QACE: QUALITY-AWARE COLLABORATIVE EVIDENCE FOR SEMI-SUPERVISED COLLABORATIVE 3D DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Collaborative 3D perception can effectively enhance the perception capability of a single agent through information sharing, yet at the same time incurs expensive multi-agent joint annotation costs. Semi-supervised learning is a promising label-efficient solution, but it still faces two unique challenges in pseudo-labels, namely cross-agent observation inequality and fine-grained attribute disagreement; specifically, a missing proposal is informative only when the corresponding agent had sufficient opportunity to observe the object, and agreement on object presence does not imply equal reliability across box attributes. To address these issues, we propose QACE, a novel semi-supervised collaborative 3D detection framework that qualifies collaborative evidence before assigning pseudo-supervision. The Validated Collaborative Evidence (VCE) qualifies support, conflict, and silence using sensing opportunity and role-specific proposal checks. The Evidence-Conditioned Attribute Trainability (EAT) converts this evidence into separate weights for center, size, yaw axis, and directed heading within candidate-level supervision permissions. Extensive experiments on OPV2V and DAIR-V2X demonstrate that QACE achieves state-of-the-art performance across different detectors and annotation budgets, highlighting its effectiveness. The code and model will be publicly available.

1 INTRODUCTION

Collaborative perception (Chen et al., 2019; Wang et al., 2020; Li et al., 2021; Xu et al., 2022a; Hu et al., 2022) effectively addresses the limitations of single-agent perception, such as limited perception range and sensitivity to occlusion, through information interaction among multiple agents such as vehicles and roadside infrastructure. However, it comes with expensive annotation costs, as collaborative perception requires multiple agents to perform joint synchronous annotation, a process that is complex and labor-intensive (Xu et al., 2022b; Yu et al., 2022).

Semi-supervised learning (Zhao et al., 2020; Wang et al., 2021; Yin et al., 2022; Liu et al., 2023) is a promising label-efficient solution, as it can fully leverage the information contained in large amounts of unlabeled data through pseudo-labeling. However, existing single-agent 3D semi-supervised methods cannot be well adapted directly to collaborative perception scenarios, and research on collaborative semi-supervised 3D perception methods remains largely unexplored.

In collaborative scenes, several agents often observe the same object, so their proposals can cross-check pseudo-labels. A simple vote keeps a pseudo-box when enough agents detect it, and existing pseudo-label selection judges each box as a whole (Wang et al., 2021; Liu et al., 2023; Han et al., 2025). Both overlook two properties of this evidence. First, a missing proposal can count against an object only when its agent had sufficient sensing opportunity to observe it. The vehicle in Fig. 1a receives 2,604, 0, and 70 LiDAR points from three agents, yet a vote would count the second agent's missing proposal against it. Second, agents may agree on an object's center but not on its size or heading (Fig. 1b), so accepting or rejecting the whole box either retains uncertain attributes or discards a useful center target.

Building on the above insights, we attribute these unique challenges to cross-agent observation inequality and fine-grained attribute disagreement and propose QACE, a novel semi-supervised col-

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.

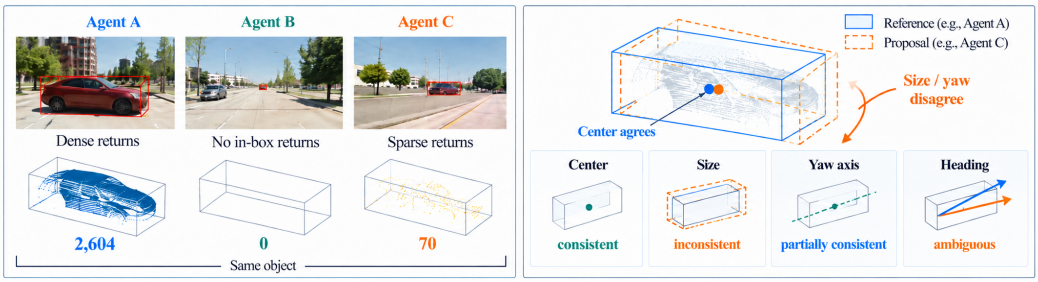

Figure 1: **Ambiguous evidence in collaborative pseudo-labeling.** (a) Unequal observation across agents. (b) Fine-grained attribute disagreement.

laborative 3D detection framework that qualifies collaborative evidence before assigning pseudo-supervision. Specifically, Validated Collaborative Evidence (VCE) uses sensing opportunity and role-specific proposal checks to distinguish support, conflict, silence, and abstention for each candidate–view pair. Views with insufficient opportunity or ambiguous correspondence provide no direct counter-evidence against the candidate. Evidence-Conditioned Attribute Trainability (EAT) then converts measured agreement and qualified conflict into separate supervision weights for center, size, yaw axis, and directed heading. Within the permissions established by candidate admission, EAT can retain reliable center supervision while reducing or disabling supervision on uncertain size or heading. Both modules operate only during training and add no inference cost. Extensive experiments on OPV2V and DAIR-V2X demonstrate that QACE achieves state-of-the-art performance across different detectors and annotation budgets, highlighting its effectiveness.

In summary, the main contributions of this paper are as follows:

- We propose QACE, the first semi-supervised collaborative 3D detection framework that links qualified cross-agent evidence to attribute-specific pseudo-supervision.
- We develop VCE, which weights support, conflict, and silence by sensing opportunity while excluding unqualified views from direct evidence.
- We design EAT, which separately weights center, size, yaw axis, and directed heading for regression-eligible candidates, so that a reliable center can retain supervision when other attributes fail their respective gates.
- Extensive experiments across multiple datasets, detectors, and annotation budgets demonstrate that QACE achieves state-of-the-art performance, highlighting its effectiveness.

2 Related Work

2.1 Collaborative 3D Detection

Collaborative detectors combine multi-agent features through pooling, learned collaboration weights, and attention (Chen et al., 2019; Wang et al., 2020; Li et al., 2021; Xu et al., 2022a). Where2comm (Hu et al., 2022) and CoSDH (Xu et al., 2025) select spatial regions for communication, while DATA (Tian et al., 2025) uses common observability to weight feature alignment. QACE applies a related idea during training, using each view's sensing opportunity to decide whether the view may act as pseudo-label evidence.

Label-efficient methods attempt to address the cost of annotating cooperative scenes. SSC3OD (Han et al., 2023) combines self-supervised pretraining with instance mining, and CoDTS (Han et al., 2025) recovers missing foreground annotations. R&B-POP (Yoo et al., 2025) trains an ego detector from nearby agents' predictions and handles viewpoint mismatch. QACE learns a fused cooperative detector jointly from fully annotated scenes and unlabeled multi-agent observations.

