InfoCom: Kilobyte-Scale Communication-Efficient Collaborative Perception with Information Bottleneck

Quanmin Wei[1, 2], **Penglin Dai**[1, 2*], **Wei Li**[1, 2], **Bingyi Liu**[3], **Xiao Wu**[1, 2]

[1] School of Computing and Artificial Intelligence, Southwest Jiaotong University
[2] Engineering Research Center of Sustainable Urban Intelligent Transportation, Ministry of Education
[3] School of Computer Science and Artificial Intelligence, Wuhan University of Technology
wqm@my.swjtu.edu.cn, penglindai@swjtu.edu.cn, liwei@swjtu.edu.cn, byliu@whut.edu.cn, wuxiaohk@gmail.com

Abstract

Precise environmental perception is critical for the reliability of autonomous driving systems. While collaborative perception mitigates the limitations of single-agent perception through information sharing, it encounters a fundamental communication-performance trade-off. Existing communication-efficient approaches typically assume MB-level data transmission per collaboration, which may fail due to practical network constraints. To address these issues, we propose InfoCom, an information-aware framework establishing the pioneering theoretical foundation for communication-efficient collaborative perception via extended Information Bottleneck principles. Departing from mainstream feature manipulation, InfoCom introduces a novel information purification paradigm that theoretically optimizes the extraction of minimal sufficient task-critical information under Information Bottleneck constraints. Its core innovations include: i) An Information-Aware Encoding condensing features into minimal messages while preserving perception-relevant information; ii) A Sparse Mask Generation identifying spatial cues with negligible communication cost; and iii) A Multi-Scale Decoding that progressively recovers perceptual information through mask-guided mechanisms rather than simple feature reconstruction. Comprehensive experiments across multiple datasets demonstrate that InfoCom achieves near-lossless perception while reducing communication overhead from megabyte to kilobyte-scale, representing 440-fold and 90-fold reductions per agent compared to Where2comm and ERMVP, respectively.

Code — https://weiquanmin.github.io/infocom

Introduction

The reliability and safety of modern autonomous driving systems are significantly dependent on precise environmental perception (Hu et al. 2023; Zhao 2024; Liu et al. 2024; Zeng et al. 2024, 2025; Ni et al. 2025; Yao et al. 2025; Chen et al. 2025). Collaborative perception addresses the limitations inherent to single-agent perception by complementary information exchange, thus enhancing perception performance (Han et al. 2023; Gao et al. 2025; Li et al. 2025; Huang et al. 2025). Related strategies have been widely adopted in safety-critical tasks, including object detection (Xia et al. 2025; Zhou et al. 2026), path planning (Qiu et al. 2022), and occupancy prediction (Song et al. 2024).

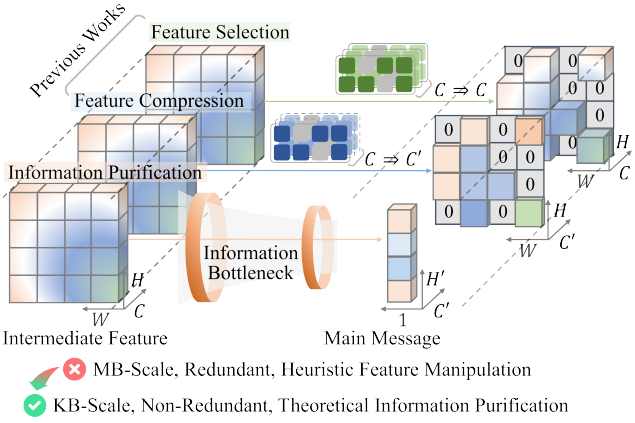


Figure 1: From redundant spatial features to essential information: InfoCom's theoretically grounded information purification ensures minimal sufficient information for perception, enabling KB-scale, near-lossless collaborative perception beyond feature manipulation.

The core trade-off of pragmatic collaboration is balancing perception performance against communication bandwidth consumption (Hu et al. 2024a). Existing communication-efficient approaches primarily fall into two categories: (1) feature selection approaches, which selectively transmit critical features, but suffer from high bandwidth consumption due to their high dimensionality (Hu et al. 2022; Zhao, Zhang, and Zou 2023); and (2) feature compression approaches mapping feature into low-dimensional spaces while preserve spatial structures (Zhang et al. 2024; Hu et al. 2024b). Despite progress, these solutions share significant limitations: they assume MB-level data transmission per collaboration and potentially underestimate practical network constraints. While 5G achieves 3.5 MB/s averages in vehicular scenarios, rates may fluctuate below 0.4 MB/s (Thornton and Dey 2024), risking incomplete perception cycles within an acceptable time (Qiu et al. 2022). More fundamentally, most methods lack a theoretical analysis characterizing the communication-performance trade-off, thereby restrict-

*Corresponding author.

ing their real-world applicability and optimization potential.

Here, we rethink what is a "good" communication-efficient mechanism for addressing the above problems. In particular, Information Bottleneck (IB) provides a mathematical framework for learning a minimal sufficient representation Z from observed data X with target Y, which balances maximization of task-relevant information $I(Z;Y)$ and minimization of redundant information $I(Z;X)$ (Tishby and Zaslavsky 2015). While IB aligns intuitively with our objective, its direct application conflicts with extreme compression requirements. This incompatibility stems from the data processing inequality $I(Z;Y) \leq I(X;Y)$ (Beaudry and Renner 2012), which creates an inherent tension between radical compression and high-precision perception (Tian et al. 2021). Consequently, this limitation reveals that existing feature-based approaches remain constrained by the high-dimensional redundancy of spatial bird's-eye view (BEV) features, compromising task-critical information under practical network constraints.

To address this, we propose a novel information purification paradigm that directly focuses on purifying minimal sufficient information using information-theoretic criteria distinct from feature-space operations. We instantiate this paradigm as *InfoCom*, an information-aware communication-efficient collaborative perception framework comprising three key modules: i) an Information-Aware Encoding module that leverages an extended IB principle to condense and preserve perception-relevant information from intermediate features, generating minimal sufficient messages; ii) a Sparse Mask Generation module that identifies spatial cues critical for collaborative decision-making through efficient filtering and quantization at negligible communication cost; and iii) a Multi-Scale Decoding module that employs mask-guided progressive reconstruction to recover perceptual information rather than simple reconstructing features.

