

# CoRA: A Collaborative Robust Architecture with Hybrid Fusion for Efficient Perception

Gong Chen<sup>1</sup>, Chaokun Zhang<sup>1\*</sup>, Pengcheng Lv<sup>2</sup>, Xiaohui Xie<sup>3</sup>

<sup>1</sup>College of Intelligence and Computing, Tianjin University, Tianjin, China

<sup>2</sup>School of Future Technology, Tianjin University, Tianjin, China

<sup>3</sup>Department of Computer Science and Technology, Tsinghua University, Beijing, China  
{gongchen01, zhangchaokun}@tju.edu.cn

## Abstract

Collaborative perception has garnered significant attention as a crucial technology to overcome the perceptual limitations of single-agent systems. Many state-of-the-art (SOTA) methods have achieved communication efficiency and high performance via intermediate fusion. However, they share a critical vulnerability: their performance degrades under adverse communication conditions due to the misalignment induced by data transmission, which severely hampers their practical deployment. To bridge this gap, we re-examine different fusion paradigms, and recover that the strengths of intermediate and late fusion are not a trade-off, but a complementary pairing. Based on this key insight, we propose CoRA, a novel collaborative robust architecture with a hybrid approach to decouple performance from robustness with low communication. It is composed of two components: a feature-level fusion branch and an object-level correction branch. Its first branch selects critical features and fuses them efficiently to ensure both performance and scalability. The second branch leverages semantic relevance to correct spatial displacements, guaranteeing resilience against pose errors. Experiments demonstrate the superiority of CoRA. Under extreme scenarios, CoRA improves upon its baseline performance by approximately 19% in AP@0.7 with more than 5x less communication volume, which makes it a promising solution for robust collaborative perception.

## Introduction

Autonomous driving perception systems are fundamentally limited by occlusions and finite sensing ranges, which pose critical safety risks (Pradeep, Bakoev, and Akhroljonova 2023). Enabled by Vehicle-to-Everything (V2X) communication, collaborative perception presents a powerful solution by sharing information across agents (Omeiza et al. 2022). However, the benefits of V2X are constrained by real-world communication bottlenecks, such as limited bandwidth and unreliability (Trabelsi, Fourati, and Chen 2024). To mitigate these overheads, the research community has converged on intermediate fusion, an approach that exchanges feature-level data. This paradigm has become the prevailing choice for SOTA methods (Hu et al. 2022; Zhang et al. 2024).

\*Corresponding author.

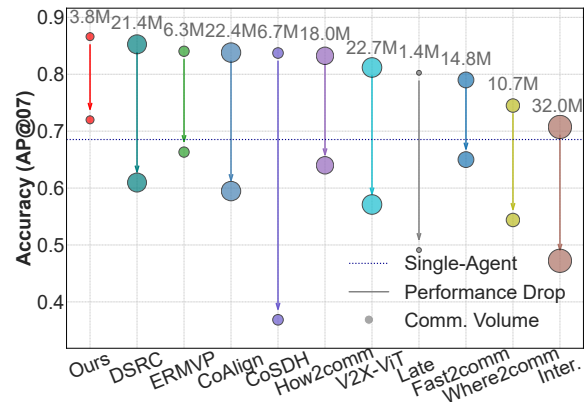


Figure 1: Performance comparison of collaborative perception methods in communication efficiency and robustness.

Existing communication-efficient methods focus mainly on selectively transmitting (Hu et al. 2022; Yang et al. 2023) or compressing features (Wang et al. 2023) to reduce communication overhead. However, a fundamental limitation of these frameworks lies in their presupposition of accurate alignment in a shared coordinate system, which critically depends on precise localization (Yazgan et al. 2024).

This assumption crumbles when unstable communication channels undermine the integrity of transmitted pose information, which in turn exacerbates pose errors and leads to severe data misalignment during multi-agent fusion (Tang et al. 2025). This highlights a challenge that extends beyond mere bandwidth constraints (Su et al. 2023). The presence of pose errors makes these efficient strategies susceptible to degradation. The severity of this issue is illustrated in Fig.1: leading efficient methods degrade in performance under pose errors, falling even below single-agent detection.

To develop a robust solution for these real-world challenges, we first re-examine the vulnerability patterns of different fusion paradigms. Intermediate Fusion leverages dense feature-level interaction to achieve a high performance ceiling when poses are accurate (Xu et al. 2022a). However, any pose-induced misalignment at the feature level can trigger mismatched interactions, resulting in an irrecoverable collapse of the shared representation. In con-

trast, late fusion exhibits greater resilience. Pose errors do not corrupt the features themselves but affect object association (Arnold et al. 2022). While this leads to the mislocalization of collaborator detections, the intrinsic integrity of each individual detection is preserved. Although its accuracy ceiling is typically lower, the integrity of information grants it superior rectifiability. Our analysis thus reveals a key insight: *the weaknesses of intermediate and late fusion are complementary*. This discovery provides the theoretical foundation for our hybrid design.

Motivated by this insight, we introduce CoRA, a parallel, dual-branch framework that unifies the complementary benefits of both intermediate and late fusion. This design decouples the pursuit of high performance from pose error robustness, overcoming the apparent trade-off prevalent in conventional methods under strict bandwidth limitations. Specifically, our contributions are embodied in two synergistic branches: 1) A high-performance feature-level fusion branch. We first introduce a receiver-centric Competitive Information Transmission (CIT) module to select and transmit only the most critical features. The module exchanges lightweight confidence maps to identify the optimal information provider for each spatial region and then requests only the critical features to fill its perceptual gaps. This approach keeps the communication overhead nearly constant regardless of the number of collaborators, thereby ensuring the system’s scalability. Next, we employ a Lightweight Collaboration (LC) module to enable efficient knowledge sharing across collaborators at the feature level for high-accuracy perception. Its core objective is to exploit the high-performance ceiling of intermediate fusion. The entire process is optimized end-to-end via a training-only guidance module, ensuring that the most valuable features are processed to ultimately maximize the global perception gain. 2) A robust object-level correction branch. A naive combination of fusion strategies is insufficient (Xu et al. 2025), as standard late fusion itself degrades severely under pose errors. To guarantee resilience against pose errors, we introduce a novel Pose-Aware Correction (PAC) module. This module mitigates spatial misalignment by leveraging the semantic relevance between ego and collaborator detections, thereby ensuring stable performance even under pose errors.

