COMMUNICATION-EFFICIENT MULTI-TASK COLLABORATIVE PERCEPTION USING JOINT REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Collaborative perception (CP) improves the environmental understanding of connected agents via the exchange of visual scene representations. While prior works focus on single-task perception, autonomous agents often need to run multiple tasks concurrently, for comprehensive scene understanding. A direct application of existing CP frameworks requires each agent to send a separate feature representation for every task, causing communication to scale poorly with both the number of tasks and agents. We study multi-task CP from a joint representation-sharing perspective. Intuitively, task-specific features produced by an agent are distinct encodings of the same local observation and hence, contain overlapping scene information. We formalize this using rate-distortion theory, comparing separate task-wise coding with joint coding of multiple task-specific feature representations. The analysis shows that, for the same task-level distortion requirements, jointly coding multiple task representations can reduce communication by leveraging the information shared among them. Motivated by this insight, we develop ORAT (**O**ne **R**epresentation for **A**ll **T**asks), a multi-task CP framework that maps multiple task-specific representations into a single *base* representation before transmission. Unlike prior task-wise communication methods that apply compression, task-relevant region selection, or receiver-driven refinement on each task-specific representation separately, ORAT applies these operations directly on that base. The receiver agent adapts the shared base representation into task-compatible features through learned task-specific projection modules, and fuses them with its local features for downstream prediction. Experiments on OPV2V and V2X-Sim datasets show that ORAT preserves perception performance for all tasks while reducing communication by factors ranging from $3.61\times$ to $4810\times$ compared to prior task-specific transmission frameworks.

1 INTRODUCTION

Collaborative perception (CP) allows multiple agents to share intermediate perception features in connected autonomous systems, improving overall scene understanding (Xu et al., 2022c; Yu et al., 2022; Xu et al., 2022a;b). Such information exchange from complementary viewpoints enables agents to uncover occluded objects, extend their effective sensor range, and improve overall performance in safely-critical tasks like 3D object detection, bird's eye view (BEV) segmentation, etc. However, a fundamental challenge in CP is the trade-off between perception performance and communication volume. The perception features to be transmitted are often high-dimensional, whereas the wireless channel available for inter-agent communication is bandwidth-limited and latency-sensitive. This poses a major challenge in the deployment of CP mechanisms in real-world systems.

To address this, recent works propose communication-efficient CP frameworks, such as transmitting only features from high-confidence regions (Hu et al., 2022), filling information gaps between agents while suppressing redundant information sharing (Hu et al., 2024), or leveraging temporal continuity to transmit only uncertain features (Mukhopadhyay et al., 2026). These approaches demonstrate good communication-performance trade-offs in single-task settings, where all agents collaborate toward one downstream objective. However, in real-life autonomous systems, an agent may need to detect dynamic objects, segment drivable lanes, realize static obstacles and track surrounding agents at the same time. Although the importance of different tasks has been studied in isolation (Li et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2021; Xu et al., 2022c;a; Hu et al., 2023; Ruan et al., 2024; Zhong et al., 2025), previous works do not consider scenarios where multiple tasks must be executed simultaneously to achieve a more comprehensive understanding of the environment. In such settings, communication efficiency must be evaluated across the entire set of perception tasks.

A straightforward application of existing frameworks to the multi-task setting is to run a separate CP pipeline for each task, where each agent transmits a distinct *feature representation* for each task. Although simple, this approach scales poorly with team size and degrades ungracefully when resources are tight. More importantly, the task-specific representations are not independent sources of information. They are different encodings of the same physical scene. Transmitting them independently communicates redundant scene information, such as geometry, occupancy, and object layout. At the same time, however, these representations may contain task-specialized cues for different objectives. For instance, they may share common scene structure such as object locations, geometry, and the background while having task-specialized cues, such as object-level features for detection or dense layout of the driving lanes for segmentation. Hence, we ask:

1. Can jointly encoding multiple task representations reduce communication compared with separate task-wise transmission under the same task-level fidelity requirements?
2. If such savings are possible, how can we design a practical CP framework that communicates one shared representation while preserving performance across multiple downstream tasks?

We address the first question through a rate-distortion analysis of multi-task representation sharing. Our analysis shows that, under the same per-task distortion constraints, joint coding can reduce communication compared to separate task-wise coding when the sender's task-specific representations are not independent. The potential saving grows with the amount of shared inter-task information, which we formalize using conditional total correlation (Watanabe, 1960). Motivated by this insight, we address the second question, by developing a novel communication-efficient CP framework ORAT (**O**ne **R**epresentation for **A**ll **T**asks), which learns a shared representation that fulfills the requirements of multiple tasks. Specifically, each agent first maps its task-specific representations into a shared *base* using a learned representation aggregator that captures common scene information across tasks, while preserving a similar level of performance as achieved from transmission of task-specific individual representations. Unlike task-wise communication methods that optimize transmission separately for each task-specific representation (Hu et al., 2022; 2024; Liu et al., 2026), ORAT utilizes compression techniques such as quantization, task-relevant filtering, and receiver-conditioned refinement directly on this shared base. The receiver agent maps this communicated representation to task-compatible features using lightweight, learned task-specific projection modules. Thus, rather than inducing separate communication streams for different tasks, ORAT enables one compact stream with a unified feature representation to support all tasks at the receiver. We evaluate ORAT on widely used CP datasets, OPV2V (Xu et al., 2022c) and V2X-Sim (Li et al., 2022). Experiments show that ORAT preserves multi-task perception performance while substantially reducing communication volume compared to task-specific individual transmissions. Briefly, our contributions are listed as follows:

- We perform a rate-distortion analysis for multi-task CP which shows that, creating a single representation encoding for the shared information across tasks, can satisfy multiple tasks while also drastically reducing the communication compared to the above naive approach.
- We introduce ORAT, a practical multi-task CP framework that first maps multiple task-specific features into a single base representation. Importantly, all communication-efficient operations, including quantization, task-relevant filtering, and receiver-conditioned refinement are performed on this base rather than on separate task-specific streams. The receiver then adapts the refined shared base into task-compatible features through lightweight task-specific projection modules, which are used for task-specific downstream inference.
- Finally, extensive experiments on multiple datasets show that collaborating through one compact base representation retains performance for multiple perception tasks while reducing communication volume ranging from $3.61\times$ to $4810\times$ compared to the benchmarks.

