BEVDiffuser: Plug-and-Play Diffusion Model for BEV Denoising with Ground-Truth Guidance

Supplementary Material

6. Model Architecture

We follow Latent Diffusion Models (LDMs) [24] to build a conditional diffusion model as our BEVDiffuser by augmenting the U-Net with cross-attention layers. The cross-attention operation is defined in Equation 10, where W_* represents learnable projection matrices unless otherwise specified, $\varphi_i(\boldsymbol{x}_t)$ denotes the intermediate embedding of $\boldsymbol{x}_t$ from the i-th layer of the U-Net, and $\tau_\theta(y)$ indicates the embedding of the condition y.

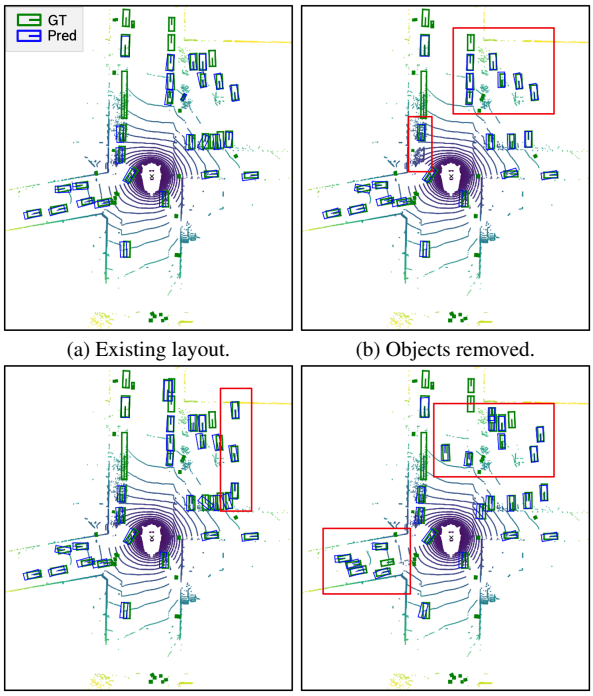
$$\begin{aligned} &cross\text{-}attn(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d}}) \cdot V \\ &Q = \varphi_i(\boldsymbol{x}_t)W_Q^i,\ K = \tau_\theta(y)W_K^i,\ V = \tau_\theta(y)W_V^i \end{aligned} \tag{10}$$

To better fuse the BEV feature map $\boldsymbol{x}_t$ and the layout condition $y = l$ and have more control over all the objects specified in the layout, we adopt the global conditioning and the object-aware local conditioning mechanism proposed by [39]. Specifically, we first use a transformer-based layout fusion module LFM as τ_θ to get a self-attended embedding o_i' for each object o_i as shown in Equation 11. In this way, o_0' contains the information of the entire layout and is then added to $\boldsymbol{x}_t$ for global conditioning, i.e., $\boldsymbol{x}_t' = \boldsymbol{x}_t + o_0'W_o$. Meanwhile, the embedding of all the objects $l' = \{o_i'\}_{i=0}^n$ is used to construct the key K_l and the value V_l of the layout for object-aware local conditioning. We adopt convolutional operations for the construction as shown by Equation 12. Similarly, we construct the query, key and value of the BEV feature as Equation 13 shows. To align the BEV feature with the layout, we divide the BEV feature map $\boldsymbol{x}_t$ equally into $k \times k$ bounding boxes, denoted by $\{b_x\}_1^{k\times k}$. We encode the bounding boxes from both BEV feature and layout, i.e., b_x and b_l, into the same embedding space using the shared weights W_b and W_p, and get the positional embedding P_x and P_l for the BEV feature and the layout, respectively (see Equation 14). P_x and P_l are utilized to generate the fused query, key and value by combining the BEV feature and the layout for the cross-attention operation, as formulated in Equation 15. $[\,\cdot\,]$ represents the concatenation operation.

$$\begin{aligned} l' = \{o_i'\}_{i=0}^n &= LFM(\{o_i\}_{i=0}^n) \\ &= self\text{-}attn(\{c_iW_c + b_iW_b\}_{i=0}^n) \end{aligned} \tag{11}$$

$$K_l,\ V_l = conv_{w_l}(l') \tag{12}$$

$$Q_x,\ K_x,\ V_x = conv_{w_x}(\varphi_i(\boldsymbol{x}_t')) \tag{13}$$



(a) Existing layout. (b) Objects removed. (c) Objects added. (d) Objects repositioned.