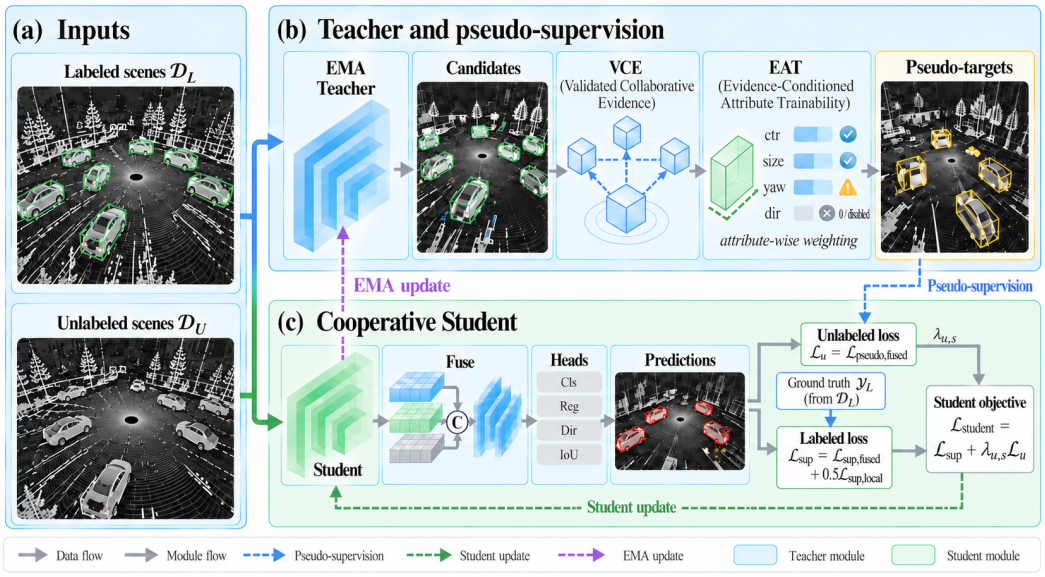

Figure 2: **QACE training pipeline.** VCE qualifies cross-agent evidence, and EAT assigns attribute-specific supervision weights to the resulting candidates. The detector learns from these pseudo-targets and fully annotated scenes.

2.2 Semi-Supervised 3D Detection

Semi-supervised 3D detection uses unlabeled observations through pseudo-label selection and consistency learning. SESS (Zhao et al., 2020) regularizes object properties, 3DIoUMatch (Wang et al., 2021) filters boxes by predicted IoU, and HSSDA (Liu et al., 2023) combines hierarchical supervision with shuffle augmentation. DQS3D (Gao et al., 2023) aligns dense predictions under transformed views, whereas NoiseDet (Chen et al., 2023) models noisy regression targets. Sparse-label methods such as SS3D (Liu et al., 2022) and HINTED (Xia et al., 2024) instead mine missing objects within partially annotated scenes.

Pseudo-label reliability also varies across tasks and box components. Soft Teacher (Xu et al., 2021) selects regression targets using box jittering, and Side-aware learning (Wang et al., 2023) estimates uncertainty for individual box faces. UA3D (Zhang et al., 2025) weights box coordinates by disagreement between detection branches in unsupervised 3D detection, while DPL (Zhang et al., 2024) separates 2D and 3D pseudo-label generation. QACE relates attribute supervision to agreement across physical sensing views, using sensing opportunity and role-specific proposal checks to qualify counter-evidence.

3 Methodology

3.1 Problem Setting and Overview

Let $\mathcal{D}_L$ and $\mathcal{D}_U$ denote labeled and unlabeled collaborative scenes. Each scene contains multi-agent LiDAR sweeps and the relative poses used to align them. PointPillars (Lang et al., 2019) encodes each sweep into BEV features, which are aligned, fused, and decoded into boxes. The detector θ learns from ground-truth boxes on $\mathcal{D}_L$ and pseudo-targets supplied by an exponential moving average (EMA) teacher $\bar{\theta}$ (Tarvainen & Valpola, 2017) on $\mathcal{D}_U$.

Aligned per-agent proposals, produced with shared parameters from different sensing inputs, are grouped and supplemented with fused detections (Section 3.2). We call an agent's observation of a candidate a *view*. VCE qualifies the evidence from each view (Section 3.3), and EAT assigns weights to $\mathcal{A} = \{\text{ctr}, \text{size}, \text{yaw}, \text{dir}\}$ (Section 3.4). Here, yaw denotes footprint orientation modulo

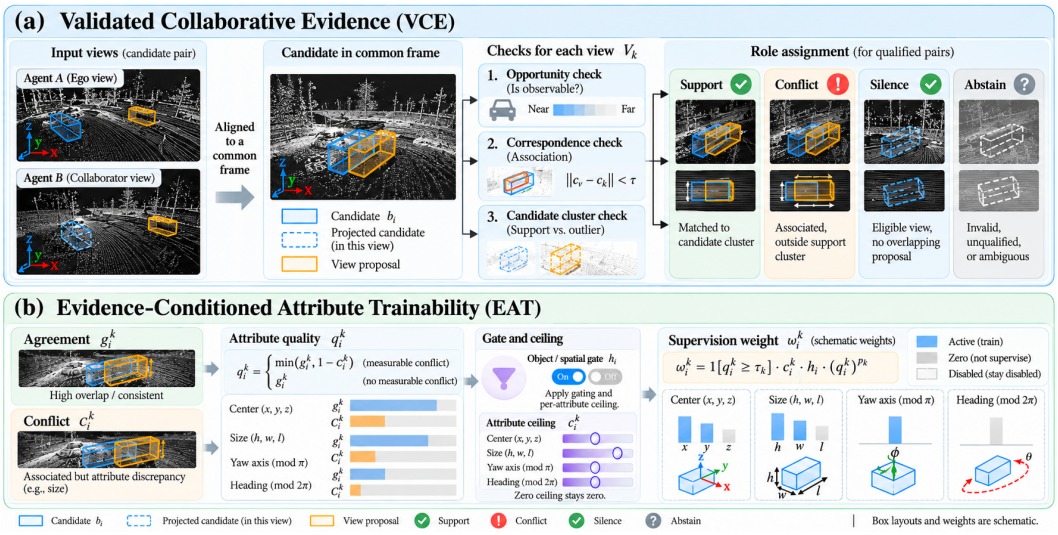

Figure 3: **View qualification and attribute-wise supervision.** (a) VCE qualifies view evidence using sensing opportunity and role-specific proposal checks. (b) EAT assigns attribute weights within candidate-level supervision ceilings using measured agreement and qualified conflict.

π; directed heading resolves front–back ambiguity modulo 2π. Fig. 2 summarizes the pipeline, and Algorithm 1 gives one training update.

3.2 Candidate Construction

Per-agent proposals are decoded at a relaxed score threshold and filtered by rotated non-maximum suppression (NMS). Aligned proposals are clustered in the common frame, retaining at most one proposal per agent. Centers and dimensions use score-weighted means, while headings use a circular mean. Score quantiles set adaptive thresholds with momentum across three pseudo-epoch stages.

Clusters supported by at least two agents form *consensus* candidates, subject to VCE admission. Single-agent (*weak*) candidates retain their source identity, and fused detections supplement the local clusters. Weak and fused-only candidates provide classification targets but lack the cross-agent measurements required for EAT regression.

After warm-up, an auxiliary head uses statistics of provisional candidates to predict object probability and XY center error. Its admissions set initial supervision ceilings; retained candidates remain subject to VCE and EAT (Appendix A.4).

3.3 Validated Collaborative Evidence

VCE determines whether a missing or conflicting proposal can provide counter-evidence. Sensing opportunity depends on the candidate's position and local point support. Conflict additionally requires an unambiguous proposal association; silence requires an eligible view with no overlapping proposal. These opportunity measures approximate sensing conditions without directly estimating free space or detection probability.

