RATE-DISTORTION OPTIMIZED PRAGMATIC COMMUNICATION FOR COLLABORATIVE PERCEPTION

Genjia Liu[1], Anning Hu[1], Yue Hu[2], Wenjun Zhang[1], Siheng Chen[1]*
[1]Shanghai Jiao Tong University, [2] University of Michigan
[1]{LGJ1zed,huanning,zhangwenjun,sihengc}@sjtu.edu.cn
[2]huyu@umich.edu

ABSTRACT

Collaborative perception emphasizes enhancing environmental understanding by enabling multiple agents to share visual information with limited bandwidth resources. While prior work has explored the empirical trade-off between task performance and communication volume, a significant gap remains in the theoretical foundation. To fill this gap, we draw on information theory and introduce a pragmatic rate-distortion theory for multi-agent collaboration, specifically formulated to analyze performance-communication trade-off in goal-oriented multi-agent systems. This theory concretizes two key conditions for designing optimal communication strategies: supplying pragmatically relevant information and transmitting redundancy-less messages. Guided by these two conditions, we propose RDcomm, a communication-efficient collaborative perception framework that introduces two key innovations: i) task entropy discrete coding, which assigns features with task-relevant codeword-lengths to maximize the efficiency in supplying pragmatic information; ii) mutual-information-driven message selection, which utilizes mutual information neural estimation to approach the optimal redundancy-less condition. Experiments on 3D detection and BEV segmentation show that RDcomm achieves state-of-the-art accuracy on datasets DAIR-V2X, OPV2V, V2XSeq, and V2V4Real, while reducing communication volume by up to 108×. Our code is available at https://github.com/gjliu9/RDcomm.

1 INTRODUCTION

Multi-agent collaborative perception enhances environmental understanding by enabling agents to jointly perceive and share information. This paradigm has shown clear advantages over single-agent sensing, particularly in overcoming occlusions and limited fields of view, and has been widely adopted in tasks such as 3D object detection Wang et al. (2020) and BEV segmentation Xu et al. (2022a).

In this field emphasizing multi-agent collaboration, a fundamental challenge remains: the trade-off between task performance and communication volume Hu et al. (2022). While sharing richer information tends to preserve collaboration quality, it introduces significant communication overhead; conversely, aggressively limiting communication may result in the loss of task-critical information, ultimately degrading overall performance. Recent works tackle this trade-off by extracting informative and compact representations from visual observations to serve as collaborative messages. One representative line of work focuses on spatial selection, aiming to transmit only task-relevant regions, such as those with high detection confidence Hu et al. (2022) or sparse observation coverage Xu et al. (2025). Another direction leverages neural compression techniques Ballé et al. (2018); Van Den Oord et al. (2017), reducing the size of transmitted features through autoencoders Shao et al. (2024); Hu et al. (2024) or channel reduction Li et al. (2021); Lu et al. (2024). Despite some empirical gains, previous approaches are heuristic in nature, relying on manually designed communication strategies or intuitive criteria. Such approaches lack theoretical grounding and provide no principled guidance on what to communicate or how to encode it under bandwidth constraints.

To fill this gap, we take an information-theoretic perspective and propose *pragmatic rate-distortion theory for multi-agent collaboration*, which explicitly models the trade-off between communication

*Corresponding author

bit-rate and task-specific pragmatic distortion. Our theoretic analysis extends Shannon's classical rate-distortion framework Shannon et al. (1959) in two key aspects. First, we introduces pragmatic distortion, a task-driven metric that reflects the impact of message degradation on downstream task performance, distinct from reconstruction-based distortions Blau & Michaeli (2019); Cover (1999). Second, our theory generalizes to distributed communication among multiple agents, where both message senders and receivers observe the environment. We thus account for inter-agent redundancy, a factor typically neglected in traditional rate–distortion analysis. Building upon these extensions, our theory ultimately characterizes the minimal communication cost required to meet a specified distortion threshold, and derives two key conditions that an optimal communication strategy should satisfy: *pragmatic-relevant* and *redundancy-less*. We envision this theoretical framework as a foundation for analyzing communication efficiency in broader multi-agent tasks.

Inspired by the two theoretical conditions, we propose **RDcomm** (**R**ate-**D**istortion guided pragmatic **comm**unication), a novel communication-efficient collaborative perception system, which optimizes both message selection and coding to reduce communication overhead while preserving collaborative complementarity and task effectiveness. Specifically, we design the two core components of RDcomm based on the two derived conditions: i) Based on the pragmatic-relevant condition, we propose a novel task entropy discrete coding module. It first utilizes learned codebooks to quantize feature vectors, and then applies variable-length coding guided by task relevance, assigning shorter codewords to more informative features. ii) Based on the redundancy-less condition, we propose a novel feature selection module leveraging mutual information neural estimation Belghazi et al. (2018). This module enables agents to perform an inter-agent handshake process to assess message redundancy by quantifying the mutual information between shared and locally observed features. We validate RDcomm on two representative perception tasks: 3D object detection and BEV semantic segmentation, using both real-world datasets DAIR-V2X Yu et al. (2022), V2XSeq Yu et al. (2023), V2V4Real Xu et al. (2023) and simulation dataset OPV2V Xu et al. (2021). Experimental results show that RDcomm reduces communication volume by up to 108 times compared against existing methods.

Our main contributions are summarized as follows:

- We introduce a pragmatic rate-distortion theory for multi-agent collaboration, which characterizes the performance-communication trade-off, and concretize two optimal conditions: i) supply pragmatic-relevant information; ii) avoid inter-agent redundancy.
- We propose RDcomm, a communication-efficient collaborative perception framework that is designed to approach the two optimal conditions with two innovations: i) task entropy discrete coding; ii) mutual-information-driven message selection. Experiments on detection and segmentation tasks demonstrate that RDcomm achieves dual superiority in both performance and communication efficiency.

2 Related works

2.1 Communication-efficient collaborative perception

In multi-agent collaboration, a key challenge is to balance task performance and communication cost Hu et al. (2022). Early collaboration transmits raw sensor data Han et al. (2023) and achieves high accuracy but suffers from heavy bandwidth usage. Late collaboration reduces bandwidth by sending final predictions, but degrades performance under noise Lu et al. (2023); Hu et al. (2022). To address this, intermediate collaboration transmits feature maps to strike a balance between performance and efficiency. Prior works mainly improve efficiency via: i) spatial selection, which transmits features at critical regions; and ii) feature compression. For spatial selection, Where2comm Hu et al. (2022) selects high-confidence regions, CodeFilling Hu et al. (2024) removes redundant collaborators, and CoSDH Xu et al. (2025) targets unobserved areas. For compression, techniques include value quantization Wang et al. (2020); Shao et al. (2024), vector quantization Hu et al. (2024), and channel reduction Li et al. (2021); Lu et al. (2024). However, these methods are primarily heuristic and lack theoretical guarantees for communication efficiency. In our work, we provide a theoretical framework grounded in rate-distortion analysis, offering explicit conditions for optimal communication.

2.2 Rate-distortion theory background

Rate-distortion theory Shannon et al. (1959) provides a fundamental framework for lossy compression by characterizing the minimum bits required to represent a signal X as a compressed representation Z under a specified distortion constraint $\mathrm{D}[X, Z] \leq \delta$. The goal is to find a probabilistic encoding

map $p(Z|X)$ that minimizes the mutual information $\mathrm{I}(X;Z)$, as formulated in (1):

$$\mathrm{Rate}(\delta) = \min_{p(Z|X)} \mathrm{I}(X;Z) \quad \text{s.t. } \mathrm{D}[X,Z] \leq \delta \tag{1}$$

In lossy compression, Z is a reconstruction of X. A typical example is a Gaussian source $X \sim \mathcal{N}(\mu, \sigma^2)$ with mean squared error (MSE) distortion, where the optimal rate under distortion level δ has a closed-form solution $R(\delta) = h(X) - \frac{1}{2}\log(2\pi e\delta)$, which reflects the total information in X minus the portion tolerable under the distortion budget. Despite its general application in visual compression Ballé et al. (2018), classical rate-distortion analysis mostly caters to fidelity-based distortion metrics Blau & Michaeli (2019) and single-source settings. In general, any distortion measure $d : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$ of the form $D[X,Z] = \mathbb{E}_{p(x,z)}[d(x,z)]$ is valid, as long as there exists $z \in \mathcal{Z}$ such that $D[X,Z]$ is finite Dubois et al. (2021). Our work extends this framework to multi-agent collaborative perception by incorporating pragmatic distortion and inter-agent redundancy.

3 Pragmatic rate-distortion theory for collaboration

We introduce the pragmatic rate-distortion theory for multi-agent collaboration. Our theoretical analysis follows three high-level steps: i) We reformulate the collaboration task objective in a rate-distortion formulation (Sec. 3.1); ii) Defining the pragmatic distortion for collaboration (Sec. 3.2); iii) We derive the minimal transmission rate under constrained distortion, and present the conditions for an optimal communication strategy (Sec. 3.3). See detailed proofs in the Appendix A.6.

3.1 Problem formulation

In collaborative tasks, our goal is to optimize model parameters and the message generation strategy, in order to achieve the minimal transmission bits under constrained task loss, that is (2):

$$\min_{Z_s,\theta} \sum_{r=1}^{N} |\{Z_{s\to r}\}_{s\neq r}| \quad \text{s.t. } \sum_{r=1}^{N} L_Y\left(Y_r, \Phi_\theta\left(X_r, \{Z_{s\to r}\}_{s\neq r}\right)\right) \leq L_{max} \tag{2}$$

where N denotes the number of agents; $\{Z_{s\to r}\}_{s\neq r}$ represents the messages sent from agent s to other agents; $|\cdot|$ measures the information volume; $\Phi_\theta(\cdot)$ is the task model parameterized by θ; X_r and Y_r denote the local observation and ground-truth label of agent r, respectively; $L_Y(\cdot)$ is the loss function associated with task Y, and $L_{\max}$ specifies the maximum tolerable task loss. We then present (3) as a formal rate-distortion optimization version of (2), which defines the minimal communication bits $\mathrm{Rate}(\delta)$ and serves as the foundational objective.

$$\mathrm{Rate}(\delta) = \min_{p(Z_{s\to r}|X_s)} \mathrm{I}(X_s; Z_{s\to r}) \quad \text{s.t. } \mathrm{D}_Y[X_s, Z_{s\to r}|X_r] \leq \delta \tag{3}$$

As shown in (3), the communication volume is captured by $\mathrm{I}(X_s; Z_{s\to r})$, which quantifies the amount of information from the original observation X_s that is preserved in the transmitted message $Z_{s\to r}$. We denote $\mathrm{D}_Y[X_s, Z_{s\to r} \mid X_r]$ as the pragmatic distortion for the collaborative task Y, which measures the degradation in task performance when transmitting $Z_{s\to r}$ instead of X_s, given the local observation X_r. See detailed discussions on our problem formulation in Appendix A.6.1.

3.2 Pragmatic distortion for collaborative perception

We then make the pragmatic distortion $\mathrm{D}_Y[X_s, Z_{s\to r} \mid X_r]$ in (3) explicit for collaborative perception. Let X_s, X_r denote the sender's and receiver's local observations, respectively, Y denotes the perception task target. $Z_{s\to r}$ is compressed from X_s. We define $\mathrm{D}_Y[X_s, Z_{s\to r}|X_r]$ in (4):

$$\mathrm{D}_Y\left[X_s, Z_{s\to r}|X_r\right] = \mathrm{B}_{risk}\left[Y|Z_{s\to r}, X_r\right] - \mathrm{B}_{risk}\left[Y|X_s, X_r\right] \tag{4}$$

where $\mathrm{B}_{\mathrm{risk}}[Y|X_r, X_s]$ denotes the Bayes Risk Dubois et al. (2021), which measures the minimum achievable prediction error for the target Y given the joint inputs X_r and X_s. We define the distortion $\mathrm{D}_Y[X_s, Z_{s\to r}|X_r]$ as the increase in Bayes Risk when predicting Y with the compressed representation $Z_{s\to r}$ instead of the original signal X_s, while conditioning on the existing local information X_r. Formally, the Bayes Risk is $\mathrm{B}_{\mathrm{risk}}[Y|X] = \inf_f \mathbb{E}_{p(X,Y)}\left[L\big(Y, f(X)\big)\right]$, where f is any predictor. For perception tasks, the loss is typically computed independently at each BEV location. Accordingly, we define the overall Bayes Risk as the average of pixel-wise Bayes Risks

over all locations, as in (5), where $i \in \mathcal{S}$ denotes a BEV coordinate within the perception range $\mathcal{S}$, and $Y_{(i)}$ and $f(X)_{(i)}$ denote the corresponding ground-truth label and model prediction.

$$\mathrm{B}_{risk}[Y|X] = \inf_{f} \mathbb{E}_{p(X,Y)}\left[L(Y, f(X))\right] = \inf_{f} \frac{1}{|Y|} \sum_{i \in \mathcal{S}} \mathbb{E}_{p(X,Y_{(i)})}\left[L(Y_{(i)}, f(X)_{(i)})\right] \tag{5}$$

We instantiate the pragmatic distortion for two representative perception tasks: BEV segmentation and 3D object detection. Specifically, for BEV segmentation, we adopt the per-pixel cross-entropy (CE) loss; for 3D object detection, we adopt the widely used CenterPoint loss Yin et al. (2021). We directly present the final derivations of the pragmatic distortion $\mathrm{D}_Y\left[X_s, Z_{s\to r}|X_r\right]$ in Tab. 1.

