# Supplementary Materials: Leveraging Vision-Centric Multi-Modal Expertise for 3D Object Detection

**Anonymous Author(s)**
Affiliation
Address
email

In this supplementary document, we present a comprehensive account of the implementation and training details in Section A. We delve into the analysis of misalignment resulting from temporal fusion and discuss the effectiveness of our proposed methods for addressing this issue in Section B. Moreover, we provide additional visualizations in Section C. Lastly, we explore the potential social impact of our research in Section D.

## A Experiment Details

### A.1 Dataset and Evaluation Metrics

We conduct our experiments on the nuScenes dataset [1], a widely used benchmark for autonomous driving tasks. The dataset encompasses diverse driving scenarios captured using cameras and LiDAR sensors, offering rich information for both visual and LiDAR-based 3D object detection. The dataset comprises 700 training scenes, 150 validation scenes, and 150 testing scenes. Each scene spans approximately 20 seconds, with key frames annotated at a 2 Hz frequency.

The two dominant metrics for the nuScenes detection task are the nuScenes Detection Score (NDS) and mean Average Precision (mAP). The mAP for nuScenes is computed based on the center distance between predictions and ground truth annotations on the ground plane. Moreover, the nuScenes dataset defines five true positive metrics (mATE, mASE, mAOE, mAVE, mAAE) for measuring translation, scale, orientation, velocity, and attribute, respectively. The NDS for nuScenes is a weighted sum of mAP and the five true positive metrics, defined as $NDS = \frac{1}{10}[5mAP + \sum_{mTP}(1 - \min(1, mTP))]$.

### A.2 Implementation Details

We conduct experiments on BEVDepth [4]. The codebase is developed upon MMDetection3D [2]. Main experiments are trained on 8 NVIDIA A100 GPUs, while ablation experiments are conducted on 8 NVIDIA V100 GPUS. For BEVDepth, the model is trained for 20 epochs with an initial learning rate of 2e-4. In the distillation process, the per-GPU batch size is set to 4, whereas during the training of the baseline model, it is set to 8. Normal data augmentations are introduced in the training process such as flip and rotate. In our apprentice models, future frames are not incorporated into the long-term temporal fusion throughout the training phase to ensure a fair comparison.

In our research, we implement distinct temporal modeling strategies for both apprentice and expert models. For the apprentice models, we incorporate a sequence of eight frames into the temporal modeling process. In contrast, the expert models integrate four future frames into the temporal modeling as demonstrated in our primary results. However, in our ablation study, we deviate from this approach and instead employ eight historical frames for temporal modeling.

Table 1: Experiment settings. * denotes that the training schedule for VCD-E is approximately one-fourth of the original schedule. This reduction was implemented to expedite the training process during the ablation study. The first group is engaged in training on the main results, whereas the second group is utilized in the ablation study.

| Method | Backbone | Image Size | mAP (%) | NDS (%) |
|---|---|---|---|---|
| VCD-E | ConvNext-B [5] | 512 x 1408 | 67.7 | 71.1 |
| VCD-A | Res-50 [3] | 256 x 704 | 41.8 | 54.2 |
| VCD-E* | ConvNext-B [5] | 256 x 704 | 54.2 | 58.8 |
| VCD-A | Res-50 [3] | 256 x 704 | 29.7 | 40.9 |

### A.3 Experiments Settings

The setting of adopted expert-apprentice pairs is depicted in Tab. 1. We categorize the distillation setting into two distinct groups. The primary group is engaged in training on the main results, whereas the second group is utilized for the ablation study.

