WHEN SILENCE MISLEADS: CENSORED OBSERVATION-AWARE VERIFICATION FOR COLLABORATIVE PERCEPTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Object-level collaborative perception lets an ego agent recover objects it cannot see, yet every fusion rule in use reads the ego's own silence as evidence: a remote proposal the ego did not also detect is discounted. We show that this silence is *censored by observability*. Among uncorroborated remote proposals, the fraction that correspond to real objects falls from 10.7% to 2.0% on real data and from 6.0% to 0.1% in simulation as the ego's opportunity to verify the claim rises, while the confidence the sender reports is flat across the same axis — so a rule that reads scores alone cannot separate an unobservable hypothesis from a refuted one. The same structure defeats calibration: scores well calibrated on ego-visible objects are over-confident on ego-invisible ones, and neither message-observable context nor corroboration counting repairs it. We formulate missing corroboration as a censored observation: an absence is evidence only where the ego was eligible to observe. COVE is the minimal operator enforcing that admissibility condition — a remote hypothesis is attenuated only when the ego was physically eligible to verify it and did not. It needs no retraining or extra message payload. On six frozen scene-level holdouts spanning real and simulated data and a heterogeneous LiDAR/camera pairing, it improves aggregate AP70 over the score-mean late-fusion rule it is applied to on every holdout ($+1.9$ to $+5.7$ pp, every interval excluding zero) and ego-invisible AP70 by up to $+20.8$ pp, significantly on the four simulated holdouts; treating silence as a negative vote costs up to -29.1 pp, and uniform down-weighting buys aggregate accuracy at the cost of the ego-invisible subset. A controlled intervention — identical proposals, association and emitted geometry, with only the score aggregation replaced by evidential aggregation — improves aggregate AP70 in all twelve bootstrapped comparisons with every paired interval excluding zero, so the effect is not a property of one scoring rule. Across six more expressive extensions, none yields a consistent held-out improvement over the minimal operator.

1 INTRODUCTION

Collaborative perception promises what a single vehicle cannot have: objects behind a truck, around a corner, past the range of its own sensor. Object-level (late) fusion is the deployment-friendly way to get it — agents exchange a few kilobytes of boxes and scores, with no shared feature space, no joint training and no architectural compatibility requirement (Chen et al., 2019; Xu et al., 2022b; Lu et al., 2024). The ego agent then faces a decision that has no counterpart in single-agent perception: *how much should it trust a hypothesis it cannot check?*

Every fusion rule in deployment answers this with the ego's own detector. A remote box that the ego also produced is reinforced; a remote box that the ego did not produce is discounted. That rule reads the ego's silence as evidence against the object. But silence is *censored by observability*: the ego is silent most often about precisely the objects it could not see, and those are the objects collaboration exists to supply. Figure 1 separates the two cases empirically. Among remote proposals the ego did not corroborate, the fraction that correspond to real objects falls from 10.7% to 2.0% on real data and from 6.0% to 0.1% in simulation as the ego's opportunity to verify the claim rises — up to a $60\times$ difference in what the same silence means — while the confidence the sender reports moves

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.

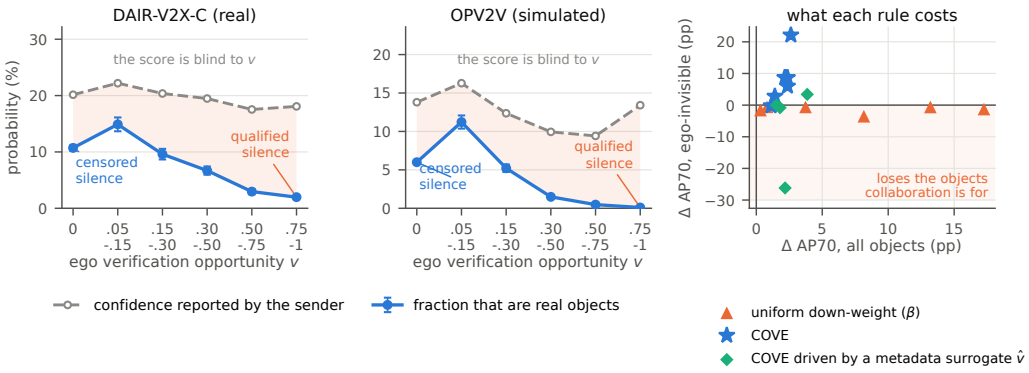


Figure 1: Silence is not one event, and treating it as one is expensive. **Left, centre:** among remote proposals that the ego did *not* corroborate, the fraction that correspond to a real object (blue) against the ego's verification opportunity v — the fraction of its rays aimed at the box that reach it rather than being intercepted earlier. Where the ego could not look ($v \approx 0$) its silence is nearly uninformative (*censored*); where it could look and still reported nothing, the proposal is almost never real (*qualified*). The sender's confidence (grey) is flat across the same axis; error bars are 95% binomial intervals. **Right: aggregate accuracy alone can reward suppressive fusion.** What each use of silence costs against naive late fusion on the frozen holdouts. Uniform down-weighting (β) buys aggregate accuracy while giving back nothing — or losing up to 3.7 pp — on the ego-invisible objects, and driving COVE with a metadata surrogate $\hat{v}$ loses up to 26.2 pp there (Section 5). Among the rules shown, only the ray-based COVE improves both axes (a second ego-side signal: Appendix M).

by less than two points. What separates informative silence from uninformative silence is not in the message.

Answering any of this first requires an evaluation set in which ego-invisible objects exist at all. Section 3.1 shows that the released cooperative labels of a standard benchmark omit a third of the objects its own infrastructure sensor labels inside the evaluation range, most of them invisible to the ego, so the population this paper is about is largely absent by construction; we therefore label it, and report the headline results under both definitions.

We formulate missing corroboration as a *censored observation* and introduce COVE, a fusion operator with an explicit abstention state. The ego computes, from its own LiDAR sweep and the received box, whether it was physically eligible to verify the claim: the fraction of its rays aimed at the box that actually reach it rather than being intercepted earlier. A remote hypothesis is attenuated only when that eligibility is high *and* the ego failed to corroborate; when the ego could not look, its silence changes nothing. The rule is one line and one parameter, requires no retraining and no extra message payload, and subsumes both uncorrected late fusion and uniform attenuation as limiting cases.

Across six frozen scene-level holdouts — real (DAIR-V2X-C) and simulated (OPV2V) LiDAR–LiDAR pairs, and a heterogeneous LiDAR-ego / camera-collaborator pairing — COVE improves aggregate AP70 over its base rule on every holdout (+1.9 to +5.7 pp) and ego-invisible AP70 by up to +20.8 pp, significantly on the four simulated holdouts and in the same direction on the two real-data ones, with the ego-visible subset improving as well (Section 5). Uniform attenuation in the style of Xu et al. (2025) buys up to +13.2 pp aggregate AP70 on real data *while losing* the ego-invisible subset — a trade-off that aggregate AP alone rewards and visible in Figure 1 (right). We also report what did not work: six elaborations — a likelihood-ratio form, a three-state extension, a multi-agent qualified jury, a displacement prior over v, learned localisation uncertainty and a learned veto-risk controller — are all no better than the one-parameter rule.

Contributions. (i) A benchmark-level diagnosis of a hidden failure mode in object-level collaborative detection: missing ego corroboration is used as negative evidence even when the ego could not observe the hypothesis, while the released cooperative labels substantially under-represent the objects on which this mistake matters most (Section 3.1). (ii) A transfer diagnosis showing, with

confidence intervals across real and simulated data and two detector seeds, that remote-proposal reliability estimated where the ego can verify does not transfer to ego-invisible objects, and that neither recalibration nor message-observable context closes the gap. (iii) An admissibility–strength factorisation of censored negative evidence and the finding that the dominant decision is whether an absence may count as evidence: the minimal COVE operator improves aggregate AP70 on every frozen holdout and under every score aggregation tested, while six more expressive variants yield no consistent held-out improvement. (iv) A controlled evaluation protocol with frozen proposals, association and geometry, outer scene-level holdouts and separate aggregate/ego-invisible reporting, exposing the trade-off by which suppressive fusion can improve aggregate AP while removing precisely the objects collaboration is meant to recover.

2 RELATED WORK

Object-level collaborative perception. Agents can share raw points, intermediate features or detections (Chen et al., 2019; Wang et al., 2020; Xu et al., 2022a; Lu et al., 2024). Object-level fusion needs neither a shared feature space nor joint training, which is why it survives heterogeneous deployment, and it is the setting we study. Its known weakness is that merging boxes and applying NMS admits low-confidence false positives from collaborators (Xu et al., 2025; Mohammadisarab et al., 2025); the standard remedy is a uniform discount by β, which we reproduce and cost out on ego-invisible objects. Uncertainty-weighted and evidential variants replace the score with a learned uncertainty (Fadili et al., 2025; Si et al., 2026); none conditions negative evidence on whether the ego could observe the object.

Observability as a reason not to update. Occupancy mapping and multi-object tracking already condition negative updates on visibility (Antonopoulos et al., 2026; Godoy et al., 2021; Krejčí et al., 2025; Motro & Ghosh, 2018), and misbehaviour detection for collective perception applies an occupied/free/unknown trichotomy (Qualcomm Incorporated, 2024; Liu et al., 2021). Our claim is not the observability principle itself: the missing result is that learned object-level collaborative detection systematically violates it, and that the standard evaluation protocol obscures the population on which the violation is most costly. On DAIR-V2X-C, 31% of in-range infrastructure-side labels are absent from the released cooperative annotations and 59% of those contain no ego LiDAR return (Section 3.1). Once that population is represented, the violation produces a measurable failure that neither recalibration nor message-observable context repairs, and the more expressive update rules we test do not consistently improve on correcting admissibility alone.