InfoCom delivers three principal advantages: i) *Extreme communication efficiency*: It maintains near-lossless perception performance while requiring only kilobyte-scale (KB) message transmission instead of the current megabyte-level (MB). ii) *Plug-and-play modularity*: Its standardized design enables seamless integration with existing collaborative perception models by replacing communication layers. iii) *Theoretical and empirical co-analysis*: Beyond achieving SOTA empirical performance, it establishes a pioneering theoretical analysis that supports communication-performance trade-offs, fundamentally advancing heuristic approaches.

To validate InfoCom, we conducted extensive experiments on three representative datasets: OPV2V (Xu et al. 2021), V2XSet (Xu et al. 2022), and DAIR-V2X (Yu et al. 2022). Experimental results show that InfoCom surpasses existing feature-based Where2comm (Hu et al. 2022) and ERMVP (Zhang et al. 2024), with 440-fold and 90-fold reduction in communication volume. Meanwhile, InfoCom enhanced the mean AP of collaborative perception models with weaker feature extraction by 1.27% while reducing bandwidth consumption from 34.3 MB to 2.7 KB.

Related Work

Communication-Efficient Collaborative Perception

Collaborative perception enables agents to share complementary information, leading to a more comprehensive understanding of the traffic environment compared to a single perception system (Zimmer et al. 2024; Hu et al. 2025; Xie et al. 2025; Tao et al. 2025). This paper concentrates on the mainstream intermediate collaboration, which involves the aggregation of intermediate features (Xiang et al. 2024; Wei et al. 2025). The core challenge of this paradigm lies in balancing the communication-performance trade-off inherent in multi-agent information exchange (Hu et al. 2022). Existing solutions primarily fall into two categories. *Feature selection approaches* conserve bandwidth by selectively transmitting critical feature segments while discarding nonessential data (Liu et al. 2020a,b; Wang et al. 2023). A representative example is Where2comm (Hu et al. 2022), which employs spatial importance weighting to select key information. *Feature compression approaches* employ encoding techniques to reduce high-dimensional features into compact representations while preserving spatial information for transmission (Hu et al. 2024b). For example, ERMVP (Zhang et al. 2024) achieves state-of-the-art communication efficiency through spatial filtering and clustering.

However, existing methods share two common limitations: i) they typically assume MB-scale bandwidth availability, whereas practical mobile networks often operate significantly below theoretical rates and render this assumption potentially unrealistic; and ii) they lack rigorous theoretical foundations for optimizing collaborative message reduction.

Information Bottleneck

Our approach builds upon the Information Bottleneck (IB) theory (Tishby, Pereira, and Bialek 2000; Tishby and Zaslavsky 2015). The objective of IB is to summarize raw observation X into a compact representation Z while maximally preserving task-relevant information Y. This optimization is formally expressed as:

$$Z = \underset{Z}{\operatorname{argmin}} -I(Z;Y) + \beta I(Z;X), \tag{1}$$

where $I(\cdot;\cdot)$ denotes mutual information and β serves as a Lagrange multiplier that balances information sufficiency $I(Z;Y)$ and minimality $I(Z;X)$. IB principle has been successfully applied in various domains, including representation learning and domain generalization (Yuan et al. 2024; Wang et al. 2025; Wu and Deng 2023). More importantly, its core concept aligns intuitively with our goal.

However, applying the standard IB to collaborative perception encounters a fundamental limitation: its optimization objective merely trades off representational minimality against sufficiency, making it challenging to achieve extreme compression and high-accuracy perception simultaneously (Tian et al. 2021). To address this, we propose two key innovations: i) an extension of the IB principle using an ultra-low-dimensional feature space for preserving critical information under extreme compression, and ii) a mask-guided multi-scale decoding mechanism that ensures near-lossless perceptual accuracy through progressively spatial cues.

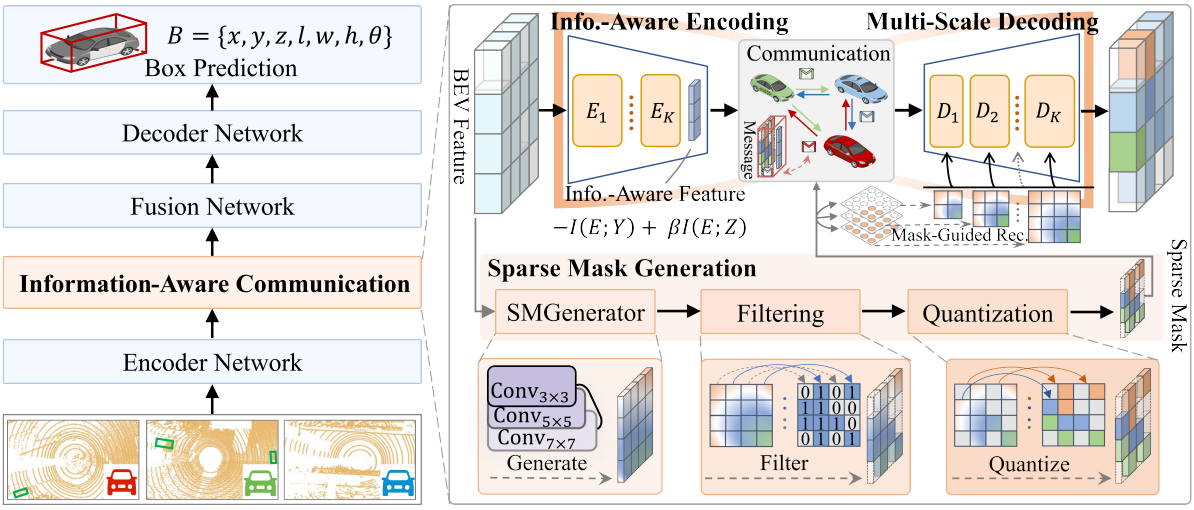


Figure 2: System overview. InfoCom is a communication-efficient collaborative perception framework based on a novel information purification paradigm, consisting of three core modules: (1) Information-Aware Encoding condenses task-critical information from high-dimensional intermediate features into minimal sufficient representations by extending the Information Bottleneck principle; (2) Sparse Mask Generation identifies essential spatial cues with minimal communication overhead; (3) Multi-Scale Decoding progressively recovers perceptual information through mask-guided reconstruction.

