CoLC: Communication-Efficient Collaborative Perception with LiDAR Completion

Supplementary Material

This supplementary material offers extra information to support the CoLC, including (1) Experimental Settings, (2) Point Selector, (3) VQ-based LiDAR Completion, (4) Quantitative Evaluation and (5) Qualitative Evaluation.

1. Experimental Settings

Datasets. We use four datasets: V2XSim, OPV2V, V2XSet and DAIR-V2X. The datasets include both simulated and real-world scenarios, and cover three types of collaboration: V2V, V2I and V2X.

V2XSim [2] is a simulated dataset for V2X collaborative perception, containing 10,000 frames captured using a 32-channel LiDAR sensor at a recording interval of 0.2 seconds. Each scene involves 2 to 5 collaborating agents. The dataset is split into 8,000 training frames, 1,000 validation frames, and 1,000 testing frames. The LiDAR sensing range is set as $x \in [-32m, 32m]$ and $y \in [-32m, 32m]$.

OPV2V [7] is a large-scale V2V collaborative perception dataset. It contains 11,464 frames with synchronized 64-beam LiDAR point clouds and RGB images. The dataset is split into 6,764 training frames, 1,981 validation frames, and 2,719 testing frames. Each scene involves up to 5 collaborating agents. The LiDAR sensing range is set as $x \in [-140.8m, 140.8m]$ and $y \in [-40m, 40m]$.

V2XSet [6] is a simulated V2X perception dataset, incorporating both roadside units (RSUs) and autonomous vehicles. It comprises 6,694 training frames, 1,920 validation frames, and 2,834 testing frames. Each frame includes synchronized 32-beam LiDAR point clouds and RGB images, annotated with 3D bounding boxes. Up to five agents collaborate in each scene. The LiDAR sensing range is set as $x \in [-140.8m, 140.8m]$ and $y \in [-40m, 40m]$.

DAIR-V2X [9] is a real-world V2I perception dataset. Each scene involves one roadside unit (RSU) and one vehicle, both equipped with LiDAR sensors. The RSU uses a 300-line LiDAR, while the vehicle is equipped with a 40-line LiDAR. The dataset is split into training, validation, and testing sets in a 5:2:3 ratio. The LiDAR sensing range is set as $x \in [-100.8m, 100.8m]$ and $y \in [-40m, 40m]$.

During training, all models are trained on the training set, and the best-performing checkpoint is selected based on validation performance. The selected model is then evaluated on the test set. We reproduce all baselines using their publicly available code on NVIDIA 4090 GPUs.

Communication Volume. Following [1], communication volume is measured by the number of transmitted elements $\mathcal{M}$ as $\log_2(\mathcal{M} \times 32/8)$, where each element is a 32-bit float (float32), and division by 8 converts bits to bytes.

- Given a feature $\mathcal{F} \in \mathbb{R}^{H \times W \times C}$, the communication volume is calculated as: $\log_2(H \times W \times C \times 32/8)$.
- For point cloud $\mathcal{X} \in \mathbb{R}^{N \times 4}$, the communication volume is: $\log_2(N \times 4 \times 32/8)$.
- For detection results consisting of bounding boxes $\mathcal{B} \in \mathbb{R}^{M \times 7}$ and classification scores $\mathcal{S} \in \mathbb{R}^{M \times 1}$, the communication volume is: $\log_2(M \times (7+1) \times 32/8)$.

Note CoSDH [5] converts features from float32 to float16, therefore, its communication volume is calculated as: $\log_2(H \times W \times C \times 16/8)$.

More Implementation Details. The pillarization $\Phi_{\text{p}}(\cdot)$ follows the PointPillars implementation in OpenCOOD, which converts point clouds into voxelized pillars and encodes them into pillar-wise features using PFN layers. During training, we use RPS at a 0.1 ratio for point sampling for efficiency, which generalizes well to the FAPS sampling used at inference (See Table 2 and Table 3). When transmitting the full point cloud, LiDAR completion becomes unnecessary. In this case, the complementary pillar fusion is also omitted, as it provides no additional performance gain (See Table 7). To mitigate point misalignment caused by pose errors or latency, we apply the Iterative Closest Point (ICP) algorithm [10] to align the received sparse point clouds with the ego view. The implementation is based on the CUDA-accelerated Cupoch library [4], using a point-to-plane metric with voxel downsampling (0.4 m). Convergence is determined by relative fitness and RMSE thresholds of 1×10^{-3}, with up to 10 iterations. ICP alignment is applied only to CoLC and early fusion baseline, which operate on raw point clouds requiring geometric alignment (See Table 9).

2. Point Selector

We adopt a lightweight MLP-based point selector to predict point-wise foreground probabilities. The network adopts the hierarchical feature extraction [3], consisting of three

Set Abstraction (SA) and three Feature Propagation (FP) blocks, each implemented with shared MLPs. The input is the raw point cloud, and the output is a saliency score indicating the probability of each point belonging to the foreground. Since all operations are shared MLPs, the model remains computationally lightweight and can be efficiently deployed on edge agents.

A natural alternative for foreground point selection is to leverage an object detector's output for bounding-box filtering. To validate the effectiveness of the point selector, we compare it against the detector-based approach. As summarized in Table 1, our MLP-based selector achieves competitive performance while being dramatically more compact. Furthermore, the selector is task-agnostic and can be readily applied to other perception tasks, particularly when neighbor agents do not perform detection.

Table 1. Results on V2XSim, where R^{fg}=0.2 and R^{bg}=0.1.

Model	AP@0.5/0.7↑	Comm↓	MB↓
Point Selector	88.89 / 79.28	15.35	1.7
PointPillars	90.44 / 80.17	15.18	30.82

3. VQ-based LiDAR Completion

Codebook Reinitialization. Due to the inherent sparsity and heterogeneity of point clouds and their corresponding pillar representations, the codebook is more susceptible to unbalanced usage (i.e., codebook collapse). To improve codebook learning, we adopt a data-driven initialization strategy. Specifically, we maintain a memory bank that stores continuous latent embeddings from the encoder at each iteration. When the codebook utilization rate drops below a predefined threshold $T_{\text{use}} = 0.3$ (i.e., fewer than 3% of codes are active) The number of iterations since the last update exceeds a predefined threshold $T_{\text{iter}} = 1000$, we reinitialize the codebook using K-means centroids computed from the memory bank. This strategy significantly improve codebook utilization and training stability.

Training Details. For each agent in the collaborative scene, we randomly sample its point cloud to obtain a sparse version, which is then paired with the original full point cloud to form a sparse-dense pair. To ensure stable optimization, we pre-train the LiDAR completion module (Algorithm 1) separately from the detection model. The completion network focuses on reconstruction and quantization objectives, while the detector optimizes task-specific losses, which may yield conflicting gradients when trained jointly. In addition, the quantization step's gradient approximation and differing optimization settings (e.g., learning rates, regularization) further hinder stable convergence.

Code Usage Histogram. We visualize the codebook usage histogram to assess whether the quantized embeddings suffer from collapse. As shown in Figure 1, the usage distribution is relatively balanced, indicating that most codes are effectively utilized during training and our (re)-initialization strategies successfully mitigate codebook collapse.

4. Quantitative Evaluation

Ablation Study. We first evaluate different point sampling strategies during training. As shown in Table 2 and 3, training completion module and detector with RPS yields comparable performance to FAPS while accelerating training, whereas applying FAPS during inference remains critical. Table 4 evaluates the adaptive fusion in CEEF with and without the occupancy mask $\mathcal{M}i^{\text{se}}$, which is used to preserve the fidelity of the initial sparse early fusion. The results show that directly applying adaptive fusion without $\mathcal{M}i^{\text{se}}$ leads to a drop in detection accuracy, as the completed pillars introduce additional noise.

Table 5 to 8 provide numerical results for Figure 5 (e) to (h) in the main content. The results demonstrate that all three modules contribute significantly to achieving a better performance-communication trade-off. Furthermore, we evaluate the effectiveness of ICP alignment on pose error. As shown in Table 9, ICP consistently improves performance by mitigating pose-induced misalignment. For fair comparison, both CoLC and the early fusion baseline use the same ICP preprocessing.

Efficiency Comparison. Table 10 presents the inference latency of collaborative perception methods on four datasets. Among the methods, Where2comm demonstrates the best inference efficiency, while CoLC, CoBEVT, and V2X-ViT exhibit comparable inference times.

Accuracy-Bandwidth Trade-Off Comparison. Figure 2 (1-a) to (1-d) compares the proposed CoLC with existing methods in terms of the trade-off between detection performance and communication bandwidth, while Table 11 presents the corresponding quantitative results. The results demonstrate that CoLC consistently achieves a superior perception-communication trade-off across all bandwidth levels and collaborative perception settings.

Robustness to Heterogeneous Scenarios. Figure 2 (2-a) to (2-d) and Table 12 evaluate CoLC under heterogeneous scenarios, where models trained in homogeneous settings are applied to agents with different architectures. We apply the models trained under the homogeneous setting to the heterogeneous scenario. To address this feature discrepancy (spatial resolution and channel dimension) when applying intermediate fusion methods, we adopt bilinear interpolation to spatially align the received features with the ego view in height and width, and apply channel-wise dropping or padding to match the channel dimension [8].