Calibration and distribution shift. Post-hoc (Guo et al., 2017) and localisation-aware (Oksuz et al., 2024) calibration assume matched calibration and deployment populations; label-shift corrections estimate a changed prior from unlabelled target data (Lipton et al., 2018), and positive-unlabelled learning models why a label is missing rather than treating absence as negative (Elkan & Noto, 2008). Our diagnosis belongs to that family — the "label" is the ego's corroboration, missing non-randomly — but the fix is an abstention state, not a reweighting.

Free-space reasoning for detection. LiDAR occlusion and in-box free space have been used to flag implausible detections against spoofing (Sun et al., 2020); COVE uses the geometry not to reject a hypothesis (Appendix C) but to decide whether the *ego's own silence* is admissible evidence.

3 VERIFICATION CENSORING IN OBJECT-LEVEL FUSION

3.1 EGO-CENTRIC GROUND TRUTH CANNOT EVALUATE REMOTE-ONLY COLLABORATION

Collaborative perception is motivated by objects the ego cannot see, yet benchmarks label objects the way a single agent would. On the $1{,}789$ validation frames of DAIR-V2X-C (Yu et al., 2022) with the cooperative annotations of Lu et al. (2023), 31% of the infrastructure-side vehicle labels inside the evaluation range ($6{,}107$ of $19{,}522$) have no counterpart in the cooperative label set, and 59% of those contain no ego LiDAR return at all. The "cooperative" ground truth is therefore largely the vehicle's ground truth: most objects behind an occluder are not in it, and a collaborator that detects one is penalised (Figure 4).

We therefore evaluate against the union of the cooperative and infrastructure-side labels of the same four vehicle classes, in the ego frame, de-duplicated at BEV IoU ≥ 0.1 or centre distance < 2.5 m, dropping boxes centred inside the ego footprint; the infrastructure-only labels add 19% to the ground truth. OPV2V (Xu et al., 2022b) labels all vehicles in range and needs no such treatment.

3.2 Setting and notation

Agents run independently trained detectors and exchange object proposals. The ego agent receives from a collaborator a hypothesis $h_i = (b_i, s_i)$, a 3D box in a common frame with a confidence, and after association emits one box and one confidence per cluster. Geometry is fixed throughout: every rule selects the same box for a cluster and differs only in the confidence it assigns, so any change in AP is attributable to scoring.

Visibility strata. For evaluation only, we stratify ground-truth objects by the number of ego LiDAR returns inside them (ignoring the bottom 0.3 m, so ground returns do not count): *ego-visible* (≥ 20 points), *remote-only* (0 points, and ≥ 20 points from some collaborator), and *marginal* otherwise. Remote-only objects are the population collaboration exists for; the stratification is never available to any fusion rule.

The event that fusion actually observes. Let $Y_i = 1$ denote that h_i is a correct hypothesis, $M_i = 1$ that the ego produced a matching proposal, and $O_i = 1$ that the ego was in a position to observe the region h_i refers to. Fusion rules see M_i, not O_i. The informative event is not $M_i = 0$ but $M_i = 0 \mid O_i = 1$:

$$\underbrace{M_i = 0}_{\text{no corroboration}} \quad = \quad \underbrace{\{O_i = 0\}}_{\text{censored: the ego could not look}} \quad \cup \underbrace{\{M_i = 0,\, O_i = 1\}}_{\text{informative absence}}.$$

Proposition 1 (Ambiguity of missing corroboration)**.** *Under censored sensing, $\{M = 0\}$ contains both unobservable hypotheses ($O = 0$) and observable-but-uncorroborated ones ($O = 1$). Any fusion rule whose update depends on M alone therefore assigns the same evidence to two observationally distinct events, and in particular applies the same negative update to a hypothesis the ego was unable to examine as to one it examined and failed to confirm.*

The decomposition exposes a structural ambiguity: the observed event $M = 0$ conflates censoring with genuine negative evidence, and empirically the two branches are neither rare nor exchangeable: on our data most remote-only objects fall in the censored branch, so a rule that penalises $M = 0$ uniformly removes recall exactly where the collaborator is the only source of information. Section 5 measures that cost.

Why recalibration does not resolve it. A natural alternative is to recalibrate the remote confidence for the shift between the two strata. Calibrators can only be fitted where labels or corroboration exist — that is, in the ego-visible stratum — and Section 5 shows that the fitted map does not transfer, that message-observable geometric context does not close the gap even where the covariates lie inside the ego-visible support, and that the only residual signal which does transfer is corroboration itself: precisely the ambiguous event of Proposition 1.

4 Cove: censored observation-aware verification

Cove is not a recalibration of the remote confidence. It is a rule for *when a missing corroboration is admissible as negative evidence at all*, and it replaces the binary "corroborated / not corroborated" distinction with three semantic actions: CORROBORATE, ABSTAIN and VETO. Only the third one changes a score.

4.1 Ego verification opportunity

We need a quantity that says whether the ego was *eligible* to verify a hypothesis — not how confident it is, nor whether it agrees — computed from the ego's own sweep and the received box, with no extra payload.

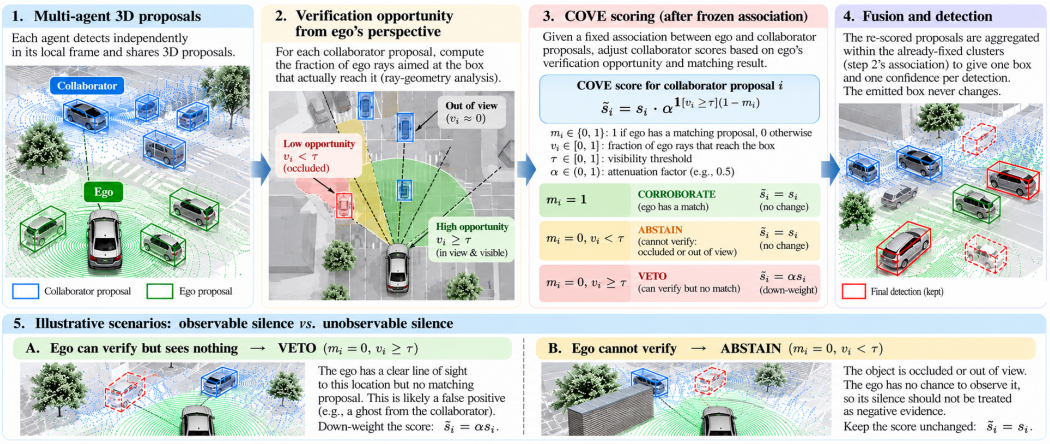


Figure 2: COVE in the order it is applied. (1) Agents detect independently and exchange 3D proposals. (2) The ego computes each proposal's verification opportunity v_i from its own sweep by ray geometry; occluded and out-of-range locations both give $v_i \approx 0$. (3) Under the frozen association of Section 5 only the confidence is rewritten: CORROBORATE and ABSTAIN leave s_i untouched, and only a *qualified* absence is attenuated. (4) Proposals are aggregated within those already-fixed clusters; the emitted box never changes. (5) The two meanings of ego silence: conventional late fusion applies the VETO branch in both scenes, whereas COVE abstains in B.

Consider the ray from the ego sensor origin to each return. A ray is *aimed* at b_i if it crosses the box footprint at a height inside the box's vertical extent (excluding the bottom 0.3 m). An aimed ray is *blocked* if its return lies before the near face of the box, and *reaching* otherwise. The verification opportunity is the fraction of aimed rays that reach the box,

$$v_i \;=\; \frac{\#\{\text{reaching rays}\}}{\#\{\text{aimed rays}\}} \in [0,1], \qquad v_i \doteq 0 \text{ when no ray is aimed at } b_i. \tag{1}$$

Absence of a return is not an observation, so an empty region beyond the sensor's range yields $v_i = 0$ (abstention), not a false veto. The computation is a slab–ray intersection per (proposal, sweep) pair.

v_i is about the *ego's geometry*, not the object, and it is not a proxy for distance: Section 5 replaces it with a regression on message-observable covariates and finds that the surrogate keeps the aggregate benefit and loses the ego-invisible one.

4.2 When is an absence admissible evidence?

Let $O_i \in \{0,1\}$ be the latent event that the ego was in a position to observe the region h_i refers to, M_i its match indicator and Y_i the correctness of the hypothesis. An ego that cannot observe a region cannot match a proposal in it, so $P(M_i{=}0 \mid Y_i, O_i{=}0) = 1$ for both values of Y_i, and the likelihood ratio carried by the observed absence is

$$\Lambda_i(O_i) \;=\; \frac{P(M_i{=}0 \mid Y_i{=}1, O_i)}{P(M_i{=}0 \mid Y_i{=}0, O_i)} \;=\; \begin{cases} q_1/q_0 < 1, & O_i = 1, \\ 1, & O_i = 0, \end{cases} \tag{2}$$

with $q_y = P(M_i{=}0 \mid Y_i{=}y, O_i{=}1)$. An absence is evidence only on $\{O=1\}$; on $\{O=0\}$ it is exactly uninformative, whatever the sender's score.

