HERO: Heterogeneous Evidential Robust Object-Level Collaborative Perception

Hao Si*, Ehsan Javanmardi, Hanlin Wu, and Manabu Tsukada

The University of Tokyo, Tokyo, Japan
{si-hao,ejavanmardi,hanlinwu,mtsukada}@g.ecc.u-tokyo.ac.jp

Abstract. Real-world collaborative perception agents typically deploy diverse sensor configurations and network architectures. While feature-level fusion methods address this heterogeneity by aligning feature spaces, they require collaborative training and large communication overhead. In contrast, object-level fusion sidesteps these constraints, as transmitted bounding boxes are directly usable across varying architectures. However, standard late fusion relies only on box coordinates and confidence scores, which provide insufficient information for robust association and leave the system vulnerable to localization noise and asynchronous delays. To address this, we propose HERO, an uncertainty-guided object-level collaborative perception method that equips proposals with evidential statistics and decouples the two fusion decisions, aggregating semantic evidence across matched proposals while selecting geometry from the most reliable source, to perform robust association and fusion under pose noise and asynchrony. Experiments on OPV2V-H and DAIR-V2X show that HERO matches the SOTA feature-level fusion methods while transmitting only about 2 KB per frame of sparse object proposals. Under localization error and latency scenarios, HERO further outperforms feature-level SOTA models.

1 Introduction

Collaborative Perception aggregates observations from multiple agents to improve 3D detection, see through occlusion, and extend the effective perception range [8, 10, 46]. Depending on the granularity of shared information, existing approaches can be categorized into these types: early fusion that transmits raw sensor data, such as point clouds or images; feature-level fusion exchanging intermediate BEV features; object-level late fusion sharing detected proposals; and hybrid approaches integrating multiple fusion levels [16, 34]. Feature-level fusion has achieved strong accuracy [24,31,45], but it typically transmits dense features and relies on architectural compatibility and collaborative training.

However, real V2X deployments are open and heterogeneous: vehicles and roadside units may use different sensing modalities and detector backbones [35, 43], and enforcing collaborative training and feature alignment is unrealistic.

* Corresponding author.

This makes object-level late fusion naturally fit open deployment constraints: exchanging sparse proposals enables detector-agnostic cooperation with KB/frame bandwidth. A naive object-level pipeline typically communicates compact messages of 3D boxes and scalar confidence [8]. The ego agent transforms all boxes into a common frame and applies NMS. This process becomes unstable in heterogeneous settings and under system perturbations. Different agents have different localization noise scales, making fixed IoU thresholds prone to false merges or missed matches; pose bias further reduces inter-box IoU and degrades association [28]; and raw confidence scores are not comparable across models, so "highest score" does not imply "most reliable" [7]. Moreover, stale observations from communication delay can corrupt both association and fusion [4].

To address these issues, we propose HERO, an object-level framework that makes cooperation robust under open heterogeneity, pose errors, and communication delays. HERO attaches evidential statistics to each proposal: Beta statistics for objectness and Normal-Inverse-Gamma (NIG) statistics for regression, serving as proxies for semantic and geometric uncertainty. Building on these signals, its key design principle is to decouple the two fusion decisions. Object existence is corroborated by evidence from multiple matched proposals, so HERO aggregates semantic evidence across agents; geometry is instead vulnerable to source-level pose and localization bias, where averaging boxes can corrupt a well-localized proposal, so HERO selects geometry from the most reliable source. As shown in Fig. 1, the same uncertainty signals also drive Mahalanobis-gated association, delay-aware freshness weighting, and detection-level pose refinement (DPR), allowing HERO to handle asynchrony and systematic pose bias without collaborative training and to support heterogeneous detector combinations.

On the OPV2V-H [19] dataset, HERO reaches AP50=0.922 and AP70=0.841, which are comparable to strong feature-level fusion baselines. On DAIR-V2X [44], HERO achieves the best performance among all evaluated methods, obtaining AP50=0.823/0.744/0.753 on the three V2I heterogeneous combinations. HERO communicates only 2 KB/frame, over 1000× lower than dense BEV feature exchange. Under the noise setting on OPV2V-H, HERO remains robust: it achieves AP70=0.763 under pose noise $\sigma = 0.6$ m and AP70=0.774 under a 500 ms asynchronous delay. Our contributions are:

- We propose an evidential proposal message and a decoupled fusion rule for object-level cooperation: object existence is aggregated by summing semantic evidence across matched proposals, while geometry is taken from the single most reliable proposal rather than averaged, avoiding the box drift caused by correlated source-level pose bias.
- We build an ego-side inference pipeline that makes this message robust under real-world heterogeneous deployment, combining Mahalanobis-gated association, delay-aware freshness weighting, and detection-level pose refinement with no cross-agent collaborative training.
- Experiments on OPV2V-H and DAIR-V2X show feature-level accuracy with KB/frame communication and stronger robustness under pose noise and asynchronous delay.

2 Related Work

Fusion paradigms. Based on the granularity of shared information, cooperative 3D detection can be categorized into early fusion, feature-level fusion, and object-level late fusion [8]. Early fusion shares raw sensor data and can approach high performance, but it imposes stringent requirements on bandwidth, spatiotemporal synchronization, and calibration [17]. Feature-level methods exchange BEV features and perform joint inference across agents; they have achieved strong accuracy [15, 32, 38, 39] and propose efficient cooperation mechanisms along the axes of when, where, and how to communicate [11, 18, 22, 42]. However, feature-level fusion often assumes compatible feature spaces and collaborative training and typically requires MB/frame communication. Late fusion exchanges compact object proposals, is naturally decoupled from detector backbones, and better matches bandwidth and cooperation constraints in open deployments. A broader bottleneck of object-level fusion is the limited information in the message: under localization errors and asynchronous delays, aggregation based on fixed thresholds is insufficient for robust association and fusion.