Sensing opportunity. For candidate i and agent a, $v_{ia} \in \{0, 1\}$ checks whether the candidate center is in range and the pose and timestamp are valid. A view with $v_{ia} = 0$ abstains. We normalize local position as $\rho_{ia} = \max\{|x_{ia}|/R_x, |y_{ia}|/R_y\}$, using the horizontal half-ranges R_x and R_y. Let n_{ia} count points represented by voxels whose centers fall within the candidate's rotated BEV footprint. Voxels span the full height range, so the count covers the entire footprint column, including points outside the object. With scale $\kappa > 0$, $\mathcal{F}_{ia}$ contains v_{ia}, $e^{-\rho_{ia}}$, and, when measured,

Table 1: **VCE evidence roles.** Each candidate–view pair receives one role. Non-supporter eligibility requires a measured count and $o_{ia} \geq \tau_{\text{elig}}$. Conflict requires unambiguous association; silence requires no overlapping proposal and preserved observability. The subset $\mathcal{M}_i$ additionally satisfies $o_{ia} \geq \tau_{\text{sil}}$.

Role	Condition	Used for
Matched support ($\mathcal{S}_i$)	Valid view with a proposal matched to the candidate cluster	Support S_i; attribute agreement g_i^k
Conflict ($\mathcal{N}_i$)	Eligible; associated overlapping proposal outside the support cluster	Denominator of S_i; conflict C_i^k
Silence ($\mathcal{Z}_i \supseteq \mathcal{M}_i$)	Eligible; no overlapping proposal; silence observable	Denominator of S_i; if $o_{ia} \geq \tau_{\text{sil}}$, silence M_i
Abstention	Invalid view, unqualified non-supporter, or ambiguous correspondence	No direct vote; computable opportunity still enters D_i

$1 - e^{-n_{ia}/\kappa}$. Their geometric mean defines sensing opportunity,

$$o_{ia} = \Big(\prod_{f \in \mathcal{F}_{ia}} f \Big)^{1/|\mathcal{F}_{ia}|}. \tag{1}$$

A measured zero gives zero opportunity. When a count is unavailable, matched supporters omit its factor; non-supporting views require a measured count to provide counter-evidence.

Evidence roles. Let $\mathcal{V}_i$ contain views with computable opportunity and $\mathcal{S}_i$ contain matched supporters. Table 1 defines the four roles, illustrated in Figure 3a. Non-supporters require $v_{ia} = 1$, a measured count, and $o_{ia} \geq \tau_{\text{elig}}$ for eligibility. Eligible views with unambiguously associated, non-supporting proposals form $\mathcal{N}_i$; eligible views without overlapping proposals form $\mathcal{Z}_i$. Other views abstain, giving $\mathcal{E}_i = \mathcal{S}_i \cup \mathcal{N}_i \cup \mathcal{Z}_i$. The stricter subset $\mathcal{M}_i = \{a \in \mathcal{Z}_i : o_{ia} \geq \tau_{\text{sil}}\}$ supplies the silence term, with $\tau_{\text{sil}} > \tau_{\text{elig}}$.

Non-supporting views require at least three represented points to enter $\mathcal{V}_i$. When an overlapping proposal is present, association requires an XY center distance no larger than the candidate's footprint diagonal. A near-tie between the two nearest matches, within the association ambiguity gap, causes abstention. These checks distinguish disagreement about a candidate from predictions of neighboring objects. Eligible views without overlapping proposals enter silence when their in-range points remain observable after augmentation.

With total computable opportunity $D_i = \sum_{a \in \mathcal{V}_i} o_{ia}$, support and silence are

$$S_i = \frac{\sum_{a \in \mathcal{S}_i} o_{ia}}{\sum_{a \in \mathcal{E}_i} o_{ia}}, \qquad M_i = \frac{\sum_{a \in \mathcal{M}_i} o_{ia}}{D_i}. \tag{2}$$

Support uses eligible opportunity, whereas silence and conflict use total computable opportunity D_i. Abstaining views contribute no direct evidence but retain computable opportunity in D_i. They can therefore dilute M_i and C_i^k without changing S_i. Support and silence are set to zero when their respective denominators vanish.

For an associated but unmatched proposal, normalized discrepancies $r_{ia}^k \in [0, 1]$ measure center, size, yaw-axis, and directed-heading disagreement (Appendix A.2). When $\mathcal{N}_i \neq \emptyset$ and $D_i > 0$, the conflict measure is

$$C_i^k = \frac{\sum_{a \in \mathcal{N}_i} o_{ia} r_{ia}^k}{D_i}. \tag{3}$$

Conflict measures the share of computable opportunity carrying disagreement. An empty $\mathcal{N}_i$ means conflict is unavailable; an associated view can instead supply a measured zero for an agreeing attribute.

Use of the evidence. Support and silence determine candidate admission. Let G_i summarize supporter agreement in BEV IoU, center, and yaw, and let $\bar{o}_i$ be mean view opportunity. A consensus candidate is admitted when

$$S_i \geq \tau_S, \qquad M_i \leq \tau_M, \qquad G_i \geq \tau_G, \tag{4}$$

Any candidate, including weak and fused-only candidates, is vetoed when $M_i \geq \tau_{\text{veto}}$ and $\bar{o}_i \geq \tau_{\bar{o}}$. Candidate reliability η_i combines teacher confidence, support S_i, cluster geometry G_i, and the candidate-source prior. Reliability sets the initial classification and regression ceilings, while qualified conflict constrains the attribute qualities used by EAT.

3.4 Evidence-Conditioned Attribute Trainability

EAT refines supervision within the permissions established by candidate admission (Figure 3b). Candidate-level geometry and reliability determine whether regression is allowed; attribute-level gates then retain, reduce, or disable each permitted component. Thus, an admitted, regression-eligible candidate can preserve center supervision while size or heading supervision is disabled.

Attribute agreement. For each supporting proposal j, let $\mathbf{p}_j$ denote its center, l_j, w_j its footprint dimensions, and ψ_j its heading. Let $\bar{\mathbf{p}}$ be the mean center and CV_d the coefficient of variation of dimension $d \in \{\text{length}, \text{width}, \text{height}\}$. The positive scales σ_c and σ_s control sensitivity to center and size dispersion; ϵ controls the elongation penalty. Attribute agreement is

$$\begin{aligned} g_i^{\text{ctr}} &= \exp\big[-\operatorname{mean}_j \|\mathbf{p}_j - \bar{\mathbf{p}}\|/\sigma_c\big], \qquad g_i^{\text{size}} = \exp\big[-\max_d \mathrm{CV}_d\,/\sigma_s\big], \\ g_i^{\text{yaw}} &= \big|\operatorname{mean}_j e^{2\mathrm{i}\psi_j}\big| \, \min\Big\{\tfrac{1}{\epsilon}\operatorname{mean}_j \tfrac{|l_j - w_j|}{l_j + w_j},\, 1\Big\}, \qquad g_i^{\text{dir}} = \big|\operatorname{mean}_j e^{\mathrm{i}\psi_j}\big|. \end{aligned} \tag{5}$$

The double angle separates yaw-axis agreement from front–back ambiguity; the elongation factor downweights nearly square footprints. These scores measure agreement among views, not the probability that an attribute is correct. Single-supporter candidates lack measured agreement and receive no geometric supervision.