Proposition 2 (Censoring-induced suppression bias)**.** *A rule that applies a fixed log-penalty* $\log\alpha < 0$ *to every* $M_i = 0$ *incurs, relative to Equation 2, an excess penalty of* $\log\alpha$ *on* $\{O_i = 0\}$ *and none on* $\{O_i = 1\}$.

Equation 2 shows that this error is systematic rather than zero-mean noise: the excess penalty is concentrated on $\{O = 0\}$, precisely where collaboration supplies information unavailable to the ego, and $P(O = 0)$ is near one on remote-only objects and near zero on ego-visible ones. So the mis-applied penalty lands almost entirely on the stratum collaboration exists to supply, while the

aggregate — dominated by ego-visible objects — can improve at the same time. Both halves of that prediction are measured in Section 5, and the second is why aggregate AP alone cannot adjudicate this design.

COVE is the minimal estimator of Equation 2: it substitutes the computable eligibility $\mathbf{1}[v_i \geq \tau]$ for O_i and applies one constant penalty there. The constant α stands in for the magnitude q_1/q_0, which no message reveals, and τ trades the two ways of mis-estimating O_i. The operator is one line because the content is three admissibility conditions, not a functional form: a corroborated hypothesis is never penalised; an uncorroborated hypothesis the ego could not observe is never penalised; only an uncorroborated hypothesis the ego could observe may be.

4.3 THE FUSION OPERATOR

Let $m_i \in \{0, 1\}$ indicate that the ego produced a proposal matched to h_i under the frozen association. COVE scores the hypothesis as

$$\tilde{s}_i \;=\; s_i \, \alpha^{\,\mathbf{1}[v_i \geq \tau]\,(1-m_i)} \tag{3}$$

with a single attenuation factor $\alpha \in (0, 1]$ and an eligibility threshold τ. The operator has exactly three regimes, shown in Figure 2(4): CORROBORATE ($m_i = 1$) and ABSTAIN ($m_i = 0$, $v_i < \tau$) leave s_i unchanged; VETO ($m_i = 0$, $v_i \geq \tau$) multiplies it by α.

Admissibility–strength factorisation. COVE separates *whether* negative evidence may be used from *how strongly* it should act. With $A_i = \mathbf{1}[v_i \geq \tau](1 - m_i)$ for admissibility, a general censored-evidence update is $\tilde{s}_i = s_i \exp(-A_i w_i)$ with evidence strength $w_i \geq 0$; Equation 3 is the case $w_i = -\log \alpha$. The constant is deliberate: it isolates the censoring correction without a learned or cross-agent-calibrated uncertainty, and Section 6 shows the elaborations of w_i we tried do not pay for themselves. Equation 3 reduces to uncorrected late fusion at $\alpha = 1$ and to "attenuate every uncorroborated remote box" at $\tau = 0$; both are baselines in Section 5, and both are worse.

Plug-in property. Equation 3 reads only (v_i, m_i) and writes only a confidence, so COVE needs no retraining, shared feature space, common architecture or extra communication, and composes with any rule that emits one box and one score per group — the property Section 5.4 uses to swap the score aggregation while holding everything else fixed.

Parameters. (α, τ) are selected on development scenes by aggregate plus remote-only AP70 over the grid $\alpha \in \{0.25, 0.5\}$, $\tau \in \{0.25, 0.5\}$, then frozen; every holdout number uses parameters chosen on the disjoint half, and every baseline with a free parameter receives a selection budget of the same size on the same scenes (Section 5).

5 EXPERIMENTS

5.1 PROTOCOL

Data and detectors. **DAIR-V2X-C** (Yu et al., 2022) (real, one vehicle and one roadside unit; 1789 validation frames, 22 sequences) with the cooperative annotations of Lu et al. (2023), and **OPV2V** (Xu et al., 2022b) (simulated, 2–5 agents; 2170 test frames, 16 scenarios). Detectors are PointPillars (Lang et al., 2019) single-agent models from HEAL (Lu et al., 2024): the released OPV2V weights, and for DAIR a model trained with the released configuration (two seeds). The heterogeneous setting pairs a LiDAR PointPillars ego with a camera Lift-Splat-Shoot collaborator (Philion & Fidler, 2020).

Frozen pipeline and holdouts. Per-agent NMS, top-100 proposals at score ≥ 0.05, association (BEV centre distance ≤ 2 m, IoU ≥ 0.3, greedy by score, one proposal per agent) and the selected box are computed once and shared by every rule; rules differ only in cluster confidence. AP is BEV average precision (VOC all-point); detections matching a ground-truth box outside the evaluated subset are ignored. Intervals are 95% sequence-level bootstrap of the paired difference. Sequences are split in half by index parity; every free parameter — (α, τ) for COVE, β for uniform attenuation, and those of every variant — is selected on the development half and the other half is evaluated

Table 1: Six frozen holdouts, paired AP70 differences (pp, 95% CI), every parameter selected on the disjoint development half. COVE is applied on score-mean late fusion: the first column pair is what that base changes against naive (max-score) late fusion, the deployed default; the second is what COVE adds to its base. Uniform attenuation, given the same budget, gains $+13.2$ aggregate on DAIR-A while losing the ego-invisible subset (up to -3.7 pp, camera-A).

		score-mean − naive (base)		+COVE − base	
setting	holdout	all objects	ego-invisible	all objects	ego-invisible
DAIR-V2X-C (real)	A	-0.5 [$-2.0, +0.1$]	$+2.4$ [$+1.1, +4.2$]	$+1.9$ [$+1.2, +3.0$]	$+0.4$ [$-2.0, +2.5$]
	B	-0.8 [$-2.2, +0.9$]	$+6.3$ [$+1.9, +19.6$]	$+3.1$ [$+2.0, +4.2$]	$+2.4$ [$-3.5, +5.3$]
OPV2V (sim)	A	-2.5 [$-3.8, -1.0$]	$+1.3$ [$-9.4, +8.4$]	$+5.1$ [$+3.3, +8.0$]	$+20.8$ [$+12.4, +31.6$]
	B	-2.2 [$-3.7, -0.9$]	$+2.9$ [$-6.4, +6.9$]	$+4.3$ [$+2.7, +6.9$]	$+5.8$ [$+2.0, +21.5$]
OPV2V, camera collaborator	A	-3.3 [$-4.5, -2.2$]	-3.5 [$-11.2, +3.4$]	$+5.7$ [$+2.7, +8.6$]	$+9.4$ [$+2.0, +12.7$]
	B	-3.0 [$-5.6, -1.0$]	-2.2 [$-9.3, +1.6$]	$+4.1$ [$+2.5, +6.8$]	$+2.0$ [$+0.3, +5.1$]

once, in both directions, giving six holdouts. The two directions reuse the same scenes with roles exchanged, so they are not independent replicates. Unit tests assert label-leakage freedom, identical cluster ids across rules and AP invariance to monotone score transforms (Appendix D).

What is compared. Unless stated otherwise, COVE is applied on *score-mean* late fusion (box from the max-score member, confidence the mean of member scores), the late-fusion reference of Si et al. (2026) and the one conventional rule that pools member evidence; naive max-score fusion, the deployed default, never pools, and Table 1 reports the two changes separately. COVE is a plug-in for object-level late fusion, not a competing end-to-end collaborative-perception architecture: in every experiment proposals, association and emitted geometry are fixed and COVE changes only the admissibility of negative evidence. Labels are the union of the cooperative and infrastructure-side annotations (Section 3.1); because that changes the evaluation population, absolute AP is not directly comparable to published systems evaluated under their native annotations; Appendix A repeats the main comparison under the original annotations, where the ordering of the fusion rules is unchanged.

5.2 THE DIAGNOSIS

Reliability does not transfer. Table 4 reports the transfer gaps; over-confidence appears in every bin (Figure 3), a second DAIR seed reproduces it, and ranking is preserved (AUROC 0.92), so the failure is invisible to score-ordering metrics. Message-observable covariates do not improve NLL or Brier score on the ego-invisible stratum ($+0.029$ [$+0.011, +0.047$] NLL on DAIR, i.e. worse), even inside the ego-visible 5–95% support; the only feature that helps is cluster size, the corroboration indicator, as a base-rate correction rather than better ranking.

Corrections that hurt where it matters. Table 4 (bottom): the obvious repairs buy aggregate accuracy by removing ego-invisible objects, while an oracle that knows each cluster's correctness gains $+37.2$ pp there.

5.3 COVE ON FROZEN HOLDOUTS

Table 1 separates what the base rule changes from what COVE adds. Score-mean aggregation costs aggregate AP70 in point estimate everywhere, significantly on OPV2V, and helps the ego-invisible subset on real data; COVE then improves aggregate and ego-visible AP70 on all six holdouts with every interval excluding zero, and ego-invisible AP70 in point estimate on all six, significantly on the four simulated holdouts ($+2.0$ to $+20.8$ pp) and positive but less precisely resolved on the two real-data ones ($+0.4$, $+2.4$ pp). Against naive late fusion, the deployed default, the two add to $+2.8$ to $+22.1$ pp ego-invisible on the LiDAR–LiDAR holdouts. Two checks (Appendices A and M): remote-only status does not determine v — such objects contain no ego return by definition, yet their claimed boxes can have high ray reachability when ego rays traverse the region — so COVE vetoes 45–78% of the remote-only false positives on all six holdouts but also 3–15% of the remote-only clusters that correspond to a real object on five (30% on DAIR-B, consistent with box displacement, which Appendix K measured only on the heterogeneous holdouts); and the marginal stratum (1–19 ego returns) is never significantly harmed and improves on five of six. A simpler in-box return-

presence test is stronger on the DAIR ego-invisible subset but weaker on the simulated and marginal strata (Appendix M): what carries the effect is observability-conditioned admissibility, and the best ego-side eligibility signal is dataset-dependent.

