Carnegie Mellon University

Penn
University of Pennsylvania

The Ohio State University

Morgan State University

CCAC Community College of Allegheny County

The University of Texas Rio Grande Valley

ViCARUS: Vision-centric Connected Autonomy for Vulnerable Road Users' Safety

PI: Vijayakumar Bhagavatula

https://orcid.org/0000-0001-7126-6381

Project Participant: Dereje Shenkut

Final Report – August 2026

DISCLAIMER

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle	5. Report Date
ViCARUS: Vision-centric Connected Autonomy for Vulnerable Road Users' Safety	August 2026
	6. Performing Organization Code
7. Author(s)	**8. Performing Organization Report No.**
Vijayakumar Bhagavatula, https://orcid.org/0000-0001-7126-6381 Dereje Shenkut	
9. Performing Organization Name and Address	**10. Work Unit No.**
Carnegie Mellon University 5000 Forbes Avenue, Pittsburgh, PA 15213	**11. Contract or Grant No.** Federal Grant # 69A3552344811
12. Sponsoring Agency Name and Address	**13. Type of Report and Period Covered**
Safety21 University Transportation Center Carnegie Mellon University 5000 Forbes Avenue Pittsburgh, PA 15213	Final Report (July 1, 2025 – June 30, 2026) **14. Sponsoring Agency Code** USDOT

15. Supplementary Notes

Safety21 Project 582. Conducted in cooperation with the U.S. Department of Transportation, Federal Highway Administration. Code and pretrained models are available at https://github.com/scdrand23/FocalComm and https://github.com/scdrand23/ReVQom.

16. Abstract

The safety of vulnerable road users (VRUs) such as pedestrians, cyclists and motorcyclists remains a pressing challenge on U.S. roads. Cooperative perception and prediction, in which connected vehicles and infrastructure exchange sensing information over Vehicle-to-Everything (V2X) communication, can see past the occlusions and range limits of any single vehicle. These capabilities disproportionately benefit VRUs, who are small, easily occluded, and severely injured in collisions, yet communication bandwidth remains the central challenge to deployment. This report presents the results of three studies conducted under the ViCARUS project, covering a collaborative perception method that prioritizes hard, safety-critical instances during communication, a learned feature codec that compresses exchanged features by more than two orders of magnitude with minimal accuracy loss, and a goal-level communication scheme for cooperative trajectory prediction that matches full-trajectory exchange at one tenth of the bandwidth. Together, the results indicate that cooperative perception and prediction for VRU safety is achievable within practical V2X bandwidth budgets.

17. Key Words	18. Distribution Statement
Vulnerable road users; VRU safety; collaborative perception; V2X communication; 3D object detection; vector quantization; feature compression; cooperative trajectory prediction; connected autonomous vehicles	No restrictions. This document is available through the National Technical Information Service, Springfield, VA 22161.

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	32	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

3GPP	3rd Generation Partnership Project
AP	Average Precision
BEV	Bird's Eye View
bpp	bits per pixel
CAV	Connected Autonomous Vehicle
CMP	Cooperative Motion Prediction
CP	Collaborative Perception
C-V2X	Cellular Vehicle-to-Everything
DAIR-V2X	Real-world vehicle-infrastructure cooperative perception dataset
EMA	Exponential Moving Average
HIM	Hard Instance Mining
IC	Infrastructure-Centric
IoU	Intersection over Union
LiDAR	Light Detection and Ranging
LTE-V2X	Long-Term Evolution Vehicle-to-Everything
mAP	mean Average Precision
minADE	minimum Average Displacement Error
minFDE	minimum Final Displacement Error
MLP	Multi-Layer Perceptron
MTR	Motion Transformer
OPV2V	Open simulated dataset for vehicle-to-vehicle cooperative perception
PoGE	Product of Goal Experts
QAFF	Query-guided Adaptive Feature Fusion
ReLU	Rectified Linear Unit
RVQ	Residual Vector Quantization
USDOT	United States Department of Transportation
V2I	Vehicle-to-Infrastructure
V2V	Vehicle-to-Vehicle
V2X	Vehicle-to-Everything
V2XPnP	Real-world V2X cooperative perception and prediction dataset
V2X-Real	Real-world multi-agent V2X cooperative perception dataset
VC	Vehicle-Centric
VRU	Vulnerable Road User

Chapter 1

Overview

The safety of vulnerable road users (VRUs), including pedestrians, cyclists, motorcyclists, and users of micromobility devices, remains one of the most pressing challenges in U.S. road transportation. VRU deaths have risen rapidly over the past decade. Pedestrian fatalities alone rose from 4,280 in 2010 to an estimated 7,508 in 2022, a 77% increase that far outpaces the 25% rise in total traffic fatalities over the same period [3]. Intersections, where complex interactions occur between vehicles and vulnerable road users (VRUs), account for approximately one quarter of all traffic fatalities and nearly one half of all traffic injuries in the United States [4]. Improving the perception and prediction capability of automated and connected vehicles around VRUs is therefore a direct lever on transportation safety. Although single-vehicle perception has advanced considerably on large-scale driving benchmarks [5, 6, 7], it exhibits fundamental shortcomings, including limited sensing range, susceptibility to occlusions, and sparse measurements on small or distant objects such as VRUs [8]. Collaborative perception (CP) addresses these limitations by allowing connected autonomous vehicles (CAVs) and smart infrastructure to exchange complementary sensing information over Vehicle-to-Everything (V2X) communication, constructing a more complete representation of the traffic environment than any single agent can achieve alone. Our prior work demonstrated both the promise of vehicle-infrastructure collaboration for pedestrian detection and the sensitivity of CP systems to communication latency, interruption, and bandwidth limits [9].

One of the central challenges between this promise and practical deployment of cooperative autonomous driving is *communication bandwidth*. To cooperate, each agent must continuously share what it perceives with its neighbors. The most effective form of sharing in current methods is intermediate fusion, in which agents exchange the bird's eye view (BEV) feature maps computed by their perception networks rather than raw sensor data or final detections. These feature maps are far too large to transmit at the rate driving requires, amounting to hundreds of megabytes per second per agent. A typical map of size $256 \times 128 \times 128$ requires roughly 16.8 MB per frame as 32-bit floats, sent at 10 Hz. Realistic C-V2X links cannot carry this load. The LTE-V2X sidelink defined in 3GPP Release 14 delivers only a few Mbps of usable throughput in practice [10], with reported rates of about 10 Mbps at a range of 10 m that fall to 5 Mbps at 100 m and 2–3 Mbps at 300 m [11]. The longer ranges are precisely the ones cooperation is meant to cover. Agents therefore cannot share everything they see and acquire, and what they choose to share matters as much as how much they transmit. Not every object in a scene is equally worth transmitting. A clearly visible vehicle adds little to a neighbor's picture, while a partially occluded pedestrian that only one agent sees is exactly the evidence cooperation exists to deliver. However, current methods spend bandwidth uniformly across the scene and optimize metrics dominated by vehicles, so the hardest and most safety-critical instances receive no special treatment. Cooperation becomes a question of spending a limited communication budget where it helps perception the most. This

trade-off is captured by the constrained optimization

$$\max_{\theta,\{\mathbf{m}_{ij}\}} \sum_{i=1}^{N} \mathcal{P}\left(f_\theta\left(\mathbf{X}_i, \{\mathbf{m}_{ji}\}_{j\in\mathcal{N}_i}\right), \mathbf{Y}_i\right) \quad \text{subject to} \sum_{i=1}^{N} \sum_{j\in\mathcal{N}_i} \|\mathbf{m}_{ij}\| \leq B, \tag{1.1}$$

where N is the number of collaborating agents, $\mathbf{X}_i$ is the sensor observation of agent i, $\mathcal{N}_i$ is the set of neighboring agents that communicate with agent i, $\mathbf{m}_{ij}$ is the message agent i transmits to agent j, and $\{\mathbf{m}_{ji}\}_{j\in\mathcal{N}_i}$ collects the messages agent i receives from its neighbors. The perception network f_θ, with learnable parameters θ, maps the local observation and the received messages to a perception output, and $\mathcal{P}$ measures the quality of that output against the ground truth $\mathbf{Y}_i$ of agent i, for example through detection average precision. The optimization is over both the network parameters θ and the messages themselves, and the constraint bounds the total size of all messages by the communication budget B. Every method studied in this report answers some aspect of Equation (1.1).

