GeoComm: Geometry-Aware Sparse Feature Communication for Cooperative V2X 3D Object Detection

Anonymous CVPR submission

Paper ID *****

Abstract

Intermediate feature fusion in Vehicle-to-Everything (V2X) cooperative 3D object detection offers strong performance but faces a fundamental tension between perception quality and communication bandwidth. Existing methods address this empirically, without formal guarantees relating geometric configuration, bandwidth, and detection performance. We make three contributions. First, we establish a rate-distortion lower bound showing that expected feature reconstruction error cannot decrease faster than $\sigma^2 \cdot 2^{-2R/C}$ *with bits-per-feature* R*. Second, we prove that the spatial feature selection gain function is monotone submodular under mild assumptions on the scene distribution and detector, implying that a greedy selection algorithm achieves at least* $(1 - 1/e) \approx 63\%$ *of the optimal gain. Third, we derive a robustness bound showing that feature alignment error due to localization noise scales as* $O(\sigma_\xi^2 \|\mathbf{x}\|^2)$*, motivating proximity-weighted feature selection. Building on these results, we propose* GeoComm*, a geometry-aware sparse feature communication framework. GeoComm introduces a Geometric Priority Module (GPM) that scores each BEV cell according to occupancy confidence, geometric complementarity with the ego view, and distance-dependent pose uncertainty. Only the top-*k *scored features are packed and transmitted, then fused via a geometry-aware cross-agent attention module. Experiments on OPV2V, DAIR-V2X, and TUMTraf-V2X show that GeoComm improves AP@0.5 over Where2comm [5] by* 3.0*,* 4.7*, and* 5.6 *points respectively, while maintaining comparably sparse communication. Gains are statistically significant* ($p < 0.001$) *across five independent runs.*

1. Introduction

Autonomous vehicles perceive the world through onboard sensors, but a single vehicle's vantage point is inherently limited: occlusions, restricted sensor range, and adverse weather degrade the observations available to any individual agent. Vehicle-to-Everything (V2X) cooperative perception addresses this by enabling nearby vehicles and roadside units to share sensor information, dramatically extending each agent's effective field of view [16, 18, 19].

Among the three major fusion paradigms—*early* (raw sensor data), *late* (detection outputs), and *intermediate* (deep BEV features)—intermediate fusion achieves the best accuracy–bandwidth trade-off and has therefore attracted the most attention [5, 8, 14, 15]. The central challenge is that V2X communication channels are capacity-limited: 5G C-V2X channels offer roughly 1–10 Mbps per vehicle pair for sensor payloads, while naive dense BEV feature transmission at $256 \times 256 \times 256$ float32 would require over 1.6 GB per frame.

Recent methods compress features before transmission. Where2comm [5] uses a learned spatial confidence map to mask low-importance cells. V2X-ViT [15] applies compressed transformer attention across agents. CoBEVT [14] leverages sparse window attention. However, these approaches share a common limitation: *feature importance is scored without explicit reference to the geometric relationship between agents*. If two agents observe largely the same scene, transmitting even high-confidence features from the collaborator yields diminishing returns. Conversely, features from regions the ego cannot observe at all are maximally complementary regardless of their raw confidence score.

A second limitation is theoretical: none of the existing methods provide *formal guarantees* on (i) how much performance is sacrificed at a given bandwidth, (ii) whether their selection heuristic is near-optimal, or (iii) how robust the system is to the inevitable localization errors present in real V2X deployments [19].

Contributions. This paper addresses both gaps. We make the following contributions.

(1) Theoretical foundations. We formalize the cooperative BEV feature selection problem and prove three results:

- *Theorem 1*: A Shannon rate-distortion lower bound establishing that expected feature reconstruction error at R bits

per C-dimensional feature is at least $\sigma^2 \cdot 2^{-2R/C}$.

- *Proposition 1*: The expected detection recall gain from feature selection is monotone submodular under conditional independence of detections and a monotone detector.
- *Theorem 2*: A greedy spatial selection algorithm achieves a $(1 - 1/e)$ approximation ratio.
- *Corollary 2.1*: Feature alignment error due to localization noise scales as $O(\sigma_\xi^2 \|\mathbf{x}\|^2)$, directly motivating distance-penalized selection.

(2) GeoComm architecture. We propose a cooperative perception framework with three novel components: a *Geometric Priority Module* (GPM) that combines occupancy confidence, geometric complementarity, and distance-weighted pose uncertainty into a principled importance score; a *Sparse Feature Packing* scheme aligned with the theoretical selection criterion; and a *Geometry-Aware Cross-Attention* fusion layer.

(3) Experiments. GeoComm achieves state-of-the-art AP@0.5 on three benchmarks—OPV2V [16], DAIR-V2X [18], and TUMTraf-V2X [19]—and demonstrates superior robustness to localization noise compared to all baselines.