Heterogeneous cooperative perception. Heterogeneity is inevitable in V2X deployments: agents may use different sensing modalities and detection networks [21, 37]. HEAL [19] explicitly formulates open heterogeneous cooperation and provides an extensible framework supporting diverse agent types, while HM-ViT [33] introduces cross-modal modeling and alignment in feature-level fusion. PHCP [27] addresses inference-time adaptation through few-shot pseudo-label self-training. STAMP [6] emphasizes scalable task- and model-agnostic cooperation; GenComm [48] constructs exchangeable information via generative communication under heterogeneity; and CodeFilling [12] improves communication efficiency via codebook-driven information filling. Unlike approaches centered on feature alignment, inference-time adaptation, generation, or reconstruction, HERO focuses on object-level message design and inference rules.

Robustness to pose errors and asynchronous delay. Pose noise and calibration bias can shift features or proposals from different agents, substantially degrading cooperation. CoAlign [20] and related pose-correction methods [29] mostly study online alignment in feature-level pipelines, while object-level methods such as VIPS [26] and OptiMatch [28] match, rectify, or temporally smooth proposals under V2X perturbations. Asynchronous delay introduces another challenge: messages are often stale upon arrival, and naive gating can over-associate or miss associations under large delays. HERO instead equips proposals with evidential Beta/NIG statistics and handles both perturbations at the object level, correcting pose via inference-time DPR and discounting stale messages via freshness decay, without retraining.

Uncertainty and evidential learning. Uncertainty-aware learning is widely used for confidence calibration, reliability estimation, and safety-critical decision making. Evidential Deep Learning (EDL) represents predictions using conjugate

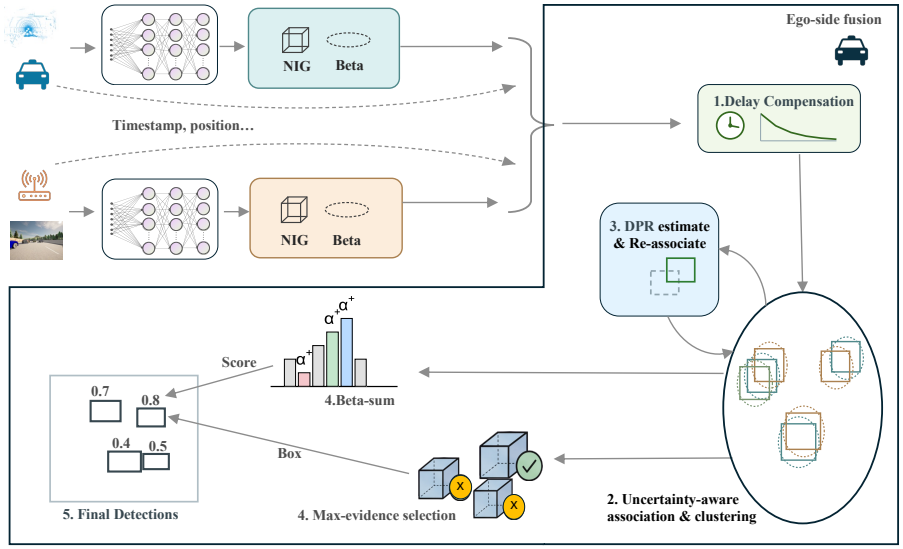

Fig. 1: Overview of HERO. Agents transmit evidential proposals with 3D boxes, Beta objectness, and NIG geometry statistics. On the ego side, HERO aligns proposals to the ego frame; freshness factors discount stale proposals; DPR uses an internal coarse association to estimate source bias; after pose correction, HERO re-runs uncertainty-aware clustering and decouples final fusion into Beta-sum semantic aggregation and max-evidence geometry selection.

distribution parameters so that a single forward pass yields both the predictive mean and an "evidence strength", and it can be extended to regression via NIG [1,23]. Meanwhile, the interpretation and calibration of EDL-derived uncertainty signals have been actively discussed, with prior work reporting potential failure modes and calibration issues in certain settings [2,25]. Motivated by these observations, we treat the Beta/NIG outputs as lightweight, interpretable proxy signals for association gating and fusion weighting. EDL based fusion has been explored in V2X perception [14, 47]; in contrast, HERO focuses on inference robustness under open heterogeneous deployment.

3 Method

HERO is an object-level collaborative perception framework for heterogeneous deployments, targeting robust 3D detection under limited bandwidth, localization errors, and communication delays. It defines a unified evidential message format that any detector can produce, and the ego agent fuses received messages without cross-agent collaborative training. Its defining choice is to decouple fusion: existence is aggregated across matched proposals, while geometry is selected from the most reliable source. Figure 1 illustrates the pipeline, Tab. 1 summarizes notation, and we detail each stage in the following subsections.

Table 1: Notation summary for evidential proposals and ego-side fusion.

Symbol	Definition
$\alpha^+, \alpha^-, S, p, u$	Beta objectness parameters, evidence strength, objectness mean, and semantic uncertainty.
$\gamma, \nu, \alpha_r, \beta_r$	NIG regression parameters for each box dimension; ν is the regression pseudo-count.
$\bar{\nu}, u_{\text{epi}}, \sigma^2_{i,k}$	Mean pseudo-count across box dimensions, epistemic-uncertainty proxy, and planar uncertainty for association.
$\tau_{\text{abs}}, \tau_{\text{mah}}$	Absolute-distance and Mahalanobis gates for proposal association.
$r_i, f_{\text{geo}}, f_{\text{sem}}$	Reliability and freshness weights for fusion; the same age signal also enters DPR.
$r_{a,c}, w_{a,c}, \Delta t_a$	DPR residual, match weight, and estimated per-agent translation correction.

3.1 Evidential Detection Head

The evidential detection head parameterizes semantic confidence and geometric uncertainty with evidential distributions, making each proposal more informative than the conventional "3D box + confidence score" representation.

