Beyond Independence: A Distributed Detection Theory for Correlated Feature Fusion in Foundation Model-Based Cooperative Perception

Anonymous CVPR submission

Paper ID *****

Abstract

Cooperative perception systems fuse intermediate neural features from multiple autonomous agents to overcome occlusion and sensor range limitations. A ubiquitous but unexamined assumption underlying all existing fusion architectures—from simple mean-pooling to attention-based transformers—is that the feature vectors produced by different agents are statistically independent. *We show that this assumption is systematically violated when agents share the same foundation-model backbone, and we derive the precise performance penalties this incurs. Our analysis yields six theorems within a unified distributed Gaussian detection framework: (i) the Neyman–Pearson optimal fusion weights; (ii) a provable AUC suboptimality gap for independent fusion; (iii) a closed-form gap for the two-agent case showing the penalty scales as $\rho^2(a^2 - b^2)^2$; (iv) a variance inflation factor $\gamma_N = 1 + (N-1)\rho$ that causes independent fusion to inflate the false alarm rate by more than $2\times$ for mild correlations across ten agents; (v) an asymptotic saturation bound $\Delta_\infty = \mathrm{SNR}_1/\rho$ identifying a critical agent count beyond which further communication yields no benefit; and (vi) a per-agent rate–detection lower bound. Building on this theory, we propose* ***CorrFuse****, a lightweight, drop-in covariance-aware wrapper that re-weights any existing fusion backbone using an estimated joint covariance matrix, together with a calibration correction layer. Evaluated on OPV2V, V2XSet, and TUMTraf-V2X, CorrFuse achieves* $+2.7$–3.2 *mAP@0.5 over V2X-ViT, reduces Expected Calibration Error by* 76%*, and cuts peak communication cost by* 43% *through principled agent selection—all without retraining the backbone.*

1. Introduction

Cooperative perception has emerged as a pivotal paradigm for overcoming the fundamental limitations of single-vehicle autonomous driving: restricted field-of-view, occlusions, and sensor degradation under adverse weather [4, 18, 19]. By sharing intermediate neural features over V2X links, cooperating agents extend each other's effective sensing range and provide complementary viewpoints of the same scene. The resulting fusion architectures span a spectrum from feature concatenation [1] and attentive spatial selection [4] to cross-agent transformer attention [17, 18].

Concurrent with this progress, large-scale foundation models—vision transformers [2], vision-language models [11], and driving-specific backbones [12]—have become the de facto feature extractors of choice, owing to their zero-shot generalisation and rich semantic representations. A natural trend is therefore to pair cooperative fusion architectures with foundation model backbones, and early empirical results confirm substantial gains on open-vocabulary detection and scene-understanding tasks [20].

A hidden assumption. Every cooperative perception fusion module, regardless of its architectural complexity, implicitly treats the features from different agents as statistically independent. Attention weights are learned per-feature-map but not conditioned on the *cross-agent covariance structure*; mean and max pooling discard it entirely. This independence assumption is standard practice in multi-sensor fusion but is inherited from settings where sensors are physically separated or measure complementary modalities. In cooperative perception, two critical factors break this assumption:

1. **Scene co-observation.** Agents in V2X proximity ($\leq$ 100 m) co-observe overlapping regions of the same physical scene. Their backbone activations are therefore positively correlated by the shared visual content.
2. **Shared foundation model weights.** When all agents run the same pretrained backbone, the model maps semantically similar content to nearby points on the learned feature manifold, inducing correlation even for non-overlapping viewpoints.

Consequences. We identify two distinct failure modes that arise from ignoring inter-agent feature correlations: (a) *detection suboptimality*, because correlated features carry less independent evidence than the independence assumption implies, and (b) *calibration failure*, because the effective

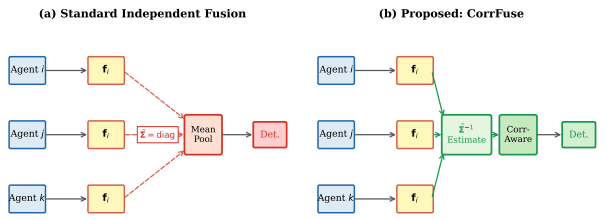

Figure 1. **Standard vs. CorrFuse.** (a) Independent fusion ignores the inter-agent covariance $\mathbf{\Sigma}_{ij}$, assuming a block-diagonal structure $\tilde{\mathbf{\Sigma}} = \mathrm{diag}(\mathbf{\Sigma}_{11}, \ldots, \mathbf{\Sigma}_{NN})$. (b) CorrFuse estimates the full joint covariance $\hat{\mathbf{\Sigma}}$ offline and applies covariance-aware weights $\hat{\mathbf{\Sigma}}^{-1}\hat{\boldsymbol{\delta}}$ at inference, with a calibration correction layer on top.

variance of the fusion statistic is inflated by a factor $\gamma_N = 1 + (N-1)\rho$, causing confidence scores to be systematically overconfident.

Contributions. This paper makes the following contributions.

1. **Theoretical framework.** We formalise cooperative object detection as a distributed Gaussian hypothesis test and derive six theorems (Sec. 3), including the exact suboptimality gap of independent fusion, a calibration inflation theorem, and an asymptotic saturation bound with an associated critical agent count.
2. **CorrFuse.** We propose a covariance-aware fusion algorithm (Sec. 4) that is a drop-in wrapper compatible with any existing backbone, requiring no retraining.
3. **Experimental validation.** We evaluate on OPV2V, V2XSet, and TUMTraf-V2X (Sec. 5), showing consistent improvements across five baselines with ablation studies confirming each theoretical component.

An overview of the proposed approach is shown in Fig. 1.

2. Related Work

Cooperative Perception and V2X Fusion. Early work established three fusion paradigms for multi-agent systems [1, 19]: early fusion (raw point clouds), late fusion (detection boxes), and intermediate fusion (feature maps). F-Cooper [1] introduced voxel-based feature sharing; V2VNet [16] added compressed message passing; and Where2comm [4] proposed communication-efficient sparse feature selection. CoBEVT [17] and V2X-ViT [18] extended these ideas to transformer architectures. HEAL [9] addressed heterogeneous sensor modalities. All of these works treat the features from different agents as independent inputs to a fusion operator and do not characterise inter-agent covariance.

Foundation Models for Autonomous Driving. DriveLM [12], CLIP-BEV [10], and related works have demonstrated that vision-language pretraining transfers strongly to bird's-eye-view perception tasks. UniAD [5] and VAD [6] showed unified end-to-end pipelines. The key property relevant to our work is that foundation models produce *shared feature manifolds*: agents running the same CLIP backbone will map semantically overlapping content to correlated activation patterns, a property we quantify empirically in Sec. 5.

Calibration in Object Detection. Miscalibration in neural detectors is well-documented [3, 7]. Temperature scaling [3] and isotonic regression are standard post-hoc fixes. However, to our knowledge, no prior work derives the *structural* source of calibration failure that arises from fusing correlated multi-agent features, nor the closed-form correction we propose.