Table 1: Pragmatic distortions differ from classical reconstruction distortion by considering task entropy $\mathrm{H}(Y|\cdot)$ and local redundancy X_r. Proofs are provided in Appendix A.6.4.

Task (loss function)	**Distortion** $\mathrm{D}_Y\left[X_s, Z_{s\to r}|X_r\right]$	
Lossy reconstruction Ballé et al. (2018) (MSE)	$\frac{1}{\lvert X_s\rvert}\left\lVert X_s - Z_{s\to r}\right\rVert_2^2, \quad \text{no } X_r$	(6)
BEV segmentation (CE)	$\frac{1}{\lvert Y\rvert}\sum_{i\in\mathcal{S}}[\mathrm{H}(Y_{(i)} \mid Z_{s\to r}, X_r) - \mathrm{H}(Y_{(i)} \mid X_s, X_r)]$	(7)
3D detection. (CenterPoint)	$\frac{1}{\lvert Y\rvert}\sum_{i\in\mathcal{S}}[\mathrm{H}(Y_{(i,c)} \mid Z_{s\to r}, X_r) - \mathrm{H}(Y_{(i,c)} \mid X_s, X_r) + \frac{1}{2}\sum_{k\in\mathcal{K}}\left(e^{\mathrm{H}(Y_{(i,k)}\mid Z_{s\to r}, X_r)-1} - e^{\mathrm{H}(Y_{(i,k)}\mid X_s, X_r)-1}\right)]$	(8)

3.3 Minimal bit-rate of collaborative message and optimal conditions

We present the trade-off between communication bit-rate and distortion by incorporating the pragmatic distortions (7) and (8) into objective (3).

Theorem 1 *(Minimal bit-rate* $\mathrm{Rate}(\delta)$ *of collaborative message under pragmatic distortion constraint, see proof in A.6.5). Consider a message sender agent* a_s *and a message receiver agent* a_r *and their observation denoted as* X_s, X_r*, where the sender compresses* X_s *as* $Z_{s\to r}$ *and transmits it to the receiver to collaborates in achieving task target* Y*. Then, the minimal transmission bit-rate* $\mathrm{Rate}(\delta) = \min_{p(Z_{s\to r}|X_s)} \mathrm{I}(X_s; Z_{s\to r})$ *s.t.* $\mathrm{D}_Y[X_s, Z_{s\to r}] \le \delta$ *is given in (10).*

$$\mathrm{Rate}(\delta) = \min_{p(Z_{s\to r}|X_s) \ \text{s.t. } \mathrm{D}_Y[X_s, Z_{s\to r}]\le\delta} \mathrm{I}(X_s; Z_{s\to r}) = \mathrm{I}(Y; X_s \mid X_r) - \delta \tag{9}$$

$$= \mathrm{H}(X_s) - \underbrace{\mathrm{H}(X_s|Y)}_{\text{information in } X_s \text{ irrelevant to } Y} - \underbrace{\mathrm{I}(Y; X_s; X_r)}_{\text{information in } X_s \text{ redundant with } X_r \text{ about } Y} - \delta \tag{10}$$

The minimal bit-rate $\mathrm{Rate}(\delta)$ can be achieved only if the following two conditions are satisfied. **Pragmatic-relevant.** The transmitted message $Z_{s\to r}$ should contain only information relevant to the receiver's task Y, as formalized in (11):

$$\mathrm{H}(Z_{s\to r}|Y) = 0 \tag{11}$$

Equation (11) implies that the uncertainty of $Z_{s\to r}$ is eliminated given a specific task target Y, indicating a unique mapping from Y to $Z_{s\to r}$. Consequently, $Z_{s\to r}$ should exclude any information unrelated to the task Y, and the task-relevant messages should be prioritized during coding.

Redundancy-less. The transmitted message $Z_{s\to r}$ should avoid maintaining information that is already contained in the receiver's observation X_r, which is expressed as (12):

$$\mathrm{I}(Z_{s\to r}; X_r) = 0 \tag{12}$$

Equation (12) shows that the mutual information between the transmitted message $Z_{s\to r}$ and the observation of receiver X_r should be eliminated. In other words, $Z_{s\to r}$ should avoid containing the redundant information, but exactly supply the information missed in X_r.

4 RDcomm: efficient communication for co-perception

Inspired by the theoretical analysis, we introduce **RDcomm**, a novel communication-efficient collaborative perception framework with three main components (Fig. 1): i) a perception pipeline that provides the basic functionalities for perception tasks; ii) a task entropy discrete coding module (Sec. 4.1), following the pragmatic-relevant condition (11), which adopts a novel variable-length coding guided by task relevance; iii) a mutual-information–driven selection module (Sec. 4.2), following the redundancy-less condition (12), which selects complementary messages for transmission.

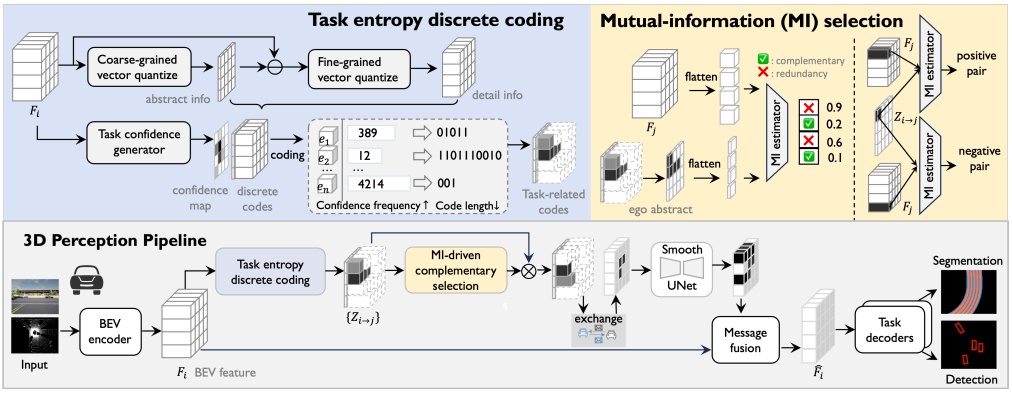


Figure 1: RDcomm features two key components: i) task entropy discrete coding for improving the pragmatic relevance of message, which assigns short codewords to the codes with high confidence frequency; ii) mutual-information-driven message selection, which measures message redundancy by mutual information estimation.

Perception pipeline. The perception pipeline couples a BEV encoder with task-specific decoders. The BEV encoder accepts either LiDAR or camera inputs and maps sensor data into a unified bird's-eye-view (BEV) representation, enabling consistent spatial alignment across agents. Denoting the observation of the sender agent by X_s, the BEV encoder $\Phi_{\text{bev}}(\cdot)$ produces a BEV feature map by $F_s = \Phi_{\text{bev}}(X_s) \in \mathbb{R}^{h\times w\times c}$, followed by task-specific decoders Φ_{task} for downstream tasks such as 3D object detection and BEV segmentation. Backbone details are provided in Appendix A.4.1. We then focus on compressing the feature F_s into a collaboration message $Z_{s\to r}$ for any receiver agent.

4.1 Task entropy discrete coding

Our first objective is to approach the pragmatic-relevant condition $\mathrm{H}(Z_{s\to r}|Y) = 0$ in (11). We approximate it by minimizing the task-conditioned entropy $\mathrm{H}(Z_{s\to r}|Y)$.

Layered vector quantization. We begin by constraining $\mathrm{H}(Z_{s\to r})$ via vector quantization inspired by Zhu et al. (2022), where the core idea is to represent each vector in F_s with the nearest embedding $\mathbf{e}_i$ in a codebook $\mathbf{B} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n] \in \mathbb{R}^{n\times d}$ with n learnable embeddings. We further implement a layered discrete auto-encoder $\Phi_{\text{vq}}(\cdot)$ to quantize the BEV feature F_s by $F_s^q = \Phi_{\text{vq}}(F_s, \mathbf{B}_{base}, \mathbf{B}_{res})$. $\mathbf{B}_{base}$ is used to approximate the basic coarse-grained information of F_s with small codebook volume n, and the residual error $F_s - \Phi_{\text{vq}}(F_s, \mathbf{B}_{base})$ is further approximated by a fine-grained codebook $\mathbf{B}_{res}$ with larger volume. This layered quantization is described in (13)(14):

$$Z_{base}^q = \arg\min_i \|\mathbf{B}_{base}[i] - f_{in}(F_s)\|_2, \quad Z_{res} = F_s - Z_{base}^q \tag{13}$$

$$Z_{res}^q = \arg\min_i \|\mathbf{B}_{res}[i] - f_{in}(Z_{res})\|_2, \quad F_s^q = f_{out}(Z_{res}^q + Z_{base}^q) \tag{14}$$

where $f_{in}(\cdot), f_{out}(\cdot)$ are MLP projectors to bridge the distribution gap between continuous vectors and codebook embeddings, and the input feature map F_s is flattened before quantization.

Task-aware priority and encoding. While quantization reduces representation length by restricting the vector-space volume, we further improve coding efficiency by introducing task bias. Specifically, we prioritize task-relevant messages for selection and encode them with shorter code lengths. Recall that our objective is to minimize the task-conditioned entropy $\mathrm{H}(Z_{s\to r}|Y) \to 0$, which can be expanded as (15).

$$\min_{Z_{s\to r}} \mathbb{E}_Y \sum_{Z_{s\to r}} [-p(Z_{s\to r}|Y)\log p(Z_{s\to r}|Y)] \tag{15}$$

Note that maximizing $p(Z_{s\to r}|Y)$ to 1 provides a sufficient solution for minimizing $\mathrm{H}(Z_{s\to r}|Y)$, and $p(Z_{s\to r}|Y) \propto p(Y|Z_{s\to r})p(Z_{s\to r})$ for a given target distribution $p(Y)$. Therefore we priority the messages with high task confidence $p(Y|Z_{s\to r})$ for transmission. We implement this with a confidence generator $\Phi_{\text{conf}}(\cdot)$ producing scores $C_s = \Phi_{\text{conf}}(F_s) \in \mathbb{R}^{h\times w}$. The confidence mask is $M_c = \mathbf{1}[C_s > \tau_c]$, where τ_c is a confidence threshold, and quantized features are selected as $F_{sc}^q = M_c \odot F_s^q$. In practice, we instantiate $\Phi_{\text{conf}}(\cdot)$ by reusing the task decoder $\Phi_{\text{task}}(\cdot)$.

We further reduce the average coding length of the quantized F_{sc}^q, where we encode each embedding $\mathbf{e}_i$ in $\{\mathbf{B}_{base}, \mathbf{B}_{res}\}$ considering the joint effect of task confidence $p(Y|\mathbf{e}_i)$ and occurrence frequency

$p(\mathbf{e}_i)$. Specifically, we define the confidence frequency $p_c(\cdot)$ for each embedding $\mathbf{e}_i$ as (16):

$$p_c(\mathbf{e}_i) = \sum_{F_s} \sum_{\{(u,v):\ \mathbf{e}_i \in \Phi_{\text{vq}}(F_s)[u,v]\}} f_{\text{filter}}\big(\Phi_{\text{conf}}(F_s)[u,v]\big) \tag{16}$$

where $p_c(\mathbf{e}_i)$ represents the total task confidence predicted from $\mathbf{e}_i$ across the entire dataset. We compute it by accumulating the confidence scores $\Phi_{\text{conf}}(F_s)[u,v]$ at spatial locations (u,v) where $F_s[u,v]$ are quantized with embedding $\mathbf{e}_i$. To mitigate the adverse effect of accumulating low-confidence values, which can inflate the confidence frequency $p_c(\mathbf{e}_i)$ across all embeddings $\mathbf{e}_i$ and reduce their distinguishability, we introduce a filtering function $f_{\text{filter}}(\cdot)$ defined as $f_{\text{filter}}(c) = 0$ if $c < \tau_{\text{filter}}$ and $f_{\text{filter}}(c) = c$ otherwise, where we set $\tau_{\text{filter}} = 0.2$ in all experiments. Based on $p_c(\mathbf{e}_i)$, we assign shorter code lengths for the embeddings with higher confidence frequency to improve coding efficiency for the task-relevant embeddings. For implementation, this work provides a straightforward yet effective solution by applying Huffman coding Huffman (2007), where we set the Huffman weight of $\mathbf{e}_i$ to be its confidence frequency $p_c(\mathbf{e}_i)$. Our task-entropy coding ultimately produces an index map $D_s \in (\{0,1\}^{l_{u,v}})^{h\times w}$ to represent the input feature $F_s \in \mathbb{R}^{h\times w\times c}$, where $\{0,1\}^{l_{u,v}}$ denotes binary strings of variable length $l_{u,v}$. $D_s[u,v] = [D_s^{base}[u,v] \,\|\, D_s^{res}[u,v]] \in \{0,1\}^{l_{u,v}}$ denotes the code at location (u,v), which consists of coarse-grained semantic information $D_s^{base}[u,v]$ and fine-grained $D_s^{res}[u,v]$.