Attribute quality. For an attribute with measured agreement, the evidence-conditioned quality is

$$q_i^k = \begin{cases} \min\{g_i^k,\, 1 - C_i^k\}, & \text{if conflict is available,} \\ g_i^k, & \text{otherwise.} \end{cases} \tag{6}$$

Equation 6 caps measured agreement at $1 - C_i^k$ when conflict is available. Regression eligibility requires measured supporter agreement.

Supervision weights. The shared gate is $h_i = \mathbf{1}[Q_i \geq \tau_Q]\,\mathbf{1}[\max\{q_i^{\text{ctr}}, g_i^{\text{assoc}}\} \geq \tau_{\text{sp}}]$. It requires both object quality Q_i and spatial support from center or association quality. Appendix A.3 defines these source-dependent qualities. When center agreement is unavailable, association quality supplies the spatial input.

For reliable consensus candidates, candidate construction sets classification ceiling $c_i^{\text{cls}} = \eta_i$ and regression ceiling η_i^2, with $\eta_i \geq \tau_\eta$. Intermediate-reliability, weak, and fused-only candidates remain classification-only. Auxiliary candidates inherit the head's admission ceilings (Appendix A.4). Each attribute ceiling c_i^k retains earlier caps; EAT cannot restore a zero ceiling. With thresholds τ_k and exponents $p_k > 0$,

$$\omega_i^{\text{cls}} = c_i^{\text{cls}}\, h_i, \qquad \omega_i^k = \mathbf{1}[q_i^k \geq \tau_k]\; c_i^k\, h_i\, (q_i^k)^{p_k}. \tag{7}$$

Supporting headings differing by nearly π set $\omega_i^{\text{dir}} = 0$ for the candidate. Yaw-axis supervision can remain active when the candidate's aggregated heading is compatible with the agreed axis (Appendix A.2). Center and size follow their own gates within the existing ceilings.

3.5 Training Objective

All anchors assigned to a pseudo-box inherit its box parameters and supervision weights. Classification uses the focal loss (Lin et al., 2017), and EAT weights scale the corresponding center, size, yaw, and direction losses. The sine-difference yaw loss is invariant to a π flip; the direction classifier resolves front–back orientation. Classification-only targets contribute to objectness learning while their geometric losses remain disabled. Box, classification, and direction losses are normalized by their respective counts of valid positive anchors.

Background targets use anchors with very low fused-teacher object probability, excluding neighborhoods of positive, ignored, or budget-dropped candidates. Scenes without positive pseudo-anchors receive no background supervision (Appendix A.3).

Table 2: **Quantitative comparison.** Oriented BEV AP (%) with 10% or 20% labeled training scenes. Within each fusion group and column, bold denotes the best result and underlining the second best. All baselines are adapted to the cooperative scene-level protocol.

	OPV2V				DAIR-V2X			
	10%		20%		10%		20%	
Method	AP@0.5	AP@0.7	AP@0.5	AP@0.7	AP@0.5	AP@0.7	AP@0.5	AP@0.7
F-Cooper								
3DIoUMatch	81.58	58.04	84.36	67.72	62.79	40.55	69.36	48.54
CoDTS	81.57	61.51	84.95	65.73	63.66	40.77	70.06	49.94
HSSDA	82.13	58.95	83.72	62.41	53.19	31.12	69.38	47.53
HINTED	81.90	56.78	86.41	67.99	56.50	37.70	63.88	45.48
QACE (ours)	**86.98**	**67.47**	**88.48**	**70.69**	**77.89**	**58.06**	**80.27**	**60.76**
AttFusion								
3DIoUMatch	86.02	74.45	88.53	78.55	59.97	45.06	65.25	50.26
CoDTS	88.71	75.66	90.17	78.88	61.57	46.80	66.74	51.22
HSSDA	88.75	73.52	91.08	77.35	60.62	44.05	67.06	52.17
HINTED	90.96	76.91	92.16	79.45	57.78	43.78	65.30	50.15
QACE (ours)	**94.25**	**82.47**	**94.92**	**84.73**	**73.20**	**59.06**	**75.41**	**61.48**
DiscoNet								
3DIoUMatch	86.75	73.78	89.01	78.21	62.47	46.93	65.68	52.44
CoDTS	88.02	75.00	90.42	78.99	62.54	47.09	67.47	52.85
HSSDA	89.24	73.24	90.35	76.96	60.36	45.24	65.00	51.23
HINTED	90.18	75.14	91.61	79.22	55.75	41.47	64.63	50.94
QACE (ours)	**93.67**	**81.92**	**94.39**	**84.40**	**73.77**	**59.55**	**75.03**	**61.28**

The detector is initialized from a supervised checkpoint. Let $\mathcal{L}_{\text{sup,fused}}$ and $\mathcal{L}_{\text{sup,local}}$ denote the fused and individual-agent losses on labeled scenes, and $\mathcal{L}_{\text{pseudo,fused}}$ the fused loss on pseudo-targets. At update s,

$$\mathcal{L}_{\text{student}} = \mathcal{L}_{\text{sup,fused}} + \lambda_{\text{single}}\mathcal{L}_{\text{sup,local}} + \lambda_{u,s}\mathcal{L}_{\text{pseudo,fused}}, \tag{8}$$

where $\lambda_{\text{single}} = 0.5$. The scheduled upper bound on $\lambda_{u,s}$ increases from 0.1 to 0.5 over the first eight pseudo-epochs. The coefficient is further capped so that the weighted pseudo-loss does not exceed half the detached fused supervised detection loss. Appendix A.5 gives the exact coefficient and EMA schedule.

Each update pairs one labeled scene with one unlabeled scene. Training runs for 30 pseudo-epochs, each defined by a fixed update budget over part of the unlabeled list. Pseudo-targets and their weights remain fixed during each gradient update. Shared reflection, rotation, and scaling transform every agent's observations and relative pose consistently. Negative silence is observable only when the augmentation preserves the view's in-range points within the region of interest. Inference uses detector θ without VCE, EAT, the auxiliary head, or the EMA teacher.

4 Experiments

4.1 Setup

Datasets. OPV2V (Xu et al., 2022b) contains simulated vehicle-to-vehicle scenes; evaluation uses its 2,170-scene test split. DAIR-V2X (Yu et al., 2022) contains real vehicle–infrastructure observations; evaluation uses 1,711 unique scenes from the public validation split. We use *scene* to denote one cooperative frame throughout.

Protocol and metrics. The primary comparisons use 10% or 20% labeled training scenes; an additional OPV2V/F-Cooper experiment uses 5%. The remaining training scenes are unlabeled, excluding a ten-frame guard around each labeled block (Appendix A.1). We evaluate F-Cooper (Chen et al., 2019), AttFusion (Xu et al., 2022b), and DiscoNet (Li et al., 2021). We report AP@0.5 and

AP@0.7 using one-to-one polygon-IoU matching, global score ranking, and the all-point precision envelope. QACE inference uses a 0.2 classification-score threshold and a 0.15 rotated-NMS IoU threshold. Appendix B describes adaptations of 3DIoUMatch, CoDTS, HSSDA, and HINTED.

Comparison protocol. Within a comparison group, methods share the detector, fusion operator, scene split, and AP implementation. Their pseudo-labeling and training procedures remain method-specific. The main table therefore compares complete systems. Section 4.3 varies VCE and EAT within QACE; Appendix C reports the 5% label budget and pseudo-label threshold comparisons.

