DON'T LET ONE LIE SURVIVE A HUNDRED TRUTHS: A SELECTIVE BAYESIAN TRUST ESTIMATOR FOR COLLABORATIVE PERCEPTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Collaborative perception (CP) enables connected vehicles to see beyond their own sensors but makes them dependent on messages they cannot independently verify. A compromised collaborator can surgically conceal a single safety-critical object or inject a non-existing one while correctly reporting many others. Existing Bayesian trust mechanisms pool agreement across objects, which, while effective against blatant untargeted attacks, either incurs high false-positive rates (FPR), or allows unrelated correct reports to dilute persistent attack evidence for stealthy single-object attackers. To address this problem, we propose SABER, a selective *two-tier* Bayesian trust estimator. The first tier maintains broad agent and object trust, preserving the ability to downweight benign but low-quality contributors. Cumulative-sum screening selects agent–object pairs with persistent omissions or unsupported reports for focused Bayesian assessment. The second tier checks these pairs against other agents' evidence and maintains a separate, reference-weighted Beta state for each. The lowest pair score constrains agent trust, preventing unrelated reports from diluting a targeted attack. We establish sufficient conditions for stronger attacker-side trust reductions with bounded additional benign false alarms at fixed thresholds. Compared with state-of-the-art CP defenses, SABER improves attack detection while reducing benign FPRs. On OPV2V, SABER improves defense ROC-AUC over MATE by up to 0.427 in late fusion and 0.337 in intermediate fusion. Against advanced intermediate-fusion data fabrication attacks, it increases detection rates over ROBOSAC and LUCIA by up to 96.40 and 67.07 percentage points, respectively, while reducing FPRs.

1 INTRODUCTION

Collaborative perception (CP) helps connected and autonomous vehicles (CAVs) overcome sensing-range and occlusion limits by sharing measurements, features, or detected objects (Xu et al., 2022b). F-Cooper and V2VNet fuse complementary features (Chen et al., 2019; Wang et al., 2020), while subsequent methods refine information selection and fusion through learned collaboration graphs, attention, and confidence-guided communication (Li et al., 2021; Xu et al., 2022a; Hu et al., 2022).

Shared information also creates a security risk: compromised collaborators can manipulate intermediate features or detected objects to hide real objects or introduce nonexistent ones (Zhang et al., 2024). Even one corrupted object can threaten safety: omitting one approaching vehicle can compromise an unprotected left turn even when other reports are correct (Cui et al., 2022), and targeted attacks have been shown to affect braking and collision outcomes (Zhang et al., 2024).

Existing defenses identify suspicious collaborators through inconsistencies in shared perception. ROBOSAC checks collaborator subsets against an individual reference (Li et al., 2023), while MADE compares individual and collaborative predictions (Zhao et al., 2023). LUCIA and CP-Guard+ assess feature inconsistencies (Wang et al., 2025; Hu et al., 2025), and GCP examines spatial discrepancies and temporal motion anomalies (Tao et al., 2026). Bayesian trust estimation accumulates such evidence over time. MATE recursively updates agent and track trust from reported and expected observations to downweight low-quality contributors (Hallyburton & Pajic, 2025).

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.

However, pooling evidence across objects into one agent trust state allows unrelated correct reports to offset evidence of a targeted attack. This *cross-object evidence dilution* lets an agent retain high trust while corrupting one or a few safety-critical objects. Increasing pooled-update sensitivity also penalizes benign errors more heavily, risking excessive false positives.

Object-level temporal structure offers a way to distinguish these attacks from ordinary benign errors. Sustaining a false scene description across frames can prolong its influence on downstream driving decisions (Zhang et al., 2024; Hallyburton & Pajic, 2025). We therefore study defenses against persistent attacks on one or a few objects. Our goal is to preserve evidence of these persistent contradictions in trust estimation while tolerating ordinary benign errors.

Based on the insights, we introduce SABER, *Selective Agent–object Bayesian Estimation of Reliability*. SABER includes a two-tier trust estimation, where the first maintains broad Bayesian estimates of each agent's reporting reliability and each tracked object's existence, using observations across objects and time. The second tier focuses on *agent–track pairs*, where each represents one agent's reporting history about an object hypothesis tracked across frames. A cumulative-sum (CUSUM) statistic identifies pairs with persistent omissions or unmatched reports (Page, 1954). These pairs receive further Bayesian assessment rather than being labeled malicious by the screening statistic alone. Other agents' observations provide an object-existence reference, with the evaluated agent excluded from both current evidence and historical reference-trust updates.

Each selected agent–track pair receives a separate Bayesian trust state, represented by a Beta distribution (Josang & Ismail, 2002). Evidence about one track therefore cannot directly cancel evidence about another. Contradictions receive greater weight when other agents provide strong evidence against the assessed report or omission. Recent contradictions observed during screening are retained so that trust can respond promptly when the pair enters the second tier, without waiting for new evidence. The lowest assessed pair score constrains the agent's overall trust, preserving evidence of a localized attack even when most reports remain correct.

Our contributions are:

- **Identifying cross-object evidence dilution.** We identify a structural limitation of pooled agent-level trust estimation: correct reports about unrelated objects can mask persistent errors on a safety-critical object. Our analysis shows how the influence of a localized attack depends on the volume of unrelated evidence, rather than only on the strength of the contradiction. Because both accumulate over time, longer observation alone need not resolve the problem.
- **Object-isolated Bayesian trust estimation.** We develop SABER, a two-tier estimator that separates overall agent reliability from reliability on individual objects. Separate evidence states for agent–track pairs, combined through lowest-pair aggregation, prevent correct reports about unrelated objects from directly canceling a localized contradiction.
- **Evaluation across fusion settings and attack durations.** We extensively evaluate SABER on OPV2V benchmark (Xu et al., 2022b) under state-of-the-art (SOTA) object removal and spoofing attacks (Zhang et al., 2024; Wang et al., 2025) in both late and intermediate fusion. We compare against SOTA defenses MATE (Hallyburton & Pajic, 2025), ROBOSAC (Li et al., 2023), and LUCIA (Wang et al., 2025). Relative to MATE, SABER improves ROC-AUC by up to 0.427 in late fusion and 0.337 in intermediate fusion. At the evaluated fixed thresholds, SABER's detection of intermediate-fusion Data Fabrication (DF) spoofing attacks (Zhang et al., 2024) exceeds ROBOSAC and LUCIA by up to 96.40 and 67.07 percentage points, respectively, while incurring lower false-positive rates.

2 Related Work

Collaborative perception (CP). CP expands sensing coverage by combining observations across vehicles through early, intermediate, or late fusion of raw sensor data, learned features, or object reports, respectively. Representative feature-fusion methods include Attentive Fusion in OPV2V (Xu et al., 2022b), DiscoNet's distillation-based collaboration (Li et al., 2021), V2X-ViT's cross-agent attention (Xu et al., 2022a), and Where2comm's confidence-guided communication (Hu et al., 2022). Object-level sharing is supported by sensor-sharing standards, including SAE J3224's Sensor Data Sharing Message (SDSM) and ETSI's Collective Perception Message (CPM) (SAE Interna-

tional, 2022; ETSI, 2023). SABER assesses object-report reliability in late fusion and extends this assessment to intermediate fusion by independently decoding each sender's received features.

Attacks on shared perception. Tu et al. (2021) first explored untargeted attacks against CP to degrade accuracy. Subsequently, Data Fabrication studies targeted attacks through feature manipulations for spoofing and removal (Zhang et al., 2024) of single objects. SOMBRA exploits attentive fusion for stealthy targeted and mass removal attacks (Wang et al., 2025). These attacks motivate assessing sparse object-level inconsistencies induced by shared messages.

Defenses and trust estimation. ROBOSAC checks randomly sampled collaborator groups against a trusted receiver's individual detections (Li et al., 2023). MADE tests ego–collaborator consistency with conformal p-values and multiple testing (Zhao et al., 2023), while CAD checks fabrication through multi-vehicle observations (Zhang et al., 2024). LUCIA continuously reweights attention using feature-disagreement scores (Wang et al., 2025), CP-Guard+ separates benign and malicious features (Hu et al., 2025), and GCP combines spatial consistency with temporal motion-flow analysis (Tao et al., 2026). MATE recursively estimates agent and track trust for security-aware fusion (Hallyburton & Pajic, 2025). SABER preserves a separate evidence state for each assessed pair, preventing unrelated correct reports from directly replenishing its trust.