Figure 7. BEV feature maps generated by our BEVDiffuser (BD^{fu}) from pure noise, conditioned on user-defined layouts. We modify an existing layout (a) from nuScenes `mini-val` dataset by randomly removing (b), adding (c), and repositioning (d) some objects, as highlighted by the red boxes. BEVDiffuser generates accurate BEV feature maps, enabling the detection head to produce predictions that closely align with the ground truth.

$$P_x = b_xW_bW_p, \quad P_l = b_lW_bW_p \tag{14}$$

$$Q = \begin{bmatrix} Q_x \\ P_x \end{bmatrix},\ K = \begin{bmatrix} K_x & K_l \\ P_x & P_l \end{bmatrix},\ V = \begin{bmatrix} V_x & V_l \end{bmatrix} \tag{15}$$

7. Implementation Details

Our implementation is built upon the official BEVFormer implementation [1] and the MMCV implementation of the BEVFusion [2]. The hyperparameter λ and λ_{BEV} are empirically tuned based on the scale of the loss. Specifically,

[1] https://github.com/fundamentalvision/BEVFormer

[2] https://github.com/open-mmlab/mmdetection3d/tree/main/projects/BEVFusion

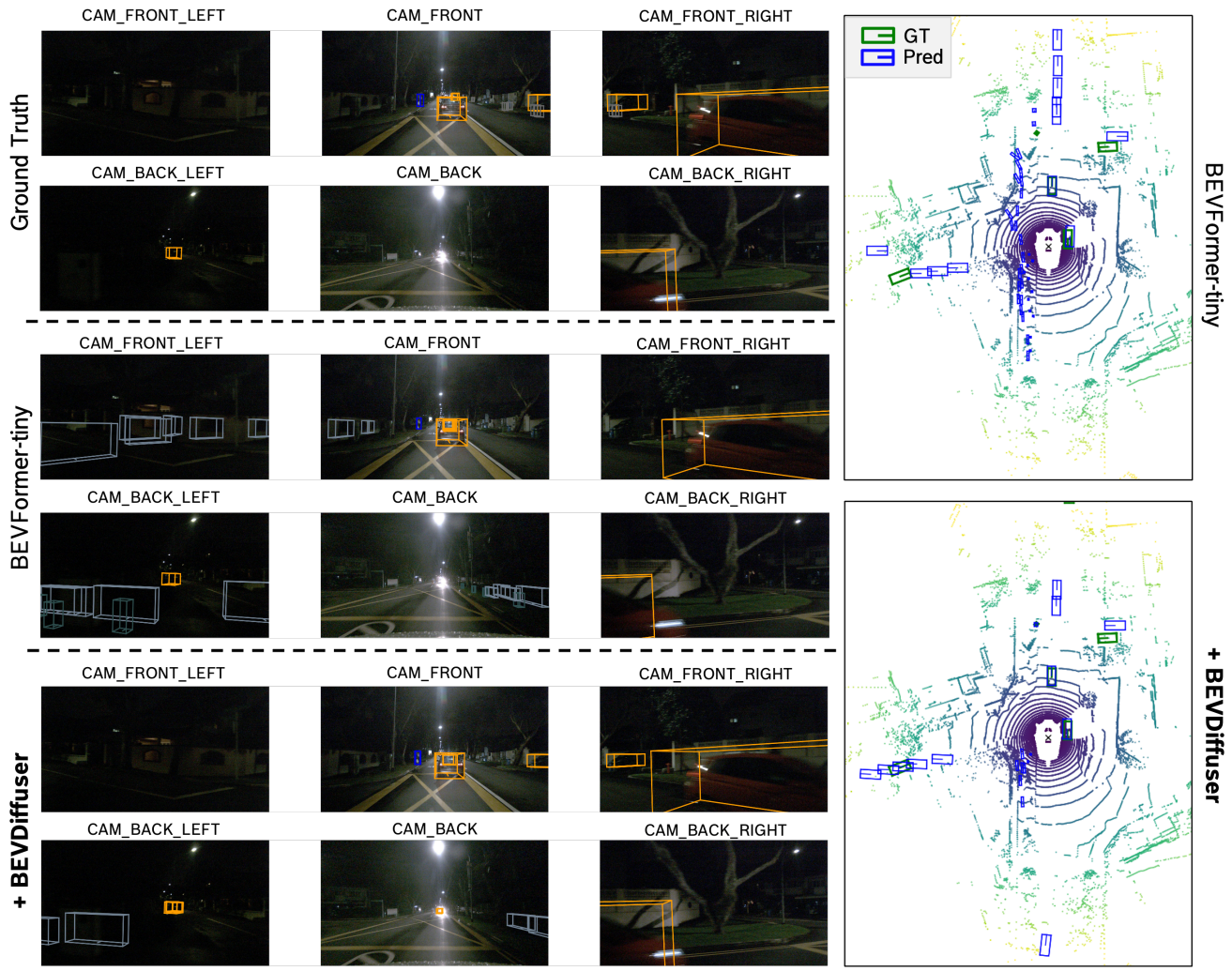


Figure 8. Visualization results of our BEVDiffuser enhanced BEVFormer-tiny on nuScenes `val` dataset. As shown in `CAM_FRONT` and `CAM_FRONT_RIGHT`, BEVDiffuser helps BEVFormer-tiny to detect the car intending to cross the road under the challenging lighting condition. Moreover, BEVDiffuser also helps to reduce hallucinations generated by BEVFormer-tiny, especially on `CAM_FRONT_LEFT`.