2 Related Works

CP methods are commonly categorized by the stage at which information is exchanged - early fusion shares raw sensor data, Chen et al. (2019); Zhang et al. (2021), intermediate fusion shares features

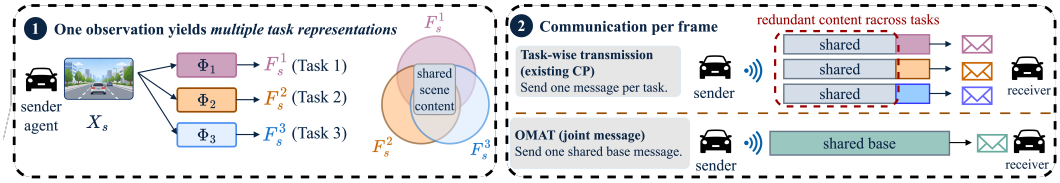


Figure 1: A sender-receiver pair with $M = 3$ tasks. The sender uses task-specific BEV encoders Φ_m to generate representations F_s^m for $m \in [3]$. Existing frameworks transmit one feature representation per task, thereby repeatedly transmitting shared content. In contrast, ORAT generates a single shared base representation, allowing the sender to transmit common scene information once.

extracted from BEV encoders (Wang et al., 2020; Xu et al., 2022a; Hu et al., 2022) and late fusion shares final perception statistics like a detection map (Rawashdeh & Wang, 2018; Lu et al., 2023). Early fusion can ensure high perception accuracy, but is communication-intensive, while late fusion requires less bandwidth, but may degrade accuracy. Thus, recent CP methods transmit intermediate feature representations to achieve an optimal trade-off (Xu et al., 2022a;b; Hu et al., 2023).

Communication-efficient CP. When2com (Liu et al., 2020) learns communication grouping among agents. How2Comm (Yang et al., 2023) adopts a mutual information-aware communication mechanism combined with spatial-channel filtering. Where2comm (Hu et al., 2022) transmits features from spatially confident regions, and CodeFilling (Hu et al., 2024) further improves efficiency through codebook-based message representation and collective information-filling selection. CoSDH (Xu et al., 2025) proposes supply-demand-aware region selection and intermediate-late hybrid fusion to achieve a superior accuracy-bandwidth trade-off. CooperTrim (Mukhopadhyay et al., 2026) adaptively selects and transmits uncertainty-aware informative features across consecutive frames. RDComm (Liu et al., 2026) provides a rate-distortion analysis of collaborative communication; introduces task-entropy coding and mutual-information driven message selection. However, these methods focus on optimizing the communication for a single downstream task. When applied to simultaneous multi-task perception, each task requires its own selected, quantized, or entropy-coded stream to be transmitted separately, causing communication cost to grow with the number of tasks.

Heterogeneous CP. HEAL (Lu et al., 2024) aligns continually emerging heterogeneous agents with different sensors and models into a unified feature space for cross-agent fusion. STAMP (Gao et al., 2025) studies scalable CP across heterogeneous agents with different modalities, model architectures, and downstream tasks. It maps the local feature of each single-task agent into a shared protocol space before communication. However, this protocol representation is optimized for a selected objective, and their ablations show that performance can drop substantially when the protocol objective is mismatched with the receiver's downstream task. In contrast to these works, ORAT learns a task-unified base representation that jointly encodes multiple task-specific features within the same agent, allowing one transmitted representation to support all the tasks at the receiver.

3 Problem Formulation

Multi-task collaborative perception. We consider a CP system with N connected agents. Each agent $i \in [N]$ has a local observation X_i, obtained from its onboard sensors such as LiDAR or cameras (Xu et al., 2022c; 2023; Yu et al., 2022; Wu et al., 2026). Unlike standard single-task CP, we consider a multi-task setting, where every agent simultaneously runs M perception tasks. Each agent i runs task-specific BEV encoder Φ_m, to produce a feature representation $F_i^m \triangleq \Phi_m(X_i)$ for $m \in [M]$. Each agent also has task-specific prediction modules Ψ_m, which uses all the received features from the neighboring agents and its own local feature to produce the final task prediction (e.g., 3D bounding boxes for object detection or BEV segmentation maps for segmentation). The ground-truth label is denoted by Y^m for $m \in [M]$. A straightforward extension of existing single-task CP methods is to perform collaboration independently for each task. Each sender s transmits a separate representation for every task. For a receiver r, the prediction for task m can be written as,

$$\hat{Y}_r^m \triangleq \Psi_m\Big(F_r^m, \{Z_{s\to r}^m\}_{s\in[N]\setminus\{r\}}\Big), \tag{1}$$

where $Z_{s\to r}^m$ denotes the task-m feature received from sender s, possibly after compression, selection, and coordinate transformation. Hence, the total communication cost is,

$$R_{\text{task}} \triangleq \sum_{r=1}^{N} \sum_{s\in[N]\setminus\{r\}} \sum_{m=1}^{M} b(Z^m_{s\to r}), \tag{2}$$

where $b(\cdot)$ defines the communication cost. As in Eq. 2, communication grows with the number of agents and tasks. Moreover, since all task-specific features are derived from the same observation X_s, transmitting them independently communicates overlapping information about the same scene.

Shared representation. To avoid task-wise repetition, we explore whether each agent can transmit a single representation that supports all tasks. Let $\mathbf{F}_s \triangleq \left(F^1_s, \ldots, F^M_s\right)$ be the set of task-specific representations produced by agent s. Instead of M separate transmissions $\{Z^m_{s\to r}\}_{m=1}^M$, agent s encodes $\mathbf{F}_s$ into a single representation $Z_{s\to r}$. The receiver r then performs task-wise prediction,

$$\hat{Y}^m_r \triangleq \Psi_m\Big(F^m_r, \{Z_{s\to r}\}_{s\in[N]\setminus\{r\}}\Big). \tag{3}$$

The idea is illustrated in Fig. 1. The goal of multi-task communication is therefore to minimize the communication cost of the shared messages while preserving the performance of all tasks:

$$\min \quad \sum_{r=1}^{N} \sum_{s\in[N]\setminus\{r\}} b(Z_{s\to r}) \quad \text{s.t.} \mathcal{L}_m\Big(Y^m, \hat{Y}^m_r\Big) \le L_m,\ \forall m \in [M], r \in [N] \tag{4}$$

where $\mathcal{L}_m$ denotes the loss for task m, and L_m is the maximum tolerable loss for that task.