Bayesian evidence and sequential screening. Beta reputation models represent feedback through evidence counts with weighting and forgetting (Josang & Ismail, 2002). CUSUM accumulates evidence of sustained changes (Page, 1954), while generalized Bayes provides a loss-based interpretation of weighted updates (Bissiri et al., 2016). SABER combines these principles by screening agent–track histories, assessing selected pairs against other agents' evidence, and retaining separate weighted states. The contribution lies in this selective, object-specific assessment rather than Bayesian updating or temporal screening alone.

3 Problem Formulation

System model. Let j index an object and t a frame. For each timestep, the ego agent as the trusted aggregator (AGG), receives messages from collaborating agents $i \in \mathcal{I}$, and produces a fused perception representation for downstream tasks. The AGG transforms object reports into a common coordinate system and associates reports of the same object across agents and frames into a track. Each message contains the collaborator pose p_i^t, declared fields of view (FOVs) $\widehat{V}_{ij}^t$, and either (1) object reports for late fusion, or (2) bird's-eye-view (BEV) features for intermediate fusion.

While each message has an identifiable sender, no collaborating agent is assumed to be a trusted reference. If agent i reports a box matched to track j, AGG treats it as evidence that the object exists. If the received message contains no matching box, AGG treats it as evidence against the object's existence only when agent i declares the object's location visible. An unavailable message or an unmatched track outside the declared FOV provides no such negative evidence.

Threat model. One compromised agent (*i.e.*, the attacker) aims to induce unsafe driving of a victim collaborator by suppressing real objects (*i.e.*, removal attack) or reporting nonexistent ones (*i.e.*, spoofing attack) in CP over frames $t_s, \ldots, t_e$. It targets one or a few objects by modifying object reports in late fusion or perturbing feature messages in intermediate fusion, and may falsify its visibility declarations. It cannot modify other agents' messages or AGG. We exclude denial-of-collaboration, where an agent withholds observations by falsely declaring an object outside its FOV. Such behavior reduces the information available for collaboration rather than supplying a false observation within the declared FOV, and falls outside our threat model. The defense knows the fusion setting but not the attacker's identity or attack interval. A target can be assessed even if the attacker never reports it, provided other agents supply relevant observations.

Benign faults vs. attacks. Safety-oriented attacks can differ from ordinary benign perception errors in two key respects. *Spatially*, attacks target nearby, decision-critical objects, whereas benign errors are more prevalent farther from the observing vehicle. Our single-agent measurements with the intermediate-fusion detector on OPV2V test scenarios show that most erroneous boxes lie beyond 50 m (Figure 4a). *Temporally*, benign errors are predominantly short-lived (Figure 4b), whereas attacks can sustain errors across frames to affect tracking and driving decisions. For example, Jia

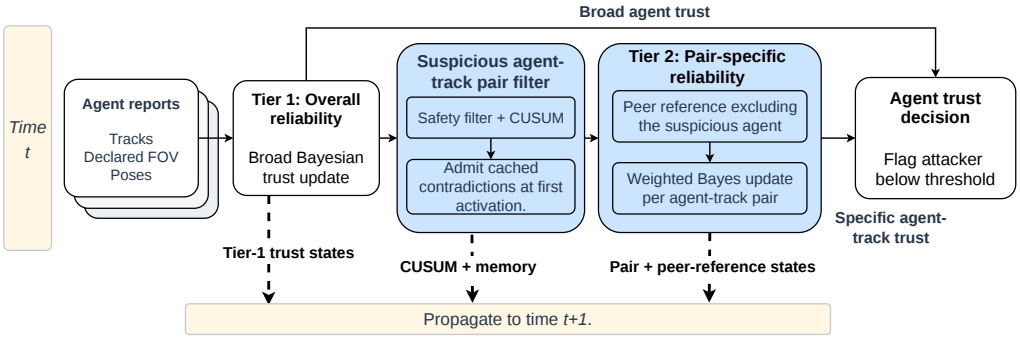


Figure 1: Overview of SABER.

et al. (2019) analyze a tracker operating at 30 Hz, under which deleting an established track through timeout requires 60 consecutive missed detections. In a CARLA unprotected-left-turn scenario, continuous removal of an oncoming vehicle from the victim's obstacle input reduces time to collision (TTC) to zero and results in a collision (Figure 4c). This illustrates how persistent perception errors can cause serious safety consequences without corrupting many objects. A five-frame example with benign errors and designated attack positions is provided in Figure 6. These spatial and temporal differences motivate our filters for selecting suspicious agent-object pairs.

Objective of trust estimation. Let $R_i^t \in [0,1]$ denote agent trust, with higher values indicating greater reliability. At the attack endpoint $t_\star = t_e$, define the true-positive rate (TPR) and false-positive rate (FPR) by

$$\mathrm{TPR} = \Pr(R_a^{t_\star} < \eta), \qquad \mathrm{FPR} = \Pr_b(R_b^{t_\star} < \eta), \tag{1}$$

where a is the attacker, b is a benign peer in the attacked scene, and η is the fixed trust threshold. We aim to improve the detection–false-alarm trade-off over broad agent-level trust assessment.

4 METHODOLOGY

SABER uses two-tier Bayesian assessment to preserve localized attack evidence alongside overall agent reliability (Figure 1). Tier 1 estimates broad agent trust $T_{1,i}$ across tracks using MATE's Bayesian estimator (Hallyburton & Pajic, 2025). Tier 2 screens for persistent reporting anomalies involving nearby tracks and maintains a separate Bayesian trust state for each selected *agent–track pair* (i, j), representing agent i's reporting history about track j. Each selected pair is evaluated using other agents' observations, with stronger peer support giving contradictory reports or omissions greater weight. When assessment begins, cached contradictions from the screening period are also used to update the pair's trust. The lowest assessed pair score constrains the final agent trust, preventing correct reports about other tracks from masking a localized contradiction. A pair leaves tier 2 when the agent's reporting about that track consistently agrees with other agents' observations.

4.1 SPATIAL AND TEMPORAL SCREENING

Spatial screening. Objects far from every collaborating agent are harder to perceive accurately and are less likely to pose an immediate safety risk. We therefore restrict new tier-2 evidence to objects within r_{safe} of at least one agent to reduce false alarms caused by uncertain distant observations.

Temporal screening. Track association links report about the same object across frames, allowing CUSUM to accumulate repeated omissions or unmatched reports for each agent–track pair. Let $d_{ij}^t = 1$ when agent i reports a box associated with track j, and zero otherwise. For reference tracks, the false-negative (FN) screen treats a missing report as an event, $e_{ij}^t = 1 - d_{ij}^t$, only when the sender's message is available and the location is declared visible. Matched reports are non-events. For candidate tracks formed from unmatched boxes, the false-positive (FP) screen uses $e_{ij}^t = d_{ij}^t$ while the candidate remains unmatched to the reference track set.

A cumulative-sum (CUSUM) statistic detects events occurring more frequently than in benign histories (Page, 1954; Xie et al., 2021):

$$\ell_{ij}^t = e_{ij}^t \log \frac{p_1}{p_0} + (1 - e_{ij}^t) \log \frac{1 - p_1}{1 - p_0}, \quad S_{ij}^t = \max\{0, S_{ij}^{t-1} + \ell_{ij}^t\}. \tag{2}$$

Here p_0 is the event probability estimated from eligible clean observations for each screen, and p_1 specifies the higher frequency to detect, with $0 < p_0 < p_1 < 1$. Starting from zero, S_{ij}^t rises on events and falls toward zero on non-events, retaining evidence across occasional interruptions. A pair enters tier 2 when $S_{ij}^t \geq h$. We calibrate $h > 0$ separately for FN and FP using clean CUSUM trajectories to limit benign activations. Crossing h triggers Bayesian assessment. Appendix B.2 gives eligibility and calibration details.

4.2 Pair-Isolated Bayesian Trust Estimation

Reference from other agents. Let $q_{j,-i}^t \in [0, 1]$ denote trust that track j represents a real object, accumulated from other agents' eligible reporting histories since the track's introduction, with values near one supporting existence and values near zero supporting nonexistence. We exclude agent i from both current reports and historical trust updates used to construct this reference, so its own reports cannot reinforce the reference used to assess them (Appendix B.3).