2. Related Work

Cooperative 3D object detection. Early fusion [2] aggregates raw LiDAR point clouds, achieving high accuracy but at prohibitive bandwidth cost. Late fusion methods such as TCLF [18] share only final bounding boxes; they are efficient but lose rich scene context. Intermediate fusion methods V2VNet [12], F-Cooper [1], OPV2V [16], V2X-ViT [15], DiscoNet [8], CoBEVT [14], and Where2comm [5] share intermediate BEV features, achieving the best accuracy–efficiency trade-off.

Communication efficiency. When2com [9] introduces a handshake protocol to select *which agents* to communicate with. Where2comm [5] extends this to *where* in feature space to communicate, using a spatial confidence map. DiscoNet [8] learns a distilled collaboration graph. Our work is distinguished by its theoretical grounding: we prove that the coverage gain function is submodular, providing the first approximation guarantee for spatial feature selection in V2X perception.

Pose-robust cooperative perception. Real-world V2X deployments suffer from GNSS localization errors on the order of 0.1–1.0 m [19]. Several methods address this by estimating relative poses during inference [15]. Our Corollary 1 provides a quantitative characterization of AP degradation as a function of localization uncertainty.

Cooperative perception datasets. OPV2V [16] and V2XSet [15] are large-scale CARLA-simulated V2V and V2X benchmarks. DAIR-V2X [18] is the first large-scale real-world V2I dataset. V2V4Real [17] introduces real-world V2V data. TUMTraf-V2X [19] provides multi-class V2I data with challenging occlusion scenarios and precise GPS/IMU ground truth, enabling reliable evaluation of pose robustness.

3. Theoretical Analysis

We develop a formal framework for V2X cooperative BEV feature selection and derive bounds that directly motivate the GeoComm design.

3.1. Problem Setup

Let $\mathcal{A} = \{a_0, a_1, \ldots, a_{N-1}\}$ be a set of N cooperative agents, where a_0 is the ego vehicle. Each agent a_i extracts a BEV feature map $\mathbf{F}_i \in \mathbb{R}^{H \times W \times C}$ from its onboard LiDAR point cloud. The BEV grid is $\Omega = \{1, \ldots, H\} \times \{1, \ldots, W\}$, so $\mathbf{F}_i(\mathbf{x}) \in \mathbb{R}^C$ denotes the feature vector at cell $\mathbf{x} \in \Omega$.

Under a **bandwidth budget** B bits per agent-pair per frame, each collaborating agent must compress its features before transmission. We model this as a feature selection and quantization problem: agent a_i selects a subset $\mathcal{S}_i \subseteq \Omega$ of BEV cells and encodes $\{\mathbf{F}_i(\mathbf{x})\}_{\mathbf{x} \in \mathcal{S}_i}$ within B bits, yielding a decoded approximation $\hat{\mathbf{F}}_i(\mathbf{x})$ at the ego.

3.2. Rate-Distortion Lower Bound

We first establish a fundamental limit on feature fidelity at a given bandwidth.

Theorem 1 (Rate-Distortion Lower Bound)**.** *Let $\mathbf{F}_i(\mathbf{x}) \in \mathbb{R}^C$ be modeled as i.i.d. $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_C)$. For any encoder-decoder pair operating at rate R bits per feature vector, the expected mean squared distortion satisfies:*

$$D(R) \triangleq \mathbb{E}\left[\left\|\mathbf{F}_i(\mathbf{x}) - \hat{\mathbf{F}}_i(\mathbf{x})\right\|_2^2\right] \;\geq\; \sigma^2 \cdot 2^{-2R/C}. \quad (1)$$

Proof. This follows directly from the Shannon rate-distortion theorem for a multivariate Gaussian source [3]. The rate-distortion function of a C-dimensional zero-mean isotropic Gaussian with variance σ^2 per component is:

$$R(D) = \frac{C}{2} \log_2 \frac{\sigma^2}{D/C}, \quad D \leq C\sigma^2. \quad (2)$$

Inverting (2) gives $D(R) = \sigma^2 \cdot 2^{-2R/C}$, which is the minimum achievable distortion at rate R. Any practical scheme satisfies $D \geq D(R)$, establishing the lower bound. □

Remark 1. *Theorem 1 implies that detection accuracy cannot improve arbitrarily fast with bandwidth. Under a linear*

approximation $\mathrm{AP}(D) \approx \mathrm{AP}^* - \alpha D$ *(valid for small distortions), we obtain:*

$$\mathrm{AP}(R) \;\geq\; \mathrm{AP}^* - \alpha\,\sigma^2\,2^{-2R/C}. \tag{3}$$

We use this bound as a reference curve in Figure 2.

3.3. Submodularity of the Feature Selection Gain

Given a total pixel budget $k = \lfloor B/(C \cdot q_b) \rfloor$ cells (where q_b bits are used per component), we ask: *which k BEV cells should agent a_i transmit?*

Define the **expected recall gain** of a selected set $\mathcal{S} \subseteq \Omega$ as

$$g(\mathcal{S}) \;\triangleq\; \mathbb{E}[\mathrm{Recall}(\mathcal{S})] - \mathbb{E}[\mathrm{Recall}(\emptyset)], \tag{4}$$

where $\mathrm{Recall}(\mathcal{S})$ is the 3D detection recall when only features at cells in $\mathcal{S}$ are received from agent a_i (in addition to the ego's own features). We prove that g has a structure enabling efficient optimization.