Extensive experiments on two benchmarks validate the superiority of our method. On OPV2V, for instance, results show that CoRA achieves a remarkable 86.58% in AP@0.7 with only 3.80 MB communication overhead under extremely adverse conditions. Compared to the baseline, CoRA increases the detection accuracy by 19.02% while slashing communication volume by 82.2%, demonstrating an exceptional balance between efficiency and robustness. The contributions of this work are summarized as follows:

- We propose CoRA, a novel dual-stream hybrid framework that decouples performance and robustness for reliable collaborative perception.
- We introduce CIT and LC for efficient feature-level fusion, and PAC for robust object-level correction, decoupling performance from robustness.
- We conduct comprehensive experiments on multiple

datasets. CoRA consistently outperforms over nine powerful methods, setting a new SOTA with minimal communication overhead.

## Related Work

### Collaborative Perception

Collaborative perception enables multiple agents to build a comprehensive environmental understanding by exchanging sensory information. This field is broadly categorized into three paradigms based on the fusion stage: early, intermediate, and late fusion. Early fusion is considered impractical due to its prohibitive communication bandwidth requirements (Gao et al. 2024). Meanwhile, although late fusion is communication-efficient, it suffers from limited performance because fusing detections at the object level discards low-level context (Lv et al. 2024). This has led to the prevalence of intermediate fusion, which offers a favorable trade-off between perception accuracy and communication overhead (Xu et al. 2023). Numerous methods have been proposed within this paradigm. V2X-ViT (Xu et al. 2022a) designs specialized attention mechanisms for interaction among heterogeneous agents; CoAlign (Lu et al. 2023) leverages graph neural networks to achieve pose-robust collaboration; DSRC (Zhang et al. 2025) employs teacher-student distillation to enhance performance in adverse weather conditions. Despite these advances, purely intermediate fusion exposes a key vulnerability: direct feature-level fusion is susceptible to cross-agent interference. Motivated by this limitation, our work proposes a more robust and practical solution that integrates the high performance of intermediate fusion with the rectifiability of late fusion.

### Multi-Agent Communication

To make intermediate fusion practical, numerous strategies have been proposed to reduce communication overhead. One major strategy involves transmitting only the most critical features. For instance, sender-centric frameworks like Where2comm (Hu et al. 2022) and How2comm (Yang et al. 2023) require each agent to select key features to minimize its transmitted volume independently. The second strategy focuses on the compression of holistic features. Methods like CORE (Wang et al. 2023) use deep learning to compress the entire feature map, while CodeFilling (Hu et al. 2024) transmits learned codes corresponding to the features for efficiency. Although these methods are effective at reducing the payload from individual agents, they share a common limitation: the communication and computation load on the receiver still scales linearly with the number of collaborators. To address this scalability issue, our work proposes a novel on-demand transmission strategy. This ensures the communication burden remains nearly constant, regardless of the number of participating agents.

## Methodology

### Problem Formulation

We consider a cooperative scenario involving a set of  $N$  agents, including ego and collaborators, denoted by  $\mathcal{A} =$

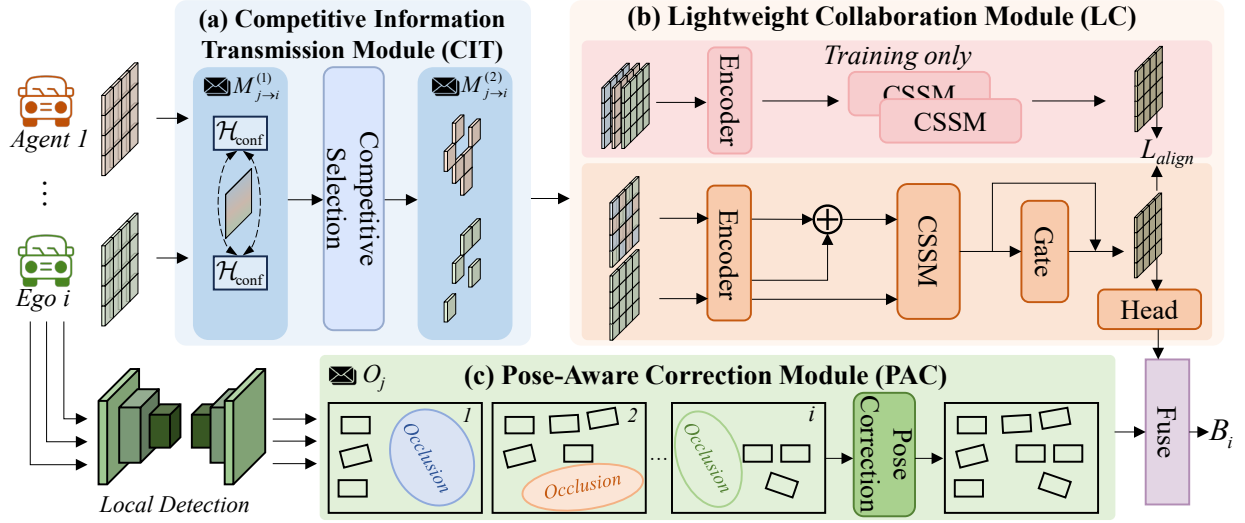


Figure 2: Overall architecture of the proposed CoRA. CoRA comprises two main branches: a feature-level fusion branch (blue & orange) and an object-level correction branch (green). The feature branch leverages CIT for competitive feature transmission and LC for post-fusion interaction. The LC’s training is guided by dense fused features from other vehicles. Meanwhile, the object-level correction branch transmits collaborator detection results to the ego and corrects them via the PAC module.