we configure λ and λ_{BEV} as follows: for BEVFormer-tiny and BEVFormer-base, $\lambda = 0.1$ and $\lambda_{BEV} = 100$; for BEVFormerV2, $\lambda = 0.05$ and $\lambda_{BEV} = 100$; and for BEVFusion, $\lambda = 0.2$ and $\lambda_{BEV} = 20$.

8. Ablation Study

We conduct an ablation study on BEVDiffuser (BD^{tiny}) to validate our design choices of layout conditioning and optimization objective, i.e. optimizing towards x_{t_0} with the task loss. Note that to optimize towards $\hat{\epsilon}_t$, we are not able to attach the task head or use the task loss. As shown in Tab. 5, without the task loss, whether we optimize towards x_{t_0} or $\hat{\epsilon}_t$, the denoising capability we obtained is quite limited, demonstrating that the task loss is critical to guarantee the denoising performance. Similarly, our layout conditioning also contributes to the superior denoising capability of BEVDiffuser, as evidenced by the inferior performance of the unconditional model.

Method	obj.	# denoising steps 1	3	5	10
Ours	x_{t_0}	**35.8/47.7**	**40.4/52.3**	**40.8/52.7**	**40.3/52.3**
−task	x_{t_0}	24.5/34.7	23.1/32.8	21.7/31.0	17.4/26.1
	$\hat{\epsilon}_t$	25.2/35.5	25.2/35.5	25.2/35.5	25.2/35.5
−cond.	x_{t_0}	25.4/35.4	25.3/35.3	25.1/35.0	24.7/34.6

Table 5. Ablation study. mAP/NDS achieved by the variants of BEVDiffuser (BD^{tiny}) with increasing denoising steps (1→10). Results validate that both the task loss and the layout conditioning contribute to the superior denoising capability of BEVDiffuser.

