From Discriminative to Generative: A Diffusion-Based Paradigm for Multi-Agent Collaborative Perception

Kexin Gong[1*], Puyi Yao[1*], Guiyang Luo[1†], Quan Yuan[1], Tiange Fu[1], Hui Zhang[2†], Jinglin Li[1]

[1]School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China
[2]School of Computer Science, Beijing Jiaotong University, Beijing, China
surooprise@gmail.com, {yaopuyi, luoguiyang, yuanquan, tgfu}@bupt.edu.cn, huizhang1@bjtu.edu.cn, jlli@bupt.edu.cn

Abstract

Collaborative perception leveraging intermediate feature fusion has emerged as a leading paradigm to significantly enhance the environmental perception capabilities of autonomous driving systems. However, existing methods typically rely on discriminative supervision guided by downstream tasks. This paradigm compels models to learn minimal, task-specific representations, which conflicts with the goal of cooperative perception to capture comprehensive information, thereby limiting generalization. To address this issue, we propose DiGS-CP, a novel two-stage generative supervised collaborative perception framework. Specifically, we introduce a diffusion-based generative task that conditions on fused object-level features to generate representations of object-level point clouds. The proposed generative supervision provides fine-grained, task-agnostic signals that encourages the fusion module to learn comprehensive representations beyond task-specific requirements. By preserving and integrating complementary information from collaborative agents, our approach overcomes the limitations of task-specific learning and enhances the generalizability of the learned features. Furthermore, our two-stage architecture requires agents to transmit only object-level features, significantly reducing communication overhead. Extensive experiments on three benchmark datasets demonstrate that DiGS-CP achieves state-of-the-art performance in 3D object detection, while maintaining low bandwidth requirements and exhibiting excellent generalization ability.

Code — https://github.com/gkx7w/DiGS-CP

Introduction

Single-agent autonomous perception is limited by occlusion and narrow field-of-view, which significantly compromise driving safety (Fang et al. 2024; Hu et al. 2025). Collaborative perception addresses these challenges by enabling information sharing among multiple agents (Su et al. 2024; Zhang et al. 2025). Depending on when information is integrated, strategies are categorized into early, intermediate, and late fusion. Early fusion (Chen et al. 2019b; Hu et al. 2022a) achieves high accuracy but has high communication overhead, making it impractical for real-world use. Late fusion (Rawashdeh and Wang 2018) reduces overhead by sharing only decision-level results, but suffers from lower performance due to limited context. Intermediate fusion, balancing bandwidth efficiency and detection performance, has become the dominant approach in recent studies (Hu et al. 2022b; Li et al. 2021; Xu et al. 2022a).

Current mainstream intermediate fusion methods employ an end-to-end training paradigm, where the fusion module is supervised by losses from the downstream detection task. This approach aligns with the concept of discriminative supervision. As defined by Ng and Jordan (Ng and Jordan 2001), discriminative models learn a direct mapping from inputs to task-specific outputs. Existing cooperative fusion methods exemplify this paradigm by training networks to minimize detection errors. Previous studies (Borse et al. 2023; Su et al. 2024) further demonstrate that discriminative approaches operate on a principle of feature selectivity, retaining only task-specific features while filtering out those lacking discriminative value for the target subset of data. Consequently, this task-specific optimization often results in poor feature generalization across different scenarios or datasets.

These characteristics present particular issues in multi-agent collaborative perception. Effective collaboration necessitates integrating complementary information from diverse viewpoints, requiring fused features to preserve rich semantic and spatial information beyond individual agent capabilities. Conversely, discriminative supervision promotes learning minimal feature sets optimized for specific tasks, inevitably sacrificing complementary information crucial for collaboration. Recognizing this conflict, prior works have introduced auxiliary tasks such as semantic segmentation or scene reconstruction (Xu et al. 2022a; Wang et al. 2023a) to enrich feature representations, but these approaches remain within the discriminative paradigm with inherently task-specific supervision signals.

To address the inherent limitations of discriminative fusion, we propose DiGS-CP, a **Di**ffusion-based **G**enerative **S**upervision for **C**ollaborative **P**erception framework. Unlike discriminative methods that learn direct mappings from inputs to task-specific outputs, our approach uses diffusion-based generative modeling as supervision to guide feature fusion. This approach requires the fusion module to preserve

[*]These authors contributed equally.
[†]Corresponding authors.

sufficient information for data generation, thereby encouraging comprehensive and generalizable feature representations essential for effective collaboration.

Our framework introduces two key innovations. First, we design a two-stage object-level feature transmission mechanism. The first stage produces initial object proposals and obtains corresponding object-level feature representations from the generated features. This design allows agents to exchange compact object-level features instead of large scene-level maps, significantly reducing communication overhead. The second stage clusters and fuses the transmitted object-level features using IoU-based matching with Distance-Weighted Fusion (DWF), which assigns higher fusion weights to closer objects due to their greater reliability.

Second, leveraging the object-level fused features from the first innovation, we introduce a generative supervision mechanism to guide the fusion process. The object-level design enables the diffusion model to focus on learning fine-grained structural representations for individual objects rather than global scene patterns. Fused object-level features are used as conditioning inputs to a diffusion model, which generates corresponding point cloud statistical feature maps for each object. The diffusion loss is backpropagated to the upstream fusion module, linking fusion quality to generation performance and providing task-agnostic supervision that promotes learning of generalizable feature representations. The diffusion model is used exclusively during training and discarded at inference.

We conduct experiments on three public benchmarks to validate our approach. Results show that our method outperforms existing state-of-the-art approaches in both 3D object detection accuracy and communication efficiency.

