RC-GeoCP: Geometric Consensus for 4D Radar-Camera Collaborative Perception

Anonymous authors
Paper under double-blind review

Abstract

Collaborative perception (CP) extends sensing range through feature sharing, but most systems remain LiDAR-centric. Camera and 4D radar sensing combines dense semantics with lower-cost range–velocity measurements, yet exploiting their complementarity across agents remains difficult. Camera depth ambiguity produces spatially diffuse BEV evidence, while sparse transmission couples the choice of evidence to its subsequent contribution. Separate acquisition and aggregation objectives can assign conflicting source preferences to the retained evidence. We propose RC-GeoCP, connecting radar-grounded representation with persistent receiver–source attribution. Geometric Structure Rectification (GSR) uses radar-conditioned deformable sampling and gated calibration to construct grounded BEV features before communication. On these features, Uncertainty-Aware Communication (UAC) combines receiver demand with semantic confidence and radar support in a pre-delivery responsibility field that selects sparse support within fixed quotas. The Consensus-Driven Assembler (CDA) maps the same field into a bounded source mixture, carrying acquisition preferences into assembly. Persistent attribution coordinates spatial selection and source weighting across sparse transmission. Across two datasets and three aggregation backbones, RC-GeoCP consistently improves detection. Its sparse model retains 98.6%/96.6% of dense AP@0.7 at 10.88/4.95 MB per link. Code will be released.

1 Introduction

Collaborative perception (CP) extends the sensing range and reliability of autonomous driving by sharing observations among vehicles and roadside infrastructure (Xu et al., 2022b; Yu et al., 2022; Xu et al., 2023b). By accessing information beyond each agent's line of sight, CP mitigates blind spots and severe occlusions (Wang et al., 2020; Li et al., 2021; Zhou et al., 2025), leading to rapid progress in intermediate feature sharing (Chen et al., 2019; Xu et al., 2022a), communication efficiency (Hu et al., 2022; 2024; Yazgan et al., 2025), and multi-agent fusion (Xiang et al., 2023; Xu et al., 2023a; Tang et al., 2025).

Most existing CP frameworks are LiDAR-centric (Liu et al., 2020a; Hu et al., 2022; Xu et al., 2022a). They benefit from LiDAR's precise 3D geometry for spatial alignment and feature-level consensus (Chen et al., 2019; Gu et al., 2023; Yang et al., 2024), but large-scale deployment is constrained by high sensor cost and weather-related degradation (Dreissig et al., 2023).

Cameras provide dense visual observations and rich semantics for object recognition and scene understanding (Mohammed et al., 2020; Li et al., 2022; Wang et al., 2024b). However, camera-only perception is sensitive to environmental conditions and viewpoint changes (Li et al., 2023b; Park et al., 2023). Moreover, monocular features suffer from inherent depth ambiguity, causing spatial smearing along the optical ray during BEV projection (Li et al., 2023b; Philion & Fidler, 2020). In collaborative settings, such ambiguity is amplified across agents, leading to geometric misalignment and degraded structural consistency (Xu et al., 2023a; Xiang et al., 2023; Wang et al., 2024a). In contrast, radar sensing offers complementary advantages (Yang et al., 2025; Huang et al., 2025). Millimeter-wave radar provides direct range and velocity measurements that often remain useful as visibility degrades (Fan et al., 2024; Nabati & Qi, 2021; Scheiner et al., 2021). These geometric cues also remain useful across viewpoints (Kim et al., 2023; Xiong et al., 2024). While recent radar-

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

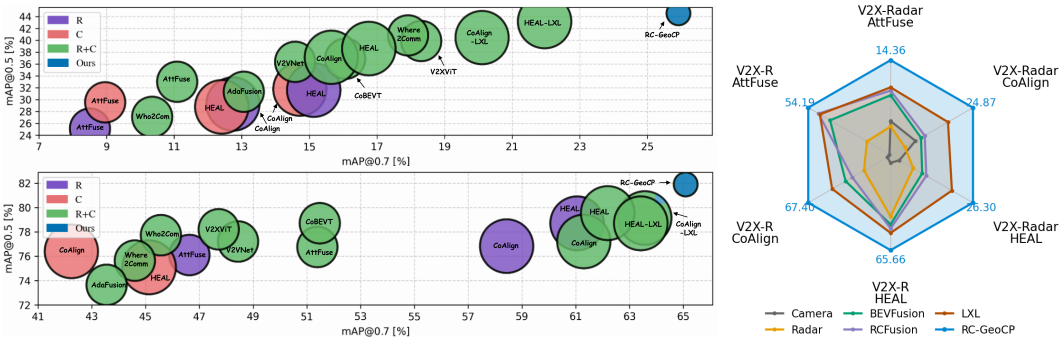

Figure 1: Accuracy–feature-payload and fusion comparisons on V2X-Radar (top left) and V2X-R (bottom left). Bubble size denotes feature payload, while the quantitative tables report complete algorithmic traffic. Color denotes the sensing modality. The radar plot compares relative AP@0.7 across AttFuse, CoAlign, and HEAL.

based CP methods leverage these properties to enhance robustness (Yazgan et al., 2025; Huang et al., 2025), radar signals are sparse and lack semantics, making them insufficient when used alone.

These observations expose two linked challenges. Camera depth ambiguity disperses local BEV evidence, while sparse collaboration couples which source-specific evidence to transmit with how it should influence the receiver (Zhang et al., 2026; Yazgan et al., 2025). Independently learned decisions can favor different sources. Selected evidence may then be attenuated during assembly, although discarded local alternatives survive only as compressed context. This motivates *geometric consensus*, a radar-supported receiver–source responsibility that persists through acquisition and assembly, linking grounded local evidence to its use after delivery.

In this work, we propose RC-GeoCP, connecting radar-grounded representation with persistent source attribution. Building on selective collaboration and aligned fusion (Liu et al., 2020b; Zhang et al., 2026; Shenkut & Bhagavatula, 2026; Lu et al., 2023), Geometric Structure Rectification (GSR) grounds camera semantics in radar-supported spatial structure. Uncertainty-Aware Communication (UAC) then combines ego uncertainty, neighbor confidence, complementarity, and radar reliability into a pre-delivery responsibility. Within fixed source quotas, it induces Top-K spatial support. After token arrival, the Consensus-Driven Assembler (CDA) maps the same responsibility to a bounded distribution over valid sources. The source preferences used for acquisition therefore also guide assembly of the retained evidence. We evaluate RC-GeoCP on V2X-Radar and V2X-R. Figure 1 summarizes its accuracy–feature-payload profile and gains across aggregation backbones, while the quantitative tables report complete algorithmic traffic. Our contributions are summarized as follows.

- We propose RC-GeoCP, a unified 4D radar–camera collaborative perception framework that connects radar-grounded representation with sparse source coordination to address camera depth ambiguity and acquisition–assembly inconsistency.
- We introduce persistent source attribution as an acquisition–assembly coordination principle. GSR grounds camera semantics, while UAC and CDA share one responsibility to couple sparse acquisition with detection-supervised source weighting.
- Across two benchmarks and three aggregation backbones, RC-GeoCP improves the accuracy–payload trade-off. Matched-payload comparisons assess scoring, replicated controls support shared attribution, and held-out evaluation and stress tests assess transfer and sensitivity.

2 Related Work

2.1 Collaborative Perception

Intermediate collaboration shares features through graph or attention models such as V2VNet (Wang et al., 2020) and V2X-ViT (Xu et al., 2022a). Communication-efficient methods select confident regions (Hu et al., 2022), reconstruct omitted information from a codebook (Hu et al., 2024), match source supply with receiver demand (Xu et al., 2025), or use Doppler-guided sparse queries (Yazgan

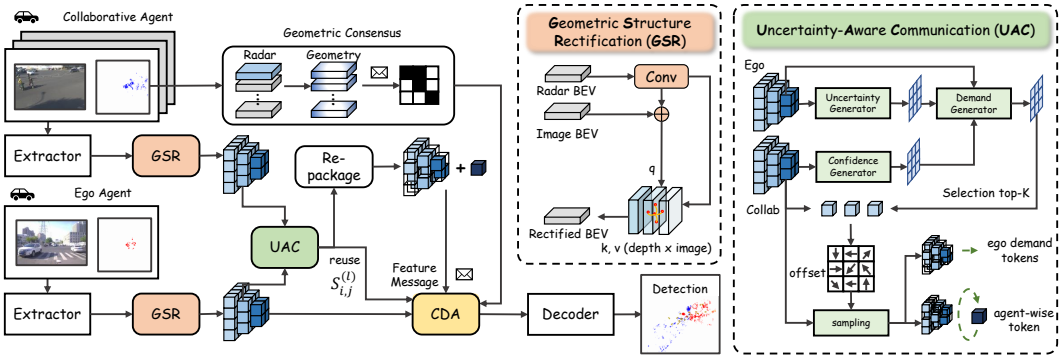

Figure 2: RC-GeoCP coordinates local grounding, acquisition, and assembly. Geometric Structure Rectification (GSR) grounds camera semantics in radar-supported spatial structure. Uncertainty-Aware Communication (UAC) uses the pre-delivery receiver–source responsibility $\mathbf{S}_{i,j}^{(l)}$ to select spatial tokens for transmission. The reuse path carries this same field into the Consensus-Driven Assembler (CDA), where it weights selected features and residual context over valid sources. Thus, the source preferences that guide acquisition also enter the assembled BEV representation.

et al., 2025). Fully sparse methods replace BEV with geometry-aware instance queries (Wang et al., 2026a). Recent work further studies rate–distortion optimization (Liu et al., 2026), temporal uncertainty (Mukhopadhyay et al., 2026), and receiver-global spatial–channel allocation (Chen et al., 2026).

CoSDH forms a binary receiver-demand–source-supply mask before max fusion. SlimComm broadcasts ego-generated Doppler and confidence queries, admits collaborators through source density, and averages their responses before deformable fusion. RC-GeoCP links these decisions through a continuous, radar-supported responsibility that determines both within-quota spatial selection and source-wise assembly. WhisperNet addresses outer global allocation, whereas RC-GeoCP fixes source quotas and coordinates acquisition with assembly within them. Matched scoring adaptations assess acquisition within this assembler, while shared-versus-independent controls test score sharing (Supplementary Tables A2– A4). Robust alignment methods such as CoAlign and TraF-Align (Lu et al., 2023; Song et al., 2025) correct spatial or temporal transforms. RC-GeoCP complements this alignment by selecting and weighting the resulting valid evidence.

2.2 Radar-based Perception

Recent 4D imaging radar advances radar–camera fusion (Zheng et al., 2026b;a; Bai et al., 2025c;b; 2026; Wu et al., 2026; Li et al., 2026; Bai et al., 2025a). CenterFusion and CRAFT associate radar returns with image frustums (Nabati & Qi, 2021; Kim et al., 2023), while RCBEVDet and RaCFormer improve single-agent BEV construction (Lin et al., 2024; Chu et al., 2025). GSR couples radar-conditioned sampling with gated calibration before collaboration, grounding the evidence used by acquisition and assembly.

Radar-based CP uses radar for robust sensing, modality-specific transmission, or sparse query placement (Yazgan et al., 2025; Huang et al., 2025; Yang et al., 2026). HRCP studies LiDAR–radar collaboration under adverse weather, whereas the camera–radar setting must address depth-diffuse visual features. RC-GeoCP grounds local camera features and carries radar-supported attribution through neighbor requests and source weighting.

3 Method

3.1 RC-GeoCP Design Overview

Problem Formulation. We consider N connected camera–radar agents. Agent i observes camera images $\mathcal{I}_i$ and 4D radar measurements $\mathcal{R}_i$, and extracts a BEV feature $\mathbf{F}_i \in \mathbb{R}^{C\times H\times W}$. In each ego-centric inference pass, neighboring messages $\{\mathbf{P}_{j\to i}\}_{j\in\mathcal{N}(i)}$ are transformed to ego agent i's

frame and aggregated for 3D detection under a total algorithmic-payload budget B_{total}. A control–request–data cycle exposes lightweight neighbor metadata before budgeted feature acquisition.

Let $\hat{\mathcal{Y}}_i = \mathcal{D}_\theta(\mathbf{F}_i^{\text{collab}})$ denote the ego prediction after GSR, UAC, and CDA. Total traffic $\mathcal{C}_i^{\text{total}} = \mathcal{C}_i^{\text{feat}} + \mathcal{C}_i^{\text{ctrl}}$ counts BEV features, control maps, and returned indices.

$$\theta^\star = \arg\min_\theta \ \mathbb{E}\left[\sum_{i=1}^{N} \mathcal{L}_{\text{det}}(\hat{\mathcal{Y}}_i, \mathcal{Y}_i)\right] \quad \text{s.t.} \quad \mathcal{C}_i^{\text{total}} \le B_{\text{total}},\ \forall i. \tag{1}$$

Equation 1 defines payload feasibility, while source–level quotas are fixed by the reported policy. After reserving $\mathcal{C}_i^{\text{ctrl}}$, UAC optimizes within-quota location ranking under the remaining feature budget B_{feat}. Neural collaborative processing is reported separately in the efficiency analysis.