In intermediate or late fusion methods, agents exchange features or outputs generated by their local models. However, when model architectures differ, such heterogeneity

input : Sparse pillar $\mathcal{P}^{\mathrm{s}}$, ground-truth dense pillar $\mathcal{P}^{\mathrm{d}}$ and its occupancy grid $\mathcal{O}^{\mathrm{d}}$, sparse encoder $\Psi^{\mathrm{s}}_{\mathrm{enc}}$, dense decoder $\Psi^{\mathrm{d}}_{\mathrm{dec}}$, codebook E, code utilization threshold $T_{\mathrm{use}} = 0.03$, iteration threshold $T_{\mathrm{iter}} = 1000$, memory bank $\mathcal{B}$

Preparation:

Initialize encoder $\Psi^{\mathrm{s}}_{\mathrm{enc}}$, decoder $\Psi^{\mathrm{d}}_{\mathrm{dec}}$, codebook E

$t_{num} \leftarrow 0$ ▷ Initialize the counter of iteration

Do training iteration:

for *each training step* $t = 1$ *to* T **do**

- $\mathbf{z}^{\mathrm{s}} \leftarrow \Psi^{\mathrm{s}}_{\mathrm{enc}}(\mathcal{P}^{\mathrm{s}})$ ▷ Encode sparse pillar
- $\mathbf{z}^{\mathrm{q}} \leftarrow \Psi_{\mathrm{vq}}(\mathbf{z}^{\mathrm{s}}, E)$ ▷ VQ with nearest neighbors
- $\mathcal{B} \leftarrow \mathcal{B} \cup \mathbf{z}^{\mathrm{s}}$ ▷ Update memory bank
- **if** *code utilization rate* $< T_{use}$ *and* ($t_{num} = 0$ *or* $t_{num} > T_{use}$) **then**
  - $E \leftarrow \mathrm{KMeans}(\mathcal{B})$ ▷ Reinitialize codebook from memory bank
  - $t_{num} \leftarrow 0$ ▷ Set the iteration counter to zero
- $t_{num} \leftarrow t_{num} + 1$
- $\hat{\mathcal{P}}^{\mathrm{d}}, \hat{\mathcal{O}}^{\mathrm{d}} \leftarrow \Psi^{\mathrm{d}}_{\mathrm{dec}}(\mathbf{z}^{\mathrm{q}})$ ▷ Get dense pillar
- Compute reconstruction loss: $\mathcal{L}_{\mathrm{rec}} = \mathrm{BCE}(\hat{\mathcal{O}}^{\mathrm{d}}, \mathcal{O}^{\mathrm{d}}) + \frac{1}{|\mathcal{O}^{\mathrm{d}}|} \sum_{i,j} \mathcal{O}^{\mathrm{d}}_{i,j} \| \hat{\mathcal{P}}^{\mathrm{d}}_{i,j} - \mathcal{P}^{\mathrm{d}}_{i,j} \|_2^2$
- Compute quantization loss: $\mathcal{L}_{\mathrm{vq}} = \| \mathrm{sg}[\mathbf{z}^{\mathrm{s}}] - \mathbf{z}^{\mathrm{q}} \|_2^2 + \beta \| \mathbf{z}^{\mathrm{s}} - \mathrm{sg}[\mathbf{z}^{\mathrm{q}}] \|_2^2$
- Backpropagate and update $\Psi^{\mathrm{s}}_{\mathrm{enc}}$, $\Psi^{\mathrm{d}}_{\mathrm{dec}}$, and E

output: sparse encoder $\Psi^{\mathrm{s}}_{\mathrm{enc}}$, dense decoder $\Psi^{\mathrm{d}}_{\mathrm{dec}}$, codebook E

Algorithm 1: Training Procedure of the VQ-based LiDAR Completion Module

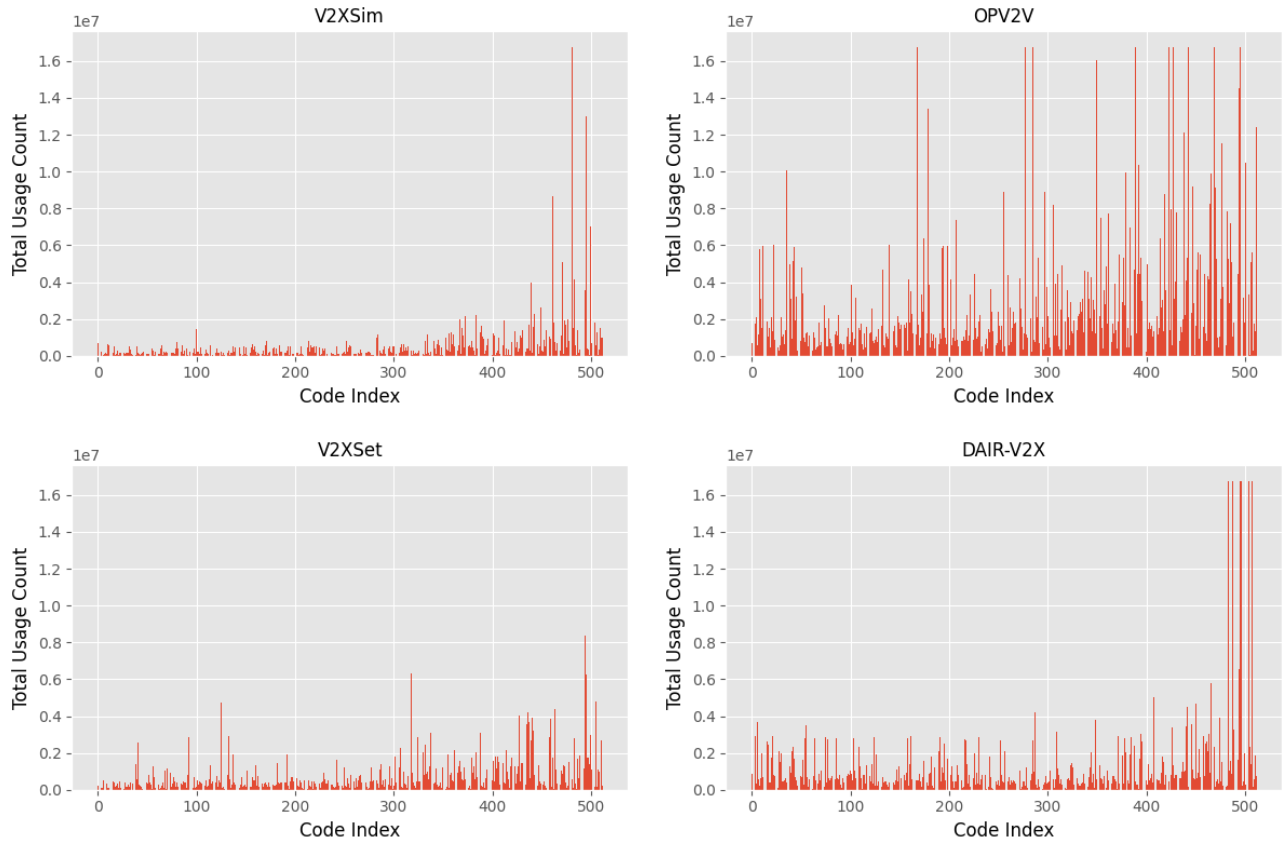


Figure 1. Codebook Usage Histogram.

Table 2. Ablation study (AP@0.5/0.7) on the sampling strategy used during detector training and inference. For fair comparison, R^{fg} and R^{bg} in FAPS are set to 0.2, and the random sampling ratio R is also set to 0.2, ensuring the same communication volume.

Training/Inference	100% LiDAR	FAPS(R^{fg}=$R^{bg}=0.2$)	RPS($R=0.2$)
RPS(R=0.1)	95.14 / 87.89	91.31 / 81.57	89.69 / 79.48
FAPS(R^{fg}=0.2, R^{bg}=0.1)	95.63 / 87.65	91.25 / 80.82	89.35 / 77.95

Table 3. Ablation study (AP@0.5/0.7) on the sampling strategy used during VQ-based completion module training. During inference, we use FAPS with R^{fg}=0.2.

Training/Inference	$R^{bg}=0.5$	$R^{bg}=0.1$	$R^{bg}=0.01$
RPS(R=0.1)	93.47 / 85.28	88.89 / 79.28	83.44 / 70.47
FAPS(R^{fg}=0.2, R^{bg}=0.1)	93.50 / 85.04	88.80 / 78.64	83.71 / 71.66

Table 4. Ablation study (AP@0.5/0.7) on occupancy mask $\mathcal{M}_i^{\text{se}}$ in the adaptive fusion.

with	100% LiDAR	$R^{bg}=0.5$	$R^{bg}=0.1$	$R^{bg}=0.01$
w/ $\mathcal{M}_i^{\text{se}}$	95.14 / 87.89	93.47 / 85.28	88.89 / 79.28	83.44 / 70.47
w/o $\mathcal{M}_i^{\text{se}}$	92.40 / 85.18	92.72 / 84.18	88.90 / 78.16	81.27 / 68.04

can introduce semantic misalignment, leading to noticeable performance degradation. In extreme cases, intermediate or late fusion methods may even underperform the no-fusion baseline. In contrast, CoLC remains robust across different sampling ratios, as its collaboration is based on raw LiDAR data, which is inherently agnostic to model heterogeneity.

Robustness to Pose Error. Figure 2 (3-a) to (3-d) and Table 13 to Table 16 evaluate the robustness of the CoLC under positional and heading perturbations. The results show that CoLC outperforms intermediate fusion methods across all noise levels in V2XSim, OPV2V and V2XSet, and consistently achieves better performance than the no fusion baseline, demonstrating its robustness to pose noises.

Robustness to Latency. Figure 2 (4-a) to (4-d) and Table 17 to Table 20 evaluate the robustness of the CoLC under latency. The results show that CoLC consistently outperforms the no-fusion baseline across all latency levels in V2XSim and DAIR-V2X. However, in OPV2V and V2XSet, where temporal asynchrony is more severe ($\geq 200ms$), early fusion methods (including CoLC) and several intermediate fusion methods exhibit performance degradation. We attribute this phenomenon to the inherent domain shift and higher scene complexity present in OPV2V and V2XSet. In the future, we plan to improve the temporal and spatial robustness of early collaborative perception in the face of real-world latency variations.

5. Qualitative Evaluation

Completion Results. Figures 3 to 6 present the pillar-level LiDAR completion results across four datasets. We provide results from multiple agent perspectives to demonstrate generalizability. The results show that our completion module performs robustly under varying LiDAR configurations (e.g., different beam numbers). Even at very low sampling ratios, the module successfully reconstructs dense pillar structures in the regions surrounding each agent.

Detection Results. Figures 7 to 10 present qualitative comparisons between CoLC and existing methods across multiple datasets. In most scenarios, CoLC demonstrates superior detection accuracy. The figures also include visualization results under varying point sampling ratios. Notably, as the sampling ratio decreases, CoLC consistently maintains correct object classifications, suggesting that FAPS effectively selects informative points while CEEF successfully reconstructs missing spatial information.

References

[1] Yue Hu, Juntong Peng, Sifei Liu, Junhao Ge, Si Liu, and Siheng Chen. Communication-efficient collaborative perception via information filling with codebook. In *CVPR*, pages 15481–15490, 2024. 1

[2] Yiming Li, Dekun Ma, Ziyan An, Zixun Wang, Yiqi Zhong, Siheng Chen, and Chen Feng. V2X-Sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving. *RAL*, 7(4):10914–10921, 2022. 1

[3] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, 2017. 1

[4] Kenta Tanaka. cupoch – robotics with gpu computing, 2020. https://github.com/neka-nat/cupoch. 1

[5] Junhao Xu, Yanan Zhang, Zhi Cai, and Di Huang. Cosdh: Communication-efficient collaborative perception via supply-demand awareness and intermediate-late hybridization. In *CVPR*, pages 6834–6843, 2025. 1

[6] Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2X-ViT: Vehicle-to-everything cooperative perception with vision transformer. In *ECCV*, pages 107–124, 2022. 1

[7] Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi Ma. OPV2V: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *ICRA*, pages 2583–2589, 2022. 1

[8] Runsheng Xu, Jinlong Li, Xiaoyu Dong, Hongkai Yu, and Jiaqi Ma. Bridging the domain gap for multi-agent perception. In *ICRA*, pages 6035–6042, 2023. 2

[9] Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, et al. DAIR-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In *CVPR*, pages 21361–21370, 2022. 1

[10] Juyong Zhang, Yuxin Yao, and Bailin Deng. Fast and robust iterative closest point. *TPAMI*, 44(7):3450–3466, 2021. 1

Table 5. Ablation study on three components in V2XSim. When $R^{bg} \neq 1.0$, the R^{fg} is fixed to 0.2.