4.2 Main Results

Table 2 compares QACE with four adapted baselines in 12 dataset–fusion–budget settings. All margins below use the strongest competing value in the same column and are expressed in percentage points.

On OPV2V, QACE achieves the highest AP at both IoU thresholds for each fusion operator and annotation budget. At 10% labels, AP@0.7 margins are 5.96, 5.56, and 6.78 points for F-Cooper, AttFusion, and DiscoNet, respectively. At 20%, the corresponding margins are 2.70, 5.28, and 5.18 points. On DAIR-V2X, QACE also achieves the highest AP at both IoU thresholds. With 10% labels, AP@0.7 is 58.06, 59.06, and 59.55 for F-Cooper, AttFusion, and DiscoNet. The margins over the strongest adapted baselines are 17.29, 12.26, and 12.46 points, respectively. With 20% labels, the margins range from 8.43 to 10.82 points.

Overlap threshold and annotation budget. The AP@0.7 margin exceeds the AP@0.5 margin in all 12 settings, using the strongest baseline separately for each metric. AP combines localization, candidate retention, and score ranking; the larger margins show an advantage under stricter overlap matching. Increasing labels from 10% to 20% improves QACE's AP@0.7 for every fusion operator on both datasets. Gains range from 2.26 to 3.22 points on OPV2V and from 1.73 to 2.70 points on DAIR-V2X.

4.3 Ablation Study

The component comparisons use OPV2V/F-Cooper with 10% labels and keep candidate construction, the auxiliary head, and optimization fixed. Table 3 compares VCE and EAT configurations. Without VCE, admission uses at least two matched agents, a matched fraction of at least 0.5 among in-range agents, and $G_i \geq \tau_G$. Silence and conflict do not enter admission, the veto, or EAT. The auxiliary head receives unweighted vote-based counterparts of the VCE features, preserving its input dimension. Without EAT, every regression-eligible candidate supervises all four attributes at its object-level ceiling.

Table 3: **VCE/EAT component comparisons** on OPV2V/F-Cooper with 10% labels. Values are oriented BEV AP (%); Δ is relative to both modules disabled.

VCE	EAT	AP@0.5 ↑	AP@0.7 ↑	ΔAP@0.5	ΔAP@0.7
		83.94	61.90	–	–
✓		84.72	63.66	+0.78	+1.76
	✓	84.39	64.25	+0.45	+2.35
✓	✓	**86.98**	**67.47**	**+3.04**	**+5.57**

Component comparisons. The joint configuration achieves the highest AP@0.7, at 67.47. Adding EAT to VCE gives 3.81 points; adding VCE to EAT gives 3.22 points. Each module improves AP@0.7 when added to the other in this setting.

Table 4 varies evidence weighting and supervision within the full system. Uniform voting keeps the eligibility rules but assigns $o_{ia} = 1$ to eligible views. The variant without silence removes the silence term from candidate admission and the veto while retaining conflict evidence. Removing the conflict role reassigns those views to abstention, changing the denominator of S_i and making C_i^k

unavailable. The shared-minimum variant replaces each attribute quality with $\min_k q_i^k$, retaining the attribute-specific threshold and exponent.

Table 4: **Design comparisons** on OPV2V/F-Cooper with 10% labels. Values are oriented BEV AP (%); Δ is full QACE minus the variant.

Variant	AP@0.5 ↑	AP@0.7 ↑	ΔAP@0.5	ΔAP@0.7
VCE evidence aggregation				
Uniform voting ($o_{ia} = 1$)	84.02	63.31	+2.96	+4.16
Without silence in admission and veto (no $\mathcal{M}_i$)	85.92	66.19	+1.06	+1.28
No conflict role (no $\mathcal{N}_i$)	86.31	65.52	+0.67	+1.95
EAT attribute supervision				
Shared minimum quality ($\min_k q_i^k$)	85.23	64.71	+1.75	+2.76
Unified yaw and heading	86.72	66.41	+0.26	+1.06
No conflict cap ($q_i^k = g_i^k$)	86.61	65.81	+0.37	+1.66
Complete QACE	**86.98**	**67.47**	–	–

VCE design. Uniform voting gives 63.31 AP@0.7, compared with 67.47 for the full model. Removing silence or the conflict role gives 66.19 and 65.52 AP@0.7, respectively. The conflict-role ablation affects candidate admission and attribute supervision together; the conflict-cap ablation below keeps the VCE roles fixed.

EAT design. Sharing the minimum attribute quality gives 64.71 AP@0.7, a 2.76-point gap from the full model. Merging yaw axis with directed heading gives 66.41 AP@0.7. Removing the conflict cap while retaining VCE's evidence roles gives 65.81.

4.4 Visualization

Fig. 4 compares CoDTS and QACE on four selected OPV2V scenes with F-Cooper and 10% labeled training scenes. Both methods use the same score and NMS thresholds. At BEV IoU 0.7, matched-box counts (CoDTS/QACE) are 7/13, 3/6, 8/10, and 17/19. Additional four-method visualizations appear in Appendix D.

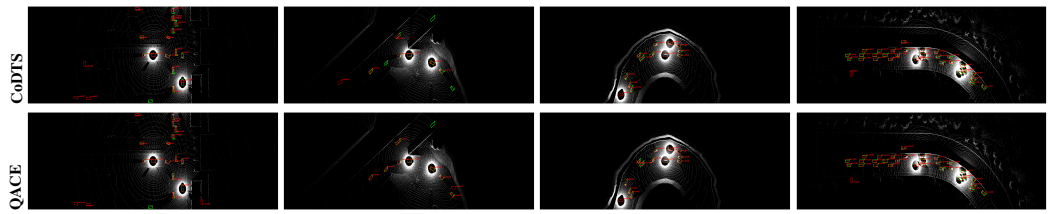

Figure 4: **Qualitative detection results on OPV2V.** Columns: test indices 1320, 1800, 360, 480. Rows: CoDTS, QACE. Green/red boxes: annotations/predictions. Both models use score threshold 0.2 and NMS IoU threshold 0.15. Full-scene panels use a common $1.35\times$ vertical display magnification; IoU is computed in the original BEV coordinates.

5 Conclusion

In this paper, we propose QACE, a framework for semi-supervised collaborative 3D detection with attribute-specific pseudo-supervision. VCE qualifies cross-agent evidence using sensing opportunity and proposal correspondence. EAT weights box attributes within the permissions set during candidate admission. Extensive experiments on OPV2V and DAIR-V2X demonstrate that QACE achieves state-of-the-art performance, while OPV2V ablations favor using both modules. Future work will explore multimodal collaborative perception with richer pseudo-label evidence.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing by improving the grammar, wording, and readability of the manuscript and providing critical feedback on the clarity of the methodology description and experimental presentation. The research ideas, methodology, implementation, experiments, data, and results were developed and produced by the authors. We reviewed, verified, and revised all AI-assisted textual edits. We take responsibility for the final content of this work, including its technical claims and experimental results.

REPRODUCIBILITY STATEMENT

Section 3 describes evidence roles, admission conditions, attribute agreement, supervision weights, and the training objective. Appendix A reports the detector, scene partitions, update sequence, and selected method settings. Appendix B summarizes the baseline adaptations. Furthermore, after this paper is accepted, we will publicly release the source code and models.