Methodology

Method Overview

Problem Formulation. Consider a collaborative perception system comprising N agents. Let X_i denote the raw observation of the i-th agent and Y_i its corresponding label. The objective of communication-efficient collaborative perception is to maximize the multi-agent perception performance under limited communication costs, that is,

$$\begin{aligned} \underset{\theta,\ \{\mathcal{P}_{j\to i}\}_{1\le i,j\le N, j\neq i}}{\text{argmax}} & \sum_{i=1}^{N} m\big(f_\theta(X_i, \{\mathcal{P}_{j\to i}\}),\ Y_i\big), \\ \text{s.t.} & \sum_{i=1}^{N}\sum_{j=1, j\neq i}^{N} c(\mathcal{P}_{j\to i}) \le B, \end{aligned} \tag{2}$$

where $\mathcal{P}_{j\to i}$ denotes the learnable message transmitted from agent j to i, f_θ is the collaborative perception model parameterized by θ, m is the perception performance metric (here focusing on 3D object detection), function c quantifies bandwidth consumption, and B denotes the system-level communication budget. In contrast to existing work that typically requires MB, our objective is to maximize the sufficiency of the message $\mathcal{P}$. while maintaining $\frac{B}{N(N-1)}$ at the KB scale.

Overall Pipeline. The overall workflow of InfoCom is illustrated in Fig. 2. To maintain compatibility with existing collaborative systems, our solution only requires replacing the communication layer and incorporating an Information Bottleneck regularizer into the training loss. Specifically, each agent extracts intermediate BEV features from environmental readings using a local encoder network and transforms them into a unified coordinate system with pre-shared pose. Subsequently, a customized Information-Aware Communication mechanism compresses messages at the transmitter and decompresses them at the receiver. This mechanism comprises: i) an Information-Aware Encoding that extracts minimal sufficient messages, ii) a Sparse Mask Generation that identifies spatial cues, and iii) a Multi-Scale Decoding that reconstructs processable features. Finally, the receiving agent aggregates multi-view features via a fusion network to generate 3D object detection results.

Information-Aware Communication

Information-Aware Encoding. We introduce an Information-Aware Encoding (IAE) module based on the extended IB principle. To resolve IB's compression-performance dilemma, we first extend the standard Markov Chain from $Y \to X \to Z$ to $Y \to X \to Z \to (E, M)$ where information-aware feature $E \in \mathbb{R}^{N\times D}$ satisfies $D \ll C \times H \times W$ and $Z \in \mathbb{R}^{N\times C\times H\times W}$ is intermediate feature. This extension creates a low-dimensional space for extreme compression while decoupling spatial cues as auxiliary information M. Accordingly, we reformulate Eq. (1) to derive the new IB objective as follows:

$$E, M = \underset{E,M}{\text{argmin}} -I(E, M; Y) + \beta I(E, M; Z). \tag{3}$$

In collaborative systems, the mapping from X to Z is handled by a fixed encoder. For the Information-Aware Encoding process $Z \to E$, we follow existing work to derive a tractable variational approximation (Alemi et al. 2017; Fu et al. 2025), which are instantiated via the proposed Information-Aware Encoder (IAEncoder). Specifically, for the i-th agent, the encoding process of IAE is represented as

$$\begin{aligned} (\mu_i, \sigma_i) &= \text{IAEncoder}(Z_i), \\ E_i &= \mu_i + \sigma_i \odot \epsilon_i,\ \epsilon_i \sim \mathcal{N}(0, I). \end{aligned} \tag{4}$$

The IB principle governs E as minimal sufficient for perception tasks, which achieves information-aware processing. Under a $\mathcal{N}(0, I)$ prior, IAEncoder generates Gaussian parameters μ and σ rather than complete E. This enables closed-form computation of the KL divergence for $I(E_i; Z_i)$ during training and straightforward implementation of the reparameterization trick for stochasticity isolation. Regarding network architecture, IAEncoder comprises three residual-like blocks designed to accommodate resource-constrained agents while ensuring robust feature extraction. Further details are provided in the Appendix.

Sparse Mask Generation. As an essential component of InfoCom, this module compensates for critical spatial priors in perception tasks with minimal communication overhead and thereby mitigates information loss under extreme compression. According to the data processing inequality, a higher dimensional compression ratio $D/(C \times H \times W)$ implies a greater risk of task-relevant information loss. Conversely, extreme compression of E frees bandwidth for transmitting auxiliary information. To this end, we introduce a Sparse Mask Generation that expands the communication unit on the sender side from solely $\mathcal{P}_i = \{E_i\}$ to $\mathcal{P}_i = \{E_i, M_i\}$. The spatial importance mask $M_i \in \mathbb{R}^{H \times W}$ is generated by the Sparse Mask Generator (SMGenerator in short) through multi-scale feature extraction, preserving perceptual cues across varying granularities:

$$M_i = W_{\text{proj}}([\text{Conv}_{s\times s}(Z_i)|s \in \{3, 5, 7\}] + Z_i), \quad (5)$$

where W_{proj} is projection layer and $[\cdot, \cdot]$ represents channel concatenation.

We note that the initial mask M_i suffers from high-entropy redundancy issue, as it is neither sparse nor compressed. Therefore, we propose a joint compression post-processing comprising two steps: filtering and quantization. In the filtering stage, our empirical observations (see Fig. 4b) indicate that only a minimal number of spatial cues benefit the task ; thus, we retain only the top-k critical positions:

$$M_i^{\text{s}} = \text{TopK}(M_i, k), \quad k = \lfloor \alpha \cdot HW \rfloor, \quad (6)$$

where α denotes retention ratio and is empirically set to 0.1.

In the quantization stage, we demonstrate that spatial cues remain effective without high-precision (see Fig. 4c):

$$M_i^{\text{q}} = \text{Clamp}\Big(\text{Round}\big(\frac{M_i^{\text{s}}}{\delta}\big), 0, 2^b - 1\Big), \ \delta = \frac{1}{2^b - 1}, \quad (7)$$

where b is uniform quantization bit-width, defaulting to 4.