FAPS	CEEF	DGDA	$R^{bg}=1.0$		$R^{bg}=0.5$		$R^{bg}=0.2$		$R^{bg}=0.1$		$R^{bg}=0.05$		$R^{bg}=0.01$	
			AP@0.5/0.7	Comm	AP@0.5/0.7	Comm	AP@0.5/0.7	Comm	AP@0.5/0.7	Comm	AP@0.5/0.7	Comm	AP@0.5/0.7	Comm
✓			94.68 / 83.61	18.37	91.35 / 79.26	17.38	87.32 / 74.17	16.07	83.95 / 70.72	15.35	81.46 / 68.49	14.85	78.17 / 64.21	14.41
✓	✓		94.98 / 85.77	18.37	93.35 / 83.33	17.38	90.64 / 80.24	16.07	88.48 / 77.03	15.35	85.15 / 73.51	14.85	81.62 / 69.20	14.41
✓	✓	✓	95.14 / 87.89	18.37	93.47 / 85.28	17.38	91.31 / 81.57	16.07	88.89 / 79.28	15.35	87.04 / 76.00	14.85	83.44 / 70.47	14.41

Table 6. Ablation study on FAPS in V2XSim. We use random point sampling (RPS) for Background points in all settings.

Methods	R^{fg}	$R^{bg}=1.0$		$R^{bg}=0.5$		$R^{bg}=0.2$		$R^{bg}=0.1$		$R^{bg}=0.05$		$R^{bg}=0.01$	
		AP@0.5/0.7	Comm	AP@0.5/0.7	Comm	AP@0.5/0.7	Comm	AP@0.5/0.7	Comm	AP@0.5/0.7	Comm	AP@0.5/0.7	Comm
FG-RPS	$=R^{bg}$	94.68 / 83.61	18.37	91.74 / 78.76	17.38	86.54 / 72.75	16.07	80.43 / 66.65	15.05	76.30 / 62.04	14.05	- / -	-
FG-All	1.0	94.68 / 83.61	18.37	92.21 / 78.95	17.75	88.22 / 74.93	17.05	84.84 / 71.39	16.89	82.39 / 67.87	16.68	78.35 / 61.82	16.56
FG-RPS	0.2	94.68 / 83.61	18.37	91.04 / 76.88	17.38	86.54 / 72.75	16.07	82.95 / 68.88	15.35	79.97 / 64.87	14.85	76.22 / 59.18	14.41
FG-FPS	0.2	94.68 / 83.61	18.37	91.35 / 79.26	17.38	87.32 / 74.17	16.07	83.95 / 70.72	15.35	81.46 / 68.49	14.85	78.17 / 64.21	14.41

Table 7. Ablation study on progressive fusion in CEEF: sparse early fusion (SEF), pillar completion (PC), and adaptive complementary fusion (ACF). When $R^{bg} \neq 1.0$, the R^{fg} is fixed to 0.2. When transmitting the full point cloud, LiDAR completion becomes unnecessary. In this case, the complementary pillar fusion is also omitted, as it provides no additional performance gain (w/ 87.83 VS w/o 87.89).

SEF	PC	ACF	$R^{bg}=1.0$		$R^{bg}=0.5$		$R^{bg}=0.2$		$R^{bg}=0.1$		$R^{bg}=0.05$		$R^{bg}=0.01$	
			AP@0.5/0.7	Comm	AP@0.5/0.7	Comm	AP@0.5/0.7	Comm	AP@0.5/0.7	Comm	AP@0.5/0.7	Comm	AP@0.5/0.7	Comm
✓			95.14 / 87.89	18.37	93.25 / 84.69	17.38	89.21 / 78.92	16.07	84.65 / 74.68	15.35	79.18 / 70.89	14.85	72.88 / 65.22	14.41
		✓	94.98 / 86.06	18.37	92.09 / 82.03	17.38	86.73 / 76.51	16.07	82.12 / 72.68	15.35	77.30 / 69.25	14.85	72.57 / 65.37	14.41
✓		✓	95.11 / 87.83	18.37	93.29 / 84.05	17.38	89.32 / 78.79	16.07	84.50 / 74.91	15.35	79.18 / 70.89	14.85	72.88 / 65.22	14.41
	✓	✓	- / -	18.37	92.66 / 82.72	17.38	89.77 / 78.79	16.07	87.32 / 76.83	15.35	85.16 / 73.62	14.85	81.95 / 69.54	14.41
✓	✓	✓	- / -	18.37	93.47 / 85.28	17.38	91.31 / 81.57	16.07	88.89 / 79.28	15.35	87.04 / 76.00	14.85	83.44 / 70.47	14.41

Table 8. Ablation study on DGDA in V2XSim. When $R^{bg} \neq 1.0$, the R^{fg} is fixed to 0.2.

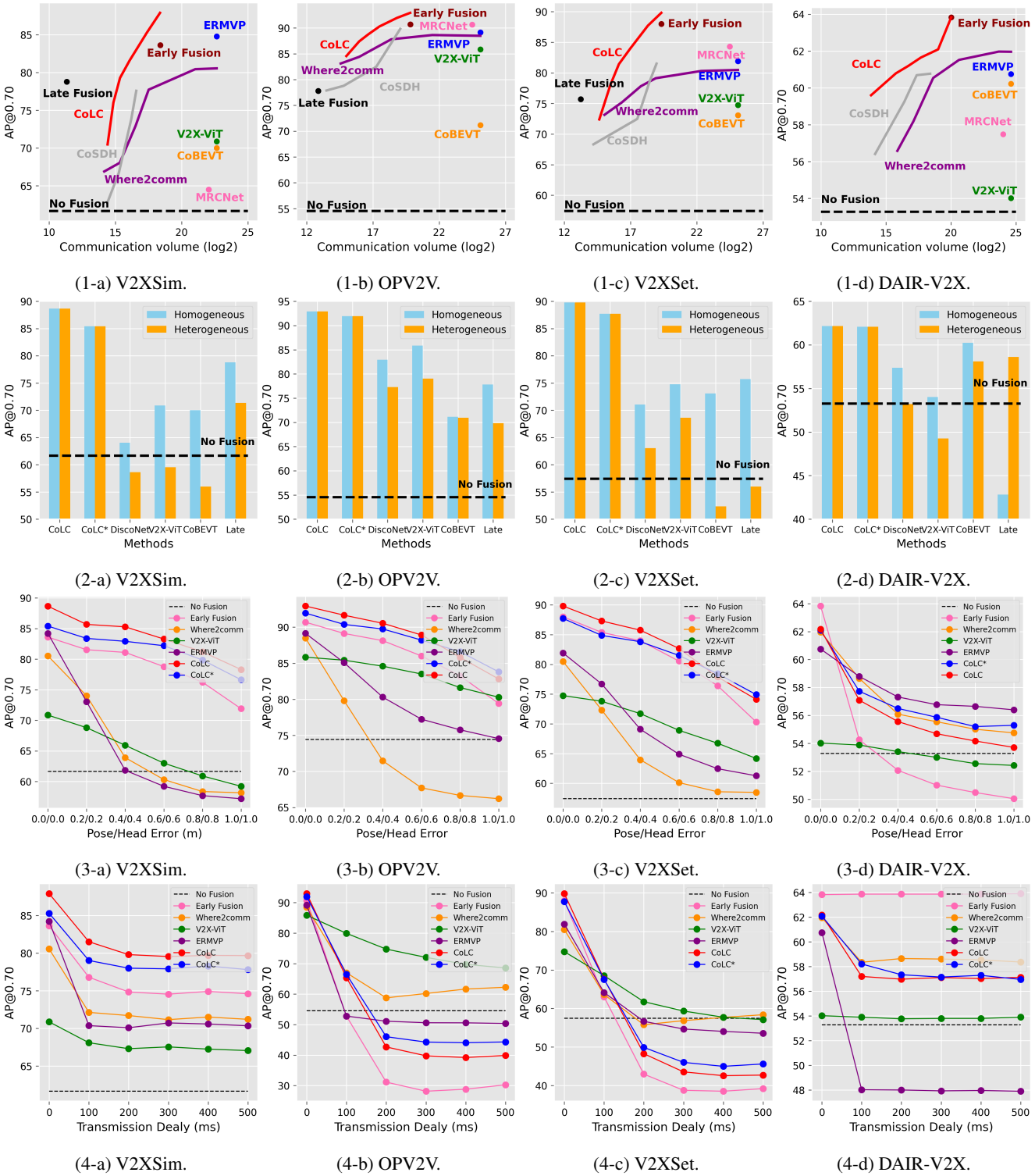
$\mathcal{L}_{\text{sda}}$	$\mathcal{L}_{\text{gda}}$	$R^{bg}=1.0$		$R^{bg}=0.5$		$R^{bg}=0.2$		$R^{bg}=0.1$		$R^{bg}=0.05$		$R^{bg}=0.01$	
		AP@0.5/0.7	Comm	AP@0.5/0.7	Comm	AP@0.5/0.7	Comm	AP@0.5/0.7	Comm	AP@0.5/0.7	Comm	AP@0.5/0.7	Comm
		94.98 / 85.77	18.37	93.35 / 83.33	17.38	90.64 / 80.24	16.07	88.48 / 77.03	15.35	85.15 / 73.51	14.85	81.62 / 69.20	14.41
✓		94.81 / 87.04	18.37	92.99 / 83.94	17.38	90.53 / 80.20	16.07	88.75 / 77.88	15.35	86.33 / 75.16	14.85	81.83 / 69.58	14.41
	✓	94.18 / 85.42	18.37	92.87 / 82.58	17.38	90.48 / 79.63	16.07	88.16 / 76.82	15.35	85.51 / 73.92	14.85	81.04 / 68.93	14.41
✓	✓	95.14 / 87.89	18.37	93.47 / 85.28	17.38	91.31 / 81.57	16.07	88.89 / 79.28	15.35	87.04 / 76.00	14.85	83.44 / 70.47	14.41

Table 9. The performance (AP@0.5/0.7) of early fusion methods with (w/) or without (w/o) ICP on pose error, evaluated on V2XSim.