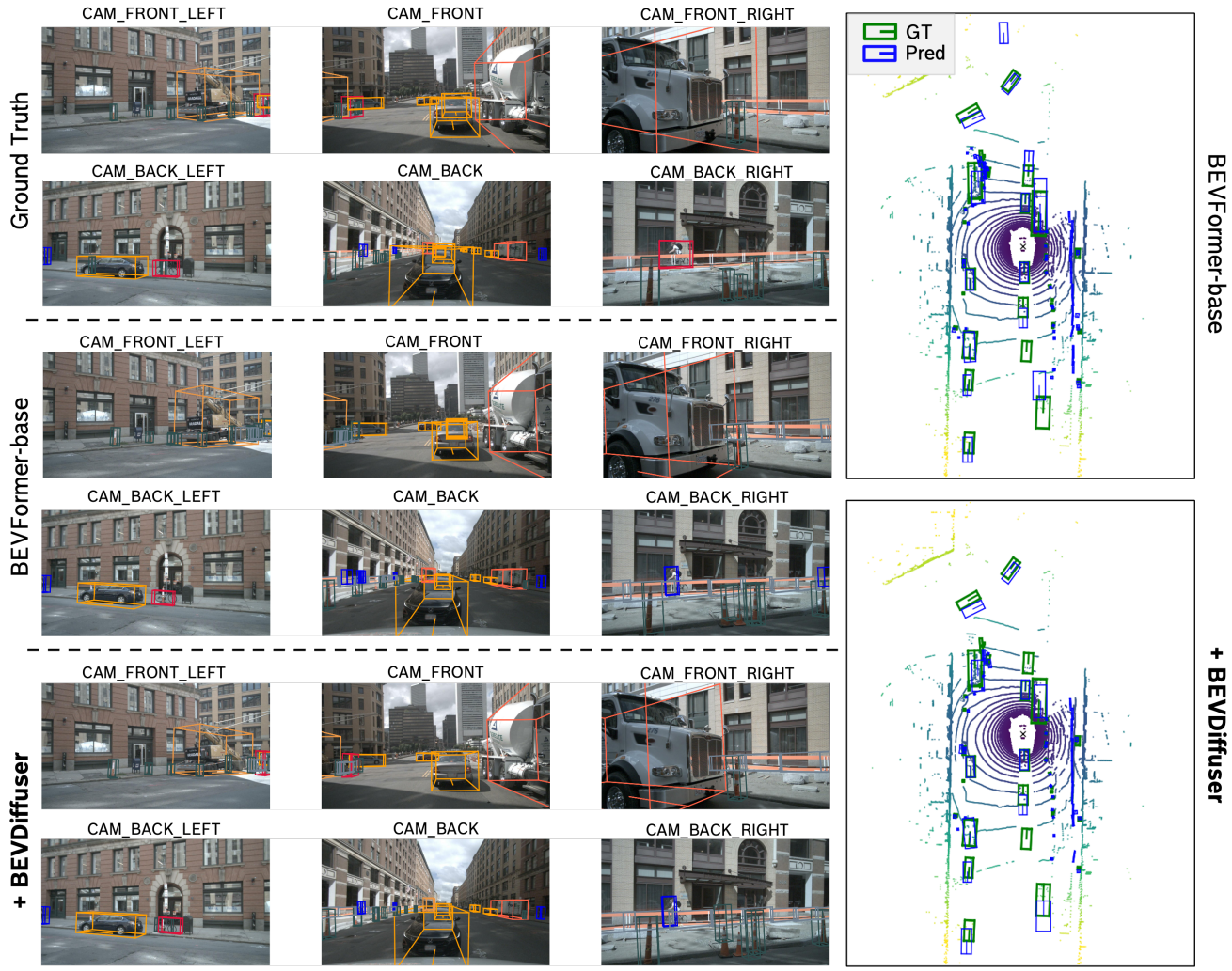


Figure 9. Visualization results of our BEVDiffuser enhanced BEVFormer-base on nuScenes `val` dataset. While BEVFormer-base shows good performance in the crowded environment, BEVDiffuser enhances its performance further, such as by detecting a human riding a bicycle in front of the autonomous vehicle, as indicated by the red bounding box in `CAM_FRONT` and `CAM_FRONT_LEFT`.

9. Additional Qualitative Results

9.1. Controllable BEV Generation

We present user-defined layout-conditioned BEV generation in Fig. 7. We modify an existing layout by randomly removing, adding, or repositioning some objects, and then condition the BEVDiffuser on the modified layouts to generate BEV feature maps. As shown in Fig. 7, BEVDiffuser is able to produce BEV feature maps that enable accurate object detection in alignment with the specified layouts, demonstrating its strong controllable generation capability. This capability facilitates easy adjustments to object presence and positioning in the BEV feature space, paving the way for large-scale data collection and driving world model development to advance autonomous driving.

9.2. 3D Object Detection

We visualize the 3D object detection results achieved by our BEVDiffuser enhanced BEVFormer-tiny, BEVFormer-base, BEVFormerV2 and BEVFusion in Fig. 8, Fig. 9, Fig. 10 and Fig. 11, respectively. We present the ground-truth and predicted 3D bounding boxes in both multi-camera images and the LiDAR top view to offer a comprehensive overview of the models' performance. As illustrated in the figures, BEVDiffuser consistently enhances the existing BEV models for object detection in complex environments and under challenging conditions by minimizing both false positives and false negatives, demonstrating its ability to improve the quality of the BEV representations.