This project addresses three aspects of this problem, and the report is organized accordingly. In Chapter 2 we ask what to focus on when selecting the content of the exchanged messages. Mainly, we argue that existing CP methods optimize aggregate detection metrics that are dominated by vehicles and exchange features uniformly across space, underserving the small, occluded instances that matter most for VRU safety. The proposed method, FocalComm [12], counters this with learnable progressive hard instance mining and query-based fusion, steering both communication and fusion toward the instances for which collaboration has the highest value. In Chapter 3 we address the encoding of the messages. Specifically, we develop ReVQom [13], a learned feature codec based on multi-stage residual vector quantization that transmits only per-pixel code indices, reducing transmission payloads by more than two orders of magnitude while preserving the spatial identity of BEV features. In Chapter 4 we extend cooperation from perception to trajectory prediction and address the semantic level of the messages. Here, we propose GoalMAP, currently under review, in which agents communicate compact Gaussian goal waypoints that capture where a road user intends to go rather than dense features or full trajectories that describe how it gets there. The receiver fuses the transmitted goals through a closed-form product of goal experts, matching or exceeding full-trajectory exchange at one tenth of its bandwidth.

Chapter 5 provides our conclusions and outlines future directions. Publications and products resulting from this project are listed at the end of the report.

Chapter 2

Hard Instance-Aware Multi-Agent Perception

2.1 Introduction

Collaborative perception promises the largest safety gains precisely where single-vehicle perception is weakest, on small, occluded, safety-critical road users. In practice, however, most CP methods are trained and evaluated to maximize vehicle detection metrics, and they exchange features uniformly across the scene. The result is that VRUs, the instances for which detection failures are most catastrophic, receive no special treatment in either the learning objective or the communication protocol. The class imbalance inherent in driving datasets compounds the problem. Vehicles dominate both the annotation counts and the loss signal, so a network trained to maximize mean average precision (mAP) can achieve strong headline numbers while systematically missing VRUs that are small, distant, or partially occluded. From the perspective of Equation (1.1), uniform feature exchange spends the communication budget B on spatial regions in proportion to their area, not in proportion to their difficulty or safety relevance. The instances that collaboration could help most, those that one agent barely sees but another sees well, receive no special treatment.

To this end, we developed **FocalComm** [12], a collaborative perception framework that focuses collaboration on *hard instances*, meaning objects that are difficult for an individual agent to detect but crucial for safety.

2.2 Related Work

Intermediate-fusion CP methods such as V2VNet [14], AttFuse on the OPV2V benchmark [15], DiscoNet [16], V2X-ViT [2], and CoBEVT [17] exchange learned BEV features and fuse them with graph, attention, or transformer mechanisms. A second line of work addresses communication efficiency. Where2comm [18] transmits only spatially confident features, What2comm [19] decouples task-relevant components, and CodeFilling [20] uses codebook-based message representations. These methods select *where* or *what* to send based on confidence or information content, but none of them reason explicitly about instance *difficulty*. Confident regions are by construction the easy ones, so confidence-driven selection can reinforce the neglect of hard, safety-critical instances rather than correct it. Difficulty-aware learning has a long history in single-agent detection, from online hard example mining [21] and focal loss [22] to cascaded refinement [23], and recent 3D detectors such as FocalFormer3D [24] and HINTED [25] mine hard instances through progressive multi-stage refinement. FocalComm brings this line of reasoning to the collaborative setting, where difficulty additionally determines what is worth communicating.

2.3 Approach

The FocalComm architecture is illustrated in Figure 2.1. Each agent, whether the ego vehicle, another connected vehicle, or an infrastructure sensor, processes its point cloud through a shared sparse voxel feature encoder Φ, producing agent-specific BEV feature maps $F_k = \Phi(X_k) \in \mathbb{R}^{H \times W \times C}$. Two components then operate on these features, corresponding to the two stages at which a collaborative perceiver makes resource decisions. A progressive Hard Instance Mining (HIM) module determines which instances deserve collaborative attention, and a Query-guided Adaptive Feature Fusion (QAFF) module determines how each agent's evidence is weighted during fusion.

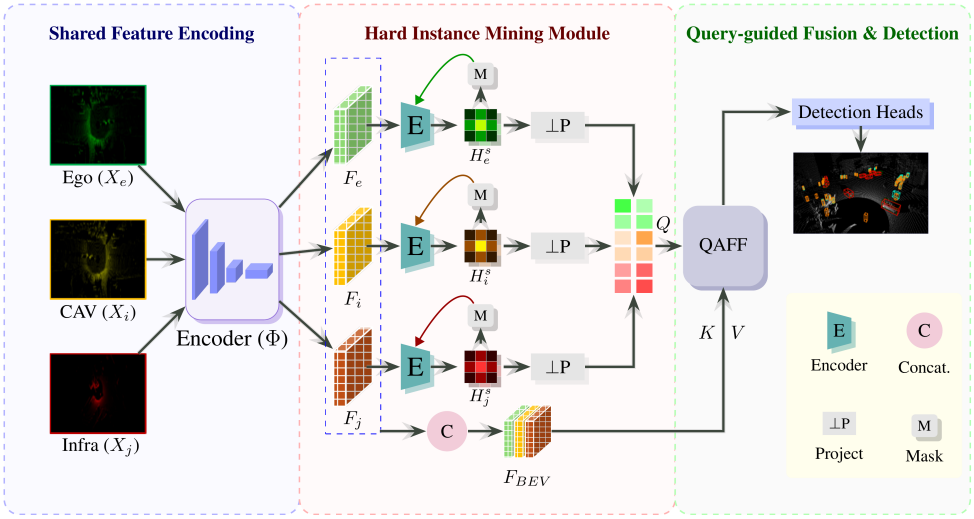


Figure 2.1: Overall architecture of FocalComm. Each agent extracts BEV features through a shared sparse voxel encoder Φ. The progressive Hard Instance Mining (HIM) module identifies hard-to-detect objects across stages by masking out instances already found, and the Query-guided Adaptive Feature Fusion (QAFF) module aggregates multi-agent features using the resulting instance-aware queries before the detection head.

Progressive hard instance mining. HIM extracts hard-instance-oriented features in n_S successive stages, with $n_S = 3$ in our implementation. Let F denote an agent's BEV feature map and let M_{s-1} denote a binary spatial mask marking all instances detected in stages up to $s-1$. Stage s operates on the masked features $F \odot (1 - M_{s-1})$, where $\odot$ is element-wise multiplication, so that regions already explained by earlier stages are suppressed and the current stage attends to what remains. Each stage runs a lightweight detector on its masked features and adds the confident findings to the mask. During training the mask is built by matching stage predictions to ground truth, and during inference by confidence thresholding whose threshold decays from stage to stage so that later stages accept weaker evidence. Easy objects are therefore claimed early, while later stages concentrate on instances with weak or ambiguous evidence such as small, distant, or partially occluded road users. The per-stage features are concatenated into instance-aware query features that carry the mining result into fusion.

Query-guided adaptive feature fusion. QAFF aggregates information across agents under the guidance of the mined queries. For each mining stage, self-attention across agents refines

that stage's queries jointly, and learned stage weights combine the refined queries into a single representation that emphasizes the most informative difficulty level for the current scene. This combined query then attends over the agents' original BEV feature maps through cross-attention, and learned per-agent weights determine each agent's contribution to the fused output, so that each hard instance draws evidence from the agents that observe it best. A validity mask over agents makes the module robust to the varying number of collaborators in a scene.

Detection head and training. The fused features feed an anchor-free transformer detection head [26], which localizes objects directly rather than through predefined anchor boxes, a property that benefits hard instances with unusual scales or occlusion patterns and supports multi-class detection of road users from pedestrians to trucks. Detection and mining are optimized jointly through

$$\mathcal{L} = \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{bbox} + \lambda_3 \mathcal{L}_{hm} + \lambda_4 \sum_{s=1}^{n_S} \mathcal{L}_{him}^{s}, \tag{2.1}$$

where $\mathcal{L}_{cls}$ is a focal classification loss on object categories, $\mathcal{L}_{bbox}$ an L1 regression loss on box parameters, $\mathcal{L}_{hm}$ a Gaussian focal loss on the detection heatmap, $\mathcal{L}_{him}^{s}$ supervises the mining at stage s, and the weights λ_1 through λ_4 balance the terms. Together, mining and query-guided fusion change the allocation in Equation (1.1), so that bandwidth and fusion capacity follow instance difficulty rather than spatial area.