Noise Level $\sigma_t/\sigma_r\,(m/^\circ)$	0.0/0.0	0.2/0.2	0.4/0.4	0.6/0.6	0.8/0.8	1.0/1.0
Early Fusion (w/ ICP)	94.68/83.61	92.06/81.55	91.98/81.11	91.30/78.79	90.15/76.24	87.27/71.91
Early Fusion (w/o ICP)	94.68/83.61	89.98/56.06	65.02/16.71	34.92/4.24	16.40/1.57	8.97/0.63
CoLC (w/ ICP)	95.14/87.89	93.17/85.69	93.08/85.29	92.25/83.30	91.05/81.21	89.10/78.31
CoLC (w/o ICP)	95.14/87.89	91.04/65.84	73.95/45.04	59.39/35.43	50.44/32.52	46.66/30.70
CoLC* (w/ ICP)	93.47/85.28	91.95/83.40	91.77/82.92	91.16/82.21	89.85/79.87	88.26/76.64
CoLC* (w/o ICP)	93.47/85.28	8984/67.84	75.84/50.35	64.22/42.68	58.00/39.61	53.92/38.64

Table 10. Detailed inference time (ms), frames per second (FPS) and AP@0.7 for CoLC and baseline methods across multiple datasets, including the breakdown of ego and neighbor agent computation time.

Datasets	V2XSim			OPV2V			V2XSet			DAIR-V2X		
Method/Metrics	Time (ms)	FPS	AP@0.7	Time (ms)	FPS	AP@0.7	Time (ms)	FPS	AP@0.7	Time (ms)	FPS	AP@0.7
Where2comm	69.7	14.328	80.54	104.7	9.545	88.48	128.2	7.7950	80.48	108.7	9.19	61.96
V2X-ViT	197.7	5.056	70.86	209.0	4.783	85.84	224.5	4.453	74.75	129	7.74	54.02
CoBEVT	84.5	11.823	69.99	230.0	4.347	71.16	236.2	4.233	73.08	116.3	8.597	60.23
ERMVP	100.5	9.9452	84.76	4783	0.209	89.14	914.02	1.094	81.91	138.1	7.239	60.75
CoLC (FAPS)	27.2	-	-	61.07	-	-	47.0	-	-	54.63	-	-
CoLC (ICP)	10.7	-	-	14.66	-	-	13.41	-	-	16.49	-	-
CoLC (CEEF+Det)	37.5	-	-	75.4	-	-	95.3	-	-	76.46	-	-
CoLC (All)	75.86	13.18	87.89	151,1	6.614	92.93	155.7	6.420	89.81	147.5	6.775	62.17



(1-a) V2XSim. (1-b) OPV2V. (1-c) V2XSet. (1-d) DAIR-V2X.

(2-a) V2XSim. (2-b) OPV2V. (2-c) V2XSet. (2-d) DAIR-V2X.

(3-a) V2XSim. (3-b) OPV2V. (3-c) V2XSet. (3-d) DAIR-V2X.

(4-a) V2XSim. (4-b) OPV2V. (4-c) V2XSet. (4-d) DAIR-V2X.

Figure 2. (1) Accuracy-Bandwidth Trade-Off. (2) Robustness to Heterogeneous Scenarios. (3) Robustness to Pose Error. (4) Robustness to Communication Latency.

Table 11. Overall performance and communication volume. The communication volume is denoted as B.

	V2XSim		OPV2V		V2XSet		DAIR-V2X	
Method/Metric	B	AP@0.5/0.7	B	AP@0.5/0.7	B	AP@0.5/0.7	B	AP@0.5/0.7
No Fusion	0	73.72/61.65	0	74.42/54.52	0	74.18/57.43	0	64.32/53.27
Late Fusion	11.32	86.87/78.78	12.86	92.51/77.82	13.24	89.05/75.72	12.51	62.05/42.82
Early Fusion	18.38	94.68/83.61	19.81	96.13/90.69	19.32	94.59/88.00	20.02	76.51/**63.83**
DiscoNet	22.65	77.46/64.03	25.11	93.45/82.95	25.11	86.28/71.03	24.62	73.21/57.38
V2X-ViT	22.65	87.77/70.86	25.11	95.01/85.84	25.11	90.28/74.75	24.62	70.62/54.02
CoBEVT	22.65	85.57/69.99	25.11	92.48/71.16	25.11	88.16/73.08	24.62	74.10/60.23
ERMVP	22.65	94.35/84.76	25.11	95.99/89.14	25.11	93.08/81.91	24.62	74.73/60.75
MRCNet	22.04	75.40/64.49	24.49	96.95/90.66	24.49	95.41/84.32	24.01	73.10/57.48
Where2comm	22.65	88.45/80.54	25.11	95.10/88.48	25.11	90.68/80.48	24.62	76.70/61.96
	21.03	88.25/80.43	21.47	95.26/88.63	22.33	90.48/80.31	23.67	76.67/61.97
	17.51	84.96/77.72	18.38	94.79/87.80	18.91	89.25/79.13	20.62	75.94/61.52
	16.52	81.05/72.90	17.41	94.06/86.45	17.78	88.33/77.82	18.63	74.47/60.54
	15.33	76.83/68.04	15.96	93.08/84.40	16.35	86.47/75.18	17.11	70.97/58.18
	14.16	76.55/66.89	14.56	92.44/83.11	15.03	85.07/73.15	15.87	68.08/56.58
CoSDH	16.58	85.33/77.57	19.06	95.63/89.82	18.96	91.65/81.51	18.40	76.62/60.77
	16.19	84.60/73.82	17.24	92.00/82.46	18.45	91.03/78.65	17.36	76.68/60.69
	15.26	79.29/66.71	14.79	88.60/78.79	17.51	86.58/72.49	16.41	75.10/59.25
	14.39	74.98/62.75	13.49	87.90/77.88	14.20	82.58/68.33	14.16	69.92/56.41
CoLC	18.37	**95.14**/**87.89**	19.81	**96.88**/**92.93**	19.32	**95.97**/**89.81**	20.02	**76.71**/62.17
	17.38	93.47/85.28	18.81	96.46/91.95	18.31	95.05/87.72	19.01	76.03/62.09
	16.07	91.31/81.57	17.49	95.59/90.29	17.00	93.27/84.04	17.69	74.76/61.65
	15.35	88.89/79.28	16.65	94.93/88.77	16.12	91.78/81.43	16.71	74.08/61.19
	14.85	87.04/76.00	15.96	94.18/87.40	15.49	89.67/78.11	15.76	73.24/60.79
	14.41	83.44/70.47	14.99	92.80/84.55	14.64	86.22/72.40	13.85	71.19/59.60

Table 12. Robustness to heterogeneous scenarios. The first two rows present the detection accuracy (AP@0.3/0.5/0.7) of PointPillars and SECOND. The first row in the same method indicates results under **homogeneous** settings, where both the ego and neighbor agents use PointPillars.The second row shows performance under **heterogeneous** settings, where the ego agent uses PointPillars and neighbors use SECOND. We highlight cases where performance significantly degrades in the heterogeneous setting.

Methods	V2XSim	OPV2V	V2XSet	DAIR-V2X
PointPillars	74.73/73.72/61.65	77.65/74.42/54.52	78.41/74.18/57.43	67.15/64.32/53.27
SECOND	80.28/78.82/73.51	85.32/84.47/79.43	82.28/80.45/70.35	73.14/70.57/60.59
Late Fusion	96.23/86.87/78.78	94.22/92.51/77.82	91.81/89.05/75.72	74.12/62.05/42.82
	95.70/84.87/71.30 ↓	92.65/89.97/69.81 ↓	67.96/66.05/55.97 ↓	76.08/72.21/58.63
DiscoNet	80.73/77.46/64.03	95.45/93.45/82.95	89.63/86.28/71.03	80.32/73.21/57.38
	72.85/70.80/58.64 ↓	88.66/87.09/77.27 ↓	79.81/76.90/63.05 ↓	72.80/66.93/53.21 ↓
V2X-ViT	90.42/87.77/70.86	97.35/95.01/85.84	93.93/90.28/74.75	78.34/70.62/54.02
	74.63/71.16/59.54 ↓	90.31/89.07/79.01 ↓	85.36/83.18/68.64 ↓	69.38/64.61/49.27 ↓
CoBEVT	88.04/85.57/69.99	95.61/92.48/71.16	91.61/88.16/73.08	80.22/74.10/60.23
	67.96/66.05/55.97 ↓	85.73/83.89/70.95	82.14/78.69/52.35 ↓	73.44/68.85/58.11
CoLC	95.63/95.14/87.89	97.18/96.88/92.93	96.59/95.97/89.81	80.91/76.71/62.17
	95.63/95.14/87.89	97.18/96.88/92.93	96.59/95.97/89.81	80.91/76.71/62.17
CoLC*	94.09/93.47/85.28	96.94/96.46/91.95	95.87/95.05/87.72	80.37/76.03/62.09
	94.09/93.47/85.28	96.94/96.46/91.95	95.87/95.05/87.72	80.37/76.03/62.09

Table 13. Robustness to pose error on V2XSim.

Dataset	V2XSim					
Method/Metric	AP@0.5/0.7					
Noise Level $\sigma_t/\sigma_r\,(m/^\circ)$	0.0/0.0	0.2/0.2	0.4/0.4	0.6/0.6	0.8/0.8	1.0/1.0
No Fusion	73.72/61.65	73.72/61.65	73.72/61.65	73.72/61.65	73.72/61.65	73.72/61.65
Early Fusion	94.68/83.61	92.06/81.55	91.98/81.11	91.30/78.79	90.15/76.24	87.27/71.91
V2X-ViT	87.77/70.86	85.37/68.81	81.76/65.93	77.01/63.01	73.42/60.92	70.02/59.23
Where2comm	88.45/80.54	86.85/74.03	79.31/63.91	73.83/60.33	69.91/58.37	67.36/58.17
ERMVP	94.35/84.76	92.05/73.06	82.44/61.84	75.73/59.23	72.05/57.70	69.29/57.20
CoLC	**95.14/87.89**	**93.17/85.69**	**93.08/85.29**	**92.25/83.30**	**91.05/81.21**	**89.10/78.31**
CoLC*	93.47/85.28	91.95/83.40	91.77/82.92	91.16/82.21	89.85/79.87	88.26/76.64

Table 14. Robustness to pose error on OPV2V.