Finally, the non-differentiability of this post-processing is resolved using the straight-through estimator like (Bengio, Léonard, and Courville 2013), $\frac{\partial \mathcal{L}}{\partial M_i} \approx \frac{\partial \mathcal{L}}{\partial M_i^{\text{q}}}$. The sparse mask M_i^{q} achieves efficient compression via extreme sparsity and low-precision representation. When integrated with the Multi-Scale Decoding, it delivers spatial priors at negligible communication cost while significantly mitigating information loss in E.

Multi-Scale Decoding. At the receiver k, the efficient reconstruction of actionable features from message units $\mathcal{P}_i = \{E_i, M_i^{\text{q}}\}, i \in \{1, \ldots, N\} \setminus \{k\}$, is a pivotal step for translating communication efficiency into perception performance. To this end, we propose Multi-Scale Decoding (MSD), which leverages the highly compressed information-aware feature E and the sparse mask M^{q} to reconstruct BEV feature progressively with a focus on perceptual information. The MSD comprises three core steps. [1]

Feature Initialization. The E is expanded into a lower-resolution initial feature map $F_{\text{init}}^0 \in \mathbb{R}^{C^0 \times H^0 \times W^0}$ via fully connected and transposed convolutional layers, where $C^0 > C$, $H^0 < H$, and $W^0 < W$. This establishes the foundation for subsequent spatial reconstruction.

Mask-Guided Modulation. Taking F_{init}^0 as the target, the dequantization mask $M^{\text{q}} \cdot \delta \in \mathbb{R}^{H \times W}$ is downsampled to the resolution $H^0 \times W^0$ via a convolutional layer, yielding M^0. The features are then modulated by the mask as $F^0 = F_{\text{init}}^0 \odot M^0$. This modulation directs the subsequent progressive reconstruction toward task-critical regions for enhancing perceptual information recovery.

Multi-Scale Reconstruction. Building upon mask-guided modulation, this phase employs cascaded decoding blocks with multi-scale masks for progressive upsampling. Let F^i denote the output of the i-th block, with resolution $C^i \times H^i \times W^i$ that satisfies $2C^i = C^{i-1}$, $H^i = 2H^{i-1}$, and $W^i = 2W^{i-1}$. After K iterations, the feature map F^K reaches the target resolution $C \times H \times W$ of intermediate feature Z.

Multi-Scale Decoding processes all messages to actionable BEV features $F_{1:N} = \{F_i^K \mid i \in \{1, \ldots, N\} \setminus \{k\}\}$ at the k-th agent. Finally, the resulting set $F_{1:N} \cup \{Z_k\}$ is fed directly into a standard fusion network to aggregate multi-source information, ensuring compatibility with collaborative perception systems.

Overall Loss

To enable end-to-end training, we follow existing works by decomposing the IB objective in Eq. (3) into a supervised loss and a regularization term (Wu and Deng 2023; Chen et al. 2023). The overall loss function is shown as follows:

$$\mathcal{L} = \mathcal{L}_{\text{detect}} + \beta \text{KL}\big(p(E|Z)\big\|r(E)\big), \quad (8)$$

where $\mathcal{L}_{\text{detect}}$ denotes the standard detection loss consistent with existing collaborative perception models (Lu et al. 2023), the $\text{KL}(\cdot)$ divergence term instantiates the IB regularization with $p(E|Z)$ representing the stochastic mapping $Z \to E$ parameterized by the IAEncoder. Under Gaussian prior assumption $r(E) = \mathcal{N}(0, I)$ for variable E, this regularization term admits a closed-form solution.

Theoretical Analysis

Here, we provide a theoretical analysis, where Lem. 1 establishes the noise bound for the compression process, and Prop. 1 demonstrates that InfoCom achieves communication efficiency, information retention, and noise suppression.

Lemma 1. ***(Noise Suppression Bound of Collaborative Features)*** *Consider an agent's observation X, intermediate feature Z, message unit $\mathcal{P} = \{E, M\}$ consisting of the information-aware feature E and the sparse mask M, and*

[1] For notational simplicity, agent ID subscripts are omitted.

Dataset	Comm. Method	Comm. Volume	AP@30	AP@50	AP@70
OPV2V (Xu et al. 2021)	No Collaboration	0	0.8665	0.8464	0.7288
	Late Collaboration	6.250 KB	0.9622	0.9499	0.8495
	Standard Colla. (Lu et al. 2023)	34.375 MB	0.9709	0.9653	0.9229
	Where2comm (Hu et al. 2022)	3.439 MB	0.9548	0.9463	0.8820
	ERMVP (Zhang et al. 2024)	0.741 MB	0.9618	0.9557	0.9127
	InfoCom (Ours)	**7.875 KB**	**0.9702**	**0.9650**	**0.9202**
V2XSet (Xu et al. 2022)	No Collaboration	0	0.8032	0.7719	0.6222
	Late Collaboration	6.250 KB	0.9321	0.9076	0.7120
	Standard Colla. (Lu et al. 2023)	34.375 MB	0.9317	0.9212	0.8426
	Where2comm (Hu et al. 2022)	3.439 MB	0.8834	0.8604	0.7417
	ERMVP (Zhang et al. 2024)	/	OOM	OOM	OOM
	InfoCom (Ours)	**7.875 KB**	**0.9360**	**0.9273**	**0.8488**
DAIR-V2X (Yu et al. 2022)	No Collaboration	0	0.7278	0.6844	0.5665
	Late Collaboration	6.250 KB	0.7993	0.6709	0.4708
	Standard Colla. (Lu et al. 2023)	24.609 MB	0.8294	0.7843	0.6353
	Where2comm (Hu et al. 2022)	2.462 MB	0.8048	0.7539	0.6070
	ERMVP (Zhang et al. 2024)	0.531 MB	0.8217	**0.7791**	0.6324
	InfoCom (Ours)	**5.922 KB**	**0.8228**	0.7789	**0.6385**

Table 1: Performance comparison on three representative collaborative perception datasets. All communication-efficient methods are built upon CoAlign, with the best and second-best results highlighted in bold and underlined, respectively.

the perceptual target Y with task-irrelevant noise Y_N satisfying $Y \perp Y_N$. Under the Markov chain $(Y, Y_N) \to X \to Z \to (E, M)$, the following inequality holds:

$$I(E, M; Y_N) \leq I(E, M; Z) - I(E, M; Y). \quad (9)$$

Furthermore, due to the entropy constraint of M,

$$I(E, M; Z) \leq I(E; Z) + H(M) \leq I(E; Z) + \log \binom{HW}{k} + k\, b, \quad (10)$$

where $k = \lfloor \alpha \cdot HW \rfloor$ is the number of retained spatial positions and b is the quantization bit width.