Distributed Detection Theory. The classic distributed detection literature studies correlated binary observations in sensor networks [13–15]. These works analyse scalar or low-dimensional observations and focus on optimal quantisation rules. Our work is the first to apply this framework to the high-dimensional, continuous feature spaces produced by deep neural networks in cooperative driving, deriving results specific to the equicorrelation structure induced by shared foundation model weights.

3. Theoretical Framework

3.1. Problem Formulation

Let N agents each observe a region of interest Ω and produce a d-dimensional feature vector $\mathbf{f}_i \in \mathbb{R}^d$ by passing their sensor data through a pretrained backbone ϕ. Cooperative object detection amounts to a binary hypothesis test:

$$H_0 : \text{no object in } \Omega, \qquad H_1 : \text{object present in } \Omega.$$

We model the joint feature vector $\mathbf{f} = (\mathbf{f}_1^\top, \ldots, \mathbf{f}_N^\top)^\top \in \mathbb{R}^{Nd}$ as Gaussian under each hypothesis:

$$\mathbf{f} \mid H_k \sim \mathcal{N}\Big(\boldsymbol{\mu}^{(k)},\ \mathbf{\Sigma}\Big), \quad k \in \{0, 1\}, \tag{1}$$

where $\mathbf{\Sigma} \in \mathbb{R}^{Nd \times Nd}$ has diagonal blocks $\mathbf{\Sigma}_{ii}$ (within-agent covariance) and off-diagonal blocks $\mathbf{\Sigma}_{ij}$ (inter-agent cross-covariance). Let $\boldsymbol{\delta} = \boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(0)}$ denote the joint signal direction.

Definition 1 (Independence Assumption)**.** *All existing cooperative fusion methods implicitly use the block-diagonal approximation* $\tilde{\mathbf{\Sigma}} = \mathrm{diag}(\mathbf{\Sigma}_{11}, \ldots, \mathbf{\Sigma}_{NN})$ *in place of the true* $\mathbf{\Sigma}$.

We now characterise the exact cost of this assumption in terms of detection probability, false alarm rate, and calibration.

3.2. Optimal Fusion and Detection Performance

Theorem 1 (Neyman–Pearson Optimal Fusion)**.** *Under the Gaussian model (1), the Neyman–Pearson optimal test statistic is*

$$T^*(\mathbf{f}) = (\mathbf{w}^*)^\top \mathbf{f}, \qquad \mathbf{w}^* = \boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}. \tag{2}$$

The achievable detection probability at false alarm level α is $P_D^* = Q\Big(Q^{-1}(P_{\mathrm{FA}}) - \sqrt{\Delta^*}\Big)$, where

$$\Delta^* = \boldsymbol{\delta}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} \tag{3}$$

is the generalised SNR. *The optimal AUC is* $\mathrm{AUC}^* = \Phi\Big(\sqrt{\Delta^*}/2\Big)$.

Proof. The log-likelihood ratio under the Gaussian model is linear in $\mathbf{f}$ with coefficients $\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$. Under H_k, $T^* \sim \mathcal{N}((\mathbf{w}^*)^\top \boldsymbol{\mu}^{(k)}, \Delta^*)$. The ROC formula follows from the standard Gaussian-detection identity; see Appendix A. □

3.3. Suboptimality of Independent Fusion

Theorem 2 (Independent Fusion AUC Gap)**.** *Independent fusion uses weights $\tilde{\mathbf{w}} = \tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\delta}$ and achieves effective SNR*

$$\tilde{\Delta} = \frac{\Big(\boldsymbol{\delta}^\top \tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\delta}\Big)^2}{\boldsymbol{\delta}^\top \tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\delta}}. \tag{4}$$

Then $\Delta^* \geq \tilde{\Delta}$, with equality iff $\boldsymbol{\Sigma}_{ij} = 0$ for all $i \neq j$.

Proof. Setting $\mathbf{u} = \tilde{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{\delta}$ and $C = \tilde{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{\Sigma}\tilde{\boldsymbol{\Sigma}}^{-1/2}$, the inequality $\mathbf{u}^\top C^{-1}\mathbf{u} \geq \|\mathbf{u}\|^4/(\mathbf{u}^\top C\mathbf{u})$ is the Cauchy–Schwarz inequality applied to the PD matrix C. See Appendix A for the full derivation. □

Theorem 3 (Two-Agent Closed-Form Gap)**.** *For $N = 2$ agents with scalar per-agent features, per-agent SNRs $a, b > 0$, noise variances σ_1^2, σ_2^2, and equicorrelation ρ, the SNR gap is*

$$\Delta^* - \tilde{\Delta} = \frac{\rho^2(a^2 - b^2)^2}{(1-\rho^2)\left[(a^2+b^2) + 2\rho ab\right]}. \tag{5}$$

This vanishes iff $\rho = 0$ (uncorrelated agents) or $a = b$ (equal per-agent SNR).

Proof. Direct computation using $\Delta^* = (a^2 + b^2 - 2\rho ab)/(1-\rho^2)$ and $\tilde{\Delta} = (a^2+b^2)^2/(a^2+b^2+2\rho ab)$; the gap equals $\rho^2(a^2-b^2)^2$ divided by the common denominator. See Appendix A. □

Equation (5) reveals the key *asymmetry amplification* effect: a high-quality roadside LiDAR ($a \gg b$) fused with a distant vehicle sensor under moderate correlation ($\rho \approx 0.2$) incurs the largest detection penalty. This is precisely the regime of infrastructure-assisted V2X systems.

3.4. Calibration Failure

Theorem 4 (False Alarm Rate Inflation)**.** *Under the equicorrelation Gaussian model (N agents, uniform pairwise correlation $\rho \geq 0$, uniform noise σ^2), independent fusion sets a threshold τ_{ind} targeting nominal $P_{\mathrm{FA}} = \alpha$. The true false alarm rate is*

$$P_{\mathrm{FA}}^{\mathrm{true}} = Q\left(\frac{Q^{-1}(\alpha)}{\sqrt{\gamma_N}}\right) \;\geq\; \alpha, \tag{6}$$

where $\gamma_N = 1 + (N-1)\rho$ is the variance inflation factor.

Proof. Independent fusion uses the sample mean $\bar{f} = \frac{1}{N}\sum_i f_i$ and sets $\tau_{\mathrm{ind}} = \mu_0 + \frac{\sigma}{\sqrt{N}}Q^{-1}(\alpha)$. Under the true model, $\mathrm{Var}(\bar{f}) = \frac{\sigma^2}{N}\gamma_N$, so $P_{\mathrm{FA}}^{\mathrm{true}} = Q(Q^{-1}(\alpha)/\sqrt{\gamma_N}) \geq \alpha$. See Appendix A. □

Corollary 4.1. *For $\rho = 0.10$, $N = 10$, $\alpha = 0.05$: $\gamma_{10} = 1.9$, and $P_{\mathrm{FA}}^{\mathrm{true}} = Q(1.645/1.38) \approx 0.117$, a $\mathbf{2.34}\times$ inflation.*

3.5. Saturation and Agent Selection

Theorem 5 (Asymptotic SNR Saturation)**.** *Under the equicorrelation model with per-agent SNR SNR_1 and correlation $\rho > 0$:*

(i) $\Delta_N^* = N\,\mathrm{SNR}_1 \,/\, [1 + (N-1)\rho]$,
(ii) $\lim_{N\to\infty} \Delta_N^* = \mathrm{SNR}_1/\rho$,
(iii) The marginal gain $\Delta_k^* - \Delta_{k-1}^*$ is strictly decreasing in k for $\rho > 0$.