Overview. Sparse assembly uses the local evidence retained by acquisition together with compressed source context. At location p, geometric consensus represents valid sources through normalized receiver-frame attribution informed by radar reliability. In Fig. 2, GSR (Sec. 3.2) first anchors visual semantics to local radar geometry, yielding $\tilde{\mathbf{F}}_i$. UAC (Sec. 3.3) then forms pre-delivery responsibility $\mathbf{S}_{i,j}^{(l)}$. Equation 8 applies it over locations to obtain discrete support, while Equation 10 applies it over sources to obtain CDA's continuous assembly distribution. These decisions share the same source preferences. Radar grounds *what* each agent represents, while persistent attribution links *where* to communicate with *which* evidence contributes after delivery.

3.2 Geometric Structure Rectification (GSR)

Camera lifting supplies dense semantic evidence whose depth uncertainty can spread responses across BEV locations. GSR grounds this evidence before cross-agent selection and aggregation. Radar-conditioned queries guide where camera features are gathered, and gated calibration adds local radar support. Together, these operations improve the spatial basis on which UAC compares candidate evidence and CDA assembles the returned sources.

We initialize the radar-grounded query field as $\mathbf{Q}_i = \mathbf{F}_i^{\text{cam}} + \Phi_{\text{init}}(\mathcal{D}(\mathbf{F}_i^{\text{rad}}))$. Here $\mathcal{D}(\cdot)$ denotes stride-matched downsampling from the radar BEV resolution to $H \times W$, and Φ_{init} is a zero-initialized 1×1 convolution that maps radar channels to the camera BEV channel C. Following 3D deformable lifting (Li et al., 2023a; Bai et al., 2025c), let $\mathbf{F}_{i,v}^{\text{frus}}$ denote the depth-weighted frustum feature of camera v. At BEV query q, GSR gathers

$$\mathbf{F}_i^{\text{rect}}(q) = \sum_{v=1}^{V}\sum_{n=1}^{K} \alpha_{n,v}(\mathbf{Q}_i(q))\, \phi\big(\mathbf{F}_{i,v}^{\text{frus}}, \mathcal{P}_{i,v}(q, z_n) + \Delta q_{n,v}(\mathbf{Q}_i(q))\big)\,, \tag{2}$$

where $\mathcal{P}_{i,v}$ projects a height-aware 3D reference point into view v, ϕ is bilinear sampling, and the normalized weight $\alpha_{n,v}$ and offset $\Delta q_{n,v}$ are both predicted from the radar-conditioned query. Thus, radar changes where camera evidence is gathered while depth-weighted image features remain the semantic carrier.

Adaptive gated calibration then incorporates radar support into the gathered semantics. The radar-conditioned query controls deformable re-sampling, while the residual branch adjusts the contribution of radar features through a learned spatial gate. The final rectified feature $\tilde{\mathbf{F}}_i$ is

$$\tilde{\mathbf{F}}_i = \mathbf{F}_i^{\text{rect}} + \sigma\left(\mathcal{G}(\mathbf{F}_i^{\text{rect}})\right) \odot \Psi(\mathbf{F}_i^{\text{rad}}), \tag{3}$$

where σ is sigmoid, $\mathcal{G}$ is a learned 1×1 spatial gate, and Ψ is stacked convolution. Equations 2–3 therefore form an ordered transformation. Sampling relocates camera semantics with radar-conditioned queries, after which calibration injects gated radar support. Together, they define the GSR operator used across the aggregation backbones in Table 1.

3.3 Uncertainty-Aware Communication (UAC)

UAC selects the local evidence available to sparse assembly under communication constraints. Each receiver-conditioned request prioritizes confident neighbor support for uncertain ego regions. In one control–request–data cycle, neighbors send lightweight confidence and radar-reliability maps, the ego returns indices, and neighbors return requested tokens with one agent-wise token. Payload

Table 1: Dense fusion-backbone isolation on V2X-Radar and V2X-R.

Agent	M	Fusion	V2X-Radar						V2X-R							
			AP@0.7 ↑			AP@0.5 ↑			Syn. AP@0.7 ↑				Asyn. AP@0.7 ↑			
			0	0–30	30–50	0	0–30	30–50	0	0–30	30–50	50–100	0	0–30	30–50	50–100
AttFuse	C	BEVDepth	8.96	16.56	3.14	29.55	51.00	9.90	31.18	43.29	26.04	23.38	22.20	36.41	16.32	11.80
	R	RadarPillarNet	8.51	14.76	3.14	25.11	38.02	15.98	46.63	66.10	44.36	28.36	41.62	59.31	41.97	22.05
	R+C	BEVFusion	11.09	20.41	3.41	33.01	55.62	16.07	51.39	70.84	51.31	34.76	42.77	62.77	36.47	27.85
	R+C	RCFusion	11.48	20.34	3.56	33.26	54.52	16.96	52.99	72.90	52.16	34.80	43.68	64.61	41.94	25.08
	R+C	LXL	11.75	21.61	2.39	33.46	55.77	16.57	52.80	72.21	52.62	34.67	43.75	63.47	40.39	26.35
	R+C	RC-GeoCP	14.36	23.85	7.63	35.02	56.25	21.91	54.19	73.29	53.11	35.28	44.12	64.99	42.53	26.01
CoAlign	C	BEVDepth	14.72	25.93	2.55	31.77	52.83	10.11	42.24	56.40	42.43	28.74	37.87	53.18	35.36	26.05
	R	RadarPillarNet	12.74	22.32	3.64	29.26	44.28	16.96	58.43	76.83	62.64	34.98	50.23	70.17	53.13	26.25
	R+C	BEVFusion	15.65	26.81	3.24	37.14	57.89	18.96	61.30	80.35	65.22	33.02	54.63	76.89	51.22	36.01
	R+C	RCFusion	16.26	28.77	4.25	37.60	57.91	19.90	60.22	78.69	64.08	40.10	56.07	79.42	55.33	35.19
	R+C	LXL	20.11	33.85	4.45	40.46	61.60	19.02	63.57	82.63	66.51	45.11	54.82	77.82	54.26	33.18
	R+C	RC-GeoCP	24.87	40.53	8.75	44.77	65.60	28.68	67.40	86.27	69.51	48.95	58.99	81.91	57.65	38.09
HEAL	C	BEVDepth	12.41	22.94	2.21	28.78	53.39	7.36	45.13	59.09	43.77	34.45	33.94	53.74	29.03	16.65
	R	RadarPillarNet	15.13	24.69	5.93	31.61	46.87	19.32	61.05	81.70	64.21	35.19	50.11	69.62	54.22	25.24
	R+C	BEVFusion	16.76	27.59	6.68	38.61	58.74	23.49	62.19	83.94	65.97	35.65	51.45	77.16	50.25	26.96
	R+C	RCFusion	17.47	29.48	6.82	38.89	56.62	25.48	62.77	82.47	68.93	38.25	57.36	80.17	58.51	32.81
	R+C	LXL	21.96	36.37	6.34	43.17	64.02	24.92	63.42	83.29	70.75	39.84	48.19	71.25	48.15	23.18
	R+C	RC-GeoCP	26.30	43.94	9.27	46.17	68.13	29.30	65.66	87.64	70.15	39.16	58.65	81.87	60.02	34.21

counts FP32 features, agent tokens, control maps, and Int32 indices. This cycle makes source preferences available to the receiver before the selected features arrive.

Given rectified features $\tilde{\mathbf{F}}_i$, UAC builds a multi-scale BEV pyramid and transmits only selected ego-demand tokens plus a compact agent-wise token for each neighbor.

Geometric responsibility for consensus seeking. UAC estimates the utility of neighbor j from control cues available before its feature message. Semantic confidence and radar support indicate usable evidence, while receiver uncertainty and learned complementarity identify demand. These cues define the responsibility evaluated under matched payload in Supplementary Table A3. The resulting field retains the receiver, source j, and spatial location indices through selection. CDA can therefore use the same attribution when weighting the delivered evidence.

Operationally, each agent predicts semantic confidence logits in its local coordinate frame. These logits are aligned to the ego frame before applying the sigmoid

$$\mathbf{C}_{j\to i}^{(l)} = \sigma\Big(\mathcal{T}_{j\to i}\Big(\Phi_{\text{conf}}^{(l)}\Big(\tilde{\mathbf{F}}_j^{(l)}\Big)\Big)\Big), \tag{4}$$

where $\mathcal{T}_{j\to i}$ warps source-frame logits into the ego frame. This one-channel message is counted as control traffic, while ego uncertainty is obtained by complementing ego confidence.

At each scale, a compact radar head follows the same logit alignment and sigmoid order, giving $\mathbf{R}_{j\to i}^{(l)} = \sigma(\mathcal{T}_{j\to i}(\Phi_{\text{rad}}^{(l)}(\mathbf{F}_j^{\text{rad}})))$. For cell p, let $\mathcal{V}_i^{(l)}(p)$ contain neighbors whose warped support is valid. The complete valid source set is $\mathcal{A}_i^{(l)}(p) = \{i\} \cup \mathcal{V}_i^{(l)}(p)$. Lightweight heads $\Phi_{\text{comp}}^{(l)}$ and $\Phi_{\text{dem}}^{(l)}$ then estimate receiver-conditioned complementarity and source demand

$$\mathbf{D}_{i,j}^{(l)} = \Phi_{\text{comp}}^{(l)}([\tilde{\mathbf{F}}_i^{(l)}, \mathbf{C}_i^{(l)}, \mathbf{U}_i^{(l)}, \mathbf{C}_{j\to i}^{(l)}]), \tag{5}$$

$$\mathbf{W}_{i,j}^{(l)} = \text{Softmax}_{j\in\mathcal{A}_i^{(l)}(p)}\Big(\Phi_{\text{dem}}^{(l)}([\tilde{\mathbf{F}}_i^{(l)}, \mathbf{D}_{i,j}^{(l)}, \mathbf{U}_i^{(l)}, \mathbf{C}_{j\to i}^{(l)}, \mathbf{R}_{j\to i}^{(l)}])\Big). \tag{6}$$

The final source score is

$$\mathbf{S}_{i,j}^{(l)} = \text{Norm}_j\Big(\mathbf{W}_{i,j}^{(l)} \odot \mathbf{C}_{j\to i}^{(l)} \odot \mathbf{R}_{j\to i}^{(l)}\Big), \tag{7}$$

We term $\mathbf{S}_{i,j}^{(l)}$ the *geometric responsibility score*. It expresses a relative preference for source j at each location of ego i by combining receiver demand, semantic usability, and radar-derived occupancy support. The complementarity head uses ego features and neighbor confidence available during the request stage, preserving the request–response protocol. Let q_j denote the unnormalized product inside Equation 7, and let Q be its sum over the valid source set. Then $\text{Norm}_j(q_j) = q_j/Q$ for a valid source and zero otherwise. This normalization provides a bounded relative responsibility for comparing locations during acquisition. Acquisition ranks only valid neighbors, whereas ego retains its dense identity path and participates in CDA.

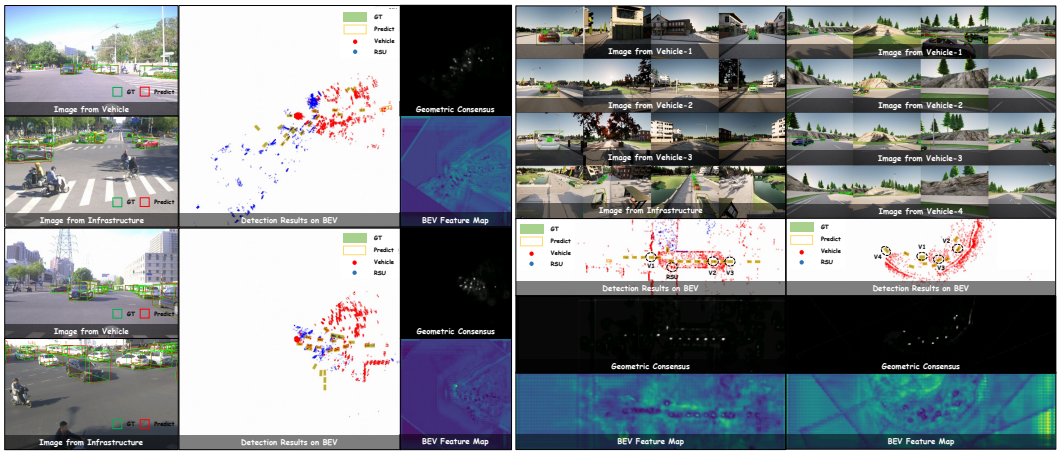

Figure 3: Qualitative results on (a) V2X-Radar and (b) V2X-R. Each scene pairs camera views with ego-frame detections, the radar-supported consensus response, and the assembled BEV feature. These aligned views relate spatial support to the resulting detections.