Semantic evidence (Beta). For objectness, we output Beta parameters (α^+, α^-) with $\alpha^{\pm} \geq 1$. $S = \alpha^+ + \alpha^-$ denotes the evidence strength. We adopt the uncertainty definition in binary EDL, $u = 2/S$ [23]:

$$S = \alpha^+ + \alpha^-, \quad p = \frac{\alpha^+}{S}, \quad u = \frac{2}{S}. \tag{1}$$

Geometric uncertainty (NIG). For box regression, for each dimension, we output NIG parameters $(\gamma, \nu, \alpha_r, \beta_r)$ [1] with $\alpha_r > 1$. The pseudo-count ν reflects the effective number of observations for that regression dimension. We use the average pseudo-count over the 7 box dimensions, $\bar{\nu}$, to form a proxy for epistemic uncertainty:

$$u_{\text{epi}} = \frac{1}{\bar{\nu} + 1}. \tag{2}$$

Small $\bar{\nu}$ often indicates weak support for the predicted box. For association, the x, y predictive variances from the NIG head provide planar uncertainty terms $\sigma^2_{i,k}$ in Eq. (4). We use these quantities as empirical reliability signals.

Message sparsification. To satisfy bandwidth budgets, we select proposals using a joint score:

$$s = (1 - u)p = \frac{(S - 2)\alpha^+}{S^2}. \tag{3}$$

This score encourages both high objectness and strong evidence. Each agent keeps Top-K candidate boxes locally and sends them to the ego agent after the per-agent NMS process.

3.2 Uncertainty-Aware Association

Conventional late fusion aggregates boxes from all agents and directly applies NMS to remove duplicates. This strategy essentially treats multi-agent cooperation as a deduplication problem: NMS keeps the box with the highest confidence score and discards the rest, permanently losing supporting evidence from peers. In contrast, we view redundancy as an opportunity for evidence aggregation. We cluster candidate boxes that may correspond to the same physical object via Mahalanobis-gated association. Proposals within the same cluster jointly support the existence and geometry of the object, enabling the ego agent to make fusion decisions on a stronger information basis.

Coordinate alignment. Received proposals are first transformed into the ego coordinate frame via the relative pose $T_{e\leftarrow a}$. Let $x_i \in \mathbb{R}^2$ denote the BEV center of proposal i. We transform both the proposal center and its planar uncertainty into the ego frame before association.

Dual gating. When two proposals i, j satisfy $\|x_i - x_j\| \leq \tau_{\text{abs}}$ in BEV distance and their Mahalanobis distance is below τ_{mah}, we consider them matchable (potentially belonging to the same physical object):

$$d^2_{\text{mah}}(i,j) = \sum_{k\in\{x,y\}} \frac{(x_{i,k} - x_{j,k})^2}{\sigma^2_{i,k} + \sigma^2_{j,k} + \epsilon}. \tag{4}$$

Here $\sigma^2_{i,k}$ is the ego-frame planar uncertainty for proposal i, and $\epsilon > 0$ is a numerical stabilizer. Mahalanobis gating adapts to stochastic uncertainty: proposals with larger localization variance allow for larger spatial deviations, while high-precision proposals require tighter matches. In our main setting, we use a fixed set of gating hyperparameters $(\tau_{\text{abs}}, \tau_{\text{mah}})$.

Greedy clustering. We sort proposals by evidence strength S and form clusters greedily. Each unassigned proposal seeds a cluster and absorbs matchable proposals, while enforcing at most one proposal per agent in each cluster. A temporary uncertainty-weighted centroid is used only to stabilize association; the final output geometry is never averaged and is determined by max-evidence selection.

3.3 Fusion: Bounding-Box Selection and Semantic Aggregation

For each cluster C, HERO treats semantic and geometric fusion differently by design. Matched proposals can provide corroborating evidence for object existence, so semantic pseudo-counts are aggregated. In contrast, transformed boxes may share source-level pose bias, so geometric averaging can propagate bias instead of canceling it. HERO therefore aggregates semantics but selects geometry. We describe each step below; the full pipeline is given in Algorithm 1.

Reliability. We construct a reliability score from the epistemic uncertainty proxy and clip it to avoid extreme weights:

$$r_i = \mathrm{clip}(1 - u_{\mathrm{epi},i},\ r_{\min},\ 1)\,, \qquad r_{\min} = 0.1. \tag{5}$$

Cross-modal evidence scale. In heterogeneous deployments, evidence scales need not be perfectly calibrated across modalities. HERO therefore uses native evidence and uncertainty as empirical reliability signals, so weak or stale proposals are down-weighted by r_i and the freshness terms rather than by hand-crafted modality penalties.

Beta-sum semantic fusion. We aggregate pseudo-counts within each cluster and weight them by reliability and a semantic freshness factor f_{sem}. From a conjugate-distribution perspective, when agents output a shared consistent scale and are conditionally independent, the addition of pseudo-counts corresponds to an ideal posterior update. In open heterogeneous deployments, these assumptions do not strictly hold, so we treat it as an interpretable fusion rule and down-weight unreliable or stale evidence via r_i and f_{sem}:

$$\alpha^{+*} = 1 + \sum_{i \in C} r_i f_{\mathrm{sem}}(\Delta t_i)\,(\alpha_i^{+} - 1), \tag{6}$$

$$\alpha^{-*} = 1 + \sum_{i \in C} r_i f_{\mathrm{sem}}(\Delta t_i)\,(\alpha_i^{-} - 1). \tag{7}$$

The fused objectness is $p^* = \alpha^{+*}/S^*$ where $S^* = \alpha^{+*} + \alpha^{-*}$. The fused confidence is $c^* = p^*(1 - u^*)$ where $u^* = 2/S^*$.

Max-evidence geometry selection. When cooperative proposals come from an agent with pose bias, proposals from the same agent often share correlated systematic errors in the ego frame rather than independent noise. Under such correlated errors, covariance-weighted geometric averaging cannot cancel the bias and may instead pull the ego's high-precision geometry toward the biased direction, which is especially harmful at high-IoU metrics such as AP70. Therefore, HERO uses selection instead of averaging: it outputs the candidate with the strongest evidence, highest reliability, and best freshness:

$$i^* = \arg\max_{i \in C} S_i \cdot r_i \cdot f_{\mathrm{geo}}(\Delta t_i), \tag{8}$$

This criterion uses evidence strength and reliability derived from regression uncertainty rather than directly comparing raw confidence scores across models and is thus less sensitive to score-scale mismatch. It is inherently robust to correlated pose bias: semantics are aggregated via Beta-sum, while geometry is provided by a single most trustworthy source.