REFERENCES

Qi Chen, Xu Ma, Sihai Tang, Jingda Guo, Qing Yang, and Song Fu. F-Cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3D point clouds. In *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing*, pp. 88–100, 2019. doi: 10.1145/3318216.3363300.

Zehui Chen, Zhenyu Li, Shuo Wang, Dengpan Fu, and Feng Zhao. Learning from noisy data for semi-supervised 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6929–6939, 2023.

Huan-ang Gao, Beiwen Tian, Pengfei Li, Hao Zhao, and Guyue Zhou. DQS3D: Densely-matched Quantization-aware Semi-supervised 3D Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21905–21915, October 2023.

Yushan Han, Hui Zhang, Honglei Zhang, and Yidong Li. SSC3OD: Sparsely supervised collaborative 3D object detection from LiDAR point clouds. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2023.

Yushan Han, Hui Zhang, Honglei Zhang, Jing Wang, and Yidong Li. CoDTS: Enhancing sparsely supervised collaborative perception with a dual teacher-student framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3366–3373, 2025.

Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 4874–4886, 2022.

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

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

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Chuandong Liu, Chenqiang Gao, Fangcen Liu, Jiang Liu, Deyu Meng, and Xinbo Gao. SS3D: Sparsely-supervised 3D object detection from point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8428–8437, 2022.

Chuandong Liu, Chenqiang Gao, Fangcen Liu, Pengcheng Li, Deyu Meng, and Xinbo Gao. Hierarchical supervision and shuffle data augmentation for 3D semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23819–23828, 2023.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Chengchang Tian, Jianwei Ma, Yan Huang, Zhanye Chen, Honghao Wei, Hui Zhang, and Wei Hong. DATA: Domain-And-Time Alignment for High-Quality Feature Fusion in Collaborative Perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28643–28652, October 2025.

Chuxin Wang, Wenfei Yang, and Tianzhu Zhang. Not every side is equal: Localization uncertainty estimation for semi-supervised 3D object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3814–3824, 2023.

He Wang, Yezhen Cong, Or Litany, Yue Gao, and Leonidas J. Guibas. 3DIoUMatch: Leveraging IoU prediction for semi-supervised 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14615–14624, 2021. doi: 10.1109/CVPR46437.2021.01438.

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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 605–621, 2020. doi: 10.1007/978-3-030-58536-5_36.

Qiming Xia, Wei Ye, Hai Wu, Shijia Zhao, Leyuan Xing, Xun Huang, Jinhao Deng, Xin Li, Chenglu Wen, and Cheng Wang. HINTED: Hard instance enhanced detector with mixed-density feature fusion for sparsely-supervised 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15321–15330, 2024.

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

Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-End Semi-Supervised Object Detection with Soft Teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3060–3069, 2021.

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 *Computer Vision – ECCV 2022*, pp. 107–124. Springer, 2022a. doi: 10.1007/978-3-031-19842-7_7.

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 *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2583–2589, 2022b. doi: 10.1109/ICRA46639.2022.9812038.

Junbo Yin, Jin Fang, Dingfu Zhou, Liangjun Zhang, Cheng-Zhong Xu, Jianbing Shen, and Wenguan Wang. Semi-supervised 3D object detection with Proficient Teachers. In *Computer Vision – ECCV 2022*, pp. 727–743, 2022. doi: 10.1007/978-3-031-19839-7_42. URL https://arxiv.org/abs/2207.12655.

Jinsu Yoo, Zhenyang Feng, Tai-Yu Pan, Yihong Sun, Cheng Perng Phoo, Xiangyu Chen, Mark Campbell, Kilian Q. Weinberger, Bharath Hariharan, and Wei-Lun Chao. Learning 3D perception from others' predictions. In *International Conference on Learning Representations (ICLR)*, 2025.

Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, and Zaiqing Nie. DAIR-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21361–21370, 2022.

Jiacheng Zhang, Jiaming Li, Xiangru Lin, Wei Zhang, Xiao Tan, Junyu Han, Errui Ding, Jingdong Wang, and Guanbin Li. Decoupled Pseudo-labeling for Semi-Supervised Monocular 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16923–16932, June 2024.

Ruiyang Zhang, Hu Zhang, and Zhedong Zheng. Harnessing Uncertainty-aware Bounding Boxes for Unsupervised 3D Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9230–9240, October 2025. doi: 10.1109/ICCV51701.2025.00862.

Na Zhao, Tat-Seng Chua, and Gim Hee Lee. SESS: Self-ensembling semi-supervised 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11079–11087, 2020.

APPENDIX

A IMPLEMENTATION DETAILS

A.1 DETECTOR, DATA, AND TRAINING UPDATE

PointPillars uses a 64-channel pillar encoder and a three-scale BEV backbone with 64/128/256 channels and 3/5/8 convolutional layers. The upsampled 128-channel outputs are concatenated and reduced to 256 channels. F-Cooper uses elementwise maximum fusion; AttFusion uses scaled dot-product attention; DiscoNet uses learned spatial collaboration weights. OPV2V uses region $[-140.8, 140.8] \times [-40, 40] \times [-3, 1]$ m and voxel size $0.4 \times 0.4 \times 4$ m. DAIR-V2X uses $[-100.8, 100.8] \times [-40, 40] \times [-3.5, 1.5]$ m and $0.4 \times 0.4 \times 5$ m. Each scene includes at most five agents.

Fixed scene lists define labeled and unlabeled data, with a ten-frame exclusion guard around each labeled block (Table 5). Each update pairs one labeled scene with one unlabeled scene. Training starts from a supervised checkpoint and runs for 30 pseudo-epochs. A pseudo-epoch is a fixed update budget covering part of the unlabeled list; it also defines the threshold schedule. Shared reflection, rotation ($\pm\pi/4$), and scaling ($[0.95, 1.05]$) are applied consistently to every agent and relative pose. Silence remains observable only when augmentation preserves the view's in-range points inside the region of interest. Algorithm 1 gives the update sequence.

Table 5: **Training-scene partitions.** L and U count labeled and unlabeled scenes. Guard counts scenes excluded by the ten-frame separation rule. Budgets refer to labeled scenes, not labeled objects.

Dataset	Budget	L	U	Guard
OPV2V	10%	634	4,947	793
OPV2V	20%	1,275	4,330	769
DAIR-V2X	10%	482	3,467	940
DAIR-V2X	20%	971	3,017	901

Algorithm 1 One QACE training update

Require: labeled scene (X_L, Y_L), unlabeled scene X_U, student θ, teacher $\bar{\theta}$

1: Teacher $\bar{\theta}$ predicts individual-agent and fused proposals on X_U
2: Form provisional candidates from aligned local clusters and fused proposals (Section 3.2)
3: **for** each candidate i **do**
4: Compute opportunity o_{ia} for every view a (Eq. 1) and assign its role (Table 1)
5: Compute S_i, M_i, C_i^k, and supporter agreement g_i^k (Eqs. 2, 3, 5)
6: Determine provisional admission and ceilings from candidate-source rules and available auxiliary predictions (Appendix A.4)
7: Apply VCE admission (Eq. 4) and the silence veto; for retained candidates, compute q_i^k and $\omega_i^{\mathrm{cls}}, \omega_i^k$ (Eqs. 6–7)
8: **end for**
9: Assign pseudo-boxes and weights to anchors; compute $\mathcal{L}_{\mathrm{pseudo,fused}}$ on X_U
10: Compute $\mathcal{L}_{\mathrm{sup,fused}}$ and $\mathcal{L}_{\mathrm{sup,local}}$ on (X_L, Y_L)
11: Update θ with $\mathcal{L}_{\mathrm{student}}$ (Eq. 8); update $\bar{\theta}$ by EMA

A.2 Candidate Construction and Opportunity

Section 3.2 describes proposal decoding, alignment, clustering, and threshold adaptation. The following specifies the opportunity and correspondence conditions used to qualify candidate evidence.