Collaboration against the ego alone. On OPV2V collaboration is worth having under every rule: ego-only reaches 67.3 AP70 against 80.9 for naive late fusion and 83.0 with COVE (Appendix A). On DAIR-V2X-C it is not: ego-only reaches 55.6, naive late fusion 45.1 and COVE 46.9 (cf. Lu et al. (2024)). Two mechanisms produce the deficit, both measurable on the same holdouts. Remote false positives pollute the ranking: uniform attenuation, which changes no geometry, lifts aggregate AP70 to 61.8 and 56.9 on DAIR-A/B against 58.7 and 51.6 for the ego alone, at the cost of the ego-invisible subset (Figure 5). Remote boxes are also worse localised: the deficit is 2.5 pp at AP50 but 8.7 pp at AP70, keeping the ego box in mixed clusters — which changes no score there — lifts aggregate AP70 to 57.7 and 56.5 (Appendix L), and the frozen clusters show why: in 44% of clusters holding both an ego and a remote proposal the emitted box is the remote one; on the same object the ego box has the higher IoU 81% of the time (0.82 vs 0.60 median); and among remote-emitted boxes 44% fall below 0.7 IoU where the ego box would not have (17% the other way). COVE corrects a different, orthogonal failure — the censoring of negative evidence — and does not repair this one. A system-level variant that keeps the ego box in mixed clusters and admits remote-only clusters through COVE (Appendix L, a system-level comparison) beats the ego alone in aggregate AP70 on DAIR-A, DAIR-B and OPV2V-A ($+1.6$, $+5.9$, $+7.6$ pp, intervals excluding zero) and is level on OPV2V-B, where preserving the ego box itself costs accuracy; COVE's aggregate gain over that base is significant on all four holdouts and its ego-invisible gain on three; the variant did not meet the pre-registered inclusion rule on DAIR-A ego-invisible, so it is reported as system-level context only. A development-selected scale on the admitted remote-only clusters is chosen at 0.9–1 on all four holdouts, so a stratum prior on abstained proposals is not what the data ask for as far as AP, which is blind to a uniform rescaling within those clusters, is concerned.

Untouched confirmation. The six holdouts above share scenes. For an independent one we froze everything and evaluated once on nine previously unexported OPV2V validation sequences (Appendix J). Against conventional late fusion COVE improves AP70 by $+1.54$ $[+1.05, +4.83]$ pp overall and $+8.82$ $[+2.94, +29.36]$ pp on ego-invisible objects. With nine sequences the interval is wide: a confirmation, not universal generalisation.

Silence as a negative vote. On OPV2V, counting every non-reporting agent as a vote against costs -29.1 $[-40.5, -17.9]$ pp ego-invisible AP70 against COVE on holdout A (-11.3 on B): the clearest measurement of Proposition 1.

Is v just an occlusion prior in disguise? (post-hoc ablation.) Occlusion correlates with range and crowding, which the message already carries. On development scenes only we regress the true v on message-observable covariates (range, bearing, aspect, box size, local density, score), select the surrogate's own (α, τ) under the same objective and grid, and freeze both. The surrogate explains 37–39% of the variance in v out-of-fold, yet Appendix I shows a dissociation: every aggregate interval against naive late fusion excludes zero, while no ego-invisible difference is positive with an interval excluding zero and three of four are negative against COVE with the true v. Metadata reproduces the aggregate gains and not the objects collaboration is for; what it misses is ego-side observability itself: ray reachability here, and on real data an in-box return count serves as well (Appendix M).

5.4 IS THE GAIN TIED TO ONE SCORING RULE? A CONTROLLED INTERVENTION

Holding the proposal set, the association and the emitted box fixed, we replace the score aggregation with the evidential semantics of Si et al. (2026) and ask whether the COVE correction is still needed. Each proposal carries Beta pseudo-counts (α_i^+, α_i^-) and an epistemic uncertainty $u_{\mathrm{epi},i}$; with reliability $r_i = \mathrm{clip}(1 - u_{\mathrm{epi},i}, 0.1, 1)$, a cluster's semantics are the reliability-weighted sum of pseudo-counts while its geometry is selected, never averaged:

$$\alpha^{\pm *} = 1 + \sum_{i \in C} r_i \, (\alpha_i^{\pm} - 1), \qquad c^* = p^*(1 - u^*), \qquad i^* = \arg\max_{i \in C} S_i r_i, \tag{4}$$

Table 2: Controlled intervention on DAIR-V2X-C: paired AP70 difference (pp, 95% CI) from adding COVE, with proposals, association and emitted geometry fixed and only score aggregation changed.

		frozen association		calibrated association	
score aggregation	subset	A	B	A	B
conventional (max score)	all	$+1.5\,[+0.8, +2.2]$	$+2.1\,[+1.4, +2.8]$	$+2.3\,[+1.2, +3.3]$	$+2.4\,[+1.5, +3.3]$
	ego-invisible	$+0.0\,[-2.1, +0.9]$	$+0.8\,[-2.2, +2.7]$	$+0.7\,[+0.1, +1.0]$	$+1.2\,[-1.7, +3.4]$
evidential $S{\cdot}r$	all	$+1.7\,[+0.3, +2.7]$	$+2.6\,[+1.4, +3.8]$	$+2.3\,[+0.9, +3.3]$	$+2.8\,[+1.5, +4.0]$
	ego-invisible	$+0.4\,[-3.3, +1.4]$	$+5.9\,[+0.5, +9.6]$	$+0.7\,[-1.5, +1.4]$	$+6.5\,[+0.5, +11.5]$
evidential $p{\cdot}r$	all	$+2.1\,[+0.5, +3.3]$	$+2.7\,[+1.6, +3.9]$	$+2.6\,[+1.1, +3.7]$	$+2.9\,[+1.6, +4.2]$
	ego-invisible	$+0.6\,[-3.3, +1.8]$	$+6.3\,[+0.6, +9.8]$	$+0.8\,[-1.5, +1.6]$	$+7.0\,[+0.8, +11.5]$

with $S^* = \alpha^{+*} + \alpha^{-*}$, $p^* = \alpha^{+*}/S^*$, $u^* = 2/S^*$; we report the canonical selector $S_i r_i$ and its probability variant $p_i r_i$. Equation 4 replaces only the map from a cluster's members to its confidence, and Equation 3 then acts on c^* exactly as it acts on a conventional score — within a family the +COVE version emits the same box as its base. We run this under the frozen association and under one whose gates come from ground-truth correspondence on development scenes alone (Appendix E).

Table 2 reports twelve paired comparisons: three aggregation rules $\times$ two association protocols $\times$ two split directions. In all twelve COVE improves aggregate AP70 with every paired 95% interval excluding zero; all twelve ego-invisible point estimates are positive and five intervals exclude zero, clearest under evidential aggregation on split B ($+5.9$ to $+7.0$ pp). The twelve are not independent confirmations: the split directions reuse scenes and the rules score the same clusters.

What this does and does not show. The evidential variant is a *controlled* reimplementation, not a competitive system (absolute AP70 30–37 against 40–50), so we claim nothing about the published system; it shows that COVE changes *when negative evidence is admissible*, not *how confidence is represented*. The two association protocols give nearly identical absolute levels (33.8 vs 34.5, 30.2 vs 30.1). It runs on DAIR-V2X-C, where our evidential head keeps the detector's ranking (0.929 vs 0.945 AUROC; Appendix G).

6 WHAT FURTHER ELABORATION BUYS

Table 5 (Appendix C) reports six elaborations a reader would reasonably propose; none gives a consistent held-out benefit, locating the difficulty in when an absence is admissible as evidence, not in a more expressive confidence model.

Why each one fails. The likelihood-ratio gate loses on every holdout because the estimated $P(\text{corroboration} \mid Y{=}1, v)$ saturates at 1; the qualified jury does not beat the ego-only test (-2.0 pp on objects seen by two or more collaborators); learned localisation uncertainty σ carries no signal about which vetoes are mistakes (0.545 and 0.484 matched on score and range, chance); and predicting the veto's own risk fails because only 0.08–0.14% of candidate vetoes are mistakes (Appendix C).

What these failures suggest. Each buys expressiveness by relaxing one of the three admissibility conditions and pays for it on exactly the objects those conditions protect.

7 LIMITATIONS AND SCOPE

COVE assumes the ego's silence about a location it could observe is evidence about the object; that fails when the collaborator's box is displaced (Appendix K, which also lists our further assumptions). Two boundaries are deliberate: COVE addresses the admissibility of negative evidence, not its magnitude, and it is frame-local, since a later observation bears on an earlier hypothesis only if both refer to the same object.

AI USE STATEMENT

We used a generative AI coding assistant for software engineering tasks: implementing and refactoring the experiment scripts, the ray–box verification routine and the plotting code, for editorial assistance in drafting and tightening the text of this paper, and to render the schematic scenes in Figure 2 (whose content — the operator, its three regimes and the pipeline order — was specified and verified by the authors against Equation 3). All research questions, the evaluation protocol, the choice of baselines and every claim made here were formulated and checked by the authors; all AI-assisted code was read and, where it produced numbers, verified against independently computed quantities. Generative AI was not used to produce or alter any experimental result, and no text generated by it was included without author review. The assistant was not used for ideation of the core contribution.