Information-theoretic intuition. Now, we study whether a shared representation provides provable communication savings compared to exchanging separate task-specific transmissions. The scene generates the task labels and observations $(\mathbf{Y}, X_s, X_r)$, where $\mathbf{Y} \triangleq \left(Y^1, \ldots, Y^M\right)$. The sender observes only X_s and transmits $Z_{s\to r}$ produced from task-specific representations $\mathbf{F}_s$. The receiver observes X_r and uses $Z_{s\to r}$ for downstream prediction. We view representation-level distortion as the information about $\mathbf{F}_s$ that is not preserved by $Z_{s\to r}$, upon accounting for the receiver's view X_r. For tractability of the theoretical analysis, let $\mathbf{F}_s, X_r$ take values in finite alphabets (e.g., through quantization). Given $Z_{s\to r}$ and X_r, let $\hat{p}(F^m_s|Z_{s\to r}, X_r)$ denote the prediction by the receiver for task m, and consider the per-task distortion under logarithmic loss $-\log \hat{p}(F^m_s|Z_{s\to r}, X_r)$. We define the distortion for task m as the minimum expected log loss over the receiver predictions,

$$D_m \triangleq H(F^m_s|Z_{s\to r}, X_r) \tag{5}$$

where the minimum is achieved when the prediction is equal to the true distribution. $H(X \mid Y)$ denotes the conditional entropy of X given Y (Shannon, 1948). Our goal is to find the minimum source representation under the fidelity criterion $D_m \le \delta_m$ for a per-task distortion budget δ_m, when the encoder (sender) observes $\mathbf{F}_s$ and the decoder (receiver) has side information X_r. Classical Wyner-Ziv lossy source coding formalizes the minimum achievable rate with decoder-side information (Wyner & Ziv, 1976). We adopt its characterization under log loss (Courtade & Weissman, 2014).

4 Communication Savings from Joint Multi-Task Encoding

This section analyzes when communicating one shared representation can be more efficient than transmitting separate task-specific representations.

Theorem 1 (Joint multi-representation rate)**.** *Suppose the sender encodes the representation for m-tasks $\mathbf{F}_s \triangleq \left(F^1_s, \ldots, F^M_s\right)$ while X_r is the observation available at the decoder. Define the per-task distortion $D_m = H(F^m_s|Z_{s\to r}, X_r)$ with a fidelity criterion $D_m \le \delta_m$ for task $m \in [M]$. Then, Wyner-Ziv rate-distortion function under logarithmic loss is given by,*

$$R_{\textit{joint}}(\delta_1, \ldots, \delta_M) \triangleq \min_{p[U|\mathbf{F}_s]:\, H(F^m_s|U,X_r)\le\delta_m \forall m\in[M]} I(\mathbf{F}_s; U|X_r) \tag{6}$$

where U is an auxiliary random variable generated via $p[U|\mathbf{F}_s]$ with a Markov condition $U - \mathbf{F}_s - X_r$. $I(\mathbf{F}_s; U \mid X_r)$ denotes the conditional mutual information (Cover & Thomas, 2006) between the sender's task representations and the encoded representation U given the receiver observation.

Remark 1. Consider the naive task-wise CP baseline where the encoder sends each task's message $Z^m_{s\to r}$ based on F^m_s (as opposed to $Z_{s\to r}$) with per-task distortion budgets δ_m for $m \in [M]$. Then, applying the rate-distortion lower bound in Thm. 1 to each task separately yields the per-task rate,

$$R^m_{\text{sep}}(\delta_m) \triangleq \Big(\min_{p[U|F^m_s]:\, H(F^m_s|U,X_r)\le\delta_m} I(F^m_s; U|X_r)\Big) = \left(H(F^m_s|X_r) - \delta_m\right)^+ \tag{7}$$

under the Markov condition $U - F_s^m - X_r$. $(H(F_s^m|X_r) - \delta_m)^+ \triangleq \max(H(F_s^m|X_r) - \delta_m, 0)$ denotes that the required rate is 0 whenever $\delta_m \geq H(F_s^m|X_r)$. Therefore, when each task is encoded separately, the total rate for all m tasks is,

$$R_{\text{sep}}(\delta_1, \ldots, \delta_M) \triangleq \sum_{m=1}^{M} R_{\text{sep}}^m(\delta_m) = \sum_{m=1}^{M} (H(F_s^m|X_r) - \delta_m)^+ \tag{8}$$

This quantity represents a lower bound on the rate required by separate task-wise communication.

Proof. For an admissible U that meets the distortion constraint $H(F_s^m|U, X_r) \leq \delta_m$,

$$I(F_s^m; U|X_r) = H(F_s^m|X_r) - H(F_s^m|U, X_r) \geq H(F_s^m|X_r) - \delta_m \tag{9}$$

where the lower bound can be achieved by a suitable auxiliary variable (e.g., time-sharing). □

Proposition 1 (Relation between joint and per-task rate-distortion functions)**.** *From the definitions of joint and per-task rate-distortion functions in Eqs. 6 and 7, we have,*

$$\max_{m \in [M]} R_{sep}^m(\delta_m) \leq R_{joint}(\delta_1, \ldots, \delta_M) \leq R_{sep}(\delta_1, \ldots, \delta_M) \tag{10}$$

Proof. The lower bound follows simply by contradiction, whereas the upper bound follows by observing that combining the m single-task coding processes, each achieving $R_{\text{sep}}^m(\delta_m)$, already provides an upper bound on the joint rate for satisfying the distortion constraints $(\delta_1, \ldots, \delta_M)$. □

Next, we consider distortion budgets $0 \leq \delta_m \leq H(F_s^m|X_r)$ for task $m \in [M]$. The regime $\delta_m > H(F_s^m|X_r)$ leads to zero communication and is outside the interesting communication-saving regime.

Theorem 2 (Communication savings)**.** *Consider a distortion budget $0 \leq \delta_m \leq H(F_s^m|X_r)$ for task $m \in [M]$, along with the separate encoding of F_s^m for each task yielding a total rate $R_{sep}(\delta_1, \ldots, \delta_M)$ from Eq. 8, and the joint encoding of $\mathbf{F}_s = (F_s^1, \ldots, F_s^M)$ yielding the joint rate $R_{joint}(\delta_1, \ldots, \delta_M)$ from Eq. 6. Define the joint auxiliary representation U^* as the minimizer of Eq. 6. If all distortion constraints are active at the optimum, i.e., $H(F_s^m|U^*, X_r) = \delta_m$ for all $m \in [M]$, then, the communication savings from using a joint representation is given by,*

$$0 \leq R_{sep}(\delta_1, \ldots, \delta_M) - R_{joint}(\delta_1, \ldots, \delta_M) = \text{TC}\big(F_s^1, \ldots, F_s^M \,\big|\, X_r\big) - \text{TC}\big(F_s^1, \ldots, F_s^M \,\big|\, U^*, X_r\big) \tag{11}$$

where TC denotes the conditional total correlation (TC) (Watanabe, 1960), which is non-negative.

The proof is provided in App. A.1. The first term in the right hand side of Eq. 11 represents the initial inter-task redundancy conditioned on the receiver's side information X_r, while the second term quantifies the residual inter-task redundancy remaining after the decoder additionally has access to the auxiliary description U^*. Their difference measures the inter-task redundancy resolved by the joint representation. The largest saving is achieved when the joint representation captures all inter-task correlations conditioned on the receiver observation, i.e., $\text{TC}\big(F_s^1, \ldots, F_s^M \,\big|\, U^*, X_r\big) = 0$.