Proof. Since (E, M) is a function of Z, the data processing inequality applied to the Markov chain $(Y, Y_N) \to Z \to (E, M)$ gives $I(E, M; Z) \geq I(E, M; Y, Y_N)$. Decomposing the mutual information yields $I(E, M; Y, Y_N) = I(E, M; Y_N) + I(E, M; Y|Y_N)$. Since $Y \perp Y_N$, we have $H(Y|Y_N) = H(Y)$, and thus $I(E, M; Y|Y_N) \geq I(E, M; Y)$. Substituting and rearranging gives $I(E, M; Y_N) \leq I(E, M; Z) - I(E, M; Y)$. Finally, decomposing $I(E, M; Z) = I(E; Z) + I(M; Z|E)$ and applying $I(M; Z|E) \leq H(M) \leq \log \binom{HW}{k} + kb$ establishes the stated bound. □

Proposition 1. ***(Theoretical Foundations of Information-Aware Communication in Collaborative Perception)*** *The InfoCom achieves communication-efficient collaborative perception through three interconnected mechanisms with theoretical foundations: (1) bandwidth reduction via filtering and quantization, (2) preservation of task-relevant information, and (3) suppression of task-irrelevant noise.*

Proof. (1) *Bandwidth reduction:* The $\mathcal{P} = \{E, M\}$ achieves substantial bandwidth reduction through explicit design choices: $E \in \mathbb{R}^D$ with $D \ll C \times H \times W$ enables dimension reduction, while the sparse mask M selects only $k \ll H \times W$ positions with b-bit quantization. Lem. 1 provides the information-theoretic foundation through the bound (10), which constrains the fundamental entropy of $\mathcal{P}$.

(2) *Perceptual preservation:* The detection loss $\mathcal{L}_{\text{detect}}$ implicitly maximizes $I(E, M; Y)$, ensuring preservation of task-relevant information for perceptual accuracy.

(3) *Noise suppression:* From Lem. 1, $I(E, M; Y_N) \leq I(E, M; Z)$–$I(E, M; Y)$; thus, minimizing $I(E, M; Z)$ via IB regularization and maximizing $I(E, M; Y)$ via detection loss jointly suppress noise. Furthermore, quantization processes provide additional regularization through the data processing inequality, enhancing robustness. □

Experiments

Experimental Settings

Datasets and Evaluation Metrics. We follow existing work (Wei et al. 2025) and conduct experiments on three representative collaborative perception datasets: the simulated OPV2V (Xu et al. 2021) and V2XSet (Xu et al. 2022), and the real-world DAIR-V2X (Yu et al. 2022). For DAIR-V2X, we follow standard practices to extend annotation coverage (Lu et al. 2023). 3D object detection performance is evaluated using Average Precision (AP) at Intersection over Union (IoU) thresholds of 0.3, 0.5, and 0.7. Communication volume is measured in human-readable units to estimate actual transmission requirements per agent.

Baselines. InfoCom's communication efficiency is validated against SOTA feature-based alternatives, Where2comm (Hu et al. 2022) and ERMVP (Zhang

		AttFuse (Xu et al. 2021)			MKD-Cooper (Li et al. 2024)		
Comm. Method	Comm. Volume	AP@30	AP@50	AP@70	AP@30	AP@50	AP@70
Standard Colla.	34.375 MB	0.9546	0.9327	0.8039	0.9575	0.9360	0.8153
Where2comm	3.438 MB	0.9173	0.9004	0.8023	0.9134	0.8974	0.7957
ERMVP	0.701 MB	0.9212	0.9041	0.7917	0.9112	0.8946	0.7848
InfoCom (Ours)	**2.718 KB**	**0.9575**	**0.9360**	**0.8153**	**0.9640**	**0.9490**	**0.8340**

Table 2: Evaluation of communication-efficient methods using alternative collaborative perception models on OPV2V dataset.

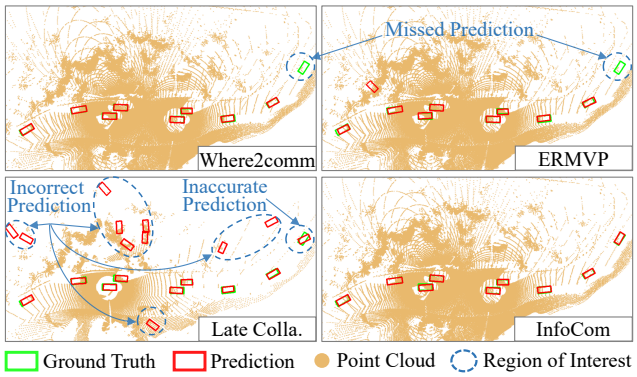


Figure 3: Qualitative comparison on OPV2V dataset.

et al. 2024). For fairness, all non-communication components and customizable settings of the base collaborative perception model remain fixed. The default base employs the multi-scale CoAlign framework (Lu et al. 2023), while single-scale models, AttFuse (Xu et al. 2021) and MKD-Cooper (Li et al. 2024), validate cross-model applicability. We concurrently report performance for standard intermediate, late, and no collaboration as baseline references. More experiment details are provided in the Appendix.

