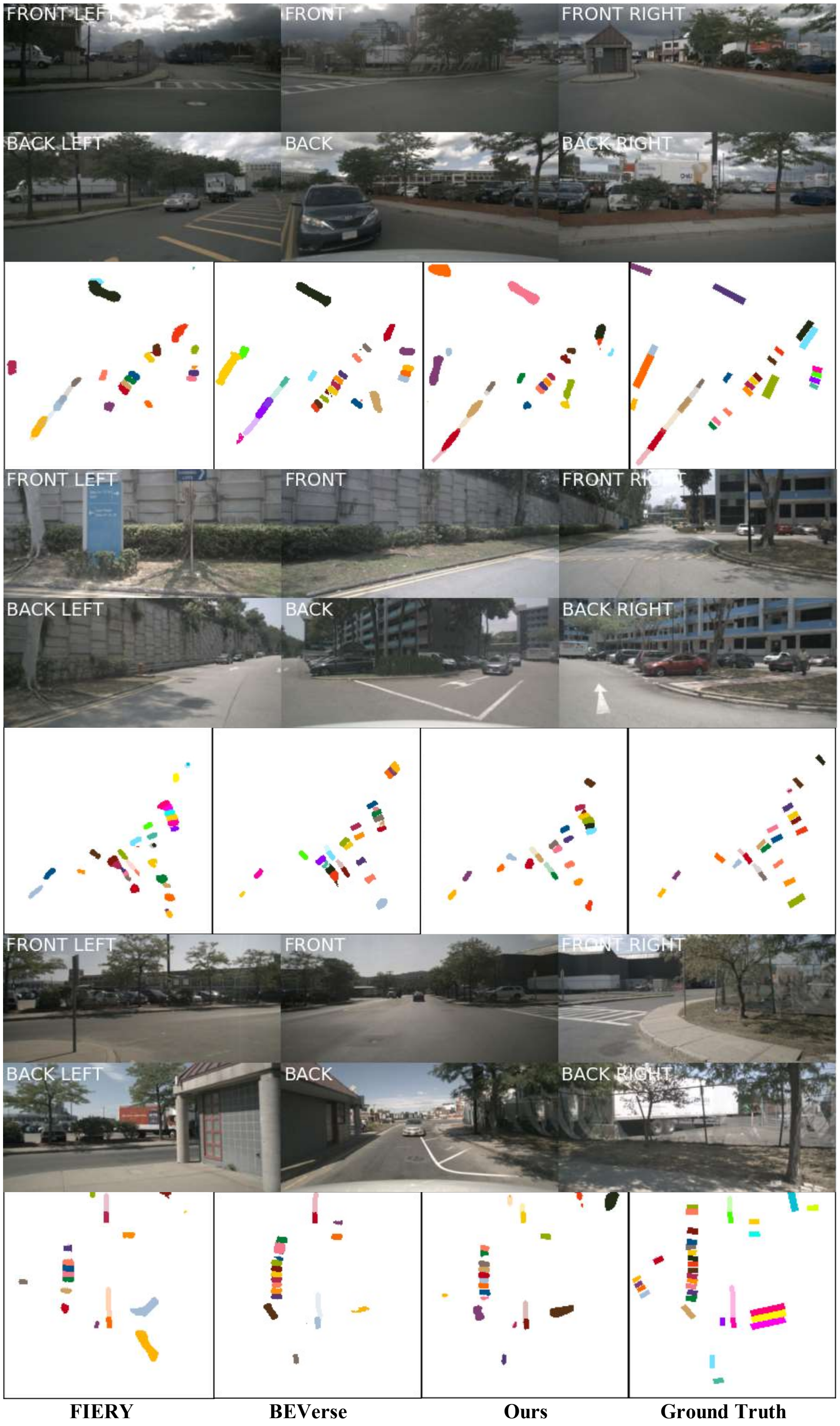

Figure 1. Comparisons between our method and previous methods.