2.4 Experiments

FocalComm is evaluated on two real-world collaborative perception datasets. **V2X-Real** [27] contains 33K LiDAR frames and over 1.2 million annotated 3D boxes collected by two connected vehicles and two smart infrastructure units, with up to four agents per scene. To our knowledge it is the only publicly available real-world multi-agent dataset with multi-class annotations that include pedestrians, which makes it the appropriate testbed for VRU-focused collaboration. **DAIR-V2X** [28] is a large-scale real-world vehicle-infrastructure dataset with 71K frames. For fair comparison, all methods use the same voxel backbone [29] and the same anchor-free detection head [26], and we report AP at IoU thresholds of 0.3 and 0.5 within a ±100 m range of the ego agent.

Table 2.1: Performance comparison on V2X-Real under vehicle-centric (VC) and infrastructure-centric (IC) setups, and on DAIR-V2X vehicle detection. Entries are AP@0.3/AP@0.5. In the no-collaboration row, VC and IC mean vehicle-only and infrastructure-only perception.

Method	**V2X-Real** ($K_{max} = 4$)								**DAIR-V2X** ($K_{max} = 2$)	
	Car VC	Car IC	Ped VC	Ped IC	Truck VC	Truck IC	mAP VC	mAP IC	AP@0.3	AP@0.5
No Collaboration	73.7/68.4	70.6/59.1	31.8/13.9	29.7/10.7	21.2/15.7	46.6/42.0	42.2/32.7	49.0/37.3	58.9	54.4
F-Cooper [1]	88.3/85.6	84.3/80.8	47.8/22.7	45.4/15.9	47.9/46.1	48.3/47.9	61.3/51.4	59.4/48.2	70.4	64.8
V2VNet [14]	87.0/84.4	85.0/81.4	34.5/13.9	36.5/15.2	40.0/36.8	44.3/41.9	53.8/45.0	55.3/46.2	69.5	63.5
AttFuse [15]	81.3/80.7	81.5/80.9	46.8/21.7	48.5/24.8	49.6/47.7	47.6/45.7	59.2/50.0	59.2/50.5	69.7	63.8
CoBEVT [17]	87.2/85.6	84.1/82.1	54.8/26.1	**52.3**/25.6	50.1/45.1	48.9/47.8	64.0/53.3	61.7/**52.9**	72.8	65.7
V2X-ViT [2]	83.9/81.1	81.4/78.2	38.5/15.2	33.5/13.3	42.5/35.6	45.4/38.9	55.0/44.0	53.4/43.5	74.5	67.6
CoAlign [30]	85.8/83.4	84.7/83.4	38.3/17.3	36.4/14.8	52.7/43.9	**53.2/51.1**	59.9/48.2	58.1/49.8	76.9	69.7
ERMVP [31]	88.5/86.4	**86.7**/84.0	53.2/25.4	50.6/23.5	42.9/41.3	41.7/38.7	61.5/51.0	59.7/48.7	69.2	63.4
FocalComm (ours)	**91.5/89.6**	86.2/**84.8**	**57.4/27.3**	51.2/**26.7**	**53.9/51.6**	49.6/47.3	**67.6/56.1**	**62.3/52.9**	**77.2**	**70.1**

Detection performance. Table 2.1 compares FocalComm against state-of-the-art collaborative methods. The best values in each column are indicated in bold font. In the vehicle-centric setting, FocalComm reaches 67.6% mAP@0.3 and 56.1% mAP@0.5, absolute improvements of 5.6% and 5.1% over the next best method. The gains concentrate

exactly where the method aims. Pedestrian detection improves to 57.4% AP@0.3, the best among all methods, and truck detection improves from 21.2% without collaboration to 53.9%. The infrastructure-centric setting shows the same pattern at 62.3% mAP@0.3, and on DAIR-V2X FocalComm generalizes with 77.2% AP@0.3 and 70.1% AP@0.5. In separate communication-scenario evaluations, FocalComm attains 64.8% mAP@0.3 in vehicle-to-vehicle and 70.2% in infrastructure-to-infrastructure configurations, where the elevated infrastructure viewpoint most benefits truck and pedestrian coverage.

Table 2.2: Ablation of FocalComm components on the V2X-Real test set, vehicle-centric setup (AP@0.3/AP@0.5).

Method	Car	Pedestrian	Truck	Overall
No Collaboration	73.7/68.4	31.8/13.9	21.2/15.7	42.2/32.7
+ Collaboration (baseline)	88.3/85.6	47.8/22.7	47.9/46.1	61.3/51.4
+ Collaboration + HIM	91.5/88.8	54.6/26.6	52.6/48.7	66.2/54.7
+ Collaboration + QAFF	91.2/88.2	52.7/23.7	48.7/45.1	65.5/54.0
Full model (both)	**91.5/89.6**	**57.4/27.3**	**53.9/51.6**	**67.6/56.1**

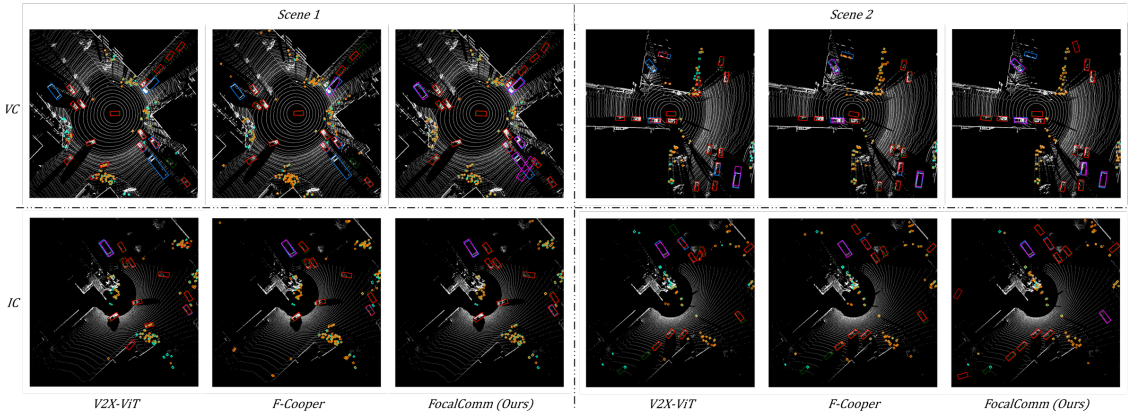


Figure 2.2: Qualitative detection results on V2X-Real, cropped to 80×80 m around the ego agent. Within each scene, columns from left to right show F-Cooper [1], V2X-ViT [2], and FocalComm. Ground truth and predictions are shown for cars (ground truth in dark green, predictions in red), pedestrians (ground truth in cyan, predictions in orange), and trucks (ground truth in light blue, predictions in magenta).

Qualitative comparison. Figure 2.2 compares detections across two scenes. In the dense intersection on the left, FocalComm detects the crowded pedestrians that V2X-ViT and F-Cooper miss, visible as orange predictions aligning with the cyan ground truth, while the baselines leave many cyan boxes unmatched. In the scene on the right, FocalComm maintains reliable detections across all three classes with few false positives. The pedestrian detections in particular illustrate what the quantitative gains mean on the ground, since each detected pedestrian is a road user that a planner would otherwise never see.

Component analysis and bandwidth study. Table 2.2 isolates the contributions. Collaboration alone lifts overall AP@0.3 from 42.2% to 61.3%. Adding HIM contributes a further 4.9% and QAFF alone 4.2%, and the full model reaches 67.6%, indicating the two components are complementary. The pedestrian column tells the VRU story most directly, improving from 31.8% with no collaboration to 57.4% with the full model. HIM concentrates the transmitted

features on hard instances, FocalComm tolerates compression well (Figure 2.3). At 8× compression, mAP@0.3 drops only from 67.6% to 65.9%, and even at an aggressive 32× the method retains 62.1% mAP@0.3. This behavior connects directly to the codec developed in Chapter 3, which pushes compression two orders of magnitude further.