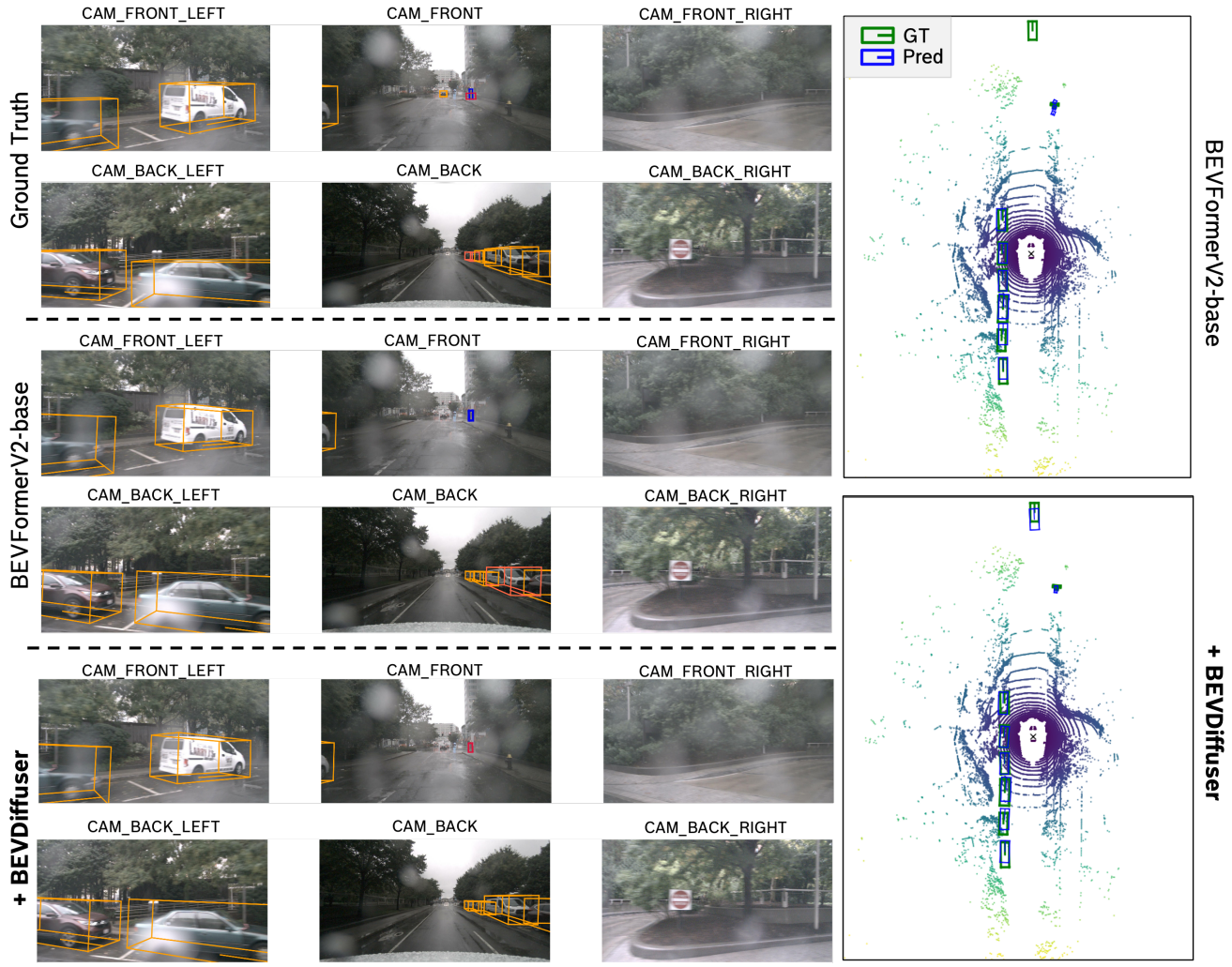


Figure 10. Visualization results of our BEVDiffuser enhanced BEVFormerV2 on nuScenes `val` dataset. In this representative example, despite the rain causing blurriness in the camera images, BEVDiffuser still enables BEVFormerV2 to reliably detect the object in front of the autonomous vehicle, as captured by the LiDAR top view.