Budgeted acquisition. The feature budget first determines neighbor source–level quotas $\mathbf{K}_i$ satisfying $\mathcal{C}_i^{\text{feat}}(\mathbf{K}_i) \le B_{\text{feat}}$. This payload function sums channel-weighted token quotas across sources and levels and includes the agent-wise tokens, as detailed in Supplementary Sec. C. We use the fixed fine/mid/coarse allocation reported in Sec. 4 and optimize location selection conditionally within each quota. UAC selects locations by solving

$$\max_{\mathbf{\Omega}_i} \sum_{l} \sum_{j \in \mathcal{N}(i)} \sum_{p \in \Omega_{i,j}^{(l)}} \mathbf{S}_{i,j}^{(l)}(p) \quad \text{s.t.} \quad \Omega_{i,j}^{(l)} \subseteq \mathcal{V}_{i,j}^{(l)}, \qquad |\Omega_{i,j}^{(l)}| = K_{i,j}^{(l)},\ \forall l,\ j \in \mathcal{N}(i). \tag{8}$$

Here, $\mathcal{V}_{i,j}^{(l)}$ is the valid warped overlap. The reported total adds confidence and radar-reliability maps and returned indices to the feature payload. For a nonempty overlap, $K_{i,j}^{(l)}$ is the allocated fraction of the level grid capped by $|\mathcal{V}_{i,j}^{(l)}|$. An empty overlap has zero quota. Each conditional subproblem is solved by selecting the $K_{i,j}^{(l)}$ largest valid entries of $\mathbf{S}_{i,j}^{(l)}$. Conditioned on fixed feasible quotas, Top-K exactly solves Equation 8. Meanwhile, Equation 10 is the unique solution of the entropy-regularized source-weighting objective. These conditional decisions use complementary axes of $\mathbf{S}$. Top-K selects locations within each source, and softmax weights sources at each location. Sharing their input responsibility carries the selection basis into assembly under the fixed outer allocation. The derivation is in Supplementary Sec. C.

Sparse message construction and learning. Before transmission, a deformable refinement operator updates the candidate features, after which only selected tokens are retained.

$$\bar{\mathbf{F}}_j^{(l)}(p) = \mathcal{D}^{(l)}\left(\tilde{\mathbf{F}}_j^{(l)}, p\right) + \tilde{\mathbf{F}}_j^{(l)}(p) \tag{9}$$

where $\mathcal{D}^{(l)}(\cdot)$ is deformable attention with offsets predicted from local rectified features.

To retain residual context outside these sparse locations, a learnable agent-wise token $\mathbf{e}_j^{(l)}$ summarizes each source feature grid. Each neighbor sends the selected refined tokens and $\mathbf{e}_j^{(l)}$ to the ego frame. The supplementary material details its attention and broadcast implementation. Local and collaborative occupancy objectives supervise semantic confidence, receiver demand, and radar-derived support, training the responsibility used by both communication and aggregation. Supplementary Sec. D details the objectives. Figure 3 first visualizes consensus outputs on both benchmarks, and Figure 4 then diagnoses their source-specific behavior.

3.4 Consensus-Driven Assembler (CDA)

CDA carries UAC's acquisition preferences into assembly by mapping the same pre-delivery responsibility into a continuous source distribution. At cell p, its source set is $\mathcal{A}_i^{(l)}(p)$ as defined above. Ego

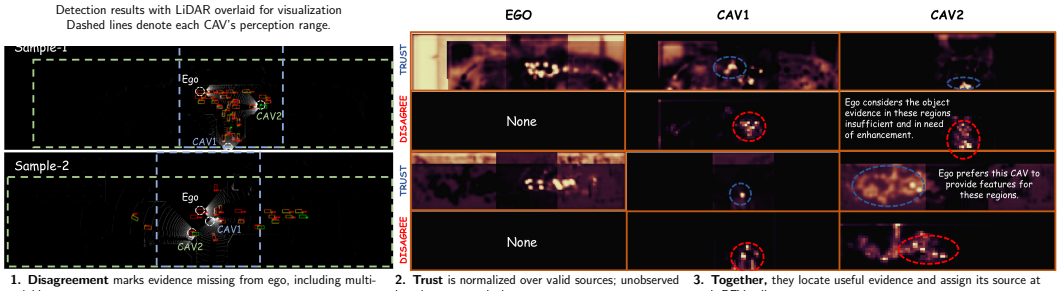

Figure 4: UAC diagnostic. Left, agent ranges contextualize evidence missing from ego. Right, disagreement marks receiver demand, while trust normalizes preference over valid sources and masks unobserved regions. Columns compare EGO, CAV1, and CAV2 at the same BEV locations. Blue and red callouts highlight trusted support and ego–source disagreement, respectively. Comparing agents at the same location reveals how receiver demand and source support shape local preferences. These source-indexed responses form the shared basis for UAC selection and CDA weighting.

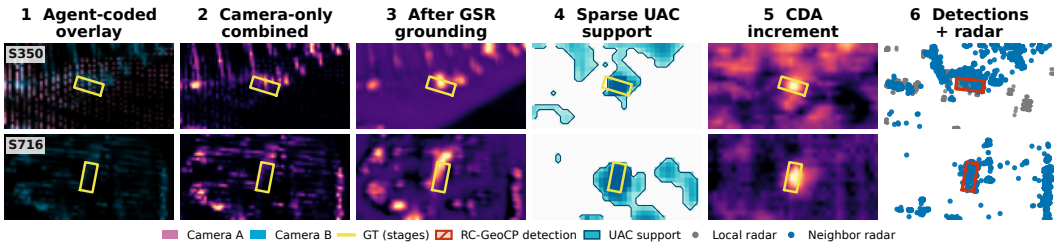

Figure 5: Two recovered cases across the RC-GeoCP stages. Columns trace camera evidence, GSR grounding, sparse UAC support, CDA's incremental response, and final radar-supported detections. Yellow/red boxes denote GTs/detections matched at a BEV IoU threshold of 0.7. Independently normalized heatmaps show spatial support at each stage.

is always valid, invalid warped sources are masked before normalization, and $\mathbf{S}_{i,i}^{(l)}$ uses the identity transform,

$$\alpha_{i,j}^{(l)}(p) = \mathrm{Softmax}_{j \in \mathcal{A}_i^{(l)}(p)}\left(\mathbf{S}_{i,j}^{(l)}(p)/\tau\right), \qquad \tau = 1. \tag{10}$$

Equation 7 provides a bounded score for cross-location ranking. Equation 10 projects it across valid sources into convex assembly weights. The source preferences used to request features therefore remain available after delivery. Since the score is bounded between zero and one with unit temperature, it changes the odds between valid sources by at most e. Valid sources retain positive weight, while invalid-source masking handles exclusion. The acquisition mask determines which spatial values arrive, whereas responsibility determines their source weights. Reusing the latter coordinates these distinct decisions.

Let $m_{i,i}^{(l)}(p)$ denote the always-active ego mask and let $m_{i,j}^{(l)}(p)$ indicate whether a valid neighbor location is requested. CDA uses the spatial feature when the corresponding mask is active and the agent-wise token for a valid, unrequested neighbor.

$$\mathbf{F}_i^{\mathrm{collab},(l)}(p) = \sum_{j \in \mathcal{A}_i^{(l)}(p)} \alpha_{i,j}^{(l)}(p)\left[m_{i,j}^{(l)}(p)\bar{\mathbf{F}}_{j\to i}^{(l)}(p) + \left(1 - m_{i,j}^{(l)}(p)\right)\mathbf{e}_j^{(l)}\right]. \tag{11}$$

Top-K membership is detached. For fixed masks, continuous CDA weights provide detection supervision to the responsibility used for acquisition. Gradients also reach selected features and agent tokens, while auxiliary losses supervise the score heads. Selected tokens retain location-specific evidence, and the agent-wise token supplies context at unrequested locations. The re-aggregated tokens are unpacked to the BEV grid. Finally, multi-scale resizing, concatenation, and convolution produce the dense BEV representation used for detection.

Figure 5 traces two cases across these stages. S350 shows cross-agent spatial dispersion, while S716 exhibits asymmetric camera support. The GSR response and selected UAC support overlap the targets, followed by CDA increments and matched detections.

Table 2: Scoring adaptations on V2X-Radar at matched total algorithmic payload.

Selector	AP@0.7	MB/link
CoSDH-style	24.91	10.88
SlimComm-style	25.17	10.88
UAC w/o comp.	21.43	10.88
Full UAC	**25.92**	10.88

Table 3: Local radar-grounding comparison under the common input/training protocol on V2X-Radar validation set.

Method	AP@0.5	AP@0.7
RCBEVDet 'CVPR24	40.61	19.29
RaCFormer 'CVPR25	41.54	21.27
GSR (ours)	**43.12**	**23.81**

4 EXPERIMENT

Evaluation protocol. We evaluate collaborative 3D detection on V2X-Radar and V2X-R using camera and 4D radar streams. We report BEV AP@0.5/AP@0.7 within each benchmark's valid sensor field of view. Supplementary Sec. B.1 specifies the data, training, and checkpoint protocol. Communication is reported in decimal MB per directed link. Total algorithmic payload covers FP32 feature values, agent tokens, control maps, and Int32 indices, using grid sizes and valid overlaps measured over each validation set. Supplementary Sec. C gives the accounting and transport-format sensitivity. Radar–camera baselines share inputs, encoders, detector, splits, sensing ranges, training schedule, and evaluation code. Every method is evaluated using its last saved checkpoint at the end of training. Within-model controls separate representation, selection, and shared source weighting. We transfer the validation-selected 40/30/20 policy unchanged to held-out V2X-Radar test. Backbone-controlled ablations use CoAlign on V2X-R and HEAL on V2X-Radar. Finally, Fig. 6 reports sensitivity to stale inputs and pose perturbations.

4.1 REPRESENTATION, COORDINATION, AND SPARSIFICATION

Representation across aggregation backbones. Table 1 examines whether radar-grounded representation and source coordination benefit different aggregation operators. Overall AP@0.7 improves with AttFuse, CoAlign, and HEAL on V2X-Radar, and with CoAlign under synchronous and asynchronous V2X-R settings. This consistency supports the use of RC-GeoCP around different fusion backbones. Relative to the corresponding BEVFusion rows, gains are larger at AP@0.7 than at AP@0.5 for all three backbones, a stricter-overlap pattern consistent with improved localization. The benefit also extends beyond the near field, with CoAlign reaching 8.75% at 30–50 m compared with 4.45% for LXL. The next comparisons examine the local grounding and selection designs that prepare this evidence for collaboration.

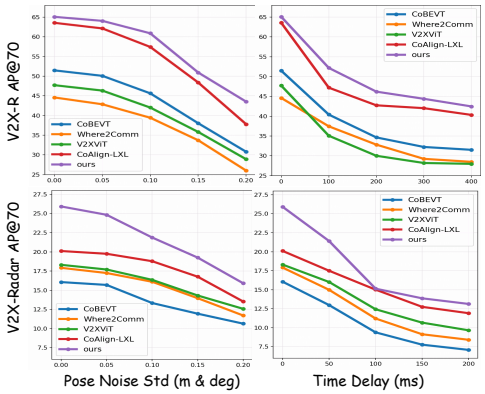

Figure 6: AP under pose error and stale neighbor inputs.

Tables 2 and 3 examine selection and local grounding under matched protocols. GSR ranks highest among the evaluated local radar–camera methods at both IoU thresholds, supporting its role in preparing spatial evidence. Full UAC also ranks first on both datasets at identical feature and total payload (Supplementary Table A3). The CoSDH-style and SlimComm-style adaptations change Top-K selection while CDA retains the UAC responsibility formulation. These comparisons favor acquisition guided by the assembly responsibility. The independent-CDA control below then tests sharing its parameterization across the two stages.

Table 4: Accuracy–payload comparison (V2X-Radar/V2X-R).