Discussion. i) The relation between coding method and the information optimization target (15): the high-confidence selection module $\Phi_{\text{conf}(\cdot)}$ reduces the task-conditioned entropy $\mathrm{H}(Z_{s\to r}|Y)$ governed by $p(Z_{s\to r}|Y) \propto p(Y|Z_{s\to r})p(Z_{s\to r})$. We instantiate this by defining the confidence frequency $p_c(\cdot)$, and use it as the weight for entropy coding. ii) The intuition behind is simple: with limited bandwidth, high-confidence messages are selected more often, making their embeddings frequent. Accordingly, we weight entropy coding using the confidence frequency $p_c(\mathbf{e}_i)$. iii) We use the accumulation of confidence rather than the average to distinguish embeddings with similar task relevance but different occurrence frequencies. iv) The confidence-frequency coding is applied post-training and remains lossless for task performance, as it only reassigns embedding indices.

4.2 Complementary Selection with Mutual Information Estimation

In this section we focus on reducing inter-agent message redundancy according to the redundancy-less condition $\mathrm{I}(Z_{s\to r}; X_r) = 0$ (12). We approximate this target by (17). Ω denotes the selection region.

$$\min_{\Omega} \mathrm{I}(\hat{F}_{sc}^{q}[\Omega]; F_r) \tag{17}$$

To obtain the redundancy-less region Ω, we perform feature selection via mutual information neural estimation: we approximate $\mathrm{I}(\hat{F}_{sc}^{q}; F_r)$ by with a learnable estimator $\Phi_{\text{MI}}(\hat{F}_{sc}^{q}, F_r)$ and select features accordingly. Note that we use the coarse-grained compression $\hat{F}_{sc}^{q} = \mathbf{B}_{base}[D_s^{base}]$ as an abstract of F_s, which is pre-handed to the receiver to estimate its redundancy with the receiver's information F_r. This abstraction $\hat{F}_{sc}^{q}$ helps estimate semantic redundancy by transmitting D_s^{base} with much smaller communication volume, which is 10 times smaller than the lossless message D_s.

Mutual information neural estimation. Consider the variable pair $\mathbf{s}, \mathbf{r} \in \mathbb{R}^c$ that are included in $\hat{F}_{sc}^{q}, F_r \in \mathbb{R}^{h\times w\times c}$ respectively, the mutual information $\mathrm{I}(\mathbf{s},\mathbf{r})$ can be estimated using (18):

$$\mathrm{I}(\mathbf{s},\mathbf{r}) := D_{KL}(\mathbb{P}_{\mathbf{s},\mathbf{r}} \parallel \mathbb{P}_{\mathbf{s}} \otimes \mathbb{P}_{\mathbf{r}}) \geq \sup_{T:\mathbb{R}^c\times\mathbb{R}^c\to\mathbb{R}} \Big\{ \mathbb{E}_{\mathbb{P}_{\mathbf{s},\mathbf{r}}}[T(\mathbf{s},\mathbf{r})] - \log(\mathbb{E}_{\mathbb{P}_{\mathbf{s}}\otimes\mathbb{P}_{\mathbf{r}}}[e^{T(\mathbf{s},\mathbf{r})}]) \Big\} \tag{18}$$

where $\mathbb{P}_{\mathbf{s},\mathbf{r}}$ denotes the joint distribution of $\mathbf{s}$ and $\mathbf{r}$, and $\mathbb{P}_{\mathbf{s}}, \mathbb{P}_{\mathbf{r}}$ denote marginals; D_{KL} is KL divergence; $T : \mathbb{R}^c \times \mathbb{R}^c \to \mathbb{R}$ is any function that projects the $(\mathbf{s},\mathbf{r})$ pair into a real number. Formulation (18) shows that mutual information $\mathrm{I}(\mathbf{s},\mathbf{r})$ actually measures the divergence between the joint distribution $\mathbb{P}_{\mathbf{s},\mathbf{r}}$ and marginal distribution $\mathbb{P}_{\mathbf{s}} \otimes \mathbb{P}_{\mathbf{r}}$. While (18) employs the KL divergence as a standard estimation, we instead adopt the GAN-style divergence Nowozin et al. (2016) following Li et al. (2020) to facilitate optimization, and use the estimation target in (19). $\sigma(\cdot)$ denotes sigmoid.

$$\hat{\mathrm{I}}(\mathbf{s},\mathbf{r}) \geq \sup_{T:\mathbb{R}^c\times\mathbb{R}^c\to\mathbb{R}} \big\{ \mathbb{E}_{\mathbb{P}_{\mathbf{s},\mathbf{r}}}[\log\sigma(T(\mathbf{s},\mathbf{r}))] + \mathbb{E}_{\mathbb{P}_{\mathbf{s}}\otimes\mathbb{P}_{\mathbf{r}}}[\log(1-\sigma(T(\mathbf{s},\mathbf{r})))] \big\} \tag{19}$$

The inequality results from approximating $T(\cdot)$ with limited representation ability, which is a subclass of projection $\mathbb{R}^c \times \mathbb{R}^c \to \mathbb{R}$. Our objective is to optimize $T(\cdot)$ to maximize the estimation

bound in (18), where $T(\cdot)$ is implemented by a learnable estimator $\Phi_{\mathrm{MI}}(\cdot)$, with the loss defined in (20):

$$\mathcal{L}_{MI} = -\frac{1}{|\mathcal{P}_{\mathbf{s},\mathbf{r}}|} \sum_{(\mathbf{s},\mathbf{r})\in\mathcal{P}_{\mathbf{s},\mathbf{r}}} \log\sigma\left(\Phi_{\mathrm{MI}}(\mathbf{s},\mathbf{r})\right) - \frac{1}{|\mathcal{P}_{\mathbf{s}}||\mathcal{P}_{\mathbf{r}}|} \sum_{\mathbf{s}\in\mathcal{P}_{\mathbf{s}},\mathbf{r}\in\mathcal{P}_{\mathbf{r}}} \log\left(1-\sigma\left(\Phi_{\mathrm{MI}}(\mathbf{s},\mathbf{r})\right)\right) \quad (20)$$

where $\mathcal{P}_{\mathbf{s},\mathbf{r}}$ denotes the set sampled from $\mathbb{P}_{\mathbf{s},\mathbf{r}}$ where $\mathbf{s},\mathbf{r}$ are visual features from two agents at the same location. $\mathcal{P}_{\mathbf{s}},\mathcal{P}_{\mathbf{r}}$ are sampled from the marginals where $\mathbf{s},\mathbf{r}$ are randomly combined. We see that the mutual information estimator $\Phi_{\mathrm{MI}}(\cdot)$ actually serves as a discriminator. It predicts whether a feature pair $(\mathbf{s},\mathbf{r})$ is drawn from the joint distribution $\mathbb{P}_{\mathbf{s},\mathbf{r}}$, i.e., two agents' observations at the same location that exhibit similar patterns, or from the product of marginals $\mathbb{P}_{\mathbf{s}}\otimes\mathbb{P}_{\mathbf{r}}$, in which observations are randomly paired and likely to differ in pattern due to non-corresponding locations.

Redundancy-less feature selection. To reduce redundancy in transmission, we prioritize the message $\mathbf{s}$ with a low mutual information score $\Phi_{\mathrm{MI}}(\mathbf{s},\mathbf{r})$. Specifically, we obtain the redundancy map $R_{s\to r} = \sigma(\Phi_{\mathrm{MI}}(\hat{F}^q_{sc}, F_r)) \in \mathbb{R}^{h\times w}$ and derive the redundancy-less selection mask $M_{MI} = \mathbf{1}[R_{s\to r} < \tau_{MI}]$, where τ_{MI} is a redundancy threshold. The feature finally sent to the receiver is $Z_{s\to r} = M_{MI}\odot\hat{F}^q_{sc}$. This selection filters out the messages in $Z_{s\to r}$ that are already covered by F_r. The total communication volume of RDcomm is computed as $|D_s\odot M_c\odot M_{MI}| + |D^{base}_s\odot M_c|$. The first term measures the volume of the selected lossless information $Z_{s\to r}$, and the second term measures the cost for identifying redundancy and is much smaller than the first term by setting $\mathbf{B}_{base}$ with small codebook size n and dimension d with limited segments.

Message smoothing and fusion. Note that $Z_{s\to r}$ is obtained under sparse masks M_C and M_{MI}. Although it preserves salient information, the sparsity may degrade semantic content. We mitigate this by applying a UNet Ronneberger et al. (2015) $\Phi_{\mathrm{smth}}(\cdot)$ to smooth and dilate $Z_{s\to r}$, propagating sparse signals to neighboring regions. The receiver then obtains the enhanced perception results as $\bar{Y}_r = \Phi_{\mathrm{task}}(\Phi_{\mathrm{fusion}}(F_r, \Phi_{\mathrm{smth}}(Z_{s\to r})))$, where $\Phi_{\mathrm{fusion}}(\cdot)$ is instantiated using the effective max-fusion operation following Hu et al. (2024).

4.3 Training

We train RDcomm in three stages. First, we train the BEV encoder Φ_{bev} and the task decoder Φ_{task} with task loss $\mathcal{L}_{task}$, which corresponds to the CenterPoint loss for 3D detection and the per-pixel cross-entropy loss for BEV segmentation. Second, we train the vector quantization module Φ_{vq} with both task loss $\mathcal{L}_{task}$ and feature reconstruction loss $\mathcal{L}_{recon} = \|F^q_s - F_s\|^2_2$. After that, the confidence frequency $p_c(\cdot)$ is updated. Finally, we train the mutual-information estimator Φ_{MI} using $\mathcal{L}_{MI}$. In the later stages of training, the thresholds τ_c, τ_{MI} are randomly varied to facilitate bandwidth adaptation. We present an analysis of the training cost in Appendix A.3.

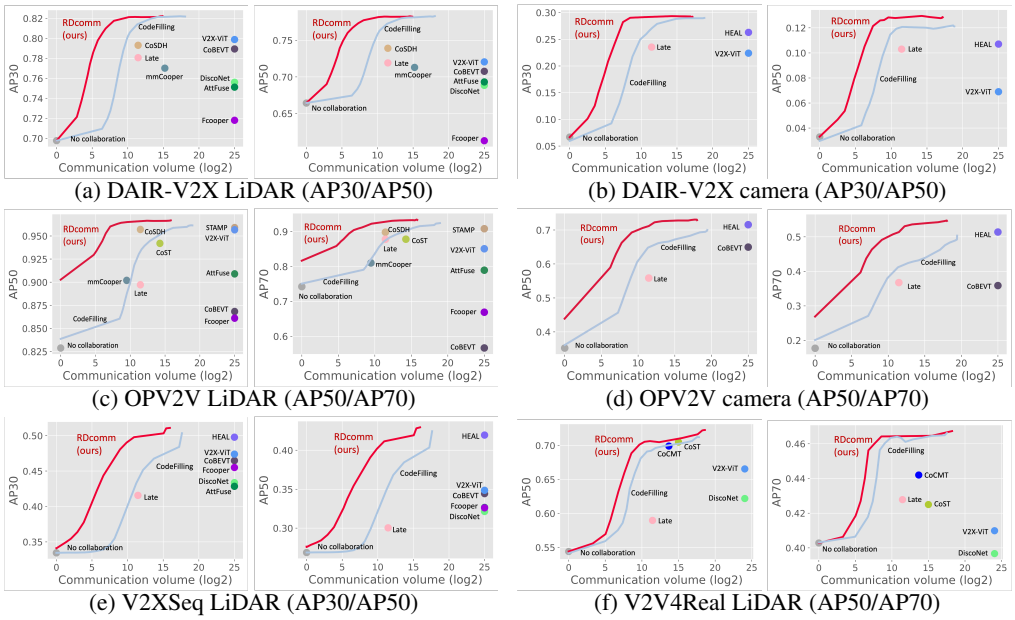
5 Experiments

To evaluate RDcomm, we conduct experiments on two representative collaborative perception tasks: 3D object detection and BEV semantic segmentation. Our evaluation spans both LiDAR and camera modalities, 2-5 collaborating agents, and varying bandwidth constraints. In RDComm, we adjust the communication volume by varying τ_c in interval $[0,1]$ while keeping $\tau_{MI} = 0.7$ to adapt to different bandwidth constraints. The effects of controlling thresholds τ_{MI}, τ_c are analyzed in Appendix A.5.1, and performance with varying numbers of collaborators is reported in Appendix A.5.2.