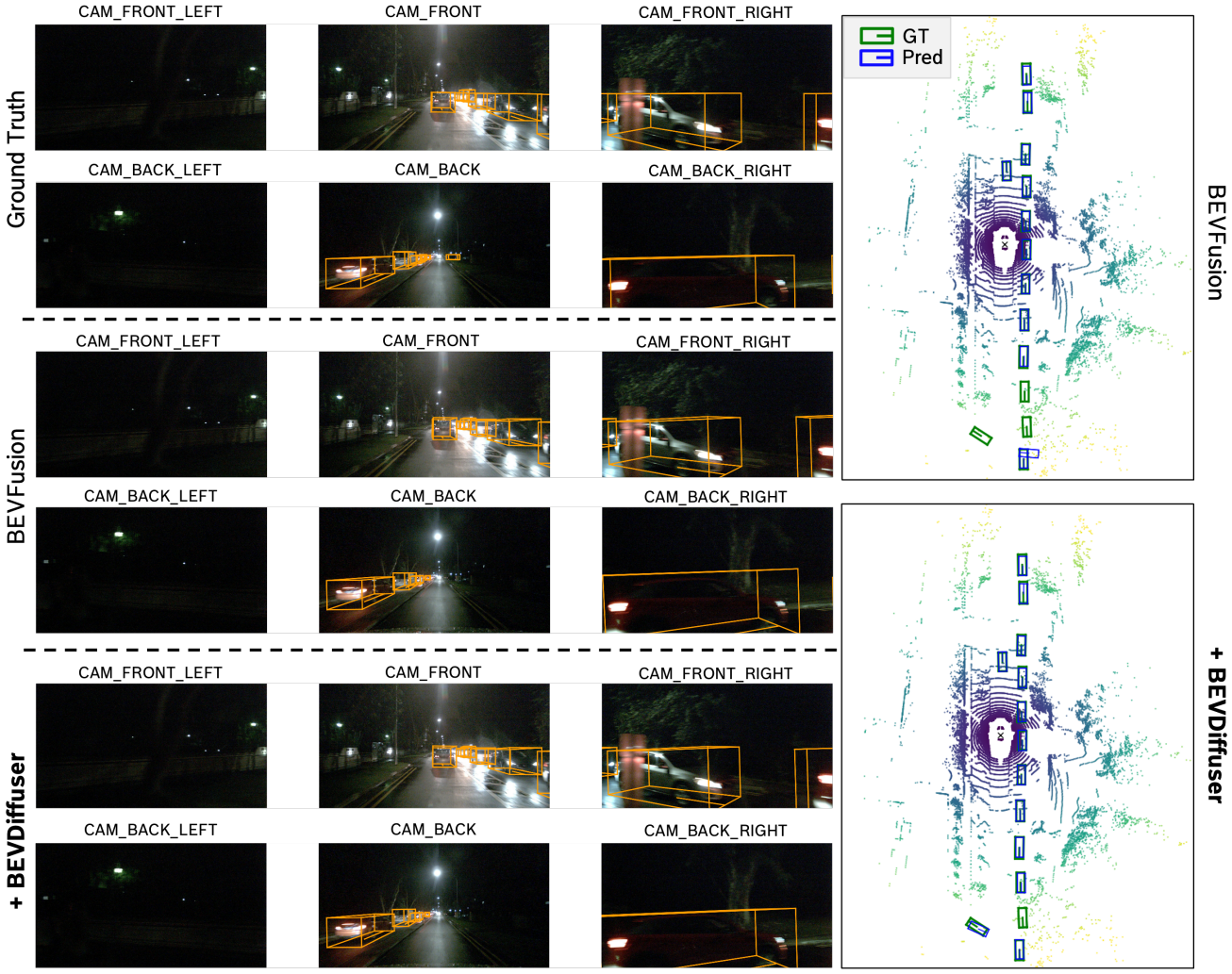


Figure 11. Visualization results of our BEVDiffuser enhanced BEVFusion on nuScenes `val` dataset. BEVFusion, which integrates both camera and LiDAR data, delivers robust performance in low-light conditions at night. BEVDiffuser further enhances BEVFusion by effectively reducing false negatives, as demonstrated in the LiDAR top view.