A Appendix

A.1 Comparison of Depth Prediction

In this section, we report the related depth estimation metric (L1 Error and AbsReal) for pixel-wise and object-wise depth of foreground objects. Besides, we also report these metrics for distant objects (> 40m). As shown in Table 1, we can find that the accuracy of object-wise depth prediction is better than pixel-wise. For distant objects, the difference is even more significant. These experiments indicate the object-wise depth is more easily estimated, specifically for some distant objects.

Table 1: Comparison of pixel-wise depth prediction and object-wise depth prediction on nuScenes `val` set.

Depth Prediction	L1 Error (m) ↓	AbsRel ↓	L1 Error$_{>40}$ (m) ↓	AbsRel$_{>40}$ ↓
Pixel-wise	4.44	0.233	12.28	0.515
Object-wise	**2.20**	**0.090**	**3.88**	**0.106**

A.2 Comparison of Efficiency

In this section, we compare the running speed, computation cost, and parameter size with StreamPETR [5] and 3DPPE [4]. We adopt the same experiment setting as the ablation studies of the main paper. For a fair comparison, the running speed of all methods is evaluated on an NVIDIA GeForce RTX 3090 with a batch size of 1. As shown in Table 2, our method brings satisfactory performance improvements under the comparable running speed, computation cost, and parameter size compared with other methods.

Table 2: Comparison of running speed, computation cost, and parameter size with other methods. For a fair comparison, the running speed of all methods is evaluated on an NVIDIA GeForce RTX 3090 with a batch size of 1.

Method	NDS↑	mAP↑	Speed (FPS)	FLOPs (G)	Params (M)
StreamPETR [5]	59.4	50.3	3.5	524.6	73.2
3DPPE [4]	60.0	51.6	3.5	558.8	79.2
OPEN	61.3	52.1	3.5	560.8	79.6

A.3 More Experiments

In this section, we conduct experiments with the ViT-L [1] backbone to demonstrate the effectiveness of OPEN and conduct experiments with the single frame as input to demonstrate the generalization of OPEN on nuScenes `val` set.

Performance of OPEN with ViT backbone. To further verify the effectiveness of OPEN. We validate our method with the ViT-L backbone with an image size of 320×800 on nuScenes `val` set. For all methods, we train 24 and 48 epochs. As shown in Table 3, when training 24 epochs, OPEN achieves 61.6% NDS and 52.3% mAP, outperforming StreamPETR [5] by 1.5% NDS and 0.8% mAP. When training 48 epochs, OPEN achieves 62.5% NDS and 54.2% mAP, outperforming StreamPETR by 1.2% NDS and 1.3% mAP. These experiments demonstrate the effectiveness of OPEN with a larger backbone.

Table 3: Comparison of other method with ViT-L backbone on nuScenes `val` set.

Method	Backbone	Input Size	Epoch	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
StreamPETR [5]	ViT-L	320 × 800	24	60.1	51.5	0.576	**0.254**	0.300	0.233	0.204
OPEN	ViT-L	320 × 800	24	**61.6**	**52.3**	**0.534**	0.259	**0.269**	**0.200**	**0.197**
StreamPETR [5]	ViT-L	320 × 800	48	61.3	52.9	0.563	**0.253**	**0.266**	0.230	0.198
OPEN	ViT-L	320 × 800	48	**62.5**	**54.2**	**0.511**	0.258	0.295	**0.205**	**0.197**

Table 4: Comparison of other method with single frame on nuScenes `val` set.

Method	Backbone	Input Size	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
PETR [3]	V2-99	900 × 1600	45.5	40.6	0.736	0.271	0.432	0.825	0.204
OPEN	V2-99	640 × 1600	**50.3**	**44.0**	**0.648**	**0.265**	**0.394**	**0.668**	**0.193**

Performance of OPEN with Single Frame. We validate our method with single frames as input on nuScenes `val` set. We adopt the V2-99 [6] backbone pre-trained on FCOS [6] with an image size of 640×1600 and 24 epochs. for comparison with PETR [3]. As shown in Table 4, even though we use a smaller input size, OPEN still achieves 50.3% NDS and 43.0% mAP, outperforming PETR by 4.8% NDS and 3.4% mAP. It is worth noting that we disable the temporal information of the object-wise depth encoder and keep the same training settings as PETR (*e.g.*, without denoising training [2], the weight of center loss is set to 1). These experiments demonstrate the generalization of OPEN.

A.4 Visualization

Comparison of Qualitative Results. To illustrate the superiority of our OPEN, we compare the qualitative detection results of StreamPETR and OPEN