The main contributions of this work are as follows:

- **We propose a novel generative supervision for collaborative perception framework.** Our approach employs a conditional diffusion model to provide fine-grained, task-agnostic supervision, promoting rich feature representations and overcoming the generalization limitations of discriminative supervision methods.
- **We design an efficient two-stage object-level feature transmission and fusion architecture.** Our approach extracts and transmits compact object-level features instead of large scene-level maps, reducing communication overhead, and introduces Distance-Weighted Fusion (DWF) for reliable feature integration by prioritizing closer detections.
- **We achieve state-of-the-art performance across multiple benchmarks.** Comprehensive experiments demonstrate that our method outperforms existing discriminative fusion approaches on three mainstream collaborative perception datasets while maintaining superior communication efficiency.

Related work

Collaborative Perception

Collaborative perception (Liu et al. 2020; Yu et al. 2025; Wang et al. 2023b; Lu et al. 2023; Xu et al. 2025a; Xiang, Xu, and Ma 2023; Luo et al. 2022; Shao et al. 2025; Luo et al. 2023) addresses the inherent limitations of single-agent perception through inter-agent information sharing. Intermediate fusion has emerged as the dominant paradigm due to its ability to balance detection accuracy and communication efficiency, garnering significant research interest. The supervision strategies for these intermediate fusion methods can be broadly categorized into two main approaches. The most common is single-task supervision, where the entire networks is trained end-to-end using only the final 3D detection loss (Chen et al. 2019a; Wang et al. 2020; Huang et al. 2025). To enrich learned representations, some another works explore multi-task supervision, which introduces auxiliary tasks such as semantic segmentation or motion prediction to be jointly optimized with the primary detection objective (Song et al. 2024; Wang et al. 2024). Whether employing single-task or multi-task learning, the training process is guided by loss functions that focus on fitting decision boundaries for specific predictive objectives rather than modeling the broader data distribution. This task-specific focus inherently leads to minimal feature representations that limit generalization capabilities across diverse scenarios.

Diffusion Models

Diffusion models (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020) are generative models based on chained stochastic processes. The forward of this chain adds Gaussian noise to data, called diffusion process. Reverse of this chain leverages U-Net architectures (Ronneberger, Fischer, and Brox 2015) to learn progressive denoising by optimizing a reweighted variational lower-bound (Dhariwal and Nichol 2021). Conditional variants (Zhang, Rao, and Agrawala 2023; Ye et al. 2023; Ho and Salimans 2022) inject guidance signals, such as class labels, text, or feature embeddings, to guide the generative process, thereby expanding their application scope. Due to their superior generation quality and training stability, diffusion models have achieved remarkable success across diverse computer vision tasks, including image synthesis (Song, Meng, and Ermon 2020), super-resolution (Shang et al. 2024; Rombach et al. 2022), and image editing (Xu et al. 2025b). These applications demonstrate that diffusion models not only generate high-quality visual content but also effectively learn latent data representations. However, their potential for guiding feature learning in collaborative perception remains unexplored.

Method

Overall Architecture

As illustrated in Figure 1, our framework addresses collaborative perception through a novel two-stage generative supervision approach. The first stage maintains standard single-agent detection to extract bird's-eye-view (BEV) features and generate initial proposals.

The second stage introduces key innovations by performing object-level feature fusion under generative supervision. To handle overlapping detections from multiple agents, we use IoU-based matching followed by Distance-Weighted Fusion (DWF). In LiDAR systems, point cloud density de-