Main Results

Comparison of Communication-Efficient Methods. As summarized in Tab. 1, experimental results reveal three superiorities inherent to InfoCom. i) *Exceptional communication efficiency*: InfoCom requires only kilobyte-level communication volume, comparable to Late Collaboration but significantly lower than other feature-based solutions. Specifically, its bandwidth consumption is over 400 times lower than Where2comm, only 1% of that of ERMVP, and over 4000 times lower than Standard Collaboration. ii) *Superior perception performance*: InfoCom maintains perception performance on par with the bandwidth-intensive Standard Collaboration despite minimal communication overhead, while significantly outperforms Where2comm. Moreover, ERMVP exhibits the smallest performance gap relative to InfoCom. iii) *Optimal communication-performance trade-off*: InfoCom demonstrates state-of-the-art efficiency in performance gain per unit bandwidth. For example, on the OPV2V dataset, InfoCom achieves 1.8×10^{-2} average performance gain per kilobyte, substantially exceeding Where2comm (3.2×10^{-5}) and ERMVP (1.7×10^{-4}).

ID	Component	Variant	Mean AP
1	/	InfoCom (Full)	0.9518
2	IAE	Simple Encoder	0.9320
3	SMG	Simple Generator	0.9379
4		w/o STE	0.8845
5	MSD	w/o Mask	0.8839
6		w/o Multi-Scale Rec.	0.9439

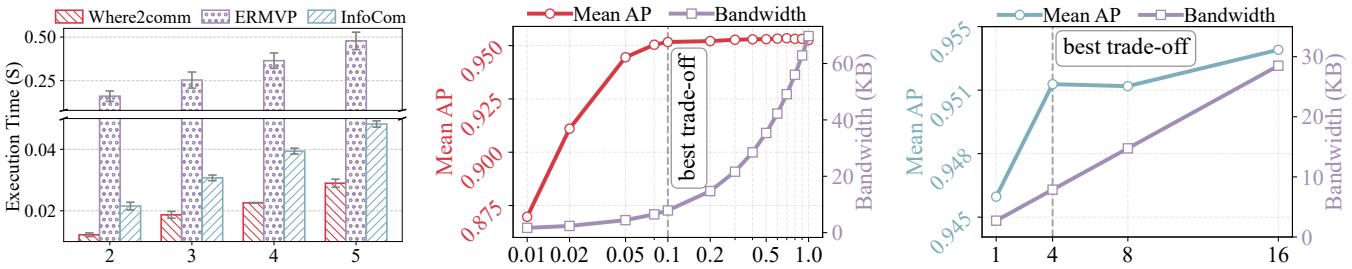
Table 3: Effects of the different components on InfoCom.

Validation of Different Collaborative Models. To evaluate compatibility, we integrated communication-efficient methods into the base AttFuse and MKD-Cooper models, with results presented in Tab. 2. The findings reaffirm core conclusions in Tab. 1 while also revealing a notable performance enhancement. For example, the InfoCom-integrated MKD-Cooper variant surpassed the original model by 1.27% in mean AP. Although AttFuse and MKD-Cooper generate suboptimal single-scale intermediate features compared to CoAlign's multi-scale representations, InfoCom's information purification mechanism enhances feature quality by simultaneously suppressing noise interference and extracting task-critical information. This approach effectively compensates for the feature constraints inherent in weaker collaborative perception backbones.

Qualitative Analysis. The visualization results in Fig. 3 further demonstrate that InfoCom achieves better 3D object detection. Specifically, it outperforms existing methods in prediction accuracy, localization precision, and false positive suppression. These intuitive results are highly consistent with the quantitative analysis presented earlier.

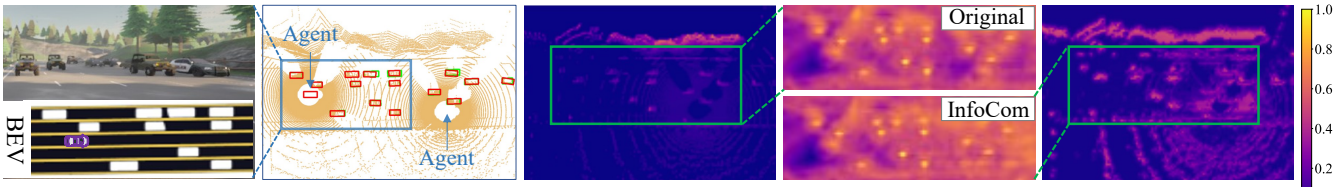
Deeper Analysis

Runtime Analysis. Using identical hardware configurations with a 3090 GPU, we evaluated additional execution time across varying agent densities on the OPV2V. The resultant time is reported in Fig. 4a, derived from 20 experimental trials. The results show that Where2comm achieves the shortest computation time, approximately half of InfoCom, whereas ERMVP incurs the most substantial computational overhead. Note that total system latency consists of both data transmission time and computation time. Where2comm's MB-level data transmission demands substantially offset its computational advantages. Conversely,



(a) Execution time w.r.t agent density (b) Trade-off w.r.t retention ratio α (c) Trade-off w.r.t quantization bit-width b

Figure 4: Runtime and trade-off analysis for InfoCom.



(a) Image (b) Detection result (c) Original feature map (d) Prediction head (e) InfoCom feature map

Figure 5: Deeper visualization analysis for InfoCom.

InfoCom's KB-level communication overhead demonstrates significant potential for overall latency optimization.

Validation of Spatial Cue Sparsity. By progressively adjusting the retention ratio of sparse mask M in Fig. 4b, we empirically demonstrate that transmitting more than 10% of spatial cues yields only marginal performance gains ($<$ 0.2% AP on average). This results from the inherent sparsity of point cloud data. Accordingly, controlling the sparsity of bandwidth-dominant M is essential for optimizing InfoCom's communication efficiency.

Feasibility of Low-Precision Representation. Fig. 4c reveals redundancy in high-precision floating-point representations of spatial cues. Experiments demonstrate that quantizing M to 4-bit precision reduces communication overhead by a factor of 8 while slightly affecting perception performance (AP fluctuation $< 0.18\%$). Critically, InfoCom's uniform quantization strategy incurs negligible computational costs and additional burden.

Feature-Level Comparative Visualization. Fig. 5 visualizes feature disparities in a two-agent scenario, depicting: (a) the perspective of the right agent from Fig. 5b, (b) collaborative perception results, and (c-e) feature comparisons between standard collaboration and InfoCom, with yellow regions indicating higher activation values. InfoCom's task-driven information purification mechanism fundamentally redefines information transmission by conveying KB-scale critical perception data rather than spatially structured key features. This approach eliminates explicit feature-level alignment constraints, resulting in significant feature map disparities evident in Figs. 5c and 5e, yet maintains uncompromised perception performance. As shown in Fig. 5d, both prediction heads exhibit consistent response intensities in target regions while equally suppressing background.