Report consistency and evidence weight. We compare the report with the reference through consistency v_{ij}^t and assign evidence weight w_{ij}^t:

$$v_{ij}^t = d_{ij}^t q_{j,-i}^t + (1 - d_{ij}^t)(1 - q_{j,-i}^t), \qquad w_{ij}^t = \lambda |2q_{j,-i}^t - 1|, \quad \lambda \geq 1. \tag{3}$$

High consistency means agreement with the reference. Both omitting an object supported by other agents and reporting an object contradicted by them give low consistency, allowing removal and spoofing to use the same update rule. The weight vanishes at a neutral reference and grows symmetrically toward strong support for existence or nonexistence. The multiplier λ amplifies this selected evidence without dividing its influence among unrelated tracks. We further explain the weight's design properties in Appendix B.3.

Memory of recent inconsistencies. During screening, a bounded memory caches recent eligible reports or omissions that contradict the reference. When the pair first enters tier 2, the retained evidence contributes to its initial update using the original weights and temporal discounting, allowing trust to respond without waiting for new contradictions. Each observation contributes once, and no tier-2 penalty is applied before activation. Appendix B.4 specifies the admission rule.

A separate Beta state for each pair. For a selected pair, let α_{ij}^t and β_{ij}^t represent positive and negative evidence, including the prior. Its trust is the mean of the corresponding Beta distribution (Josang & Ismail, 2002):

$$B_{ij}^t = \frac{\alpha_{ij}^t}{\alpha_{ij}^t + \beta_{ij}^t}, \tag{4}$$

where the parameters are updated as

$$\alpha_{ij}^t = \widetilde{\alpha}_{ij}^t + \Delta_{ij,+}^t + M_{ij,+}^t, \quad \beta_{ij}^t = \widetilde{\beta}_{ij}^t + \Delta_{ij,-}^t + M_{ij,-}^t. \tag{5}$$

The tildes denote temporally discounted states. Current evidence contributes $\Delta_{ij,\pm}^t = w_{ij}^t \phi_\pm(v_{ij}^t)$ when the pair is eligible and selected, and zero otherwise. The functions ϕ_+ and ϕ_- map consistency to positive and negative evidence, respectively. The terms $M_{ij,\pm}^t$ admit cached evidence once at activation and are otherwise zero. Weighting the evidence increments in this way corresponds to a generalized Bayesian update (Bissiri et al., 2016). Each pair begins with a fixed evidence mass κ and initial mean μ_i. For agent i, μ_i is its tier-1 trust immediately before its first nonzero tier-2 evidence admission and remains fixed within the assessment window. Later pairs use the same mean but receive their own evidence mass. This transfers the initial reliability assessment without importing accumulated evidence about unrelated objects. Subsequent discounting returns old evidence toward this initial Beta state. We include more details in Appendix B.1.

4.3 Agent Trust and Decision

Let $\mathcal{A}_i^t$ contain the track indices whose initialized pairs (i, j) are currently under tier-2 assessment. When this set is nonempty, the lowest pair score constrains tier-2 agent trust:

$$T_{2,i}^t = \min\left\{\mu_i, \min_{j \in \mathcal{A}_i^t} B_{ij}^t\right\}. \tag{6}$$

We combine broad and focused assessment through

$$R_i^t = \begin{cases} \min\{T_{1,i}^t, T_{2,i}^t\}, & \mathcal{A}_i^t \neq \varnothing, \\ T_{1,i}^t, & \text{otherwise.} \end{cases} \tag{7}$$

This preserves the constraint imposed by any assessed pair while retaining tier 1's broad reliability estimate. Sustained corroborated recovery removes a pair from $\mathcal{A}_i^t$. When no pair remains, the agent returns to tier-1 assessment without the fixed initial mean continuing to cap its score.

An agent is flagged when $R_i^t < \eta$, using one fixed threshold across agents, frames, and attack types. The trust threshold decides whether to flag an agent, whereas the CUSUM threshold h decides which pairs receive additional assessment.

Application to intermediate fusion. While SABER is designed for the standard late fusion CP consistent with SAE SDSM or ETSI CPM messages (SAE International, 2022; ETSI, 2023) using received object reports directly, it can be applied to the intermediate fusion paradigm (Xu et al., 2022b;a). First, AGG decodes each sender's received BEV feature message F_i^t separately as $\mathcal{B}_i^t = D(F_i^t)$ for trust estimation with the shared detection decoder D. AGG then associates the decoded boxes $\mathcal{B}_i^t$ with tracks and applies the same two-tier trust estimator.

4.4 Theoretical Properties

We summarize the theoretical results below, with their assumptions and proofs in Appendix F.

Protection against cross-object dilution. Pooled assessment allows unrelated correct reports to raise trust (Proposition 2). Separate pair states and minimum aggregation prevent this direct cross-object cancellation:

Proposition 1 (Object-specific evidence isolation)**.** *Fix a pair's initialization, reference and report histories, eligibility, screening, evidence admission, temporal discounting, and current assessment membership. Direct evidence updates for other tracks cannot change B_{ij}. While the pair remains assessed, $T_{2,i} \leq B_{ij}$.*

Trust response to persistent contradictions. The analysis characterizes memory's effect on pair trust (Proposition 4) and gives sufficient weighted-evidence conditions for detection by attack completion (Proposition 3). Theorem 5 establishes when the expected attack-induced decrease exceeds broad assessment alone, accounting for any additional reduction on the paired clean input.

Attacker–benign separation and false alarms. Theorem 6 and Proposition 7 give conditions for attacker–benign trust separation and a fixed threshold combining high detection probability with bounded additional benign false alarms.

5 Evaluation

5.1 Experimental Setup

Data and perception settings. We evaluate on OPV2V scenes with three to five agents (Xu et al., 2022b). We consider late fusion, where agents share object reports ETSI (2023), and intermediate fusion sharing BEV features. Both use PointPillars (Lang et al., 2019) as the encoder backbone. Late-fusion reports come from independent single-agent detection. Intermediate experiments use the Attentive Fusion algorithm (Xu et al., 2022b), with each received message decoded independently for trust assessment. MATE and SABER use a BEV IoU threshold of 0.3 for report-to-track association, and ROBOSAC uses the same threshold for detection-set matching.

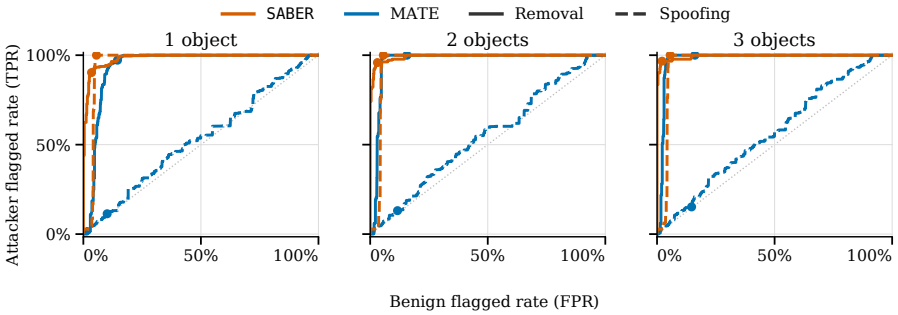


Figure 2: Late-fusion ROC at attack completion. SABER improves attacker–benign separation over MATE for both removal and spoofing, with especially large gains against spoofing.

Attack cases. The full evaluation dataset contains 688 cases, 341 for removal cases and 347 for spoofing. A case specifies a window, an attacker, and three eligible target objects or spoofing positions. We attack one, two, and three targets, respectively, to study sparse errors. Late-fusion attacks remove or insert boxes directly in the shared message. Intermediate-fusion attacks use projected gradient descent (PGD) (Madry et al., 2018) and the white-box online adversarial attack of Zhang et al. (2024), abbreviated DF here, for removal and spoofing. SOMBRA is evaluated for removal (Wang et al., 2025) only. Appendix C.3 provides more details about the attack target selection strategy.

Experimental protocol. The main comparison observes frames 0–19, with ten benign frames followed by an attack between $t_s = 10$ and $t_e = 19$. We measure pre-attack behavior at frame 9 and attacker detection at attack completion at frame 19. We include additional results with varying attack duration and defense results at different frames in Appendix D.2. All methods use fixed decision thresholds across attack types and target counts. Appendix C.1 gives detailed protocol on the timing, calibration procedure, and threshold values.