5 Designing the ORAT Framework

The rate-distortion analysis in Section 4 shows that joint encoding reduces communication when task-specific representations contain shared scene information not already available to the receiver. We develop a novel framework, ORAT to realize this principle. The framework is described for M tasks and N connected agents. All agents run the same set of tasks, without loss of generality.

5.1 Sender-Side Shared Representation Encoding

Shared-base generation. We consider M encoder models Φ_m (one for each task) for $m \in [M]$. Given the local observation X_i of agent i, the task encoder Φ_m produces an intermediate BEV representation $F_i^m \triangleq \Phi_m(X_i), \forall m \in [M]$. To communicate efficiently, we introduce a learned aggregator module $\mathcal{G}$, which maps the M task representations of agent i onto a single *base* feature,

$$F_i^B \triangleq \mathcal{G}(F_i^1, \ldots, F_i^M) \in \mathbb{R}^{C \times H \times W} \tag{12}$$

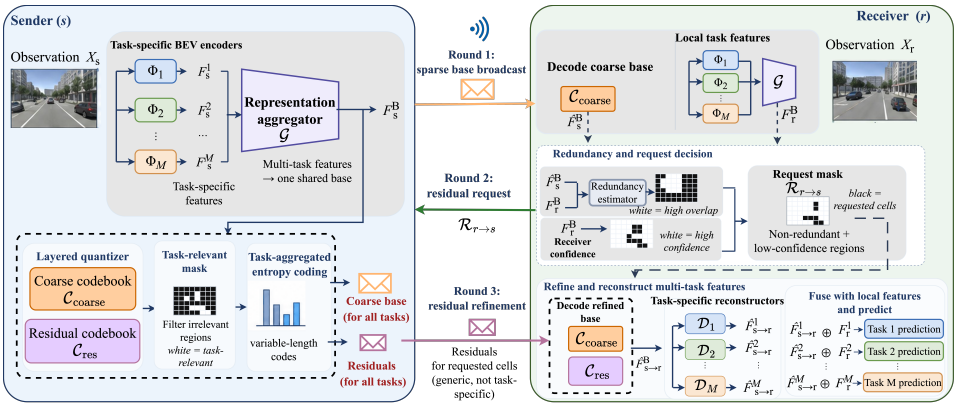


Figure 2: **ORAT**. The sender maps all M tasks' representations into a single base feature F_s^B, masks irrelevant regions, applies task-aggregated entropy coding, and broadcasts coarse base. The receiver estimates redundancy to identify low-overlap regions. The sender transmits residuals for the requested cells. Finally, the receiver projects the refined base $\hat{F}_{s\to r}^B$ to task-specific features $\hat{F}_{s\to r}^m$ using reconstructors $\{\mathcal{D}_m\}_{m\in[M]}$ and fuses them with its local features for downstream prediction.

The task-unified base representation F_i^B encodes shared scene content once while allowing the receiver to generate task-compatible features for effective downstream task prediction.

Layered quantization. We implement a layered vector quantizer $\mathcal{Q}$ (Zhu et al., 2022; Liu et al., 2026), to quantize the base feature map $F_i^B \in \mathbb{R}^{C\times H\times W}$ for agent $i \in [N]$. We use a learned projection network, $f_{\text{in}}(.)$ to generate $Z_i \triangleq f_{\text{in}}(F_i^B)$, which maps the base feature to the quantization space. We define $\mathcal{C}_{\text{coarse}}$ as the always-transmitted coarse code, and $\mathcal{C}_{\text{res}}$ as a *task-blind fidelity residual* refining the base. The codebooks are learned during training, each containing a set of learnable embeddings (codewords). For each $h \in [H]$, $w \in [W]$, we define

$$k_i^C[h,w] \triangleq \arg\min_k \|Z_i[:,h,w] - \mathcal{C}_{\text{coarse}}[k]\|_2^2 \tag{13}$$

Hence, the coarse quantized feature can be written as, $Z_{i,\text{coarse}}^q[:,h,w] \triangleq \mathcal{C}_{\text{coarse}}\left[k_i^C[h,w]\right]$. Next, we define residuals as, $Z_{i,\text{res}} = Z_i - Z_{i,\text{coarse}}^q$ and the selected codeword indices as,

$$k_i^R[h,w] \triangleq \arg\min_k \|Z_{i,res}[:,h,w] - \mathcal{C}_{\text{res}}[k]\|_2^2 \tag{14}$$

Hence, the residual quantization is performed as, $Z_{i,\text{res}}^q[:,h,w] = \mathcal{C}_{\text{res}}\left[k_i^R[h,w]\right]$.

Task-relevant masking. We leverage a spatial confidence map to prune out the regions before transmission where an agent has low confidence or does not hold critical information (e.g., background). For each task m, the frozen prediction head produces a confidence map $C_{i,m} \in [0,1]^{H\times W}$ for $i \in [N]$. These task-specific confidence maps then produce a single aggregate confidence score,

$$\bar{C}_i[h,w] \triangleq \max_m C_{i,m}[h,w] \tag{15}$$

Thus, a cell $[h,w]$ produces a high score if it holds critical information for at least one downstream task. From the quantized base as described earlier, only the cells with aggregate confidence above a certain threshold ϵ_c, i.e., the regions $\mathcal{V}_i \triangleq \{(h,w) : \bar{C}_i[h,w] \geq \epsilon_c\}$ are transmitted.

Entropy coding. The codeword indices are further compressed using entropy coding with Huffman tables (Huffman, 1952), (Liu et al., 2026). The entropy table should prioritize codewords that are frequently selected in task-relevant regions. We define $u(h,w)$ as the selected codeword index at location (h,w). For each task m, we compute a confidence-weighted codeword distribution,

$$p_c^{(m)}(e_k) = \frac{\sum_{(h,w)\in\mathcal{T}} \mathbf{1}[u(h,w)=k]\; C_m[h,w]}{\sum_{(h,w)\in\mathcal{T}} C_m[h,w]}, \tag{16}$$

where e_k denotes the k-th codeword from the codebook. $\mathcal{T}$ includes cells from all training frames. The codebooks use the *task-marginal* frequency,

$$\bar{p}_c(e_k) \triangleq \sum_{m\in[M]} w_m\, p_c^{(m)}(e_k), \tag{17}$$

where w_m is the weight assigned to task $m \in [M]$. The table is created from $\bar{p}_c$. Agent i then broadcasts the (Huffman-coded) coarse codebook indices $\mathcal{Q}_i^C \triangleq \text{Huff}\left(\left\{k_i^C[h,w]\right\}_{(h,w)\in\mathcal{V}_i}\right)$.