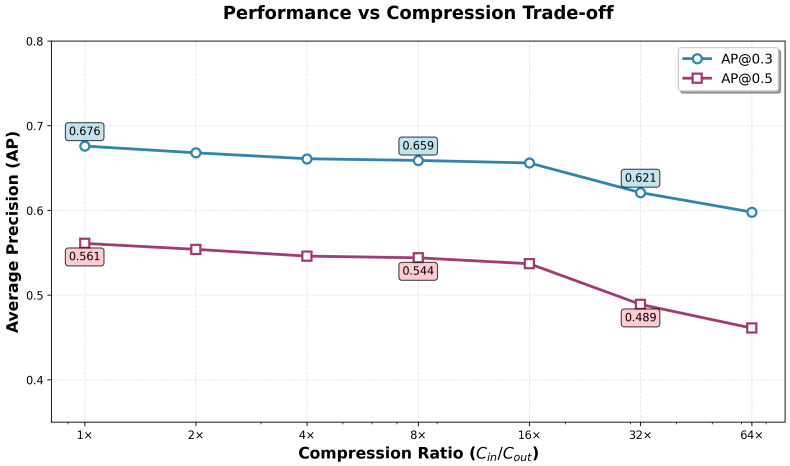


Figure 2.3: Detection performance versus feature compression on V2X-Real. FocalComm holds nearly full accuracy up to 8× compression and degrades gracefully beyond it.

2.5 Summary

FocalComm demonstrates that *what* agents communicate matters as much as how much they communicate. By mining hard instances and fusing them with query-based attention, collaborative perception can be steered toward the failure modes that dominate VRU risk. In Chapter 3 we address the complementary question of *how* to compress the exchanged features aggressively enough for realistic V2X links.

Chapter 3

Communication-Efficient Feature Compression via Residual Vector Quantization

3.1 Introduction

Chapter 2 established which instances deserve the communication budget. This chapter addresses the complementary question of how the message itself should be encoded. Among the fusion strategies available to collaborative perception, intermediate fusion preserves the richest scene evidence short of raw-sensor exchange, and it underlies most current systems [14, 15, 2, 17]. Under this design the message $\mathbf{m}_{ij}$ in Equation (1.1) is a dense BEV feature map, and the link analysis in Chapter 1 shows that full-precision feature exchange exceeds the throughput of deployed V2X radios by more than two orders of magnitude. The encoding of the message, rather than the selection of its content, is therefore the binding constraint this chapter removes.

To this end, we developed **ReVQom** [13], a learned feature codec that compresses BEV features while preserving their spatial identity. A 1×1 bottleneck reduces the channel dimension without mixing spatial locations, and multi-stage residual vector quantization (RVQ) [32] encodes each location as a small set of integer codebook indices. Sender and receiver hold identical codebooks, so only the indices travel over the air (Figure 3.1). The payload falls from 8192 bits per pixel for 32-bit features to 6–30 bits per pixel, a compression of 273× to 1365×, without any change to the fusion backbone.

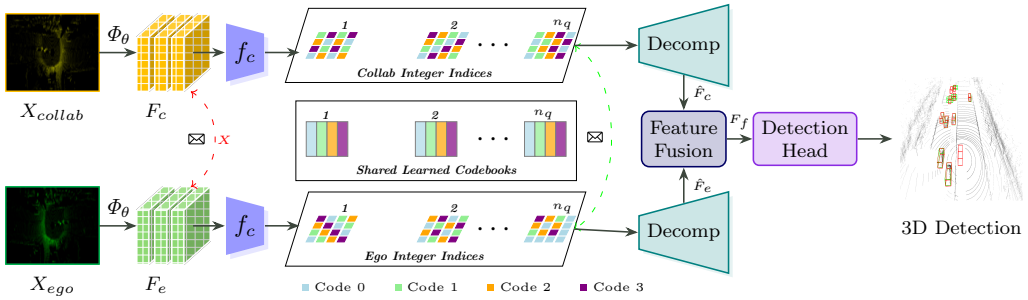


Figure 3.1: ReVQom overview. Each agent extracts BEV features, applies a 1×1 bottleneck and n_q-stage residual vector quantization, and transmits only per-pixel code indices ($\approx HWn_q \log_2 K$ bits) instead of full features ($32 \times C \times H \times W$ bits). With a shared codebook, the receiver decodes the indices, reconstructs the features, and proceeds to multi-agent fusion and detection.

3.2 Related Work

Existing work reduces the communication load of collaborative perception primarily by deciding what to send. Where2comm restricts transmission to the locations its confidence maps mark as informative [18], What2comm decouples and forwards task-relevant feature components [19], and more recent systems exchange codebook indices selected by information filling [20] or summarize scenes as point clusters [33]. These designs save bandwidth by discarding content, and the most aggressive among them give up the dense spatial layout that BEV fusion consumes. How to compress the retained features themselves has received less attention, and the reported ratios remain well below the requirements identified in Chapter 1.

A separate body of work addresses the practical conditions of cooperation, including fusion under communication latency [34], end-to-end cooperative driving [35], and collaboration among heterogeneous agents whose sensors and models differ [36, 37, 38, 39, 40]. These methods presuppose that the exchanged representation is affordable and are complementary to the codec developed here. On the representation side, vector-quantized autoencoders [41] and residual vector quantization from neural audio coding [32] show that learned discrete codes approach the fidelity of continuous features at a small fraction of the rate. ReVQom carries this principle to multi-agent BEV features, where the quantizer must additionally respect spatial structure and remain consistent across agents.

3.3 Approach

ReVQom instantiates the message $\mathbf{m}_{ij}$ of Equation (1.1) as per-pixel code indices, following the protocol in Table 3.1. The sender applies a 1×1 convolution with group normalization to reduce the BEV map $\mathbf{F} \in \mathbb{R}^{H \times W \times C}$ to $C_r \ll C$ channels. Because the convolution has unit spatial support, every location is transformed independently and the spatial identity of the map is preserved. The reduced features are then quantized in n_q residual stages. Stage i selects the nearest entry of its codebook, subtracts it from the running residual, and passes the remainder onward, so later stages refine what earlier stages could not express. The transmitted message consists of the selected indices alone, at a rate of $R = n_q \log_2 K$ bits per spatial location.

The receiver sums the codebook entries named by the received indices, applies a learned affine correction, and expands the result back to C channels before standard multi-agent fusion. Codebooks are learned during training through exponential moving average (EMA) updates with a preset decay α and are stored identically on every agent, so deployment requires no codebook transmission. Training adds two regularizers to the detection loss, a commitment term [41] that keeps encoder outputs close to their quantized versions and an orthogonality term [42] that decorrelates the rows of the encoder weight matrix. The combined objective is

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \beta_{\text{commit}} \|\text{sg}[\mathbf{z}_q] - \mathbf{z}\|_2^2 + \lambda_{\text{ortho}} \|\mathbf{W}_e \mathbf{W}_e^T - \mathbf{I}_{C_r}\|_F^2, \tag{3.1}$$

where $\mathcal{L}_{\text{task}}$ is the detection loss, comprising a focal classification loss and a bounding box regression loss, $\mathbf{z}$ is the reduced feature vector at a spatial location before quantization, $\mathbf{z}_q$ is its quantized version obtained by summing the selected codebook entries across the n_q stages, and $\text{sg}[\cdot]$ denotes the stop-gradient operation, which blocks gradients from flowing into its argument so that the commitment term moves the encoder toward the codebook rather than the reverse. The matrix $\mathbf{W}_e \in \mathbb{R}^{C_r \times C}$ is the weight of the 1×1 channel-reduction convolution, $\mathbf{I}_{C_r}$ is the $C_r \times C_r$ identity matrix, and $\|\cdot\|_F$ is the Frobenius norm, so the orthogonality term penalizes correlated rows in the

encoder. The scalar weights β_{commit} and λ_{ortho} balance the two regularizers against the detection loss and are set to 0.05 and 10^{-4} in our experiments.