$\{1, 2, \dots, N\}$ . Each agent  $j \in \mathcal{A}$  captures local sensor data  $X_j$  and its own pose data  $T_j$ . To facilitate collaboration under strict bandwidth limits, a collaborator  $j$  generates a compact message  $M_j$  from its local data  $X_j$  and transmits this message to the ego  $i$ . The goal of ego  $i$  is to learn a collaborative perception model  $\mathcal{F}_{\text{coop}}$ . This model takes the ego’s own local data  $X_i$  and the set of messages  $\{M_{j \rightarrow i}\}_{j \neq i}$  received from other agents as input, to output the final detection predictions  $B_i$ . Formally, the problem is to design  $\mathcal{F}_{\text{coop}}$ :

$$B_i = \mathcal{F}_{\text{coop}}(X_i, \{M_{j \rightarrow i}\}_{j \in \mathcal{A}}) \quad \text{s.t. } |M_{j \rightarrow i}| \leq C \quad (1)$$

where the message  $M_{j \rightarrow i}$  is generated from the data  $X_j$  of agent  $j$ .  $|M_{j \rightarrow i}|$  represents the communication cost of the message, which is constrained by a budget  $C$ . However, a fundamental challenge arises in real-world applications: the poses are imperfect. We only have access to noisy poses  $\hat{T}_i = T_i \oplus \Delta T_i$ , where  $\Delta T_i$  is an unknown pose error.

## Overall Architecture

As illustrated in Fig. 2, our framework for each ego  $i$  utilizes a parallel dual-stream architecture to decouple performance from robustness. The feature-level branch first employs a two-stage transmission strategy: collaborators  $j$  ( $j \neq i$ ) first use Competitive Information Transmission (CIT) module to exchange lightweight confidence maps ( $M_{j \rightarrow i}^{(1)}$ ) to identify perception demands, and transmit only sparse features ( $M_{j \rightarrow i}^{(2)}$ ). The Lightweight Collaboration (LC) module then fuses these sparse features. LC is designed for information integration with dense feature distillation during training. Concurrently, the object-level branch serves as a compensation mechanism. We propose a novel Pose-Aware Correction (PAC) module to rectify misalignments involving the received detection outputs  $O_j$ . Finally, the outputs from both branches are fused to produce the final prediction  $B_i$ .

## Competitive Information Transmission Module

Prevailing methods for collaborative perception often adopt a sender-centric communication strategy. For instance, in ERMVP (Zhang et al. 2024), each collaborator compresses and transmits features based on its self-generated or an individual ego’s confidence map. While this approach reduces the payload from each agent, the feature volume processed by the ego scales significantly with the number of collaborators. To overcome this bottleneck, we introduce the CIT module, which shifts the paradigm from sender-centric broadcasting to a receiver-centric, on-demand protocol, ensuring a consistently low communication overhead.

CIT operates through a two-stage communication mechanism. In the first stage, each collaborator  $j$  processes its sensor data with a local encoder  $\mathcal{E}$  to produce a feature map  $F_j \in \mathbb{R}^{C \times H \times W}$ . Subsequently, a confidence head  $\mathcal{H}_{\text{conf}}$  generates a map  $M_{j \rightarrow i}^{(1)} \in \mathbb{R}^{1 \times H \times W}$  as the initial message transmitted to the ego  $i$ .

$$M_{j \rightarrow i}^{(1)} = \mathcal{H}_{\text{conf}}(F_j), \quad j \neq i \quad (2)$$

Upon receiving  $\{M_{j \rightarrow i}^{(1)}\}$  from all collaborators, ego  $i$  determines the optimal information provider for each spatial location. It first computes its own demand coefficient matrix  $D_i$  and uses it to generate a relevance score  $S_j$  for each collaborator  $j$  to enhance regions that the ego cannot perceive:

$$D_i = 1 - \sigma(\mathcal{H}_{\text{conf}}(F_i)) \quad (3)$$

$$S_j = D_i \odot M_{j \rightarrow i}^{(1)}, \quad j \neq i \quad (4)$$

where  $\sigma(\cdot)$  is the sigmoid function and  $\odot$  denotes the multiplication by elements. A winner-take-all strategy then per-

forms pixel-wise competition on all relevance maps to determine each pixel's most confident collaborator index.

$$I_{\text{win}} = \underset{j \neq i}{\operatorname{argmax}} (\operatorname{Stack}(\{S_j\})) \quad (5)$$

where  $I_{\text{win}} \in \mathbb{N}^{H \times W}$  assigns each pixel to its best-providing agent. Based on  $I_{\text{win}}$ , the ego  $i$  generates a set of binary request masks  $Q_j$ , which are both exclusive and highly sparse.

$$Q_j(h, w) = \begin{cases} 1 & \text{if } I_{\text{win}}(h, w) = j \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

where  $Q_j \in \{0, 1\}^{1 \times H \times W}$ . Each unique request mask  $Q_j$  is then transmitted back to the corresponding collaborator  $j$ .

In the second stage, each collaborator  $j$  uses its request mask  $Q_j$  to extract and transmit only the critical sparse features  $M_{j \rightarrow i}^{(2)} = F_j \odot Q_j$ .

Due to the spatial disjointness of the request masks, ego  $i$  directly sums the received features, yielding a consolidated collaborative feature  $F_{\text{coll}} = \sum_{j \neq i} M_{j \rightarrow i}^{(2)}$ . This on-demand mechanism transforms the communication burden to a near-constant overhead, thus enhancing system scalability.

### Lightweight Collaboration Module

We further introduce the LC module to effectively fuse this collaborative feature  $F_{\text{coll}}$  with the ego feature  $F_i$ . As depicted in Fig. 3, LC leverages Collaborative State Space Model (CSSM) based on Mamba (Gu and Dao 2024) to capture inter-vehicle dependencies and employs a Gating Unit to ensure adaptive feature integration. Meanwhile, a training-only module is used to guide feature distillation.