Dataset	OPV2V					
Method/Metric	AP@0.5/0.7					
Noise Level $\sigma_t/\sigma_r\,(m/^\circ)$	0.0/0.0	0.2/0.2	0.4/0.4	0.6/0.6	0.8/0.8	1.0/1.0
No Fusion	74.42/54.52	74.42/54.52	74.42/54.52	74.42/54.52	74.42/54.52	74.42/54.52
Early Fusion	96.13/90.69	96.00/89.13	95.83/88.13	95.45/86.01	94.28/83.25	92.61/79.43
V2X-ViT	95.01/85.84	94.65/85.42	93.61/84.59	92.15/83.51	90.27/81.63	88.52/80.29
Where2comm	95.10/88.48	93.59/79.80	89.46/71.47	85.74/67.72	83.73/66.67	82.60/66.22
ERMVP	95.99/89.14	94.90/85.07	91.64/80.30	88.60/77.24	86.32/75.77	84.10/74.55
CoLC	96.88/92.93	96.82/91.66	96.71/90.54	96.44/88.92	95.21/85.82	93.66/82.82
CoLC*	96.46/91.95	96.22/90.38	96.07/89.75	95.65/88.21	95.25/86.60	93.39/83.81

Table 15. Robustness to pose error on V2XSet.

Dataset	V2XSet					
Method/Metric	AP@0.5/0.7					
Noise Level $\sigma_t/\sigma_r\,(m/^\circ)$	0.0/0.0	0.2/0.2	0.4/0.4	0.6/0.6	0.8/0.8	1.0/1.0
No Fusion	74.18/57.43	74.18/57.43	74.18/57.43	74.18/57.43	74.18/57.43	74.18/57.43
Early Fusion	94.59/88.00	94.20/85.38	94.02/83.98	92.85/80.55	90.81/76.42	86.16/70.33
V2X-ViT	90.28/74.75	89.65/73.80	87.89/71.72	85.19/68.92	82.64/66.75	79.56/64.19
Where2comm	90.68/80.48	88.56/72.33	83.01/63.97	78.46/60.13	75.57/58.59	74.21/58.49
ERMVP	93.08/81.91	91.26/76.71	86.07/69.10	81.66/64.90	77.45/62.47	74.79/61.28
CoLC	95.97/89.81	95.30/87.30	94.95/85.75	94.08/82.68	91.59/77.89	89.13/74.13
CoLC*	95.05/87.72	94.24/84.87	94.12/83.81	93.21/81.51	91.66/78.44	88.90/74.92

Table 16. Robustness to pose error on DAIR-V2X.

Dataset	DAIR-V2X					
Method/Metric	AP@0.5/0.7					
Noise Level $\sigma_t/\sigma_r\,(m/^\circ)$	0.0/0.0	0.2/0.2	0.4/0.4	0.6/0.6	0.8/0.8	1.0/1.0
No Fusion	64.32/53.27	64.32/53.27	64.32/53.27	64.32/53.27	64.32/53.27	64.32/53.27
Early Fusion	76.51/63.83	68.80/54.26	65.26/52.07	63.21/51.01	61.61/50.48	60.42/50.04
V2X-ViT	70.62/54.02	70.15/53.88	68.93/53.41	67.70/53.00	66.73/52.55	66.06/52.42
Where2comm	76.77/61.96	74.75/58.65	70.62/56.09	68.46/55.55	67.51/55.01	66.63/54.75
ERMVP	74.73/60.75	73.44/58.79	70.95/57.32	69.39/56.77	68.78/56.65	68.11/56.40
CoLC	76.71/62.17	71.35/57.09	69.30/55.56	67.32/54.69	66.30/54.17	65.31/53.71
CoLC*	76.03/62.09	71.59/57.73	69.68/56.50	68.07/55.87	67.06/55.20	66.52/55.29

Table 17. Robustness to latency on V2XSim.

Dataset	V2XSim					
Method/Metric	AP@0.5/0.7					
Latency Level (ms)	0	100	200	300	400	500
No Fusion	73.72/61.65	73.72/61.65	73.72/61.65	73.72/61.65	73.72/61.65	73.72/61.65
Early Fusion	94.68/83.61	92.65/76.79	88.89/74.83	88.18/74.54	88.45/74.91	88.25/74.60
V2X-ViT	87.77/70.86	85.44/68.09	84.57/67.30	84.33/67.54	84.10/67.24	83.60/67.05
Where2comm	88.45/80.54	86.03/72.12	85.24/71.70	84.94/71.16	85.27/71.50	84.92/71.20
ERMVP	94.35/84.76	90.00/70.36	89.17/70.08	88.90/70.72	88.63/70.58	88.40/70.34
CoLC	**95.14**/**87.89**	91.67/81.50	89.08/79.79	88.60/79.54	88.59/79.69	88.22/79.66
CoLC*	93.47/85.28	90.32/77.34	88.31/75.64	87.88/76.27	88.01/76.12	87.67/76.22

Table 18. Robustness to latency on OPV2V.

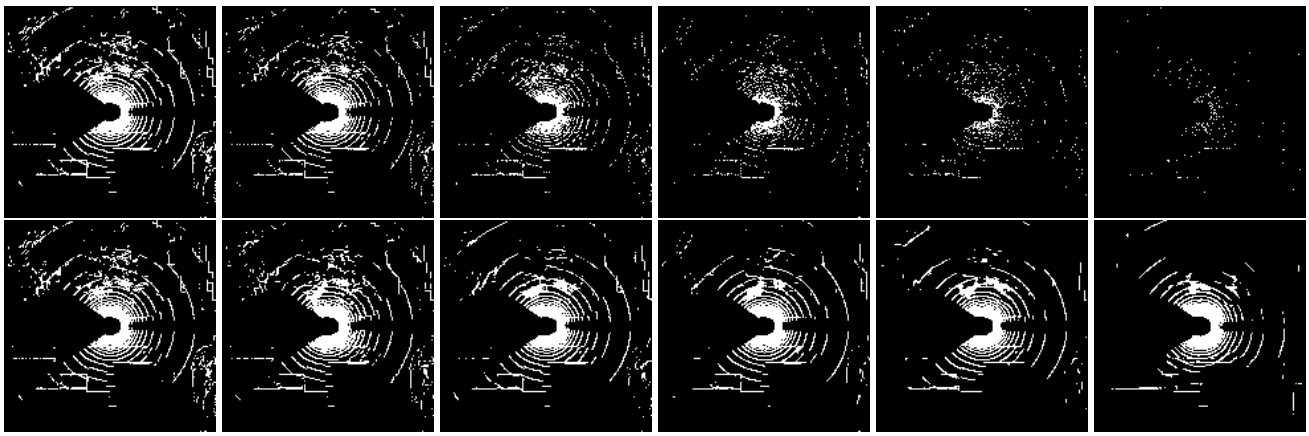
Dataset	OPV2V					
Method/Metric	AP@0.5/0.7					
Latency Level (ms)	0	100	200	300	400	500
No Fusion	74.42/54.52	74.42/54.52	74.42/54.52	74.42/54.52	74.42/54.52	74.42/54.52
Early Fusion	96.13/90.69	93.09/52.75	60.23/31.15	47.22/28.16	45.81/28.82	45.63/30.26
V2X-ViT	95.01/85.84	93.96/79.91	88.56/74.75	83.68/72.04	80.23/69.75	77.65/68.57
Where2comm	95.10/88.48	91.75/66.98	81.30/58.77	77.04/60.19	76.65/61.66	76.57/62.24
ERMVP	95.99/89.14	81.53/52.79	75.67/51.11	72.54/50.61	72.07/50.58	71.69/50.37
CoLC	96.88/92.93	94.45/65.37	69.52/42.68	57.35/39.77	54.21/39.21	52.92/39.89
CoLC*	96.46/91.95	93.99/66.48	71.77/46.03	61.46/44.27	58.44/44.06	56.98/44.32

Table 19. Robustness to latency on V2XSet.

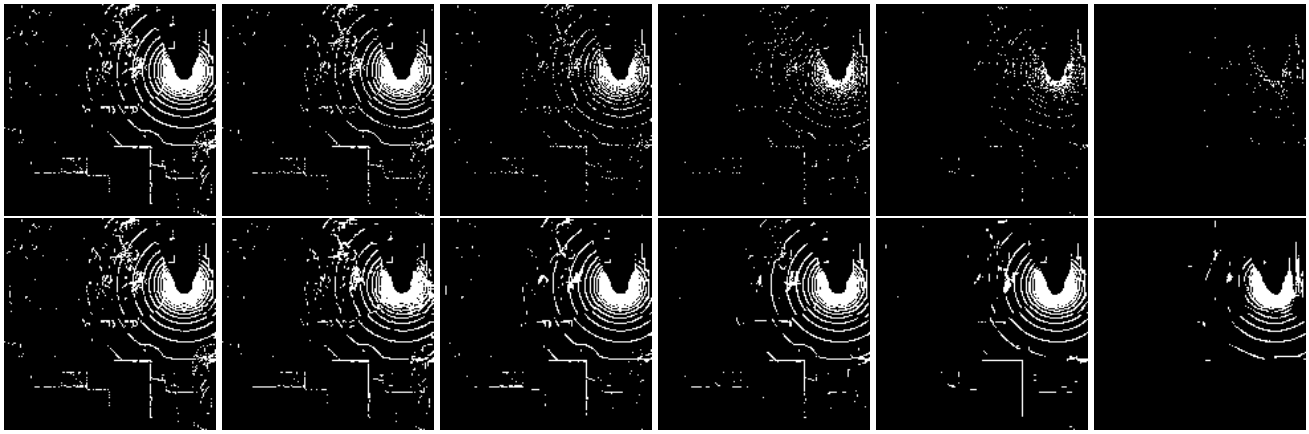
Dataset	V2XSet					
Method/Metric	AP@0.5/0.7					
Latency Level (ms)	0	100	200	300	400	500
No Fusion	74.18/57.43	74.18/57.43	74.18/57.43	74.18/57.43	74.18/57.43	74.18/57.43
Early Fusion	94.59/88.00	92.30/63.01	70.97/42.96	56.59/38.76	53.56/38.52	52.74/39.20
V2X-ViT	90.28/74.75	88.80/68.52	83.90/61.76	78.09/59.34	74.85/57.73	73.09/57.11
Where2comm	90.68/80.48	87.17/63.63	79.60/55.78	74.65/56.84	74.09/57.68	74.12/58.38
ERMVP	93.08/81.91	83.97/64.14	77.91/56.66	72.44/54.67	70.93/54.05	69.88/53.57
CoLC	95.97/89.81	93.42/67.67	76.29/48.26	62.30/43.54	57.61/42.56	56.17/42.71
CoLC*	95.05/87.72	92.20/67.55	77.64/49.90	65.46/46.01	61.07/44.98	59.88/45.60

Table 20. Robustness to latency on DAIR-V2X.