Opportunity and roles. Opportunity is evaluated in each agent's local frame (Eq. 1). Matched supporters are retained when a count is unavailable. Eligibility requires τ_{elig}; negative silence additionally requires τ_{sil} and preserved observability after augmentation. Silence denotes an eligible view without an overlapping proposal, not verified object absence. Appendix A.5 gives the opportunity and attribute-quality thresholds.

Heading targets and agreement. The circular mean in candidate construction supplies the heading target; Eq. 5 separately measures directed and axial agreement. Axial agreement evaluates orientation modulo π, so it can remain high when supporting headings disagree on front–back direction. Yaw-axis supervision requires sufficient axial quality and an aggregated target compatible with that orientation. Once these conditions hold, the sine-difference loss can supervise the axis while the direction weight is zero. Center and size retain their own permissions and gates.

Attribute discrepancies. Associated but unmatched proposals can constrain individual attributes. Center discrepancy is the XY distance normalized by the boxes' mean footprint diagonal. Size discrepancy is the mean absolute log ratio of their three dimensions, and yaw-axis discrepancy is the angular difference modulo π. Clipped linear ramps map these discrepancies to $r_{ia}^{k} \in [0, 1]$. Directed conflict is binary and occurs when the heading difference exceeds $\pi/2$. Equation 3 aggregates all four measurements, including measured zeros, over total computable opportunity.

Scope of counter-evidence. Abstention removes a view from direct support and counter-evidence; any computable opportunity remains in D_i. It can therefore alter admission or attribute gates through normalization. With exactly two physical views, regression eligibility requires both to be supporters. Both views then contribute support, leaving no view to supply conflict or silence; EAT uses their measured agreement. Conflict-conditioned regression requires at least one additional qualified physical view; fused predictions do not constitute additional physical views.

Admission and reliability. Cluster geometry G_i combines axis-aligned BEV IoU with exponential center and yaw agreement among supporting proposals. Reliability η_i combines teacher confidence, support S_i, geometry G_i, and the candidate-source prior. High reliability yields classification and regression ceilings η_i and η_i^2; intermediate reliability yields classification only. Weak and fused-only candidates retain their classification-only permission.

A.3 Gate and Losses

Gate. Auxiliary candidates use the head's object probability for Q_i. Candidates with an unambiguous fused reference use the geometric mean of candidate and reference confidence. Candidates without an auxiliary or fused reference use the geometric mean of teacher confidence, support, and cluster geometry. Association quality g_i^{assoc} takes the larger of cluster quality and fused-reference match quality. Cluster association quality is the mean of support and geometry. Fused-reference match quality is the mean of IoU, center, and yaw agreement. Each attribute ceiling c_i^k is the minimum of the object regression ceiling and any attribute weight already assigned during candidate construction.

Losses. Section 3.5 specifies anchor inheritance and loss normalization. Background supervision uses anchors with very low fused-teacher object probability in scenes containing at least one positive pseudo-anchor. Neighborhoods around positive, ignored, and budget-dropped candidates are excluded from these background targets. Classification, regression, and direction losses use separate permissions after anchor assignment.

A.4 Auxiliary Heads

The detector's IoU head predicts localization quality from detached features. It is trained on labeled predictions with target $2\,\mathrm{IoU}_{\mathrm{BEV}} - 1$. Its mapped output $q_{\mathrm{IoU}} \in [0, 1]$ serves as a candidate feature and a regression-admission condition.

The candidate-quality head is a $27 \to 32 \to 2$ multilayer perceptron with 962 parameters. Its detached inputs comprise teacher score, IoU quality, vote ratio, cluster geometry, attribute consistencies, VCE statistics, normalized box geometry, and availability flags. Algorithm 1 evaluates these features for provisional candidates before auxiliary admission. Disabling VCE replaces its statistics with unweighted vote-based counterparts, preserving the input dimension. The head predicts object probability and XY center error. It is trained with a separate optimizer on labeled scenes, using rotated-IoU-based positive and negative labels.

After warm-up, the head admits candidates with high object probability, low predicted center error, and observed spatial agreement, provided that no existing box covers them. Object probability sets the classification ceiling. It also sets the regression ceiling when q_{IoU} reaches the regression-admission threshold; otherwise that ceiling is zero. All admitted candidates then pass through VCE and EAT. Their attribute weights still require measured cross-agent agreement, so a candidate with one supporter remains classification-only. Uncertain candidates protect ignored regions. Final detection scores come from the student's classification head.

A.5 Optimization

The teacher follows $\bar{\theta} \leftarrow m_s\bar{\theta} + (1 - m_s)\theta$ with $m_s = \min\{1 - 1/(s+1), 0.99\}$. For pseudo-epoch e_s, let $u(e_s) = 0.1 + 0.4\min(e_s/8, 1)$, $\hat{L}_s^{\mathrm{sup}}$ the detached fused supervised loss without its IoU term, and $\hat{L}_s^{\mathrm{ps}}$ the detached fused pseudo-loss. The pseudo-loss coefficient is $\lambda_{u,s} = u(e_s)$ if $\hat{L}_s^{\mathrm{ps}} = 0$ and $\min\{u(e_s), 0.5\hat{L}_s^{\mathrm{sup}}/\hat{L}_s^{\mathrm{ps}}\}$ otherwise. This caps the pseudo-loss contribution at half the detached fused supervised detection loss.

Method settings. VCE uses opportunity thresholds $\tau_{\mathrm{elig}} = 0.65$ for view eligibility and $\tau_{\mathrm{sil}} = 0.75$ for negative silence. EAT uses attribute-quality thresholds of 0.5 for center and size, and 0.7 for yaw axis and directed heading.

B Baseline Adaptations

We implement each baseline's supervision strategy using the common collaborative detector and scene-level label splits. CoDTS and HINTED were originally developed for sparse object annotations within scenes. We adapt their supervision strategies to the scene-level semi-supervised protocol used here.

3DIoUMatch (Wang et al., 2021) filters EMA proposals by fused confidence, predicted BEV IoU, and ego-only confidence, applies lower-half suppression, and excludes uncertain background anchors; its IoU head is trained jointly with the feature extractor.

CoDTS (Han et al., 2025) combines a fixed supervised teacher and an EMA teacher; EMA proposals fill unoccupied anchor locations, and accepted boxes supervise their best-overlap anchors and all anchors above the positive threshold.

HSSDA (Liu et al., 2023) groups fused EMA predictions into confidence levels by Jenks breaks of labeled positive-anchor scores and applies a common 2×2 BEV patch permutation across agents.

HINTED (Xia et al., 2024) passes regular and locally downsampled point clouds through shared encoder and fusion modules, mixes the final BEV features with normalized gates, and supervises each representation.