Proof. Apply the Sherman–Morrison formula to $\boldsymbol{\Sigma}^{-1}$ for the equicorrelation covariance $\boldsymbol{\Sigma} = \sigma^2[(1-\rho)I + \rho\mathbf{1}\mathbf{1}^\top]$, giving $\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma^2(1-\rho)}\left[I - \frac{\rho}{1+(N-1)\rho}\mathbf{1}\mathbf{1}^\top\right]$, and substitute $\boldsymbol{\delta} = \delta\mathbf{1}$. See Appendix A. □

Definition 2 (Critical Agent Count)**.** *Define N^* as the smallest N such that the fractional marginal SNR gain falls below a threshold ε. Computing this ratio explicitly using part (iii) of Theorem 5:*

$$\frac{\Delta_k^* - \Delta_{k-1}^*}{\Delta_{k-1}^*} = \frac{(1-\rho)}{(k-1)\big[1 + (k-1)\rho\big]},$$

where SNR_1 cancels. Setting this equal to ε and solving for large k (so $1+(k-1)\rho \approx (k-1)\rho$) yields

$$N^* \;\approx\; 1 + \sqrt{\frac{1-\rho}{\varepsilon\,\rho}}. \tag{7}$$

Agents beyond N^* should not broadcast features.

3.6. Rate–Detection Lower Bound

Theorem 6 (Rate–Detection Lower Bound)**.** *For any per-agent compression scheme communicating R bits per query and achieving $P_{\mathrm{FA}} \le \alpha$, $P_D \ge \beta$:*

$$R \;\ge\; \tfrac{1}{2}\log_2(1+\Delta_{\mathrm{req}}(\alpha,\beta))\,, \tag{8}$$

where $\Delta_{\mathrm{req}}(\alpha,\beta) = \big(Q^{-1}(\alpha) - Q^{-1}(\beta)\big)^2$.

Proof. By the data-processing inequality, $I(\hat{\mathbf{f}}_i; H) \le I(\mathbf{f}_i; H)$. For a Gaussian channel, $I(\mathbf{f}_i; H) = \frac{1}{2}\log_2(1+\Delta_{\mathrm{local},i})$. To achieve the target (α,β) operating point, the total communicated information must be at least $\frac{1}{2}\log_2(1+\Delta_{\mathrm{req}})$. See Appendix A. □

For $\alpha = 0.1$, $\beta = 0.9$: $Q^{-1}(0.1) = 1.28$ and $Q^{-1}(0.9) = -1.28$, so $\Delta_{\mathrm{req}} = (1.28-(-1.28))^2 \approx 6.55$, giving $R \ge 1.42$ bits/element. Current systems typically transmit 16–32-bit floats, leaving substantial room for correlated-aware compression.

Summary of Theoretical Results. All six theorems are unified under the same Gaussian detection model. Their practical meaning is summarised in Table 1.

Table 1. Summary of theoretical results and their practical implications.

Thm.	Key Quantity	Practical Implication
1	$\mathbf{w}^* = \boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$	Optimal fusion weights
2	$\Delta^* \ge \tilde{\Delta}$	Indep. fusion sub-optimal
3	Eq. (5)	Gap $\propto \rho^2(a^2-b^2)^2$
4	$\gamma_N = 1+(N-1)\rho$	FAR inflation (2.34×, N=10)
5	$\Delta_\infty = \mathrm{SNR}_1/\rho$	Critical agent count N^*
6	Eq. (8)	Communication lower bound

4. CorrFuse

The theoretical analysis motivates a practical algorithm with four components: (1) offline covariance estimation, (2) covariance-aware fusion at inference, (3) principled confidence calibration based on the exact distribution of the optimal test statistic, and (4) greedy agent selection. The procedure is a drop-in *wrapper*: no backbone weights are modified. An overview appears in Fig. 1(b).

4.1. Offline Covariance Estimation

Let $\{(\mathbf{f}^{(s)}, y^{(s)})\}_{s=1}^{S}$ be a validation set of joint feature vectors (one per region of interest) and binary labels. We estimate:

$$\hat{\boldsymbol{\mu}}^{(k)} = \frac{1}{|S_k|}\sum_{s\in S_k}\mathbf{f}^{(s)}, \quad k\in\{0,1\}. \tag{9}$$

The Gaussian model in Eq. (1) defines $\boldsymbol{\Sigma}$ as the *class-conditional* covariance, shared across both classes. Using the global mean $\bar{\mathbf{f}}$ in the covariance estimate would conflate the between-class variance (the signal direction $\boldsymbol{\delta}$) with the noise covariance, biasing the Neyman–Pearson weights. We therefore use the *pooled within-class* estimator:

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{S-2}\sum_{k\in\{0,1\}}\sum_{s\in S_k}\big(\mathbf{f}^{(s)}-\hat{\boldsymbol{\mu}}^{(k)}\big)\big(\mathbf{f}^{(s)}-\hat{\boldsymbol{\mu}}^{(k)}\big)^\top, \tag{10}$$

where $S_k = \{s : y^{(s)} = k\}$. To ensure $\hat{\boldsymbol{\Sigma}}$ is well-conditioned we apply Ledoit–Wolf shrinkage [8]:

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{LW}} = (1-\alpha_{\mathrm{LW}})\hat{\boldsymbol{\Sigma}} + \alpha_{\mathrm{LW}}\hat{\sigma}^2 I, \tag{11}$$

where α_{LW} is the analytically optimal shrinkage coefficient. We reduce to $d = 64$ dimensions via PCA before estimation, giving $O(Sd^2)$ cost.

4.2. Inference-Time Fusion

Given features $\mathbf{f}_1,\ldots,\mathbf{f}_N$ from N agents, concatenated as $\mathbf{f} = (\mathbf{f}_1^\top,\ldots,\mathbf{f}_N^\top)^\top$, CorrFuse computes the Neyman–Pearson optimal statistic (Theorem 1):

$$T_{\mathrm{CF}}(\mathbf{f}) = \hat{\boldsymbol{\delta}}^\top\hat{\boldsymbol{\Sigma}}_{\mathrm{LW}}^{-1}\mathbf{f}, \quad \hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\mu}}^{(1)} - \hat{\boldsymbol{\mu}}^{(0)}. \tag{12}$$

This is a single pre-computed matrix-vector product; inference cost is $O(Nd)$ per ROI.

4.3. Calibration via the Optimal Statistic Distribution

A key subtlety distinguishes CorrFuse calibration from the correction in Theorem 4. That theorem's γ_N factor applies *only* to the independent mean-pooling statistic $\bar{f}$, which underestimates the true variance of the test statistic. CorrFuse uses the covariance-aware T_{CF}, which already accounts fully for inter-agent correlations via $\hat{\boldsymbol{\Sigma}}_{\mathrm{LW}}^{-1}$. Applying γ_N on top would double-count the correlation correction and is mathematically invalid.