Algorithm 1 HERO ego-side fusion

Input: per-agent proposals, relative poses, delays
Output: fused detections $\{(b, c^*)\}$
1: Align all proposals to the ego frame
2: Compute freshness and reliability scores
3: Estimate per-agent pose bias by DPR and realign proposals (Eq. (11))
4: Cluster proposals with dual gates and one-proposal-per-agent constraint (Eq. (4))
5: **for** each cluster C **do**
6: **Geometry:** select the max-evidence box (Eq. (8))
7: **Semantics:** aggregate objectness by Beta-sum (Eq. (7))
8: Emit the selected box with fused confidence c^*
9: **end for**

3.4 Delay Compensation

To handle communication delay, HERO down-weights stale information without re-training. The same age signal enters association gating and DPR through motion-variance inflation, and enters semantic aggregation and bounding-box selection through freshness factors:

Freshness decay. Object categories are relatively stable at short time scales, while object locations are not. A vehicle observed 500 ms ago is still a vehicle, but it may have moved several meters. We use two decay factors to model this asymmetry:

$$f_{\text{geo}}(\Delta t) = \frac{1}{1 + \left(\frac{v_{\max} \cdot \Delta t}{d_{\text{ref}}}\right)^2}, \qquad (9)$$
$$f_{\text{sem}}(\Delta t) = \sqrt{f_{\text{geo}}(\Delta t)}.$$

Here $v_{\max}$ is the maximum expected agent speed, and d_{ref} is a reference localization scale (on the order of a typical object width). f_{sem} decays slower than f_{geo}, reflecting the prior that "categories are more stable than positions". Both equal 1 at $\Delta t = 0$, so they do not affect the zero-delay setting.

Delay-aware association gating. Even with correct pose transforms, object motion introduces offsets between predictions and current locations. For a proposal pair (i, j) with a time gap Δt_{ij}, we relax Mahalanobis gating by adding a saturated motion variance in the denominator of Eq. (4), whose cap prevents delayed matching from becoming overly permissive under long latency:

$$\sigma^2_{\text{gate}} = \min\left((v_{\max}\Delta t_{ij})^2, \sigma^2_{\text{cap}}\right). \qquad (10)$$

3.5 Detection-Level Pose Refinement

Pose errors introduce consistency bias: proposals from the same agent share a common planar bias in the ego frame. We estimate and correct this bias from

matched clusters, without collaborative training. DPR is not an isolated add-on; it reuses uncertainty proxies and the age-based down-weighting principle from association in Sec. 3.2 and fusion in Sec. 3.3.

To ensure identifiability, DPR uses ego proposals as anchors and runs a two-pass inference procedure. We first associate proposals under the current poses and collect clusters containing both the ego and agent a. For each matched cluster c, we define the BEV residual $r_{a,c} = x_c^{(e)} - x_c^{(a)}$ and compute reliability weights $w_{a,c}$ from evidence, uncertainty, and age. In the translation-only case, the per-agent correction is the weighted residual mean:

$$\Delta t_a = \frac{\sum_c w_{a,c}\, r_{a,c}}{\sum_c w_{a,c}}. \tag{11}$$

HERO also supports an optional yaw-aware correction for far-range alignment. After applying the estimated correction to all proposals from agent a, we run association and fusion again. This adds one lightweight extra association pass at inference time and does not introduce training.

4 Experiments

We evaluate HERO on the simulated OPV2V-H dataset and the real-world DAIR-V2X dataset. Results show that HERO achieves accuracy comparable to strong feature-level baselines while requiring only KB/frame proposal communication and no collaborative training. We further analyze robustness under pose noise and asynchronous delays and report the trade-off between communication cost and accuracy. Finally, we validate key design choices via ablations on core modules and fusion rules.

4.1 Experimental Setup

Datasets. We conduct experiments on OPV2V-H [19] and DAIR-V2X [44]. OPV2V-H extends the OPV2V benchmark [40]. It is collected with CARLA [5] and OpenCDA [36] to support heterogeneous multi-agent cooperative 3D detection, covering multi-modal sensors and diverse detector backbones. DAIR-V2X is a real-world vehicle-infrastructure cooperative (V2I) dataset with both vehicle and roadside agents in urban traffic scenes.

Heterogeneous Setup. On OPV2V-H, we use four heterogeneous agent models: two LiDAR encoders based on PointPillars [13] and SECOND [41], denoted as L_P and L_S, alongside two camera models. These camera models feature EfficientNet [30] and ResNet [9] backbones, designated as C_E and C_R respectively. On DAIR-V2X, we fix the vehicle-side agent as L_P and evaluate three vehicle–infrastructure heterogeneous combinations with the roadside unit.

Table 2: Quantitative comparison on the OPV2V-H dataset. The table presents AP@0.5 and 0.7 across different heterogeneous setups, evaluated progressively as more agents are incorporated into collaboration.

Method	$L_P + C_E$		$L_P + C_E + L_S$		$L_P + C_E + L_S + C_R$	
	AP50↑	AP70↑	AP50↑	AP70↑	AP50↑	AP70↑
F-Cooper [3]	0.842	0.686	0.843	0.686	0.843	0.686
DiscoNet [15]	0.865	0.716	0.870	0.720	0.869	0.720
AttFusion [40]	0.864	0.696	0.867	0.697	0.868	0.694
CoBEVT [38]	0.877	0.734	0.897	0.750	0.900	0.754
HEAL [19]	0.875	0.782	0.922	**0.853**	0.923	**0.855**
STAMP [6]	0.829	0.709	0.837	0.725	0.831	0.718
CodeFilling [12]	0.882	0.751	0.916	0.786	0.916	0.787
GenComm [48]	**0.892**	0.733	**0.924**	0.774	**0.925**	0.775
HERO(ours)	0.881	**0.788**	0.922	0.840	0.922	0.841

Table 3: Performance of two-agent heterogeneous collaboration on DAIR-V2X and OPV2V-H. We evaluate the AP@0.5 across various pairwise combinations.