Experimental setup of collaborative 3D detection. We evaluate on four representative collaborative perception datasets: three real-world datasets DAIR-V2X Yu et al. (2022), V2XSeq Yu et al. (2023), V2V4Real Xu et al. (2023), and a simulated dataset OPV2V Xu et al. (2021). DAIR-V2X is a vehicle-to-infrastructure dataset with 9K frames of 2-agent collaboration between a vehicle and a roadside unit (RSU). Each agent is equipped with a LiDAR and a 1920×1080 camera, where the RSU uses a 300-channel LiDAR and the vehicle a 40-channel LiDAR. V2XSeq is a sequential perception dataset, which includes more than 15,000 frames captured from 95 vehicle-to-infrastructure scenarios. V2V4Real is a vehicle-to-vehicle dataset. It includes a total of 20K frames of LiDAR point cloud with 240K annotated 3D bounding boxes. OPV2V is a vehicle-to-vehicle dataset simulated with CARLA Dosovitskiy et al. (2017), containing 12K frames. Our experiments involve up to 3 agents, each equipped with a 64-channel LiDAR and four RGB cameras at 800×600 resolution. We evaluate both LiDAR and camera modalities and report Average Precision (AP) at IoU thresholds of 30%, 50%, and 70%. We set the perception range to 204.8m×102.4m for DAIR-V2X, OPV2V, V2XSeq and 140.8m*76.8m for V2V4Real, following Lu et al. (2024); Wang et al. (2025).

Experimental setup of collaborative BEV segmentation. We evaluate BEV semantic segmentation on OPV2V dataset following CoBEVT Xu et al. (2022a). Each agent predicts a BEV semantic occupancy map with camera inputs, ground truth classes include dynamic vehicles, drivable area, and lane. We involve collaboration among up to 5 agents. Performance is measured by Intersection-over-Union (IoU) between predictions and ground-truth BEV labels. Perception range is 100m×100m.

5.1 QUANTITATIVE ANALYSIS



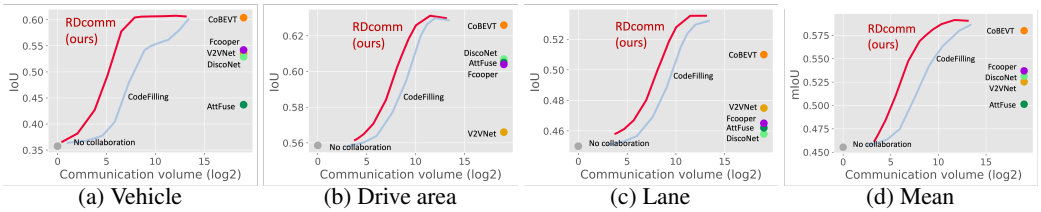
(a) DAIR-V2X LiDAR (AP30/AP50)
(b) DAIR-V2X camera (AP30/AP50)
(c) OPV2V LiDAR (AP50/AP70)
(d) OPV2V camera (AP50/AP70)
(e) V2XSeq LiDAR (AP30/AP50)
(f) V2V4Real LiDAR (AP50/AP70)

Figure 2: RDcomm achieves the best performance–communication trade-off in 3D detection, across DAIRV2X/OPV2V/V2XSeq/V2V4Real datasets with LiDAR/camera input modalities.

Benchmark comparison. Fig. 2 and Fig. 3 compare RDcomm with previous collaborative perception methods in terms of the trade-off between perception performance and communication volume on 3D detection and BEV segmentation. The baselines include the state-of-the-art collaborative perception methods CodeFilling Hu et al. (2024), CoSDH Xu et al. (2025), STAMP Gao et al. (2025), CoST Tang et al. (2025), CoCMT Wang et al. (2025), mmCooper Liu et al. (2025), V2X-ViT Xu et al. (2022b), CoBEVT Xu et al. (2022a), DiscoNet Li et al. (2021), AttFuse Xu et al. (2021), V2VNet Wang et al. (2020), Fcooper Chen et al. (2019), HEAL Lu et al. (2024), no collaboration, and Late Fusion (directly share the final perception results). We see that: i) RDcomm consistently delivers a superior perception-communication trade-off across all bandwidth settings for both detection and segmentation; the gains persist across LiDAR/camera modalities and multiple semantic classes. ii) Under extreme bandwidth constraints, RDcomm achieves larger gains than prior methods: for detection, +11.49/+19.82% (LiDAR/camera) on DAIR-V2X and +12.01/+22.92% on OPV2V with a 50K times reduction relative to uncompressed features; for segmentation, +5.69% mIoU at a 1K times reduction. iii) RDcomm outperforms previous communication-efficient methods with significantly reduced communication cost in detection: 15/13 times less (LiDAR/camera) on DAIR-V2X, 30/108 times less on OPV2V, 32 times less on V2XSeq, 4 times less on V2V4Real, and 8 times less for OPV2V segmentation.

Ablation on coding method. Fig. 4a compares the proposed task entropy coding in RDcomm against: i) classic entropy coding weighted by occurrence frequency Huffman (2007); ii) fixed-length coding Hu et al. (2024), where code length is $\log 2$ of the codebook volume. We see that task entropy coding saves 83/57%(detection/segmentation) in communication volume compared to fixed-length coding, and 30/25% compared to occurrence-driven entropy coding. The reason is that our task entropy coding prioritizes pragmatic-rich codes by assigning them shorter codewords, whereas classic entropy coding may waste short codewords on high-frequency but pragmatically weak codes.

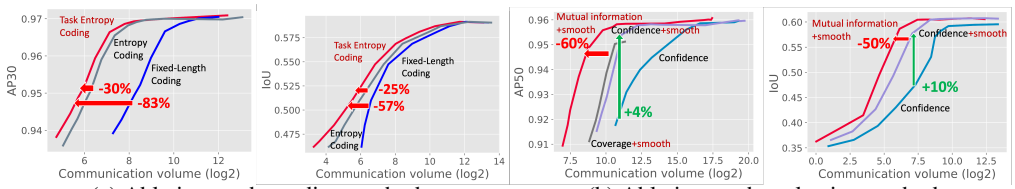
Ablation on selection method. Fig. 4b compares our mutual-information-driven (MI) redundancy selection against other redundancy selection: i) confidence-based Hu et al. (2024); ii) LiDAR-coverage-based Xu et al. (2025). Our MI selection reduces communication volume by 60/50% (detection/segmentation) relative to these baselines. The gain arises since we leverages a pragmatic



(a) Vehicle (b) Drive area (c) Lane (d) Mean

Figure 3: RDcomm achieves the best performance–communication trade-off in BEV segmentation.

yet lightweight abstraction to identify redundancy, providing richer cues than one-dimensional confidence or coverage signals. Besides, the smoothing module improves AP50 by 4% and IoU by 10%, demonstrating its effectiveness in mitigating sparsity under high selection rates.



(a) Ablation on the coding method (b) Ablation on the selection method

Figure 4: Ablation study on OPV2V detection/segmentation, evaluating coding and selection methods.

The approximation of optimal conditions. Fig. 5 illustrates RDComm's impact on conditional entropy $\mathrm{H}(Z_{s\to r}|Y)$ in (11) and the mutual information $\mathrm{I}(Z_{s\to r}; X_r)$ in (12). The results show that: i) As RDcomm removes 0-90% redundant information via mutual-information-driven selection, $\mathrm{I}(Z_{s\to r}; X_r)$ drops from 2.16 to 0.74 with negligible impact on detection accuracy, indicating effective redundancy removal of RDcomm. ii) As task-irrelevant information is reduced by 0-99% through task-entropy-based discrete coding, the conditional entropy $\mathrm{H}(Z_{s\to r} \mid Y)$ decreases from 0.39 to 0.08, with only marginal performance loss, demonstrating effective preservation of task-relevant cues. The evaluation is on OPV2V LiDAR detection, see more details in Appendix A.5.4.

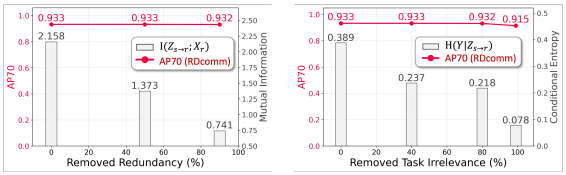


Figure 5: Effects on $\mathrm{H}(Z_{s\to r}|Y)$, $\mathrm{I}(Z_{s\to r}; X_r)$.

Table 2: Robustness against transmission latency and pose noise. We highlight **first**/second-place performances.

Method	V2V4Real AP50↑					DAIR-V2X AP50↑				
	Ideal	Latency (ms) 200	Latency (ms) 400	Pose noise (m/°) 0.2/0.2	Pose noise (m/°) 0.4/0.4	Ideal	Latency (ms) 200	Latency (ms) 400	Pose noise (m/°) 0.2/0.2	Pose noise (m/°) 0.4/0.4
Late fusion	0.540	0.474	0.462	0.500	0.459	0.719	0.625	0.612	0.632	0.608
Disconet Li et al. (2021)	0.622	0.527	0.502	0.576	0.483	0.688	0.651	0.625	0.656	0.637
V2X-ViT Xu et al. (2022b)	0.665	0.587	0.569	0.625	0.561	0.720	0.719	0.705	0.721	0.709
CoST Tang et al. (2025)	0.705	0.618	0.559	0.650	0.558	0.743	0.708	0.648	0.716	0.679
RDcomm (ours)	**0.726**	**0.631**	**0.593**	**0.672**	**0.581**	**0.782**	**0.754**	**0.731**	**0.768**	**0.724**
No collaboration	0.516					0.664				

Robustness to transmission latency and pose noise. Tab. 2 reports the robustness of RDcomm under transmission latency and pose noise on V2V4Real and DAIR-V2X datasets. We evaluate latency at 200 ms and 400 ms. Pose noise follows CoAlign Lu et al. (2023), where Gaussian noise is added to both location and orientation with zero mean and standard deviations of 0.2m/0.2° and 0.4m/0.4°. We see that RDcomm outperforms baselines and the no-collaboration setting across all tested conditions. We owe this gain to two aspects. First, the UNet smoothing module helps propagate visual evidence from perturbed regions to their correct spatial locations. Second, to boost model robustness, we incorporate communication constraints and noise into the training process, enhancing the generalizability of RDcomm to disturbed scenarios.

Inference cost. We evaluate the parameter size and inference time of the fusion/communication module while excluding the perception backbone. The results show that the proposed RDcomm method is both lightweight and efficient: the communication module requires only 3.75 MB of GPU memory and 14.88 ms per inference. At the same time, RDcomm achieves strong detection performance with this lightweight design, demonstrating its suitability for practical deployment.

Table 3: Parameter size, average inference time, and accuracy on OPV2V LiDAR detection with RDcomm.

Method	Params	Infer time	AP50↑
V2X-ViT Xu et al. (2022b)	20.58 MB	87.19 ms	0.850
CoBEVT Xu et al. (2022a)	9.37 MB	23.25 ms	0.566
mmCooper Liu et al. (2025)	3.83 MB	21.72 ms	0.810
RDcomm (ours)	**3.75 MB**	**14.88 ms**	**0.933**

Table 4: Segmentation performance under different compressors. bpp: bits-per-pixel.