Method	Radar Val		Radar Test		V2X-R Val		Total
	@0.5	@0.7	@0.5	@0.7	@0.5	@0.7	MB/link
Who2com	27.10	10.35	17.51	4.28	77.73	45.57	16.78/9.01
AttFuse	33.01	11.09	24.78	5.75	76.76	51.39	16.78/9.01
CoAlign	37.14	15.65	29.40	12.11	77.21	61.30	29.36/15.77
CoBEVT	36.89	16.06	31.90	10.73	78.71	51.48	16.78/9.01
HEAL	38.61	16.76	27.80	10.10	79.54	62.19	29.36/15.77
Where2comm	40.83	17.92	33.19	11.10	75.64	44.60	16.78/9.01
V2XViT	39.88	18.31	34.81	12.35	78.21	47.72	16.78/9.01
RC-GeoCP	44.55	25.92	42.61	18.77	81.90	65.09	**10.88/4.95**

Test-set generalization and cross-method comparison. Selected on validation, the sparse policy reaches 18.77 AP@0.7 on held-out V2X-Radar test. Keeping its quota policy fixed tests the selected

sparse operating point on unseen sequences. V2X-R results cover validation, while Table 4 compares accuracy and payload across methods. Figs. 3–5 trace the same contribution chain from grounded local evidence to source selection and assembled detections. The visualizations locate support at each stage, while recovering 112/453 ego misses connects collaboration to objects absent from ego detections. Controlled RGB degradation and pose/timing analyses then assess sensitivity as visual evidence or cross-agent correspondence deteriorates (Supplementary Table A11 and Fig. 6).

4.2 Ablation Study

Grounding and coordination. Table 5 connects local grounding to collaborative assembly. Baseline, +GSR, and dense GSR+CDA exchange the same three-level feature pyramid. GSR adds 7.05 V2X-Radar AP@0.7 without increasing payload. Within GSR, radar-guided sampling adds 0.51/1.07 AP over calibration alone (Supplementary Table A5), showing the additional value of correcting where camera features are gathered. Supplementary Table A4 then examines how this evidence is coordinated. The independent-CDA control matches inputs, head capacity, supervision, and UAC acquisition, but learns separate assembly scores. Sharing the responsibility improves three-seed mean AP@0.7 by 2.72/4.84 on V2X-R/V2X-Radar. This contrast supports coupling acquisition and assembly through a common responsibility that shares source preferences and detection supervision. The factorial controls further favor radar reliability in both stages over either single-stage variant, with positive cell-mean interactions describing their joint response.

Table 5: AP@0.7 ablation (V2X-R/V2X-Radar).

Variant	MB/link	V2X-R	V2X-Radar
Baseline	15.77/29.36	61.30	16.76
+ GSR	15.77/29.36	64.99	23.81
+ GSR + CDA, dense	15.77/29.36	67.40	26.30
Full w/o radar control	4.77/10.54	63.10	22.07
Full RC-GeoCP	4.95/10.88	**65.09**	**25.92**

Table 6: FP32 payload sensitivity (V2X-R/V2X-Radar).

Policy	MB/link	V2X-R ↑		V2X-Radar ↑	
		@0.7	@0.5	@0.7	@0.5
U-25%	4.15/8.12	60.95	78.34	25.88	44.50
40/30/20	4.95/10.88	65.09	81.90	25.92	44.55
U-50%	6.01/15.54	62.88	79.83	25.93	44.54
Dense	15.77/29.36	67.40	79.85	26.30	46.17

Table 7: Neural cost and FP32 payload (V2X-R/V2X-Radar).

Variant	Params	C-FLOPs	ms	MB/link
Baseline	59.8	0.1	1.0	15.77/29.36
+GSR	62.1	0.1	0.9	15.77/29.36
Dense GSR+CDA	63.0	262.2	78.7	15.77/29.36
RC-GeoCP	63.8	102.8	30.8	4.95/10.88
LXL-V2VNet	75.4	982.1	165.7	9.01/16.78

Algorithmic payload and neural compute. RC-GeoCP trades additional collaborative computation for lower inter-agent traffic. UAC cuts payload from 15.77/29.36 to 4.95/10.88 MB/link while retaining 96.6%/98.6% of dense AP@0.7 on V2X-R/V2X-Radar (Table 6), and reduces collaborative processing from 262.2 GFLOPs/78.7 ms to 102.8 GFLOPs/30.8 ms (Table 7). Across sparse policies, V2X-Radar is nearly saturated, whereas V2X-R favors 40/30/20 over uniform ratios. The outer allocation therefore affects the operating point alongside within-quota ranking. With 0.8M more parameters than dense GSR+CDA, RC-GeoCP lowers compute by processing fewer collaborative locations. Under FP16 wire-format accounting, the fixed-quota message occupies 2.50/5.51 MB/link. Reported AP uses FP32 inference. Timing covers collaborative neural processing, excluding local GSR and network exchange. Supplementary Sec. C specifies transport accounting and latency scope.

5 Conclusions

In this work, we present RC-GeoCP for 4D radar–camera collaborative perception. It links radar-grounded BEV construction to persistent receiver–source attribution across sparse acquisition and assembly. GSR corrects local camera-feature dispersion, while one pre-delivery responsibility induces UAC's discrete support and CDA's continuous source mixture. This formulation treats spatial selection and source weighting as complementary decisions over one field, carrying acquisition preferences into the fusion of evidence. Radar geometry informs both representation and cross-agent coordination. Across two benchmarks, RC-GeoCP improves the accuracy–payload trade-off and transfers its selected policy to held-out V2X-Radar test data. Replicated controls support shared attribution, while stress tests assess its operating conditions. Future work will study adaptive global allocation and extend radar-grounded collaborative representations to vision–language–action models for cooperative driving.

AI USE STATEMENT

Generative AI tools were used for literature organization, language editing, and conversion of the manuscript to the ICLR template. The authors verified all citations, technical descriptions, analyses, and revisions, and take responsibility for the final content of this work.

REFERENCES

Xiaokai Bai, Jiahao Cheng, Songkai Wang, Yixuan Luo, Lianqing Zheng, Xiaohan Zhang, Si-Yuan Cao, and Hui-Liang Shen. SD4R: Sparse-to-Dense Learning for 3D Object Detection with 4D Radar. In *IEEE International Conference on Intelligent Transportation Systems*, pp. 4362–4368, 2025a. doi: 10.1109/ITSC60802.2025.11423715.

Xiaokai Bai, Qin Yang, Zili Zhou, Fuyi Zhang, Zhe Wu, Si-Yuan Cao, Lianqing Zheng, Beinan Yu, Fang Wang, Jie Bai, and Hui-Liang Shen. LGDD: Local-Global Synergistic Dual-Branch 3D Object Detection Using 4D Radar. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 13318–13325. IEEE, 2025b. doi: 10.1109/IROS60139.2025.11246504.

Xiaokai Bai, Zhu Yu, Lianqing Zheng, Xiaohan Zhang, Zili Zhou, Xue Zhang, Fang Wang, Jie Bai, and Hui-Liang Shen. SGDet3D: Semantics and Geometry Fusion for 3D Object Detection Using 4D Radar and Camera. *IEEE Robotics and Automation Letters*, 10(1):828–835, 2025c. doi: 10.1109/LRA.2024.3513041.

Xiaokai Bai, Chenxu Zhou, Lianqing Zheng, Jianan Liu, Si-Yuan Cao, Xiaohan Zhang, Yiming Li, Zhengzhuang Zhang, and Hui-Liang Shen. RaGS: Unleashing 3D Gaussian Splatting from 4D Radar and Monocular Cue for 3D Object Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4983–4992, 2026.

Gong Chen, Chaokun Zhang, and Xinyan Zhao. WhisperNet: A Scalable Solution for Bandwidth-Efficient Collaboration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32154–32163, 2026.

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 *ACM/IEEE Symposium on Edge Computing*, pp. 88–100, 2019. doi: 10.1145/3318216.3363300.

Xiaomeng Chu, Jiajun Deng, Guoliang You, Yifan Duan, Houqiang Li, and Yanyong Zhang. Racformer: Towards high-quality 3d object detection via query-based radar-camera fusion. In *IEEE/CVF conference on Computer Vision and Pattern Recognition Conference*, pp. 17081–17091, 2025.

Mariella Dreissig, Dominik Scheuble, Florian Piewak, and Joschka Boedecker. Survey on LiDAR Perception in Adverse Weather Conditions. In *IEEE Intelligent Vehicles Symposium*, pp. 1–8, 2023.

Lili Fan, Junhao Wang, Yuanmeng Chang, Yuke Li, Yutong Wang, and Dongpu Cao. 4D mmWave Radar for Autonomous Driving Perception: A Comprehensive Survey. *IEEE Transactions on Intelligent Vehicles*, 9(4):4606–4620, 2024. doi: 10.1109/TIV.2024.3380244.

Jiaming Gu, Jingyu Zhang, Muyang Zhang, Weiliang Meng, Shibiao Xu, Jiguang Zhang, and Xiaopeng Zhang. FeaCo: Reaching Robust Feature-Level Consensus in Noisy Pose Conditions. In *ACM International Conference on Multimedia*, pp. 3628–3636, 2023.

Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-Efficient Collaborative Perception via Spatial Confidence Maps. *Advances in neural information processing systems*, 35:4874–4886, 2022.

Yue Hu, Juntong Peng, Sifei Liu, Junhao Ge, Si Liu, and Siheng Chen. Communication-Efficient Collaborative Perception via Information Filling with Codebook. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15481–15490, 2024.

Xun Huang, Jinlong Wang, Qiming Xia, Siheng Chen, Bisheng Yang, Xin Li, Cheng Wang, and Chenglu Wen. V2X-R: Cooperative LiDAR-4D Radar Fusion with Denoising Diffusion for 3D Object Detection. In *IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 27390–27400, 2025.

Youngseok Kim, Sanmin Kim, Jun Won Choi, and Dongsuk Kum. CRAFT: Camera-Radar 3D Object Detection with Spatio-Contextual Fusion Transformer. In *AAAI Conference on Artificial Intelligence*, volume 37, pp. 1160–1168, 2023.

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*, pp. 12697–12705, 2019.

Hongyang Li, Hao Zhang, Zhaoyang Zeng, Shilong Liu, Feng Li, Tianhe Ren, and Lei Zhang. DFA3D: 3D Deformable Attention For 2D-to-3D Feature Lifting. In *IEEE/CVF International Conference on Computer Vision*, pp. 6684–6693, 2023a.

Xiaozhi Li, Huijun Di, Jian Li, Feng Liu, and Wei Liang. M3Detection: Multi-Frame Multi-Level Feature Fusion for Multi-Modal 3D Object Detection with Camera and 4D Imaging Radar. *IEEE Transactions on Intelligent Transportation Systems*, 2026. doi: 10.1109/TITS.2026.3695618.

Yiming Li, Shunli Ren, Pengxiang Wu, Siheng Chen, Chen Feng, and Wenjun Zhang. Learning Distilled Collaboration Graph for Multi-Agent Perception. *Advances in Neural Information Processing Systems*, 34:29541–29552, 2021.

Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. BEVDepth: Acquisition of Reliable Depth for Multi-View 3D Object Detection. In *AAAI Conference on Artificial Intelligence*, pp. 1477–1485, 2023b.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. BEVFormer: Learning Bird's-Eye-View Representation from Multi-camera Images via Spatiotemporal Transformers. In *European Conference on Computer Vision*, pp. 1–18, 2022.

Zhiwei Lin, Zhe Liu, Zhongyu Xia, Xinhao Wang, Yongtao Wang, Shengxiang Qi, Yang Dong, Nan Dong, Le Zhang, and Ce Zhu. RCBEVDet: Radar-camera Fusion in Bird's Eye View for 3D Object Detection. In *IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 14928–14937, 2024.

Genjia Liu, Anning Hu, Yue Hu, Wenjun Zhang, and Siheng Chen. Rate-Distortion Optimized Pragmatic Communication for Collaborative Perception. In *International Conference on Learning Representations*, 2026.

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*, pp. 4106–4115, 2020a.

Yen-Cheng Liu, Junjiao Tian, Chih-Yao Ma, Nathan Glaser, Chia-Wen Kuo, and Zsolt Kira. Who2com: Collaborative Perception via Learnable Handshake Communication. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6876–6883. IEEE, 2020b.

Yifan Lu, Quanhao Li, Baoan Liu, Mehrdad Dianati, Chen Feng, Siheng Chen, and Yanfeng Wang. Robust Collaborative 3D Object Detection in Presence of Pose Errors. In *IEEE International Conference on Robotics and Automation*, pp. 4812–4818. IEEE, 2023.