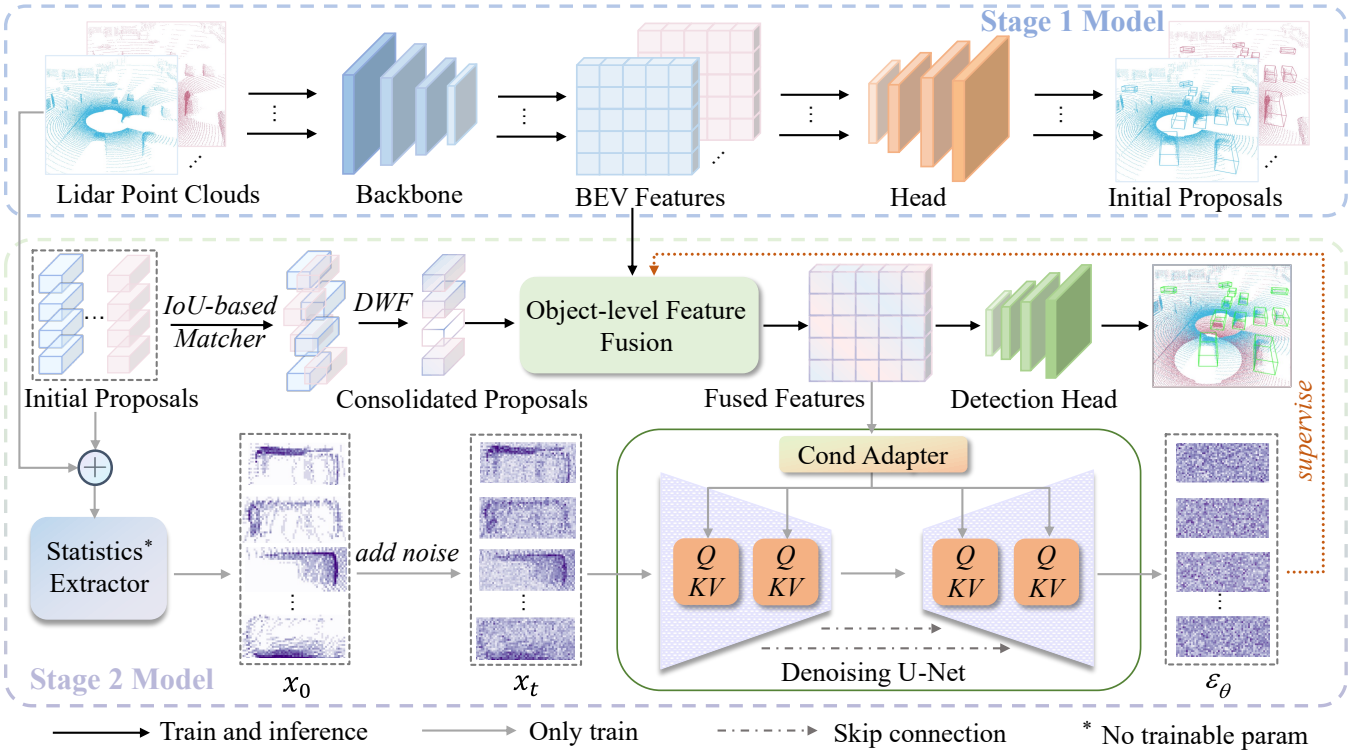


Figure 1: Overview of our two-stage generative supervision framework. Stage 1: Individual agents employ standard detectors to generate initial object proposals from their respective sensor data. Stage 2: Object proposals are consolidated through IoU-based matching followed by DWF, and corresponding object-level features are fused and input to a conditional denoising U-Net that provides generative supervision for the fusion module during training and is removed at inference.

creases with distance, leading to sparse and noisy measurements for distant objects that reduce detection reliability. Our DWF strategy addresses this by fusing both bounding boxes and object-level features from the first stage using distance-based exponential weights that prioritize closer detections. This improves fusion accuracy and localization compared to confidence-only methods. The fused object-level features from the fusion module are then processed using a generative supervision mechanism, where a conditional diffusion model generates point cloud statistical features. Unlike discriminative supervision, which targets task-specific decision boundaries, our generative approach encourages the fusion module to capture broader geometric relationships, resulting in more comprehensive feature representations and improved generalization capabilities.

During inference, the first stage remains unchanged while the diffusion model branch is entirely removed from the second stage. The pre-trained and information-rich features from the fusion module are then directly fed into the detection head for result refinement.

Two-Stage Cooperative Perception Pipeline

Stage 1: Single-Agent Object Detection We consider a collaborative perception scenario with N agents equipped with LiDARs. For each agent $i \in \{1, \ldots, N\}$, the LiDAR sensor captures raw point cloud data $\mathcal{P}_i$, which is processed through a standard 3D detection backbone to extract BEV features $\mathbf{F}_i = \text{Backbone}(\mathcal{P}_i)$. The feature map is subsequently fed into a lightweight detection head to generate preliminary object predictions. Each prediction result $\mathbf{d}_i$ consists of eight elements:

$$\mathbf{d}_i = (x_i, y_i, z_i, l_i, w_i, h_i, \theta_i, s_i), \tag{1}$$

where (x_i, y_i, z_i) represents the bounding box center coordinates, (l_i, w_i, h_i) denote the length, width, and height respectively, θ_i is the rotation angle, and s_i is the confidence score. We apply confidence thresholding with τ to filter predictions, retaining only detections satisfying $s_i > \tau$, followed by Non-Maximum Suppression (NMS) to eliminate redundant detections. After these operations, each agent i produces a detection set $\mathcal{R}_i = \{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_m\}$ containing m objects. These detection results $\mathcal{R}_i$ and the object-level BEV features $\mathbf{f}$ extracted using the detection boxes serve as inputs to the second-stage collaborative fusion process.

Stage 2: Object-level Feature Fusion Since different agents may detect overlapping objects, we use an IoU-based matcher to cluster all first-stage detections from multiple agents and obtain unified scene-level boxes. Detection results are first projected into a common coordinate system. The IoU-based matcher applies a greedy clustering strategy:

detection boxes with IoU exceeding a threshold are considered the same object and merged into cluster. This results in M clusters, each containing detections of the same physical object observed from different agent viewpoints.

Within each cluster C_j, we apply Distance-Weighted Fusion (DWF) to account for varying detection reliability. The fusion weights are computed as:

$$w_k = s_k \cdot \exp(-\alpha \cdot d_k), \qquad w_k^* = \frac{w_k}{\sum_{i=1}^{n} w_i}. \quad (2)$$

where d_k is the distance from the originating agent's sensor to the center of detection k, s_k is the confidence score, and $\alpha > 0$ is a decay factor that penalizes distant detections.

The fused bounding boxes (including position, size, and rotation angles) are obtained through weighted averaging of all detections in the cluster based on the normalized fusion weights w_k^*. These fused detection results $\tilde{\mathbf{d}}_j$ constitute the unified scene detection list $\tilde{\mathcal{R}} = \{\tilde{\mathbf{d}}_1, \ldots, \tilde{\mathbf{d}}_M\}$. For each detection, the corresponding object-level BEV features transmitted from the first stage are fused using the same DWF weights through weighted aggregation. The Object-level Feature Fusion module follows the approach from BEVFusion (Liu et al. 2023) to produce the final fused representations. The fusion process can be expressed as:

$$\bar{\mathbf{f}} = g(\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n; \phi), \quad (3)$$

where $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n$ are the object-level BEV features after DWF-based weighted aggregation, $g(\cdot; \phi)$ denotes the fusion module parameterized by ϕ, and $\bar{\mathbf{f}}$ is the fused feature representation from multiple agents.

Generative Supervision via Conditional Diffusion

Object-level Point Statistics Extraction To enable effective supervision of the fusion process, we utilize point cloud statistical features that accurately capture the 3D geometric structure of objects. Using the unified scene detection list $\tilde{\mathcal{R}}$ obtained from the IoU-based matcher and DWF, we trace each unified object $\tilde{\mathbf{d}}_j$ back to its corresponding original detection box cluster C_j. For each cluster, we aggregate the point clouds enclosed by all boxes in cluster across their respective scenes to construct the complete object-level point set: $\mathcal{P}_j = \bigcup_{d_k \in C_j} \mathcal{P}(d_k)$.