Method	mIoU↑	bpp↓
VQVAE-32 Zhu et al. (2022)	0.38 (82%)	10
VQVAE-128 Zhu et al. (2022)	0.40 (87%)	14
RDcomm-128 (ours)	0.44 (95%)	4
No compression	0.46 (100%)	4096

Lossless bit-rate. We compare the bit-rate of RDcomm with the optimal bit-rate $\mathrm{Rate}(\delta)$ in (10) under lossless pragmatic compression, i.e., $\delta = 0$. To reduce the error in estimating $\mathrm{Rate}(0)$, we exclude the receiver's information X_r in the experiment; the optimal rate is then $\mathrm{I}(Y; X_s)$, which is tightly upper-bounded by $\mathrm{H}(Y) \approx \log_2(4)$, i.e., 2 bits-per-pixel (bpp). Here we define "lossless" as a performance drop of less than 5%. Tab. 4 compares RDcomm's BEV segmentation performance with the no-compression scheme. RDcomm attains 95% of mIoU with 4 bpp, close to the 2 bpp upper bound. Here bpp describes BEV feature. We also report the performance of residual VQVAE Zhu et al. (2022) (codebook sizes 32/128, segment number 2), where RDcomm uses a codebook size of 128. RDcomm delivers higher accuracy at a lower rate, indicating that effective pragmatic compression cannot be achieved by merely tuning codebook size, requires selective transmission and shorter codewords for task-relevant codes.

Cost for transmitting abstract. Tab. 5 reports the share of bandwidth consumed by transmitting the pragmatic abstraction $\hat{F}_{sc}^{q}$ on DAIR-V2X detection. We observe that abstraction transmission accounts for only 9%–11% of the total communication volume, yet is effective to identify redundancy as showed in Fig. 4b.

Table 5: Allocation of communication.

AP30	0.82	0.79	0.76
total bits	9054	449	166
abstract bits	882(9%)	49(11%)	19(11%)

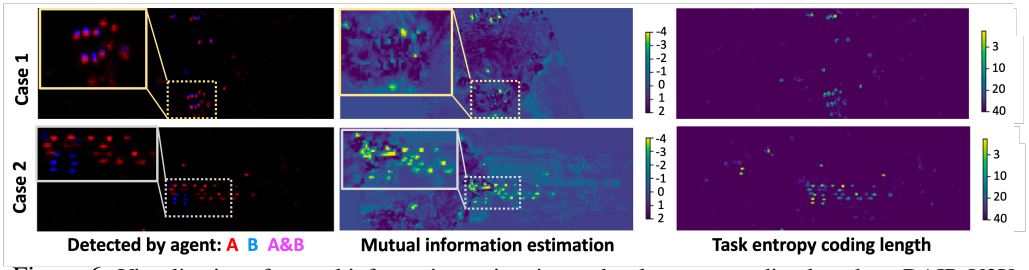


Figure 6: Visualization of mutual information estimation and task entropy coding length on DAIR-V2X.

5.2 Qualitative Analysis

Fig. 6 visualizes mutual-information estimates and task-entropy code lengths in two cases. In case 1, both agents A and B detect the same vehicles, forming high–mutual-information regions. In case 2, A and B detect different vehicles, yielding low–mutual-information regions that are prioritized for sharing. We also observe that task-entropy coding assigns short codewords to task-relevant regions. Long codewords are assigned to background areas, which are omitted when bandwidth is limited.

6 Conclusions

This work investigates the trade-off between task performance and communication volume from an information-theoretic perspective. We formulate a pragmatic rate–distortion theory for collaborative perception, deriving the optimal bit rate for message transmission and two necessary conditions for optimal compression: pragmatic-relevant and redundancy-less. Guided by these two conditions, we propose RDcomm, a communication-efficient collaborative perception method with two novel components: i) task entropy discrete coding and ii) mutual-information-driven message selection. Experiments covering both detection and segmentation show that RDcomm achieves state-of-the-art perception–communication trade-offs across both LiDAR and camera modalities.

Limitations and future work. Our study focuses on perception tasks. Future directions include extending the framework to broader tasks such as navigation, manipulation, and scene captioning, and incorporating additional modalities such as motion and language.

REFERENCES

Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. *arXiv preprint arXiv:1802.01436*, 2018.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pp. 531–540. PMLR, 2018.

Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *International Conference on Machine Learning*, pp. 675–685. PMLR, 2019.

Qi Chen, Xu Ma, Sihai Tang, Jingda Guo, Qing Yang, and Song Fu. 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*, pp. 88–100, 2019.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pp. 1–16. PMLR, 2017.

Yann Dubois, Benjamin Bloem-Reddy, Karen Ullrich, and Chris J Maddison. Lossy compression for lossless prediction. *Advances in Neural Information Processing Systems*, 34:14014–14028, 2021.

Xiangbo Gao, Runsheng Xu, Jiachen Li, Ziran Wang, Zhiwen Fan, and Zhengzhong Tu. Stamp: Scalable task and model-agnostic collaborative perception. *arXiv preprint arXiv:2501.18616*, 2025.

Yushan Han, Hui Zhang, Huifang Li, Yi Jin, Congyan Lang, and Yidong Li. Collaborative perception in autonomous driving: Methods, datasets, and challenges. *IEEE Intelligent Transportation Systems Magazine*, 15(6):131–151, 2023.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. *Advances in neural information processing systems*, 35:4874–4886, 2022.

Yue Hu, Juntong Peng, Sifei Liu, Junhao Ge, Si Liu, and Siheng Chen. Communication-efficient collaborative perception via information filling with codebook. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15481–15490, 2024.

David A Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, 2007.

Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12697–12705, 2019.

Maosen Li, Siheng Chen, Ya Zhang, and Ivor Tsang. Graph cross networks with vertex infomax pooling. *Advances in neural information processing systems*, 33:14093–14105, 2020.

Yiming Li, Shunli Ren, Pengxiang Wu, Siheng Chen, Chen Feng, and Wenjun Zhang. Learning distilled collaboration graph for multi-agent perception. *Advances in Neural Information Processing Systems*, 34:29541–29552, 2021.

Bingyi Liu, Jian Teng, Hongfei Xue, Enshu Wang, Chuanhui Zhu, Pu Wang, and Libing Wu. mmcooper: A multi-agent multi-stage communication-efficient and collaboration-robust cooperative perception framework. *arXiv preprint arXiv:2501.12263*, 2025.

Yifan Lu, Quanhao Li, Baoan Liu, Mehrdad Dianati, Chen Feng, Siheng Chen, and Yanfeng Wang. Robust collaborative 3d object detection in presence of pose errors. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4812–4818. IEEE, 2023.

Yifan Lu, Yue Hu, Yiqi Zhong, Dequan Wang, Yanfeng Wang, and Siheng Chen. An extensible framework for open heterogeneous collaborative perception. *arXiv preprint arXiv:2401.13964*, 2024.

Guiyang Luo, Hui Zhang, Quan Yuan, and Jinglin Li. Complementarity-enhanced and redundancy-minimized collaboration network for multi-agent perception. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 3578–3586, 2022.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. *Advances in neural information processing systems*, 29, 2016.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pp. 194–210. Springer, 2020.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Claude E Shannon et al. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec*, 4(142-163):1, 1959.

Jiawei Shao, Teng Li, and Jun Zhang. Task-oriented communication for vehicle-to-infrastructure cooperative perception. In *2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6. IEEE, 2024.

Wanfang Su, Lixing Chen, Yang Bai, Xi Lin, Gaolei Li, Zhe Qu, and Pan Zhou. What makes good collaborative views? contrastive mutual information maximization for multi-agent perception. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 17550–17558, 2024.

Zongheng Tang, Yi Liu, Yifan Sun, Yulu Gao, Jinyu Chen, Runsheng Xu, and Si Liu. Cost: Efficient collaborative perception from unified spatiotemporal perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1120–1129, 2025.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Rujia Wang, Xiangbo Gao, Hao Xiang, Runsheng Xu, and Zhengzhong Tu. Cocmt: Communication-efficient cross-modal transformer for collaborative perception. *arXiv preprint arXiv:2503.13504*, 2025.

Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyuan Zeng, and Raquel Urtasun. 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*, pp. 605–621. Springer, 2020.

Junhao Xu, Yanan Zhang, Zhi Cai, and Di Huang. Cosdh: Communication-efficient collaborative perception via supply-demand awareness and intermediate-late hybridization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6834–6843, 2025.

Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Liu, and Jiaqi Ma. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2583–2589, 2021.

Runsheng Xu, Zhengzhong Tu, Hao Xiang, Wei Shao, Bolei Zhou, and Jiaqi Ma. CoBEVT: Cooperative bird's eye view semantic segmentation with sparse transformers. *CoRL*, 2022a.

Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In *European conference on computer vision*, pp. 107–124. Springer, 2022b.

Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, et al. V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13712–13722, 2023.

Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11784–11793, 2021.

Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, et al. DAIR-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. *In Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition (CVPR)*, 2022.

Haibao Yu, Wenxian Yang, Hongzhi Ruan, Zhenwei Yang, Yingjuan Tang, Xu Gao, Xin Hao, Yifeng Shi, Yifeng Pan, Ning Sun, et al. V2x-seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5486–5495, 2023.

Xiaosu Zhu, Jingkuan Song, Lianli Gao, Feng Zheng, and Heng Tao Shen. Unified multivariate gaussian mixture for efficient neural image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17612–17621, 2022.

A Appendix

A.1 Statements

LLM Usage. We used an LLM (ChatGPT) solely for language refinement, such as improving grammar, clarity, and readability of sentences. The research ideas, methodology, experiments, and overall writing structure were entirely developed by the authors.

A.2 Related works

Mutual information neural estimation. Mutual information measures the statistical dependence between two variables, yet it is historically difficult to compute, especially in high-dimensional settings. To address this, mutual information neural estimation Belghazi et al. (2018) provides a general-purpose solution by reformulating mutual information as the Kullback–Leibler divergence between the joint distribution and the product of marginals, and then maximizing a variational lower bound to obtain an estimate. Beyond direct estimation, contrastive learning is often interpreted through the lens of mutual information. Deep InfoMax Hjelm et al. (2018) maximizes mutual information between the input and the output of an encoder; Contrastive Predictive Coding Oord et al. (2018) employs the InfoNCE loss, which serves as a variational lower bound on the mutual information between context representations and future latent variables; and SimCLR Chen et al. (2020) similarly builds upon InfoNCE as the core objective for learning invariant visual representations.

Mutual information estimation has also been adopted in collaborative perception. CMiMC Su et al. (2024) focuses on obtaining an effective feature fusion module, it introduces a feature fusion module that maximizes mutual information before and after fusion to preserve each collaborator's local information; CRCNet Luo et al. (2022) also focuses on obtaining an effective fusion module, it transmits full feature maps and then minimizes inter-view mutual information to encourage complementary representations. These approaches typically use mutual information estimation as an auxiliary training signal, without employing the estimator during inference, and seldom analyze the meaning or utility of the estimated values. In contrast, our work focuses on communication efficiency and incorporates mutual information from a rate–distortion perspective. We show that minimizing mutual information between collaborators provides an optimality condition for efficient communication, which explicitly motivates the estimation of mutual information between collaboration pairs,

and guides the redundancy-less feature selection on regions with low inter-agent mutual information estimates.

A.3 TRAINING DETAILS

Training cost analysis. To further illustrate the training process described in Sec. 4.3, we present a detailed breakdown of the training cost for RDcomm's three-stage pipeline.

- Stage 1 ($\sim$68%) trains the BEV encoder Φ_{bev} and task decoder Φ_{task}, taking approximately 25 GPU·h on an RTX 3090 (DairV2X camera detection). This is comparable to existing collaborative perception methods, as we implement all baselines with the same BEV encoder and task decoder configuration.
- Stage 2 ($\sim$8%) trains the vector quantization module Φ_{vq} as a compressor. This stage converges quickly, taking approximately 3 GPU·h on an RTX 3090.
- Stage 3 ($\sim$24%) trains the mutual information estimator Φ_{MI}, taking about 9 GPU·h on an RTX 3090.

We also compare the training cost of the proposed RDcomm with that of existing methods on the DAIR-V2X camera-based detection task. RDcomm employs a three-stage training scheme, with 40 epochs in Stage 1, 20 epochs in Stage 2, and 80 epochs in Stage 3. For comparison, we train V2XViT Xu et al. (2022b) for 40 epochs, DiscoNet Li et al. (2021) for 40 epochs each for its teacher and student models, and CodeFilling Hu et al. (2024) for 40 epochs for the backbone followed by an additional 30 epochs for its compression module. Overall, the training cost of RDcomm is comparable to these baselines while achieving superior detection performance.

Table 6: Comparison of training cost in DAIR-V2X camera-based detection.

Method	Training cost (3090 GPU*h)	AP30 (DAIR-V2X camera)
V2XViT Xu et al. (2022b)	$\sim$30	0.224
DiscoNet Li et al. (2021)	$\sim$50	0.288
CodeFilling Hu et al. (2024)	$\sim$30	0.289
RDcomm (ours)	$\sim$37	0.293

Discussion on the training of mutual information estimation. Our method is not sensitive to potential bias in estimating absolute mutual information value, we only need the estimator to assign the redundant messages with a relatively lower value than the non-redundant ones. While it is well known that estimating absolute MI values in high-dimensional spaces is challenging, our framework does not depend on obtaining accurate absolute mutual information measurements. Instead, the MI estimator serves as a distribution discriminator, allowing us to distinguish feature pairs that are relatively more "independent" (i.e., sampled from the independent distribution and considered complementary) from those that are more "joint" (i.e., sampled from the joint distribution and considered redundant). We do not rely on obtaining accurate estimates of the absolute mutual information between the two variables; we only require that redundant feature pairs be assigned relatively lower values by the estimator.