Method	DAIR-V2X (AP50↑)			OPV2V-H (AP50↑)		
	$L_P + L_S$	$L_P + C_E$	$L_P + C_R$	$L_P + L_S$	$L_P + C_E$	$L_P + C_R$
F-Cooper [3]	0.545	0.627	0.611	0.868	0.844	0.828
DiscoNet [15]	0.634	0.576	0.621	0.895	0.861	0.867
AttFusion [40]	0.661	0.649	0.623	0.891	0.863	0.839
CoBEVT [38]	0.692	0.638	0.660	0.934	0.860	0.865
HEAL [19]	0.770	0.659	0.682	**0.942**	0.875	0.862
STAMP [6]	0.683	0.630	0.612	0.838	0.829	0.818
CodeFilling [12]	0.706	0.671	0.662	0.926	0.882	0.877
GenComm [48]	0.642	0.641	0.641	0.936	**0.892**	**0.891**
HERO(ours)	**0.823**	**0.744**	**0.753**	0.933	0.881	0.880

Training and baselines. Each agent is trained independently in two stages: focal/SmoothL1 warm-up for stable detectors, followed by Beta-EDL [23] and NIG [1] losses for evidential semantic and geometric prediction. No cooperative training is used; at inference, the ego-side fusion engine consumes evidential proposals. We compare with feature-level baselines AttFusion [40], F-Cooper [3], DiscoNet [15], and CoBEVT [38]. We also include heterogeneous methods HEAL [19], STAMP [6], CodeFilling [12], and GenComm [48]. All methods use the same splits and pose-noise/delay protocols, with the BEV evaluation range and communication range held identical across methods within each table or curve; no method is retrained or tuned for stress settings.

4.2 Quantitative evaluation

Main performance comparison. Tab. 2 summarizes agent-scaling results: with 4 agents, HERO reaches AP50=0.922 and AP70=0.841, comparable to feature-level baselines. Tab. 3 shows two-agent results: HERO is best on DAIR-V2X and

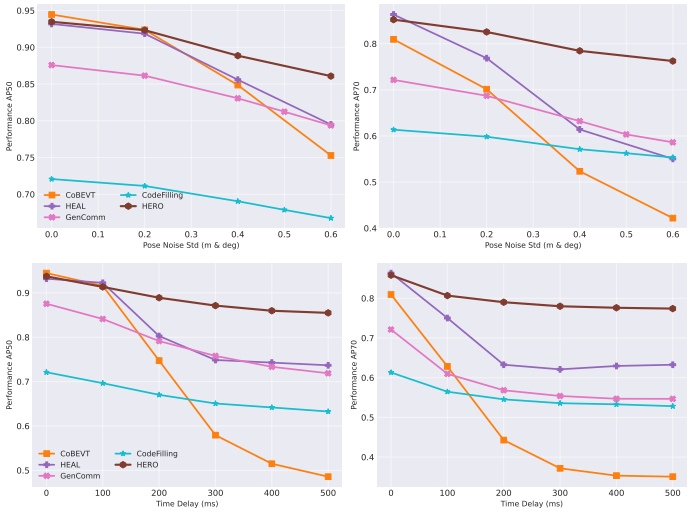

Fig. 2: Robustness against pose error and transmission delay on OPV2V-H. HERO maintains competitive accuracy as system imperfections increase, whereas baseline methods suffer severe performance degradation. Top: pose noise; bottom: transmission delay. In each row, the left plot reports AP50 and the right reports AP70; pose noise is in meters/degrees and delay in milliseconds.

remains competitive on OPV2V-H, especially for camera-related heterogeneous combinations.

Robustness to Pose Noise and Asynchronous Delay. Fig. 2 shows robustness trends under pose noise and asynchronous delay. Overall, performance degrades as noise or delay increases, while HERO degrades more gracefully. At representative severe settings, HERO achieves AP70=0.763 under pose noise level 0.6 (meters/degrees) and AP70=0.774 under 500 ms delay.

DAIR/OPV2V gap analysis. To explain why HERO gains more on DAIR-V2X, we analyze the proposal distribution. A source-selection conflict denotes an eligible cluster where semantic-evidence-only selection would choose a high-evidence but geometrically weaker source, while reliability-weighted selection suppresses it. Tab. 4 shows that such conflicts are much more common on DAIR-V2X than on OPV2V-H, and removing reliability weighting drops DAIR AP70 sharply. The pair-wise drop reflects not only conflict frequency but also the high-IoU impact of each correction.

Object-level comparison. We further compare HERO with object-level fusion baselines under the same per-agent detections and stress protocol. As shown in Tab. 5, matching and rectification improve over naive late fusion under perturbations, but HERO remains consistently higher, indicating that the gain is not solely from using an object-level interface.

Table 4: DAIR/OPV2V discrepancy analysis. ΔAP70 denotes the AP70 change when reliability weighting is removed, so negative values indicate drops. The conflict rate is the fraction of eligible clusters where reliability-aware selection overrides a high-evidence but geometrically unreliable source.

Pair	AP50/AP70	ΔAP70 w/o reliability	DAIR conflict	OPV2V-H conflict
$L_P + L_S$	0.824 / 0.704	−0.203	31.2%	1.4%
$L_P + C_E$	0.744 / 0.639	−0.108	72.6%	3.5%
$L_P + C_R$	0.753 / 0.647	−0.030	27.8%	2.8%

Table 5: Object-level comparison on OPV2V-H in AP50. All methods use the same per-agent detections. Naive late fusion applies global NMS; OptiMatch [28] adds object matching and pose correction; VIPS [26] adds graph matching, lightweight tracking-by-detection, and frame rectification. The OptiMatch and VIPS rows follow their core matching/rectification ideas rather than official full-system reproductions. The clean AP gap to Tab. 2 comes from fixed four-agent assignment here versus sequential assignment there.