We perform voxelization on $\mathcal{P}_j$ and compute four types of statistical features for each non-empty voxel v containing points $\{p_1, p_2, \ldots, p_n\}$ with their centroid $\bar{p}_v$: (1) element-wise mean absolute deviation $\frac{1}{n}\sum_{i=1}^{n} |p_i - \bar{p}_v|$, (2) element-wise second central moment $\frac{1}{n}\sum_{i=1}^{n}(p_i - \bar{p}_v)^2$, (3) point count n, and (4) maximum pairwise distance $\max_{i,j} \|p_i - p_j\|_2$. The first two features are 3-dimensional (xyz), while the latter are scalars, resulting in an 8-channel feature representation per voxel. These features form a compact yet information-rich statistical feature map that serves as the ground truth $\mathbf{x}_0$ for the diffusion model.

Conditional Denoising Diffusion Model Our generative supervision is implemented via a conditional denoising diffusion model. The framework comprises a predetermined forward noising process and a learnable noise prediction process.

In the forward process, Gaussian noise is progressively added to each sample $\mathbf{x}_0$ from the statistical feature map distribution $q(\mathbf{x}_0)$ via a predefined Markov chain with T timesteps. At any timestep $t \in \{1, \ldots, T\}$, the noisy sample $\mathbf{x}_t$ is generated according to:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}), \quad (4)$$

where $\{\beta_t\}_{t=1}^{T}$ represents the noise variance parameters controlling the noise magnitude at each step. For computational efficiency, we employ the reparameterization trick to directly sample $\mathbf{x}_t$ at any timestep t from $\mathbf{x}_0$:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad (5)$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, and $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ represents standard Gaussian noise.

In the noise prediction process, our goal is to predict the noise ϵ added during the forward process, which generates noisy data $\mathbf{x}_t$ from the original data $\mathbf{x}_0$. This is done using a noise prediction network $\epsilon_\theta(\mathbf{x}_t, t, \bar{\mathbf{f}})$, where the conditioning signal $\bar{\mathbf{f}}$ is generated by the object-level feature fusion module as defined in Eq. (3), abbreviated here as $\bar{\mathbf{f}} = g(\phi)$.

This network adopts a U-Net-based architecture that injects the condition $\bar{\mathbf{f}}$ through cross-attention layers across multiple resolution levels, including downsampling, bottleneck, and upsampling modules. In each cross-attention layer, the query matrix $\mathbf{Q}$ is derived from the intermediate feature representations of the noisy input $\mathbf{x}_t$, while the key matrix $\mathbf{K}$ and value matrix $\mathbf{V}$ are obtained by linearly projecting the conditional feature vector $\bar{\mathbf{f}}$ using a conditional adapter. The cross-attention mechanism is computed as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_{\text{key}}}}\right)\mathbf{V}, \quad (6)$$

where d_{key} is the dimension of the key vectors. This enables the U-Net to dynamically attend to relevant parts of the conditioning features $\bar{\mathbf{f}}$ during the denoising process.

We optimize the noise prediction network ϵ_θ and the fusion module parameters ϕ jointly by minimizing the mean squared error between the predicted and ground-truth noise. The loss function is defined as:

$$\mathcal{L}_{\text{Diff}}(\theta, \bar{\mathbf{f}} = g(\phi)) = \mathbb{E}_{t, \mathbf{x}_0, \epsilon}\left[\|\epsilon_\theta(\mathbf{x}_t, t, \bar{\mathbf{f}}) - \epsilon\|_2^2\right], \quad (7)$$

The gradient of this loss with respect to the fusion module parameters ϕ, $\nabla_\phi \mathcal{L}_{\text{Diff}}$, is computed via the chain rule. By defining the noise residual as $r = \epsilon_\theta - \epsilon$, the gradient can be expressed more compactly as:

$$\nabla_\phi \mathcal{L}_{\text{Diff}}(\theta, \bar{\mathbf{f}}) = \mathbb{E}_{t, \mathbf{x}_0, \epsilon}\left[2r \cdot \frac{\partial \epsilon_\theta(\mathbf{x}_t, t, \bar{\mathbf{f}})}{\partial \bar{\mathbf{f}}} \frac{\partial \bar{\mathbf{f}}}{\partial \phi}\right]. \quad (8)$$

The gradient formulation in Eq. (8) reveals a direct feedback loop: it first measures how changes in the fused features $\bar{\mathbf{f}}$ affect the noise prediction (via the $\partial\epsilon_\theta/\partial\bar{\mathbf{f}}$ term), and then propagates this signal back to the fusion module parameters ϕ. By optimizing the loss in Eq. (7), the diffusion model is equivalent to pushing the reverse conditional distribution $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \bar{\mathbf{f}})$ to approach $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$. Consequently, the diffusion loss $\mathcal{L}_{\text{Diff}}$ provides a powerful generative supervision signal, driving the fusion module to learn high-quality, information-rich representations.