To improve the stability of training the mutual information estimator, we adopt re-balanced feature sampling to construct positive and negative pairs. Specifically, for all pairs, we ensure that the ratio of foreground to background regions is approximately 5:1. This is because we aim to focus the MI estimation on foreground regions, which contain more task-relevant information, while excessive training on background regions could disturb the estimator's ability to distinguish redundancy between task-related features.

A.4 MODEL STRUCTURE

A.4.1 SINGLE-AGENT PERCEPTION PIPELINE

The perception pipeline comprises two components: a BEV encoder and a task-specific decoder. The BEV encoder projects sensor inputs into bird's-eye-view (BEV) representations, enabling consistent spatial alignment and collaboration across different views. Task-specific decoders are then applied for downstream detection or segmentation tasks.

BEV Encoder. Our framework supports either LiDAR or camera inputs, where we commonly denote the observation of the ith agent as X_i. We obtain BEV feature as $F_i = \Phi_{enc}(X_i) \in \mathbb{R}^{h \times w \times c}$, where $\Phi_{enc}(\cdot)$ denotes the complete BEV encoder. F_i then serves as the information source for selection and compression. For LiDAR inputs, we adopt the widely used PointPillars encoder Lang et al. (2019) to extract BEV features from point clouds. For camera inputs, we employ the Lift-Splat-Shoot Philion & Fidler (2020) module following Lu et al. (2024), which lifts image features into 3D frustums and aggregates them into the BEV plane via learned depth distributions. For both LiDAR and camera modality, the extracted BEV features are further processed by a 2D convolutional ResNeXt-based backbone Lu et al. (2024).

Decoder. Based on the BEV feature F_i, we incorporate task-specific decoders. For 3D object detection, the decoder consists of a classification head, a box regression head, and a direction estimation head to predict object bounding boxes, following Lu et al. (2024); Hu et al. (2022). For BEV semantic segmentation, we employ a MLP as decoder to produce dense, per-pixel semantic predictions, following Xu et al. (2022a).

A.5 Experiment analysis

A.5.1 Thresholds of spatial selection

We provide further analysis and ablation on the two thresholds: τ_c in Sec. 4.1 and τ_{MI} in Sec. 4.2. Recall that $\tau_c \in [0, 1]$ is the threshold to select high-confidence regions, features with confidence above τ_c will be transmitted; $\tau_{MI} \in [0, 1]$ is for selecting low-redundancy regions, features with mutual-information estimation value $\sigma(\Phi_{MI})$ below τ_{MI} will be transmitted. Our strategy to set threshold for different bandwidth limit is: set a fixed $\tau_{MI} = 0.7$ and adjust τ_c to adapt to various bandwidth limit, this strategy selects the high-confidence regions and then remove the redundant ones.

In Tab. 7 and Tab. 8 we study the individual effect of τ_{MI} and τ_c on the perception–communication trade-off by varying each hyperparameter. The evaluations are conducted on OPV2V LiDAR detection. In Tab. 7, we examine performance under different τ_{MI} while fixing $\tau_c = 0.005$. We see that as τ_{MI} decreases, the selector filters out more redundancy, reducing the proportion of selected regions. Remarkably, even when only 1% of regions are selected, AP70 drops by only 1% (from 0.934 to 0.920), demonstrating that our mutual-information-driven selection effectively identifies the most critical regions for collaboration. In Tab. 8, we evaluate performance under different τ_c with $\tau_{MI} = 0$. We see that as τ_c increases, the proportion of selected regions decreases. Specifically, when $\tau_c = 0.01$, the selected regions account for only 3% of the total area, yet AP50 drops only slightly (0.967 to 0.964). This demonstrate the sparsity of task-relevant information in the features.

Table 7: OPV2V LiDAR detection performance under different threshold τ_{MI}, where $\tau_c = 0.005$.

τ_{MI}	AP50	AP70	Selection ratio
0.9000	0.967	0.934	48.944%
0.7000	0.966	0.933	18.607%
0.5000	0.966	0.929	2.782%
0.2500	0.966	0.925	1.711%
0.1000	0.965	0.920	1.006%
0.0500	0.963	0.911	0.579%
0.0200	0.960	0.897	0.342%
0.0067	0.955	0.878	0.151%
0.0025	0.948	0.858	0.046%
0.0009	0.919	0.824	0.004%
0.0003	0.904	0.817	0.000%
0.0000	0.903	0.816	0.000%

Table 8: OPV2V LiDAR detection performance under different threshold τ_c, where $\tau_{MI} = 1$.

τ_c	AP50	AP70	Selection ratio
0.000	0.966	0.933	94.75%
0.001	0.966	0.933	86.31%
0.003	0.966	0.931	64.57%
0.005	0.966	0.928	48.94%
0.007	0.964	0.922	8.69%
0.010	0.964	0.915	3.62%
0.030	0.958	0.886	0.43%
0.050	0.954	0.873	0.22%
0.070	0.950	0.867	0.15%
0.100	0.946	0.861	0.10%
0.200	0.935	0.850	0.05%
0.400	0.921	0.836	0.03%
0.700	0.911	0.826	0.01%
1.000	0.903	0.816	0.00%

A.5.2 PERFORMANCE IN MULTI-AGENT COLLABORATION

We further investigate the performance under varying numbers of collaborating agents on OPV2V LiDAR detection (communication <1 KB). Tab. 9 compares RDcomm with CodeFilling Hu et al. (2024), a method specifically designed for modeling multi-agent redundancy. We see that RDcomm consistently outperforms CodeFilling across all agent-number settings, demonstrating both the efficiency of our communication strategy and its scalability to multi-agent scenarios.

Table 9: Performance under different numbers of collaborator.

OPV2V LiDAR (AP50/AP70), communication <1 KB				
Method	2 agents	3 agents	4 agents	5 agents
CodeFilling Hu et al. (2024)	0.874/0.796	0.918/0.855	0.927/0.864	0.929/0.867
RDcomm (ours)	**0.959/0.899**	**0.965/0.925**	**0.967/0.932**	**0.968/0.933**

A.5.3 EFFECTIVENESS OF MUTUAL INFORMATION NEURAL ESTIMATION

We further compare two types of mutual information estimators: i) a KL-divergence–based estimator Belghazi et al. (2018), and ii) the GAN-style divergence estimator Nowozin et al. (2016) (the one adopted in our method). We evaluate their effectiveness by examining the downstream perception performance under each estimator. It is worth noting that classical statistical MI estimators (e.g., kNN-based estimators) cannot be integrated into the proposed RDcomm framework. These methods estimate mutual information between two variables only after observing a large set of samples, and therefore cannot evaluate whether a single message should be selected to reduce mutual information. In contrast, our approach reduces the MI between two variables by selecting feature pairs that are likely to be drawn from the independent distribution.

Table 10: Comparison between two types of mutual information neural estimator.

Method	OPV2V LiDAR (AP50)			DAIR-V2X LiDAR (AP50)		
	no bandwidth limit	<1KB	<0.06KB	no bandwidth limit	<1KB	<0.06KB
Random selection	0.967	0.906	0.902	0.782	0.674	0.662
Confidence selection Hu et al. (2022)	0.967	0.94	0.914	0.782	0.756	0.737
RDcomm selector with KL divergence Belghazi et al. (2018)	0.967	0.948	0.922	0.782	0.762	0.743
RDcomm selector with GAN-style divergence Nowozin et al. (2016)	0.967	**0.965**	**0.954**	0.782	**0.777**	**0.760**

Here we report the LiDAR detection performance (AP50) in both OPV2V and DairV2X datasets under two communication-volume constraints (<1 KB and <0.06 KB). We observe that: i) the mutual-information–driven selector, using either the KL-based or GAN-style divergence estimator, consistently outperforms confidence-based selection Hu et al. (2022) and random selection, indicating that our selector is effective at identifying the critical complementary messages; ii) the GAN-style divergence estimator achieves better performance than the KL-based estimator. We hypothesize that this advantage arises from the symmetry of the GAN-style divergence objective (19) with respect to positive and negative sample pairs, while the KL-based estimator adopts an asymmetric formulation (18), which may make optimization more difficult and hinder stable convergence.

A.5.4 DETAILS OF ESTIMATING THE APPROXIMATION OF OPTIMAL CONDITIONS.

We illustrate the details in estimating $\mathrm{I}(Z_{s\to r}; X_r)$ and $\mathrm{H}(Z_{s\to r} \mid Y)$ in Fig. 5(a)(b). In our experiments, both $Z_{s\to r}$ and X_r are 64-dimensional BEV feature vectors, and the label Y is a 9-dimensional vector used for 3D detection. Such high-dimensional feature spaces pose substantial challenges for estimating entropy and mutual information, as traditional statistical estimators may suffer from the curse of dimensionality and could lead to highly biased results. Trying to mitigate this issue, we employ dense sampling and collect 100K triplets $(Z_{s\to r}, X_r, Y)$ to reduce estimation variance. For mutual information, we estimate $\mathrm{I}(Z_{s\to r}; X_r)$ by computing the average pairwise mutual information across each feature dimension using a kNN-based estimator. The conditional entropy $\mathrm{H}(Z_{s\to r} \mid Y)$ is approximated via the decomposition $\mathrm{H}(Z_{s\to r}) - \mathrm{I}(Z_{s\to r}; Y)$, where $\mathrm{I}(Z_{s\to r}; Y)$ is also estimated using kNN-based methods. All evaluations are performed on the OPV2V LiDAR detection benchmark.

A.6 THEORY

In this section, we provide: i) further discussion on our problem formulation; ii) proof of the proposed propositions and theories.

A.6.1 DISCUSSION ON PROBLEM FORMULATION

Note that our theory formulation equation 2 equation 3 is not an ad hoc assumption, but consistent with the learning objects in pragmatic compression . Table 11 reveals that the optimization objective equation 3 is dual-equivalent to several prior approaches Hu et al. (2022); Ballé et al. (2018), as they all share the same Lagrangian target $\min \mathcal{D} + \lambda\mathcal{R}$, which is a weighted sum of distortion $\mathcal{D}$ and communication rate $\mathcal{R}$ with weight λ. In the Section 3.2 we make the task distortion $\mathrm{D}_Y[X_s, Z_s \mid X_r]$ explicit.

Table 11: Problem formulations of bandwidth-constrained collaboration. $\mathcal{R}$: bit-rate, $\mathcal{D}$: distortion/loss.

Formulation Type	Optimization target	Distortion/loss $\mathcal{D}$
Constrained task loss Hu et al. (2022)	$\min \mathcal{D}$ s.t. $\mathcal{R} \leq \delta$	task loss function
Task-compression joint loss Ballé et al. (2018)	$\min \mathcal{D} + \lambda\mathcal{R}$	task loss function
Pragmatic rate-distortion (ours)	$\min \mathcal{R}$ s.t. $\mathcal{D} \leq \delta$	pragmatic distortion

As an extreme case, consider early collaboration Han et al. (2023), where agents directly transmit raw sensor data (i.e., $Z_s = X_s$). In this case, the communication volume becomes $\mathrm{I}(X_s; Z_s) = \mathrm{H}(X_s)$, corresponding to the full information content of X_s, and the distortion $\mathrm{D}_Y[X_s, X_s|X_r]$ is zero.

A.6.2 DISCUSSION ON PRAGMATIC DISTORTION

From Tab. 8 we see that: i) for collaborative BEV segmentation, pragmatic distortion is expressed as the gap in conditional entropy, based on the Bayes risk $\mathrm{B}_{risk}[Y|X] = \frac{1}{|Y|}\sum_{i\in\mathcal{S}} \mathrm{H}\left(Y_{(i)}|Z_s, X_r\right)$; ii) for collaborative 3D detection, the distortion further incorporates an exponential term of conditional entropy, where $\mathcal{K} = \{loc, size, ori\}$ denotes the set of regression losses (location, size, and orientation), which contribute more uncertainty than classification; iii) compared to the widely used MSE distortion in image reconstruction Ballé et al. (2018), the pragmatic distortion defined in our theory differs in two aspects: first, it emphasizes task uncertainty rather than fidelity, second, it accounts for the receiver's information (e.g., X_r) to analyze redundancy.