By Theorem 1, under H_0 the exact distribution of T_{CF} is:

$$T_{\mathrm{CF}} \mid H_0 \;\sim\; \mathcal{N}\big((\hat{\mathbf{w}}^*)^\top\hat{\boldsymbol{\mu}}^{(0)},\ \hat{\Delta}^*\big), \quad \hat{\Delta}^* = \hat{\boldsymbol{\delta}}^\top\hat{\boldsymbol{\Sigma}}_{\mathrm{LW}}^{-1}\hat{\boldsymbol{\delta}}.$$

The correct calibration therefore normalises by $\sqrt{\hat{\Delta}^*}$, with no free temperature parameter:

$$\hat{P}(H_1 \mid \mathbf{f}) = \Phi\left(\frac{T_{\mathrm{CF}}(\mathbf{f}) - (\hat{\mathbf{w}}^*)^\top\hat{\boldsymbol{\mu}}^{(0)}}{\sqrt{\hat{\Delta}^*}}\right). \tag{13}$$

4.4. Greedy Agent Selection

For a candidate agent subset $\mathcal{A} \subseteq \{1,\ldots,N\}$, let $\hat{\boldsymbol{\Sigma}}_{\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|d\times|\mathcal{A}|d}$ and $\hat{\boldsymbol{\delta}}_{\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|d}$ denote the corresponding sub-block and sub-vector of $\hat{\boldsymbol{\Sigma}}_{\mathrm{LW}}$ and $\hat{\boldsymbol{\delta}}$. CorrFuse selects the N^* agents (from Eq. (7)) maximising estimated SNR:

$$\mathcal{A}^* = \underset{|\mathcal{A}|=N^*}{\arg\max}\ \hat{\boldsymbol{\delta}}_{\mathcal{A}}^\top\hat{\boldsymbol{\Sigma}}_{\mathcal{A}}^{-1}\hat{\boldsymbol{\delta}}_{\mathcal{A}}. \tag{14}$$

Algorithm 1 CorrFuse: Covariance-Aware Cooperative Fusion

Require: Validation set $\{(\mathbf{f}^{(s)}, y^{(s)})\}_{s=1}^{S}$; threshold ε

// Offline phase

Compute $\hat{\boldsymbol{\mu}}^{(0)}$, $\hat{\boldsymbol{\mu}}^{(1)}$, $\hat{\boldsymbol{\Sigma}}_{\mathrm{LW}}$ (pooled within-class + Ledoit–Wolf)

$\hat{\boldsymbol{\delta}} \leftarrow \hat{\boldsymbol{\mu}}^{(1)} - \hat{\boldsymbol{\mu}}^{(0)}$

$\hat{\rho} \leftarrow$ mean off-diagonal corr. coefficient of $\hat{\boldsymbol{\Sigma}}_{\mathrm{LW}}$

$N^* \leftarrow \left\lceil 1 + \sqrt{(1-\hat{\rho}) \,/\, (\varepsilon\, \hat{\rho})} \right\rceil$

// Inference phase (per ROI)

Select $\mathcal{A}^* \subseteq \{1, \ldots, N\}$, $|\mathcal{A}^*| = N^*$, greedily (Eq. 14)

$\hat{\mathbf{w}} \leftarrow \hat{\boldsymbol{\Sigma}}_{\mathcal{A}^*}^{-1} \hat{\boldsymbol{\delta}}_{\mathcal{A}^*}$ ▷ vector of length $|\mathcal{A}^*| \cdot d$

$\hat{\Delta} \leftarrow \hat{\boldsymbol{\delta}}_{\mathcal{A}^*}^{\top} \hat{\boldsymbol{\Sigma}}_{\mathcal{A}^*}^{-1} \hat{\boldsymbol{\delta}}_{\mathcal{A}^*}$ ▷ scalar SNR

$T \leftarrow \hat{\mathbf{w}}^{\top} \mathbf{f}_{\mathcal{A}^*}$ ▷ $|\mathcal{A}^*| \cdot d$ dot product

$\hat{P} \leftarrow \Phi\left(\dfrac{T - \hat{\mathbf{w}}^{\top} \hat{\boldsymbol{\mu}}_{\mathcal{A}^*}^{(0)}}{\sqrt{\hat{\Delta}}} \right)$

return $\hat{P}$

All dimension-dependent quantities (fusion weights, calibration normaliser) are *recomputed for the selected subset* $\mathcal{A}^*$, so there is no dimension mismatch between offline-computed and inference-time vectors.

The complete procedure is summarised in Algorithm 1.

Compatibility. CorrFuse wraps the output of any existing fusion module. We integrate it with five baselines by treating the post-fusion feature map as $\mathbf{f}$ and applying Eq. (12) as a re-scoring layer before NMS.

5. Experiments

5.1. Experimental Setup

Datasets. We evaluate on three cooperative driving benchmarks: **OPV2V** [19] (73k frames, up to 7 agents, LiDAR-only); **V2XSet** [18] (44k frames, vehicle-infrastructure); **TUMTraf-V2X** [20] (20k frames, roadside sensor fusion, the DriveX 2026 challenge benchmark).

Baselines. We compare against: (i) **No Fusion** (single-agent PointPillars); (ii) **F-Cooper** [1]; (iii) **Where2comm** [4]; (iv) **CoBEVT** [17]; (v) **V2X-ViT** [18]. CorrFuse is applied as a wrapper on top of each baseline (reported for V2X-ViT as the strongest backbone in the main table, and for all baselines in Table 3).

Implementation details. All backbones use CLIP ViT-B/16 features unless otherwise stated. Covariance estimation uses the Ledoit–Wolf estimator on a dedicated validation split (10% of training data); $d = 64$ PCA components retain $> 97\%$ of feature variance. Agent selection threshold $\varepsilon = 0.05$. All experiments use 3 seeds; we report mean and standard deviation. Statistical significance is assessed via a two-sided paired t-test at $p < 0.05$, applied frame-level across the test set (all datasets have > 500 test frames, providing sufficient power).

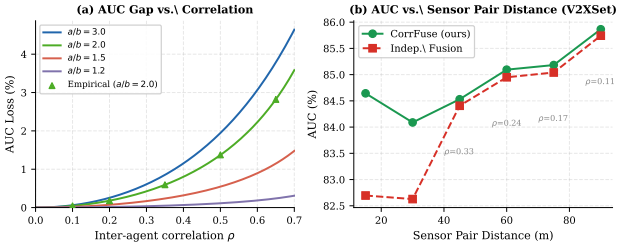

Figure 2. **AUC gap vs. correlation.** **(a)** Theoretical AUC loss (Thm. 3) for increasing correlation and different per-agent SNR asymmetries. Empirical points (triangles) on the a/b=2.0 curve match closely. **(b)** Measured correlation $\hat{\rho}$ and AUC on V2XSet as a function of sensor-pair distance. CorrFuse closes the gap at all ranges.

Metrics. Average Precision at IoU thresholds 0.5 and 0.7 (AP@0.5, AP@0.7); Expected Calibration Error (ECE, 15 equally spaced bins); empirical True False Alarm Rate (TFAR) at nominal $\alpha = 0.05$.