Dataset	DAIR-V2X					
Method/Metric	AP@0.5/0.7					
Latency Level (ms)	0	100	200	300	400	500
No Fusion	64.32/53.27	64.32/53.27	64.32/53.27	64.32/53.27	64.32/53.27	64.32/53.27
Early Fusion	76.51/63.83	76.50/63.87	76.49/63.87	76.52/63.87	76.50/63.87	76.51/63.91
V2X-ViT	70.62/54.02	70.06/53.90	70.05/53.77	70.08/53.80	70.02/53.80	70.20/53.90
Where2comm	76.77/61.96	74.62/58.35	74.89/58.65	74.80/58.61	74.75/58.54	74.67/58.37
ERMVP	74.73/60.75	67.88/48.02	67.92/48.00	67.91/47.93	67.90/47.96	67.87/47.90
CoLC	76.71/62.17	71.64/57.20	71.49/56.98	71.49/57.11	71.63/57.03	71.44/57.13
CoLC*	76.03/62.09	72.79/58.21	71.61/57.34	71.51/57.15	71.62/57.29	71.56/56.96

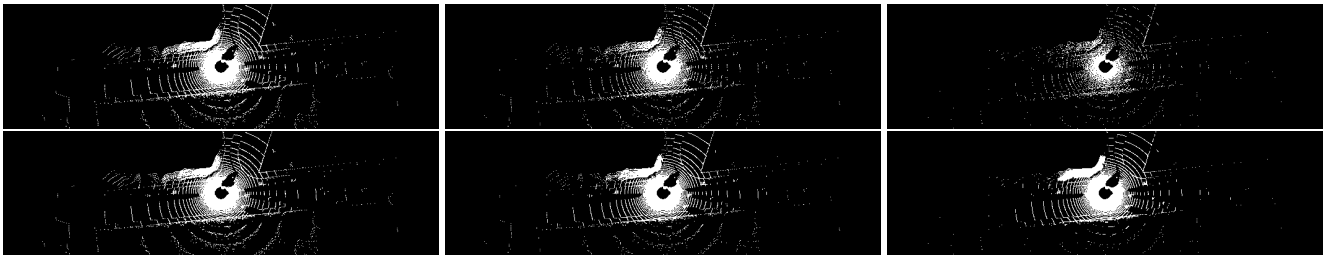


(1-a) 100% LiDAR (1-b) $R^{bg} = 0.5$ (1-c) $R^{bg} = 0.2$ (1-d) $R^{bg} = 0.1$ (1-e) $R^{bg} = 0.05$ (1-f) $R^{bg} = 0.01$

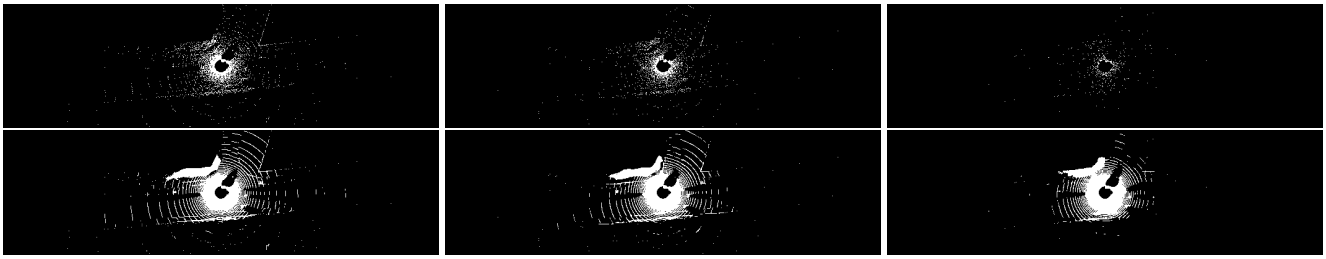


(2-a) 100% LiDAR (2-b) $R^{bg} = 0.5$ (2-c) $R^{bg} = 0.2$ (2-d) $R^{bg} = 0.1$ (2-e) $R^{bg} = 0.05$ (2-f) $R^{bg} = 0.01$

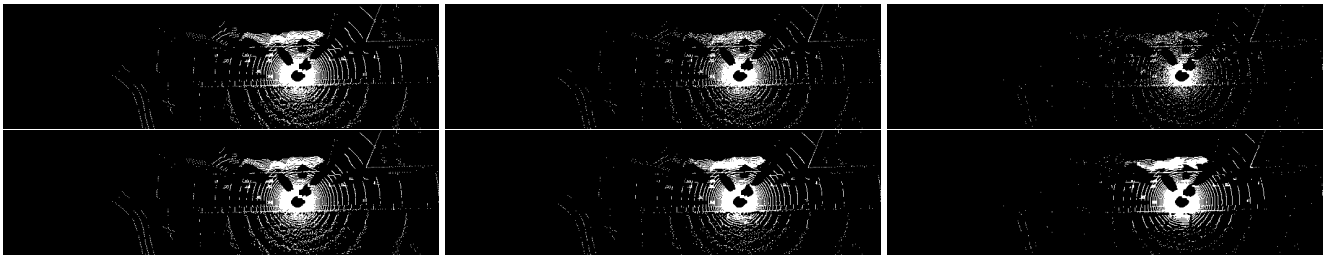
Figure 3. Pillar completion results on V2XSim. The foreground sampling ratio R^{fg}=0.2. The top row shows the input sparse pillars under varying sampling ratios. Bottom row displays the corresponding reconstructed pillars produced by the VQ-based LiDAR completion module. 1 and 2 are different agents.



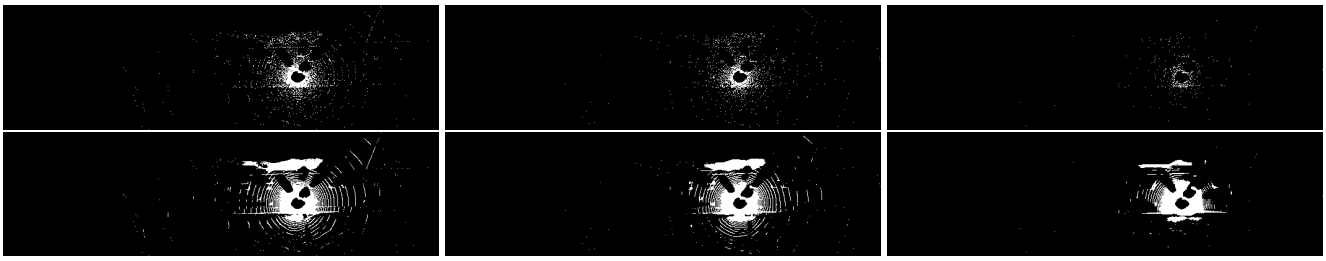
(1-a) 100% LiDAR (1-b) $R^{bg} = 0.5$ (1-c) $R^{bg} = 0.2$



(1-d) $R^{bg} = 0.1$ (1-e) $R^{bg} = 0.05$ (1-f) $R^{bg} = 0.01$

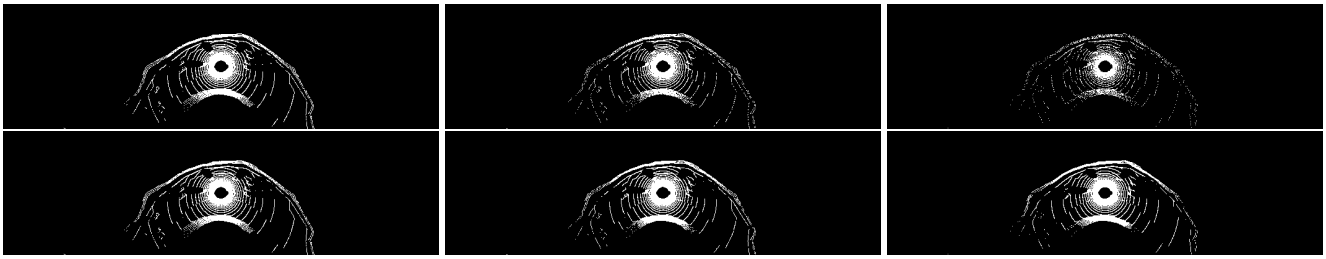


(2-a) 100% LiDAR (2-b) $R^{bg} = 0.5$ (2-c) $R^{bg} = 0.2$

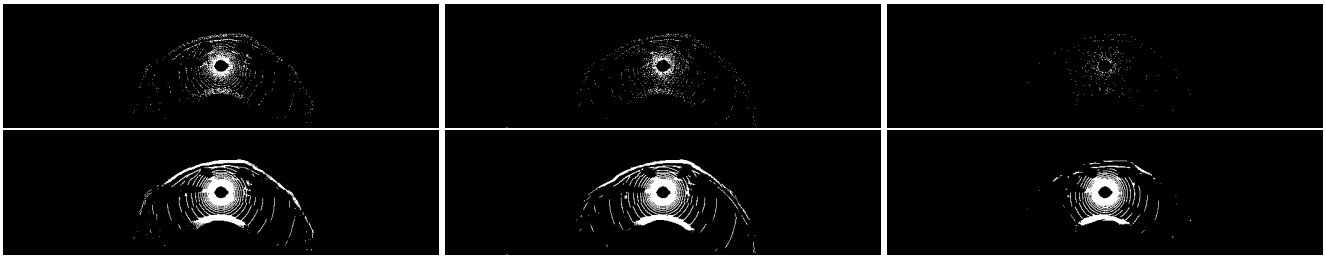


(2-d) $R^{bg} = 0.1$ (2-e) $R^{bg} = 0.05$ (2-f) $R^{bg} = 0.01$

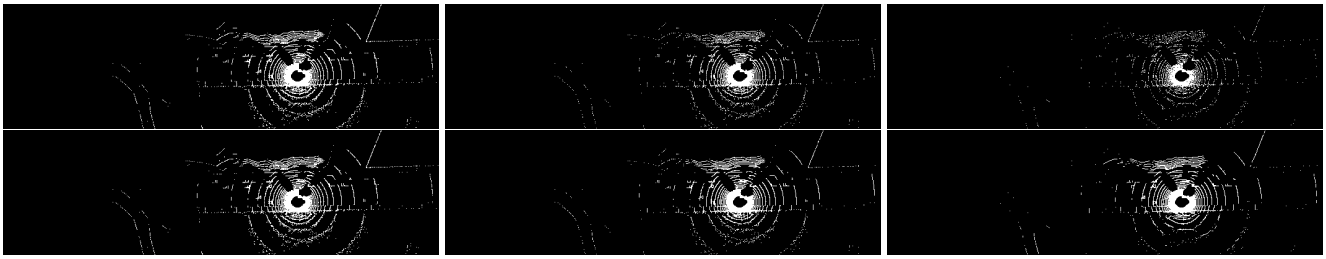
Figure 4. Pillar completion results on OPV2V. The foreground sampling ratio R^{fg}=0.2. The top row shows the input sparse pillars under varying sampling ratios. Bottom row displays the corresponding reconstructed pillars produced by the VQ-based LiDAR completion module. 1 and 2 are different agents.