C Additional Results

C.1 Five-Percent Annotation Results

Table 6 reports the 5%-label comparison. Relative to the strongest baseline in each column, QACE has gains of 3.31 AP@0.5 and 4.86 AP@0.7 points. Maximum recall increases by 3.46 points at IoU 0.5 and 6.31 at IoU 0.7. Far-range AP@0.7 increases by 5.49 points.

Table 6: **OPV2V/F-Cooper with 5% labeled scenes.** All AP and maximum recall values are reported as percentages. AP@0.3 denotes BEV AP at IoU 0.3; MR@0.5 and MR@0.7 denote maximum recall at IoU 0.5 and 0.7, respectively. Near, Mid, and Far cover $[0, 30)$, $[30, 50)$, and $[50, \infty)$ m from the ego agent, within the evaluation range. Bold denotes the best result per column.

	AP ↑			MR ↑		AP@0.7 by distance ↑		
Method	AP@0.3	AP@0.5	AP@0.7	MR@0.5	MR@0.7	Near	Mid	Far
3DIoUMatch	82.18	71.94	47.39	80.28	64.45	65.72	39.95	22.99
CoDTS	86.37	79.85	58.70	83.64	67.93	77.55	53.70	28.73
HSSDA	89.53	81.84	56.42	86.51	67.40	76.21	49.32	26.15
HINTED	86.19	79.04	56.15	83.48	66.29	74.83	52.46	23.97
QACE (ours)	**92.54**	**85.15**	**63.56**	**89.97**	**74.24**	**80.57**	**58.08**	**34.22**

C.2 Pseudo-label Threshold Policy

Table 7 evaluates threshold policies on DAIR-V2X/F-Cooper with 10% labels. Scheduled adaptation reaches 77.89 AP@0.5 and 58.06 AP@0.7, exceeding the stronger fixed policy by 0.88 and 1.55 points. The comparison varies threshold adaptation and stage scheduling together.

Table 7: **Pseudo-label threshold policies** on DAIR-V2X/F-Cooper with 10% labeled scenes. Values are oriented BEV AP (%); fixed policies are compared with the combined scheduled adaptive policy.

Threshold policy	AP@0.5 ↑	AP@0.7 ↑
Fixed 0.25	76.73	56.25
Fixed 0.35	77.01	56.51
Stage-scheduled adaptive (QACE)	**77.89**	**58.06**

D Qualitative Detection Results

Figures 5 and 6 compare CoDTS, HSSDA, HINTED, and QACE on eight selected OPV2V test scenes with F-Cooper fusion and 10% labeled training scenes. Each row uses identical inputs and annotations, with a common score threshold of 0.2 and NMS threshold of 0.15. The full-scene views retain the original point clouds, boxes, and confidence labels.

In the top row of Figure 5, QACE matches all 11 annotated vehicles at BEV IoU ≥ 0.7, with no unmatched predictions; each baseline matches five vehicles and retains 5–18 unmatched predictions. In the second row, QACE matches 20 of 28 vehicles, compared with 14–16 for the baselines, while retaining seven unmatched predictions compared with 10–16.

The additional scenes in Figure 6 show consistent frame-level differences. In its bottom row, QACE detects a vehicle 69.1 m from the ego vehicle with BEV IoU 0.77; none of the three baselines retains a prediction overlapping that annotation. Across all eight displayed scenes, QACE has more matches and no more unmatched predictions than any baseline.

We selected these success cases from 217 regularly sampled test frames using fixed checkpoints and thresholds. Section 4.4 presents the CoDTS/QACE comparison used in the main text.

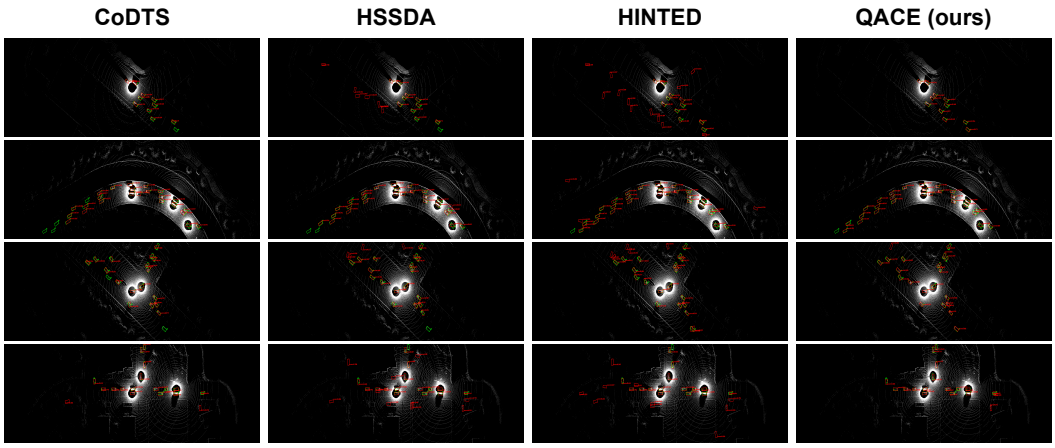

Figure 5: **Full-scene comparisons on OPV2V (I).** Columns show CoDTS, HSSDA, HINTED, and QACE; rows show four selected test scenes. Green boxes are annotations and red boxes are retained predictions. All panels preserve the same full-scene field of view; only uniform white export margins are removed. A common $1.35\times$ vertical display magnification is applied to every panel; IoU is computed in the original BEV coordinates.

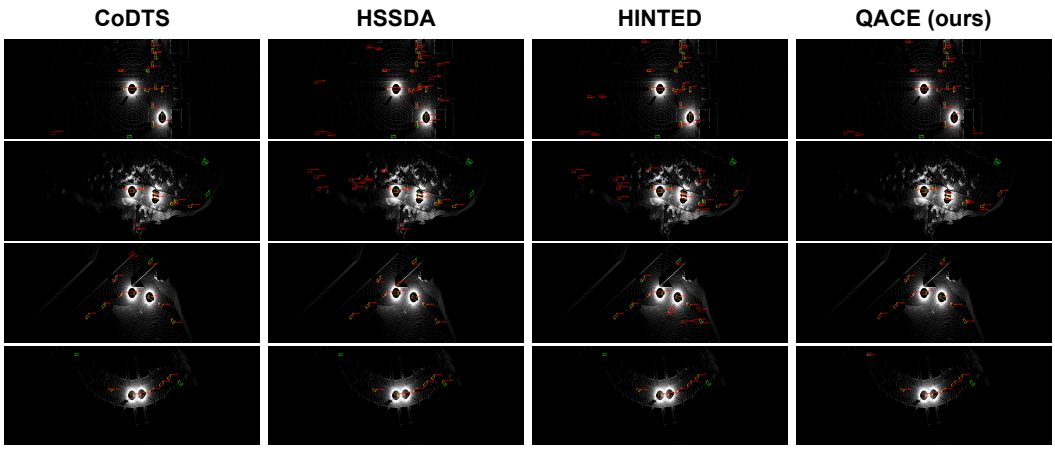

Figure 6: **Full-scene comparisons on OPV2V (II).** Four additional test scenes use the same column order, detection thresholds, and display scaling as Figure 5. The complete scene is shown for each method.