Ablation Study

Tab. 3 summarizes the ablation study on InfoCom's three key components: IAE, SMG, and MSD. The first row, representing the complete framework, serves as the baseline. Rows 2, 3, and 5 evaluate simplified implementations of individual components. These results demonstrate a positive contribution for each key element. Specifically, rows 2 and 3 indicate that the simplified variants of IAE and SMG underperform relative to the complete modules. Row 4 reveals that the non-differentiability introduced by quantization and filtering operations during joint compression post-processing necessitates an appropriate gradient estimation technique. Row 5 indicates that spatial cues effectively mitigate information loss under aggressive compression. Finally, row 6 shows that replacing Multi-Scale Decoding with single-masked reconstruction in the final stage leads to marginal performance degradation.

Conclusion

We propose InfoCom, a novel communication-efficient framework for collaborative perception. Compared to existing methods, it delivers three distinct advantages: reducing communication volume from megabytes to kilobytes by condensing perception-critical information rather than manipulating redundant spatial features; providing theoretical analysis based on information principles to address the empirical limitations of heuristic designs; and enabling plug-and-play functionality via a standardized modular architecture. These advantages originate from three complementary innovations: Information-Aware Encoding, Sparse Mask Generation, and Multi-Scale Decoding. Comprehensive evaluations on OPV2V, V2XSet, and DAIR-V2X datasets consistently demonstrate superior communication efficiency and perception performance.

Ethical Statement

We do not foresee ethical concerns posed by our method, but concede that both ethical and unethical applications of autonomous driving techniques may benefit from the improvements induced by our work. Care must be taken, in general, to ensure positive ethical and societal consequences of autonomous driving.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62172342, Grant 62372387; Key R&D Program of Guangxi Zhuang Autonomous Region, China (Grant No. AB22080038, AB22080039); The Open Fund of the Engineering Research Center of Sustainable Urban Intelligent Transportation, Ministry of Education, China (Project No. KCX2024-KF07).

References

Alemi, A. A.; Fischer, I.; Dillon, J. V.; and Murphy, K. 2017. Deep Variational Information Bottleneck. In *International Conference on Learning Representations (ICLR)*.

Beaudry, N. J.; and Renner, R. 2012. An intuitive proof of the data processing inequality. *Quantum Information & Computation*, 12(5-6): 432–441.

Bengio, Y.; Léonard, N.; and Courville, A. 2013. Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation. *arXiv preprint arXiv:1308.3432*.

Chen, J.; Deng, W.; Peng, B.; Liu, T.; Wei, Y.; and Liu, L. 2023. Variational Information Bottleneck for Cross Domain Object Detection. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2231–2236.

Chen, Z.; Yang, J.; Chen, L.; Li, F.; Feng, Z.; Jia, L.; and Li, P. 2025. RailVoxelDet: An Lightweight 3D Object Detection Method for Railway Transportation Driven by on-Board LiDAR Data. *IEEE Internet of Things Journal*.

Fu, X.; Gao, Y.; Yang, B.; Wu, Y.; Qian, H.; Sun, Q.; and Li, X. 2025. Bi-Directional Multi-Scale Graph Dataset Condensation via Information Bottleneck. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, 16674–16681.

Gao, X.; Xu, R.; Li, J.; Wang, Z.; Fan, Z.; and Tu, Z. 2025. STAMP: Scalable Task-And Model-agnostic Collaborative Perception. In *International Conference on Learning Representations (ICLR)*.

Han, Y.; Zhang, H.; Li, H.; Jin, Y.; Lang, C.; and Li, Y. 2023. Collaborative Perception in Autonomous Driving: Methods, Datasets, and Challenges. *IEEE Intelligent Transportation Systems Magazine*, 15: 131–151.

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 *AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, 23203–23211.

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.; Pang, X.; Qin, X.; Eldar, Y. C.; Chen, S.; Zhang, P.; and Zhang, W. 2024a. Pragmatic Communication in Multi-Agent Collaborative Perception. *arXiv preprint arXiv:2401.12694*.

Hu, Y.; Peng, J.; Liu, S.; Ge, J.; Liu, S.; and Chen, S. 2024b. Communication-Efficient Collaborative Perception via Information Filling with Codebook. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15481–15490.

Hu, Y.; Yang, J.; Chen, L.; Li, K.; Sima, C.; Zhu, X.; Chai, S.; Du, S.; Lin, T.; Wang, W.; et al. 2023. Planning-oriented Autonomous Driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17853–17862.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 27390–27400.

Li, J.; Liu, X.; Li, B.; Xu, R.; Li, J.; Yu, H.; and Tu, Z. 2025. CoMamba: Real-time Cooperative Perception Unlocked with State Space Models. In *International Conference on Intelligent Robots and Systems (IROS)*.

Li, Z.; Liang, H.; Wang, H.; Zhao, M.; Wang, J.; and Zheng, X. 2024. MKD-Cooper: Cooperative 3D Object Detection for Autonomous Driving via Multi-teacher Knowledge Distillation. *IEEE Transactions on Intelligent Vehicles*, 1490–1500.

Liu, Y.; Gan, M.; Zeng, H.; Liu, L.; Dong, Y.; and Cao, Z. 2024. Hydra: Accurate Multi-Modal Leaf Wetness Sensing with mm-Wave and Camera Fusion. In *International Conference on Mobile Computing and Networking*, 800–814.

Liu, Y.-C.; Tian, J.; Glaser, N.; and Kira, Z. 2020a. When2com: Multi-Agent Perception via Communication Graph Grouping. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4106–4115.

Liu, Y.-C.; Tian, J.; Ma, C.-Y.; Glaser, N.; Kuo, C.-W.; and Kira, Z. 2020b. Who2com: Collaborative Perception via Learnable Handshake Communication. In *IEEE International Conference on Robotics and Automation (ICRA)*, 6876–6883.

Lu, Y.; Li, Q.; Liu, B.; Dianat, M.; Feng, C.; Chen, S.; and Wang, Y. 2023. Robust Collaborative 3D Object Detection in Presence of Pose Errors. In *IEEE International Conference on Robotics and Automation (ICRA)*, 4812–4818.