## B The Analysis of Temporal Fusion

### B.1 The Misalignment of Motion Objects

As highlighted in preceding studies [6], long-term temporal fusion may face misalignment issues in motion estimation, which can be discerned through a reduction in performance on metrics like mATE. Let's consider a moving object and analyze the impact of inaccurate motion estimation on its position in the fused frame. We will assume that the environment is static, except for the moving object. Let the position of the moving object in the world coordinate system be represented by $\boldsymbol{P}_i^w = (x_i^w, y_i^w, z_i^w, 1)^T$ in each of the $N$ frames captured at times $t_1, t_2, \ldots, t_N$. The actual motion of the moving object between frames is represented by $\boldsymbol{M}_i^{obj}$, and the estimated motion is represented by $\hat{\boldsymbol{M}}_i^{obj}$. The difference between the estimated and actual motion of the object can be denoted as:

$$\Delta \boldsymbol{M_i}^{obj} = \boldsymbol{M_i}^{obj} - \hat{\boldsymbol{M}}_i^{obj}. \tag{1}$$

As we have already computed the transformation matrix $\boldsymbol{T}_i$ based on the estimated ego motion, we can calculate the transformed object position in the current frame, considering its actual motion, as:

$$\boldsymbol{P}_i^{w'} = \boldsymbol{T}_i \boldsymbol{M}_i^{obj} \boldsymbol{P}_i^w. \tag{2}$$

The error in the transformed object position can be computed as:

$$\boldsymbol{e}_i^{obj} = \boldsymbol{P}_i^{w'} - \hat{\boldsymbol{P}}_i^{w'}. \tag{3}$$

In the long-term fusion process, we integrate the information from all $N$ frames. Assuming we use a fusion function $F$, the fused position in the current frame can be represented as:

$$\boldsymbol{P}_{fusion}^{obj} = F(\boldsymbol{P}_1^{w'}, \boldsymbol{P}_2^{w'} \ldots, \boldsymbol{P}_N^{w'}). \tag{4}$$

The inaccuracies in the motion estimation of the moving object for each frame can propagate through the fusion function and result in a misaligned object in the fused frame. The overall error in the fused position can be represented as a function of the errors in each frame:

$$\boldsymbol{e}_{fusion}^{obj} = G(\boldsymbol{e}_1^{obj}, \boldsymbol{e}_2^{obj} \ldots, \boldsymbol{e}_N^{obj}), \tag{5}$$

where $G$ represents a function that combines the errors from each frame. The fused position of the moving object will be less accurate due to these motion estimation errors, leading to a decline in object detection performance in the long-term setting. To address the issue mentioned earlier, we introduce the trajectory-based distillation module, which compensates for the misalignment of moving objects. We will provide further details in the subsequent discussion.

Table 2: The performance gains of different trajectory length for trajectory-based distillation. As the trajectory length increases, the benefits derived from the distillation process become more pronounced.

| Trajectory Length | Distill | mAP (%) | NDS (%) |
|---|---|---|---|
| - | ✗ | 29.7 | 40.9 |
| 0 | ✓ | 31.8 | 42.1 |
| 1 | ✓ | 33.1 | 44.5 |
| 3 | ✓ | 34.6 | 45.6 |
| 5 | ✓ | **35.4** | **45.9** |
| 9 | ✓ | 33.9 | 44.7 |

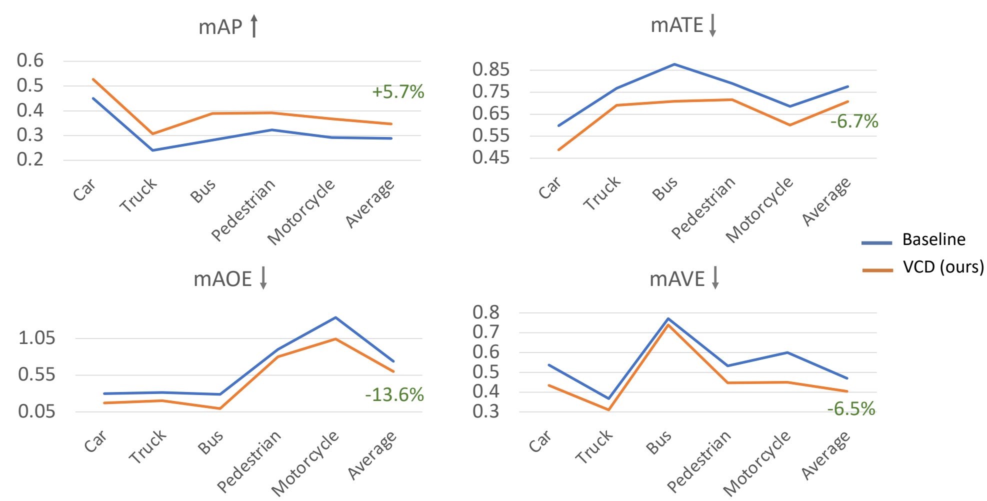


Figure 1: Effects of VCD on movable objects. Our distillation framework VCD consistently improves dynamic objects across a range of metrics.