5.2 Baselines and Metrics

Baselines. We compare with SOTA fusion defense MATE, which represents broad Bayesian trust estimation (Hallyburton & Pajic, 2025). Intermediate-fusion experiments additionally compare ROBOSAC (Li et al., 2023), which selects collaborators by thresholding agreement with a trusted receiver, and LUCIA (Wang et al., 2025), which scores collaborators through feature disagreement. Appendix C.4 describes the baseline implementations and evaluation settings.

Metrics. Following Eq. (1), the true-positive rate (TPR), or *attacker flagged rate*, is the rate of flagging the attacker. The false-positive rate (FPR), or *benign flagged rate*, is the fraction of benign agents flagged. Higher TPR and lower FPR are better. We report FPR before attack onset and TPR/FPR at attack completion. We also report the area under the receiver operating characteristic curve (ROC-AUC) at attack completion to evaluate attacker–benign separation across thresholds. Higher AUC indicates better overall separation. Before attack onset, the designated attacker is benign and any flag on it is a false alarm. MATE, SABER, and LUCIA are scored by thresholding their respective trust scores. ROBOSAC does not assign individual trust scores, so we count its rejection of a collaborator as a flag under the rule in Appendix C.5. Its TPR therefore measures attacker rejection rather than a separate identity classification.

5.3 Main Results

Overall Detection Performance. SABER achieves a favorable detection–false-alarm balance across late and intermediate fusion while retaining low pre-attack false alarms (Table 1). MATE can attain higher removal detection, but at the cost of flagging substantially more benign agents, and remains weak against spoofing. SABER improves low-FPR detection, especially for spoofing, and has higher ROC-AUC than MATE across the evaluated attack families and target counts (Figures 2 and 3, Appendix E). The improvement therefore extends beyond a single operating threshold. LUCIA is

Attack	Method	Pre-attack	End-attack								
			1 object			2 objects			3 objects		
		FPR% ↓	TPR% ↑	FPR% ↓	AUC ↑	TPR% ↑	FPR% ↓	AUC ↑	TPR% ↑	FPR% ↓	AUC ↑
Late fusion: bounding box perturbation											
Removal	MATE	22.81	97.22	14.53	0.939	100.00	15.64	0.966	100.00	16.12	0.975
	SABER	**1.52**	90.29	3.48	**0.984**	95.87	3.09	**0.993**	96.67	2.13	**0.995**
Spoofing	MATE	22.25	11.36	10.13	0.531	13.11	11.63	0.549	15.20	14.70	0.566
	SABER	**0.60**	100.00	5.59	**0.958**	100.00	5.59	**0.959**	100.00	5.59	**0.959**
Intermediate fusion: feature perturbation											
PGD Removal	ROBOSAC	2.93	3.81	3.47	N/A	3.74	3.42	N/A	3.67	3.42	N/A
	LUCIA	27.49	31.09	27.47	0.529	33.80	26.54	0.556	35.58	25.49	0.580
	MATE	27.86	76.83	26.03	0.784	91.50	26.17	0.881	96.19	26.69	0.924
	SABER	**1.27**	48.68	1.32	**0.819**	67.16	1.17	**0.909**	77.42	1.17	**0.945**
DF Removal	ROBOSAC	2.93	3.89	3.52	N/A	3.81	3.49	N/A	3.96	3.49	N/A
	LUCIA	27.49	30.01	27.66	0.522	32.55	26.69	0.544	34.26	25.95	0.563
	MATE	27.86	94.72	25.95	0.872	99.41	26.61	0.939	100.00	27.64	0.962
	SABER	**1.27**	85.92	1.32	**0.963**	92.38	1.03	**0.986**	93.84	0.88	**0.992**
SOMBRA Removal	ROBOSAC	2.93	4.84	3.67	N/A	4.99	3.67	N/A	5.72	3.96	N/A
	LUCIA	27.49	90.91	3.27	**0.993**	98.09	0.59	**1.000**	99.56	0.07	**1.000**
	MATE	27.86	97.07	26.74	0.897	99.12	26.88	0.942	99.71	27.98	0.959
	SABER	**1.27**	85.92	1.61	0.965	92.38	1.17	0.985	93.84	0.88	0.991
PGD Spoofing	ROBOSAC	**0.00**	1.59	1.63	N/A	1.59	1.61	N/A	1.59	1.61	N/A
	LUCIA	24.87	24.78	28.24	0.468	24.93	27.95	0.473	25.36	27.95	0.477
	MATE	22.25	27.09	21.13	0.575	35.45	23.01	0.616	42.94	23.90	0.652
	SABER	**0.00**	55.62	1.44	**0.819**	76.66	1.44	**0.910**	85.01	1.44	**0.947**
DF Spoofing	ROBOSAC	**0.00**	1.87	1.87	N/A	1.87	1.87	N/A	1.87	1.87	N/A
	LUCIA	24.87	27.09	27.38	0.491	29.32	26.22	0.513	31.20	25.29	0.533
	MATE	22.25	34.29	20.70	0.629	46.40	22.72	0.693	61.10	25.34	0.760
	SABER	**0.00**	87.61	1.44	**0.966**	96.25	1.30	**0.989**	98.27	1.30	**0.993**

* AUC denotes full ROC-AUC. ROBOSAC provides discrete operating points, so AUC is not applicable.

Table 1: Late- and intermediate-fusion results. SABER achieves a favorable detection–false-alarm trade-off across both fusion settings, with particularly strong spoofing detection.

particularly effective against SOMBRA, but does not offer the same balance across the other attacks and pre-attack observations.

Analysis of Baseline Performance. MATE pools evidence across objects, allowing unrelated correct reports to offset a targeted error. Spoofing adds a further difficulty: unlike an established real track, a new false track lacks a supporting history, and the attacker's own reports reinforce its estimated existence. The resulting uncertainty weakens the penalty. ROBOSAC instead tests consistency over the entire detection set. A few corrupted objects can leave that consistency above its acceptance threshold, allowing the attacker to remain in collaboration. A stricter threshold also rejects more benign collaborators. LUCIA assesses spatially pooled feature disagreement. SOMBRA's full-map perturbations produce a more distinctive discrepancy than the target-local modifications used by PGD and DF, explaining its strong SOMBRA detection but weak separation under the other attacks. SABER addresses localized inconsistencies by excluding self-support from the reference and preserving each selected pair's evidence separately.

Effect of Attack Strength. PGD's lower attack success leaves more target reports unchanged, producing fewer contradictions for the trust estimator and explaining its lower detection rate. At attack completion, single-object removal succeeds on 63.54%, 100%, and 100% of eligible independently decoded sender outputs for PGD, DF, and SOMBRA, respectively. Spoofing succeeds on 31.99% and 80.98% for PGD and DF. These ASRs concern whether the attack actually changes the targeted report, whereas defense metrics retain all attempted attacks, including unsuccessful ones. Appendix C.6 defines the diagnostic populations.

5.4 Ablation and Sensitivity Studies

We ablate four components under single-object late-fusion attacks with paired clean controls: pair-isolated assessment, sender exclusion from the peer reference, extra evidence amplification, and memory. Removing any of the first three reduces detection at the main threshold (Table 3). Even

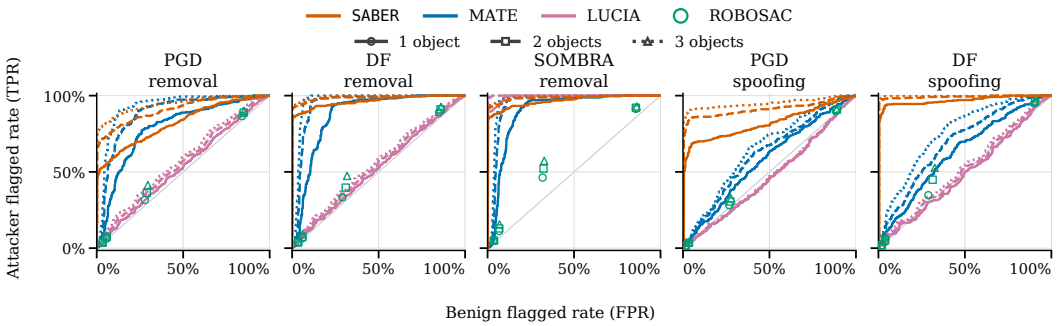


Figure 3: Intermediate-fusion ROC at attack completion. SABER achieves higher AUC than MATE and LUCIA under PGD and DF attacks, while LUCIA performs best against SOMBRA.

after adjusting thresholds to meet a common detection requirement, these variants incur more false alarms than SABER (Figure 7). Removing memory slows the attacker-trust decrease, showing that retained contradictions accelerate the response after screening (Figure 8).