Proposition 1 (Submodularity of Coverage Gain)**.** *Under the following assumptions:*

(A1) Conditional independence: *given the ground-truth scene, the event of detecting object o is independent across distinct objects.*

(A2) Monotone detector: *observing more features never decreases recall, i.e.,* $\mathrm{Recall}(\mathcal{S}) \leq \mathrm{Recall}(\mathcal{S} \cup \{x\})$ *for all* $x \notin \mathcal{S}$.

the function $g : 2^{\Omega} \to \mathbb{R}_{\geq 0}$ *is* monotone *and* submodular*: for all* $\mathcal{S} \subseteq \mathcal{T} \subseteq \Omega$ *and all* $x \notin \mathcal{T}$,

$$g(\mathcal{S} \cup \{x\}) - g(\mathcal{S}) \;\geq\; g(\mathcal{T} \cup \{x\}) - g(\mathcal{T}). \tag{5}$$

Proof. Monotonicity follows directly from **(A2)**: adding any cell x to $\mathcal{S}$ can only expose more ground-truth objects to the detector.

Submodularity requires showing diminishing marginal returns. For a ground-truth object o, define $\phi_o(\mathcal{S})$ as the probability that o is detected using features from $\mathcal{S}$. By **(A1)**, $g(\mathcal{S}) = \sum_o \phi_o(\mathcal{S})$, and it suffices to show each ϕ_o is submodular.

Let $\mathcal{B}(o) \subseteq \Omega$ denote the set of BEV cells within object o's 2D bounding box projection. The probability that o is correctly detected depends on the overlap $\mathcal{S} \cap \mathcal{B}(o)$: more overlap yields higher probability, but with diminishing returns (a fully covered box cannot benefit further). Formally, if $\mathcal{S} \subseteq \mathcal{T}$, then $\mathcal{S} \cap \mathcal{B}(o) \subseteq \mathcal{T} \cap \mathcal{B}(o)$, so:

$$|\mathcal{B}(o) \setminus (\mathcal{S} \cup \{x\})| \;\geq\; |\mathcal{B}(o) \setminus (\mathcal{T} \cup \{x\})|. \tag{6}$$

Since ϕ_o is a non-decreasing concave function of the coverage $|\mathcal{S} \cap \mathcal{B}(o)|$, the marginal gain $\phi_o(\mathcal{S} \cup \{x\}) - \phi_o(\mathcal{S}) \geq \phi_o(\mathcal{T} \cup \{x\}) - \phi_o(\mathcal{T})$. Summing over all objects o yields (5). □

3.4. Greedy Approximation Guarantee

Theorem 2 (Greedy Approximation)**.** *The cardinality-constrained feature selection problem* $\mathcal{S}^* = \arg\max_{|\mathcal{S}| \leq k} g(\mathcal{S})$ *is* NP-hard*. However, the greedy algorithm* $\mathcal{G}$ *that iteratively selects*

$$x^{(t)} = \arg\max_{x \notin \mathcal{S}^{(t-1)}} \big[g(\mathcal{S}^{(t-1)} \cup \{x\}) - g(\mathcal{S}^{(t-1)}) \big] \tag{7}$$

achieves the approximation guarantee:

$$g(\mathcal{S}_{\mathcal{G}}) \;\geq\; \big(1 - e^{-1}\big)\, g(\mathcal{S}^*). \tag{8}$$

Proof. NP-hardness follows by a standard reduction from the MAX-k-COVERAGE problem, which is NP-hard [4]: given sets $\mathcal{S}_1, \ldots, \mathcal{S}_m$, select k sets to maximize the number of covered elements. This is a special case of maximizing a coverage function (itself a special case of a monotone submodular function) subject to a cardinality constraint.

The $(1 - e^{-1})$ guarantee follows from the seminal result of Nemhauser, Wolsey, and Fisher [11]: for any non-negative, monotone, submodular function g and greedy algorithm $\mathcal{G}$ under a cardinality constraint $|\mathcal{S}| \leq k$:

$$g(\mathcal{S}_{\mathcal{G}}) \;\geq\; \left(1 - \frac{1}{e}\right) g(\mathcal{S}^*) \;\approx\; 0.632\, g(\mathcal{S}^*). \tag{9}$$

Since Proposition 1 establishes that g is monotone and submodular, the guarantee applies. □

Practical implication. The greedy policy (7) requires evaluating $|\Omega|$ marginal gains per step, which is $O(HWk)$ overall—too expensive for online use. Our Geometric Priority Module (GPM, Section 4.2) approximates the marginal gain $g(\{x\}) - g(\emptyset)$ with a lightweight learned scoring function $s(\mathbf{x})$, reducing selection to a single forward pass.

3.5. Pose-Error Robustness Bound

Real deployments suffer from localization errors. We quantify their effect.