Table 3.1: RVQ-based collaboration protocol

Sender Agent (Encoding)	
$\mathbf{F} \in \mathbb{R}^{H \times W \times C}$	▷ *input BEV features*
$\mathbf{F}_r = \text{GroupNorm}(\text{Conv}_{1\times1}(\mathbf{F})) \in \mathbb{R}^{H \times W \times C_r}, \quad C_r \ll C$	▷ *channel reduction*
$\mathbf{r}^{(0)} = \mathbf{F}_r$	▷ *initialize residual*
For $i = 0, 1, \ldots, n_q - 1$:	
$k^{(i)} = \arg\min_{k \in \{1,\ldots,K\}} \|\mathbf{r}^{(i)} - \mathbf{e}_k^{(i)}\|_2^2$	▷ *nearest codebook entry*
$\mathbf{q}^{(i)} = \mathbf{e}_{k^{(i)}}^{(i)}$	▷ *quantized vector*
$\mathbf{r}^{(i+1)} = \mathbf{r}^{(i)} - \mathbf{q}^{(i)}$	▷ *update residual*
$\mathbf{e}_{k^{(i)}}^{(i)} \leftarrow (1-\alpha)\mathbf{e}_{k^{(i)}}^{(i)} + \alpha \mathbf{r}^{(i)}$	▷ *EMA update*
$\mathcal{I} = \{k^{(0)}, k^{(1)}, \ldots, k^{(n_q-1)}\}$	▷ *transmit indices*
$R = n_q \log_2 K$ bits per spatial location	▷ *bitrate*
Receiver Agent (Decoding)	
$\mathcal{I} = \{k^{(0)}, k^{(1)}, \ldots, k^{(n_q-1)}\}$	▷ *receive indices*
$\mathbf{q}^{(i)} = \mathbf{e}_{k^{(i)}}^{(i)}$ for $i = 0, 1, \ldots, n_q - 1$	▷ *codebook lookup*
$\mathbf{z}_q = \sum_{i=0}^{n_q-1} \mathbf{q}^{(i)}$	▷ *accumulate quantized vectors*
$\mathbf{z}'_q = \text{ReLU}(\text{Conv}_{1\times1}(\mathbf{z}_q))$	▷ *post-affine transformation*
$\hat{\mathbf{F}} = \text{ReLU}(\text{Conv}_{1\times1}(\text{GroupNorm}(\mathbf{z}'_q)))$	▷ *channel expansion*
$\mathbf{F}_f = \gamma(\hat{\mathbf{F}}, \mathbf{F}_{\text{local}})$	▷ *multi-agent fusion*

3.4 Experiments

We evaluate ReVQom on the real-world DAIR-V2X vehicle-infrastructure dataset [28] and on the simulated OPV2V benchmark [15], reporting vehicle detection AP at IoU thresholds of 0.3 and 0.5. The codec plugs into a CoBEVT fusion backbone [17] without architectural changes. Unless stated otherwise, experiments use channel reduction ratio $C_{rr} = 16$, $n_q = 3$ quantization stages, EMA decay $\alpha = 0.8$, and codebook sizes K between 4 and 1024, which span 6 to 30 bits per pixel against 8192 for uncompressed 32-bit features.

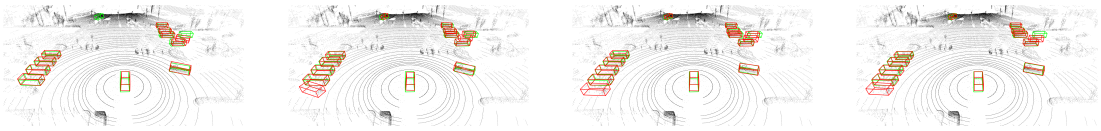
Table 3.2 supports three conclusions. Compression at the right operating point costs no accuracy. ReVQom-S at 18 bits per pixel exceeds its own uncompressed CoBEVT backbone (0.747 against 0.728 AP@0.3) despite 455× compression, and ReVQom-M reaches the best overall 0.753 AP@0.3 at 341×, which indicates that the discrete bottleneck also regularizes the features it compresses. Degradation under extreme rates is graceful rather than abrupt. At 6 bits per pixel and 1365× compression, ReVQom-μ retains 0.690 AP@0.3, above several uncompressed baselines. The behavior transfers across domains. On OPV2V, ReVQom-S matches uncompressed fusion at AP@0.3 (0.946 against 0.947) and concedes 2.9% at AP@0.5.

The qualitative results in Figure 3.2 show the same pattern at the level of individual scenes. Detections at 6 bits per pixel already capture most vehicles, and additional rate mainly tightens box alignment rather than recovering missed objects. Figure 3.3 explains why so few bits suffice. Background occupies 96–98% of spatial locations and collapses onto a single code, while the remaining codes specialize to road and object structure. BEV features are sparse in exactly the way an index-based code exploits.

The ablations in Figure 3.4 identify the operating point used above. Accuracy peaks at

Table 3.2: Comparison of collaborative perception methods on DAIR-V2X and OPV2V. ReVQom achieves 273×–1365× compression while maintaining competitive performance.

Method	K	bpp	AP@0.3	AP@0.5	Compr.
DAIR-V2X Dataset					
No Collaboration	-	0	0.589	0.544	-
F-Cooper [1]	-	8192	0.704	0.648	1×
V2VNet [14]	-	4096	0.695	0.635	2×
AttFuse [15]	-	2048	0.697	0.638	4×
CoBEVT [17]	-	8192	0.728	0.657	1×
V2X-ViT [2]	-	6144	0.745	0.676	1.3×
Where2comm [18]	-	512	0.701	0.634	16×
ReVQom-μ	4	6	0.690	0.558	1365×
ReVQom-T	16	12	0.699	0.609	683×
ReVQom-S	64	18	0.747	0.651	455×
ReVQom-M	256	24	**0.753**	0.666	341×
ReVQom-L	1024	30	0.725	0.636	273×
OPV2V Dataset					
CoBEVT [17]	-	8192	0.947	0.895	1×
ReVQom-S	64	18	0.946	0.869	455×



(a) 6 bits ($K = 4$) (b) 18 bits ($K = 64$) (c) 24 bits ($K = 256$) (d) 8192 bits (Raw)

Figure 3.2: Detection results at increasing bit rates, viewed from the ego vehicle, with ground truth in green and predictions in red. Six bits per pixel already captures most vehicles, and higher rates mainly refine box alignment.

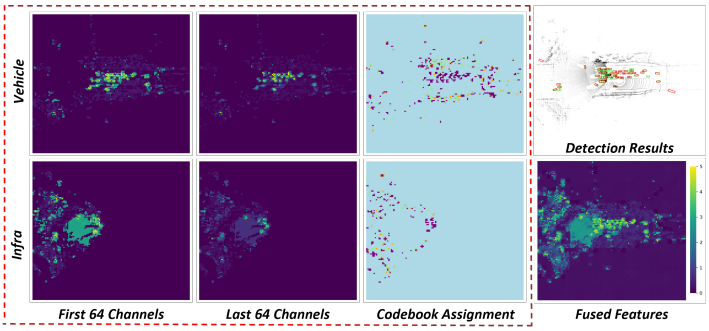


Figure 3.3: Learned codebook assignments ($K = 4$, first quantizer stage) for vehicle (top) and infrastructure (bottom) agents. One code dominates background regions and the remaining codes concentrate on foreground objects, matching the feature value distribution and the resulting detections.

$n_q = 3$ stages and $C_{rr} = 16$ and falls quickly once channel reduction passes that point, which shows that spatial quantization is more forgiving than channel compression. An EMA decay of $\alpha = 0.8$ outperforms the value of 0.99 common in static vector quantization, consistent with the

shifting viewpoint statistics of multi-agent driving scenes. Among codebook sizes, $K = 64$ offers the strongest rate-accuracy trade-off, reaching 99.2% of the $K = 256$ accuracy at 75% of its rate.

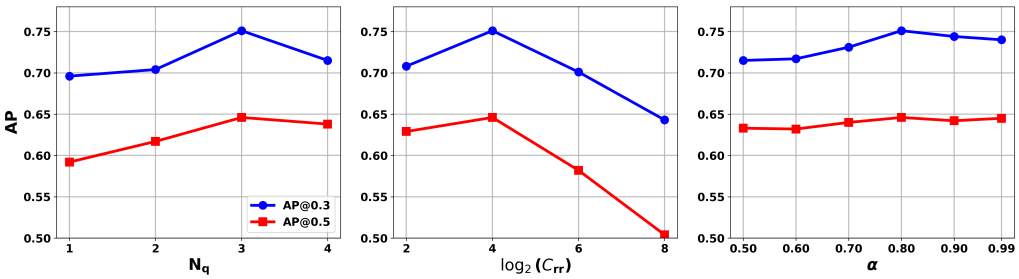


Figure 3.4: Ablations over the number of quantization stages (N_q), channel reduction rate (C_{rr}, log scale), and EMA decay rate (α). AP@0.3 in blue, AP@0.5 in red.