Abdul Sajeed Mohammed, Ali Amamou, Follivi Kloutse Ayevide, Sousso Kelouwani, Kodjo Agbossou, and Nadjet Zioui. The Perception System of Intelligent Ground Vehicles in All Weather Conditions: A Systematic Literature Review. *Sensors*, 20(22):6532, 2020.

Shilpa Mukhopadhyay, Amit Roy-Chowdhury, and Hang Qiu. COOPERTRIM: Adaptive Data Selection for Uncertainty-Aware Cooperative Perception. In *International Conference on Learning Representations*, 2026.

Ramin Nabati and Hairong Qi. CenterFusion: Center-based Radar and Camera Fusion for 3D Object Detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1527–1536, 2021.

Jinhyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris Kitani, Masayoshi Tomizuka, and Wei Zhan. Time Will Tell: New Outlooks and A Baseline for Temporal Multi-View 3D Object Detection. In *International Conference on Learning Representations*, 2023.

Jonah Philion and Sanja Fidler. Lift, Splat, Shoot: Encoding Images from Arbitrary Camera Rigs by Implicitly Unprojecting to 3D. In *European Conference on Computer Vision*, pp. 194–210, 2020.

Nicolas Scheiner, Florian Kraus, Nils Appenrodt, Jürgen Dickmann, and Bernhard Sick. Object Detection for Automotive Radar Point Clouds – A Comparison. *AI Perspectives*, 3(1):6, 2021.

Dereje Shenkut and Vijayakumar Bhagavatula. FocalComm: Hard Instance-Aware Multi-Agent Perception. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6277–6286, 2026. doi: 10.1109/WACV61042.2026.00607.

Zhiying Song, Lei Yang, Fuxi Wen, and Jun Li. TraF-Align: Trajectory-aware Feature Alignment for Asynchronous Multi-agent Perception. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12048–12057, 2025.

Mingxing Tan and Quoc Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *International Conference on Machine Learning*, pp. 6105–6114. PMLR, 2019.

Zongheng Tang, Yi Liu, Yifan Sun, Yulu Gao, Jinyu Chen, Runsheng Xu, and Si Liu. CoST: Efficient Collaborative Perception From Unified Spatiotemporal Perspective. In *IEEE/CVF International Conference on Computer Vision*, pp. 1120–1129, 2025.

Jiahao Wang, Zhongwei Jiang, Wenchao Sun, Jiaru Zhong, Haibao Yu, Yuner Zhang, Chenyang Lu, Chuang Zhang, Lei He, Shaobing Xu, and Jianqiang Wang. SparseCoop: Cooperative Perception with Kinematic-Grounded Queries. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(12):9876–9884, 2026a. doi: 10.1609/aaai.v40i12.37952.

Jiahao Wang, Zikun Xu, Yuner Zhang, Zhongwei Jiang, Chenyang Lu, Shuocheng Yang, Yuxuan Wang, Jiaru Zhong, Chuang Zhang, Shaobing Xu, and Jianqiang Wang. Long-SCOPE: Fully Sparse Long-Range Cooperative 3D Perception. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11599–11609, 2026b.

Tianhang Wang, Fan Lu, Zehan Zheng, Guang Chen, and Changjun Jiang. RCDN: Towards Robust Camera-Insensitivity Collaborative Perception via Dynamic Feature-based 3D Neural Modeling. *Advances in Neural Information Processing Systems*, 37:22350–22369, 2024a.

Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyuan Zeng, and Raquel Urtasun. V2VNet: Vehicle-to-Vehicle Communication for Joint Perception and Prediction. In *European conference on computer vision*, pp. 605–621. Springer, 2020.

Yuqi Wang, Yuntao Chen, Xingyu Liao, Lue Fan, and Zhaoxiang Zhang. Panoocc: Unified occupancy representation for camera-based 3d panoptic segmentation. In *IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 17158–17168, 2024b.

Yuzhi Wu, Li Xiao, Jun Liu, Guangfeng Jiang, and Xiang-Gen Xia. MLF-4DRCNet: Multi-Level Fusion with 4D Radar and Camera for 3D Object Detection in Autonomous Driving. *Information Fusion*, 134:104377, 2026. doi: 10.1016/j.inffus.2026.104377.

Hao Xiang, Runsheng Xu, and Jiaqi Ma. HM-ViT: Hetero-modal Vehicle-to-Vehicle Cooperative Perception with Vision Transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 284–295, 2023.

Weiyi Xiong, Jianan Liu, Tao Huang, Qing-Long Han, Yuxuan Xia, and Bing Zhu. LXL: LiDAR Excluded Lean 3D Object Detection with 4D Imaging Radar and Camera Fusion. *IEEE Transactions on Intelligent Vehicles*, 9(1):79–92, 2024. doi: 10.1109/TIV.2023.3321240.

Junhao Xu, Yanan Zhang, Zhi Cai, and Di Huang. CoSDH: Communication-Efficient Collaborative Perception via Supply-Demand Awareness and Intermediate-Late Hybridization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Conference*, pp. 6834–6843, 2025.

Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2X-ViT: Vehicle-to-Everything Cooperative Perception with Vision Transformer. In *European conference on computer vision*, pp. 107–124. Springer, 2022a.

Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi Ma. OPV2V: An Open Benchmark Dataset and Fusion Pipeline for Perception with Vehicle-to-Vehicle Communication. In *International Conference on Robotics and Automation*, pp. 2583–2589. IEEE, 2022b.

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 *Proceedings of the 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 989–1000. PMLR, 2023a.

Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, Hongkai Yu, Bolei Zhou, and Jiaqi Ma. V2V4Real: A Real-World Large-Scale Dataset for Vehicle-to-Vehicle Cooperative Perception. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13712–13722, 2023b. doi: 10.1109/CVPR52729.2023.01318.

Kun Yang, Dingkang Yang, Ke Li, Dongling Xiao, Zedian Shao, Peng Sun, and Liang Song. Align before Collaborate: Mitigating Feature Misalignment for Robust Multi-Agent Perception. In *European Conference on Computer Vision*, pp. 282–299. Springer, 2024.

Lei Yang, Xinyu Zhang, Jun Li, Chen Wang, Jiaqi Ma, Zhiying Song, Tong Zhao, Ziying Song, Li Wang, Mo Zhou, Yang Shen, Kai Wu, and Chen Lv. V2X-Radar: A Multi-modal Dataset with 4D Radar for Cooperative Perception. *Advances in Neural Information Processing Systems*, 38, 2025. doi: 10.52202/085713-3789.

Yuquan Yang, Hui Zhang, Wenyu Lu, Ziyin Zhang, Chuanming Zhang, and Xiaohua Xu. Hybrid Robust Collaborative Perception with LiDAR-4D Radar Fusion under Adverse Weather Conditions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24237–24247, 2026.

Melih Yazgan, Qiyuan Wu, Iramm Hamdard, Shiqi Li, and J. Marius Zöllner. SlimComm: Doppler-Guided Sparse Queries for Bandwidth-Efficient Cooperative 3-D Perception. In *IEEE/CVF International Conference on Computer Vision Workshops*, pp. 1803–1812, 2025. doi: 10.1109/ICCVW69036.2025.00190.

Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, and Zaiqing Nie. DAIR-V2X: A Large-Scale Dataset for Vehicle-Infrastructure Cooperative 3D Object Detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21361–21370, 2022. doi: 10.1109/CVPR52688.2022.02067.

Hongkun Zhang, Yan Wu, and Zhengbin Zhang. What2Keep: A Communication-Efficient Collaborative Perception Framework for 3D Detection via Keeping Valuable Information. *Computer Vision and Image Understanding*, 263:104572, 2026. doi: 10.1016/j.cviu.2025.104572.

Lianqing Zheng, Jianan Liu, Runwei Guan, Long Yang, Shouyi Lu, Yuanzhe Li, Xiaokai Bai, Jie Bai, Zhixiong Ma, Hui-Liang Shen, and Xichan Zhu. Doracamom: Joint 3D Detection and Occupancy Prediction with Multi-view 4D Radars and Cameras for Omnidirectional Perception. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026a. doi: 10.1109/TCSVT.2026.3657111.

Lianqing Zheng, Long Yang, Qunshu Lin, Wenjin Ai, Minghao Liu, Shouyi Lu, Jianan Liu, Hongze Ren, Jingyue Mo, Xiaokai Bai, Jie Bai, Zhixiong Ma, and Xichan Zhu. OmniHD-Scenes: A Next-Generation Multimodal Dataset for Autonomous Driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(6):7032–7049, 2026b. doi: 10.1109/TPAMI.2026.3663672.

Zewei Zhou, Hao Xiang, Zhaoliang Zheng, Seth Z. Zhao, Mingyue Lei, Yun Zhang, Tianhui Cai, Xinyi Liu, Johnson Liu, Maheswari Bajji, Xin Xia, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. V2XPnP: Vehicle-to-Everything Spatio-Temporal Fusion for Multi-Agent Perception and Prediction. In *IEEE/CVF International Conference on Computer Vision*, pp. 25399–25409, 2025. doi: 10.1109/ICCV51701.2025.02356.

APPENDIX

A OVERVIEW

This supplementary material provides additional details and analyses to complement the main paper. Section B defines the evaluation metrics and protocol. Section C details GSR, UAC, and CDA, including related-method positioning and controls for shared attribution. Section D formulates the detection, depth, and uncertainty-aware communication (UAC) objectives. Section E describes the pretraining strategy, token warm-up schedule, and implementation settings.

Section F reports RGB stress testing, timestamp asynchrony, and pose perturbation evaluations, together with fine-grained TP/FP/FN statistics across distance ranges. Section G presents additional qualitative results on V2X-Radar and V2X-R, relating camera observations and cross-agent spatial support to the assembled features and final detections.

B EVALUATION METRICS

We evaluate predicted 3D boxes using intersection-over-union (IoU) of their oriented BEV footprints. Within each frame, predictions are greedily matched to ground truth in descending confidence order at threshold t. Average precision (AP) integrates the precision–recall envelope over all predictions. AP@0.5 and AP@0.7 use IoU thresholds 0.5 and 0.7, respectively.

B.1 DATASETS AND EVALUATION PROTOCOL

We evaluate on two public V2X benchmarks with camera and 4D radar streams. V2X-Radar provides real vehicle-infrastructure radar-camera scenes and splits its cooperative subset into 30/5/5 train/validation/test sequences. We configure the camera and radar streams in V2X-R across its 8,084/829/3,166 train/validation/test frames for radar-camera CP.

Evaluation follows the physical sensing coverage of each benchmark. Specifically, V2X-R uses the camera-field-of-view BEV range, whereas V2X-Radar uses its valid radar sensing range. We use RadarPillarNet for radar encoding and EfficientNet for image encoding. Further implementation and optimization settings are provided in Section E.

Common input and training protocol. We re-train baselines with the sensing modalities listed in the tables. Radar–camera comparisons share sensor inputs, encoder outputs, detector, data split, sensing range, optimization schedule, and checkpoint save/evaluation frequency with RC-GeoCP. Camera-only and radar-only rows retain their specified input stream. This setting compares local fusion, communication, and collaborative aggregation under the stated controls. Backbone-controlled ablations use CoAlign on V2X-R and HEAL on V2X-Radar. For every method, cross-method results use the last saved checkpoint at the end of the scheduled training run. This common selection rule complements the shared input and training protocol. Within-model controls examine representation and coordination, while held-out evaluation tests transfer of the fixed sparse policy.

V2X-Radar includes real weather and illumination variation but no paired RGB-degradation protocol for its cooperative split, while V2X-R applies its released fog and snow simulation to LiDAR rather than RGB (Yang et al., 2025; Huang et al., 2025). Deterministic RGB corruptions therefore isolate sensitivity to visual degradation while radar inputs, geometry, and labels remain fixed.

C ADDITIONAL METHOD DETAILS

The main paper defines the shared responsibility and the modules that construct and use it. This section details their sampling, masking, and multi-scale assembly operations, following the notation and inference order of Sections 3.2–3.4.