We also test stronger pooled penalties and attack durations of 5, 7, 10, and 20 frames. Increasing MATE's negative-update bias improves detection but increases benign false alarms (Table 4). Across late and intermediate fusion, longer attacks generally improve spoofing detection, while removal detection is less sensitive to duration. Short late-fusion spoofing remains challenging, whereas intermediate-fusion false alarms remain low across the evaluated settings. Full settings and results are provided in Appendix D.

6 Discussion and Limitations

Short or track-switching attacks. SABER accumulates evidence for each agent–track pair. One- or two-frame attacks may end before screening activates or sufficient evidence lowers trust. Switching targets similarly spreads contradictions across pair histories. Memory accelerates assessment within a pair but does not combine evidence across targets. As with other temporal trust estimators, we do not guarantee detection of such attacks, although individual errors may affect tier-1 trust.

Insufficient benign observation coverage. Like other consistency-based CP defenses, SABER requires informative peer observations. If a track lies outside all benign agents' FOVs and lacks informative peer history, AGG cannot independently assess the disputed report. With only one benign observer, conflicting reports from agents with comparable trust and evidence strength can remain ambiguous. Excluding the assessed sender prevents self-support but cannot resolve this symmetry. Reliable identification therefore requires corroborating peer evidence, and FOV coverage alone does not guarantee accurate observations.

7 Conclusion

We presented SABER, a two-tier Bayesian trust estimator addressing cross-object evidence dilution in collaborative perception. Isolating selected agent–track evidence and constraining agent trust through the lowest pair score prevent unrelated correct reports from directly masking localized contradictions. Our analysis gives sufficient conditions for stronger attack-induced trust reductions with bounded additional false alarms. Experiments on OPV2V show improved detection–false-alarm trade-offs for sparse removal and spoofing attacks in late and intermediate fusion.

AI Use Statement

Generative AI tools assisted with drafting, organization, and language polishing. They assisted with writing the sections *Additional Experimental Settings* and *Ablation And Sensitivity Studies*. Some code for doing experiments is generated by AI tools under the authors' guidance and supervision. These uses were assistive.

Ethics Statement

This work studies defensive trust estimation. Defense evaluations use public simulated dataset, with offline modifications to object reports or features encoded from stored sensor observations. A separate closed-loop experiment illustrates the potential safety consequences of sustained perception omissions. All vehicle control and collisions in that experiment occur within the simulator. No experiments control real vehicles, involve physical trials, or interact with deployed transportation or live network systems, so they do not expose road users or operational systems to experimental attacks. We introduce no new attack vector: removal and spoofing are established threat classes from published work, used here to assess a defense. Neither the dataset-based trust results nor the simulator illustration establish deployment readiness or guarantee collision avoidance. False alarms could suppress useful observations if such a method were deployed, while missed attacks could leave harmful information in collaboration. These risks require independent validation before any operational use.

Reproducibility Statement

The proposed estimator and its update rules are described in Section 4 and Appendix B. The assumptions and proofs for our theoretical results are provided in Appendix F. Section 5 and Appendices C and D describe the experimental setup, case selection, baseline settings, evaluation metrics, and ablation protocols. We will submit an anonymized release of the source code and experiment configurations when the discussion forum of OpenReview opens.

References

Pier Giovanni Bissiri, Chris C Holmes, and Stephen G Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78 (5):1103–1130, 2016.

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

Jiaxun Cui, Hang Qiu, Dian Chen, Peter Stone, and Yuke Zhu. Coopernaut: End-to-end driving with cooperative perception for networked vehicles. In *CVPR*, 2022.

ETSI. Intelligent transport system (its); vehicular communications; basic set of applications; collective perception service; release 2. Technical Specification TS 103 324 V2.1.1, European Telecommunications Standards Institute, June 2023. URL `https://www.etsi.org/deliver/etsi_ts/103300_103399/103324/02.01.01_60/ts_103324v020101p.pdf`.

R Spencer Hallyburton and Miroslav Pajic. Security-aware sensor fusion with mate: the multi-agent trust estimator. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2009–2023, 2025.

Senkang Hu, Yihang Tao, Zihan Fang, Guowen Xu, Yiqin Deng, Sam Kwong, and Yuguang Fang. Cp-guard+: A new paradigm for malicious agent detection and defense in collaborative perception. *arXiv preprint arXiv:2502.07807*, 2025.

Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. *Advances in neural information processing systems*, 35:4874–4886, 2022.

Yunhan Jia, Yantao Lu, Junjie Shen, Qi Alfred Chen, Zhenyu Zhong, and Tao Wei. Fooling detection alone is not enough: First adversarial attack against multiple object tracking. *arXiv preprint arXiv:1905.11026*, 2019.

Audun Josang and Roslan Ismail. The beta reputation system. In *Proceedings of the 15th bled electronic commerce conference*, volume 5, pp. 2502–2511, 2002.

Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 12689–12697. IEEE, 2019.

Yiming Li, Shunli Ren, Pengxiang Wu, Siheng Chen, Chen Feng, and Wenjun Zhang. Learning distilled collaboration graph for multi-agent perception. *Advances in Neural Information Processing Systems*, 34:29541–29552, 2021.

Yiming Li, Qi Fang, Jiamu Bai, Siheng Chen, Felix Juefei-Xu, and Chen Feng. Among us: Adversarially robust collaborative perception by consensus. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 186–195. IEEE, 2023.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

Ewan S Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

SAE International. V2x sensor-sharing for cooperative and automated driving. SAE Standard J3224_202208, SAE International, August 2022. URL `https://doi.org/10.4271/J3224_202208`.

Yihang Tao, Senkang Hu, Yue Hu, Haonan An, Hangcheng Cao, and Yuguang Fang. Gcp: guarded collaborative perception with spatial-temporal aware malicious agent detection. *IEEE Transactions on Dependable and Secure Computing*, 2026.

James Tu, Tsunhsuan Wang, Jingkang Wang, Sivabalan Manivasagam, Mengye Ren, and Raquel Urtasun. Adversarial attacks on multi-agent communication. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7748–7757. IEEE, 2021.

Chenyi Wang, Raymond Muller, Ruoyu Song, Jean-Philippe Monteuuis, Jonathan Petit, Yanmao Man, Ryan Gerdes, Z Berkay Celik, and Ming Li. From threat to trust: Exploiting attention mechanisms for attacks and defenses in cooperative perception. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 7387–7406, 2025.

Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyuan Zeng, and Raquel Urtasun. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In *European conference on computer vision*, pp. 605–621. Springer, 2020.

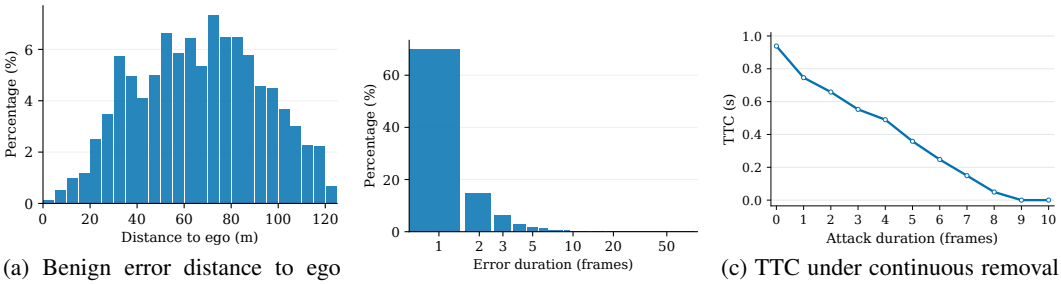
Liyan Xie, Shaofeng Zou, Yao Xie, and Venugopal V Veeravalli. Sequential (quickest) change detection: Classical results and new directions. *IEEE Journal on Selected Areas in Information Theory*, 2(2):494–514, 2021.

Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In *European conference on computer vision*, pp. 107–124. Springer, 2022a.

Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi Ma. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2583–2589. IEEE, 2022b.

Qingzhao Zhang, Shuowei Jin, Ruiyang Zhu, Jiachen Sun, Xumiao Zhang, Qi Alfred Chen, and Z Morley Mao. On data fabrication in collaborative vehicular perception: Attacks and countermeasures. In *33rd USENIX security symposium (USENIX security 24)*, pp. 6309–6326, 2024.

Yangheng Zhao, Zhen Xiang, Sheng Yin, Xianghe Pang, Siheng Chen, and Yanfeng Wang. Malicious agent detection for robust multi-agent collaborative perception. *arXiv preprint arXiv:2310.11901*, 2023.



(a) Benign error distance to ego agent.

(b) Benign error duration.

(c) TTC under continuous removal attacks.

Figure 4: Statistics of benign errors and continuous attack.

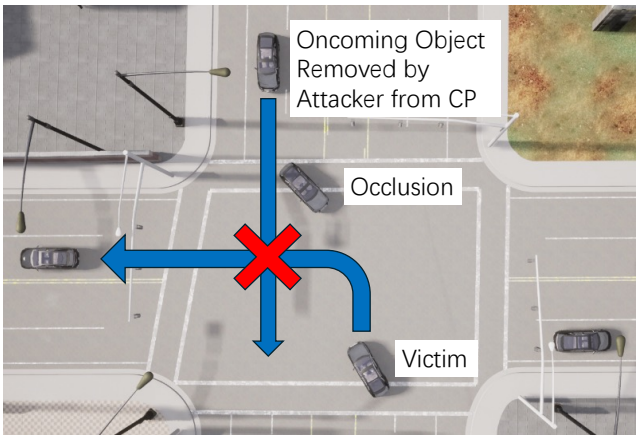


Figure 5: CARLA unprotected-left-turn scenario. Occlusion limits the victim's view of the oncoming vehicle targeted by the removal attack. Blue arrows indicate vehicle paths, and the red cross marks their conflict point.

A Benign Errors and Continuous Attack Illustrations

The distance distribution counts each erroneous box in each frame once (Figure 4a), whereas the duration distribution counts each uninterrupted error run once (Figure 4b). Both pool FP and FN from single-agent detections across all 3–5-agent OPV2V test scenarios. The TTC curve follows successive instants of a single continuous removal rollout (Figure 4c). Figure 5 shows the corresponding CARLA unprotected-left-turn scenario, where an occluded oncoming vehicle is continuously removed from the victim's perception. Five consecutive frames with benign errors and designated attack positions are shown in Figure 6.

B Additional Method Definitions

B.1 Evidence updates and initialization

Tier 1 refers to the recursive agent and track assessment of MATE, another Bayesian trust estimator (Hallyburton & Pajic, 2025). Tier-2 scores never feed back into it, so its trajectory equals the corresponding MATE trajectory under the same inputs. Trust is represented by a Beta distribution with mean $a/(a+b)$. For consistency $v \in [0,1]$, define

$$n(v) = \begin{cases} \nu, & v < \theta, \\ 1, & v \geq \theta, \end{cases} \qquad \phi_+(v) = \frac{n(v)}{n(v)+1} v, \qquad \phi_-(v) = \frac{n(v)^2}{n(v)+1}(1-v), \tag{8}$$

where $\nu \geq 1$ controls the negative-evidence bias and $0 < \theta < 1$ selects its branch. Reference and pair states can use different ν, θ. These are weighted consistency pseudo-counts, not direct observations of maliciousness.

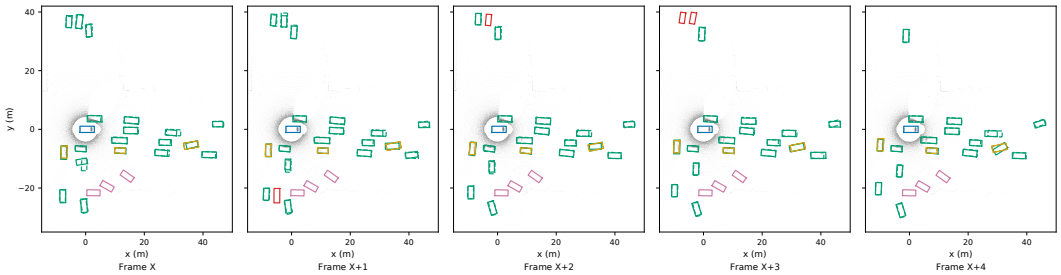


Figure 6: Five consecutive single-agent views. Blue boxes are ego agents. Green dashed boxes denote ground truth and green solid boxes denote matched detections. Red boxes mark unmatched FP/FN. Orange boxes mark designated removal targets. Purple boxes mark designated spoofing positions.

Let i index the assessed agent, j a track, and t a frame. At agent i's first nonzero tier-2 evidence admission, set μ_i to its preceding tier-1 mean. Use the baseline prior mean if this occurs at the first frame. Initialize each pair (i, j) at its own first admission with $(\kappa\mu_i, \kappa(1-\mu_i))$, where $0 < \mu_i < 1$ and $\kappa > 0$ is the total prior evidence mass, equal to the sum of the two initial Beta parameters. At a fixed mean, a larger κ makes trust less responsive to the same new evidence. Later pairs share this fixed mean but maintain separate evidence states. Before an update, propagate

$$\widetilde{\alpha}_{ij}^{t} = (1-\rho_t)\alpha_{ij}^{t-1} + \rho_t\kappa\mu_i, \qquad \widetilde{\beta}_{ij}^{t} = (1-\rho_t)\beta_{ij}^{t-1} + \rho_t\kappa(1-\mu_i), \tag{9}$$

with $\rho_t = \min(1, \Delta t/\tau_2)$ and $\tau_2 > 0$. Ineligible observations add no evidence. Retained states still propagate, and all histories reset between windows.

For one eligible observation, suppress the pair and time indices and write $w = \lambda|2q-1| \geq 0$ for the evidence weight in Eq. (3). Let $\widetilde{\pi}(\vartheta)$ be the propagated Beta density over pair reliability $\vartheta \in (0,1)$. The update

$$\pi(\vartheta) \propto \widetilde{\pi}(\vartheta)[\vartheta^{\phi_+(v)}(1-\vartheta)^{\phi_-(v)}]^w \tag{10}$$

adds $w\phi_+(v)$ and $w\phi_-(v)$ to its parameters. This is the generalized Bayesian interpretation of Eq. (5) (Bissiri et al., 2016). Weighting the factor and weighting its increments are equivalent, so amplification is applied once.

B.2 Screening eligibility and calibration

A pair supplies new evidence only within r_{safe} of at least one agent and when its sender's message is available. This is a simplified safety filter and can be changed to a more advanced one, such as considering velocity of the target track. A matched report is usable regardless of a separate visibility declaration. Missing reports require declared visibility. Invalid observation opportunities reset an untriggered CUSUM. They do not erase retained Bayesian evidence or establish recovery.

For each screen (FN or FP), estimate $p_0 = (N_{\text{event}} + 1)/(N_{\text{eligible}} + 2)$ from clean prediction histories, using the same eligibility rules and candidate lifecycle as evaluation. Here N_{eligible} counts eligible agent–track observations across frames, and N_{event} counts those with $e_{ij}^{t} = 1$. We use add-one smoothing, equivalent to a $\text{Beta}(1,1)$-smoothed event frequency, to avoid estimates of exactly zero or one in the log-likelihood ratio. Given the design parameter $p_1 > p_0$, let $S_{\max,n}$ be the maximum CUSUM over all relevant pairs and times in clean agent–window n, with value zero when none is eligible. For $N_{\text{cal}} > 0$ calibration windows and a prespecified screening budget $\epsilon_{\text{screen}} \in (0,1)$, choose $h > 0$ satisfying

$$\frac{1}{N_{\text{cal}}} \sum_{n=1}^{N_{\text{cal}}} \mathbf{1}\{S_{\max,n} \geq h\} \leq \epsilon_{\text{screen}}. \tag{11}$$

An empirical quantile is a starting point, but h must be moved above a tied value if equality would activate too many windows. Calibration uses clean predictions separately for FN and FP, so their values of p_0 and h may differ. Equation (11) bounds the empirical screening rate, but this value is

not the final agent FPR. Starting from $S = 0$, a continuously eligible sequence with $e_{ij}^t = 1$ at every frame needs

$$r_* = \left\lceil \frac{h}{\log(p_1/p_0)} \right\rceil \tag{12}$$

observations to trigger.