5.2 Receiver-Driven Refinement and Task-Specific Reconstruction

Redundancy estimation and residual request. After receiving the coarse base, the receiver r estimates overlap between sender's coarse base and its own base representation F_r^B. It first decodes a coarse proxy, $\hat{F}_s^B \triangleq \mathcal{C}_{\text{coarse}}[\mathcal{Q}_s^C]$ of the base representation from sender s. Next, for each spatial cell (h, w), it computes a redundancy score using a learnable estimator $\mathcal{M} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$,

$$\rho_{s,r}[h, w] \triangleq \sigma\left(\mathcal{M}\left(\hat{F}_s^B[:, h, w], F_r^B[:, h, w]\right)\right) \in [0, 1], \tag{18}$$

where $\sigma(\cdot)$ is the sigmoid function. A larger $\rho_{s\to r}[h, w]$ implies stronger overlap between the sender and receiver at that spatial cell. Let $\mathcal{P}_{s,r}$ denote the distribution of co-located sender-receiver feature pairs (at same spatial location), and $\mathcal{P}_s \otimes \mathcal{P}_r$ denote the product of their marginals. We train $\mathcal{M}$ to minimize the following binary classification objective (Li et al., 2020; Liu et al., 2026):

$$\mathcal{L}_{\text{MI}} \triangleq -\mathbb{E}_{(a,b)\sim\mathcal{P}_{s,r}}\left[\log \sigma(\mathcal{M}(a, b))\right] - \mathbb{E}_{(a,b)\sim\mathcal{P}_s\otimes\mathcal{P}_r}\left[\log\left(1 - \sigma(\mathcal{M}(a, b))\right)\right]. \tag{19}$$

The receiver therefore requests for fine-grained information at regions $\mathcal{R}_{r\leftarrow s}$, where content would be unique (non-redundant) and its own confidence is below a certain threshold ϵ_c:

$$\mathcal{R}_{r\leftarrow s} \triangleq \left\{(h, w) \;:\; \rho_{s,r}[h, w] < \tau_{\text{red}} \;\wedge\; \bar{C}_r[h, w] < \epsilon_c\right\}, \tag{20}$$

where τ_{red} is the redundancy threshold and $\bar{C}_r$ is the confidence map of receiver r, as in Eq. 15. The sender transmits residual indices $\mathcal{Q}_{s\to r}^R \triangleq \text{Huff}\left(\left\{k_s^R[h, w]\right\}_{(h,w)\in\mathcal{R}_{r\leftarrow s}}\right)$. Thus, the coarse base communicates scene content broadly, while the residual refines only receiver-specific regions.

Task-specific mapping and prediction. After receiving $\mathcal{Q}_{s\to r}^R$, the receiver r recovers the sender's refined base $\hat{F}_{s\to r}^B$ by mapping the indices to the associated codewords from the codebooks. Then, task-specific reconstructors $\mathcal{D}_m$ map the sender's base into the task-specific features, $\hat{F}_{s\to r}^m \triangleq \mathcal{D}_m(\hat{F}_{s\to r}^B), \forall m \in [M]$. Finally, the receiver fuses these projected features with its own local task feature and performs prediction for each task: $\hat{Y}_r^m = \Psi_m\left(F_r^m, \{\hat{F}_{s\to r}^m\}_{s\in[N]\setminus\{r\}}\right)$.

5.3 Training and Inference

Our framework consists of multiple stages of training protocols as described below:

Stage 1 (task backbones): First, we pretrain BEV encoders, $\Phi_m(.)$, the fusion-prediction module $\Psi_m(.)$ for tasks $m \in [M]$. These backbone modules would be kept frozen in later stages of training.
Stage 2 (codec): In this stage, we train the representation aggregator $\mathcal{G}$, the vector quantizer $\mathcal{Q}$, and per-task reconstructors $\mathcal{D}_m$. The goal is to minimize the following loss:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{M}\alpha_m\,\mathcal{L}_{\text{task}}^{(m)}\left(\hat{Y}_i^m, Y^m\right) + \beta\frac{1}{N}\sum_{i=1}^{N}\left\|\hat{F}_i^B - F_i^B\right\|_2^2 + \gamma\sum_{i=1}^{N}\sum_{m=1}^{M}\left\|F_i^m - \mathcal{D}_m(\hat{F}_i^B)\right\|_2^2 + q\mathcal{L}_{\text{Q}}$$

The first term in the loss function above is the task-specific prediction loss. The second term is the base reconstruction loss – it forces the quantized-then-dequantized base $\hat{F}_i^B$ to stay close to the original pre-quantization base F_i^B. The third term encourages the shared base to preserve sufficient information to recover task-specific representations. $\mathcal{L}_{\text{Q}}$ is the standard vector-quantization loss (van den Oord et al., 2017). The hyperparameters α_m, β, γ, q are used to balance these objectives.
Stage 3 (Redundancy estimation): We next train the redundancy estimator $\mathcal{M}$ while freezing the above modules. Finally, Huffman tables are created using the codeword distribution in Eq. 17.

After training, the learned modules are fixed and deployed at each agent for inference. Each agent uses them to construct and transmit the compact base representation, perform receiver-conditioned residual refinement, and transform the received base from other agents into task-compatible features for downstream prediction. Our framework is illustrated in Fig. 2.

6 Experiments

In our experiments, we seek to answer the following questions:

- Can ORAT support multiple downstream tasks using a single shared base representation?

- How well does ORAT generalize across datasets, tasks, and different sensing modalities?
- How does ORAT compare with the benchmarks in perception-communication trade-off?
- How do the compression components, including base quantization, entropy coding, and residual refinement, affect communication efficiency and task performance?

Datasets. We conduct experiments on two representative CP datasets: V2X-Sim (Li et al., 2022) and OPV2V (Xu et al., 2022c). OPV2V is a simulated vehicle-to-vehicle CP dataset collected with OpenCDA (Xu et al., 2021) and CARLA (Dosovitskiy et al., 2017), containing $11,464$ synchronized frames across diverse driving scenes. Each vehicle is equipped with a 64-channel LiDAR and 4 RGB cameras. V2X-Sim is a synthetic vehicle-to-infrastructure CP dataset built with CARLA-SUMO co-simulation, containing $10,000$ frames with multiple vehicles and a roadside unit. Each vehicle contains 6 RGB cameras and 32-channel LiDAR. We consider collaboration among up to 5 agents.

Backbone models and perception tasks. For OPV2V dataset, we utilize the open-source pretrained backbone models from Xu et al. (2022a) and evaluate three tasks: i) object detection, ii) dynamic segmentation, and iii) static segmentation. For the V2X-Sim dataset, we use open-source backbone models from Li et al. (2021) to evaluate two tasks: object detection and segmentation. The set of tasks is chosen according to the available ground-truth annotations in each dataset.