Method	Publication	OPV2V		DAIR-V2X	
		AP@0.5	AP@0.7	AP@0.5	AP@0.7
No Fusion	-	79.78%	67.16%	66.51%	55.46%
Early Fusion	-	95.05%	88.98%	74.52%	59.22%
Late Fusion	-	94.70%	87.18%	67.86%	50.47%
When2com	CVPR 2020	91.75%	81.77%	64.08%	49.14%
V2VNet	ECCV 2020	96.66%	87.16%	66.63%	45.61%
DiscoNet	NeurIPS 2021	90.93%	78.90%	73.58%	58.45%
Where2comm	NeurIPS 2022	86.85%	78.58%	76.13%	60.16%
AttFuse	ICRA 2022	94.31%	82.03%	73.80%	56.86%
FPVRCNN	RAL 2022	85.80%	84.00%	65.50%	50.50%
V2X-ViT	ECCV 2022	95.87%	89.88%	76.68%	57.57%
CoBEVT	CoRL 2022	91.40%	86.20%	63.90%	51.70%
CoAlign	ICRA 2023	96.63%	91.63%	75.26%	60.19%
ERMVP	CVPR 2024	95.99%	89.14%	74.73%	60.75%
MRCNet	CVPR 2024	96.55%	90.66%	73.10%	57.48%
CoSDH	CVPR 2025	**96.83%**	<u>92.99%</u>	<u>76.75%</u>	<u>63.85%</u>
DiGS-CP	-	<u>96.75%</u>	**93.53%**	**79.18%**	**64.86%**

Table 1: Main results on the OPV2V and DAIR-V2X datasets. We report Average Precision (AP) at IoU thresholds of 0.5 and 0.7. The best results are in **bold** and the second best are <u>underlined</u>.

Experiments

Experimental setup

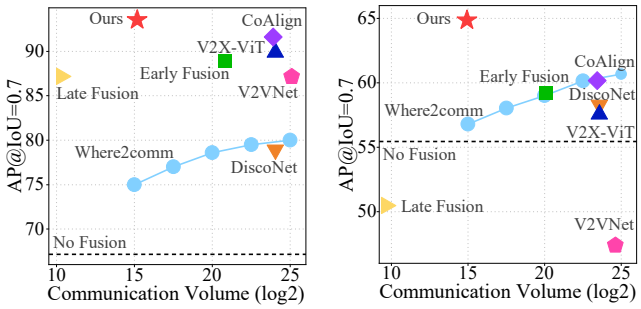
Dataset. We conduct experiments on three large-scale datasets. **OPV2V**(Xu et al. 2022c) is a joint dataset from OpenCDA(Xu et al. 2021) and CARLA (Dosovitskiy et al. 2017), providing 12K frames of 3D LiDAR point clouds and RGB images, with 230K annotated 3D bounding boxes. **V2XSet**(Xu et al. 2022b) is a simulated dataset that includes LiDAR data from vehicles and infrastructure across 73 scenarios with 2 to 5 connected agents, totaling 11K frames. **DAIR-V2X**(Yu et al. 2022) is a comprehensive real-world dataset with 9K cooperative frames, each featuring one vehicle and one roadside unit.

Implementation details. We employ PointPillars (Lang et al. 2019) as the backbone network for the first stage. The U-Net architecture comprises three downsampling layers, one bottleneck layer, and three upsampling layers. Cross-attention modules are integrated into the last two downsampling layers, the bottleneck layer, and the first two upsampling layers. For the diffusion model, we set the total diffusion steps to $T = 1000$ and adopt a linear noise schedule. We optimize the model using AdamW (Loshchilov and Hutter 2017) with an initial learning rate of 1×10^{-4} and cosine annealing scheduling with warm-up. The implementation is based on PyTorch and trained on NVIDIA RTX 4090 GPUs.

Quantitative Results

Main results. We conduct comprehensive performance comparisons of our proposed DiGS-CP against baseline and state-of-the-art methods on two mainstream collaborative perception datasets, OPV2V and DAIR-V2X. As shown in Table 1, our method demonstrates superior performance across most key metrics. On the challenging real-world DAIR-V2X dataset, DiGS-CP achieves 64.86% AP@0.7, significantly outperforming discriminative fusion methods including CoAlign (60.19%) and CoSDH (63.85%) by 4.67% and 1.01%, respectively. This validates the robustness of our generative supervision framework when handling complex real-world scenarios. On the OPV2V dataset, while our method achieves marginally lower performance than CoSDH on AP@0.5, it reaches 93.53% on the more stringent AP@0.7 metric, surpassing CoSDH by 0.54% and demonstrating superior localization accuracy. These results confirm that DiGS-CP achieves optimal detection performance through generative supervision and object-level fusion, with consistent advantages at higher AP thresholds (AP@0.7) validating the effectiveness of learning high-quality feature representations.

Performance-bandwidth trade-off analysis. Figure 2 illustrates the relationship between AP@0.7 performance and communication bandwidth requirements for various methods on the OPV2V and DAIR-V2X datasets. Late Fusion exhibits minimal communication overhead but limited performance; Early Fusion achieves better performance at the expense of extremely high communication costs; intermediate fusion methods such as V2VNet and CoAlign demonstrate performance improvements while still requiring transmission of high-dimensional BEV features, resulting in substantial communication burden. Our proposed DiGS-CP method is positioned in the optimal upper-left region across both datasets, simultaneously achieving the highest perception accuracy and low communication overhead. This demon-



(a) OPV2V (b) DAIR-V2X

Figure 2: Performance-bandwidth trade-off comparison between our method and existing approaches.

Method	V2XSet		OPV2V → V2XSet	
	AP@0.5	AP@0.7	AP@0.5	AP@0.7
Where2comm	0.8918	0.7966	0.8256 ↓6.62%	0.6754 ↓12.12%
DiscoNet	0.8577	0.7263	0.8164 ↓4.13%	0.5694 ↓15.69%
V2X-ViT	0.8903	0.7902	0.8213 ↓6.90%	0.6329 ↓15.73%
V2VNet	0.8691	0.7675	0.8397 ↓2.94%	0.5892 ↓17.83%
Ours	**0.9040**	**0.8301**	**0.8953 ↓0.87%**	**0.8199 ↓1.02%**

Table 2: Cross-dataset generalization results on V2XSet. Performance drops (%) are shown for cross-domain results.

strates that our two-stage object-level feature transmission strategy enables lightweight feature exchange to outperform high-bandwidth methods.