First, to prioritize features from high-confidence regions, we element-wise multiply  $F_{\text{coll}}$  and  $F_i$  with their respective confidence maps,  $S_{\text{coll}}$  and  $S_i$ . Notably,  $S_{\text{coll}}$  is generated by aggregating the confidence maps from collaborators.

$$\hat{F}_{\text{coll}} = F_{\text{coll}} \odot S_{\text{coll}}, \quad \hat{F}_i = F_i \odot S_i \quad (7)$$

Next,  $\hat{F}_{\text{coll}}$  is fed into an attention block (Xu et al. 2022b) to harmonize the aggregated features from heterogeneous agents, thereby mitigating potential misalignments. Subsequently, both  $\hat{F}_{\text{coll}}$  and  $\hat{F}_i$  are processed by convolutional operations. The outputs of two branches,  $Z_{\text{coll}}$  and  $Z_i$ , are then fused through element-wise addition, yielding  $Z_{\text{fused}}$ .

CSSM then facilitates efficient inter-vehicle interaction. Similar to modern SSM methods (Liu et al. 2024), CSSM receives three key inputs:  $Z_{\text{fused}}$  serves as the input matrices, a linear transformation of  $Z_{\text{fused}}$  parameterizes the step size, and  $Z_i$  acts as the output matrices. This design enables content-aware contextual modeling, yielding  $X_{\text{ssm}}$ :

$$X_{\text{ssm}} = \text{CSSM}(Z_{\text{fused}}, \text{Linear}(Z_{\text{fused}}), Z_i) \quad (8)$$

Finally, we employ a Gating Unit for fusion.  $X_{\text{ssm}}$  is processed to generate a spatial gate  $g \in \mathbb{R}^{1 \times H \times W}$  via a convolutional operation. Currently,  $X_{\text{ssm}}$  is fed into MLP, and its output is then modulated by the gate  $g$  to get the output  $F_{\text{out}}$ .

$$g = \sigma(\text{DWConv}(\text{Conv}(X_{\text{ssm}}))) \quad (9)$$

$$F_{\text{out}} = \text{Conv}(\text{MLP}(X_{\text{ssm}}) \odot g) \quad (10)$$

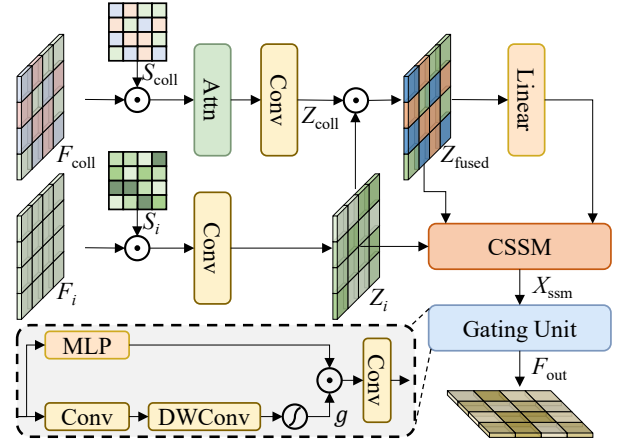


Figure 3: Illustration of the lightweight collaboration module.

**Feature Distillation.** To compensate for potential information loss from the sparse input  $F_{\text{coll}}$ , we employ a feature distillation mechanism during training. As shown in the top-right corner of the Fig. 2, a parallel teacher branch processes complete, non-sparse features to generate a guidance map  $F_{\text{teacher}}$ .  $F_{\text{out}}$  is encouraged to mimic this target representation via an alignment loss  $\mathcal{L}_{\text{align}}$ :

$$\mathcal{L}_{\text{align}} = \|F_{\text{out}} - F_{\text{teacher}}\|_2^2 \quad (11)$$

### Pose-Aware Correction Module

To enhance robustness against localization errors inherent in real-world scenarios, we introduce the PAC module, a pose correction module that operates on the object-level outputs from collaborating agents. As shown in Fig. 4, PAC receives the detection results from each collaborator  $j$ , which are represented as classification maps  $C_j \in \mathbb{R}^{N_{\text{cls}} \times H \times W}$  and regression maps  $R_j \in \mathbb{R}^{N_{\text{reg}} \times H \times W}$ . Instead of direct fusion, PAC leverages the ego's results  $\{C_i, R_i\}$  for relevance guidance.

As shown on the top of Fig. 4, the alignment process begins by addressing association uncertainty. We first employ convolutional operations to select high-confidence results from collaborator outputs. Next, we generate a descriptor for each detection by incorporating Positional Embedding (PE) of its corresponding box parameters. These descriptors from both ego  $i$  and collaborator  $j$  are then processed to compute a cross-agent attention map  $A_j \in \mathbb{R}^{1 \times H \times W}$ , which quantifies the matching degree at each spatial location.

$$A_j = \sigma(f_{\text{attn}}(\text{Concat}(\text{PE}(O_i), \text{PE}(O_j)))) \quad (12)$$

where  $O_i$  and  $O_j$  denote the detection maps, and  $f_{\text{attn}}$  generates the attention map by computing the similarity. The resulting attention map then modulates the collaborator's output, yielding relevance-scored outputs:

$$C'_j = C_j \odot A_j, \quad R'_j = R_j \odot A_j \quad (13)$$

Concurrently, the module performs an explicit semantic misalignment caused by pose errors. We concatenate the detection maps from agents and process them through a shared