Method	Clean	Pose noise		Delay	
		0.2	0.4	100 ms	200 ms
Naive late fusion	0.892	0.782	0.485	0.784	0.414
OptiMatch [28]	0.871	0.832	0.705	0.810	0.547
VIPS [26]	0.868	0.828	0.714	0.854	0.827
HERO(ours)	**0.935**	**0.923**	**0.889**	**0.913**	**0.889**

Communication Cost. Fig. 3 compares communication costs between object-level and feature-level methods. Compared to feature-level methods that exchange dense BEV features, HERO requires only about 2 KB/frame while achieving a similar AP50, over 1000× lower than dense BEV feature exchange.

4.3 Qualitative evaluation

Fig. 4 presents qualitative comparisons on the OPV2V-H dataset. Compared with STAMP, CodeFilling, and GenComm, HERO effectively captures a higher proportion of objects in challenging scenarios, leading to an increased recall of ground truth instances. Furthermore, HERO produces bounding boxes that are more closely aligned with the ground truth, with visibly smaller center and heading deviations in these examples.

4.4 Ablation studies

We conduct module ablations under the setting where noise and delay are both present. Tab. 6 reports that both delay compensation and DPR provide consistent gains; delay compensation is particularly important for high-IoU metrics.

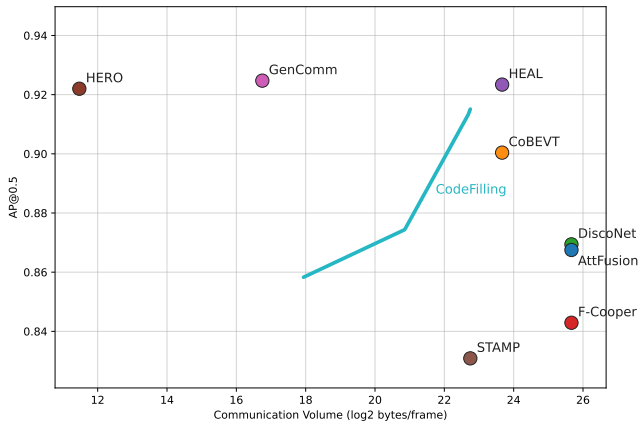

Fig. 3: Communication–accuracy trade-off on OPV2V-H. Communication is measured as the total bytes transmitted per frame: object-level methods send a sparse set of object proposals, whereas feature-level methods send dense BEV feature maps. HERO reaches comparable AP50 with far lower bandwidth.

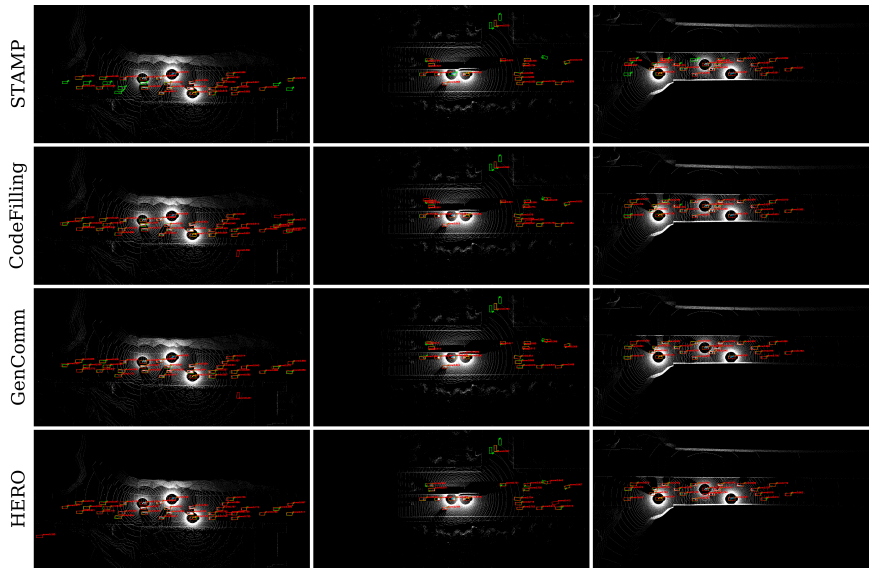

Fig. 4: Qualitative comparison on OPV2V-H. Rows correspond to STAMP, CodeFilling, GenComm, and HERO(ours); columns correspond to scenarios. Green boxes denote ground truth, and colored boxes with score labels denote model predictions.

Tab. 7 then tests the decoupled fusion rule. The two rule ablations jointly justify decoupling: combining proposals across agents helps semantics but is brittle for geometry. Beta-sum improves AP70 over score averaging, whereas mean-

box sharply degrades high-IoU accuracy and PoE only matches the stress AP70 while losing clean accuracy. Thus the beneficial operation differs between semantics and geometry, so no symmetric rule is optimal for both. The coupled max-score rule is strongest on clean data but collapses under noise and delay, while HERO's decoupled beta-sum/max-evidence rule remains stable.

Table 6: Ablation of DPR and delay compensation on OPV2V-H under pose noise 0.6 and delay 200 ms.

DPR	Delay compensation	AP50 / AP70
		0.737 / 0.661
✓		0.816 / 0.710
	✓	0.825 / 0.736
✓	✓	**0.831** / **0.737**

Table 7: Ablation of fusion rules (AP50/AP70). Clean is the ideal setting; N+D is pose noise 0.6 with 200 ms delay. For geometry, max-evidence selects a single box, mean-box and PoE average boxes, and max-score keeps the highest-scoring one; for semantics, beta-sum aggregates evidence while score-mean averages confidences, with or without reliability weighting.

(a) Bounding box fusion rules.

Geometric	Clean	N+D
max-evidence	0.902 / 0.810	**0.831** / **0.737**
mean-box	0.890 / 0.783	0.824 / 0.661
PoE	0.891 / 0.784	0.828 / 0.737
max-score	**0.906** / **0.826**	0.626 / 0.496

(b) Semantic fusion and reliability.