Cross-domain generalization. To evaluate the contribution of generative supervision to feature generalization, we conduct cross-domain experiments by training on OPV2V and evaluating on V2XSet. Table 2 shows that baseline methods exhibit substantial performance degradation during cross-domain transfer. On AP@0.7, Where2comm, DiscoNet, V2X-ViT, and V2VNet experience drops of 12.12%, 15.69%, 15.73%, and 17.83%, respectively, indicating overfitting to source domain distributions. In contrast, our DiGS-CP demonstrates superior generalization with only 0.87% and 1.02% drops on AP@0.5 and AP@0.7, significantly outperforming all baselines. This advantage stems from generative supervision encouraging the model to learn essential geometric relationships rather than task-specific decision boundaries, achieving enhanced generalization under domain shift.

Qualitative Results

Qualitative analysis of generative supervision. Figure 3 presents a comparison between ground-truth statistical feature maps and the corresponding generated results from our diffusion model. The visualization clearly demonstrates that our model generates high-fidelity feature maps conditioned on the fused features. As illustrated in the enlarged insets, the generated results accurately capture complex geometric structures of objects, including the L-shaped vehicle profile (left) and dense point distributions at vehicle front and rear regions (right). The close correspondence between generated and ground-truth feature maps provides compelling evidence for the effectiveness of our supervision mechanism. This validates that generative supervision offers a refined and structure-aware learning objective that encourages the model to develop comprehensive understanding of object geometry, thereby guiding the fusion module to preserve essential spatial structural information and achieve information-dense feature representations.

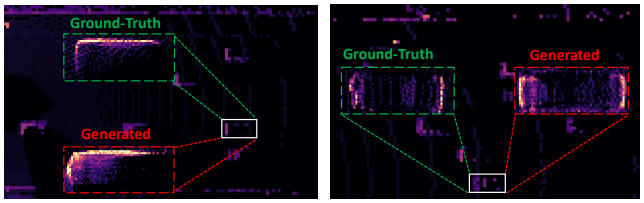


Figure 3: Qualitative results of our generative supervision.

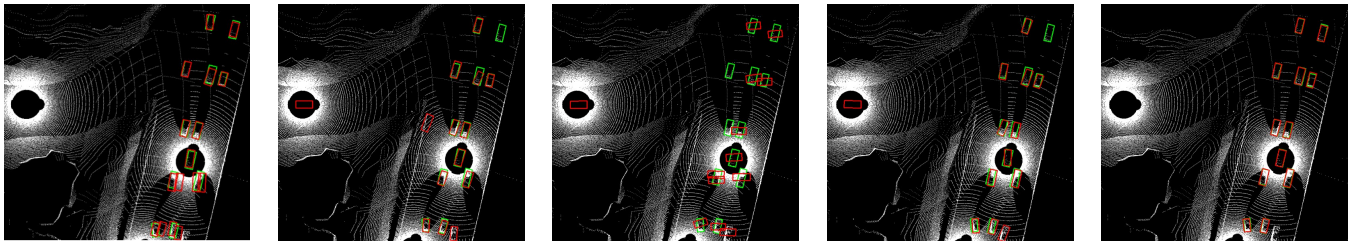
Configuration	AP@0.3	AP@0.5	AP@0.7
Supervision Paradigm			
SOTA (Discriminative)	-	**96.83%**	92.99%
Ours (Discriminative Only)	96.22%	95.35%	89.07%
Ours (Generative Supervision)	**97.16%**	96.75%	**93.53%**
Conditional Guidance			
w/ Mean-based Condition	96.00%	95.63%	92.39%
w/ Concatenation-based	96.12%	95.74%	92.63%
w/ Cross-Attention (Ours)	**97.16%**	**96.75%**	**93.53%**
Fusion Weighting Strategy			
w/ Confidence-Only Weighting	96.95%	96.50%	93.12%
w/ DWF (Ours)	**97.16%**	**96.75%**	**93.53%**

Table 3: Ablation studies on our proposed components on the OPV2V dataset.

Qualitative analysis of detection results. We visualize detection results on a challenging scene featuring distant vehicles and partial occlusions, comparing our method against several state-of-the-art approaches. As illustrated in Figure 4, baseline methods exhibit significant limitations when handling such complex scenarios. V2XNet (b), DiscoNet (c), and CoAlign (d) fail to detect multiple distant vehicles, resulting in notable missed detections. V2X-ViT (a) produces redundant detection boxes. Conversely, our DiGS-CP method (e) successfully detects all vehicles in the scene, including distant and closely clustered objects that pose detection challenges. These qualitative results provide visual confirmation that generative supervision enables the fusion module to learn robust and precise feature representations.

Ablation Studies

Efficacy of generative supervision. We first conduct ablation experiments to validate our generative supervision effectiveness. As presented in the first group of Table 3, we compare our complete method against two discrimina-



(a) V2X-ViT (b) V2XNet (c) DiscoNet (d) CoAlign (e) Ours

Figure 4: Qualitative comparison of our method with state-of-the-art approaches on a challenging scenario from the OPV2V dataset. Green boxes represent the ground-truth, while red boxes indicate the detection results.

tive supervision baselines: the current state-of-the-art discriminative method and our framework without generative supervision. Results show our discriminative baseline achieves 95.35% AP@0.5, slightly below the discriminative SOTA's 96.83% due to architectural differences. Incorporating generative supervision yields substantial improvements: our complete model achieves 93.53% AP@0.7—a 4.46% gain over our baseline and surpassing the discriminative SOTA's 92.99%. This comparison highlights generative supervision's advantages in both high-precision detection and feature learning over discriminative methods.