3.5 Summary

ReVQom reduces the per-agent message of intermediate fusion to 6–30 bits per pixel while matching or exceeding uncompressed accuracy at 18–24 bits per pixel. At the 18-bit operating point a message occupies roughly 37 KB per frame, near 3 Mbps at 10 Hz, which sits within the link rates surveyed in Chapter 1. Feature-level cooperation therefore becomes compatible with the radios vehicles carry today. The evaluation is limited to LiDAR-based benchmarks, quantization costs accuracy at strict IoU thresholds, and codebook synchronization under packet loss remains open. Chapter 4 raises the bandwidth question one level of abstraction higher and asks what should be communicated when the task moves from perception to prediction.

Chapter 4

Goal-Oriented Multi-Agent Cooperative Trajectory Prediction

4.1 Introduction

While Chapters 2 and 3 address cooperative perception, safe driving also requires anticipating the future motion of surrounding road users. Cooperative motion prediction has received far less attention than cooperative perception, and existing methods largely treat it as a downstream extension of perception-style communication [43, 44, 45]. This view misses a requirement that is unique to prediction. A receiving vehicle must forecast every road user that could affect its plan, including those that appear nowhere in its own tracking history. Such unseen targets dominate real cooperative driving data. On the real-world V2XPnP benchmark [46], two thirds of receiver-target pairs are unseen by the receiver and only a quarter are co-observed. Unseen targets are also the safety-critical ones, because a planner cannot avoid a road user for which no forecast exists. Existing cooperative prediction methods transmit either dense BEV feature maps, at megabytes per agent per cycle, or complete multi-modal trajectory forecasts, at several kilobytes per target [43, 47]. Neither primitive isolates the quantity the receiver actually lacks. For an unseen road user, the receiver already has its own map and its own view of the scene. What it cannot infer is where that road user is going. GoalMAP communicates that intent and nothing else potentially offering communication efficiency.

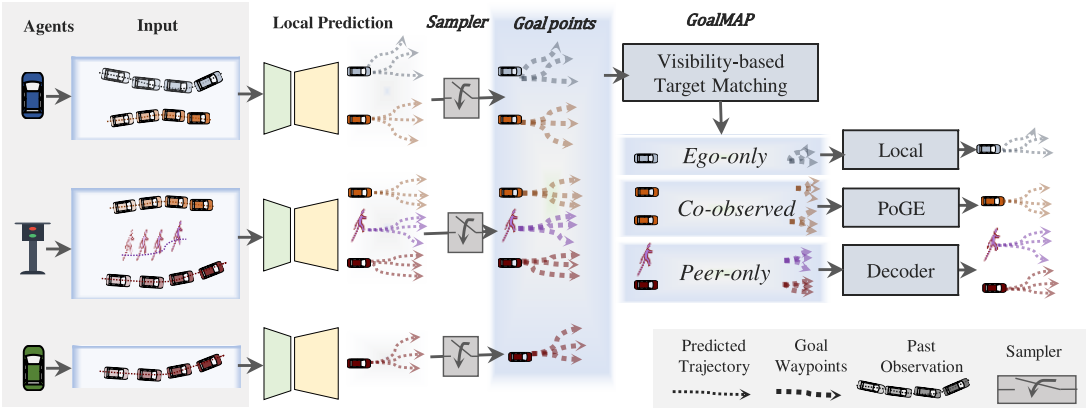


Figure 4.1: GoalMAP pipeline. Each agent runs its own frozen predictor and transmits per-mode Gaussian goal waypoints. Visibility-based matching routes each receiver-target pair to one of three regimes. Local targets keep the receiver's own prediction, co-observed targets fuse the receiver's modes with carriers regenerated from sender goals, and unseen targets are decoded from sender goals alone. Only goal Gaussians are transmitted.

4.2 Approach

4.2.1 Goals as the Communication Primitive

Goal-based decomposition is well established in single-agent prediction, both for vehicles [48, 49, 50, 51] and for pedestrians, where stepwise goal-guided encoding has proven effective [52], evidence that compact goal estimates carry most of the intent information a forecast needs. Building on this observation, each agent transmits, per predicted mode, a small set of bivariate Gaussian goal waypoints that its MTR-style predictor [53] already produces internally, sampled at $S = 5$ goal steps over a $T = 50$-step horizon. The message is 624 bytes per target, roughly one tenth of a full multi-modal trajectory and four orders of magnitude below dense BEV features (Table 4.1).

Table 4.1: Communication primitives at $K = 6$ modes, $T = 50$ future steps, and $S = 5$ goal steps in float32. Trajectory and goal messages carry five Gaussian parameters per point plus one probability per mode.

Primitive	Contents	Payload
Dense BEV features	feature map (per agent)	4.7 MB
Full trajectory	$KT \times 5 + K$	6,024 B
Goal Gaussians (ours)	$KS \times 5 + K$	**624 B**

4.2.2 Carrier Regeneration and the Product of Goal Experts

The receiver regenerates a full trajectory carrier for each mode by piecewise-linear interpolation from the target's reported position through the transmitted waypoints. Cooperation is then cast as a product of goal experts (PoGE) in the sense of [54]. For a given target, the candidate pool contains every predicted mode from every agent that observes it, and each mode i in the pool is described by its S goal waypoints, where waypoint s has mean position $\mu_i^s \in \mathbb{R}^2$ and covariance Σ_i^s transmitted by the predicting agent. Each candidate mode i is scored by the agreement of its goal waypoints with the goal densities contributed by all other agents,

$$A_i = \frac{1}{|\mathcal{J}_i|} \sum_{j \in \mathcal{J}_i} \sum_{s=1}^{S} \log \mathcal{N}\big(\mu_i^s;\, \mu_j^s, \Sigma_j^s\big), \tag{4.1}$$

where $\mathcal{J}_i$ is the set of modes contributed by agents other than the one that produced candidate i, $|\mathcal{J}_i|$ is its size, and $\mathcal{N}(\mu_i^s; \mu_j^s, \Sigma_j^s)$ is a bivariate Gaussian density with mean μ_j^s and covariance Σ_j^s, that is, the goal distribution that mode j predicts for step s, evaluated at the point μ_i^s, the waypoint that candidate i predicts for the same step. The value is high when candidate i's waypoint falls where mode j expects the target to be, and low when the two disagree, with the covariance Σ_j^s setting how much disagreement mode j tolerates. The score A_i is therefore the average log-likelihood of candidate i's waypoints under the goal distributions of every other agent, so a candidate that other agents also consider plausible receives a high score. When $\mathcal{J}_i$ is empty, as for a target seen by only one sender, A_i is set to zero. The fused score adds this agreement to the transmitted mode probability, and the final K modes are selected by farthest-endpoint non-maximum suppression. Both operations are closed form and add zero trainable cooperation parameters, so GoalMAP runs on top of any pretrained goal-based predictor without retraining. Co-observed and unseen targets run through the same operator and differ only in whether the

receiver's own modes join the candidate pool. For an unseen target observed by a single sender, the method reduces exactly to adopting that sender's prediction.

4.3 Results

Evaluation follows the CMP protocol [43], reporting minADE@5s over $K = 6$ modes on the real-world V2XPnP benchmark [46] and the simulated OPV2V benchmark [15], with receiver-observed and unseen targets reported separately. The strongest bandwidth-hungry baseline, direct adoption, transmits each sender's full multi-modal trajectory at 6 KB per target.