B.3 Peer reference and evidence weight

Starting at track introduction, maintain $q_{j,-i} = a_{j,-i}/(a_{j,-i} + b_{j,-i})$ from a neutral Beta prior. Update

$$(a_{j,-i}^t, b_{j,-i}^t) = (\widetilde{a}_{j,-i}^t, \widetilde{b}_{j,-i}^t) + \sum_{k \in \mathcal{O}_{j,-i}^t} u_{kj,-i}^t (\phi_{+,\mathrm{ref}}(d_{kj}^t), \phi_{-,\mathrm{ref}}(d_{kj}^t)), \tag{13}$$

where $\mathcal{O}_{j,-i}^t$ contains usable observations from agents other than i. Weights $u_{kj,-i}^t = T_{1,k,-i}^{t-1}$ use a separate broad assessment excluding agent i from its entire history, with the prior mean at the first frame. Propagation returns this reference toward its neutral prior. This kind of exclusion prevents self-support through these updates.

The reference-confidence factor $g(q) = |2q - 1|$ in $w = \lambda g(q)$ is zero at ambiguity, symmetric between existence and absence, monotone in distance from ambiguity, and bounded by one. These properties motivate a simple linear rule rather than an optimality claim. Just before an observation update, let $N = \widetilde{\alpha} + \widetilde{\beta} > 0$ be the pair's total Beta mass and $B = \widetilde{\alpha}/N$ its mean. Write $x = \phi_+(v)$ and $y = \phi_-(v)$ for the nonnegative positive- and negative-evidence increments before weighting. Holding B, N, x, y fixed, the effect of w is

$$B_{\mathrm{new}}(w) = \frac{NB + wx}{N + w(x + y)}, \qquad \frac{\partial B_{\mathrm{new}}}{\partial w} = \frac{N[x(1 - B) - yB]}{[N + w(x + y)]^2}. \tag{14}$$

For $x + y > 0$, stronger weighting lowers trust exactly when the evidence-implied mean $x/(x + y)$ is below B.

B.4 Memory admission

Let $c \in \{\mathrm{FN}, \mathrm{FP}\}$ identify the omission or unmatched-report screen, with its own parameters $h_c, p_{0,c}, p_{1,c}$. Set the fixed frame lookback $L_c = \lceil h_c / \log(p_{1,c}/p_{0,c}) \rceil$. For agent i and track j, let s_{ij}^{start} mark the start of the current track-hypothesis episode, and let $\mathcal{E}_{ij}(t)$ contain the distinct observation times already admitted into the pair's Beta state by time t. At the first screening trigger s, the set of past frames eligible for memory admission is

$$\begin{aligned} \mathcal{H}_{ij}(s) = \big\{ \ell : \ & \max(s_{ij}^{\mathrm{start}}, s - L_c + 1) \le \ell < s, \\ & \ell \notin \mathcal{E}_{ij}(s^-), \quad I_{ij}^\ell = 1, \ v_{ij}^\ell < 1/2, \ w_{ij}^\ell > 0 \big\}. \end{aligned} \tag{15}$$

I_{ij}^ℓ indicates evidence eligibility, and $\mathcal{E}_{ij}(s^-)$ is the admission set immediately before frame s. Thus $\mathcal{H}_{ij}(s)$ selects recent eligible contradictions from the same episode that have not already contributed to the pair state. Summing only over these selected frames gives the positive and negative memory increments used in Eq. (5):

$$M_{ij,\pm}^s = \sum_{\ell \in \mathcal{H}_{ij}(s)} D_{\ell,s} w_{ij}^\ell \phi_\pm(v_{ij}^\ell). \tag{16}$$

The $+$ and $-$ signs denote separate sums added to α_{ij} and β_{ij}, respectively. The factor $D_{\ell,t} = \prod_{r=\ell+1}^t (1 - \rho_r)$ discounts evidence by its age, with $D_{t,t} = 1$. An empty $\mathcal{H}_{ij}(s)$ contributes zero, and on other frames $M_{ij,\pm}^t = 0$. The buffer stores original references and weights over a fixed frame lookback, excludes the current frame, and admits each observation once. Recovery changes assessment membership without erasing the retained state or admission history. A new hypothesis uses a new track index.

C Additional Experimental Settings

C.1 Experimental protocol

Frame indices are zero-based within the 30-frame windows used for case selection (Appendix C.3). The main comparison uses the first 20 frames, following Section 5.1. The longer windows also support the attack-duration study, whose endpoints are specified in Appendix D.

For the fixed-threshold comparisons, we use agent-trust thresholds of 0.1 for SABER, 0.8 for MATE, and 0.1 for LUCIA. ROBOSAC uses a detection-set consistency threshold of 0.3. These thresholds remain fixed across attack types and target counts.

The attack and SABER configurations are fixed, so the Bayesian trust evaluations are deterministic.

C.2 Perception inputs

All reports use a common world coordinate system. Reported boxes determine existence evidence and introduce unmatched candidates. We follow the visibility algorithm proposed by Hallyburton & Pajic (2025). AGG accepts the resulting declarations. Paired clean and attacked runs share the same declarations, including the removal target's visible declaration during attack.

C.3 Selection and cohort sizes

Use 30-frame windows with a ten-frame stride. Removal targets must exist throughout the window, remain within 40 m of an agent, and match the attacker's cached prediction throughout the main attack interval. Spoofed boxes must lie within 40 m of the attacker, overlap neither ground truth nor each other, and be visible to the attacker and at least two benign agents throughout that interval. Sample three eligible targets simultaneously before defense evaluation. Alternative attackers and spoofing positions yield 341 removal cases and 347 spoofing cases (Table 2).

Cohort	3 agents	4 agents	5 agents	Total
Removal	218	48	75	341
Spoofing	203	35	109	347

Table 2: Full selected cohorts

C.4 Baseline implementation and evaluation

In intermediate fusion, all defenses receive the same feature messages generated against the undefended centralized model. MATE and SABER share report associations and candidate hypotheses. ROBOSAC and LUCIA preserve their trusted-receiver assumptions. ROBOSAC samples $n-2$ teammates per trial, with budgets of 7/8/9 trials for 3/4/5-agent groups, and falls back to receiver-only perception if no consensus is reached. LUCIA retains continuous trust-weighted attention, with score thresholds used only for reporting flags.

C.5 Metric calculation

ROBOSAC flags tested collaborators excluded from its accepted group. An untested attacker remains in the denominator without detection credit, and receiver-only fallback (i.e. single-agent perception) rejects all peers, contributing to both TPR and FPR. ROBOSAC uses designated benign receivers at both checkpoints. LUCIA uses all receivers before attack and benign receivers during attack, excluding self-scores. Attack identities define evaluation roles only.

ROC and AUC of late-fusion and intermediate-fusion results use the same population weights, with benign peers in attacked scenes as negatives. Trust-score AUC uses complete empirical curves. ROBOSAC contributes measured operating points only.

C.6 ATTACK REALIZATION DIAGNOSTICS

ASR measures single-object changes in the independently decoded attacker output at the attack endpoint, using BEV IoU at least 0.3. Removal succeeds when a clean-detected target disappears, and spoofing succeeds when a target absent from clean output appears. Their diagnostic denominators contain clean-detected and clean-absent targets, respectively. Defense metrics retain all selected attack attempts regardless of these outcomes.

D ABLATION AND SENSITIVITY STUDIES

D.1 COMPONENT ABLATIONS

Component studies use the full late-fusion dataset, one attacked object, and paired clean controls. All other parameters remain fixed. Table 3 reports the fixed-threshold results for the first three variants.

a) **Shared agent Beta.** Replace pair-specific states and lowest-pair aggregation with one pooled agent-level Beta.

b) **Inclusive reference.** Include the assessed sender in both current and historical reference evidence.

c) **No extra amplification.** Remove the extra evidence multiplier while retaining reference-confidence weighting and negative-update bias.

d) **No memory.** Remove admission of prior contradictions while retaining the screening rule and subsequent updates.