Analysis of conditional guidance strategies. The second group in Table 3 compares our cross-attention mechanism with two simpler baselines: concatenation-based conditioning, which concatenates the conditional vector with U-Net intermediate features, and mean-based conditioning, which computes the element-wise mean of the conditional features and input features. Results demonstrate that while simpler strategies yield some performance gains, cross-attention achieves superior performance. This advantage stems from cross-attention's ability to dynamically query and leverage the most relevant information from the guidance condition, enabling more precise conditional generation and more effective supervision.

Efficacy of Distance-Weighted Fusion. We validate the effectiveness of our proposed Distance-Weighted Fusion strategy. As shown in the third group of Table 3, compared to a weighting scheme that uses only confidence scores, our DWF strategy, which incorporates a distance-based penalty, achieves superior performance across all metrics. This demonstrates that DWF effectively down-weights the influence of distant, less reliable bounding boxes during the fusion process, leading to more accurate final proposals.

Impact of confidence threshold τ. Figure 5 (a) illustrates the performance curves across three AP metrics as the confidence threshold varies from 0.1 to 0.4. The figure clearly demonstrates that all metrics exhibit an initial increase followed by a decrease, peaking at $\tau = 0.2$. This phenomenon reveals a balance: low thresholds (e.g., 0.1) retain excessive low-quality detections with noise and false positives, while high thresholds (e.g., 0.3 or above) over-filter detections, resulting in the loss of valid detection boxes, particularly for distant or partially occluded objects. Therefore, $\tau = 0.2$ provides the optimal trade-off between retaining valid detections and filtering out noise.

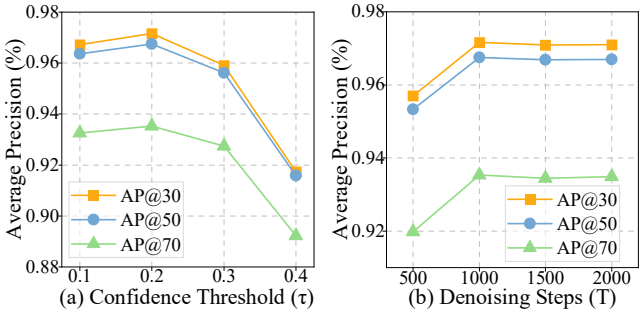


Figure 5: Ablation studies on key hyperparameters on the OPV2V dataset. **(a)** Impact of the confidence threshold (τ) for initial proposals. **(b)** Impact of the total number of diffusion timesteps (T).

Impact of total diffusion steps T. Figure 5 (b) illustrates the impact of diffusion steps T on performance. As T increases from 500 to 1000, all AP metrics improve substantially, indicating that more steps enable finer noise scheduling and better capture of the data distribution. However, performance plateaus or declines when T exceeds 1000, suggesting an optimal range exists. Considering the performance-efficiency trade-off, we adopt $T = 1000$.

Conclusion

In this paper, we propose DiGS-CP, a two-stage collaborative perception framework. This framework introduces a generative supervision mechanism that uses fused features as conditions to guide a diffusion model in generating object-level point cloud statistical feature maps. This provides fine-grained, task-agnostic supervision for the fusion process, enabling the fusion module to generate comprehensive features for collaboration and improving their generalization ability. Experimental results demonstrate that DiGS-CP achieves state-of-the-art performance across three benchmark datasets while maintaining low communication overhead. This work opens a new research direction for collaborative perception, demonstrating the significant potential of generative supervision in multi-agent collaboration.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFB4301900, in part by the Natural Science Foundation of China under Grant 62472048, and in part by Beijing Natural Science Foundation under Grant L242081.

References

Borse, S.; Das, D.; Park, H.; Cai, H.; Garrepalli, R.; and Porikli, F. 2023. Dejavu: Conditional regenerative learning to enhance dense prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19466–19477.

Chen, Q.; Ma, X.; Tang, S.; Guo, J.; Yang, Q.; and Fu, S. 2019a. F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3D point clouds. In *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing*, 88–100.

Chen, Q.; Tang, S.; Yang, Q.; and Fu, S. 2019b. Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, 514–524. IEEE.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; and Koltun, V. 2017. CARLA: An open urban driving simulator. In *Conference on robot learning*, 1–16. PMLR.

Fang, Z.; Hu, S.; An, H.; Zhang, Y.; Wang, J.; Cao, H.; Chen, X.; and Fang, Y. 2024. PACP: Priority-aware collaborative perception for connected and autonomous vehicles. *IEEE Transactions on Mobile Computing*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Hu, H.; Liu, Z.; Chitlangia, S.; Agnihotri, A.; and Zhao, D. 2022a. Investigating the impact of multi-lidar placement on object detection for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2550–2559.

Hu, S.; Tao, Y.; Xu, G.; Deng, Y.; Chen, X.; Fang, Y.; and Kwong, S. 2025. Cp-guard: Malicious agent detection and defense in collaborative bird's eye view perception. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 23203–23211.

Hu, Y.; Fang, S.; Lei, Z.; Zhong, Y.; and Chen, S. 2022b. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. *Advances in neural information processing systems*, 35: 4874–4886.

Huang, X.; Wang, J.; Xia, Q.; Chen, S.; Yang, B.; Li, X.; Wang, C.; and Wen, C. 2025. V2X-R: Cooperative LiDAR-4D Radar Fusion with Denoising Diffusion for 3D Object Detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 27390–27400.

Lang, A. H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; and Beijbom, O. 2019. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12697–12705.