### B.2 The Effectiveness of Trajectory-based Distillation

The results presented in Table 2 indicate that as the trajectory length increases, the benefits derived from the distillation process become more pronounced. The temporal fusion length for this experiment is set at eight. However, when the trajectory length exceeds five, there is a noticeable decrease in accuracy. We hypothesize that this decrease may be attributed to the model's distracted attention towards distant motions. The density of traffic can lead to distant motion locations being occupied by other objects, which may not necessarily require additional trajectory supervision. This suggests that the application of excessive trajectory supervision in such scenarios could be unnecessary and inefficient.

### B.3 The Improvements of Dynamic Objects

In this section, we present visualizations to demonstrate the improvements achieved in dynamic objects. Particularly noteworthy is the significant enhancement in the representation of dynamic objects through trajectory-based distillation, thereby highlighting the effectiveness of the trajectory-based module. As depicted in Fig. 1, our distillation framework consistently enhances dynamic objects across various metrics.

## C Visualization

We have performed several visualizations in Fig. 2 to showcase the advancements achieved by our distillation framework. Our findings indicate that our models excel in accurately predicting 3D bounding boxes for the target objects.

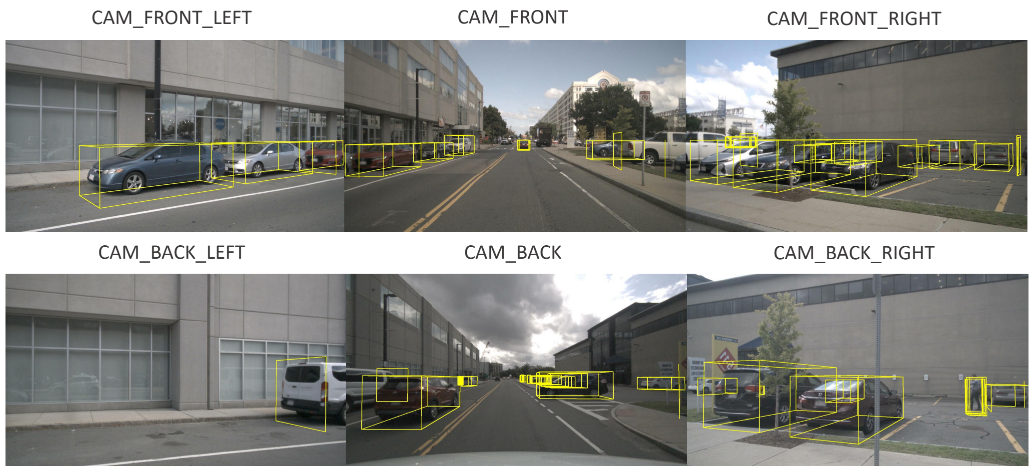


Figure 2: Visualization of the predictions for 3D object detection generated by the VCD-A.

# D Broader Impact

Our research introduces a novel perspective for multi-modal methodologies and a fresh distillation paradigm for camera-only techniques. We believe that it can establish a robust baseline for the broader scientific community. However, while our methods contribute to the enhancement of autonomous driving, they are not yet capable of addressing more complex corner cases. Consequently, these limitations could potentially introduce risks in real-world autonomous systems.

# References

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[2] MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. `https://github.com/open-mmlab/mmdetection3d`, 2020. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[4] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. BEVDepth: Acquisition of reliable depth for multi-view 3d object detection. *arXiv preprint arXiv:2206.10092*, 2022. 1

[5] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986, 2022. 2

[6] Jinhyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris Kitani, Masayoshi Tomizuka, and Wei Zhan. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. *arXiv preprint arXiv:2210.02443*, 2022. 2