Ni, C.; Zhao, G.; Wang, X.; Zhu, Z.; Qin, W.; Huang, G.; Liu, C.; Chen, Y.; Wang, Y.; Zhang, X.; et al. 2025. ReconDreamer: Crafting World Models for Driving Scene Reconstruction via Online Restoration. In *Computer Vision and Pattern Recognition Conference*, 1559–1569.

Qiu, H.; Huang, P.-H.; Asavisanu, N.; Liu, X.; Psounis, K.; and Govindan, R. 2022. AutoCast: Scalable Infrastructure-less Cooperative Perception for Distributed Collaborative Driving. In *Annual International Conference on Mobile Systems, Applications and Services (ACM MobiSys)*, 128–141.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17996–18006.

Tao, Y.; Hu, S.; Fang, Z.; and Fang, Y. 2025. Directed-CP: Directed Collaborative Perception for Connected and Autonomous Vehicles via Proactive Attention. In *IEEE International Conference on Robotics and Automation (ICRA)*.

Thornton, S.; and Dey, S. 2024. Multi-Modal Data and Model Reduction for Enabling Edge Fusion in Connected Vehicle Environments. *IEEE Transactions on Vehicular Technology*.

Tian, X.; Zhang, Z.; Lin, S.; Qu, Y.; Xie, Y.; and Ma, L. 2021. Farewell to Mutual Information: Variational Distillation for Cross-Modal Person Re-Identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1522–1531.

Tishby, N.; Pereira, F. C.; and Bialek, W. 2000. The Information Bottleneck Method. *arXiv preprint physics/0004057*.

Tishby, N.; and Zaslavsky, N. 2015. Deep Learning and the Information Bottleneck Principle. In *IEEE Information Theory Workshop (ITW)*, 1–5.

Wang, B.; Zhang, L.; Wang, Z.; Zhao, Y.; and Zhou, T. 2023. Core: Cooperative Reconstruction for Multi-Agent Perception. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 8710–8720.

Wang, J.; Yang, L.; Wei, Y.; Si, J.; Guo, C.; Sun, Q.; Li, X.; and Fu, X. 2025. An Out-Of-Distribution Membership Inference Attack Approach for Cross-Domain Graph Attacks. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Wei, Q.; Dai, P.; Li, W.; Liu, B.; and Wu, X. 2025. CoPEFT: Fast Adaptation Framework for Multi-Agent Collaborative Perception with Parameter-Efficient Fine-Tuning. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, 23351–23359.

Wu, A.; and Deng, C. 2023. TIB: Detecting Unknown Objects Via Two-Stream Information Bottleneck. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1): 611–625.

Xia, Q.; Lin, W.; Xiang, H.; Huang, X.; Chen, S.; Dong, Z.; Wang, C.; and Wen, C. 2025. Learning to Detect Objects from Multi-Agent LiDAR Scans without Manual Labels. In *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 1418–1428.

Xiang, H.; Zheng, Z.; Xia, X.; Xu, R.; Gao, L.; Zhou, Z.; Han, X.; Ji, X.; Li, M.; Meng, Z.; et al. 2024. V2X-Real: A Largs-Scale Dataset for Vehicle-to-Everything Cooperative Perception. In *European Conference on Computer Vision (ECCV)*, 455–470.

Xie, Q.; Zhou, X.; Hong, T.; Hu, W.; Qu, W.; and Qiu, T. 2025. Towards Communication-Efficient Cooperative Perception via Planning-Oriented Feature Sharing. *IEEE Transactions on Mobile Computing*, 24(4): 2551–2563.

Xu, R.; Xiang, H.; Tu, Z.; Xia, X.; Yang, M.-H.; and Ma, J. 2022. V2X-ViT: Vehicle-to-Everything Cooperative Perception with Vision Transformer. In *European Conference on Computer Vision (ECCV)*, 107–124.

Xu, R.; Xiang, H.; Xia, X.; Han, X.; Liu, J.; and Ma, J. 2021. OPV2V: An Open Benchmark Dataset and Fusion Pipeline for Perception with Vehicle-to-Vehicle Communication. In *International Conference on Robotics and Automation (ICRA)*, 2583–2589.

Yao, S.; Guan, R.; Peng, Z.; Xu, C.; Shi, Y.; Yue, Y.; Lim, E. G.; Seo, H.; Man, K. L.; Zhu, X.; et al. 2025. Exploring Radar Data Representations in Autonomous Driving: A Comprehensive Review. *IEEE Transactions on Intelligent Transportation Systems*, 26(6): 7401–7425.

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

Yuan, H.; Sun, Q.; Fu, X.; Ji, C.; and Li, J. 2024. Dynamic Graph Information Bottleneck. In *ACM Web Conference (WWW)*, 469–480.

Zeng, S.; Chang, X.; Liu, X.; Pan, Z.; and Wei, X. 2024. Driving with Prior Maps: Unified Vector Prior Encoding for Autonomous Vehicle Mapping. *arXiv preprint arXiv:2409.05352*.

Zeng, S.; Chang, X.; Xie, M.; Liu, X.; Bai, Y.; Pan, Z.; Xu, M.; and Wei, X. 2025. FutureSightDrive: Thinking Visually with Spatio-Temporal CoT for Autonomous Driving. *arXiv preprint arXiv:2505.17685*.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12575–12584.

Zhao, B.; Zhang, W.; and Zou, Z. 2023. BM2CP: Efficient Collaborative Perception with LiDAR-Camera Modalities. In *Conference on Robot Learning (CoRL)*, 1022–1035.

Zhao, Z. 2024. BALF: Simple and Efficient Blur Aware Local Feature Detector. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 3362–3372.

Zhou, J.; Dai, P.; Wei, Q.; Liu, B.; Wu, X.; and Wang, J. 2026. Pragmatic Heterogeneous Collaborative Perception via Generative Communication Mechanism. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zimmer, W.; Wardana, G. A.; Sritharan, S.; Zhou, X.; Song, R.; and Knoll, A. C. 2024. TUMTraf V2X Cooperative Perception Dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22668–22677.