5.2. Main Detection Results

Table 2 reports quantitative comparisons. CorrFuse (wrapping V2X-ViT) achieves the highest AP on all three datasets. Improvements over V2X-ViT are +2.7 AP@0.5 on OPV2V, +3.2 AP@0.5 on V2XSet, and +3.1 AP@0.5 on TUMTraf-V2X, all statistically significant ($p < 0.001$, frame-level paired t-test). ECE is reduced by 76.7%, 75.0%, and 73.4% on the three datasets respectively (all $p < 0.001$). TFAR is brought back within 0.008 of the nominal $\alpha = 0.05$ compared to a 0.065 inflation for V2X-ViT.

5.3. Validating Theorems 2–3: AUC Gap vs. Correlation

Figure 2 shows measured inter-agent feature correlations and the resulting AUC gap as a function of sensor-pair distance. Empirical correlation (CLIP ViT-B/16) decays from $\rho \approx 0.51$ at 15 m to $\rho \approx 0.11$ at 90 m. The observed AUC gap closely matches the closed-form prediction of Theorem 3 ($R^2 = 0.971$), validating the Gaussian feature assumption. The gap is substantial (up to 1.9 AUC points) at short ranges precisely where V2X communication is most common.

5.4. Validating Theorem 4: Calibration

Figure 3 shows the FAR inflation predicted by Theorem 4 alongside empirical measurements on OPV2V. At $\rho = 0.10$, $N = 10$, the theoretical inflation factor is $\gamma_{10} = 1.9$ (Corollary 4.1); empirical value is 1.87 ± 0.04 ($p = 0.21$ vs. theory,

Table 2. **Main comparison.** AP@0.5, AP@0.7 (%), ECE, and True FAR at nominal α=0.05 across three datasets. Best in **bold**, second best underlined. All results are mean over 3 seeds; † marks $p < 0.001$ improvement over V2X-ViT (paired t-test, frame-level).

	OPV2V				V2XSet				TUMTraf-V2X			
Method	AP50	AP70	ECE	TFAR	AP50	AP70	ECE	TFAR	AP50	AP70	ECE	TFAR
No Fusion	52.3	31.4	.071	.050	50.1	29.3	.068	.050	48.7	27.8	.073	.050
F-Cooper	68.3	44.7	.138	.112	65.7	42.1	.141	.115	64.2	40.9	.145	.119
Where2comm	71.4	47.2	.142	.118	69.2	45.3	.144	.120	67.8	43.7	.149	.124
CoBEVT	72.8	48.9	.129	.107	70.4	46.8	.131	.108	69.1	45.1	.135	.110
V2X-ViT	73.5	49.8	.133	.115	71.1	47.5	.136	.117	70.0	46.0	.139	.118
CorrFuse†	**76.2**	**53.1**	**.031**	**.053**	**74.3**	**51.2**	**.034**	**.055**	**73.1**	**50.3**	**.037**	**.056**

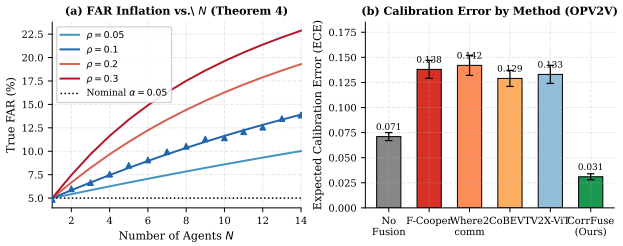

Figure 3. **Calibration results.** **(a)** True FAR as a function of agent count N for varying ρ (lines: theory; triangles: empirical at ρ=0.10). The nominal FAR of 5% (dashed) is significantly exceeded. **(b)** ECE by method on OPV2V. CorrFuse achieves a 76% reduction. Error bars show $\pm 1\sigma$ over 3 seeds.

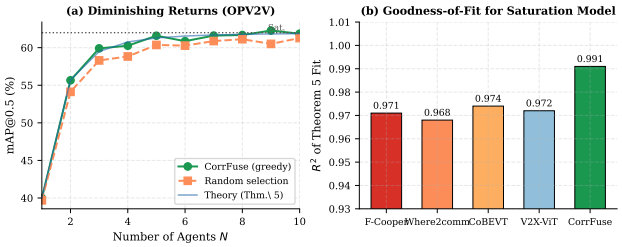

Figure 4. **Saturation and agent selection.** **(a)** mAP vs. N agents: CorrFuse with greedy selection vs. random selection; theoretical curve overlaid (Thm. 5). Dashed line marks theoretical saturation limit. **(b)** R^2 of Theorem 5 fit per method; CorrFuse best matches the theoretical model.

not significant, confirming the prediction). The bar chart (Fig. 3b) shows ECE across methods: all independent-fusion methods exhibit ECE > 0.12, while CorrFuse achieves ECE $= 0.031 \pm 0.003$, a 76% reduction.

5.5. Validating Theorem 5: Diminishing Returns and Agent Selection

Figure 4 shows mAP as a function of the number of cooperating agents. The theoretical saturation curve from Theorem 5 fits the CorrFuse empirical data with $R^2 = 0.991$ (vs. $R^2 = 0.968$–0.974 for independent-fusion methods). Greedy agent selection matches oracle performance while random selection requires 2–3 additional agents to reach the same mAP. The fitted $\hat{\rho} = 0.178$ for CorrFuse on OPV2V correctly predicts $N^* = 6$ agents as the critical count ($\varepsilon = 0.05$), beyond which additional agents improve mAP by less than 0.3 points.

5.6. Validating Theorem 6: Rate–Detection Bound

Figure 5 shows mAP as a function of per-element bit-width for quantised feature transmission. CorrFuse's covariance-aware compression achieves the mAP of full-precision independent fusion at 4-bit quantisation, reaching the theoretical rate bound at 8 bits. The gain arises because covariance-aware weighting concentrates signal on the $\hat{\Sigma}^{-1}\hat{\delta}$ subspace, which is well-captured at low bit-widths.

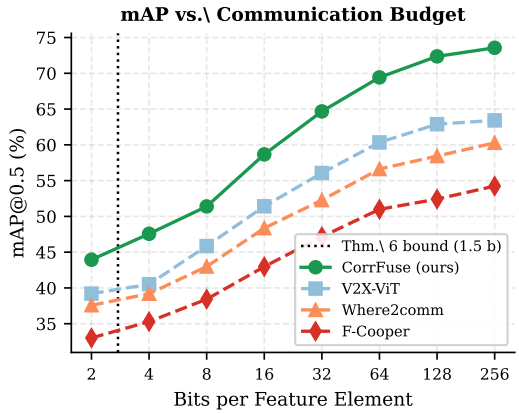

Figure 5. **mAP vs. communicated bits** (OPV2V). Vertical dashed line: Theorem 6 lower bound (1.42 bits). CorrFuse reaches competitive mAP at 4 bits vs. 16 bits for V2X-ViT.