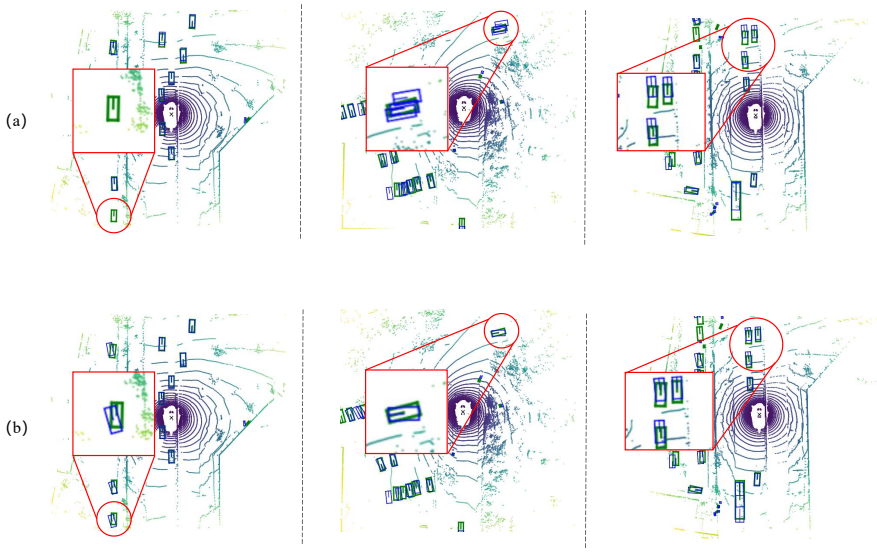

Fig. 1: Comparison of StreamPETR (a) and OPEN (b) on the nuScenes `val` set. Blue and green boxes are the prediction and ground truth boxes. In the first column, OPEN can successfully detect the hard-detect object. In the second column, OPEN can distinguish the false positive better. In the third column, OPEN can achieve more accurate location ability for distant objects.

on nuScenes `val` set. As shown in Figure 1, in the first column, OPEN can successfully detect the hard-detect object. Then, in the second column, OPEN can distinguish the false positive better. Finally, in the third column, OPEN can achieve more accurate location ability for distant objects. These results illustrate the superiority of OPEN.

Qualitative Results. We show the qualitative detection results of OPEN in Figure 2 on multi-view images and BEV space. The 3D predicted bounding boxes are drawn with different colors for different classes. The first row and second row are the prediction of OPEN and the ground truth, respectively.

A.5 Limitations

Our method utilizes global attention in the transformer decoder, which has larger computational costs when facing larger-scale scenarios. In the future, we plan to design effective local attention for the transformer decoder to improve the efficiency of OPEN.

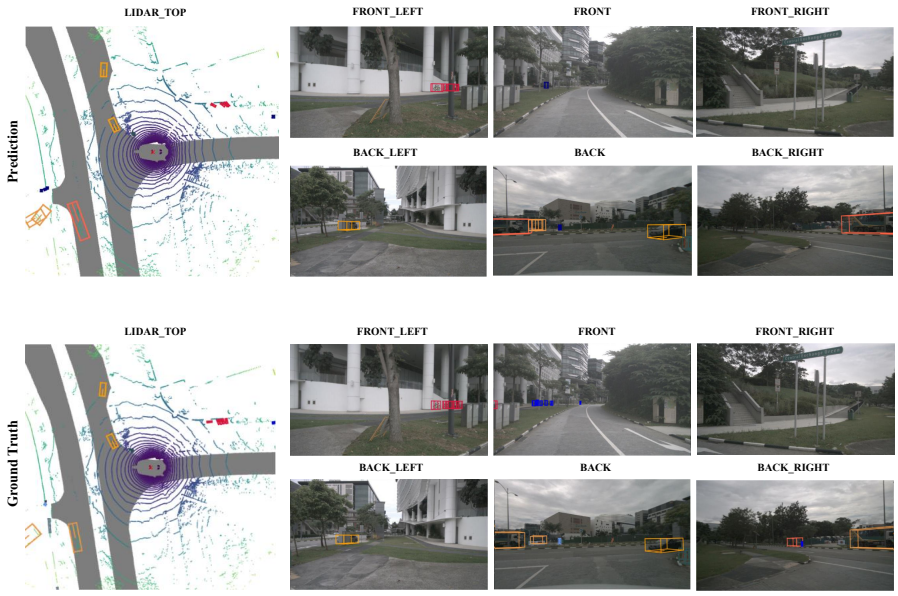

Fig. 2: Qualitative detection results on multi-view images and BEV space on the nuScenes `val` set. The 3D predicted bounding boxes are drawn with different colors for different classes.

A.6 Potential Negative Impact

OPEN improves 3D detection performance by introducing object-wise depth information. However, depth prediction usually requires more computation costs, leading to higher requirements for the hardware of autonomous driving.

References

1. Fang, Y., Sun, Q., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva-02: A visual representation for neon genesis. arXiv preprint arXiv:2303.11331 (2023)
2. Li, F., Zhang, H., Liu, S., Guo, J., Ni, L.M., Zhang, L.: Dn-detr: Accelerate detr training by introducing query denoising. In: CVPR (2022)
3. Liu, Y., Wang, T., Zhang, X., Sun, J.: Petr: Position embedding transformation for multi-view 3d object detection. In: ECCV. pp. 531–548. Springer (2022)
4. Shu, C., Deng, J., Yu, F., Liu, Y.: 3dppe: 3d point positional encoding for transformer-based multi-camera 3d object detection. In: ICCV. pp. 3580–3589 (2023)
5. Wang, S., Liu, Y., Wang, T., Li, Y., Zhang, X.: Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In: ICCV (2023)
6. Wang, T., Zhu, X., Pang, J., Lin, D.: Fcos3d: Fully convolutional one-stage monocular 3d object detection. In: ICCV. pp. 913–922 (2021)