ETHICS STATEMENT

This work uses two public datasets of road traffic, DAIR-V2X-C and OPV2V, under their respective licences; both are collected for autonomous-driving research and DAIR-V2X-C is released with privacy-preserving processing by its authors. We introduce no new data collection and no human subjects. The method concerns which evidence a vehicle should admit when fusing detections from other agents, and its intended effect is to make a system less likely to discard objects it cannot itself observe. We note one dual-use consideration: any rule that changes how much an ego agent trusts its collaborators interacts with adversarial or faulty senders, and we do not study that threat model here (Appendix K).

REPRODUCIBILITY STATEMENT

The evaluation protocol was fixed before the experiments reported here and is described in Section 5 and Appendix D: frozen proposal export, frozen association and geometry, outer scene-level holdouts in both directions, and development-half selection of every free parameter under one prespecified objective. Appendix D lists the thresholds and the unit tests asserting that no ego-invisible label enters a fitted component; detector checkpoints and the software environment will be released with the code; Appendix A gives the ground-truth definitions and repeats the headline results under the released cooperative labels; Appendix E and F give the association gates and the selected operating points. Post-hoc analyses are labelled as such where they appear. Code and the derived proposal caches needed to reproduce every table will be released.

REFERENCES

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In *Advances in Neural Information Processing Systems*, 2020.

Markos Antonopoulos, Anastasia Bolovinou, Bill Roungas, Elena Daskalaki, and Angelos Amditis. Beyond line of sight: Hybrid validation of V2X collective perception in complex scenarios. *arXiv preprint arXiv:2607.00874*, 2026.

Qi Chen, Xu Ma, Sihai Tang, Jingda Guo, Qing Yang, and Song Fu. F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3d point clouds. In *ACM/IEEE Symposium on Edge Computing*, pp. 88–100, 2019.

Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 213–220, 2008.

Maryem Fadili, Louis Lecrosnier, Steve Pechberti, and Redouane Khemmar. Evaluation of an uncertainty-aware late fusion algorithm for multi-source bird's eye view detections under controlled noise. *arXiv preprint arXiv:2507.03381*, 2025.

Jorge Godoy, Víctor Jiménez, Antonio Artuñedo, and Jorge Villagra. A grid-based framework for collective perception in autonomous vehicles. *Sensors*, 21(3):744, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, pp. 1321–1330, 2017.

Jan Krejčí, Oliver Kost, Yuxuan Xia, Lennart Svensson, and Ondřej Straka. Occlusion-aware multi-object tracking via expected probability of detection. *arXiv preprint arXiv:2511.20239*, 2025.

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

Zachary C Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning (ICML)*, 2018.

Xiruo Liu, Lily Yang, Ignacio Alvarez, Kathiravetpillai Sivanesan, Arvind Merwaday, Fahim Sheikh, Susruth Sarangi, Angela Perez, and Javier Gutierrez. MISO-V: Misbehavior detection for collective perception services in vehicular communications. In *IEEE Intelligent Vehicles Symposium (IV)*, 2021.

Yifan Lu, Quanhao Li, Baoan Liu, Mehrdad Dianati, Chen Feng, Siheng Chen, and Yanfeng Wang. Robust collaborative 3d object detection in presence of pose errors. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

Yifan Lu, Yue Hu, Yiqi Zhong, Dequan Wang, Yanfeng Wang, and Siheng Chen. An extensible framework for open heterogeneous collaborative perception. In *International Conference on Learning Representations (ICLR)*, 2024.

Amir Mohammadisarab, Miguel Sepulcre, Luca Lusvarghi, Sergei S. Avedisov, Mohammad Irfan Khan, Takayuki Shimizu, Onur Altintas, and Javier Gozalvez. Fusion or confusion? Potential and challenges in fusion of onboard sensors and V2X data in cooperative perception. In *IEEE Conference on Standards for Communications and Networking (CSCN)*, 2025.

Michael Motro and Joydeep Ghosh. Measurement-wise occlusion in multi-object tracking. In *International Conference on Information Fusion (FUSION)*, 2018.

Kemal Oksuz, Selim Kuzucu, Tom Joy, and Puneet K. Dokania. MoCaE: Mixture of calibrated experts significantly improves object detection. *Transactions on Machine Learning Research*, 2024.

Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European Conference on Computer Vision (ECCV)*, 2020.

Qualcomm Incorporated. Misbehavior detection using data consistency checks for collective perception messages, 2024. US Patent Application 2024/0323657 A1.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, 2018.

Hao Si, Ehsan Javanmardi, Hanlin Wu, and Manabu Tsukada. HERO: Heterogeneous evidential robust object-level collaborative perception. In *European Conference on Computer Vision (ECCV)*, 2026.

Jiachen Sun, Yulong Cao, Qi Alfred Chen, and Z. Morley Mao. Towards robust LiDAR-based perception in autonomous driving: General black-box adversarial sensor attack and countermeasures. In *USENIX Security Symposium*, 2020.

Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyuan Zeng, and Raquel Urtasun. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In *European Conference on Computer Vision (ECCV)*, 2020.

Junhao Xu, Yanan Zhang, Zhi Cai, and Di Huang. CoSDH: Communication-efficient collaborative perception via supply-demand awareness and intermediate-late hybridization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In *European Conference on Computer Vision (ECCV)*, 2022a.

Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi Ma. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2022b.

Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, and Zaiqing Nie. Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

A GROUND-TRUTH DEFINITIONS

Table 3 compares the released cooperative labels with the union definition of Section 5.

Table 3: AP50/AP70 (%) under the released cooperative labels and under the union definition.

		union GT		released cooperative GT	
setting	rule	AP50	AP70	AP50	AP70
DAIR-V2X-C	ego only	68.0	55.6	81.8	68.0
	naive late	63.2	45.1	58.0	41.0
	COVE	65.5	46.9	61.1	43.3
OPV2V	ego only	84.5	67.3	84.5	67.3
	naive late	89.7	80.9	89.7	80.9
	COVE	92.8	83.0	92.8	83.0

Under the released definition on DAIR-V2X-C the ego-only detector scores far above any fusion rule, because objects the vehicle cannot see are largely absent from the labels: late fusion can then only add false positives on that stratum. This reproduces the ordering reported by Lu et al. (2024), where late fusion scores below no fusion on DAIR-V2X-C, and it is the reason the ego-invisible population must be labelled before the question of this paper can be asked at all. On the ego-visible subset COVE against its score-mean base gains $+2.9$ $[+1.5, +5.4]$ and $+4.8$ $[+2.7, +7.4]$ (DAIR-A/B), $+7.9$ $[+5.2, +12.5]$ and $+7.4$ $[+5.2, +10.0]$ (OPV2V-A/B), $+9.2$ $[+3.8, +13.7]$ and $+7.3$ $[+4.6, +11.1]$ pp AP70 (camera A/B). On the marginal stratum the same differences are $+0.7$ $[+0.2, +1.6]$, -0.3 $[-2.6, +0.8]$, $+17.4$ $[+11.8, +23.4]$, $+11.6$ $[+6.3, +21.3]$, $+6.7$ $[+3.8, +10.0]$ and $+5.0$ $[+3.2, +7.8]$. Among remote-only clusters on the holdouts COVE vetoes 45, 55, 57, 47, 66 and 78% of the false positives and 12, 30, 7, 5, 3 and 15% of the remote-only clusters whose emitted box matches a ground-truth object of any stratum at IoU ≥ 0.7 (DAIR-A/B, OPV2V-A/B, camera A/B; `scripts/diag_strata.py`); this is a cluster-level rate and is not the same quantity as the mistaken-veto share of Appendix K, which counts AP70-eligible ego-invisible objects.

B THE DIAGNOSIS IN FULL

Table 4: The diagnosis. *Top*: a calibrator fitted where the ego can verify does not transfer to ego-invisible proposals, against the same calibrator fitted on that stratum (an oracle, fit size matched); positive is worse. *Bottom*: each correction's effect on ego-invisible AP70 vs raw-score fusion.

transfer gap (ego-visible fit → ego-invisible)	ΔNLL	ΔECE
DAIR-V2X-C	$+0.052$ [$+0.012, +0.100$]	$+0.103$ [$+0.066, +0.111$]
DAIR-V2X-C, second detector seed	$+0.061$ [$+0.020, +0.108$]	—
OPV2V	$+0.169$ [$+0.105, +0.233$]	$+0.078$ [$+0.038, +0.094$]
correction applied inside fusion	**DAIR**	**OPV2V**
context-conditioned calibrator	-11.5	-15.0
attenuate every uncorroborated cluster ($\tau = 0$)	$-4.7, -7.0$	$-28.1, -6.3$
oracle that knows each cluster's correctness	—	$+37.2$ [$+17.1, +48.0$]

C NEGATIVE RESULTS IN DETAIL

Table 5: Six elaborations of Equation 3, each with COVE's selection budget, evaluated once on the same holdouts; ΔAP70 on the ego-invisible subset against COVE (range over holdouts).

elaboration	what it adds	ΔAP70 ego-invisible
likelihood-ratio gate	a learned evidence weight	-0.3 to -7.0
multi-agent qualified jury	every agent's opportunity votes	$-5.6, -2.1$
three-state evidence	UNKNOWN/DISAGREE/CONTRADICT	ego-inv. 2.1 vs 19.1 (OPV2V-B)
population displacement prior	v marginalised over a prior	-1.1 to -2.2
per-proposal learned σ	evidential localisation uncertainty	chance-level separation
learned veto risk	abstain on the riskiest vetoes	$+0.0, -0.9$ (aggregate $+0.2, -1.1$)
for reference: unweighted consensus (silence as a vote)		-29.1

Population-level localisation prior. Verification opportunity marginalised over a displacement prior $\mathcal{N}(0, \sigma)$ along the sender's line of sight, with σ per sender and range bin estimated on the development half, and corroboration softened to $\exp(-d^2/2\sigma^2)$. Paired ΔAP70 (pp) against the tuned gate on the same holdout:

setting	holdout	all objects	ego-invisible
DAIR-V2X-C	A	-0.9 [$-1.6, -0.6$]	-1.6 [$-3.7, -0.8$]
	B	-2.0 [$-3.1, -0.8$]	-2.2 [$-5.0, +1.2$]
OPV2V	A	$+0.0$ [$-0.1, +0.1$]	-1.8 [$-5.6, -0.5$]
OPV2V, camera	A	-0.9 [$-1.4, -0.6$]	-1.8 [$-2.5, -0.9$]
	B	-1.7 [$-2.8, -0.9$]	-1.1 [$-1.8, -0.4$]

The prior estimated from IoU-matched proposals is 1.0–1.4 m for every sender; estimated from the displacement to the nearest ground-truth centre within 6 m it is 4.2–5.8 m for the camera collaborator. Neither estimate turns the marginalisation into a gain, and a softened rule without any prior performs the same, so the loss is attributable to softening the decision, not to the prior.