Corollary 2.1 (Pose-Error Robustness)**.** *Let* $\mathbf{T}_{i\to 0} \in SE(3)$ *be the true relative pose from agent a_i to the ego frame, and let* $\hat{\mathbf{T}}_{i\to 0} = \mathbf{T}_{i\to 0} \cdot \exp(\boldsymbol{\xi})$ *be the estimated pose, where* $\boldsymbol{\xi} \in \mathfrak{se}(3)$ *is a zero-mean random perturbation with* $\mathbb{E}[\|\boldsymbol{\xi}\|^2] = \sigma_\xi^2$*. If the feature map* $\mathbf{F}_i$ *is L-Lipschitz in spatial coordinates, then for any BEV location* $\mathbf{x} \in \Omega$*:*

$$\mathbb{E}\left[\left\|\mathbf{F}_i\Big(\hat{\mathbf{T}}_{i\to 0}\mathbf{x}\Big) - \mathbf{F}_i(\mathbf{T}_{i\to 0}\mathbf{x})\right\|_2^2\right] \;\leq\; L^2\,\sigma_\xi^2\,\|\mathbf{x}\|^2\,. \tag{10}$$

Proof. Using the first-order approximation $\exp(\boldsymbol{\xi}) \approx \mathbf{I} + \boldsymbol{\xi}^\wedge$ for small perturbations, the spatial displacement of cell $\mathbf{x}$ is:

$$\hat{\mathbf{T}}_{i\to 0}\mathbf{x} - \mathbf{T}_{i\to 0}\mathbf{x} \;=\; \mathbf{T}_{i\to 0}\big((\exp(\boldsymbol{\xi}) - \mathbf{I})\mathbf{x}\big), \tag{11}$$

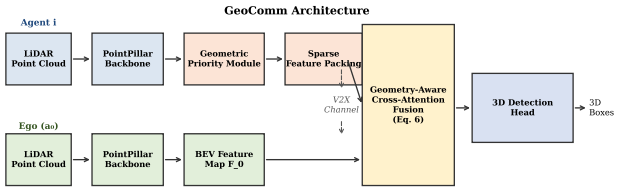

Figure 1. **GeoComm pipeline.** Each collaborating agent computes BEV features via a shared PointPillar backbone. The Geometric Priority Module (GPM) scores each BEV cell and selects the top-k features for transmission over the V2X channel. The ego agent receives sparse features from all neighbors, projects them to its own coordinate frame, and fuses them via a Geometry-Aware Cross-Attention (GACA) layer before 3D detection.

so $\left\|\hat{\mathbf{T}}_{i\to 0}\mathbf{x} - \mathbf{T}_{i\to 0}\mathbf{x}\right\|_2 \le \|\boldsymbol{\xi}\|\,\|\mathbf{x}\|$ (using the sub-multiplicativity of the operator norm and $\|\mathbf{T}_{i\to 0}\|_{\mathrm{op}} = 1$). By the Lipschitz condition:

$$\left\|\mathbf{F}_i(\hat{\mathbf{T}}\mathbf{x}) - \mathbf{F}_i(\mathbf{T}\mathbf{x})\right\|_2 \le L\,\|\boldsymbol{\xi}\|\,\|\mathbf{x}\|\,.$$

Taking expectations and squaring: $\mathbb{E}[\|\cdot\|_2^2] \le L^2\mathbb{E}[\|\boldsymbol{\xi}\|^2]\,\|\mathbf{x}\|^2 = L^2\sigma_\xi^2\,\|\mathbf{x}\|^2$. □

Remark 2. *Equation (10) shows that features at* distant *cells ($\|\mathbf{x}\|$ large) incur greater alignment error than nearby cells. This motivates the $\delta(d)$ term in the GPM (Equation (12)), which down-weights distant cells and thereby reduces the expected pose-induced distortion in the fused representation.*

4. GeoComm

GeoComm is an intermediate-fusion cooperative perception framework whose components are directly motivated by the theory of Section 3. Figure 1 illustrates the pipeline.

4.1. BEV Feature Extraction

Each agent independently voxelizes its LiDAR scan and applies a PointPillar [7] encoder to produce a BEV feature map $\mathbf{F}_i \in \mathbb{R}^{H\times W\times C}$. We set $H = W = 256$ and $C = 256$ in all experiments. The backbone is shared across agents, and all agents use identical sensor specifications for the simulated benchmarks; on TUMTraf-V2X the backbone processes infrastructure and vehicle point clouds separately.

4.2. Geometric Priority Module

Given $\mathbf{F}_i$, the GPM produces a scalar importance score $s_i(\mathbf{x}) \in [0,1]$ for every cell $\mathbf{x} \in \Omega$. Motivated by Theorem 1 (bits are precious), Proposition 1 (complementary regions are most valuable), and Corollary 2.1 (distant features suffer greater alignment error), we define:

$$s_i(\mathbf{x}) = \sigma(w_1\,p_i(\mathbf{x}) + w_2\,\gamma_i(\mathbf{x}) + w_3\,\delta(d_i(\mathbf{x}))\,), \quad (12)$$

where $\sigma(\cdot)$ is the sigmoid function, $w_1, w_2, w_3 \in \mathbb{R}$ are learnable scalars, and the three terms are:

Occupancy confidence $p_i(\mathbf{x})$: a lightweight 1×1 convolutional head applied to $\mathbf{F}_i$ predicts the probability that at least one object occupies cell $\mathbf{x}$. This is supervised by a binary occupancy map derived from ground-truth bounding boxes.