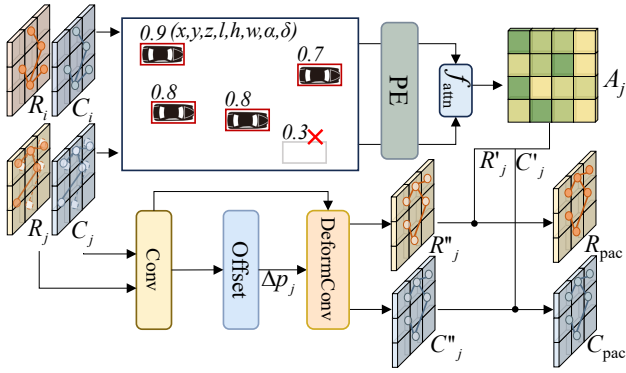


Figure 4: Illustration of the pose-aware correction module.

encoder  $f_{\text{offset}}$  to predict a dense 2D offset field  $\Delta p_j$ . This field represents the estimated spatial displacement for each cell of the collaborator’s feature map.

$$\Delta p_j = f_{\text{offset}}(\text{Concat}(C_i, R_i, C_j, R_j)) \quad (14)$$

We then leverage  $\Delta p_j$  to perform alignment using deformable convolutions (Dai et al. 2017). This operation re-samples the collaborator’s detection maps at the corrected locations specified by  $\Delta p_j$  to rectify the spatial errors:

$$C''_j = \text{DeformConv}(C_j, \Delta p_j) \quad (15)$$

$$R''_j = \text{DeformConv}(R_j, \Delta p_j) \quad (16)$$

Finally, we fuse these two corrected representations to produce the final collaborator detection maps.

### Adaptive Final Fusion

We employ an adaptive fusion mechanism to recalibrate prediction confidence scores of predictions from each branch. We first concatenate classification maps ( $C_{lc}$ ,  $C_{pac}$ ) and use a convolutional network to produce uncertainty maps ( $U_{lc}$ ,  $U_{pac}$ ), and then adjust initial confidence  $\delta$  using associated uncertainty maps  $U_{lc/pac}$  for a recalibrated score.

Finally, predictions from both the feature-level and object-level branches are combined into a single, unified pool. A 3D Non-Maximum Suppression (NMS) operation is then applied to this pool to prune redundant predictions, yielding the ego’s definitive collaborative output  $B_i$ .

## Experiments

### Setup

**Dataset.** To validate the effectiveness of CoRA, we utilize OPV2V (Xu et al. 2022b) and DAIR-V2X (Yu et al. 2022), two of the most widely used datasets for collaborative perception. Specifically, OPV2V is a large-scale simulated dataset focusing on V2V scenarios, simulating complex multi-vehicle traffic scenarios and communication occlusions. DAIR-V2X is a large-scale real-world dataset that emphasizes V2I scenarios, providing vehicle-side and road-side data in real traffic environments to compensate for single-vehicle perception blind spots. These two datasets cover both simulated and real-world scenarios, as well as

the two main collaborative types, enabling a comprehensive and effective evaluation of the model’s effectiveness.

**Experimental Setup.** We adopted the basic settings from OpenCOOD (Xu et al. 2022b). All models were trained on a single RTX 4090 GPU with 24GB. We used the Adam optimizer, set the training epochs to 30, and the batch size to 2. The vehicle communication range was limited to 70 meters. For all models, PointPillar (Lang et al. 2019) was chosen as the backbone network to extract 2D features from point clouds, with each voxel’s side length set to 0.4 meters. Performance was evaluated using Average Precision at IoU 0.5 and 0.7 (AP@0.5 / AP@0.7), standard 3D object detection metrics that measure both bounding box accuracy and spatial overlap.

### Quantitative Evaluation

**Comparison of Detection Performance.** As shown in Tab. 1, under ideal conditions, our method achieves state-of-the-art performance on both benchmarks, reaching 86.58% AP@0.7 on OPV2V and 63.61% AP@0.7 on DAIR-V2X, respectively. Notably, this top-tier performance is achieved with extremely low communication overheads of only 3.80 MB for OPV2V and 2.84 MB for DAIR-V2X, significantly reducing communication volume by 82.2% compared to other high-performance methods like DSRC and V2VNet. When low pose error is introduced, our method demonstrates superior robustness, surpassing SOTA performance with 78.17% AP@0.7 on OPV2V and 59.77% AP@0.7 on DAIR-V2X. Notably, the incorporation of the post-correction branch might lead to a slight performance perturbation in environments with minimal pose error compared to the intermediate-only branch. This trade-off is acceptable for enhancing efficiency and leveraging the robust advantages of the post-fusion branch.

**Comparison of Robustness.** As detailed in Tab. 1 and Fig. 6, Our model demonstrates exceptional robustness against pose errors. Under the high pose error level of 0.6/0.6  $\sigma_t/\sigma_r$ , our approach maintains high accuracy on OPV2V, outperforming strong methods like CoAlign by a significant 17.0% in AP@0.7. This resilience is further validated in Fig. 6, which shows performance scaling with vehicle count. For instance, while CoSDH exhibits competitive performance under ideal conditions with low communication overhead, its performance degrades gradually as more vehicles with perturbed data are introduced. In sharp contrast, our method consistently maintains its SOTA performance, leveraging collaborators even in noisy multi-agent settings.

**Latency Sensitivity Analysis.** Communication latency remains a challenge for communication-efficient methods. We thus evaluate these methods under the challenging combination of latency and pose noise. Tab. 2 reveals a critical distinction: while the performance of competing methods degrades under delay, CoRA demonstrates far superior resilience. Its ability to maintain AP@0.7 of 0.3651 at 300ms latency highlights its robustness to latency.