Per-proposal learned uncertainty. We trained evidential heads — Beta objectness and normal-inverse-gamma box regression — and replaced the population prior with each proposal's own predicted localisation uncertainty σ. The uncertainty is genuinely informative about localisation: on remote proposals matched one-to-one to ground truth, $\rho(\sigma, \|\Delta\|) = 0.697$ and median σ rises monotonically across error bands, though it is under-dispersed relative to the residuals. It is nevertheless useless here, because the operator needs to know not which boxes are badly localised but which of its vetoes are mistakes, and those are a different population. Compared directly σ appears to point the wrong way, but the comparison is confounded — a veto of a real object is by construction a

better proposal, and score alone separates the two populations with AUROC 0.805. Matching each mistaken veto to correct vetoes of the same score and range bin removes the difference entirely: $P(\sigma_{\text{mistaken}} > \sigma_{\text{correct}}) = 0.545$ at IoU ≥ 0.5 and 0.484 at IoU ≥ 0.7, both indistinguishable from chance.

Learned veto risk. The remaining target is the risk of the decision itself: for each candidate veto predict $R = P(\text{real object at IoU} \geq 0.7 \mid v_i \geq \tau, m_i{=}0, x_i)$ from deployment-observable quantities, and abstain rather than veto on the riskiest fraction, with the abstention budget selected on development like every other parameter. It does not help — against COVE on the heterogeneous holdouts, $+0.2$ $[+0.1, +0.4]$ and -1.1 $[-2.4, +0.0]$ pp aggregate AP70 — for a reason visible before any model is fitted: only 0.14% and 0.08% of candidate vetoes on the development scenes correspond to an AP70-eligible object. With the features considered here, veto-risk prediction was ineffective under a false-veto rate this extreme; we do not claim that no feature set could work.

Likelihood-ratio operator, three-state evidence, qualified jury. These three are summarised in Table 5; the likelihood-ratio form loses on every holdout because the estimated $P(\text{corroboration} \mid Y{=}1, v)$ saturates, the three-state extension does not survive stratification by visibility, and the jury does not beat the ego-only test even on objects seen by two or more collaborators. Their per-holdout paired differences are included with the released code and results.

D IMPLEMENTATION AND REPRODUCIBILITY

The evaluation protocol was frozen before any gate was run and is versioned with the code; changes are logged with the reason and the population statistics that motivated them. We verify that the sparse voxelisation used on our hardware is bit-identical to the reference implementation on four stress cases (uniform, single-voxel overflow, max-voxel overflow, range boundary). Unit tests cover label leakage, candidate freezing across rules, scene-level splitting, the monotone-invariance of AP under strictly increasing score transforms, and the geometry of the ray–box eligibility test.

E ASSOCIATION PROTOCOLS FOR THE CONTROLLED EVIDENTIAL EXPERIMENT

The evidential fusion rules we reimplement associate proposals with a dual gate: an absolute separation $\|\Delta\| \leq \tau_{\text{abs}}$ and a Mahalanobis test $d^2 = \sum_k \Delta_k^2/(\sigma_{ik}^2 + \sigma_{jk}^2) \leq \tau_{\text{mah}}$ using the predicted normal-inverse-gamma variances. The published thresholds are not reported. Choosing them to maximise development AP would let downstream performance leak into the association, so we instead set them from ground-truth correspondence alone. On development scenes we collect the pairs of proposals from two different agents that are assigned to the same ground-truth object, and take τ_{abs} as their 99th percentile separation and τ_{mah} as the 99th percentile of d^2. Equivalently — and this is the reading we prefer — one rescales the covariance by a single factor c so that those pairs attain the textbook χ^2_2 gate, after which $\tau_{\text{mah}} = 9.21$; the two are the same clustering, since scaling Σ by c^2 and gating at 9.21 is gating the raw statistic at $9.21c^2$.

The calibration is informative in itself. True matched pairs are close in space (median separation 0.43–0.48 m, 99th percentile 1.19–1.61 m across datasets and splits), while the required covariance scale is $c = 3.7$–3.8 on DAIR-V2X-C and $c = 2.6$–2.8 on OPV2V: our independently trained evidential heads are under-dispersed by roughly a factor of three, consistently across both outer splits. Applied naively, the textbook gate therefore admits only 38–64% of true matched pairs, fragmenting clusters. This is a property of our reproduction, not a diagnosis of the published method: we do not have its trained uncertainty model, and a correctly scaled covariance recovers a sensible association regime (Table 2, where the calibrated and frozen protocols agree to within 0.7 pp).

We additionally implemented the dual gate with thresholds swept on development AP over a symmetric grid ($\tau_{\text{abs}} \in \{2, 3\}$ m, $\tau_{\text{mah}} \in \{9, 25, 100, \infty\}$). That variant substantially underperformed standard late fusion on the same proposals, so we do not use it for any headline comparison and report it only as a reproduction diagnostic.

F SELECTED OPERATING POINTS

For completeness we list the development-selected (α, τ) used by the ablation of Section 5. The ray-based signal selects $(0.5, 0.5)$, $(0.5, 0.25)$, $(0.25, 0.25)$ and $(0.25, 0.5)$ on DAIR-A, DAIR-B, OPV2V-A and OPV2V-B; the metadata surrogate selects $(0.5, 0.5)$ on all four. We report this for reproducibility and do not read a mechanism into it: four splits are too few to support a claim about how sharply each signal can be thresholded.

G EVIDENTIAL COMPONENTS

Evidential components. The evidential heads of Section 5.4 are our own, obtained by warm-starting the released detectors with a Beta objectness head and a normal-inverse-gamma box-regression head that reproduce the message format of Si et al. (2026); they are not the original authors' models. On DAIR-V2X-C the Beta probability retains the detector's ranking (0.929 vs 0.945 AUROC), which is the property the controlled comparison relies on.

Training objective. Si et al. (2026) specify the evidential objectives by reference — a Beta evidential classification loss (Sensoy et al., 2018) on objectness and a normal-inverse-gamma likelihood (Amini et al., 2020) on the box targets, following a focal/SmoothL1 warm-up — without stating their weights, and at the time of writing no implementation of that paper is publicly available. We therefore implement the same Beta-EDL and normal-inverse-gamma objective *forms* that Si et al. (2026) reference, with independently chosen training weights and regularization coefficients. The Beta term is the Bayes risk of the squared error under the predictive Beta together with the Kullback–Leibler penalty on misleading evidence, $\tilde{\alpha} = y + (1-y)\alpha$; the regression term is the normal-inverse-gamma negative log-likelihood with the evidence regularizer of Amini et al. (2020). The two terms carry weights 1.0 and 0.5 against the detector losses, with 0.1 on the Kullback–Leibler penalty and 0.01 on the evidence regularizer, and anchor-class weighting as in the released detector configuration. Our Beta regularization uses a fixed coefficient rather than reproducing all scheduling details of the original evidential training recipe, and Amini et al. (2020) note that the regression coefficient has no dataset-independent setting. The scale of the learned evidence is therefore a property of our heads rather than of the aggregation rule under test. This is precisely why the comparison in Section 5.4 holds the proposals, the association and the emitted geometry fixed and varies only the aggregation: the contrast it reports is attributable to the rule, not to the evidence scale, which is shared by both arms.

H THE CONTROLLED EXPERIMENT ON OPV2V

For the controlled comparison we require an evidential head whose Beta probability preserves the detector's own ranking, so that replacing the aggregation rule is the only intervention. This holds for our DAIR-V2X-C head and for our OPV2V LiDAR head (0.962 vs 0.995 AUROC), but not for our OPV2V camera head (0.631 vs 0.879), so the heterogeneous LiDAR/camera evidential pairing is not used as a controlled baseline. Proposal-level ranking alone was not sufficient to yield a usable fused evidential baseline under our independently trained OPV2V evidence scale, so the evidential controlled intervention is restricted to DAIR-V2X-C. The conventional-semantics rows on OPV2V, computed on the evidential-head proposals and therefore at different absolute levels from Table 1, behave as in the main table: COVE improves naive late fusion from 59.0 to 62.5 AP70 (ego-invisible $14.1 \rightarrow 17.7$) on holdout A and from 63.1 to 65.3 ($0.8 \rightarrow 1.0$) on holdout B.