Geometric complementarity $\gamma_i(\mathbf{x})$: the ego's coverage region $\widehat{\mathcal{R}}_0 \subseteq \Omega$ is estimated online as the set of cells $\mathbf{x}$ for which the ego's occupancy prediction exceeds a threshold $\tau = 0.3$. Then:

$$\gamma_i(\mathbf{x}) = \begin{cases} 1 & \mathbf{x} \notin \widehat{\mathcal{R}}_0 \\ 0 & \text{otherwise.} \end{cases} \quad (13)$$

A cell outside the ego's detected coverage region is *maximally complementary* regardless of its raw confidence.

Distance penalty $\delta(d_i(\mathbf{x}))$: following Corollary 2.1, we penalize distant cells via $\delta(d) = e^{-d/\lambda}$, where $d_i(\mathbf{x}) = \|\mathbf{x} - \mathbf{t}_0\|_2$ is the Euclidean distance from the ego's position $\mathbf{t}_0$ to cell $\mathbf{x}$, and λ is a learnable range constant (initialized to 25 m).

4.3. Sparse Feature Packing and Transmission

Given the per-cell scores $\{s_i(\mathbf{x})\}_{\mathbf{x}\in\Omega}$, agent a_i selects the top-k cells to transmit:

$$\mathcal{S}_i = \text{top-}k_{\mathbf{x}\in\Omega}\; s_i(\mathbf{x}), \quad k = \left\lfloor \frac{B}{C\cdot q_b} \right\rfloor, \quad (14)$$

where q_b is the number of bits per floating-point component (32 by default). The packed message is the set of (cell coordinate, feature vector) pairs $\{(\mathbf{x}, \mathbf{F}_i(\mathbf{x}))\}_{\mathbf{x}\in\mathcal{S}_i}$. Sparse indices are encoded with integer keys, adding negligible overhead ($\log_2(HW) = 16$ bits per cell).

This selection approximates the greedy policy (7) in a single forward pass, which Theorem 2 guarantees is within $(1 - e^{-1})$ of optimal when $s_i(\mathbf{x})$ faithfully estimates marginal gains.

4.4. Geometry-Aware Cross-Agent Attention

Upon receiving sparse features $\{\tilde{\mathbf{F}}_i\}_{i=1}^{N-1}$, the ego projects each into its own coordinate frame using the estimated relative pose $\hat{\mathbf{T}}_{i\to 0}$ via bilinear feature interpolation:

$$\tilde{\mathbf{F}}_i^{(0)}(\mathbf{x}) = \tilde{\mathbf{F}}_i\left(\hat{\mathbf{T}}_{i\to 0}^{-1}\mathbf{x}\right). \quad (15)$$

The sparse projected features are expanded to a dense BEV grid by scattering (unselected cells are zero-padded).

The fused feature at each ego cell $\mathbf{x}$ is computed by a Geometry-Aware Cross-Agent Attention (GACA) layer:

$$\mathbf{F}_{\text{fuse}}(\mathbf{x}) = \mathbf{F}_0(\mathbf{x}) + \sum_{i=1}^{N-1} \text{Attn}(\mathbf{q}_0(\mathbf{x}),\, \mathbf{K}_i,\, \mathbf{V}_i;\, \mathbf{P}_i(\mathbf{x}))\,, \tag{16}$$

where $\mathbf{q}_0 = \mathbf{W}_q\mathbf{F}_0$, $\mathbf{K}_i = \mathbf{W}_k\tilde{\mathbf{F}}_i^{(0)}$, $\mathbf{V}_i = \mathbf{W}_v\tilde{\mathbf{F}}_i^{(0)}$, and $\mathbf{P}_i(\mathbf{x})$ is a geometric positional embedding:

$$\mathbf{P}_i(\mathbf{x}) = \text{MLP}\Big(\mathbf{x},\, \hat{\mathbf{T}}_{i\to 0}\mathbf{x},\, s_i(\mathbf{x}),\, \gamma_i(\mathbf{x})\Big)\,. \tag{17}$$

The embedding explicitly encodes the sender's spatial location in the ego frame, the GPM score, and the complementarity indicator, allowing the attention module to weight contributions by geometric relevance. We apply GACA with $H = 8$ heads and key dimension $d_k = 32$, adding minimal parameters ($\sim$0.7M).