Baselines. For collaborative baselines, we consider representative single-task CP frameworks, including DiscoNet (Li et al., 2021), CoBEVT (Xu et al., 2022a), Where2Comm (Hu et al., 2022), CodeFilling (Hu et al., 2024), and RDComm (Liu et al., 2026). Since these methods are originally designed for single-task communication, we apply them for each task separately in our multi-task setting and sum their communication costs across tasks. We also consider STAMP (Gao et al., 2025), which enables collaboration across agents with heterogeneous tasks by adapting each agent's task feature to a shared representation before communication. The shared space is trained with a selected *task objective*, referred to as the *protocol* objective in STAMP. Thus, STAMP communicates features adapted for a chosen task objective, whereas ORAT aggregates multiple task-specific representations into a single task-unified base that supports all downstream tasks at the receiver.

Table 1: Task performance (V2X-Sim).

Framework	AP@50 %	IoU %
No Collaboration	50	37
DiscoNet	70	41
Where2Comm	66	42
CodeFilling	64	41
RDComm	64	44
ORAT	**69**	**42**

Evaluation metrics. For segmentation, we report intersection-over-union (IoU) between predicted and ground-truth segmentation maps. For object detection, we report average precision (AP) at IoU threshold 50% (AP@50). Following prior works (Hu et al., 2022; 2024; Liu et al., 2026), we report the communication in Mbits on a $\log_2$ scale.

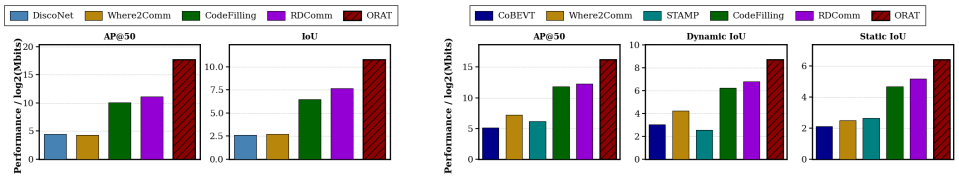
We now evaluate ORAT in terms of multi-task perception performance, communication efficiency, and a series of ablation studies on its key design components.

Perception-communication trade-off. Tables 1 and 2 compare ORAT against state-of-the-art CP methods. As we observe, ORAT achieves comparable perception performance with transmission of a single shared representation. Fig. 3 evaluates the perception-communication trade-off using the ratio between task performance (e.g., AP@50, IoU metrics) and communication volume (Mbits on $\log_2$ scale), where higher values imply better efficiency. As we observe, ORAT achieves the best trade-off, reducing communication by $3.61\times$ over RDComm, $5.43\times$ over CodeFilling, $70.73\times$ over STAMP, $2842\times$ over Where2Comm, $3371\times$ over DiscoNet, and $4810\times$ over CoBEVT. We further note that experiments with OPV2V dataset provide a more challenging setting since the task representations are derived from different sensor modalities: LiDAR features for object detection and camera features for segmentation. The competitive performance on OPV2V for all tasks in Table 2 implies that ORAT can effectively encode task representations from multiple modalities into a single base while preserving performance across downstream tasks.

Table 2: Task performance (OPV2V).

Framework	AP@50 % (LiDAR)	Dyn. IoU % (Camera)	Stat. IoU % (Camera)
No Collaboration	76	26	21
CoBEVT	92	54	38
Where2Comm	92	54	32
CodeFilling	91	48	36
RDComm	90	50	38
STAMP	75	30	38
ORAT	**91**	**49**	**36**

Qualitative example. In Fig. 4, we present a sample frame from the OPV2V evaluation set. As we observe, collaboration through a single base representation from all the neighboring agents supports multiple downstream tasks: detection, dynamic segmentation, and static segmentation.



(a) V2X-Sim (b) OPV2V

Figure 3: Perception-communication trade-off for different CP frameworks.

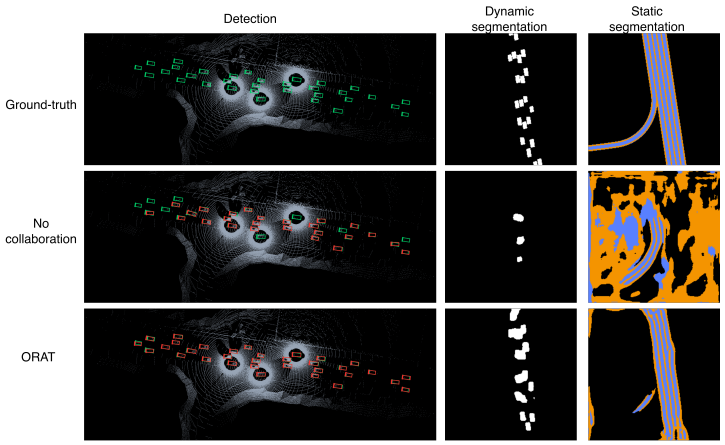
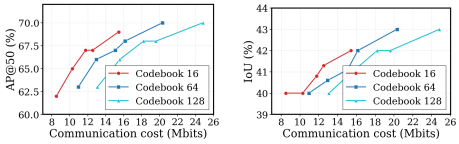


Figure 4: Perception performance of ORAT across three different tasks (for detection, the green and red boxes represent the ground-truth annotations and the predictions, respectively).

Codebook size. We study the impact of codebook size on the performance in Fig. 5. Across different codebook sizes, ORAT consistently preserves perception accuracy while incurring very low communication overhead. Larger codebooks provide finer-grained quantization and therefore improve perception performance, whereas smaller codebooks further reduce communication. Thus, collaboration through the quantized base preserves performance across all tasks.



(a) Detection AP (b) Segmentation IoU

Figure 5: Varying codebook size (V2X-Sim).

Residual sharing and entropy coding. Fig. 6 first studies the impact of the redundancy threshold τ_{red} used to determine the residual-request regions in Eq. 20., which is performed on the shared base representation, rather than separately on task-specific features. The percentage above each bar group implies the overall performance across tasks retained relative to the highest performance, when residual are received for all the regions. As evident, receiver-side redundancy estimation suppresses unnecessary residual transmission. We further compare our task-aggregated entropy coding with fixed-length coding where the code length is $\log_2$ (Hu et al., 2024) of the codebook size. By assigning shorter codes to frequent and task-relevant codewords, entropy coding of the base reduces communication by up to 31.4%.

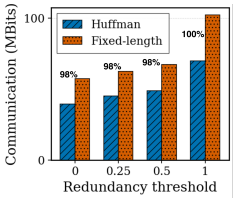


Figure 6: Impact of τ_{red} and entropy coding (OPV2V).

7 Conclusion

In this work, we analyze multi-task CP through a rate-distortion lens and show that joint coding of task-specific representations can reduce communication under matched per-task distortion constraints, by leveraging shared information across tasks. Motivated by this, we develop ORAT, a practical multi-task CP framework that maps multiple task representations into a single task-unified base representation. This base is communicated across the network, which is later adapted to task-compatible features in the receiver-side, supporting multiple downstream tasks. Our results suggest that ORAT is a promising framework for deploying multi-task CP under bandwidth constraints.