**Computational Cost Analysis.** Fig. 5 highlights the efficiency of our core fusion module. CoRA exhibits exceptional scalability. As the number of agents increases, their GFLOPs remain remarkably stable, increasing by a mere

Pose error Level $\sigma_t/\sigma_r$ (m/°)	Comm.	0/0	0.2/0.2	0.4/0.4	0.6/0.6
OPV2V					
Method/Dataset					
Single	\	0.8078/0.6853	0.8078/0.6853	0.8078/0.6853	0.8078/0.6853
V2VNet (Wang et al. 2020)	22.74	0.9175/0.8221	0.9110/0.7566	0.8759/0.5779	0.7807/0.3950
V2X-ViT (Xu et al. 2022a)	21.35	0.9035/0.8119	0.8844/0.7336	0.8283/0.5711	0.7546/0.4266
Where2comm (Hu et al. 2022)	10.74	0.8937/0.7889	0.8880/0.7159	0.8454/0.5439	0.7629/0.3904
CoAlign (Lu et al. 2023)	21.35	0.9132/0.8381	0.9090/0.7682	0.8635/0.5947	0.7810/0.4849
ERMVP (Zhang et al. 2024)	6.34	0.9139/0.8404	0.9085/0.7602	0.8588/0.6632	0.8303/0.5655
MRCNet (Hong et al. 2024)	21.35	0.8775/0.7673	0.8509/0.7365	0.8063/0.6804	0.7330/0.6113
DSRC (Zhang et al. 2025)	21.35	0.9183/0.8526	0.9133/0.7698	0.8663/0.6097	0.7753/0.4738
CoSDH (Xu et al. 2025)	6.65	0.8952/0.8373	0.8589/0.6165	0.6922/0.3685	0.5453/0.2825
MDD (Huang et al. 2025)	32.03	0.8007/0.6817	0.8032/0.6156	0.7757/0.4677	0.7028/0.3394
<b>CoRA</b>	3.80	<b>0.9341/0.8658</b>	<b>0.9297/0.7817</b>	<b>0.8858/0.7199</b>	<b>0.8451/0.6544</b>
DAIR-V2X					
Method/Dataset					
Single	\	0.6250/0.4457	0.6250/0.4457	0.6250/0.4457	0.6250/0.4457
V2VNet (Wang et al. 2020)	11.71	0.6644/0.4037	0.6493/0.3879	0.6273/0.3675	0.5984/0.3508
V2X-ViT (Xu et al. 2022a)	11.42	0.7046/0.5240	0.6959/0.5197	0.6774/0.5119	0.6601/0.5036
Where2comm (Hu et al. 2022)	5.71	0.6735/0.5317	0.6012/0.4493	0.6245/0.3301	0.4588/0.1650
CoAlign (Lu et al. 2023)	11.42	0.7772/0.6284	0.7595/0.5961	0.7274/0.5812	0.7064/0.5752
ERMVP (Zhang et al. 2024)	3.43	0.7042/0.5766	0.6740/0.5546	0.6544/0.5316	0.6228/0.5214
MRCNet (Hong et al. 2024)	11.42	0.6648/0.5388	0.6506/0.5110	0.6148/0.4886	0.5935/0.4777
DSRC (Zhang et al. 2025)	11.42	0.7852/0.6360	0.7698/0.5889	0.7085/0.5512	0.6754/0.5425
CoSDH (Xu et al. 2025)	3.22	0.7675/0.6350	0.7472/0.5756	0.6925/0.5410	0.6607/0.5267
MDD (Huang et al. 2025)	17.13	0.7495/0.5817	0.7201/0.5395	0.6688/0.5092	0.6437/0.4982
<b>CoRA</b>	2.84	<b>0.7858/0.6361</b>	<b>0.7691/0.5977</b>	<b>0.7455/0.5862</b>	<b>0.7312/0.5803</b>

Table 1: Comparison of 3D object detection AP@0.5/AP@0.7 under varying magnitudes of pose error on OPV2V and DAIR-V2X. Pose errors are introduced into the pose information, quantified by a translation standard deviation  $\sigma_t$  [m] / rotation standard deviation  $\sigma_r$  [°]. Comm. represents the average communication reception volume, in MB.

Methods\Delay	0ms	100ms	200ms	300ms	400ms
AP@0.5					
F-cooper	0.7298	0.6288	0.4230	0.3069	0.2746
Where2comm	0.7629	0.6876	0.5585	0.4650	0.4322
CORE	0.6573	0.5443	0.3447	0.2271	0.2077
ERMVP	0.8303	0.6126	0.4219	0.3007	0.2819
<b>Ours</b>	<b>0.8433</b>	<b>0.7828</b>	<b>0.6261</b>	<b>0.5088</b>	<b>0.4620</b>
AP@0.7					
F-cooper	0.3189	0.2319	0.1514	0.1462	0.1579
Where2comm	0.3904	0.3161	0.2333	0.2367	0.2289
CORE	0.2289	0.1447	0.0781	0.0709	0.0822
ERMVP	0.5655	0.2709	0.1705	0.1705	0.1898
<b>Ours</b>	<b>0.6519</b>	<b>0.5683</b>	<b>0.4383</b>	<b>0.3651</b>	<b>0.3519</b>

Table 2: Comparison under varying communication latency with a fixed high pose error (0.6m/0.6°) on OPV2V.

1.4x. This sharply contrasts with the surge of over 22x observed in methods like V2VNet. Our approach offers a significant memory advantage, with usage growing by just 1.72x compared to 4.72x for the next-best method.

### Ablation Studies

Tab. 3 further illustrates the contributions of the core components CIT, LC, and PAC within our framework. Our baseline is a single-agent object detection model with resnet backbone. Under ideal conditions, the integration of the

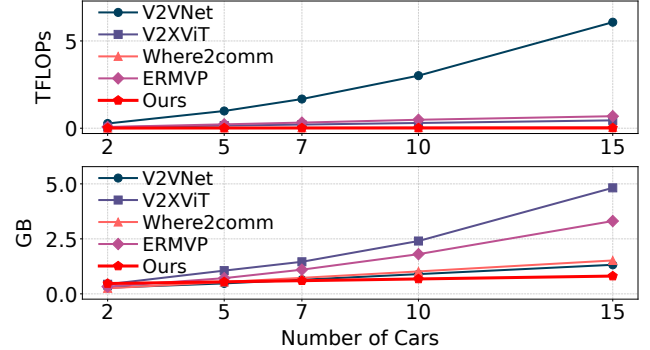


Figure 5: Comparison of resource consumption across models. Note that we only compare their core modules.

lightweight feature-level collaboration modules (CIT and LC) significantly boosts performance over the baseline, with gains of 16.06%/19.02% in AP@0.5/AP@0.7. In the presence of pose errors, PAC offered enhanced robustness at object-level, yielding an increase of 12.05%/14.05% in AP@0.5/AP@0.7 while mitigating the performance degradation typically seen in intermediate fusion.