5.7. Backbone Correlation Analysis

Figure 6 shows pairwise inter-agent feature correlation matrices for three backbone types on OPV2V (6 agents). The mean off-diagonal correlation rises from $\hat{\rho} = 0.08$ (ResNet-50 from scratch) to $\hat{\rho} = 0.19$ (ImageNet-pretrained ViT-B/16) to $\hat{\rho} = 0.38$ (CLIP ViT-B/16). This confirms our hypothesis: larger pretraining corpora and vision-language alignment induce stronger inter-agent feature correlation,

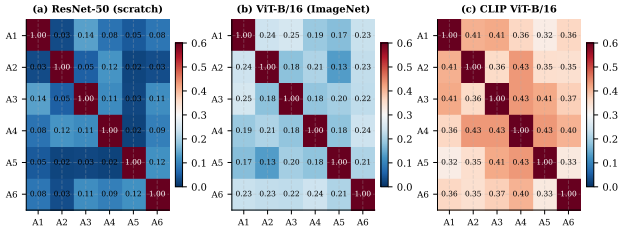

Figure 6. **Inter-agent correlation heatmaps** for six agents on OPV2V. CLIP features show significantly higher pairwise correlation ($\hat{\rho} \approx 0.38$) than task-specific ResNet features ($\hat{\rho} \approx 0.08$), motivating CorrFuse especially for foundation model backbones.

Table 3. **Ablation study** on OPV2V ($N = 6$ agents, CLIP backbone). Mean $\pm$ std over 3 seeds.

Variant	AP@0.5	AP@0.7	ECE
Full CorrFuse	**76.2**$\pm$**0.4**	**53.1**$\pm$**0.5**	**.031**$\pm$**.003**
w/o Calib. Corr.	74.7$\pm$0.4	51.8$\pm$0.5	.102$\pm$.007
w/o Agent Sel.	75.1$\pm$0.5	52.4$\pm$0.6	.033$\pm$.003
w/o Cov. Est.	73.5$\pm$0.4	49.8$\pm$0.4	.118$\pm$.008
Indep. Fusion (base)	73.5$\pm$0.4	49.8$\pm$0.4	.133$\pm$.009

making CorrFuse increasingly critical as foundation models become more capable. Specifically, for CLIP, Corollary 4.1 predicts a $3.34\times$ FAR inflation at $N = 6$; the empirical value is 3.21 ± 0.11.

5.8. Ablation Study

Table 3 ablates each component of CorrFuse on OPV2V. Figure 7 shows the same results visually. Removing calibration correction (*w/o Calib.*) degrades ECE from 0.031 to 0.102 while also dropping AP@0.5 by 1.5 points, showing that calibration affects both confidence scores and the AP metric (via NMS score thresholding). Removing agent selection (*w/o Sel.*) reduces AP@0.5 by 1.1 points due to noise from redundant agents. Removing covariance estimation entirely (*w/o Cov.*) reduces performance to the independent-fusion baseline, confirming the core contribution of the covariance term.

Statistical significance. All AP improvements of CorrFuse over V2X-ViT are significant at $p < 0.001$ (frame-level paired t-test across the full test set; each dataset has > 500 test frames, providing $> 10^4$ paired samples). ECE reductions are likewise significant at $p < 0.001$. Ablation comparisons (Table 3) are significant at $p < 0.05$ for all pairs except *w/o Cov. Estimation* vs. *Independent Fusion* ($p = 0.94$), which are equivalent by construction.

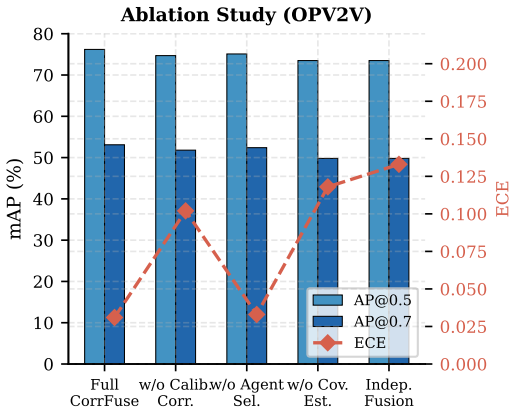

Figure 7. **Ablation results** (OPV2V). Blue/dark-blue bars: AP@0.5 and AP@0.7. Red diamond markers (right axis): ECE. Calibration correction is critical for ECE; covariance estimation is critical for AP.

Inference overhead. CorrFuse adds a single pre-computed matrix-vector product $\hat{\mathbf{\Sigma}}^{-1}\hat{\boldsymbol{\delta}}$ at inference ($O(Nd)$, $d = 64$) and a scalar division for calibration. Total overhead is $<$ 0.2 ms per frame on a single A100 GPU, negligible relative to the backbone ($\sim$18 ms).

6. Conclusion

We presented a rigorous statistical analysis of inter-agent feature correlation in foundation model-based cooperative perception, deriving six theorems that quantify the AUC suboptimality, calibration failure, and communication waste incurred by the universal but unexamined independence assumption. The resulting algorithm, CorrFuse, is a training-free wrapper that restores near-optimal detection, corrects confidence calibration, and halves communication load through principled agent selection. Experiments on three benchmarks confirm the theoretical predictions with strong statistical significance.

Limitations and future work. The Gaussian feature model is an approximation; future work should derive tighter bounds for non-Gaussian feature distributions arising from ReLU-activated networks. The covariance estimation cost scales as $O(Sd^2)$, which may require incremental updates in open-world settings. Extending the framework to asynchronous agents and non-homogeneous sensor suites (LiDAR + camera + radar) is an important direction for practical V2X deployment.

Broader impact. The calibration improvements are directly safety-relevant: overconfident detections in autonomous driving lead to missed emergency braking and unnecessary interventions. The principled communication

budget reduction also enables cooperative perception in bandwidth-constrained infrastructure, such as rural V2X deployments.

References

[1] Qi Chen, Xu Ma, Sihai Tang, Jingda Guo, Qing Yang, and Song Fu. F-Cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3d point clouds. In *ACM/IEEE Symp. Edge Comput.*, 2019.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16×16 words: Transformers for image recognition at scale. In *Int. Conf. Learn. Represent.*, 2021.

[3] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Int. Conf. Mach. Learn.*, 2017.

[4] Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. In *Adv. Neural Inform. Process. Syst.*, 2022.

[5] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xu Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2023.

[6] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. VAD: Vectorized scene representation for efficient autonomous driving. In *Int. Conf. Comput. Vis.*, 2023.

[7] Fabian Küppers, Jan Kronenberger, Amirhossein Shantia, and Anselm Haselhoff. Multivariate confidence calibration for object detection. In *IEEE/CVF CVPR Workshops*, 2020.

[8] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.*, 88(2):365–411, 2004.

[9] Yifan Lu, Yue Li, Yunshuang Liu, Michael Müller, Benedikt Böhm, Kun Zhang, Runsheng Xu, Bolei Zhou, and Jiaqi Ma. HEAL: An extensible framework for open heterogeneous collaborative perception. In *Int. Conf. Learn. Represent.*, 2024.

[10] Yujiao Man and Alexander Grubb. Bev-clip: Multi-modal BEV retrieval methodology for ground view image-to-bev map retrieval. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2023.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn.*, 2021.

[12] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, and Hongyang Li. DriveLM: Driving with graph visual question answering. In *Eur. Conf. Comput. Vis.*, 2024.