A.6.3 PROOF OF PROPOSITION 1: BAYES RISK $\mathrm{R}[Y|X]$ FOR PERCEPTION TASKS

Proposition 1 *(Bayes risk $\mathrm{R}[Y|X]$ for perception tasks, see proof in A.6.3). Given an observation input X and a perception task target Y, we focus on the Bayes risk $\mathrm{R}[Y|X]$ to measure the difficulty of predicting Y from X. In object detection task, the detection results and corresponding target label are denoted as $\hat{Y}, Y \in \mathbb{R}^{h\times w\times(8+K)}$, where the $(8+K)$ channels stand for classification heatmap $\hat{Y}_c, Y_c \in \mathbb{R}^{h\times w\times K}$, offset $\hat{Y}_o, Y_o \in \mathbb{R}^{h\times w\times 3}$, size $\hat{Y}_s, Y_s \in \mathbb{R}^{h\times w\times 3}$, rotation $\hat{Y}_r, Y_r \in \mathbb{R}^{h\times w\times 2}$. Here $[h, w]$ denotes the BEV perception range. The total loss is shown in equation 21:*

$$L_{total} = L_{heatmap} + \lambda_2 L_{offset} + \lambda_3 L_{size} + \lambda_4 L_{rotation} \tag{21}$$

where $\lambda_2, \lambda_3, \lambda_4$ are the loss weights. Consider N objects involved in ground truth, the heatmap loss optimizes the model to classify the foreground object from background, where we adopt the focal loss $L_{focal}(y, \hat{y}) = -\sum_{k=1}^{K} \alpha_k (1-\hat{y}_k)^\gamma y_k \log \hat{y}_k$, where α_k, γ are hyper-parameters in focal loss, here we consider a simplified situation that $\alpha_k = 1, \gamma = 0$ and the loss degenerates into cross entropy loss L_{ce}. L_{offset} ,L_{size} , and $L_{rotation}$ are L1 loss. Specifically, we consider the situation that the elements in Y_o follow Gaussian distribution $Y_{o(i,j)}|X \sim \mathcal{N}(\mu_o(X), \sigma_o^2(X))$, $Y_{s(i,j)}|X \sim \mathcal{N}(\mu_s(X), \sigma_s^2(X)), Y_{r(i,j)}|X \sim \mathcal{N}(\mu_r(X), \sigma_r^2(X))$, and the number of objects is $\overline{N_{obj}}$. The Bayes risk of object detection with centerpoint detection loss is given as equation 22:

$$\mathrm{R}[Y|X] = \sum_{i\leq h, j\leq w} \mathrm{H}(Y_{c(i,j)}|X) + \overline{N_{obj}}\sqrt{2/\pi}\left(\lambda_2\sigma_o(X) + \lambda_3\sigma_s(X) + \lambda_4\sigma_r(X)\right) \tag{22}$$

When the task Y refers to occupancy prediction, the regression terms are put off and the Bayes risk $\mathrm{R}[Y|X]$ solely consists of the terms of conditional entropy, as shown in equation 23.

$$\mathrm{R}[Y|X] = \sum_{i \le h, j \le w} \mathrm{H}(Y_{c(i,j)}|X) = \sum_{i \le h, j \le w} \mathrm{H}(Y_{(i,j)}|X) \tag{23}$$

In each communication round, messages are transmitted between connected agents as shown in equation 2, where the connection is established by the pre-defined collaboration principle. We denote the observation of message sender/reciever as X_s, X_r, the perception target as Y. The message $\mathcal{P}_{s\to r}$ is obtained via $\mathcal{P}_{s\to r} = C(X_s)$, where $C(\cdot)$ is a compressor that reduces the transmission bit-rate. The pragmatic distortion is defined in equation 24, where Y is the perception target, $\mathrm{R}[Y|X]$ denotes the Bayes risk when predicting Y from X, $\mathrm{R}[Y|X_r, X_s]$ denotes the Bayes risk when predicting Y from the fused information of X_r, X_s.

$$\mathrm{D}_Y\left[X_s, Z_s\right] = \mathrm{R}\left[Y \mid X_r, Z_s\right] - \mathrm{R}\left[Y \mid X_r, X_s\right] \tag{24}$$

To analyze this distortion in perception tasks, we need to:

- Give the specific formulation of Bayes risk $\mathrm{R}[Y|X_r]$(single perception) and $\mathrm{R}[Y|X_r, X_s]$(collaborative perception) in detection 3D task with centerpoint loss(for example).
- Reformulate the distortion D_Y by introducing the task related Bayes risk $\mathrm{R}[Y|X_r, X_s]$.
- Reformulate the distortion D_Y by introducing the supply-request information.

Definition 1 *(Bayes risk). Let $X \in \mathcal{X}$ be the input variable (features, observed data), $Y \in \mathcal{Y}$ be the target variable (labels), $P(X, Y)$ denote the joint probability distribution of X and Y, $L(Y, \hat{Y})$ be the loss function quantifying the discrepancy between a prediction $\hat{Y} = f(X)$ and the true value Y, and $f : \mathcal{X} \to \mathcal{Y}$ be a predictive model. The Bayes risk is defined as the infimum of the expected loss over all possible decision functions, as shown in equation 25:*

$$\mathrm{R} = \inf_f \mathbb{E}_{X,Y}\left[L(Y, f(X))\right] \tag{25}$$

Bayes risk is the minimum achievable loss by an ideally trained model. It captures unavoidable uncertainty in the data, such as the ambiguity due to overlapping classes in classification tasks or stochastic noise in target variables for regression tasks. For any model f, the expected loss satisfies $\mathbb{E}[L(Y, f(X))] \geq \mathrm{R}_{Bayes}$. With a given loss function, the Bayes risk completely depends on the data distribution $P(X, Y)$, it indicates the "difficulty" of learning the projection $f : \mathcal{X} \to \mathcal{Y}$. Due to its property to characterize data distribution, we utilize the differences of Bayes risk to measure the pragmatic distortion.

Bayes risk for perception tasks In this section, we derive the Bayes risk of perception tasks with specific loss functions.

First, we review the formulation of centerpoint loss. Suppose that the observation from camera or LiDAR can be represented by 3D voxel feature $X \in \mathbb{R}^{D\times h\times w\times C}$, the detection results and corresponding target label are denoted as $\hat{Y}, Y \in \mathbb{R}^{h\times w\times(8+K)}$, where the $(8 + K)$ channels stand for classification heatmap $\hat{Y}_c, Y_c \in \mathbb{R}^{h\times w\times K}$, offset $\hat{Y}_o, Y_o \in \mathbb{R}^{h\times w\times 3}$, size $\hat{Y}_s, Y_s \in \mathbb{R}^{h\times w\times 3}$, rotation $\hat{Y}_r, Y_r \in \mathbb{R}^{h\times w\times 2}$. The total loss is:

$$L_{\text{total}} = L_{\text{heatmap}} + \lambda_2 L_{\text{offset}} + \lambda_3 L_{\text{size}} + \lambda_4 L_{\text{rotation}} \tag{26}$$

where $, \lambda_2, \lambda_3, \lambda_4$ are the loss weights. Consider N objects in ground truth, the heatmap loss optimizes the model to classify the foreground object from background, we utilize focal loss $L_{\text{focal}}\left(y, \hat{y}\right) = -\sum_{k=1}^{K} \alpha_k \left(1 - \hat{y}_k\right)^{\gamma} y_k \log \hat{y}_k$, where α_k, γ are hyper-parameters in focal loss, here we consider a simplified situation that $\alpha_k = 1, \gamma = 0$ and the loss degenerates into cross entropy loss L_{ce}. L_{offset} ,L_{size} , and L_{rotation} are L1 loss.

Now we derive the Bayes risk in 3D object detection with centerpooint detection loss. To simplify the formulation, we approximately decompose the total Bayes risk into the sum of Bayes risk on each location as shown in equation 27:

$$\mathrm{R}[Y|X] = \inf_f \mathbb{E}_{X,Y}\left[L(Y, f(X))\right] = \inf_f \sum_{i\le h, j\le w} \mathbb{E}_{X,Y_{(i,j)}}\left[L(Y_{(i,j)}, f(X)_{(i,j)})\right] \tag{27}$$

We regard the perception task at each region as independent tasks, and we define the located Bayes risk for perception tasks in equation 28:

$$\mathrm{R}[Y_{(i,j)}|X] = \inf_f \mathbb{E}_{X,Y_{(i,j)}}\left[L(Y_{(i,j)}, f(X)_{(i,j)})\right] \tag{28}$$

We derive the Bayes risk for two primarily used loss function involved in perception tasks: focal loss, MSE loss, and L1 loss.

Focal loss. The true distribution of $P(Y_c, X)$ satisfies $p(Y_{c(i,j,k)} = 1|X) = p_{i,j,k}$, and we have $P(Y_c|X) = \prod_{i,j,k} p(Y_{c(i,j,k)}|X)$ since different classes and locations are independent. The Bayes optimal prediction is the true conditional distribution: $\hat{Y}^*_{c(i,j,k)} = p_{(i,j,k)}$, then we have:

$$\begin{aligned}
\mathrm{R}[Y_{c(i,j)}|X] &= \mathbb{E}_{X,Y_{c(i,j)}} L_{ce}(Y_{c(i,j)}, p_{(i,j)}) && (29)\\
&= \mathbb{E}_X \sum_{Y_{c(i,j)}} p(Y_{c(i,j)}|X) L_{ce}(Y_{c(i,j)}, p_{(i,j)}) && (30)\\
&= \mathbb{E}_X \sum_{k=1}^{K} -p(Y_{c(i,j,k)} = 1|X) \log p(Y_{c(i,j,k)} = 1|X) && (31)\\
&= \mathrm{H}(Y_{c(i,j)}|X) && (32)
\end{aligned}$$

MSE loss. Given a specific X, we assume that the elements in offset target Y_o follow Gaussian distribution $Y_{o(i,j)}|X \sim \mathcal{N}(\mu_o(X), \sigma_o^2(X))$, and the number of objects is $\overline{N_{obj}}$. This assumption is reasonable, since minimizing MSE loss can be regarded as MLE(Maximum likelihood estimation) when $Y_{o(i,j)}|X \sim \mathcal{N}(\mu_o(X), \sigma_o^2(X))$. The Bayes optimal prediction is $f(X) = \hat{Y}^*_{o(i,j)} = \mu_o(X)$. Put this into equation 27, the Bayes risk is derived as $\overline{N_{obj}}\sqrt{2/\pi}\sigma_o(X)$. Similarly, we can derive the Bayes risk for the size and rotation targets by assuming their distributions follow a Gaussian distribution $Y_{s(i,j)}|X \sim \mathcal{N}(\mu_s(X), \sigma_s^2(X)), Y_{r(i,j)}|X \sim \mathcal{N}(\mu_r(X), \sigma_r^2(X))$. Combining the Bayes risk of the individual loss function described in equation 26, we obtain the Bayes risk of object detection with centerpoint detection loss as equation 33:

$$\mathrm{R}_{centerpoint}[Y|X] = \sum_{i\le h, j\le w} \mathrm{H}(Y_{c(i,j)}|X) + \overline{N_{obj}}\sqrt{2/\pi}\left(\lambda_2\sigma_o(X) + \lambda_3\sigma_s(X) + \lambda_4\sigma_r(X)\right) \tag{33}$$