I METADATA SURROGATE FOR THE VERIFICATION SIGNAL

Aggregate AP alone would even prefer the surrogate on OPV2V-B (76.5 naive, 78.6 COVE, 80.3 COVE$(\hat{v})$) while the ego-invisible subset reads 10.4, 19.1 and 13.8.

Table 6: A metadata surrogate for the verification signal reproduces the aggregate gain and none of the ego-invisible gain. Paired AP70 differences (pp, 95% CI); COVE($\hat{v}$) is the same rule driven by $\hat{v}$ regressed from message-observable covariates, with its own development-selected (α, τ); both are applied on score-mean late fusion. Post-hoc ablation.

comparison	holdout	all objects	ego-invisible
COVE($\hat{v}$) vs naive late	DAIR-A	$+1.5$ [$+0.1, +2.7$]	-0.0 [$-1.6, +1.7$]
	DAIR-B	$+1.8$ [$+0.5, +3.4$]	-0.8 [$-6.6, +7.3$]
	OPV2V-A	$+2.2$ [$+1.1, +3.9$]	-26.2 [$-36.7, -11.0$]
	OPV2V-B	$+3.9$ [$+0.9, +7.8$]	$+3.4$ [$-16.0, +12.4$]
COVE($\hat{v}$) vs COVE (true v)	DAIR-A	$+0.1$ [$-0.2, +0.6$]	-2.8 [$-5.5, -0.4$]
	DAIR-B	-0.6 [$-1.3, +0.1$]	-9.5 [$-19.0, -5.4$]
	OPV2V-A	-0.4 [$-0.6, -0.1$]	-48.3 [$-57.3, -25.1$]
	OPV2V-B	$+1.7$ [$-0.3, +5.3$]	-5.3 [$-27.3, +1.5$]

J PRE-REGISTERED UNTOUCHED HOLDOUT

The six holdouts of Table 1 split each dataset's scenes by index parity and exchange the development and evaluation roles, so no number is produced on scenes whose parameters were chosen there, but the two directions share scenes. For one independent evaluation we used the OPV2V *validation* split, which had never been exported or inspected. Before it was exported we fixed, in a record committed with the code: the operator and its three regimes; the association protocol; the development source for every free parameter (the already-spent OPV2V test split, used in full); the evaluation procedure (the whole untouched split, evaluated once, with sequence-level bootstrap); and the reading of the outcome. The evaluation was then run once per association protocol and is reported as it came out. Under the calibrated association the same comparison gives $+1.70$ [$+1.22, +4.87$] pp aggregate and $+10.19$ [$+3.19, +29.56$] pp ego-invisible AP70, consistent with the frozen-association numbers in the main text. The parallel implementation used for it was verified to reproduce the reference implementation's tables and confidence intervals exactly before it was used.

K FURTHER SCOPE NOTES

Assumptions and scope (full). Eligibility is computed from ego LiDAR returns, so a camera-only ego would need a different estimate, which we have not evaluated; we assume calibrated poses and cooperative senders, and do not study adversarial or faulty ones. Both datasets are intersection-scale and the ego detector is PointPillars throughout, with one additional seed only on DAIR-V2X-C. Our ground-truth definition for ego-invisible objects departs from the released cooperative labels, so absolute AP is not comparable to published numbers and every claim is a paired difference under an identical pipeline; Appendix A repeats the headline numbers under the released definition. We compare against a re-implementation of the late-fusion attenuation rule of Xu et al. (2025), not their full system, and our evidential heads are our own rather than the original authors' models (Appendix G).

Where the abstention test is reliable. COVE assumes the ego's silence about a location it could observe is evidence about the object. The assumption fails when the collaborator's box is displaced: the ego verifies the wrong place, correctly reports nothing, and a real object is vetoed. We measured the cost. Mistaken vetoes are rare but unevenly distributed — 0.43% of the AP70-eligible ego-invisible true positives on one heterogeneous holdout and 9.2% on the other. An oracle that restores exactly those, with every other component frozen, recovers at most $+0.2$ pp aggregate AP70 anywhere, so they do not drive the aggregate result; but on the holdout where COVE loses ego-invisible accuracy against naive late fusion it recovers $+1.1$ pp, more than the -0.2 pp deficit itself. That deficit tracks the threshold rather than the mechanism: a holdout sensitivity sweep (reported, never selected on) gives ego-invisible AP70 of 10.0 at $\tau = 0.5$ against 9.3 at $\tau = 0.25$, and the development half selected $\tau = 0.25$ there. A scalar objective summing aggregate and ego-invisible AP can be won by its aggregate term, and the cost then falls on the subset the operator exists to protect. This sensitivity tracks the development-selected operating point, indicating that model selection contributes materially to the observed deficit.

Finally, observability-conditioned updating is established in neighbouring sensing problems; our contribution is to show that learned object-level collaborative detection violates it systematically, to identify why standard benchmark annotations obscure the cost of that violation, and to show that correcting the admissibility of negative evidence recovers the effect across the scoring rules and holdouts we test. We do not claim that COVE is optimal; rather, six more expressive extensions of the update provide no consistent held-out improvement under the same selection budget.

L EGO-PRESERVING FUSION: A SYSTEM-LEVEL VARIANT

Section 5 shows that on DAIR-V2X-C late fusion is below the ego alone in aggregate AP70 because mixed clusters often emit the worse-localised remote box. A system-level variant asks whether collaboration under COVE can beat the ego alone: a cluster containing an ego proposal emits the ego box with the ego score, which no collaborator may alter; a remote-only cluster emits its max-score member with its score multiplied by a cross-agent factor γ; and COVE acts on those remote-only clusters. Because the emitted box changes in mixed clusters, this is *not* part of the fixed-geometry controlled comparison of Tables 1–2. The reading rule was fixed before any holdout was evaluated and is recorded with the code: each arm is tuned independently on the development half under a matched budget of 16 operating points (ego-preserving: γ over 16 values in $[0.25, 1]$; with COVE: $(\alpha, \tau) \in \{0.25, 0.5\}^2 \times \gamma \in \{0.25, 0.5, 0.75, 1\}$), the holdout is evaluated once, and the variant would be used to support the main claim only if COVE beat the ego-preserving base on ego-invisible AP70 with a paired interval excluding zero without losing aggregate AP70. Three of the four holdouts meet the rule and DAIR-A does not on the ego-invisible criterion (Table 7); we therefore report it as a system-level scope analysis rather than as evidence for the main claim. The development half selected $\gamma = 1$ for the COVE arm on every holdout and $\gamma \in \{0.9, 1\}$ for the ego-preserving arm.

Table 7: Ego-preserving fusion on the four LiDAR–LiDAR holdouts. Top: AP70 all objects / ego-invisible for each cell. Bottom: paired differences (pp, 95% CI). The emitted box changes in mixed clusters, so this is a system-level comparison, not the fixed-geometry one; here COVE sits on max-score fusion, and Table 2's max-score row is computed on the evidential-head proposals, hence its slightly different values.

holdout	ego only	max-score	max-score +COVE	ego-preserving	ego-preserving +COVE
DAIR-A	58.7 / 0.0	48.6 / 8.5	50.0 / 8.5	57.7 / 12.9	60.3 / 13.6
DAIR-B	51.6 / 0.0	41.0 / 14.8	43.3 / 15.5	56.5 / 22.8	57.5 / 39.6
OPV2V-A	65.5 / 0.0	85.3 / 52.8	89.0 / 78.5	69.8 / 45.4	73.1 / 80.1
OPV2V-B	69.1 / 0.0	76.5 / 10.4	79.2 / 14.5	64.3 / 11.5	68.4 / 17.6

	+COVE − ego-preserving		ego-preserving +COVE − ego only		
	all	ego-invisible	all	ego-invisible	rule
DAIR-A	$+2.7$ [$+1.6, +4.0$]	$+0.7$ [$-3.4, +4.1$]	$+1.6$ [$+0.5, +3.2$]	$+13.6$ [$+7.9, +21.4$]	fail
DAIR-B	$+1.0$ [$+0.4, +1.7$]	$+16.8$ [$+9.1, +21.3$]	$+5.9$ [$+3.3, +8.1$]	$+39.6$ [$+21.8, +54.1$]	pass
OPV2V-A	$+3.3$ [$+2.4, +4.5$]	$+34.7$ [$+21.7, +41.5$]	$+7.6$ [$+3.3, +11.5$]	$+80.1$ [$+53.8, +88.2$]	pass
OPV2V-B	$+4.2$ [$+2.6, +6.5$]	$+6.0$ [$+2.1, +17.7$]	-0.7 [$-5.1, +2.2$]	$+17.6$ [$+8.6, +52.4$]	pass

Preserving the ego box by itself recovers most of the DAIR deficit (-1.1 and $+4.9$ pp against the ego alone on A and B) and COVE adds aggregate accuracy on top on every holdout; on OPV2V preserving the ego box *costs* aggregate accuracy (-4.9 pp on B), because simulated remote boxes are the better-localised ones, which is the same mechanism seen from the other side. A γ sweep on the holdout with COVE's (α, τ) frozen from development — reported, never selected on — shows the advantage of COVE over the ego-preserving base is positive over the whole range on OPV2V (aggregate $+0.0$ to $+5.7$ pp, ego-invisible $+0.1$ to $+27.4$ pp, increasing monotonically in γ) and on DAIR is within ± 0.5 pp for $\gamma \leq 0.6$ before growing at $\gamma = 1$, which is expected, since a small γ already discounts the remote-only clusters that COVE acts on. The sweep holds both arms at the same γ, whereas Table 7 uses each arm's own development-selected γ, so its end point need not equal the tabulated difference.