Semantic	Clean	N+D
beta-sum (+rel)	**0.902** / **0.810**	**0.831** / **0.737**
beta-sum (w/o rel)	0.887 / 0.804	0.832 / 0.732
score-mean (+rel)	0.854 / 0.755	0.768 / 0.673
score-mean (w/o rel)	0.846 / 0.752	0.746 / 0.648

4.5 Reliability analysis

We further validate the evidential signals that HERO relies on for association, source selection, and evidence aggregation. As summarized in Tab. 8, both semantic and geometric uncertainty act as useful reliability cues. Note that AUC_S is computed within each proposal type, so the higher camera AUC_S reflects stronger within-pool ordering, not higher overall camera reliability. These results justify using evidential signals for gating and reliability-weighted selection.

Table 8: Evidential signals as reliability cues. (a) Semantic evidence ranks proposal correctness within each type, and uncertainty-aware scoring reduces row-wise ECE. (b) NIG geometry uncertainty ranks IoU 0.7 localization success across pairs and datasets.

(a) Semantic reliability on OPV2V-H.

Proposal	S_{mean}	$u^{\text{epi}}_{\text{mean}}$	$\text{AUC}_S \uparrow$	ECE raw↓	ECE +uncert.↓
LiDAR	8.97	0.418	0.796	0.192	0.190
Camera	7.10	0.648	0.837	0.358	0.238
Fused output	–	–	–	0.298	0.222

(b) Geometry uncertainty ranking (AUC↑).

Dataset	$L_P + C_E$	$L_P + L_S$	$L_P + C_R$
DAIR-V2X	0.937	0.884	0.945
OPV2V-H	0.962	0.969	0.962

5 Conclusion

This paper proposes HERO, an object-level framework for heterogeneous collaborative 3D detection. HERO embeds interpretable evidence and uncertainty signals into each proposal, so the ego agent performs uncertainty-aware association and fusion without cross-agent collaborative training and without sharing intermediate features. This keeps communication at the KB/frame level while retaining robustness under pose noise and delay. On OPV2V-H and DAIR-V2X, HERO matches strong feature-level baselines, benefits from additional agents, and degrades more gracefully under perturbations.

Limitations. The limited representational capacity of object-level information prevents the ego vehicle from using fine-grained agent features for stronger contextual verification; under open heterogeneity and noise settings, object-level pipelines are more prone to long-tailed, low-confidence false positives, which become more evident in streaming-like per-frame decision settings. Future work may incorporate stronger confidence modeling and calibration, temporal consistency constraints, or lightweight verification mechanisms to further suppress long-tail false positives and improve reliability.

Acknowledgements

This work was supported by JST, CRONOS, Japan Grant Number JPMJCS 24K8.

References

1. Amini, A., Schwarting, W., Soleimany, A., Rus, D.: Deep evidential regression. Advances in neural information processing systems **33**, 14927–14937 (2020)
2. Bengs, V., Hüllermeier, E., Waegeman, W.: Pitfalls of epistemic uncertainty quantification through loss minimisation. Advances in Neural Information Processing Systems **35**, 29205–29216 (2022)
3. Chen, Q., Ma, X., Tang, S., Guo, J., Yang, Q., Fu, S.: F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3d point clouds. In: Proceedings of the 4th ACM/IEEE Symposium on Edge Computing. pp. 88–100 (2019)

4. Dao, M.Q., Berrio, J.S., Frémont, V., Shan, M., Héry, E., Worrall, S.: Practical collaborative perception: A framework for asynchronous and multi-agent 3d object detection. IEEE Transactions on Intelligent Transportation Systems **25**(9), 12163–12175 (2024)
5. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: Carla: An open urban driving simulator. In: Conference on robot learning. pp. 1–16. PMLR (2017)
6. Gao, X., Xu, R., Li, J., Wang, Z., Fan, Z., Tu, Z.: Stamp: Scalable task and model-agnostic collaborative perception. arXiv preprint arXiv:2501.18616 (2025)
7. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning. pp. 1321–1330. PMLR (2017)
8. Han, Y., Zhang, H., Li, H., Jin, Y., Lang, C., Li, Y.: Collaborative perception in autonomous driving: Methods, datasets, and challenges. IEEE Intelligent Transportation Systems Magazine **15**(6), 131–151 (2023)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
10. Hu, S., Fang, Z., Deng, Y., Chen, X., Fang, Y.: Collaborative perception for connected and autonomous driving: Challenges, possible solutions and opportunities. IEEE Wireless Communications (2025)
11. Hu, Y., Fang, S., Lei, Z., Zhong, Y., Chen, S.: Where2comm: Communication-efficient collaborative perception via spatial confidence maps. Advances in neural information processing systems **35**, 4874–4886 (2022)
12. Hu, Y., Peng, J., Liu, S., Ge, J., Liu, S., Chen, S.: Communication-efficient collaborative perception via information filling with codebook. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15481–15490 (2024)
13. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12697–12705 (2019)
14. Li, W., Ma, L., Chang, H., He, X., Huang, L.: Efficient collaborative perception with integrated uncertainty estimation via evidence regression. IEEE Transactions on Intelligent Transportation Systems (2025)
15. Li, Y., Ren, S., Wu, P., Chen, S., Feng, C., Zhang, W.: Learning distilled collaboration graph for multi-agent perception. Advances in Neural Information Processing Systems **34**, 29541–29552 (2021)
16. Liu, B., Teng, J., Xue, H., Wang, E., Zhu, C., Wang, P., Wu, L.: mmcooper: A multi-agent multi-stage communication-efficient and collaboration-robust cooperative perception framework. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 28396–28406 (2025)
17. Liu, S., Gao, C., Chen, Y., Peng, X., Kong, X., Wang, K., Xu, R., Jiang, W., Xiang, H., Ma, J., et al.: Towards vehicle-to-everything autonomous driving: A survey on collaborative perception. arXiv preprint arXiv:2308.16714 (2023)
18. Liu, Y.C., Tian, J., Glaser, N., Kira, Z.: When2com: Multi-agent perception via communication graph grouping. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 4106–4115 (2020)
19. Lu, Y., Hu, Y., Zhong, Y., Wang, D., Wang, Y., Chen, S.: An extensible framework for open heterogeneous collaborative perception. arXiv preprint arXiv:2401.13964 (2024)