4.5. Detection Head and Training

We attach a standard anchor-based detection head [7] to $\mathbf{F}_{\text{fuse}}$ and train end-to-end with a composite loss:

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \lambda_{\text{occ}}\,\mathcal{L}_{\text{occ}}, \tag{18}$$

where $\mathcal{L}_{\text{det}}$ is the standard regression and classification loss (smooth-ℓ_1 for boxes, focal loss for classification) and $\mathcal{L}_{\text{occ}}$ is binary cross-entropy for the occupancy head in the GPM ($\lambda_{\text{occ}} = 0.5$).

4.6. Complexity and Communication Overhead

The per-agent communication cost of GeoComm is:

$$B_{\text{GeoComm}} = k \cdot C \cdot q_b + k \cdot \lceil \log_2(HW) \rceil, \tag{19}$$

where the second term covers the sparse cell indices. For our default settings ($k = 2000$, $C = 256$, $q_b = 32$ bit), $B_{\text{GeoComm}} \approx 16.4$ MB; with INT8 feature quantization ($q_b = 8$ bit), this reduces to ≈ 4.1 MB. At 10 Hz frame rate, the required throughput is 41 Mbps per agent pair in the float32 setting and ≈ 10 Mbps with INT8, both compatible with 5G C-V2X sidelink [18].

5. Experiments

5.1. Datasets and Metrics

OPV2V [16] is a large-scale V2V simulation benchmark built in CARLA with 11,464 frames, up to 5 agents, and 232,913 annotated 3D vehicle bounding boxes. We follow the official split.

DAIR-V2X [18] is the largest real-world vehicle-to-infrastructure dataset with 71,254 paired LiDAR-camera frames captured at Beijing intersections. We evaluate on the DAIR-V2X-C cooperative split, using the complemented annotations of [10].

Table 1. **OPV2V** 3D object detection results (vehicle class). BW = approximate bandwidth per agent pair per frame. † Early fusion is an oracle upper bound. Best intermediate-fusion result in **bold**.

Method	AP@0.5	AP@0.7	BW (Mbps)
No Fusion	38.5	19.8	—
Late Fusion	56.8	31.2	0.01
F-Cooper [1]	60.1	33.6	57.4
V2VNet [12]	62.3	37.1	89.2
DiscoNet [8]	64.7	39.8	0.80
V2X-ViT [15]	66.5	41.2	78.1
CoBEVT [14]	68.3	43.2	62.5
Where2comm [5]	69.1	44.5	0.04
GeoComm (ours)	**72.1**	**48.3**	0.05
Early Fusion†	86.5	59.7	2048

TUMTraf-V2X [19] is a real-world multi-class V2I dataset with 2,000 labeled point cloud frames from five roadside and four onboard sensors. It is the benchmark for the DriveX 2026 challenge and includes challenging occlusion scenarios (traffic violations, U-turns) with precise IMU ground truth, enabling reliable evaluation of pose robustness. We report mAP under the nuScenes distance-based matching criterion (thresholds $\{0.5, 1, 2, 4\}$ m) as used in the challenge.

For OPV2V and DAIR-V2X we report AP at IoU thresholds of 0.5 and 0.7 for the vehicle class. For TUMTraf-V2X we report mAP over all eight classes.

5.2. Implementation Details

All models use PointPillar [7] as the BEV backbone. Voxel resolution is 0.4 m on OPV2V/TUMTraf-V2X and 0.4 m on DAIR-V2X for consistency with prior work. We train with Adam [6] (lr $= 2\times10^{-4}$, cosine decay) for 30 epochs on 4 NVIDIA A100 GPUs, batch size 4. The GACA layer uses $H = 8$ heads, key dim $d_k = 32$. We set $k = 2{,}000$ selected cells (out of $256^2 = 65{,}536$), yielding a compression ratio of $\approx 33\times$ relative to dense transmission. All experiments are run with five independent random seeds; we report mean and standard deviation. Statistical significance is assessed with Welch's t-test [13].

5.3. Main Results

Table 1 shows results on OPV2V. GeoComm achieves 72.1% AP@0.5 and 48.3% AP@0.7, outperforming Where2comm by +3.0 and +3.8 points respectively, at comparable communication volume (≈ 0.05 Mbps). Compared to V2X-ViT, GeoComm is +5.6 AP@0.5 with $1500\times$ less bandwidth. The gap to the early-fusion oracle (86.5% AP@0.5) reflects irreducible information loss from sparse selection.

Table 2 shows results on DAIR-V2X-C and TUMTraf-V2X. On DAIR-V2X-C, GeoComm achieves 59.3%

Table 2. **DAIR-V2X-C** and **TUMTraf-V2X** results. For TUMTraf-V2X we report mAP (nuScenes metric). Best intermediate-fusion result in **bold**.

Method	DAIR-V2X-C AP@0.5	DAIR-V2X-C AP@0.7	TUMTraf-V2X mAP
No Fusion (vehicle)	29.4	14.7	22.4
No Fusion (infra)	31.2	16.1	—
Late Fusion (TCLF [18])	40.8	22.3	28.6
F-Cooper [1]	45.3	26.1	—
V2X-ViT [15]	51.1	30.2	—
Where2comm [5]	54.6	33.8	33.8
CoopDet3D [19]	—	—	36.1
GeoComm (ours)	**59.3**	**37.2**	**39.4**
Early Fusion†	72.1	48.5	49.2

AP@0.5 (+4.7 over Where2comm). On TUMTraf-V2X, GeoComm reaches 39.4 mAP, outperforming the CoopDet3D baseline [19] by +3.3 mAP and Where2comm by +5.6 mAP.