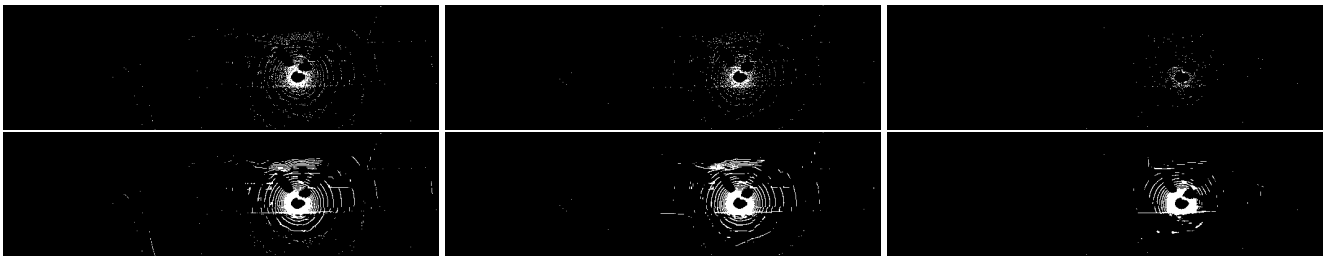
(1-a) 100% LiDAR (1-b) $R^{bg} = 0.5$ (1-c) $R^{bg} = 0.2$



(1-d) $R^{bg} = 0.1$ (1-e) $R^{bg} = 0.05$ (1-f) $R^{bg} = 0.01$

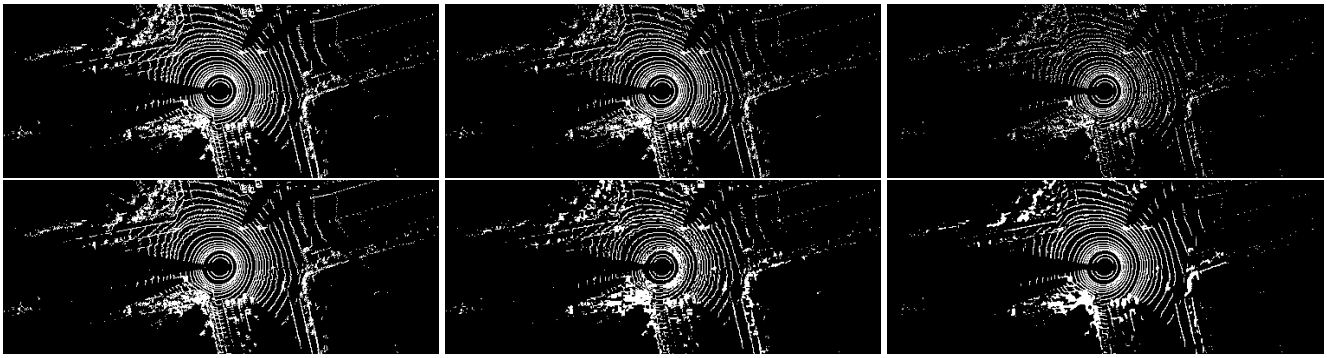


(2-a) 100% LiDAR (2-b) $R^{bg} = 0.5$ (2-c) $R^{bg} = 0.2$

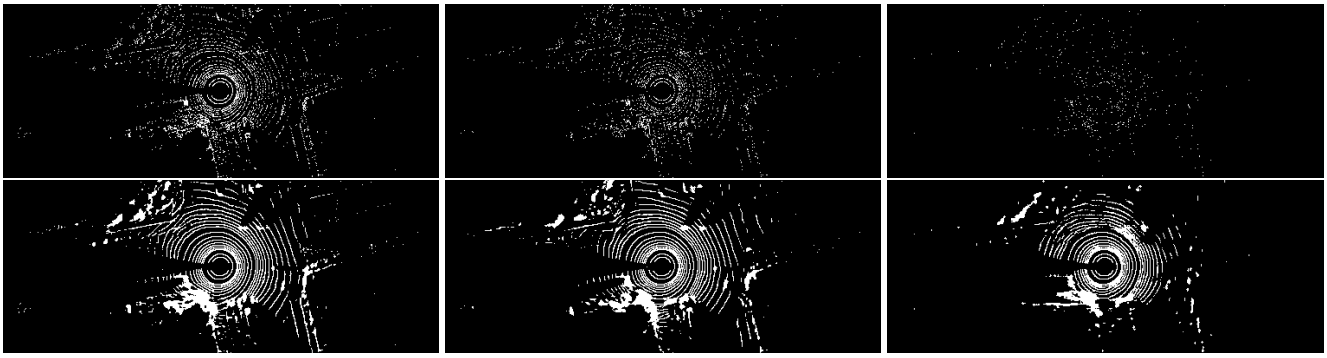


(2-d) $R^{bg} = 0.1$ (2-e) $R^{bg} = 0.05$ (2-f) $R^{bg} = 0.01$

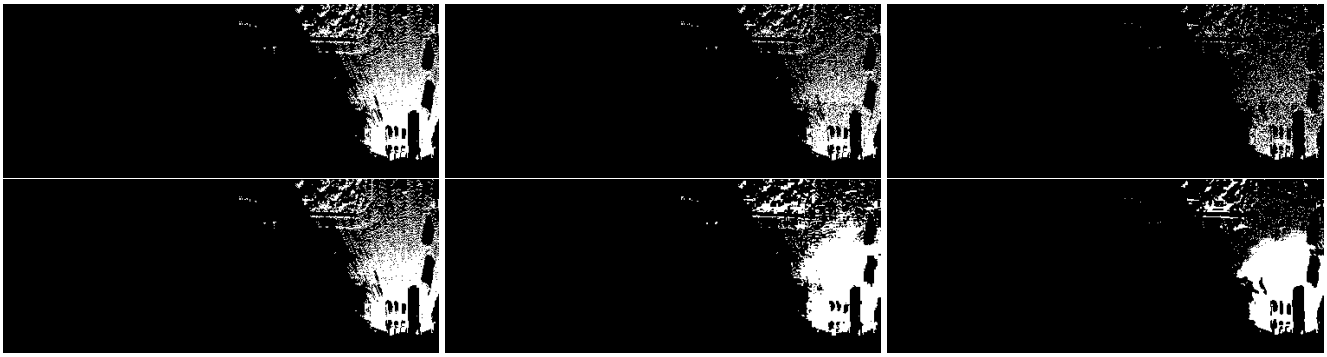
Figure 5. Pillar completion results on V2XSet. The foreground sampling ratio R^{fg}=0.2. The top row shows the input sparse pillars under varying sampling ratios. Bottom row displays the corresponding reconstructed pillars produced by the VQ-based LiDAR completion module. 1 and 2 are different agents.



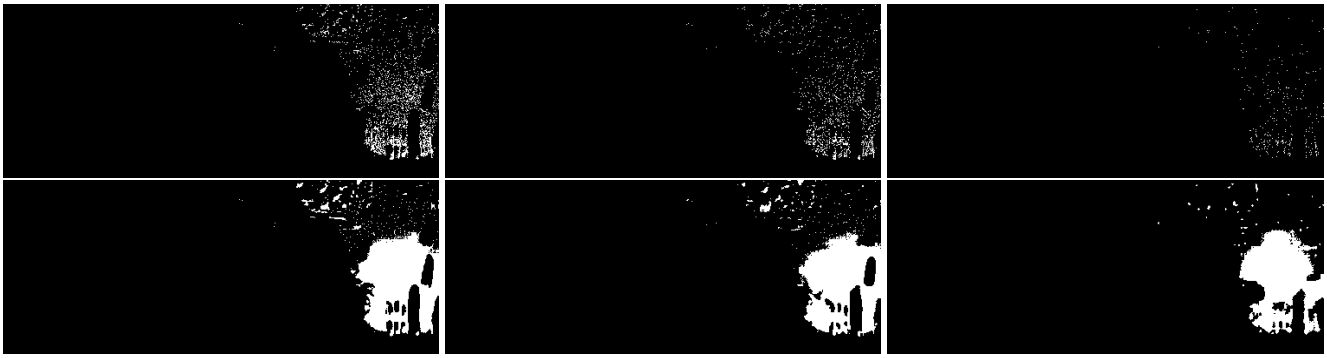
(1-a) 100% LiDAR (1-b) $R^{bg} = 0.5$ (1-c) $R^{bg} = 0.2$