L1 loss. Given a specific X, we assume that the elements in target Y_o follow Laplace distribution $p(Y_o \mid X) = \frac{1}{2b_{o|X}} \exp\left(-\frac{|Y-\mu_o(X)|}{b_{o|X}}\right)$. This assumption is reasonable, since minimizing L1 loss can be regarded as MLE(Maximum likelihood estimation) when $Y_{o(i,j)}|X \sim \frac{1}{2b_{o|X}} \exp\left(-\frac{|Y-\mu_o(X)|}{b_{o|X}}\right)$. The Bayes optimal prediction is $f(X) = \hat{Y}^*_{o(i,j)} = \mathrm{median}(Y_{o(i,j)} \mid X) = \mu_o(X)$. Put this into equation 27, the Bayes risk is derived as:

$$\mathrm{R}[Y_{o(i,j)}|X] = \mathbb{E}_{X,Y_{o(i,j)}} L_{l1}(Y_{o(i,j)}, \mu_o(X)) \quad \text{definition} \tag{34}$$

$$= \mathbb{E}_X \int_{-\infty}^{\infty} |Y_{o(i,j)} - \mu_o(X)| \cdot \frac{1}{2b_{o|X}} \exp\left(-\frac{|Y_{o(i,j)} - \mu_o(X)|}{b_{o|X}}\right) dY_{o(i,j)} \tag{35}$$

$$= \mathbb{E}_X \frac{1}{2b_{o|X}} \int_{-\infty}^{\infty} |z| \exp\left(-\frac{|z|}{b_{o|X}}\right) dz \quad z = Y_{o(i,j)} - \mu_o(X) \tag{36}$$

$$= \mathbb{E}_X \frac{1}{2b_{o|X}} 2b_{o|X}^2 \tag{37}$$

$$= \mathbb{E}_X b_{o|X} \tag{38}$$

$$= b_{o|X} \quad b_{o|X} \text{ is a constant} \tag{39}$$

On the other hand, when $p(Y_{o(i,j)} = y|X = x) = \frac{1}{2b_{o|X}} \exp\left(-\frac{|y-\mu_o(x)|}{b_{o|X}}\right)$, we can formulate the conditional entropy $\mathrm{H}(Y_{o(i,j)}|X)$ as:

$$\mathrm{H}(Y_{o(i,j)}|X) = \mathbb{E}_{x\sim X} \mathrm{H}(Y_{o(i,j)}|X = x) \tag{40}$$

$$= \mathbb{E}_{x\sim X} \int_{-\infty}^{\infty} -p(y) \log p(y)\, dy \tag{41}$$

$$= \mathbb{E}_{x\sim X} \int_{-\infty}^{\infty} -\frac{1}{2b_{o|X}} \exp\left(-\frac{|y - \mu_o(x)|}{b_{o|X}}\right) (\log \frac{1}{2b_{o|X}} - \frac{|y - \mu_o(x)|}{b_{o|X}})\, dy \tag{42}$$

$$= \mathbb{E}_{x\sim X} \log(2b_{o|X}) + 1 \tag{43}$$

$$= \log(2b_{o|X}) + 1 \quad b_{o|X} \text{ is a constant} \tag{44}$$

Combining equation 39 with equation 44, we have:

$$\mathrm{R}[Y_{o(i,j)}|X] = \frac{1}{2} \mathrm{e}^{\mathrm{H}(Y_{o(i,j)}|X)-1} \tag{45}$$

A.6.4 Proof of Tab. 1: pragmatic distortion for collaborative perception.

In this section, we derive the pragmatic distortion in collaborative perception. To achieve this, we start from decomposing the contribution of ego agent and other agents.

Now we derive the pragmatic distortion in collaborative perception task. Consider a simple scenario with 2 collaborators, and the observations/features of message sender and receiver are X_s and X_r, and the sender compresses X_s into Z_s to transmit, we define the pragmatic distortion as shown in equation 46, which measures the increase of Bayes risk after replacing the collaboration message X_s with Z_s:

$$\mathrm{D}_Y [X_s, Z_s] = \mathrm{R}[Y \mid X_r, Z_s] - \mathrm{R}[Y \mid X_r, X_s] \tag{46}$$

We give a specific formulation by replacing the Bayes risk in equation 46 with the Bayes risk of centerpoint loss in equation 33, as shown in equation 47 :

$$\mathrm{D}_{Y_{(i,j)}} [X_s, Z_s] = \mathrm{H}(Y_{c(i,j)}|X_r, Z_s) - \mathrm{H}(Y_{c(i,j)}|X_r, X_s) + \tag{47}$$

$$\frac{1}{2}\lambda_2(\mathrm{e}^{\mathrm{H}(Y_{o(i,j)}|X_r,Z_s)-1} - \mathrm{e}^{\mathrm{H}(Y_{o(i,j)}|X_r,X_s)-1}) + \tag{48}$$

$$\frac{1}{2}\lambda_3(\mathrm{e}^{\mathrm{H}(Y_{s(i,j)}|X_r,Z_s)-1} - \mathrm{e}^{\mathrm{H}(Y_{s(i,j)}|X_r,X_s)-1}) + \tag{49}$$

$$\frac{1}{2}\lambda_4(\mathrm{e}^{\mathrm{H}(Y_{r(i,j)}|X_r,Z_s)-1} - \mathrm{e}^{\mathrm{H}(Y_{r(i,j)}|X_r,X_s)-1}) \tag{50}$$

We consider a degraded version by ignoring the regression loss, which is suitable for semantic occupancy prediction task, as shown in equation 51:

$$\mathrm{D}_{Y_{(i,j)}}\left[X_s, Z_s\right] = \mathrm{H}(Y_{(i,j)}|X_r, Z_s) - \mathrm{H}(Y_{(i,j)}|X_r, X_s) \tag{51}$$

A.6.5 Proof of Theorem 1: optimal bit-rate of collaborative message

In this section, we derive the optimal transmission bit-rate in collaborative perception task. Consider the same collaboration situation described in Tab. 1 with 2 collaborators, and the observations/features of message sender and receiver are X_s and X_r, and the sender compresses X_s into Z_s to transmit. Our goal is to derive the minimum bit-rate needed to transmit Z_s while guaranteeing a limited pragmatic distortion, as shown in equation 52:

$$\mathrm{Rate}(\delta) = \min_{p(Z_s|X_s)} \mathrm{I}(X_s; Z_s) \quad \text{s.t. } \mathrm{D}_Y[X_s, Z_s] \leq \delta. \tag{52}$$

For occupancy prediction, put pragmatic distortion equation 51 into the constraint in equation 52, we have equation 55:

$$\begin{aligned}
\mathrm{D}_Y[X_s, Z_s] &= \mathrm{H}(Y|X_r, Z_s) - \mathrm{H}(Y|X_r, X_s) && (53)\\
&= [\mathrm{H}(Y|X_r) - \mathrm{I}(Y; Z_s|X_r)] - [\mathrm{H}(Y|X_r) - \mathrm{I}(Y; X_s|X_r)] && (54)\\
&= \mathrm{I}(Y; X_s|X_r) - \mathrm{I}(Y; Z_s|X_r) \leq \delta && (55)
\end{aligned}$$

This inequality condition also satisfies for object detection task, which corresponds to the distortion defined in equation 50 by considering two approximation:

- First-order approximation.

$$\mathrm{e}^{\mathrm{H}(Y_{o(i,j)}|X_r,Z_s)-1} - \mathrm{e}^{\mathrm{H}(Y_{o(i,j)}|X_r,X_s)-1} \geq \frac{1}{\mathrm{e}}(\mathrm{H}(Y_{o(i,j)}|X_r, Z_s) - \mathrm{H}(Y_{o(i,j)}|X_r, X_s)) \tag{56}$$

- Decomposition of joint entropy, with the assumption that the existing of object(Y_c) is independent with the location (Y_o), size (Y_s), and heading (Y_r).

$$\mathrm{H}(Y_c, Y_o, Y_s, Y_r) = \mathrm{H}(Y_c) + \mathrm{H}(Y_o) + \mathrm{H}(Y_s) + \mathrm{H}(Y_r) \tag{57}$$

Given that, We reformulate equation 52 as shown in equation 64:

$$\begin{aligned}
\mathrm{Rate}(\delta) =& \min_{p(Z_s|X_s)\ \text{s.t. } \mathrm{D}_Y[X_s,Z_s]\leq\delta} \mathrm{I}(X_s; Z_s) && (58)\\
\geq& \min_{p(Z_s|X_s)\ \text{s.t. } \mathrm{D}_Y[X_s,Z_s]\leq\delta} \mathrm{I}(X_s; Z_s|X_r) && (59)\\
\geq& \min_{p(Z_s|X_s)\ \text{s.t. } \mathrm{D}_Y[X_s,Z_s]\leq\delta} \mathrm{I}(Y; Z_s|X_r) && (60)\\
\geq& \min_{p(Z_s|X_s)\ \text{s.t. } \mathrm{D}_Y[X_s,Z_s]\leq\delta} \mathrm{I}(Y; X_s|X_r) - \delta && (61)\\
=& \mathrm{I}(Y; X_s \mid X_r) - \delta \quad (\text{no } Z_s) && (62)\\
=& \mathrm{H}(X_s) - \Big[\mathrm{H}(X_s) - \mathrm{I}(Y; X_s)\Big] - \Big[\mathrm{I}(Y; X_s) - \mathrm{I}(Y; X_s \mid X_r)\Big] - \delta && (63)\\
=& \mathrm{H}(X_s) - \underbrace{\mathrm{H}(X_s|Y)}_{\text{information in } X_s \text{ irrelevant to } Y} - \underbrace{\mathrm{I}(Y; X_s; X_r)}_{\text{information in } X_s \text{ redundant with } X_r \text{ about } Y} - \delta && (64)
\end{aligned}$$

Specifically, we set $\mathrm{Rate}(\delta) = 0$ when $\delta \geq \mathrm{I}(Y; X_s \mid X_r)$. We make assumption that the variables follow the Markov chain $Y \leftrightarrow X_s \leftrightarrow Z_s$ and $X_r \leftrightarrow X_s \leftrightarrow Z_s$. Next, we will explain the reasoning behind each inequality and the conditions for these inequality to achieve equality.

The first inequality equation 59 is satisfied when Markov chain $X_r \leftrightarrow X_s \leftrightarrow Z_s$ holds. This is because equation 65:

$$\mathrm{I}(Z_s; X_s, X_r) = \mathrm{I}(Z_s; X_r) + \mathrm{I}(Z_s; X_s|X_r) = \mathrm{I}(Z_s; X_s) + \mathrm{I}(Z_s; X_r|X_s) \tag{65}$$

The Markov chain $X_r \leftrightarrow X_s \leftrightarrow Z_s$ leads to $\mathrm{I}(Z_s; X_r|X_s) = 0$. Then we have equation 66:

$$\mathrm{I}(Z_s; X_s) = \mathrm{I}(Z_s; X_r) + \mathrm{I}(Z_s; X_s|X_r) \geq \mathrm{I}(Z_s; X_s|X_r) \tag{66}$$

Here we can see the equality condition for the first inequality equation 59 is that, $\mathrm{I}(Z_s; X_r) = 0$, which means Z_s, the compressed version of X_s, should not have redundant information in X_r.

The second inequality equation 60 is satisfied since due to DPI(Data Processing Inequality) given the Markov chain $Y \leftrightarrow X_s \leftrightarrow Z_s$. This is because equation 67:

$$\mathrm{I}(Z_s; X_s, Y) = \mathrm{I}(Z_s; Y) + \mathrm{I}(Z_s; X_s|Y) = \mathrm{I}(Z_s; X_s) + \mathrm{I}(Z_s; Y|X_s) \tag{67}$$

The Markov chain $Y \leftrightarrow X_s \leftrightarrow Z_s$ leads to $\mathrm{I}(Z_s; Y|X_s) = 0$. Then we have equation 68:

$$\mathrm{I}(Z_s; X_s) = \mathrm{I}(Z_s; Y) + \mathrm{I}(Z_s; X_s|Y) \geq \mathrm{I}(Z_s; Y) \tag{68}$$

We can see that the equality condition for the second inequality equation 60 is that, $\mathrm{I}(Z_s; X_s|Y) = 0$. We can derive that equation 69:

$$\mathrm{I}(Z_s; X_s|Y) = \mathrm{H}(Z_s|Y) - \mathrm{H}(Z_s|X_s, Y) = 0 \tag{69}$$

We can see from equation 69 that $\mathrm{H}(Z_s|Y) = \mathrm{H}(Z_s|X_s, Y)$, since Z_s is a compressed version of X_s, the uncertainty $\mathrm{H}(Z_s|X_s, Y)$ is 0, therefore $\mathrm{H}(Z_s|Y) = 0$. This implies that Z_s is completely task-relative, it does not contains information unrelated to the task Y.

The third inequality is derived from equation 55, and the equality condition is achieved when the distortion budget is sufficiently utilized.

A.6.6 Discussion on Redundancy vs Synergy

We would like to argue that $I(Y; X_s; X_r)$ in (10) is non-negativity in the collaborative perception scenario and discuss the reason. According to the definition, $I(Y; X_s; X_r) = I(X_r; X_s) - I(X_r; X_s|Y)$, in our case, Y presents the ground truth signal, and X_r, X_s present noisy observations of signal Y from different views ($X_r = f_r(Y)$ and $X_s = f_s(Y)$), where X_r and X_s contain shared information in Y. $I(X_r; X_s) - I(X_r; X_s|Y)$ is positive, since knowing signal Y will reduce the mutual information between X_r, X_s. In collaborative perception scenarios, we can formulate $X_r = f_r(Y)$ and $X_s = f_s(Y)$. An simlified example is $X_r = Y, X_s = 2Y$, then we have $I(X_r; X_s|Y) = 0$, and $I(Y; X_s; X_r)$ is positive.

One classical condition for $I(Y; X_s; X_r)$ to be negative is when Y emerges from the interaction between $X_s; X_r$; for example $Y = X_s \oplus X_r$ (XOR). In such case, knowing Y increases the correlation between X_s and X_r. However, in our scenario, the observations X_r, X_s depend on the ground truth Y but not vice versa, thus differs from this class of synergy condition.