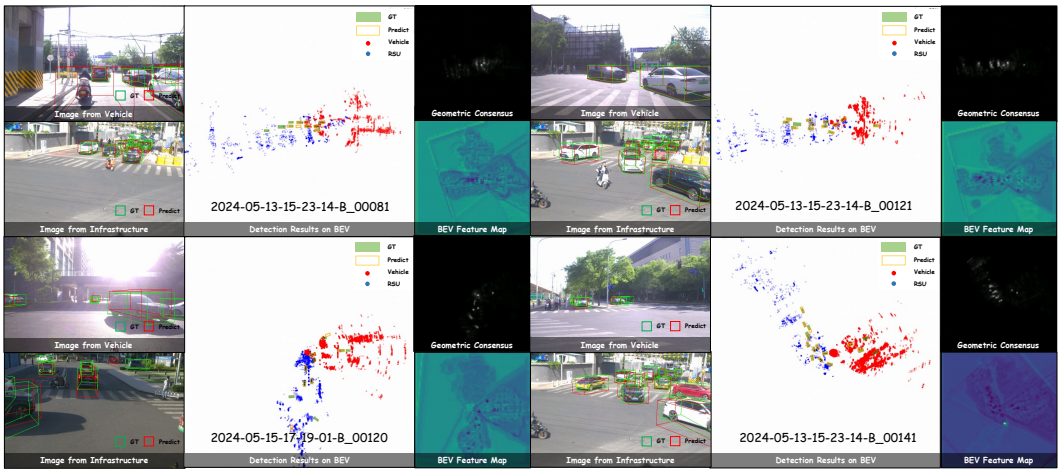

Figure A1: V2X-Radar examples pairing vehicle/infrastructure views with BEV detections, consensus responses, and assembled features (Yang et al., 2025).

C.1 GSR Sampling Operator

For camera view v, the image backbone produces $\mathbf{F}_{i,v}^{\text{img}} \in \mathbb{R}^{C\times H_I\times W_I}$ and a categorical depth volume $\hat{\mathbf{D}}_{i,v}$. Their pointwise product defines the candidate frustum feature $\mathbf{F}_{i,v}^{\text{frus}}(d,u) = \hat{\mathbf{D}}_{i,v}(d,u)\mathbf{F}_{i,v}^{\text{img}}(u)$. This lifting step provides the semantic carrier. GSR then changes how it is gathered into BEV.

At BEV location q, the radar-conditioned query $\mathbf{Q}_i(q)$ predicts the offsets and normalized weights in Equation 2 for K vertical reference points. The stride-matching operator $\mathcal{D}$ from the main paper aligns radar and camera BEV resolution before channel projection and query formation. Offsets and weights are predicted from that radar-conditioned query, normalized over the sampled points, and shared across feature channels. Our configured instantiation uses one cross-attention layer, one image feature level, and eight sampling points per BEV query. The sampled response is reshaped to the BEV grid, concatenated with the original lifted camera BEV, and projected before the gated radar residual is applied.

Two initialization details keep this correction conservative. The 1×1 radar-to-query projection is zero-initialized, so training begins from the camera query, and the spatial-gate bias is initialized to -2, corresponding to an initial radar residual weight of about 0.12. Radar features are bilinearly resized only when their BEV stride differs. This initialization starts from the camera query and allows radar conditioning to develop during training. Table 3 evaluates sampling and gated calibration together as the GSR operator defined above.

C.2 UAC Mask and Payload Construction

UAC operates on a three-level BEV pyramid with feature dimensions 64, 128, and 256. At each level, a 1×1 head predicts occupancy logits in the agent's native frame. Features, confidence logits, and radar-reliability logits share the same affine warp. The sigmoid is applied to the aligned logits. A shared validity mask excludes padded cells, whose logits are set to -5, from selection and fusion.

The demand head combines ego uncertainty, neighbor confidence, learned complementarity, and neighbor radar reliability. Scores are normalized over the valid source set $\mathcal{A}_i^{(l)}(p) = \{i\} \cup \mathcal{V}_i^{(l)}(p)$ at each BEV cell, where $\mathcal{V}_i^{(l)}(p)$ contains valid neighbors. If q_j is the product of these terms, the responsibility is $q_j / \sum_{k\in\mathcal{A}_i^{(l)}(p)} q_k$ for valid sources and zero otherwise. Detached Top-K ranking within each valid neighbor overlap forms $\Omega_{i,j}^{(l)}$, while the ego keeps a dense identity path into CDA. Detection gradients pass through selected values, agent-wise tokens, and continuous CDA weights, while auxiliary losses train the score heads and discrete indices receive no gradient.

For source–level quotas $\mathbf{K}_i$, the global feature-budget constraint is

$$\mathcal{C}_i^{\text{feat}}(\mathbf{K}_i) = \sum_l \sum_{j \in \mathcal{N}(i)} c_l K_{i,j}^{(l)} + \mathcal{C}_{\text{agent}}^{\text{feat}} \le B_{\text{feat}}, \tag{12}$$

where c_l is the payload of one level-l feature token. Conditional Top-K selection is performed only after a feasible quota allocation has been fixed.

The one-cycle protocol sends confidence and radar-reliability controls, returns selected indices, and then sends requested features. At the 40%/30%/20% fine/mid/coarse operating point, features and agent tokens use 4.525/10.068 MB per directed link on V2X-R/V2X-Radar. Control maps and Int32 indices add 0.426/0.816 MB, for FP32 totals of 4.95/10.88 MB. Dense three-level exchange uses 15.77/29.36 MB, the common payload of Baseline, +GSR, and dense GSR+CDA in Tables 5 and 7, and of CoAlign and HEAL in Table 4. Single-level exchange instead uses 9.01/16.78 MB. The sparse policy reduces these three-level totals by 68.6%/62.9% and the single-level totals by 45.1%/35.1%. Values use measured grid sizes and valid overlaps over each validation set. Removing radar controls reduces control traffic while retaining the same feature quota, giving 4.77/10.54 MB for Full w/o radar control.

Casting floating-point messages to FP16 while retaining Int32 indices gives 2.50/5.51 MB without changing selection. FP32 totals per ego frame are 13.21/10.88 MB because V2X-R averages 2.67 neighbors versus one on V2X-Radar. Table 7 separately reports collaborative GFLOPs and RTX A6000 neural latency, excluding network goodput, scheduling, contention, and protocol round trips. Table A1 compares separately trained variants under sparse training and dense evaluation. During training, the learned token summarizes each source grid with an agent- and level-specific four-head query and supplies unselected valid locations. Dense testing uses spatial features at every valid location, making the fallback term in Equation 11 inactive. The token rows therefore compare the model endpoints learned with different forms of training-time context. Removing radar reliability lowers AP@0.7 by 1.75/1.33 on V2X-R/V2X-Radar, while removing ego demand lowers it by 3.21/1.56. Table A4 complements these dense endpoints with replicated sparse controls.

Table A1: Sparse-training variants evaluated with dense exchange (point estimates).

Variant	V2X-R ↑		V2X-Radar ↑	
	@0.7	@0.5	@0.7	@0.5
No residual	65.90	78.96	24.78	44.08
Learned token	67.40	79.85	26.30	46.17
Mean pooling	67.01	80.12	25.92	44.55
No radar reliability	65.65	79.81	24.97	44.38
Full source control	67.40	79.85	26.30	46.17
No ego demand	64.19	76.20	24.74	43.40

C.3 CDA Prior Construction

Section 3.4 already defines the CDA source weighting and sparse-to-dense assembly. Here we specify how their priors enter the tensor path. At every level, the 64-channel radar BEV is resized to the feature grid and passed through a compact 64–32–16–1 reliability head. Reliability follows the same logit-warp–sigmoid order as semantic confidence. The shared UAC validity mask excludes out-of-view sources from acquisition and assembly.

For each BEV cell, the learned source demand is first normalized across agents. It is then modulated by aligned semantic confidence and radar reliability and normalized across valid sources. The resulting bounded score ranks Top-K acquisition and enters CDA as a source-wise logit prior. It thereby induces discrete spatial support along the location axis and a continuous source distribution along the agent axis. Ego-conditioned demand identifies complementary content, the confidence map indicates usable semantic evidence, and the radar head estimates occupancy support from radar features. Invalid sources are masked before the final source-wise softmax. Valid sources retain positive weights, with their relative contributions set by the shared responsibility.

Conditional decisions under fixed geometric responsibility. The global feature budget constrains the source–level quota allocation, which is fixed to the reported sparse policy in our experiments. Conditional on one such feasible allocation, fix a pyramid level, a receiver–source pair, and a quota of K locations. If a selected set contains p_a while an unselected location p_b satisfies $\mathbf{S}(p_b) > \mathbf{S}(p_a)$, exchanging the two strictly increases the acquisition objective in Equation 8. Repeating this exchange yields the Top-K set, the exact optimizer for this conditional location-selection problem. The reported policy fixes the outer allocation across sources and levels before conditional selection.

Table A2: Positioning of persistent source attribution against related communication principles.

Method	Selection principle	Radar role	Geometric-consensus function		
			Local grounding	Receiver seeking	Acquire–assemble
CoSDH	Supply–demand matching	None	–	✓	–
SlimComm	Doppler-guided queries	Query cue	–	✓	–
WhisperNet	Global spatial–channel allocation	None	–	✓	–
RDcomm	Task MI and rate–distortion	None	–	✓	–
SparseCoop	Kinematic instance queries	None	–	✓	–
Long-SCOPE	Geometry-guided long-range queries	None	–	✓	–
HRCP	Modality-specific transmission	Reliability gate	✓	–	–
RC-GeoCP	Ego demand + geometric reliability	Control criterion	✓	✓	✓

For assembly at location p, consider

$$\max_{\boldsymbol{\alpha}\in\Delta} \sum_{j\in\mathcal{A}(p)} \alpha_j \mathbf{S}_j(p) + \tau\mathcal{H}(\boldsymbol{\alpha}), \qquad \tau = 1, \tag{13}$$

where $\mathcal{A}(p)$ is the complete valid source set including ego, Δ is its probability simplex, and $\mathcal{H}$ is Shannon entropy. The stationarity condition gives $\log\alpha_j = \mathbf{S}_j/\tau + \text{const}$, hence $\boldsymbol{\alpha} = \mathrm{Softmax}_j(\mathbf{S}/\tau)$. Strict concavity of entropy makes this solution unique. With unit temperature and responsibility normalized between zero and one, responsibility-induced odds between valid sources are bounded by e. The score therefore supplies a bounded source preference, while invalid-source masking implements exclusion. Under fixed scores and quotas, acquisition and assembly are conditional decisions over the spatial and source axes of the same field. Their common input retains the acquisition basis when the selected evidence is assembled.

The selected spatial features and broadcast agent-wise fallback are weighted by these source-specific scores and summed over agents. The three fused pyramid outputs are then upsampled to a common resolution, concatenated, and processed by the final convolutional block before detection. This section specifies the tensor construction, valid-region handling, and initialization that instantiate the method equations in Sections 3.2–3.4.

C.4 Positioning and Coordination Evidence

Table A2 separates three established directions (Xu et al., 2025; Yazgan et al., 2025; Chen et al., 2026; Liu et al., 2026; Wang et al., 2026a;b; Yang et al., 2026). They are receiver-conditioned or information-efficient selection, geometry-guided sparse querying, and modality reliability. RC-GeoCP studies how source preferences persist from acquisition into assembly. Its continuous pre-delivery responsibility selects locations within fixed source quotas along the spatial axis and induces CDA's assembly distribution along the source axis. Thus, *geometric consensus* denotes radar-supported source attribution shared by these decisions. The reported policy fixes outer allocation so the controls can examine selection and assembly within that allocation.

Table A3: Feature- and total-payload-matched scoring adaptations (AP@0.7).

Selector	AP@0.7		Total
	V2X-R	V2X-Radar	MB/link
CoSDH-style	63.88	24.91	4.95/10.88
SlimComm-style	64.03	25.17	4.95/10.88
UAC w/o comp.	62.46	21.43	4.95/10.88
Full UAC	**65.09**	**25.92**	4.95/10.88

Table A3 compares scoring adaptations with matched detector, token dimensions and count, training schedule, and request format. Feature values and agent tokens occupy 4.525/10.068 MB per link on V2X-R/V2X-Radar, with totals including controls and indices as defined in Section C.2. CoSDH-style and SlimComm-style scores replace only Top-K acquisition. CDA retains the original UAC responsibility formulation and Equation 10, with parameters learned in each model. These rows assess acquisition rules within the UAC-based assembler. Table A4 then compares shared and independent score parameterizations to examine coordination across acquisition and assembly.

Table A4: AP@0.7 ablation at the 40/30/20 allocation, with feature-plus-token payload fixed to 4.525/10.068 MB per link on V2X-R/V2X-Radar. The on/off rows form the 2×2 acquisition–assembly factorization.