(1-d) $R^{bg} = 0.1$ (1-e) $R^{bg} = 0.05$ (1-f) $R^{bg} = 0.01$

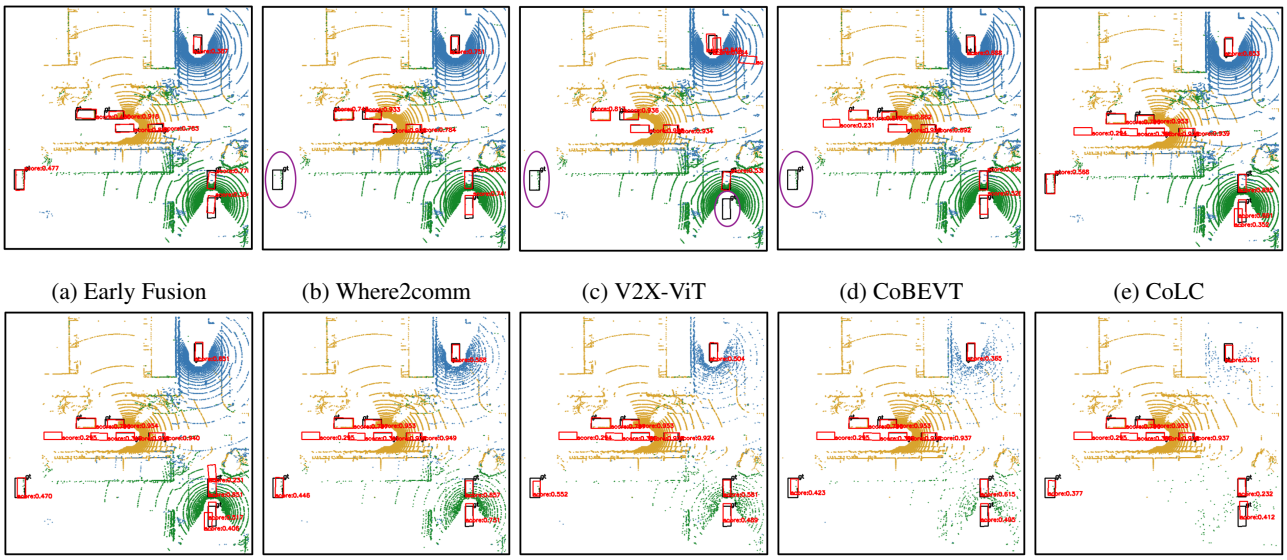


(2-a) 100% LiDAR (2-b) $R^{bg} = 0.5$ (2-c) $R^{bg} = 0.2$



(2-d) $R^{bg} = 0.1$ (2-e) $R^{bg} = 0.05$ (2-f) $R^{bg} = 0.01$

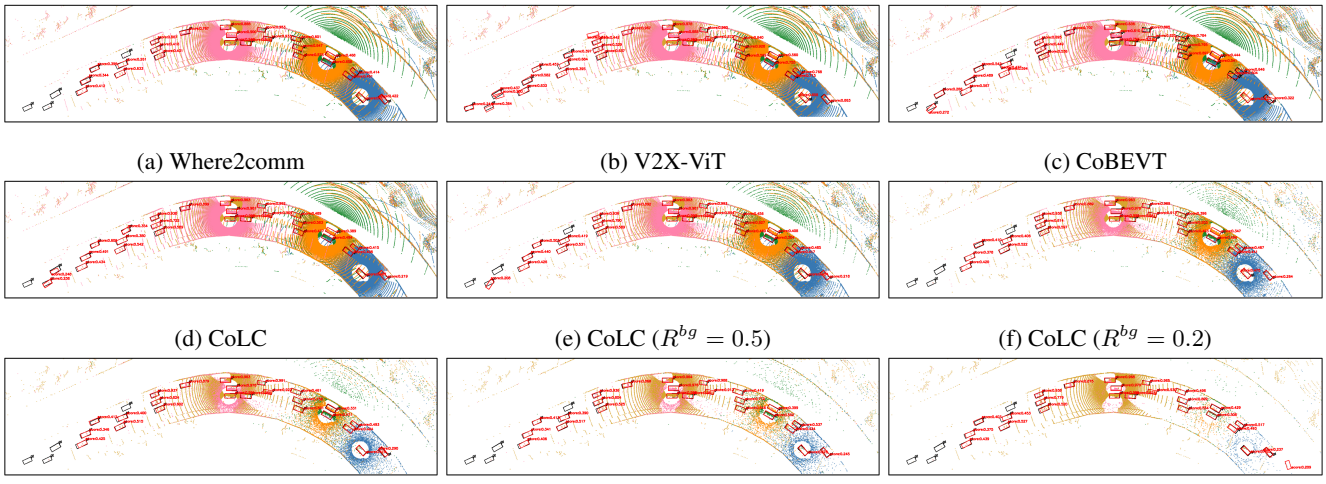
Figure 6. Pillar completion results on DAIR-V2X. The foreground sampling ratio R^{fg}=0.2. The top row shows the input sparse pillars under varying sampling ratios. Bottom row displays the corresponding reconstructed pillars produced by the VQ-based LiDAR completion module. 1 and 2 are different agents.



(a) Early Fusion (b) Where2comm (c) V2X-ViT (d) CoBEVT (e) CoLC

(f) CoLC ($R^{bg} = 0.5$) (g) CoLC ($R^{bg} = 0.2$) (h) CoLC ($R^{bg} = 0.1$) (i) CoLC ($R^{bg} = 0.05$) (j) CoLC ($R^{bg} = 0.01$)

Figure 7. Qualitative comparison of Detection on V2XSim. In (f)-(j), the foreground sampling ratio R^{fg} is fixed at 0.2. Black and red boxes represent ground truth and model predictions, respectively.

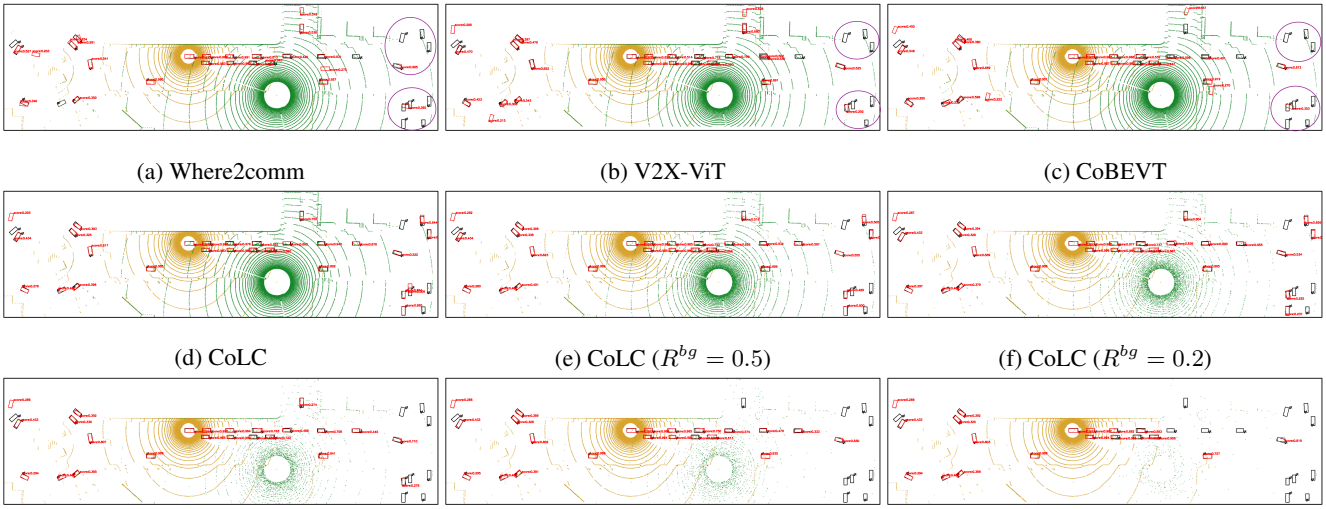


(a) Where2comm (b) V2X-ViT (c) CoBEVT

(d) CoLC (e) CoLC ($R^{bg} = 0.5$) (f) CoLC ($R^{bg} = 0.2$)

(g) CoLC ($R^{bg} = 0.1$) (h) CoLC ($R^{bg} = 0.05$) (i) CoLC ($R^{bg} = 0.01$)

Figure 8. Qualitative comparison of Detection on OPV2V. In (e)-(i), the foreground sampling ratio R^{fg} is fixed at 0.2. Black and red boxes represent ground truth and model predictions, respectively.

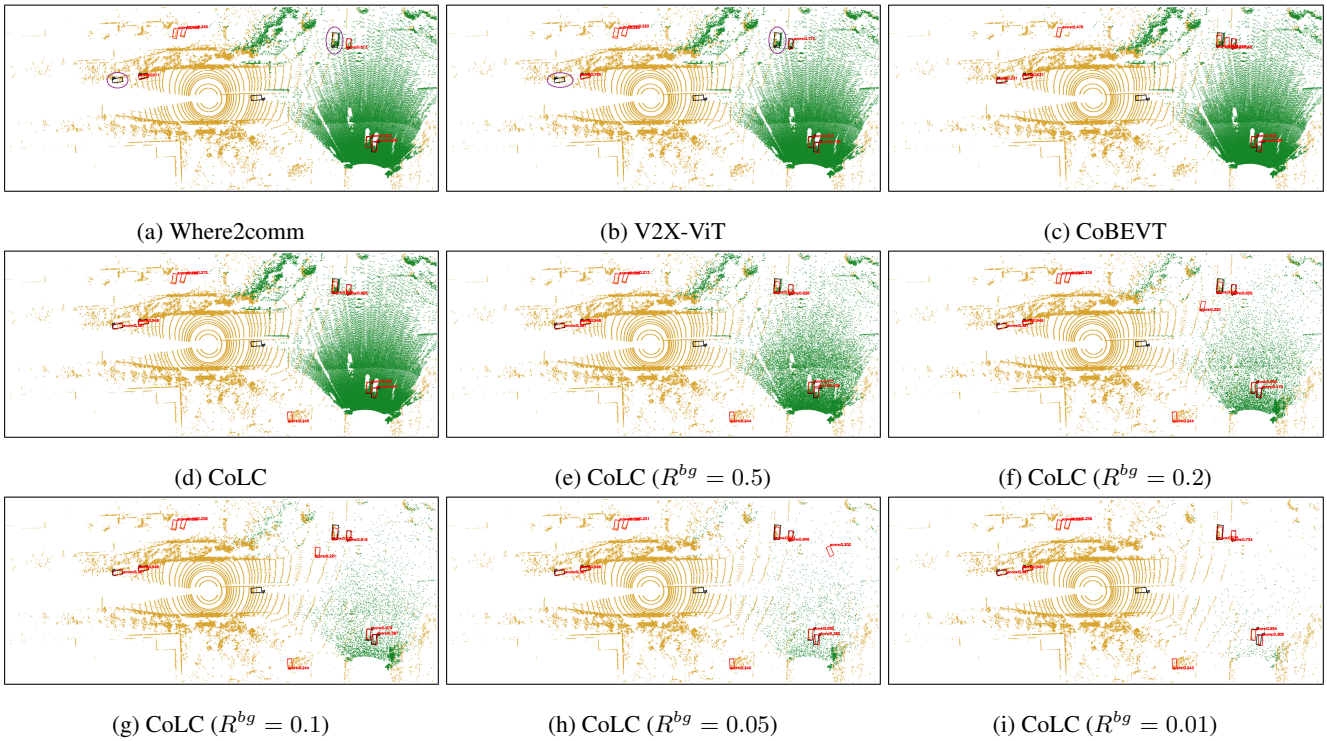


(a) Where2comm (b) V2X-ViT (c) CoBEVT

(d) CoLC (e) CoLC ($R^{bg} = 0.5$) (f) CoLC ($R^{bg} = 0.2$)

(g) CoLC ($R^{bg} = 0.1$) (h) CoLC ($R^{bg} = 0.05$) (i) CoLC ($R^{bg} = 0.01$)

Figure 9. Qualitative comparison of Detection on V2XSet. In (e)-(i), the foreground sampling ratio R^{fg} is fixed at 0.2. Black and red boxes represent ground truth and model predictions, respectively.



(a) Where2comm (b) V2X-ViT (c) CoBEVT

(d) CoLC (e) CoLC ($R^{bg} = 0.5$) (f) CoLC ($R^{bg} = 0.2$)

(g) CoLC ($R^{bg} = 0.1$) (h) CoLC ($R^{bg} = 0.05$) (i) CoLC ($R^{bg} = 0.01$)

Figure 10. Qualitative comparison of Detection on DAIR-V2X. In (e)-(i), the foreground sampling ratio R^{fg} is fixed at 0.2. Black and red boxes represent ground truth and model predictions, respectively.