M A SECOND EGO-LIDAR OBSERVABILITY PROXY: THE IN-BOX RETURN COUNT

The ego-invisible stratum is defined by ego returns and v is computed from ego rays, so one may ask whether the ray geometry is doing anything that a simpler ego-LiDAR quantity would not. The control replaces v by the number of ego returns inside the remote box, keeping everything else identical (score-mean base, frozen clusters, cluster-level gate on remote-only clusters, development-half selection, one holdout evaluation, paired sequence bootstrap). Two semantics were offered to the selection — veto if the count is at least k (*presence*: the ego sensed something there and did not detect it) or below k (*empty*) — over $k \in \{1, 2, 5, 10, 20, 50, 100, 200\} \times \alpha \in \{0.25, 0.5\}$, a larger budget than COVE's four points. The reading rule, fixed beforehand and recorded with the code, counted the result as a positive real-data answer only if COVE beat the selected count gate on ego-invisible AP70 with an interval excluding zero on a DAIR holdout and never lost to it significantly. It did not: the selection chose *presence* with $k = 1$ everywhere, and on both real-data holdouts the count gate beats COVE on the ego-invisible subset (Table 8). That column must be read with its construction in mind: ego-invisible status is defined by the absence of ego returns inside the *ground-truth* box, whereas the count gate measures returns inside the *remote proposal* box. The two quantities are strongly coupled but not identical — a displaced proposal may contain ego returns where the ground-truth box has none — so the selected presence gate is structurally biased towards preserving the zero-return stratum without being unable to veto it. The aggregate and marginal comparisons are not subject to this direct coupling. On aggregate AP70 the two are level on DAIR-A (-0.8 $[-2.5, +0.2]$) and COVE is ahead on DAIR-B ($+0.6$ $[+0.1, +1.4]$), so on real data the result is a trade-off rather than a strictly better rule. On the simulated and heterogeneous holdouts the count gate barely moves the ego-invisible subset while COVE gains up to $+20.8$ pp there, and COVE is better on the marginal stratum on every holdout. We read this as two complementary observability proxies rather than as a refutation of the admissibility condition, which both implement: the count gate can only veto proposals whose boxes contain ego returns and therefore tends to preserve hypotheses over zero-return regions, which also makes it less sensitive to displaced remote boxes on real data; the ray test can veto a false positive in free space, which is what the simulated false positives are, and it is the one that distinguishes an occluded region from an empty one. The main text's claim is about the admissibility condition; which ego-side quantity should instantiate it on real data is left open by this result.

Table 8: The in-box return count as the gate, against the ray-based v, both on score-mean late fusion. Absolute ego-invisible AP70 (base / +COVE / +count) and paired differences of COVE against the count gate (pp, 95% CI).

holdout	selected count gate	ego-inv. base / +COVE / +count	all: base / +COVE / +count	COVE − count, ego-inv.	all	marginal
DAIR-A	presence, $k=1$, $\alpha=0.5$	10.9 / 11.3 / 14.5	48.1 / 50.0 / 50.8	-3.2 $[-14.5, -0.6]$	-0.8 $[-2.5, +0.2]$	$+1.5$ $[+0.2, +2.3]$
DAIR-B	presence, $k=1$, $\alpha=0.5$	21.1 / 23.5 / 26.9	40.2 / 43.4 / 42.7	-3.4 $[-9.3, -0.3]$	$+0.6$ $[+0.1, +1.4]$	$+1.1$ $[+0.4, +2.4]$
OPV2V-A	presence, $k=1$, $\alpha=0.5$	54.1 / 74.9 / 54.6	82.8 / 87.9 / 83.0	$+20.3$ $[+11.6, +31.3]$	$+4.9$ $[+3.1, +7.9]$	$+17.3$ $[+11.7, +23.3]$
OPV2V-B	presence, $k=1$, $\alpha=0.5$	13.3 / 19.1 / 15.5	74.3 / 78.6 / 75.5	$+3.6$ $[+0.0, +21.2]$	$+3.1$ $[+1.2, +5.8]$	$+10.1$ $[+3.7, +21.0]$
camera-A	presence, $k=1$, $\alpha=0.25$	11.0 / 20.4 / 11.3	56.1 / 61.8 / 56.9	$+9.1$ $[+1.1, +12.5]$	$+4.9$ $[+1.9, +7.8]$	$+5.7$ $[+2.6, +9.6]$
camera-B	presence, $k=1$, $\alpha=0.25$	7.3 / 9.3 / 9.7	57.7 / 61.8 / 59.9	-0.4 $[-2.4, +4.7]$	$+1.9$ $[+0.3, +4.7]$	$+3.3$ $[+1.4, +6.8]$

N ADDITIONAL FIGURES

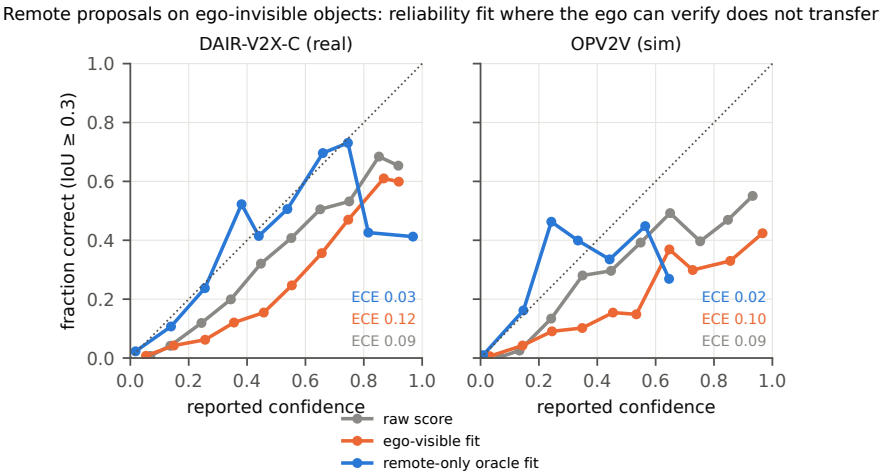


Figure 3: Remote proposals, stratified by whether the ego can see the object. A calibrator fitted where the ego can verify (orange) is over-confident on ego-invisible objects in every bin; a calibrator fitted on the ego-invisible stratum itself (blue, an oracle unavailable at deployment) is close to the diagonal. The ranking ability of the scores is preserved (AUROC 0.92); it is the meaning of the number that breaks.

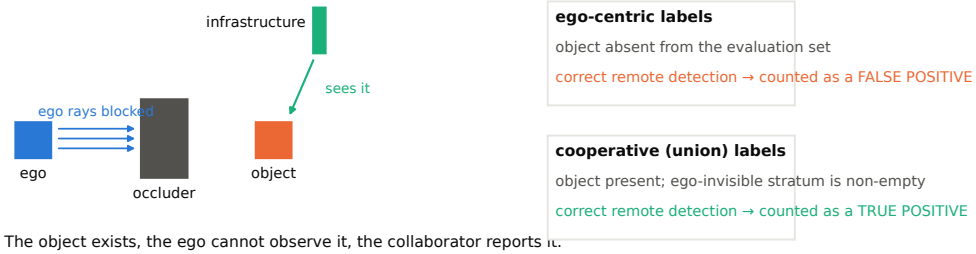


Figure 4: If an object the ego cannot observe is absent from the evaluation labels, a collaborator that reports it correctly is scored as a false positive, and the very objects collaboration exists for cannot be measured. On DAIR-V2X-C the released cooperative labels omit 31% of the in-range infrastructure-side labels, most of them without a single ego return, so this is the situation by default.

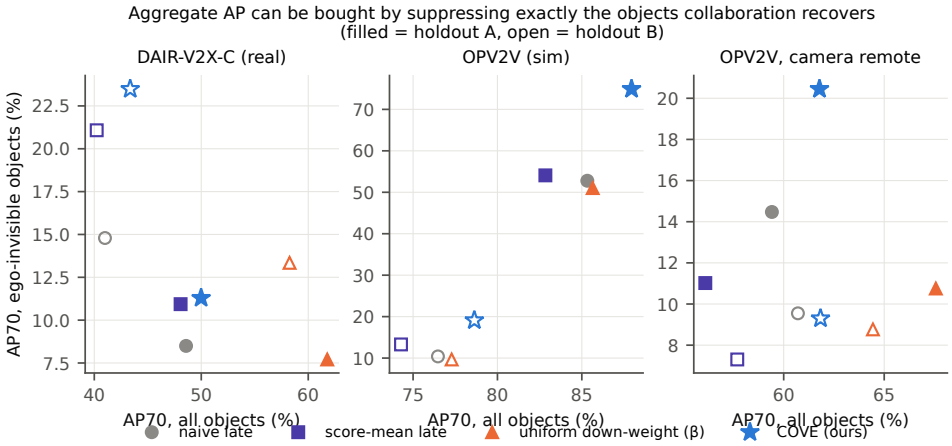


Figure 5: Aggregate versus ego-invisible AP70 on the frozen holdouts (filled: holdout A, open: B). Uniform attenuation moves right and down — it buys aggregate accuracy by suppressing the objects that only the collaborator can see. CoVE moves right and up.