**Effectiveness of CIT.** Tab. 4 compares the performance of a winner-take-all CIT strategy against MaxOut (Chen et al. 2019) and an alternative Top-2 selection CIT method. The results demonstrate the effectiveness of our strategy.

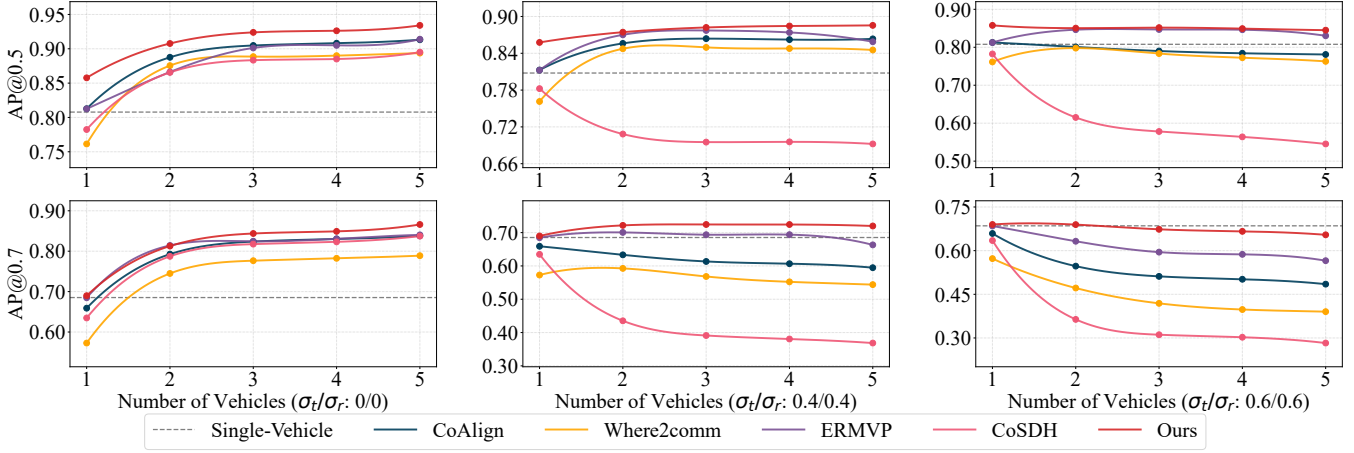


Figure 6: Comparison across varying numbers of collaborators and pose error levels ( $\sigma_t/\sigma_r$ ).

$CIT$	$LC + \mathcal{L}_{align}$	$PAC$	0/0	0.4/0.4
✓			0.6250/0.4457	0.6250/0.4457
✓	✓		0.7749/0.6036	0.6858/0.5184
✓	✓	✓	0.7826/0.6288	0.6933/0.5287
✓			0.7856/0.6359	0.7004/0.5371
✓	✓	✓	0.7786/0.6073	0.7270/0.5733
✓	✓	✓	<b>0.7858/0.6361</b>	<b>0.7455/0.5862</b>

Table 3: Ablation study on DAIR-V2X under ideal/noisy conditions.  $\mathcal{L}_{align}$  denotes the alignment loss introduced within the LC module.

Module / Dataset	OPV2V	DAIR-V2X
MaxOut	0.9120/0.8406	0.7395/0.5610
CIT (Top-1)	<b>0.9341/0.8661</b>	<b>0.7856/0.6359</b>
CIT (Top-2)	0.9325/0.8652	0.7839/0.6211

Table 4: Performance Comparison of Core Modules.

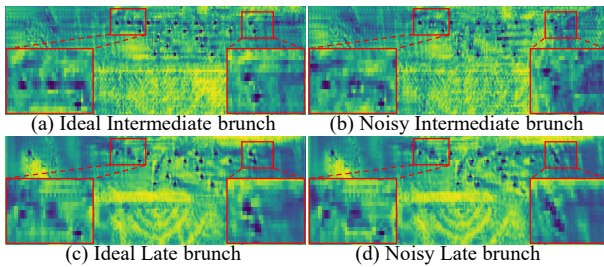


Figure 7: Visualization of both feature and object-level branches of CoRA under ideal and noisy conditions.

### Qualitative Evaluation

Fig. 7 shows the complementary design of our method. Under ideal conditions (a), (c), the feature-level branch yields higher-precision localization. In contrast, in the presence of pose error (b), (d), the features of the feature-level branch suffer from severe artifacts, whereas the object-level branch

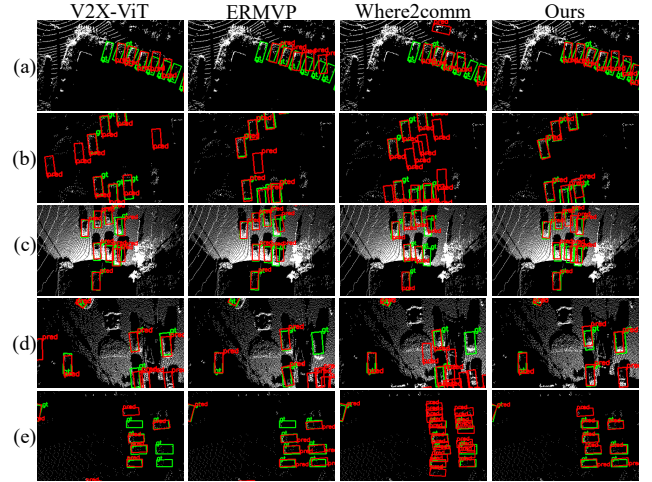


Figure 8: Visualization of detections in diverse scenarios.

demonstrates high robustness, producing a much cleaner and more coherent representation. Fig. 8 provides visualizations of CoRA’s superior detection results in diverse scenarios.