Variant	V2X-R	V2X-Radar
Standard Top-K	60.61	20.74
UAC w/o complementarity	62.46±0.21	21.43±0.17
Independent CDA score	62.37±0.08	21.08±0.20
No geo. control (Acq. off, Asm. off)	63.10±0.15	22.07±0.53
Geo. acquisition only (on, off)	63.90±0.17	23.75±0.25
Geo. assembly only (off, on)	63.87±0.33	23.60±0.19
Full RC-GeoCP (on, on)	**65.09±0.27**	**25.92±0.22**
Interaction Δ_{int}	+0.42	+0.64

The independent-CDA control retains the UAC acquisition criterion and matches inputs, head capacity, and supervision, but learns a separate assembly score for Equation 10. Full exceeds this control by 2.72/4.84 in three-seed mean AP@0.7 on V2X-R/V2X-Radar, supporting shared source preferences and detection supervision. The factorial controls then locate radar's contribution within this coordination. Full exceeds both one-factor variants on each dataset, supporting radar-informed acquisition and assembly together. Their +0.42/+0.64 cell-mean interactions describe the joint response. Section E specifies the three-seed summaries and the scope of interaction uncertainty.

GSR component factorization. Table A5 separately toggles radar-conditioned sampling in Equation 2 and gated calibration in Equation 3 within the same local fusion pipeline. Sampling alone improves AP@0.7 by 2.75/3.23 points over No GSR on V2X-R/V2X-Radar, while calibration alone yields larger gains of 3.18/5.98 points. Their ordered composition achieves the best 64.99/23.81 AP, further improving the calibration-only variant by 0.51/1.07 points and the sampling-only variant by 4.94/3.82 points. This pattern identifies gated radar calibration as the larger individual contributor, while radar-guided resampling provides an additional benefit in the full GSR operator.

Table A5: Internal GSR factorization under the common protocol (AP@0.7).

Variant	Samp.	Calib.	V2X-R	V2X-Radar
No GSR			61.30	16.76
Sampling	✓		64.05	19.99
Calibration		✓	64.48	22.74
Full GSR	✓	✓	**64.99**	**23.81**

D Loss and Calculation

D.1 Training Objective

We optimize the network with a multi-task objective combining detection, depth estimation, and supervision of the shared responsibility. The total loss is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{det}} + \lambda_{\text{dep}}\mathcal{L}_{\text{dep}} + \lambda_{\text{uac}}\mathcal{L}_{\text{uac}}, \tag{14}$$

where $\mathcal{L}_{\text{det}}$ is the standard PointPillars-style (Lang et al., 2019) detection loss, $\mathcal{L}_{\text{dep}}$ is the depth loss following (Li et al., 2023b; Bai et al., 2025c), and $\mathcal{L}_{\text{uac}}$ supervises the control signals shared by acquisition and assembly.

The detection loss combines the following terms.

$$\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{cls}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} + \lambda_{\text{dir}}\mathcal{L}_{\text{dir}}, \tag{15}$$

where $\mathcal{L}_{\text{cls}}$ is the classification loss, $\mathcal{L}_{\text{reg}}$ is the regression loss, $\mathcal{L}_{\text{dir}}$ is the direction classification loss. Following (Yang et al., 2025), here we set λ_{reg} =2.0 and λ_{dir} =0.2, which represents the standard configuration commonly used in similar tasks.

In addition to detection and UAC supervision, we introduce an auxiliary depth classification objective to improve BEV lifting stability and 3D localization accuracy. Depth is discretized into D bins and predicted as a per-pixel categorical distribution. We then adopt focal loss for depth supervision.

$$\mathcal{L}_{\text{dep}} = \frac{1}{B}\sum \text{FL}\left(\hat{\mathbf{D}}, \mathbf{D}\right), \tag{16}$$

where $\hat{\mathbf{D}} \in \mathbb{R}^D$ denotes the predicted depth logits and $\mathbf{D}$ is the discretized ground-truth depth label. The overall depth loss weight is set to λ_{dep} =10.0.

To regularize the control signals used by UAC and CDA, we introduce occupancy-based supervision for semantic confidence and radar-derived geometric reliability. The objective is defined below.

$$\mathcal{L}_{\text{uac}} = \sum_{l=1}^{2} \omega_l \left(\mathcal{L}_{\text{occ-s}}^{(l)} + \mathcal{L}_{\text{occ-c}}^{(l)} + \lambda_{\text{geo}}\mathcal{L}_{\text{geo}}^{(l)}\right), \tag{17}$$

where l indexes pyramid level and ω_l denotes the corresponding pyramid weight. The single-agent term $\mathcal{L}_{\text{occ-s}}^{(l)}$ supervises each agent in its own coordinate frame using local occupancy ground truth,

Table A6: AP@0.7 on V2X-Radar under timestamp delays and pose perturbations.

Method	Timestamp delay (ms) ↓					Pose Std. (m/deg) ↓				
	0	50	100	150	200	0.00	0.05	0.10	0.15	0.20
CoBEVT	16.06	12.98	9.39	7.77	7.08	16.06	15.69	13.33	11.94	10.64
Where2Com	17.92	14.99	11.21	9.14	8.41	17.92	17.24	16.10	13.96	11.70
V2XViT	18.31	16.02	12.43	10.67	9.66	18.31	17.70	16.35	14.31	12.57
CoAlign-LXL	20.11	17.50	15.01	12.75	11.91	20.11	19.75	18.78	16.76	13.52
RC-GeoCP (**Ours**)	**25.92**	**21.41**	**15.14**	**13.87**	**13.13**	**25.92**	**24.83**	**21.87**	**19.25**	**15.90**

Table A7: AP@0.7 on V2X-R under timestamp delays and pose perturbations.

Method	Timestamp delay (ms) ↓					Pose Std. (m/deg) ↓				
	0	100	200	300	400	0.00	0.05	0.10	0.15	0.20
CoBEVT	51.48	40.41	34.65	32.24	31.52	51.48	50.10	45.65	38.05	30.79
Where2Com	44.60	37.45	32.83	29.24	28.49	44.60	42.87	39.44	33.72	26.01
V2XViT	47.72	35.07	30.02	28.20	27.98	47.72	46.35	42.03	35.83	28.90
CoAlign-LXL	63.57	47.20	42.75	42.03	40.33	63.57	62.15	57.44	48.40	37.80
RC-GeoCP (**Ours**)	**65.09**	**52.19**	**46.20**	**44.39**	**42.47**	**65.09**	**64.07**	**60.91**	**50.94**	**43.53**

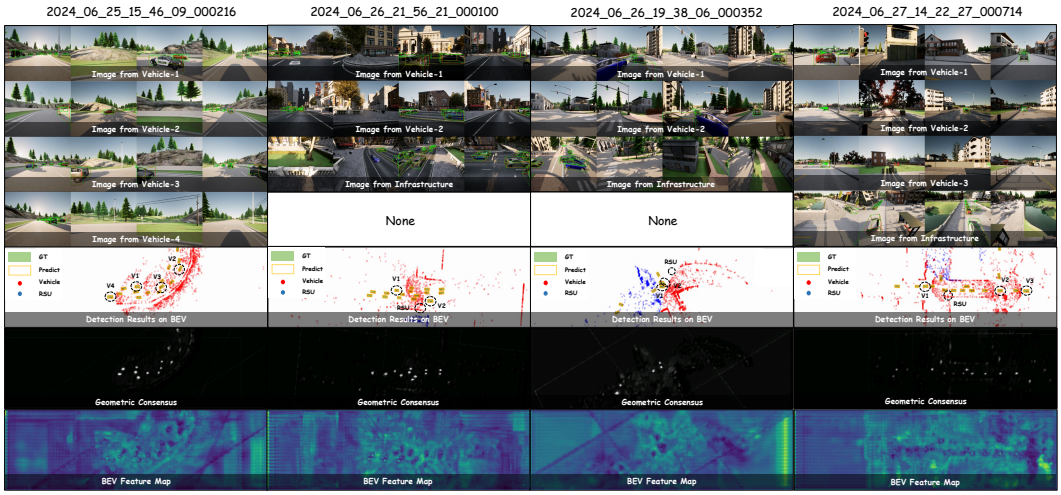

Figure A2: V2X-R examples relating multi-agent camera views to BEV detections, consensus responses, and assembled features (Huang et al., 2025).

providing a local confidence objective. The collaborative term $\mathcal{L}^{(l)}_{\text{occ-c}}$ supervises the ego-frame prediction after neighboring predictions are aligned and modulated by ego demand weights. This trains receiver-conditioned demand through the same occupancy task. Finally, $\mathcal{L}^{(l)}_{\text{geo}}$ uses the occupancy targets to train the radar-derived reliability score. Combining this support with receiver demand gives acquisition and assembly a shared attribution field grounded in task supervision. The coefficient λ_{geo} =0.1 controls its relative strength. All occupancy-related losses adopt a focal-and-dice combination to handle extreme foreground sparsity and stabilize region-level optimization.

D.2 Post-processing and Anchor Configuration

For decoding anchors and applying NMS, we use a standard voxel-based post-processing method. Ground-truth boxes and anchors are defined within each CAV's configured spatial range. Each BEV location is assigned two anchors with yaw angles of 0 and 90 degrees. The anchor box dimensions are fixed to 3.9, 1.6, and 1.56 meters, and the feature stride is set to 2. The box parameterization follows the order `hwl` during encoding and decoding. Positive and negative anchors are determined according to 3D IoU thresholds. An anchor is assigned as positive if IoU is at least 0.6 and negative if IoU is at most 0.45. Anchors with intermediate IoU are ignored during training. To resolve the orientation ambiguity of symmetric boxes, we use a two-bin direction classification head with an offset of 0.7853 radians. During inference, predicted boxes with confidence scores below 0.10 are discarded. We retain at most 150 boxes per frame and apply Non-Maximum Suppression (NMS) with a threshold of 0.05. These settings control confidence filtering and duplicate suppression across

Table A8: TP/FP/FN comparison at IoU 0.7 on V2X-Radar.

Range (m)	RC-GeoCP			CoAlign			HEAL			V2X-ViT		
	TP↑	FP↓	FN↓	TP↑	FP↓	FN↓	TP↑	FP↓	FN↓	TP↑	FP↓	FN↓
Overall	**2742**	**3634**	**5183**	2279	14720	5646	2443	4818	5482	1948	25062	5977
(0, 30)	**2172**	**1753**	**2178**	1793	6778	2557	1980	2402	2370	1571	8327	2779
(30, 50)	**476**	**977**	**1811**	356	4617	1931	387	1397	1900	310	8014	1977
(50, 100)	81	**917**	1207	**116**	3317	**1172**	65	1030	1223	55	8686	1233

Table A9: TP/FP/FN comparison at IoU 0.7 on V2X-R.

Range (m)	RC-GeoCP			CoAlign			HEAL			V2X-ViT		
	TP↑	FP↓	FN↓	TP↑	FP↓	FN↓	TP↑	FP↓	FN↓	TP↑	FP↓	FN↓
Overall	**6210**	6701	**2546**	6037	6619	2719	6038	**5882**	2718	5024	3419	3732
(0, 30)	**3148**	1700	**374**	2969	1903	553	3135	2623	387	2678	**979**	844
(30, 50)	1409	1440	577	1423	1385	563	**1499**	1415	**487**	1253	**822**	733
(50, 100)	**1445**	2655	**1103**	1439	2365	1109	1249	1719	1299	965	**1376**	1583

the compared methods. Evaluation on V2X-Radar and V2X-R follows the physical limitations of the sensor field of view (FoV). The framework retains ground-truth boxes located only within the valid forward-facing sensing range relative to the vehicle. This filtering process ensures that ground-truth labels remain consistent with the active sensing area of the collaborative agents.

E Training Details

Pre-training strategy.

Training follows two stages. We first optimize EfficientNet (Tan & Le, 2019) for depth estimation to initialize the camera representation before collaborative learning. After depth pretraining converges, its weights initialize the image backbone of the detection network. The entire framework is then jointly optimized for collaborative detection with depth and UAC supervision. This schedule provides a depth-aware starting point for GSR and subsequently trains local features and receiver–source attribution for the detection task.

Computing infrastructure. All experiments are conducted on a workstation running Ubuntu 20.04.6 LTS with two AMD EPYC 9554 64-Core CPUs and NVIDIA RTX A6000 GPUs, each with 48 GB memory. The implementation is based on PyTorch 1.12.0 with CUDA 11.3, OpenCOOD, and MMDetection3D. We use MMCV-Full 1.7.0, MMDetection 2.27.0, and spconv-cu113.

Randomness control. For reproducibility, the training and evaluation scripts fix the random seed for Python, NumPy, PyTorch, and CUDA random number generators. We further enable deterministic cuDNN behavior and initialize data-loader workers from the base seed, following the seed utilities provided in the released code. Single-run results use seed 42. Tables A4 and A10 identify the replicated configurations and report three-seed summaries.

Table A10: Three-seed AP@0.7. Model rows are mean ± standard deviation over the same seeds, and Δ is the difference of means. The baseline is CoAlign/HEAL on V2X-R/V2X-Radar, and Full uses the reported 40/30/20 sparse policy.