Reproducibility Statement

To support reproducibility of our results, we provide the details of our experimental setup including the open-source datasets and pretrained checkpoints, the training protocol, and hyperparameter values used in our experiments, in Section 6 and Appendix A.2.

AI Use Statement

We used AI to refine our writing for clarity and readability of the paper. AI-assisted tools were also used to review and understand some of the relevant prior works. The core motivation, methodology, the experimental results and the writing structure are solely developed by the authors.

Ethics Statement

This work studies communication-efficient multi-task collaborative perception for connected autonomous agents. Our experiments use publicly available datasets and open-source pretrained checkpoints, and do not involve human subjects or private user data. To the best of our knowledge, this work does not introduce any ethical concerns beyond those commonly associated with research on collaborative perception systems.

References

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

Thomas A. Courtade and Tsachy Weissman. Multiterminal source coding under logarithmic loss. *IEEE Transactions on Information Theory*, 60(1):740–761, 2014.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006.

Alexey Dosovitskiy, Germán Ros, Felipe Codevilla, Antonio López, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning (CoRL)*, 2017.

Xiangbo Gao, Runsheng Xu, Jiachen Li, Ziran Wang, Zhiwen Fan, and Zhengzhong Tu. STAMP: scalable task- and model-agnostic collaborative perception. In *The Thirteenth International Conference on Learning Representations, ICLR*, 2025.

Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. In *Thirty-sixth Conference on Neural Information Processing Systems (Neurips)*, 2022.

Yue Hu, Yifan Lu, Runsheng Xu, Weidi Xie, Siheng Chen, and Yanfeng Wang. Collaboration helps camera overtake lidar in 3d detection. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

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

David A. Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, 1952. doi: 10.1109/JRPROC.1952.273898.

Maosen Li, Siheng Chen, Ya Zhang, and Ivor W. Tsang. Graph cross networks with vertex infomax pooling. In *Advances in Neural Information Processing Systems (Neurips)*, 2020.

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

Yiming Li, Dekun Ma, Ziyan An, Zixun Wang, Yiqi Zhong, Siheng Chen, and Chen Feng. V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving. *IEEE Robotics and Automation Letters*, 7(4):10914–10921, 2022.

Genjia Liu, Anning Hu, Yue Hu, Wenjun Zhang, and Siheng Chen. Rate-distortion optimized communication for collaborative perception. *International Conference on Learning Representations (ICLR)*, 2026.

Yen-Cheng Liu, Junjiao Tian, Nathaniel Glaser, and Zsolt Kira. When2com: Multi-agent perception via communication graph grouping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

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, 2023.

Yifan Lu, Yue Hu, Yiqi Zhong, Dequan Wang, Yanfeng Wang, and Siheng Chen. An extensible framework for open heterogeneous collaborative perception. In *International Conference on Learning Representations (ICLR)*, 2024.

Shilpa Mukhopadhyay, Amit Roy-Chowdhury, and Hang Qiu. Coopertrim: Adaptive data selection for uncertainty-aware cooperative perception. In *The Fourteenth International Conference on Learning Representations*, 2026.

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

Hongzhi Ruan, Haibao Yu, Wenxian Yang, Siqi Fan, and Zaiqing Nie. Learning cooperative trajectory representations for motion forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3): 379–423, 1948.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Neural Information Processing Systems*, 2017.

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 *European Conference on Computer Vision (ECCV)*, 2020.

Satosi Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development*, 4(1):66–82, 1960.

Bo Wu, Ruoshen Mo, Justin Yue, Yanyu Zhang, Janice Nguyen, Guoyuan Wu, Amit Roy-Chowdhury, Matthew J. Barth, and Hang Qiu. Cooperscene: Multi-modal cooperative autonomy benchmark with c-v2x communication characterization. In *ECCV*. Springer, 2026.

Aaron D. Wyner and Jacob Ziv. The rate-distortion function for source coding with side information at the decoder. *IEEE Transactions on Information Theory*, 22(1):1–10, 1976.

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 (CVPR)*, 2025.

Runsheng Xu, Yi Guo, Xu Han, Xin Xia, Hao Xiang, and Jiaqi Ma. OpenCDA: An open cooperative driving automation framework integrated with co-simulation. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 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. In *Conference on Robot Learning (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. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022b.

Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi Ma. OPV2V: an open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *International Conference on Robotics and Automation, ICRA 2022, Philadelphia, PA, USA, May 23-27, 2022*, pp. 2583–2589. IEEE, 2022c.

Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, Hongkai Yu, Bolei Zhou, and Jiaqi Ma. 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 (CVPR)*, pp. 13712–13722, June 2023.

Dingkang Yang, Kun Yang, Yuzheng Wang, Jing Liu, Zhi Xu, Rongbin Yin, Peng Zhai, and Lihua Zhang. How2comm: Communication-efficient and collaboration-pragmatic multi-agent perception. In *Advances in Neural Information Processing Systems*, volume 36, pp. 25151–25164, 2023.

Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, and Zaiqing Nie. 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)*, pp. 21329–21338, 2022.

Xumiao Zhang, Anlan Zhang, Jiachen Sun, Xiao Zhu, Y. Ethan Guo, Feng Qian, and Z. Morley Mao. EMP: Edge-assisted multi-vehicle perception. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking (MobiCom)*, pp. 545–558, 2021.

Jiaru Zhong, Jiahao Wang, Jiahui Xu, Xiaofan Li, Zaiqing Nie, and Haibao Yu. Cooptrack: Exploring end-to-end learning for efficient cooperative sequential perception. *ICCV*, 2025.

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 (CVPR)*, pp. 17612–17621, June 2022.

A Appendix

A.1 Proof of Theorem 2

Proof. By definition of total correlation, $\mathrm{TC}\big(F_s^1,\dots,F_s^M \,\big|\, X_r\big) \triangleq \sum_{m=1}^{M} H(F_s^m|X_r) - H(\mathbf{F}_s|X_r)$. Then,

$$R^m_{\text{sep}}(\delta_1,\dots,\delta_M) - R^m_{\text{joint}}(\delta_1,\dots,\delta_M) \tag{21}$$

$$= \sum_{m=1}^{M} (H(F_s^m|X_r) - \delta_m) - I(\mathbf{F}_s; U^*|X_r) \tag{22}$$

$$= \sum_{m=1}^{M} (H(F_s^m|X_r) - \delta_m) - (H(\mathbf{F}_s|X_r) - H(\mathbf{F}_s|U^*, X_r)) + \sum_{m=1}^{M} (H(F_s^m|U^*, X_r) - H(F_s^m|U^*, X_r)) \tag{23}$$

$$= \mathrm{TC}\big(F_s^1,\dots,F_s^M \,\big|\, X_r\big) - \mathrm{TC}\big(F_s^1,\dots,F_s^M \,\big|\, U^*, X_r\big) - \sum_{m=1}^{M} (\delta_m - H(F_s^m|U^*, X_r)) \tag{24}$$

where the last term in Eq. 24 vanishes if $\delta_m = H(F_s^m|U^*, X_r)$ for $m \in [M]$. Consequently, when all per-task distortion budgets are used, the communication saving is exactly equal to the reduction in the (conditional) total correlation. □

A.2 Implementation Details

The experiments were conducted on a single NVIDIA RTX 6000 Ada Generation GPU with 48 GB of memory. The machine was allocated 8 CPU cores and 96 GB of system RAM, running CUDA 12.9.