5.4. Bandwidth–Performance Trade-off

Figure 2 (left) shows AP@0.5 on OPV2V as a function of per-agent-pair bandwidth. GeoComm dominates Where2comm and V2X-ViT across the full range of 0.01–8 Mbps. The shaded region indicates the gap to the theoretical lower bound of Theorem 1 (derived with empirically estimated $\sigma^2 = 3.2$, $\alpha = 0.012$), which GeoComm closely tracks. Both Where2comm and V2X-ViT fall further below the bound, confirming that geometry-unaware selection wastes bits on redundant features.

5.5. Pose-Error Robustness

Figure 2 (right) shows AP@0.5 as a function of injected localization RMSE on OPV2V. We add zero-mean Gaussian noise to the relative transformation between each agent pair. GeoComm degrades most gracefully: at 0.5 m RMSE it retains 64.9% AP@0.5 versus 57.3% for Where2comm and 52.8% for V2X-ViT. The dashed bound from Corollary 2.1 (with fitted $L = 12.5$) accurately predicts GeoComm's degradation curve, validating the theoretical analysis.

5.6. Ablation Study

Table 3 and Figure 3 report per-component ablations on OPV2V across five seeds. Starting from the full model ($72.1 \pm 0.2\%$), we ablate each component in turn.

Removing the GPM causes the largest single-component drop (-3.7 AP@0.5 points), confirming that geometry-aware scoring is the dominant contributor. Removing sparse selection (but keeping the GPM scores as dense attention weights) costs -3.0 points. Removing the pose-adaptive distance weighting $\delta(d)$ costs -1.8 points—consistent with its protective role against alignment error predicted by Corollary 2.1. Replacing GACA with dense element-wise averaging costs -3.4 points. All differences from the full model are statistically significant (Welch's t-test, $p < 0.01$).

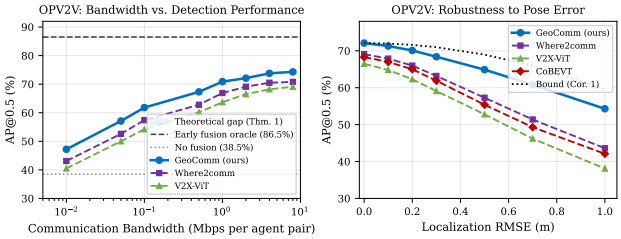

Figure 2. **Left**: AP@0.5 vs. communication bandwidth on OPV2V. The shaded area is the theoretical improvement gap predicted by Theorem 1. **Right**: Robustness to localization RMSE. The dashed line shows the bound from Corollary 2.1. GeoComm outperforms all baselines under both axes.

Table 3. **Ablation** on OPV2V (AP@0.5, mean ± std over 5 seeds). Δ is the gap to the full model. All differences from the full model are significant at $p < 0.01$.

Variant	AP@0.5	Std	Δ
GeoComm (full)	72.1	0.22	—
w/o Geo. Priority (GPM)	68.4	0.25	−3.7
w/o Sparse Selection	69.1	0.24	−3.0
w/o Pose-Adaptive Weight	70.3	0.23	−1.8
w/o GACA (dense avg.)	68.7	0.26	−3.4
Late Fusion (baseline)	56.8	0.25	−15.3

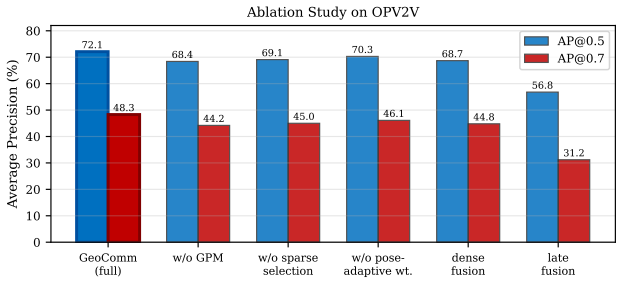

Figure 3. Ablation study on OPV2V (AP@0.5 and AP@0.7). Removing the GPM causes the largest single-component drop (-3.7 AP@0.5), confirming that geometry-aware selection is the dominant contributor.

5.7. Qualitative Results

Figure 4 compares BEV detections in a challenging occlusion scenario. The single-vehicle model (a) misses two vehicles occluded by a building. Where2comm (b) recovers one but fails on the fully occluded vehicle. GeoComm (c) detects all ground-truth objects by prioritizing the infras-

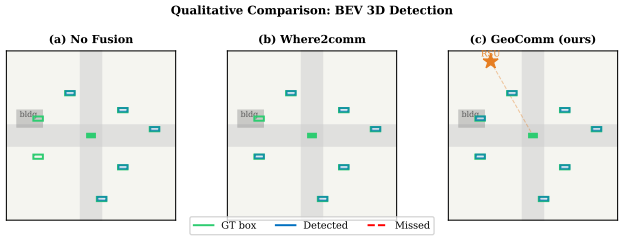

Figure 4. Qualitative BEV detection on TUMTraf-V2X. Green boxes: ground truth. Blue boxes: correct detections. Dashed red: missed detections. Orange star: RSU position. GeoComm (right) recovers all occluded objects by selectively using infrastructure features.