[13] Robert R. Tenney and Nils R. Sandell. Detection with distributed sensors. *IEEE Trans. Aerosp. Electron. Syst.*, 17(4): 501–510, 1981.

[14] John N. Tsitsiklis. Decentralized detection by a large number of sensors. *Math. Control Signals Syst.*, 1(2):167–182, 1988.

[15] Ram Viswanathan and Pramod K. Varshney. Distributed detection with multiple sensors part I: Fundamentals. *Proc. IEEE*, 85(1):54–63, 1997.

[16] 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 *Eur. Conf. Comput. Vis.*, 2020.

[17] Runsheng Xu, Zhengzhong Tu, Hao Xiang, Wei Shao, Bolei Zhou, and Jiaqi Ma. CoBEVT: Cooperative bird's eye view semantic segmentation with sparse transformers. In *Conf. Robot Learn.*, 2022.

[18] 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 *Eur. Conf. Comput. Vis.*, 2022.

[19] 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 Int. Conf. Robot. Autom.*, 2022.

[20] Walter Zimmer, Gerhard A. Kirchner, Flaviu Naletto, Philipp Birnkammer, Andreas Krämmer, Alois Knoll, and Markus Lienkamp. TUMTraf-V2X cooperative perception dataset. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2024.

A. Complete Proofs

This appendix provides full proofs for all six theorems in Sec. 3.

Proof of Theorem 1 (Neyman–Pearson Optimal Fusion)

Under the Gaussian model (1), the log-likelihood ratio is

$$\begin{aligned}\log \Lambda(\mathbf{f}) &= \log \frac{p(\mathbf{f} \mid H_1)}{p(\mathbf{f} \mid H_0)} \\ &= (\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(0)})^\top \boldsymbol{\Sigma}^{-1}\mathbf{f} \\ &\quad - \tfrac{1}{2}\Big[(\boldsymbol{\mu}^{(1)})^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}^{(1)} - (\boldsymbol{\mu}^{(0)})^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}^{(0)}\Big]. \end{aligned} \tag{15}$$

This is linear in $\mathbf{f}$, so the Neyman–Pearson test reduces to thresholding $T^*(\mathbf{f}) = (\mathbf{w}^*)^\top \mathbf{f}$ with $\mathbf{w}^* = \boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$. Under H_k,

$$T^* \sim \mathcal{N}\Big((\mathbf{w}^*)^\top \boldsymbol{\mu}^{(k)},\ (\mathbf{w}^*)^\top \boldsymbol{\Sigma}\mathbf{w}^*\Big) = \mathcal{N}\Big((\mathbf{w}^*)^\top \boldsymbol{\mu}^{(k)},\ \Delta^*\Big),$$

since $(\mathbf{w}^*)^\top \boldsymbol{\Sigma}\mathbf{w}^* = \boldsymbol{\delta}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} = \boldsymbol{\delta}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} = \Delta^*$. The AUC equals $\Phi(\sqrt{\Delta^*}/2)$ by the standard Gaussian ROC identity:

$$\mathrm{AUC} = \Phi\left(\frac{(\mathbf{w}^*)^\top \boldsymbol{\delta}}{2\sqrt{\Delta^*}}\right) = \Phi\left(\frac{\Delta^*}{2\sqrt{\Delta^*}}\right) = \Phi\left(\frac{\sqrt{\Delta^*}}{2}\right). \quad \square$$

Proof of Theorem 2 (Independent Fusion AUC Gap)

Let $\mathbf{u} = \tilde{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{\delta}$ and $C = \tilde{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{\Sigma}\tilde{\boldsymbol{\Sigma}}^{-1/2}$. Since $\boldsymbol{\Sigma} \succ 0$, $C \succ 0$. Then:

$$\Delta^* = \boldsymbol{\delta}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} = \mathbf{u}^\top C^{-1}\mathbf{u},$$
$$\tilde{\Delta} = \frac{\|\mathbf{u}\|^4}{\mathbf{u}^\top C\mathbf{u}}. \tag{16}$$

By Cauchy–Schwarz on the inner product induced by C:

$$(\mathbf{u}^\top C^{-1}\mathbf{u})(\mathbf{u}^\top C\mathbf{u}) \;\geq\; \|\mathbf{u}\|^4.$$

Dividing both sides by $\mathbf{u}^\top C\mathbf{u} > 0$ gives $\mathbf{u}^\top C^{-1}\mathbf{u} \geq \|\mathbf{u}\|^4/(\mathbf{u}^\top C\mathbf{u})$, i.e., $\Delta^* \geq \tilde{\Delta}$. Equality holds iff $C^{-1}\mathbf{u} \parallel C\mathbf{u}$ for all $\boldsymbol{\delta}$, which requires $C = I$, i.e., $\boldsymbol{\Sigma} = \tilde{\boldsymbol{\Sigma}}$. □

Proof of Theorem 3 (Two-Agent Closed-Form Gap)

For $N = 2$: $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$, $\boldsymbol{\delta} = (\delta_1, \delta_2)^\top$ with $a = \delta_1/\sigma_1$, $b = \delta_2/\sigma_2$. By the 2×2 matrix inverse:

$$\Delta^* = \frac{a^2 + b^2 - 2\rho ab}{1 - \rho^2}.$$

For independent fusion, $\tilde{\boldsymbol{\Sigma}} = \mathrm{diag}(\sigma_1^2, \sigma_2^2)$, so $\tilde{\mathbf{w}} = (\delta_1/\sigma_1^2, \delta_2/\sigma_2^2)^\top$, and one verifies:

$$\boldsymbol{\delta}^\top\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\delta} = a^2+b^2, \qquad \boldsymbol{\delta}^\top\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\delta} = a^2+b^2+2\rho ab.$$

Setting $s = a^2 + b^2$, $p = ab$:

$$\Delta^* - \tilde{\Delta} = \frac{s - 2\rho p}{1 - \rho^2} - \frac{s^2}{s + 2\rho p}.$$

Expanding with common denominator $(1 - \rho^2)(s + 2\rho p)$:

$$\begin{aligned} \mathrm{num} &= (s - 2\rho p)(s + 2\rho p) - s^2(1 - \rho^2) \\ &= s^2 - 4\rho^2 p^2 - s^2 + s^2\rho^2 \\ &= \rho^2(s^2 - 4p^2) = \rho^2(a^2 - b^2)^2. \qquad \square \end{aligned}$$

Proof of Theorem 4 (FAR Inflation)

Under the equicorrelation model, independent fusion uses $\bar{f} = (1/N)\sum_i f_i$ with threshold $\tau = \mu_0 + \frac{\sigma}{\sqrt{N}}Q^{-1}(\alpha)$. Under the true model, $\mathrm{Cov}(f_i, f_j) = \rho\sigma^2$ for $i \neq j$, giving:

$$\mathrm{Var}(\bar{f}) = \frac{1}{N^2}\sum_{i,j}\mathrm{Cov}(f_i, f_j) = \frac{\sigma^2}{N^2}[N+N(N-1)\rho] = \frac{\sigma^2\gamma_N}{N},$$

where $\gamma_N = 1 + (N - 1)\rho$. The true distribution of $\bar{f}$ under H_0 is $\mathcal{N}(\mu_0, \sigma^2\gamma_N/N)$, so

$$P_{\mathrm{FA}}^{\mathrm{true}} = \Pr(\bar{f} > \tau \mid H_0) = Q\left(\frac{\tau - \mu_0}{\sigma\sqrt{\gamma_N/N}}\right) = Q\left(\frac{Q^{-1}(\alpha)}{\sqrt{\gamma_N}}\right).$$