20. Lu, Y., Li, Q., Liu, B., Dianati, M., Feng, C., Chen, S., Wang, Y.: Robust collaborative 3d object detection in presence of pose errors. arXiv preprint arXiv:2211.07214 (2022)
21. Luo, T., Yuan, Q., Luo, G., Xia, Y., Yang, Y., Li, J.: Plug and play: A representation enhanced domain adapter for collaborative perception. In: European Conference on Computer Vision. pp. 287–303. Springer (2024)
22. Mukhopadhyay, S., Roy-Chowdhury, A., Qiu, H.: Coopertrim: Adaptive data selection for uncertainty-aware cooperative perception. arXiv preprint arXiv:2602.13287 (2026)
23. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. Advances in neural information processing systems **31** (2018)
24. Shao, C., Luo, G., Yuan, Q., Chen, Y., Liu, Y., Gong, K., Li, J.: Hetecooper: Feature collaboration graph for heterogeneous collaborative perception. In: European Conference on Computer Vision. pp. 162–178. Springer (2024)
25. Shen, M., Ryu, J.J., Ghosh, S., Bu, Y., Sattigeri, P., Das, S., Wornell, G.: Are uncertainty quantification capabilities of evidential deep learning a mirage? Advances in Neural Information Processing Systems **37**, 107830–107864 (2024)
26. Shi, S., Cui, J., Jiang, Z., Yan, Z., Xing, G., Niu, J., Ouyang, Z.: Vips: Real-time perception fusion for infrastructure-assisted autonomous driving. In: Proceedings of the 28th annual international conference on mobile computing and networking. pp. 133–146 (2022)
27. Si, H., Javanmardi, E., Tsukada, M.: You share beliefs, i adapt: Progressive heterogeneous collaborative perception. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27521–27530 (2025)
28. Song, Z., Wen, F., Zhang, H., Li, J.: A cooperative perception system robust to localization errors. In: 2023 IEEE Intelligent Vehicles Symposium (IV). pp. 1–6. IEEE (2023)
29. Song, Z., Xie, T., Zhang, H., Liu, J., Wen, F., Li, J.: A spatial calibration method for robust cooperative perception. IEEE Robotics and Automation Letters **9**(5), 4011–4018 (2024)
30. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
31. Tang, Z., Liu, Y., Sun, Y., Gao, Y., Chen, J., Xu, R., Liu, S.: Cost: efficient collaborative perception from unified spatiotemporal perspective. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1120–1129 (2025)
32. Wang, T.H., Manivasagam, S., Liang, M., Yang, B., Zeng, W., Urtasun, R.: V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In: European conference on computer vision. pp. 605–621. Springer (2020)
33. Xiang, H., Xu, R., Ma, J.: 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)
34. Xu, J., Zhang, Y., Cai, Z., Huang, D.: Cosdh: communication-efficient collaborative perception via supply-demand awareness and intermediate-late hybridization. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6834–6843 (2025)
35. Xu, R., Chen, W., Xiang, H., Xia, X., Liu, L., Ma, J.: Model-agnostic multi-agent perception framework. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 1471–1478. IEEE (2023)

36. Xu, R., Guo, Y., Han, X., Xia, X., Xiang, H., Ma, J.: Opencda: an open cooperative driving automation framework integrated with co-simulation. In: 2021 IEEE International Intelligent Transportation Systems Conference (ITSC). pp. 1155–1162. IEEE (2021)
37. Xu, R., Li, J., Dong, X., Yu, H., Ma, J.: Bridging the domain gap for multi-agent perception. arXiv preprint arXiv:2210.08451 (2022)
38. Xu, R., Tu, Z., Xiang, H., Shao, W., Zhou, B., Ma, J.: Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers. arXiv preprint arXiv:2207.02202 (2022)
39. Xu, R., Xiang, H., Tu, Z., Xia, X., Yang, M.H., Ma, J.: V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In: European conference on computer vision. pp. 107–124. Springer (2022)
40. Xu, R., Xiang, H., Xia, X., Han, X., Li, J., Ma, J.: Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 2583–2589. IEEE (2022)
41. Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. Sensors **18**(10), 3337 (2018)
42. Yang, D., Yang, K., Wang, Y., Liu, J., Xu, Z., Yin, R., Zhai, P., Zhang, L.: How2comm: Communication-efficient and collaboration-pragmatic multi-agent perception. Advances in Neural Information Processing Systems **36**, 25151–25164 (2023)
43. Yazgan, M., Graf, T., Liu, M., Fleck, T., Zöllner, J.M.: A survey on intermediate fusion methods for collaborative perception categorized by real world challenges. In: 2024 IEEE Intelligent Vehicles Symposium (IV). pp. 2226–2233. IEEE (2024)
44. Yu, H., Luo, Y., Shu, M., Huo, Y., Yang, Z., Shi, Y., Guo, Z., Li, H., Hu, X., Yuan, J., et al.: Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21361–21370 (2022)
45. Zhang, J., Yang, K., Wang, Y., Wang, H., Sun, P., Song, L.: Ermvp: Communication-efficient and collaboration-robust multi-vehicle perception in challenging environments. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12575–12584 (2024)
46. Zhao, L., Malikopoulos, A.A.: Enhanced mobility with connectivity and automation: A review of shared autonomous vehicle systems. IEEE Intelligent Transportation Systems Magazine **14**(1), 87–102 (2020)
47. Zhao, Z., Zhao, C., Chen, K., Ji, Y., Du, Y.: Belt-fusion: Bayesian evidential late fusion for trustworthy v2x perception. IEEE Transactions on Intelligent Transportation Systems **27**(1), 725–741 (2025)
48. Zhou, J., Dai, P., Wei, Q., Liu, B., Wu, X., Wang, J.: Pragmatic heterogeneous collaborative perception via generative communication mechanism. arXiv preprint arXiv:2510.19618 (2025)