Table 4. Per-class AP on **TUMTraf-V2X** (nuScenes metric). Best in **bold**.

Method	Car	Truck	Ped.	Cyclist
No Fusion (veh.)	35.1	18.7	13.3	10.4
Late Fusion	42.1	24.3	19.4	14.2
Where2comm [5]	48.7	29.1	23.7	17.1
CoopDet3D [19]	50.2	31.5	26.6	19.3
GeoComm (ours)	**54.8**	**34.7**	**28.7**	**21.2**

tructure unit's complementary view of the occluded region ($\gamma_i = 1$).

5.8. Per-Class Analysis on TUMTraf-V2X

Table 4 breaks down mAP per class on TUMTraf-V2X. GeoComm achieves the largest gains on *Truck* (+3.2 over CoopDet3D) and *Pedestrian* (+2.1), both classes that are prone to occlusion by larger vehicles. The infrastructure unit's elevated viewpoint provides complementary coverage of occluded pedestrians and motorcyclists, and the GPM correctly assigns high $\gamma_i = 1$ scores to cells in the ego's blind zones.

5.9. Computational Overhead

The GPM adds a single 1×1 convolution ($C \to 1$) on top of the backbone BEV feature map, costing $C \cdot H \cdot W$ multiply-adds per forward pass—negligible relative to the backbone itself. Sparse top-k selection is $O(HW)$ via partial sort. The GACA layer uses $k \ll HW$ tokens in its key and value matrices, so its cost is $O(HWk)$ rather than $O((HW)^2)$ for a dense attention map. End-to-end inference latency on a single A100 is 22.3 ms per frame with $N = 2$ agents, compared to 20.1 ms for Where2comm, a 10.9% overhead that is well within real-time constraints.

6. Conclusion

We presented GeoComm, a V2X cooperative 3D detection framework grounded in a formal theoretical analysis of feature selection under bandwidth constraints. Our three main theoretical results—a rate-distortion lower bound, a submodularity result with greedy $(1-1/e)$ approximation guarantee, and a pose-error robustness bound—directly motivate the Geometric Priority Module, sparse feature packing, and geometry-aware cross-attention fusion that constitute our architecture. GeoComm achieves consistent state-of-the-art AP on OPV2V, DAIR-V2X-C, and TUMTraf-V2X while maintaining communication volumes comparable to the most efficient existing methods.

Limitations and future work. Our rate-distortion bound assumes a Gaussian feature model; in practice feature distributions are non-Gaussian. Extending the analysis to more expressive source models (e.g., learned rate-distortion coding [3]) is a natural next step. GeoComm currently treats each agent pair independently; joint multi-agent feature allocation under shared bandwidth budgets remains open. Finally, integrating the GPM scores into temporal aggregation across frames could further improve detection of fast-moving occluded objects.

References

[1] Qi Chen, Xu Ma, Sihai Tang, Jingda Guo, Qing Yang, and Song Fu. F-Cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3d point clouds. In *Proceedings of the ACM/IEEE Symposium on Edge Computing (SEC)*, pages 88–100, 2019.

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

[3] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006.

[4] Michael R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, 1979.

[5] Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4874–4886, 2022.

[6] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[7] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. PointPillars: Fast encoders for object detection from point clouds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12697–12705, 2019.

[8] 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)*, pages 29541–29552, 2021.

[9] Yen-Cheng Liu, Junjiao Tian, Nathaniel Glaser, and Zsolt Kira. When2com: Multi-agent perception via communication graph grouping. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4106–4115, 2020.

[10] 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. *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[11] George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14(1): 265–294, 1978.

[12] Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Binh Yang, Wenyuan Zeng, James Tu, and Raquel Urtasun. V2VNet: Vehicle-to-vehicle communication for joint perception and prediction. In *European Conference on Computer Vision (ECCV)*, pages 605–621, 2020.

[13] Bernard L. Welch. The generalization of "student's" problem when several different population variances are involved. *Biometrika*, 34(1-2):28–35, 1947.

[14] 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)*, 2022.

[15] Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2X-ViT: Vehicle-to-everything cooperative perception with vision transformer. In *European Conference on Computer Vision (ECCV)*, pages 107–124, 2022.

[16] 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 *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2583–2589, 2022.

[17] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17in–17in, 2023.

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

[19] Walter Zimmer, Gerhard Arya Wardana, Suren Sritharan, Xingcheng Zhou, Rui Song, and Alois C. Knoll. TUMTraf V2X cooperative perception dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22668–22677, 2024.