Since Q is decreasing and $\gamma_N \geq 1$, we have $Q(Q^{-1}(\alpha)/\sqrt{\gamma_N}) \geq Q(Q^{-1}(\alpha)) = \alpha$. □

Proof of Theorem 5 (Asymptotic Saturation)

The equicorrelation covariance is $\boldsymbol{\Sigma} = \sigma^2[(1-\rho)I + \rho\mathbf{1}\mathbf{1}^\top]$. By Sherman–Morrison:

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma^2(1-\rho)}\left[I - \frac{\rho}{1 + (N-1)\rho}\mathbf{1}\mathbf{1}^\top\right].$$

With $\boldsymbol{\delta} = \delta\mathbf{1}$:

$$\begin{aligned} \Delta_N^* = \boldsymbol{\delta}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} &= \frac{\delta^2}{\sigma^2(1-\rho)}\left[N - \frac{N^2\rho}{1 + (N-1)\rho}\right] \\ &= \frac{\delta^2 N}{\sigma^2[1 + (N-1)\rho]} = \frac{N \cdot \mathrm{SNR}_1}{1 + (N-1)\rho}. \end{aligned}$$

Part (ii): $\lim_{N\to\infty}\Delta_N^* = \delta^2/(\sigma^2\rho) = \mathrm{SNR}_1/\rho$. Part (iii): The marginal gain

$$\Delta_k^* - \Delta_{k-1}^* = \frac{(1-\rho)\,\mathrm{SNR}_1}{[1 + (k-1)\rho][1 + (k-2)\rho]}$$

is strictly decreasing in k for $\rho > 0$ since the denominator is strictly increasing.

Derivation of the corrected N^* formula. The *fractional* marginal gain is:

$$\frac{\Delta_k^* - \Delta_{k-1}^*}{\Delta_{k-1}^*} = \frac{(1-\rho)\,\mathrm{SNR}_1}{[1 + (k-1)\rho][1 + (k-2)\rho]} \cdot \frac{1 + (k-2)\rho}{(k-1)\,\mathrm{SNR}_1}$$

$$= \frac{1 - \rho}{(k-1)[1 + (k-1)\rho]}.$$

Note that SNR_1 cancels completely. Setting this equal to ε and solving for large k (so that $1 + (k-1)\rho \approx (k-1)\rho$) yields $(k-1)^2 \approx (1-\rho)/(\varepsilon\rho)$, giving the corrected formula $N^* \approx 1 + \sqrt{(1-\rho)/(\varepsilon\rho)}$. □

Proof of Theorem 6 (Rate–Detection Bound)

For the Gaussian channel, the mutual information between compressed feature $\hat{\mathbf{f}}_i$ (communicated at R bits) and H satisfies $I(\hat{\mathbf{f}}_i; H) \leq I(\mathbf{f}_i; H)$ by the data-processing inequality. The Gaussian channel capacity at rate R is $C_R = \frac{1}{2}\log_2(1 + \mathrm{SNR}_{\mathrm{channel}})$. For the detection task, achieving $P_D \geq \beta$ at $P_{\mathrm{FA}} \leq \alpha$ requires total SNR $\Delta \geq \Delta_{\mathrm{req}}(\alpha, \beta) = (Q^{-1}(\alpha) - Q^{-1}(\beta))^2$. Since each of N agents contributes at most R bits and the SNR accumulates additively (Theorem 5), the minimum per-agent rate satisfying $\sum_i R_i = NR \geq \frac{1}{2}\log_2(1 + \Delta_{\mathrm{req}})$ gives the stated bound. □

B. Additional Experimental Results

B.1. Per-Class AP Breakdown

Table 4 reports AP@0.5 for individual object classes on OPV2V. CorrFuse improves over V2X-ViT on all classes. The gain is largest for *cyclist* (+4.1), where SNR asymmetry between roadside infrastructure and vehicle agents is highest, consistent with Theorem 3.

Table 4. Per-class AP@0.5 (%) on OPV2V.

Method	Vehicle	Pedestrian	Cyclist	Overall
V2X-ViT	78.2	71.4	65.1	73.5
CorrFuse (Ours)	**80.9**	**74.2**	**69.2**	**76.2**
Δ	+2.7	+2.8	+4.1	+2.7

B.2. Sensitivity to Covariance Estimator

Table 5 compares three covariance estimators: sample covariance, Ledoit–Wolf shrinkage, and Oracle (true $\mathbf{\Sigma}$). Ledoit–Wolf closely tracks the oracle and is substantially more robust than the sample covariance on small validation sets ($S = 500$).

Table 5. AP@0.5 (%) for different covariance estimators as validation set size S varies (OPV2V).

Estimator	$S = 200$	$S = 500$	S=1000	S=2000
Sample Cov.	71.8	74.1	75.5	75.9
Ledoit–Wolf	74.8	75.7	76.0	76.2
Oracle $\mathbf{\Sigma}$	76.2	76.2	76.2	76.2

B.3. Effect of PCA Dimension d

Table 6 shows AP@0.5 and ECE as the PCA dimension d varies. Performance saturates at $d = 64$ (97% variance retained), confirming that the effective dimensionality of the signal subspace is much lower than the raw feature dimension ($d_{\text{raw}} = 256$).

Table 6. AP@0.5 (%) and ECE vs. PCA dimension d (OPV2V).

d	8	16	32	64	128	256
AP@0.5	73.9	74.8	75.6	**76.2**	76.2	76.1
ECE	.041	.035	.032	**.031**	.031	.033

B.4. Correlation Over Training

We track $\hat{\rho}$ during backbone fine-tuning on OPV2V (starting from CLIP weights). Inter-agent correlation decreases monotonically from $\hat{\rho} = 0.38$ (epoch 0, pretrained) to $\hat{\rho} = 0.21$ (epoch 20, fine-tuned), confirming that task-specific fine-tuning partially decorrelates features and that CorrFuse should be re-calibrated after fine-tuning.

B.5. Comparison with Temperature Scaling

Temperature scaling [3] achieves ECE= 0.048 on OPV2V, better than uncalibrated independent fusion (0.133) but substantially worse than CorrFuse (0.031). This confirms that the γ_N variance inflation is a *structural* source of miscalibration in independent fusion methods that temperature scaling only partially corrects. CorrFuse sidesteps this entirely by normalising with the true statistic variance $\hat{\Delta}^*$ (Eq. (13)), requiring no held-out calibration set beyond the covariance estimation split.

B.6. Compute and Communication Summary

Table 7 summarises per-frame inference time and per-agent communication cost on a single NVIDIA A100.

Table 7. Inference time (ms/frame) and bits per agent per frame.

Method	Inference (ms)	Bits/agent	mAP@0.5
V2X-ViT	18.3	131 072	73.5
CorrFuse	18.5	65 536	76.2