Li, Y.; Ren, S.; Wu, P.; Chen, S.; Feng, C.; and Zhang, W. 2021. Learning distilled collaboration graph for multi-agent perception. *Advances in Neural Information Processing Systems*, 34: 29541–29552.

Liu, Y.-C.; Tian, J.; Glaser, N.; and Kira, Z. 2020. When2com: Multi-agent perception via communication graph grouping. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 4106–4115.

Liu, Z.; Tang, H.; Amini, A.; Yang, X.; Mao, H.; Rus, D. L.; and Han, S. 2023. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, 2774–2781. IEEE.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lu, Y.; Li, Q.; Liu, B.; Dianati, M.; Feng, C.; Chen, S.; and Wang, Y. 2023. Robust collaborative 3d object detection in presence of pose errors. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 4812–4818. IEEE.

Luo, G.; Shao, C.; Cheng, N.; Zhou, H.; Zhang, H.; Yuan, Q.; and Li, J. 2023. EdgeCooper: Network-aware cooperative LiDAR perception for enhanced vehicular awareness. *IEEE Journal on Selected Areas in Communications*, 42(1): 207–222.

Luo, G.; Zhang, H.; Yuan, Q.; and Li, J. 2022. Complementarity-enhanced and redundancy-minimized collaboration network for multi-agent perception. In *Proceedings of the 30th ACM International Conference on Multimedia*, 3578–3586.

Ng, A.; and Jordan, M. 2001. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14.

Rawashdeh, Z. Y.; and Wang, Z. 2018. Collaborative automated driving: A machine learning-based method to enhance the accuracy of shared information. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 3961–3966. IEEE.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Shang, S.; Shan, Z.; Liu, G.; Wang, L.; Wang, X.; Zhang, Z.; and Zhang, J. 2024. Resdiff: Combining cnn and diffusion model for image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 8975–8983.

Shao, C.; Yuan, Q.; Luo, G.; Hu, Y.; Wang, D.; Liu, Y.; Pan, R.; Chen, B.; and Li, J. 2025. NegoCollab: A Common Representation Negotiation Approach for Heterogeneous Collaborative Perception. *arXiv preprint arXiv:2510.27647*.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. pmlr.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, R.; Liang, C.; Cao, H.; Yan, Z.; Zimmer, W.; Gross, M.; Festag, A.; and Knoll, A. 2024. Collaborative semantic occupancy prediction with hybrid feature fusion in connected automated vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17996–18006.

Su, W.; Chen, L.; Bai, Y.; Lin, X.; Li, G.; Qu, Z.; and Zhou, P. 2024. What makes good collaborative views? contrastive mutual information maximization for multi-agent perception. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 17550–17558.

Wang, B.; Zhang, L.; Wang, Z.; Zhao, Y.; and Zhou, T. 2023a. Core: Cooperative reconstruction for multi-agent perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8710–8720.

Wang, T.; Chen, G.; Chen, K.; Liu, Z.; Zhang, B.; Knoll, A.; and Jiang, C. 2023b. Umc: A unified bandwidth-efficient and multi-resolution based collaborative perception framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8187–8196.

Wang, T.; Kim, S.; Wenxuan, J.; Xie, E.; Ge, C.; Chen, J.; Li, Z.; and Luo, P. 2024. Deepaccident: A motion and accident prediction benchmark for v2x autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5599–5606.

Wang, T.-H.; Manivasagam, S.; Liang, M.; Yang, B.; Zeng, W.; and Urtasun, R. 2020. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part II 16*, 605–621. Springer.

Xiang, H.; Xu, R.; and Ma, J. 2023. HM-ViT: Hetero-modal vehicle-to-vehicle cooperative perception with vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 284–295.

Xu, J.; Zhang, Y.; Cai, Z.; and Huang, D. 2025a. CoSDH: Communication-Efficient Collaborative Perception via Supply-Demand Awareness and Intermediate-Late Hybridization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 6834–6843.

Xu, P.; Fan, Q.; Kou, F.; Qin, S.; Gu, H.; Zhao, R.; Ling, C.; and Wang, B. 2025b. Textualize Visual Prompt for Image Editing via Diffusion Bridge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 21779–21787.

Xu, R.; Guo, Y.; Han, X.; Xia, X.; Xiang, H.; and Ma, J. 2021. Opencda: an open cooperative driving automation framework integrated with co-simulation. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, 1155–1162. IEEE.

Xu, R.; Tu, Z.; Xiang, H.; Shao, W.; Zhou, B.; and Ma, J. 2022a. CoBEVT: Cooperative bird's eye view semantic segmentation with sparse transformers. *arXiv preprint arXiv:2207.02202*.

Xu, R.; Xiang, H.; Tu, Z.; Xia, X.; Yang, M.-H.; and Ma, J. 2022b. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In *European conference on computer vision*, 107–124. Springer.

Xu, R.; Xiang, H.; Xia, X.; Han, X.; Li, J.; and Ma, J. 2022c. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *2022 International Conference on Robotics and Automation (ICRA)*, 2583–2589. IEEE.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Yu, H.; Luo, Y.; Shu, M.; Huo, Y.; Yang, Z.; Shi, Y.; Guo, Z.; Li, H.; Hu, X.; Yuan, J.; et al. 2022. Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21361–21370.

Yu, H.; Yang, W.; Zhong, J.; Yang, Z.; Fan, S.; Luo, P.; and Nie, Z. 2025. End-to-end autonomous driving through v2x cooperation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 9598–9606.

Zhang, J.; Wang, Y.; Qian, L.; Sun, P.; Li, Z.; Jiang, S.; Liu, M.; and Song, L. 2025. Dsrc: Learning density-insensitive and semantic-aware collaborative representation against corruptions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 9942–9950.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.