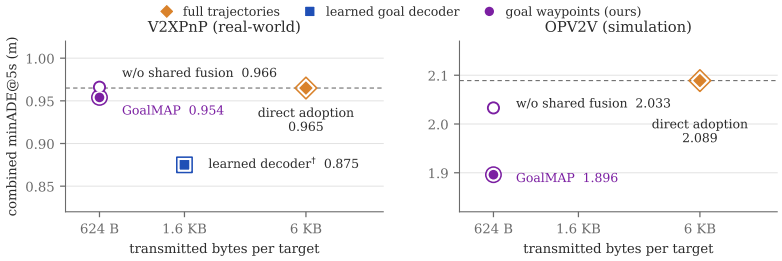


Figure 4.2: Combined minADE against transmitted bytes per target on a log scale. GoalMAP reaches direct-adoption accuracy on V2XPnP and improves on it on OPV2V at 624 bytes per target, one tenth of a full multi-modal trajectory.

On V2XPnP (Table 4.2), GoalMAP matches full-trajectory direct adoption on unseen targets (0.948 vs. 0.946 minADE) at one tenth of the bandwidth and improves receiver-observed targets by 4.0% on the co-observed subset. On OPV2V (Table 4.2), it improves every branch over direct adoption, reducing unseen-target minADE by 6.6% and co-observed minADE by 13.7%. The gains track the amount of cooperative evidence. With a single sender the operator reduces to direct adoption by construction, while with three or more senders observing the same target the unseen-target improvement reaches 18.7% (Figure 4.2). A carrier-fidelity ablation confirms that trajectories regenerated from the transmitted waypoints match full transmitted trajectories in fusion quality to within 0.1% minADE, so the 10× bandwidth reduction is essentially free.

Table 4.2: Cooperative prediction on the real-world V2XPnP benchmark (top) and the simulated OPV2V benchmark (bottom), reported as minADE@5s in meters, lower is better. Bandwidth is the per-target prediction payload. Unseen targets have no forecast at all without cooperation. On OPV2V, tracks come from CoBEVT-fused perception for all rows, so the bandwidth column counts the prediction payload alone.

Method	Bandwidth	Recv-observed	Unseen	Combined
V2XPnP (real-world)				
Receiver only (no cooperation)	0	0.987	n/a	n/a
Direct adoption (full trajectories)	6 KB	0.987	**0.946**	0.965
GoalMAP (ours)	**624 B**	**0.962**	0.948	**0.954**
OPV2V (simulated)				
Receiver only (no pred. cooperation)	0	1.854	n/a	n/a
Direct adoption (full trajectories)	6 KB	1.854	2.558	2.089
GoalMAP (ours)	**624 B**	**1.648**	**2.389**	**1.896**

4.4 Limitations

A paper describing GoalMAP is currently under review. Several limitations remain and point to natural extensions. The frame transform aligns goal means without rotating covariances, which degrades under large relative headings. Piecewise-linear carriers underestimate curvature during sharp maneuvers. Communication latency and packet loss are not modeled, and connecting the framework to our earlier study of latency effects on collaborative perception [9] is a direct next step.

Chapter 5

Conclusions and Future Work

5.1 Summary of Findings

This project investigated bandwidth-constrained multi-agent cooperation for the safety of vulnerable road users along three axes. **FocalComm** (Chapter 2) showed that steering collaboration toward hard instances through learnable progressive hard instance mining and query-based fusion improves 3D detection by up to 5.6% mAP on real-world V2X datasets, with the largest gains on VRU classes. **ReVQom** (Chapter 3) showed that residual vector quantization with pre-shared codebooks compresses per-agent messages by 273×–1365× (to 6–30 bits per pixel) while matching or exceeding uncompressed fusion accuracy at 18–24 bits per pixel. **GoalMAP** (Chapter 4) showed that communicating intent (Gaussian goal waypoints) rather than state (dense features or full trajectories) supports cooperative trajectory prediction at extremely lower rate per target, matching full-trajectory exchange on unseen targets and improving co-observed accuracy by 4.0% on real-world data and 13.7% in simulation. Taken together, these results point to a consistent conclusion. The bandwidth bottleneck of V2X cooperation is best addressed not by compression alone but by deciding what to send (hard instances), how to encode it (learned, spatially faithful quantization), and at which semantic level to communicate (goals rather than full trajectories). At 18 bits per pixel, a ReVQom-compressed BEV message is approximately 37 KB per frame, around 3 Mbps at 10 Hz, and a GoalMAP message is under 1 KB per target. Both fit within the capacity of realistic C-V2X links, in contrast to the tens of megabytes per frame required by uncompressed feature sharing.

5.2 Future Work

Several directions follow naturally from this project's findings. FocalComm decides which features matter and ReVQom decides how to encode them, so combining hard-instance mining with residual quantization, for example by spending more quantization stages on mined regions, would concentrate the communication budget on the VRUs that need it most. Extending goal-level cooperation from vehicle targets to pedestrians and cyclists is a natural next step for GoalMAP, building on our visual-linguistic pedestrian trajectory prediction work [52], and the same intent-level messages could feed cooperative maneuver planning around VRUs. Aligning discrete representations across sensing modalities would further allow LiDAR-equipped vehicles to collaborate with camera-only roadside infrastructure at the intersections where VRU collisions concentrate. Finally, codebook synchronization, packet loss, and adaptive bitrate selection under varying channel quality remain open, as does validation on physical C-V2X testbeds, where the safety benefit to VRUs will ultimately be measured.

PUBLICATIONS AND PRODUCTS

Papers.

- D. Shenkut and V. Bhagavatula, "FocalComm: Hard Instance-Aware Multi-Agent Perception," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026.
- D. Shenkut and B.V.K. Vijaya Kumar, "Residual Vector Quantization for Communication-Efficient Multi-Agent Perception," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2026.
- D. Shenkut and V. Bhagavatula, "Goal-Oriented Multi-Agent Cooperative Motion Prediction," submitted (under review).

Code and Models. The FocalComm implementation, configurations, and pretrained checkpoints are publicly available at `https://github.com/scdrand23/FocalComm`. The ReVQom implementation and pretrained checkpoints are publicly available at `https://github.com/scdrand23/ReVQom`. Both are released under the MIT license. The GoalMAP implementation will be released with its publication upon acceptance.

BIBLIOGRAPHY

[1] Q. Chen, X. Ma, S. Tang, J. Guo, Q. Yang, and S. Fu, "F-Cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3D point clouds," in *Proceedings of the IEEE/ACM Symposium on Edge Computing (SEC)*, pp. 88–100, 2019.

[2] R. Xu, H. Xiang, Z. Tu, X. Xia, M.-H. Yang, and J. Ma, "V2X-ViT: Vehicle-to-everything cooperative perception with vision transformer," in *European Conference on Computer Vision (ECCV)*, pp. 107–124, 2022.

[3] Governors Highway Safety Association, "Pedestrian traffic fatalities by state: 2022 preliminary data," tech. rep., Governors Highway Safety Association, 2023. Accessed: July 31, 2023.

[4] J. B. Shaw, R. J. Porter, M. R. Dunn, J. Soika, and I. B. Huang, "The safe system paradigm: Reducing fatalities and injuries at the nation's intersections," *Public Roads*, 2022. Winter 2022 issue. Federal Highway Administration, U.S. Department of Transportation.

[5] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 3354–3361, 2012.

[6] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 11621–11631, 2020.

[7] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 2446–2454, 2020.

[8] S. Liu, C. Gao, Y. Chen, X. Peng, X. Kong, K. Wang, R. Xu, W. Jiang, H. Xiang, J. Ma, and M. Wang, "Towards vehicle-to-everything autonomous driving: A survey on collaborative perception," *arXiv preprint arXiv:2308.16714*, 2023.

[9] D. Shenkut and B. V. K. Vijaya Kumar, "Impact of latency and bandwidth limitations on the safety performance of collaborative perception," in *Proceedings of the IEEE International Conference on Computer Communications and Networks (ICCCN)*, pp. 1–8, 2024.

[10] M. H. C. Garcia, A. Molina-Galan, M. Boban, J. Gozalvez, B. Coll-Perales, T. Şahin, and A. Kousaridas, "A tutorial on 5G NR V2X communications," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 3, pp. 1972–2026, 2021.

[11] T. Huang, J. Liu, X. Zhou, D. C. Nguyen, M. R. Azghadi, Y. Xia, Q.-L. Han, and S. Sun, "V2X cooperative perception for autonomous driving: Recent advances and challenges," *arXiv preprint arXiv:2310.03525*, 2023.