Since we use open-source trained checkpoints, COBEVT (Xu et al., 2022a) and DiscoNet (Li et al., 2021) for task-specific backbones, we do not perform Stage 1 training as mentioned in Section 5.3. Instead, we keep these backbones frozen and train only the lightweight modules introduced for communication efficiency. In Stage 2, we train the representation aggregator $\mathcal{G}$, the vector quantizer $\mathcal{Q}$, and the task-specific reconstructor modules $\{\mathcal{D}_m\}_{m=1}^{M}$ for 50 epochs. In Stage 3, we train the redundancy estimator $\mathcal{M}$ for 20 epochs. All the stages of training are performed with Adam optimizer. The learning rate is set to 0.001. Unless stated otherwise, the size of the codebook is set to 16, where each codeword has a length of 128.

Aggregator/reconstructor architecture. The representation aggregator $\mathcal{G}$ is implemented as a fully convolutional network. Each agent extracts task-specific BEV features from frozen perception backbones. An aggregation encoder first projects each task feature to a common width of C (the number of channels) using lightweight convolutional stems, concatenates the projected features, and processes them with ConvNeXt blocks to produce a shared base representation F_B. The base is further mapped to a lower-dimensional code space, quantized using layered vector quantization, and projected back to the base feature space after decoding in the receiver-side.

The receiver uses task-specific projection modules (reconstructors) to convert the decoded shared base into task-compatible features. Specifically, the task-specific reconstructors share the same topology: several ConvNeXt blocks operating at the base width C, followed by a 1×1 convolution that maps the base feature to the native task-feature dimension.

Inference speed. We note that the additional modules introduced in the pipeline is lightweight, having minimal impact on the overall inference speed. Computations involved in per-frame inference for all tasks incur a minimal latency of 23.6 ms while substantially reducing the bandwidth usage.

Table 3: Effect of the confidence threshold ϵ_c on perception performance (OPV2V dataset).

ϵ_c	AP@50 (%)	Dynamic IoU (%)	Static IoU (%)	Selected cells (%)
0.05	91	49	35	100
0.10	91	49	35	99
0.25	91	49	35	99
0.40	87	47	34	89
0.50	82	43	32	28.79
0.90	81	39	30	19.71

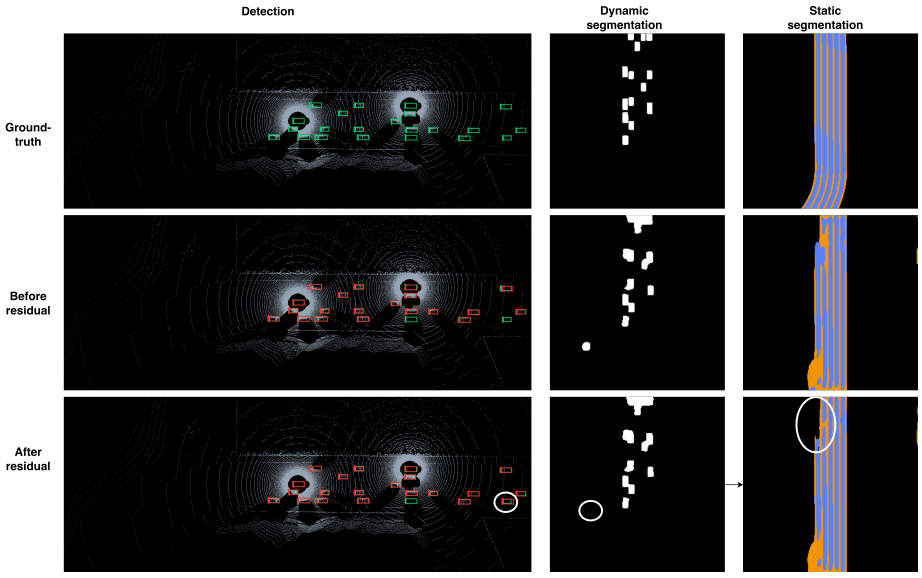
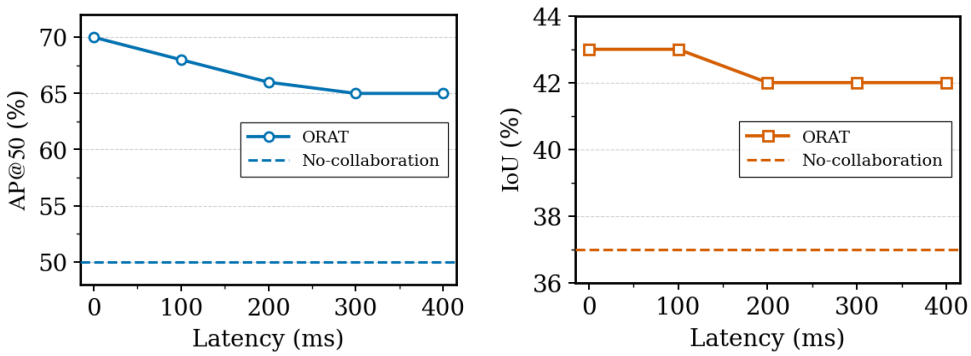


Figure 7: **Visualization of a sample frame from OPV2V dataset.** The improvements introduced by residual refinement are highlighted with white circles.



(a) AP@50 (b) IoU

Figure 8: Performance analysis under varying time delay (V2X-Sim).

A.3 Additional Experiments

Impact of confidence threshold. In Table 3, we study the impact of varying confidence threshold ϵ_c, utilized to filter-out low confidence regions. As ϵ_c increases, more regions are discarded, leading to gradual degradation in the performance for all tasks.

Visualization of residual impact. Fig. 7 presents a qualitative example from the OPV2V dataset. We observe how receiver-requested residuals refine downstream predictions for all tasks. The residual information improves predictions in regions where the receiver has limited field of view, implying that selective refinement helps recover complementary scene details from neighboring agents.

Robustness under time delay. In Fig. 8, we study the robustness of ORAT under delayed delivery of transmission bits. We vary the communication latency from 0 to 500 milliseconds (ms). As expected, perception performance decreases as latency increases because the received features become less temporally aligned with the receiver's current observation. Nevertheless, ORAT consistently outperforms the no-collaboration baseline, implying that the shared representation remains useful even under moderate communication delay.