Variant	V2X-R	V2X-Radar
Common-protocol baseline	61.30 ± 0.34	16.76 ± 0.29
GSR + CDA (dense)	67.40 ± 0.28	26.30 ± 0.24
Full w/o radar control	63.10 ± 0.15	22.07 ± 0.53
Full RC-GeoCP	65.09 ± 0.27	25.92 ± 0.22
Δ radar control	+1.99	+3.85

Across three independently trained seeds, Full RC-GeoCP reaches 65.09 ± 0.27/25.92 ± 0.22 AP@0.7 on V2X-R/V2X-Radar. Radar control changes the aggregate mean by 1.99/3.85 AP@0.7 across the two datasets. All four acquisition–assembly cells in Table A4 report three-seed mean ± standard deviation. The factorial contrast is computed from these cell means. Its paired uncertainty is unavailable because the seed-level pairing was not retained. Table A10 summarizes endpoint statistics and differences of means. Its means also provide the corresponding AP@0.7 endpoints cited in the main text.

Optimization and selection settings. Both datasets use batch size 4 and 48 training epochs. The camera lift uses 98 uniform depth bins over 2–100 m. We optimize with Adam at learning rate 0.0012 and weight decay 0.0001. Cosine annealing uses 500 linear warm-up iterations with warm-

up ratio 0.1. The depth and UAC loss weights are 10 and 0.1, and the three pyramid auxiliary weights are 0.4/0.2/0.1. UAC samples each per-level retained ratio from [0.25, 0.75] during training after warm-up. The distributed training loop saves checkpoints and evaluates validation-set overall AP@0.3/AP@0.5/AP@0.7 every four epochs. Cross-method evaluation loads the last saved checkpoint from each completed training run, following Section B.1. For sparse evaluation, we select 40%/30%/20% at the fine/mid/coarse levels from the validation-only sensitivity sweep in Table 6 as a shared cross-dataset accuracy–payload compromise. This choice avoids the V2X-R drop at uniform 25% while V2X-Radar is saturated across the listed sparse settings. Uniform 25% and 50% expose the trade-off, while 40/30/20 serves as a validation-selected operating point with fixed outer allocation.

Token warm-up strategy. To stabilize sparse-token training, we linearly reduce the effective transmission ratio from dense exchange to the target ratio. For target r_{target}, epoch e uses

$$r_{\text{eff}}(e) = 1 - \frac{e}{T_{\text{warm}}}\left(1 - r_{\text{target}}\right), \quad 0 \le e < T_{\text{warm}}, \tag{18}$$

which decreases from 1 to r_{target}. We set the warm-up period to 12 epochs. Thereafter, the training ratio is sampled from [0.25, 0.75]. Evaluation uses the fixed ratios stated above. The dense-to-sparse curriculum gives demand learning access to broad source support before training with sparse requests.

F Experiments

Perturbation protocol. Timestamp delay replaces each non-ego observation with the frame at the stated delay while retaining the current ego frame and applying no motion compensation. Pose perturbation adds independent zero-mean Gaussian noise to agent x/y positions and yaw. The tabulated scalar is the position standard deviation in meters and yaw standard deviation in degrees. All methods use the same perturbed inputs and seed 42.

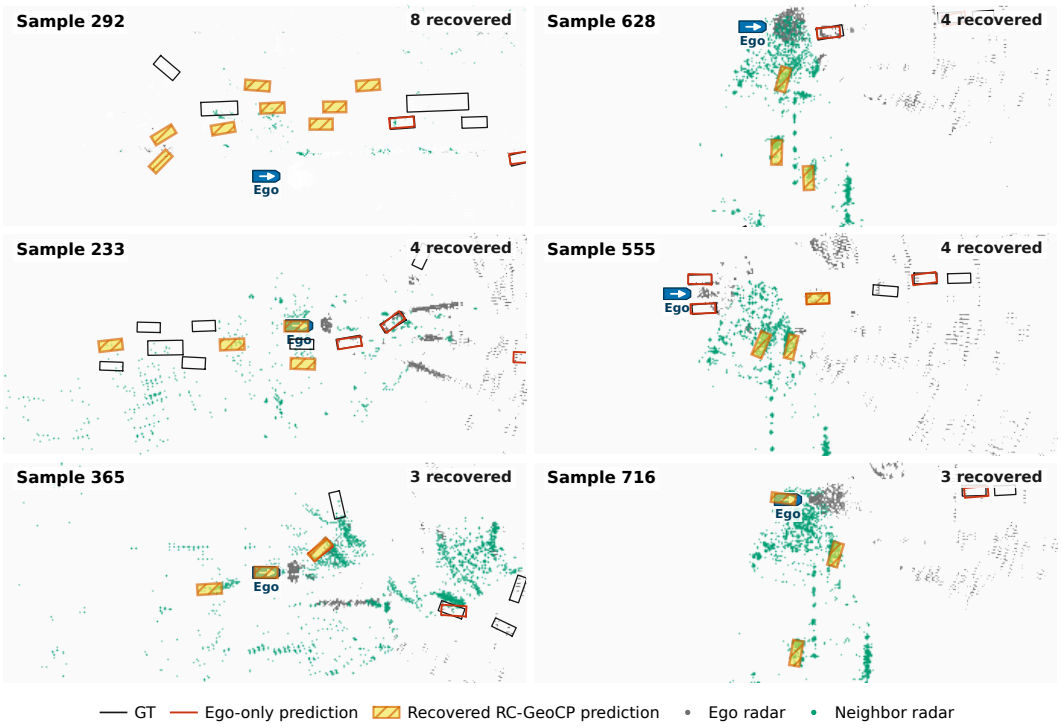

Figure A3: Six RC-GeoCP recovery examples on V2X-Radar. Each yellow-hatched box with an orange border is an RC-GeoCP prediction whose matched GT is missed by Ego-only inference and recovered at a BEV IoU of at least 0.7. Red outlines show Ego-only predictions in the same scene. The overlapping recovered-GT outline and unrelated collaborative predictions are omitted to avoid duplicate marks. Gray/green points denote Ego/neighbor radar returns, other GT boxes are black, and the blue vehicle marks the Ego pose and $+x$ heading.

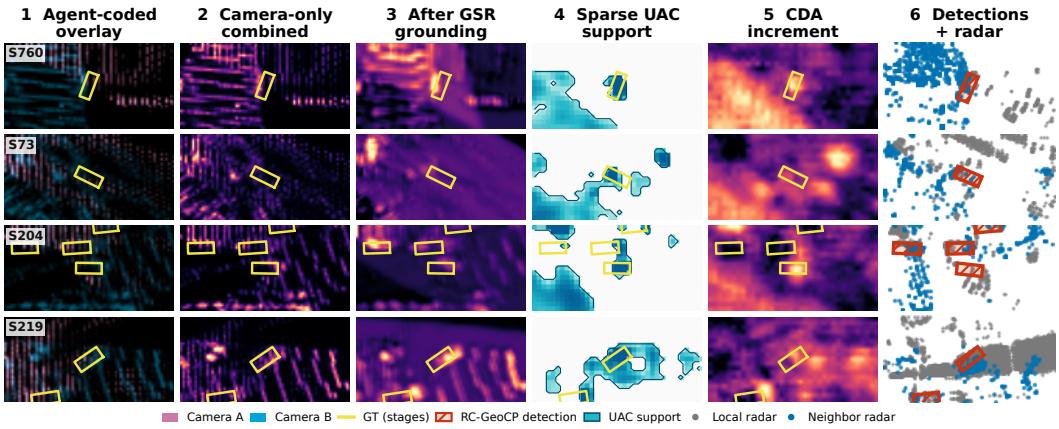

Figure A4: Four target-centered RC-GeoCP recoveries using identical 28 m × 14 m crops and targets at least 15 m from the local agent. Columns show the ego-aligned agent-coded camera overlay with A in magenta, B in cyan, and their overlap in white, followed by the combined response, per-agent GSR, sparse UAC support, and CDA increment over ego-only assembly, and detections over local/neighbor radar (gray/blue). Yellow boxes in columns 1–5 are GTs matched by visible final detections at a BEV IoU of at least 0.7. Column 6 shows detections in red and omits unmatched predictions. Heatmaps are independently normalized, so amplitudes are not comparable across columns.

F.1 RGB Stress Test

We evaluate frozen checkpoints on all 829 V2X-R validation frames with deterministic RGB corruptions. Low light applies gamma 1.5 followed by intensity scaling of 0.35. Synthetic fog uses an 11-pixel mean blur with transmission 0.35 and white atmospheric light. Radar inputs, geometry, labels, and frame order remain fixed. The clean column uses the same nominal operating point as Table 4. RC-GeoCP retains 87.3%/65.4% of clean AP, compared with 78.1%/40.5% for Who2com. The absolute margin grows from 19.52 AP on clean inputs to 21.20/24.13 AP under low light/fog (Table A11).

Tables A6 and A7 evaluate robustness under increasing input asynchrony and pose perturbation. These sweeps measure sensitivity to stale neighbor observations and perturbed alignment with a fixed protocol. On V2X-Radar (Yang et al., 2025), RC-GeoCP achieves 25.92% AP@0.7 under ideal synchronization. At 200 ms delay, it reaches 13.13%, 1.22 points above CoAlign-LXL (Lu et al., 2023; Xiong et al., 2024). Under a pose error of 0.20, it reaches 15.90%, 3.33 points above V2X-ViT. It has the highest absolute AP at every reported level.

Table A11: AP@0.7 under controlled RGB degradation on all 829 V2X-R validation frames. Parentheses give retention relative to each method's clean AP.

Method	Clean	Low light	Synthetic fog
Who2com	45.57	35.60 (78.1%)	18.47 (40.5%)
RC-GeoCP	**65.09**	**56.80 (87.3%)**	**42.60 (65.4%)**

A similar trend is observed on V2X-R (Huang et al., 2025), where overall accuracy is higher yet robustness remains crucial. At 400 ms timestamp delay, RC-GeoCP reaches 42.47% AP@0.7, outperforming CoAlign-LXL by 2.14 points and CoBEVT (Xu et al., 2023a) by 10.95 points. Under pose error 0.20, it reaches 43.53%, 5.73 points above CoAlign-LXL and 14.63 above V2X-ViT (Xu et al., 2022a). These sweeps show an absolute AP advantage over the evaluated perturbations, while clean-normalized retention depends on the perturbation type and severity.

To examine the detection profile across ranges, Tables A8 and A9 report TP/FP/FN at IoU=0.7. Matching is recomputed within each range after filtering predictions and GT, while the overall row uses the full configured coverage. The displayed bins therefore need not sum to the overall counts. On V2X-Radar, RC-GeoCP has the highest overall TP and the lowest overall FP and FN. The difference is concentrated within 0–50 m. At 50–100 m, CoAlign has higher TP and lower FN. On V2X-R, RC-GeoCP has the highest overall TP and lowest FN, but a higher raw FP count. HEAL leads the medium-range TP/FN counts, whereas RC-GeoCP leads long-range TP/FN. This is a recall-led profile, while AP still evaluates the full precision–recall ranking.

F.2 Object-Level Collaborative Recovery

Figure A3 shows Ego misses recovered by RC-GeoCP. Yellow hatching marks the matched collaborative prediction, green points show neighbor radar, and a missing or displaced red Ego box identifies the original miss. The blue vehicle and arrow define Ego-frame $+x$. Across 64 uniformly sampled V2X-Radar validation frames containing 546 GT objects, Ego detects 93, while RC-GeoCP recovers 112/453 misses (24.7%) and loses two prior hits. Six examples are drawn from this evaluation population using recovery count, target-local neighbor support, temporal separation, and full metric-FoV visibility. Every shown frame has at least three recoveries and no regression.

F.3 Stage-wise Evidence Evolution

Figure A4 adds four recovery traces to the two main-paper examples. Columns 1–2 show agent-specific camera splats and their combination. Column 3 aligns the independently rectified, still-local GSR features for comparison. UAC shows selected neighbor support, CDA its increment over ego-only assembly, and the final prediction matches the GT. Cases satisfy distance, recovery, camera-support, and UAC-localization criteria. Independent normalization makes these maps a view of where each stage responds. The 2×2 study and the independent-CDA control quantify the effect of the corresponding design choices.

G More Visualization Results

Figures A1 and A2 show vehicle/infrastructure views, BEV detections, consensus maps, and fused features on both datasets. Independently normalized responses localize object clusters and road structure under viewpoint change and partial occlusion. They illustrate spatial support, while controlled ablations and the all-object recovery count provide quantitative evidence.