### Conclusion

In this paper, we propose a low-communication and robust collaborative perception method for challenging communication environments. Our approach is distinguished by a novel dual-branch architecture that synergizes intermediate and late fusion. The feature-level branch incorporates a receiver-centric transmission module that ensures near-constant communication overhead, paired with an interaction module for efficient fusion. The object-level branch introduces a novel robust module that corrects misalignments in collaborator detections. Experiments demonstrate that CoRA outperforms current approaches across various scenarios with low communication overhead.

## Acknowledgements

This research is supported in part by the S&T Program of Hebei Province (Beijing-Tianjin-Hebei Collaborative Innovation Special Program) under Grant 25240701D.

## References

- Arnold, E.; Dianati, M.; de Temple, R.; and Fallah, S. 2022. Cooperative perception for 3D object detection in driving scenarios using infrastructure sensors. *IEEE Transactions on Intelligent Transportation Systems (TITS)*, 23(3): 1852–1864.
- Chen, Q.; Ma, X.; Tang, S.; Guo, J.; Yang, Q.; and Fu, S. 2019. 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.
- Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 764–773.
- Gao, X.; Zhang, X.; Lu, Y.; Huang, Y.; Yang, L.; Xiong, Y.; and Liu, P. 2024. A survey of collaborative perception in intelligent vehicles at intersections. *IEEE Transactions on Intelligent Vehicles (TIV)*, 1–20.
- Gu, A.; and Dao, T. 2024. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. In *First Conference on Language Modeling (CoLM)*.
- Hong, S.; Liu, Y.; Li, Z.; Li, S.; and He, Y. 2024. Multi-Agent Collaborative Perception via Motion-Aware Robust Communication Network. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15301–15310.
- Hu, Y.; Fang, S.; Lei, Z.; Zhong, Y.; and Chen, S. 2022. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 4874–4886.
- Hu, Y.; Peng, J.; Liu, S.; Ge, J.; Liu, S.; and Chen, S. 2024. Communication-efficient collaborative perception via information filling with codebook. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15481–15490.
- 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 (CVPR)*, 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 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12697–12705.
- Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; and Liu, Y. 2024. Vmamba: Visual state space model. *Advances in Neural Information Processing Systems (NeurIPS)*, 37: 103031–103063.
- 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.
- Lv, S.; Qin, Y.; Gan, W.; Xu, Z.; and Shi, L. 2024. A systematic literature review of vehicle-to-everything in communication, computation and service scenarios. *International Journal of General Systems*, 53(7-8): 1042–1072.
- Omeiza, D.; Webb, H.; Jirotko, M.; and Kunze, L. 2022. Explanations in autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems (TITS)*, 23(8): 10142–10162.
- Pradeep, A.; Bakoev, M.; and Akhroljonova, N. 2023. A reliability analysis of self-driving vehicles: Evaluating the safety and performance of autonomous driving systems. In *2023 15th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, 1–5. IEEE.
- Su, S.; Li, Y.; He, S.; Han, S.; Feng, C.; Ding, C.; and Miao, F. 2023. Uncertainty quantification of collaborative detection for self-driving. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 5588–5594. IEEE.
- Tang, T.; Zhang, C.; Chen, G.; et al. 2025. RoCooper: Robust Cooperative Perception Under Vehicle-to-Vehicle Communication Impairments. In *IEEE INFOCOM 2025-IEEE Conference on Computer Communications*, 1–10. IEEE.
- Trabelsi, N.; Fourati, L. C.; and Chen, C. S. 2024. Interference management in 5G and beyond networks: A comprehensive survey. *Computer Networks*, 239: 110159.
- Wang, B.; Zhang, L.; Wang, Z.; Zhao, Y.; and Zhou, T. 2023. CORE: Cooperative reconstruction for multi-agent perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 8710–8720.
- 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 605–621. Springer.
- Xu, J.; Zhang, Y.; Cai, Z.; and Huang, D. 2025. CoSDH: Communication-efficient collaborative perception via supply-demand awareness and intermediate-late hybridization. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 6834–6843.
- Xu, R.; Xia, X.; Li, J.; Li, H.; Zhang, S.; Tu, Z.; Meng, Z.; Xiang, H.; Dong, X.; Song, R.; et al. 2023. V2V4Real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13712–13722.
- Xu, R.; Xiang, H.; Tu, Z.; Xia, X.; Yang, M.-H.; and Ma, J. 2022a. V2X-ViT: Vehicle-to-everything cooperative perception with vision transformer. In *Proceedings of the European Conference on Computer Vision (ECCV)*.
- Xu, R.; Xiang, H.; Xia, X.; Han, X.; Li, J.; and Ma, J. 2022b. 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.

Yang, D.; Yang, K.; Wang, Y.; Liu, J.; Xu, Z.; Yin, R.; Zhai, P.; and Zhang, L. 2023. How2Comm: Communication-efficient and collaboration-pragmatic multi-agent perception. *Advances in Neural Information Processing Systems (NeurIPS)*, 36: 25151–25164.

Yazgan, M.; Graf, T.; Liu, M.; Fleck, T.; and Zöllner, J. M. 2024. A survey on intermediate fusion methods for collaborative perception categorized by real world challenges. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, 2226–2233. IEEE.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21361–21370.

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*, 9942–9950.

Zhang, J.; Yang, K.; Wang, Y.; Wang, H.; Sun, P.; and Song, L. 2024. ERMVP: Communication-efficient and collaboration-robust multi-vehicle perception in challenging environments. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12575–12584.