[12] D. Shenkut and V. Bhagavatula, "FocalComm: Hard instance-aware multi-agent perception," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026.

[13] D. Shenkut and B. V. K. Vijaya Kumar, "Residual vector quantization for communication-efficient multi-agent perception," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2026.

[14] T.-H. Wang, S. Manivasagam, M. Liang, B. Yang, W. Zeng, and R. Urtasun, "V2VNet: Vehicle-to-vehicle communication for joint perception and prediction," in *European Conference on Computer Vision (ECCV)*, pp. 605–621, 2020.

[15] R. Xu, H. Xiang, X. Xia, X. Han, J. Li, and J. Ma, "OPV2V: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2022.

[16] Y. Li, S. Ren, P. Wu, S. Chen, C. Feng, and W. Zhang, "Learning distilled collaboration graph for multi-agent perception," in *Advances in Neural Information Processing Systems*, 2021.

[17] R. Xu, Z. Tu, H. Xiang, W. Shao, B. Zhou, and J. Ma, "CoBEVT: Cooperative bird's eye view semantic segmentation with sparse transformers," in *Proc. Conf. Robot Learning (CoRL)*, pp. 989–1000, 2023.

[18] Y. Hu, S. Fang, Z. Lei, Y. Zhong, and S. Chen, "Where2comm: Communication-efficient collaborative perception via spatial confidence maps," in *Advances in Neural Information Processing Systems*, vol. 35, pp. 4874–4886, 2022.

[19] K. Yang, D. Yang, J. Zhang, H. Wang, P. Sun, and L. Song, "What2comm: Towards communication-efficient collaborative perception via feature decoupling," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023.

[20] Y. Hu, J. Peng, S. Liu, J. Ge, S. Liu, and S. Chen, "Communication-efficient collaborative perception via information filling with codebook," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[21] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Las Vegas, NV), pp. 761–769, IEEE, 2016.

[22] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, (Venice, Italy), pp. 2980–2988, IEEE, 2017.

[23] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6154–6162, IEEE, 2018.

[24] Y. Chen, Z. Yu, Y. Chen, S. Lan, A. Anandkumar, J. Jia, and J. M. Alvarez, "FocalFormer3D: Focusing on hard instance for 3D object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, IEEE, 2023.

[25] Q. Xia, W. Ye, H. Wu, S. Zhao, L. Xing, X. Huang, J. Deng, X. Li, C. Wen, and C. Wang, "HINTED: Hard instance enhanced detector with mixed-density feature fusion for sparsely-supervised 3D object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15321–15330, IEEE, 2024.

[26] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai, "TransFusion: Robust LiDAR-camera fusion for 3D object detection with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, 2022.

[27] H. Xiang, Z. Zheng, X. Xia, R. Xu, L. Gao, Z. Zhou, X. Han, X. Ji, M. Li, Z. Meng, L. Jin, M. Lei, Z. Ma, Z. He, H. Ma, Y. Yuan, Y. Zhao, and J. Ma, "V2X-Real: A large-scale dataset for vehicle-to-everything cooperative perception," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 455–470, 2024.

[28] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan, and Z. 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. 21361–21370, 2022.

[29] Y. Yan, Y. Mao, and B. Li, "SECOND: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, 2018.

[30] Y. Lu, Q. Li, B. Liu, M. Dianati, C. Feng, S. Chen, and Y. Wang, "Robust collaborative 3D object detection in presence of pose errors," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4812–4818, 2023.

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

[32] D. Lee, C. Kim, S. Kim, M. Cho, and W.-S. Han, "Autoregressive image generation using residual quantization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[33] Z. Ding, J. Fu, S. Liu, H. Li, S. Chen, H. Li, S. Zhang, and X. Zhou, "Point cluster: A compact message unit for communication-efficient collaborative perception," in *The Thirteenth International Conference on Learning Representations*, 2025.

[34] Z. Lei, S. Ren, Y. Hu, W. Zhang, and S. Chen, "Latency-aware collaborative perception," in *Proc. European Conf. Computer Vision*, 2022.

[35] J. Cui, H. Qiu, D. Chen, P. Stone, and Y. Zhu, "Coopernaut: End-to-end driving with cooperative perception for networked vehicles," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2022.

[36] R. Xu, J. Li, X. Dong, H. Yu, and J. Ma, "Bridging the domain gap for multi-agent perception," in *Proc. IEEE Int. Conf. Robotics and Automation*, 2023.

[37] Y. Lu, Y. Hu, Y. Zhong, D. Wang, Y. Wang, and S. Chen, "An extensible framework for open heterogeneous collaborative perception," in *The Twelfth International Conference on Learning Representations*, 2024.

[38] C. Shao, G. Luo, Q. Yuan, Y. Chen, Y. Liu, K. Gong, and J. Li, "Hetecooper: Feature collaboration graph for heterogeneous collaborative perception," in *Proc. European Conf. Computer Vision*, pp. 162–178, 2024.

[39] J. Zhou, P. Dai, Q. Wei, B. Liu, X. Wu, and J. Wang, "Pragmatic heterogeneous collaborative perception via generative communication mechanism," in *Advances in Neural Information Processing Systems*, 2025.

[40] C. Shao, Q. Yuan, G. Luo, Y. Hu, D. Wang, Y. Liu, R. Pan, B. Chen, and J. Li, "NegoCollab: A common representation negotiation approach for heterogeneous collaborative perception," in *Advances in Neural Information Processing Systems*, 2025.

[41] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," *Advances in Neural Information Processing Systems*, pp. 6306–6315, 2017.

[42] N. Bansal, X. Chen, and Z. Wang, "Can we gain more from orthogonality regularizations in training deep CNNs?," in *Advances in Neural Information Processing Systems*, 2018.

[43] Z. Wang, Y. Wang, Z. Wu, H. Ma, Z. Li, H. Qiu, and J. Li, "CMP: Cooperative motion prediction with multi-agent communication," *IEEE Robotics and Automation Letters*, 2025.

[44] H. Ruan, H. Yu, W. Yang, S. Fan, and Z. Nie, "Learning cooperative trajectory representations for motion forecasting," in *Advances in Neural Information Processing Systems*, vol. 37, 2024.

[45] X. Zhang, Z. Zhou, Z. Wang, Y. Ji, Y. Huang, and H. Chen, "Co-MTP: A cooperative trajectory prediction framework with multi-temporal fusion for autonomous driving," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[46] Z. Zhou, H. Xiang, Z. Zheng, S. Z. Zhao, M. Lei, Y. Zhang, T. Cai, X. Liu, J. Liu, M. Bajji, X. Xia, Z. Huang, B. Zhou, and J. Ma, "V2XPnP: Vehicle-to-everything spatio-temporal fusion for multi-agent perception and prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[47] K. Wu, J. Qiao, and Y. Zhang, "CoPAD: Multi-source trajectory fusion and cooperative trajectory prediction with anchor-oriented decoder in V2X scenarios," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025.

[48] C. Wang, Y. Wang, M. Xu, and D. J. Crandall, "Stepwise goal-driven networks for trajectory prediction," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2716–2723, 2022.

[49] J. Gu, C. Sun, and H. Zhao, "DenseTNT: End-to-end trajectory prediction from dense goal sets," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15303–15312, 2021.

[50] K. Mangalam, H. Girase, S. Agarwal, K.-H. Lee, E. Adeli, J. Malik, and A. Gaidon, "It is not the journey but the destination: Endpoint conditioned trajectory prediction," in *European Conference on Computer Vision (ECCV)*, pp. 759–776, 2020.

[51] L. Zhang, P.-H. Su, J. Hoang, G. C. Haynes, and M. Marchetti-Bowick, "Map-adaptive goal-based trajectory prediction," in *Proceedings of the 2020 Conference on Robot Learning*, vol. 155 of *Proceedings of Machine Learning Research*, pp. 1371–1383, 2021.

[52] D. Shenkut and B. V. K. Vijaya Kumar, "Visual-linguistic reasoning for pedestrian trajectory prediction," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, pp. 771–778, 2025.

[53] S. Shi, L. Jiang, D. Dai, and B. Schiele, "Motion transformer with global intention localization and local movement refinement," in *Advances in Neural Information Processing Systems*, vol. 35, pp. 6531–6543, 2022.

[54] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.