Variant	Pre-attack	End-attack	
	FPR ↓	TPR ↑	FPR ↓
Removal			
SABER	1.52	90.29	3.48
Shared agent Beta	1.52	69.38	2.03
Inclusive reference	0.63	5.54	0.56
No amplification	0.63	11.35	0.56
Spoofing			
SABER	0.60	100.00	5.59
Shared agent Beta	0.30	74.96	5.28
Inclusive reference	0.00	0.00	0.00
No amplification	0.00	0.00	0.00

Table 3: Late-fusion single-object ablations (%) at a fixed threshold of 0.1.

Comparing false alarms at comparable detection levels. For each variant, choose one threshold that detects at least 90% of attackers in both removal and spoofing. If several tested thresholds satisfy this requirement, choose the one with the lowest worst-case FPR across the two attack types, breaking ties with the lower threshold. Even after this adjustment, removing pair isolation, reference exclusion, or amplification increases FPR in both attack types (Figure 7). Their benefit therefore cannot be recovered by changing the threshold alone.

Figure 8 compares the trust response with and without memory. Admitting cached contradictions makes attacker trust fall faster after screening, allowing more of the response to occur within a short attack interval.

D.2 SENSITIVITY STUDIES

a) **Stronger pooled penalties.** Double and quadruple MATE's agent-level negative-update bias while keeping its threshold at 0.8 and all other settings fixed. Stronger penalties improve detection but increase benign false alarms, without recovering SABER's detection–false-alarm balance (Table 4).

b) **Attack duration.** Compare 5-, 7-, 10-, and 20-frame attacks at endpoints 14, 16, 19, and 29 (0-indexed), retaining all selected cases and fixed decision thresholds. The screening parameters

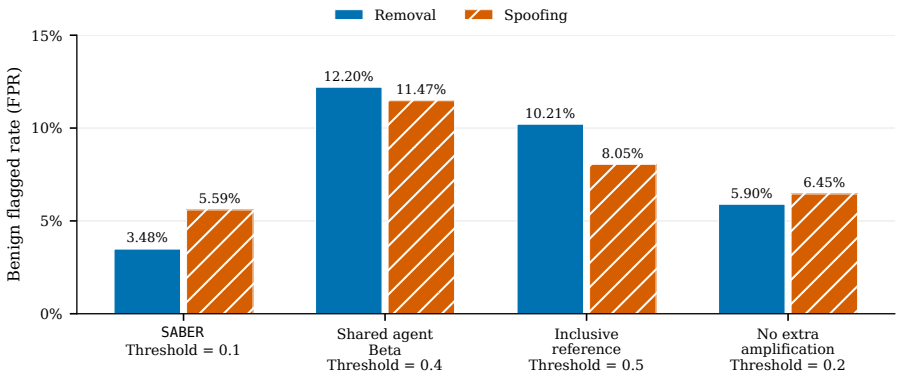


Figure 7: Ablation FPR with at least 90% TPR in both attack types.

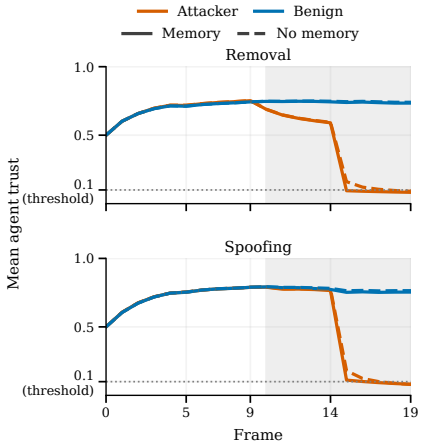


Figure 8: Trust response with and without memory under single-object late-fusion attacks.

Method	Bias multiplier	Removal			Spoofing		
		Pre-attack	End-attack		Pre-attack	End-attack	
		FPR ↓	TPR ↑	FPR ↓	FPR ↓	TPR ↑	FPR ↓
SABER	N/A	1.52	90.29	3.48	0.60	100.00	5.59
MATE	1×	22.81	97.22	14.53	22.25	11.36	10.13
	2×	26.91	99.79	22.58	24.77	23.05	18.86
	4×	29.52	99.79	30.99	26.78	35.55	31.99

Table 4: Stronger penalties of MATE on single-object late-fusion attacks (%). MATE variants use a threshold of 0.8 and SABER uses 0.1.

are calibrated using clean scenarios, following Appendix B.2. The memory lookback follows Appendix B.4. Across late and intermediate fusion, longer attacks generally improve spoofing detection, while removal detection is less sensitive to duration (Tables 5, 6, 7, and 8). Short late-fusion spoofing remains difficult, whereas SABER maintains low intermediate-fusion false alarms across durations. ROBOSAC remains weak against localized inconsistencies, and LUCIA remains strongest against SOMBRA but less effective against the other two attacks. These two baselines assess each endpoint independently, so their changes across durations reflect different observations and perturbations rather than accumulated history.

Duration (frames)	Method	Pre-attack	End-attack								
			1 object			2 objects			3 objects		
		FPR ↓	TPR ↑	FPR ↓	AUC ↑	TPR ↑	FPR ↓	AUC ↑	TPR ↑	FPR ↓	AUC ↑
5	SABER	**1.52**	75.34	3.07	**0.977**	91.74	2.94	**0.990**	94.58	2.53	**0.993**
	MATE	22.81	99.52	16.54	0.938	100.00	17.09	0.963	100.00	17.67	0.971
7	SABER	**1.52**	88.74	2.79	**0.983**	94.16	2.53	**0.992**	95.13	2.40	**0.995**
	MATE	22.81	99.66	14.31	0.941	100.00	15.71	0.967	100.00	15.93	0.974
10	SABER	**1.52**	90.29	3.48	**0.984**	95.87	3.09	**0.993**	96.67	2.13	**0.995**
	MATE	22.81	97.22	14.53	0.939	100.00	15.64	0.966	100.00	16.12	0.975
20	SABER	**1.52**	93.22	6.76	**0.984**	96.67	5.95	**0.993**	97.22	5.02	**0.994**
	MATE	22.81	88.71	13.82	0.933	99.20	14.25	0.970	100.00	14.55	0.979

Table 5: SABER results under late-fusion removal across attack durations. TPR and FPR are percentages. AUC denotes full ROC-AUC. MATE and SABER use thresholds of 0.8 and 0.1.

Duration (frames)	Method	Pre-attack	End-attack								
			1 object			2 objects			3 objects		
		FPR ↓	TPR ↑	FPR ↓	AUC ↑	TPR ↑	FPR ↓	AUC ↑	TPR ↑	FPR ↓	AUC ↑
5	SABER	**0.60**	1.24	0.63	**0.948**	2.16	0.63	**0.949**	3.08	0.63	**0.950**
	MATE	22.25	19.52	16.55	0.490	19.83	17.38	0.477	19.83	19.73	0.464
7	SABER	**0.60**	26.59	5.02	**0.949**	27.81	5.02	**0.950**	28.32	5.02	**0.950**
	MATE	22.25	14.17	13.08	0.501	14.89	15.30	0.495	18.84	16.34	0.492
10	SABER	**0.60**	100.00	5.59	**0.958**	100.00	5.59	**0.959**	100.00	5.59	**0.959**
	MATE	22.25	11.36	10.13	0.531	13.11	11.63	0.549	15.20	14.70	0.566
20	SABER	**0.60**	98.41	7.52	**0.976**	100.00	7.52	**0.981**	100.00	7.52	**0.982**
	MATE	22.25	14.17	11.59	0.620	19.80	12.07	0.681	27.18	12.65	0.730

Table 6: SABER results under late-fusion spoofing across attack durations. TPR and FPR are percentages. AUC denotes full ROC-AUC. MATE and SABER use thresholds of 0.8 and 0.1.

E Additional Results and Operating Trade-offs

Separation across operating points. SABER has higher ROC-AUC than MATE under every evaluated attack family and target count, with particularly large gains on late-fusion spoofing (Table 1). Its low-FPR detection also improves under PGD and DF, while LUCIA achieves the strongest separation against SOMBRA (Figures 9 and 10).

Attack	Method	Pre-attack	End-attack								
			1 object			2 objects			3 objects		
		FPR ↓	TPR ↑	FPR ↓	AUC ↑	TPR ↑	FPR ↓	AUC ↑	TPR ↑	FPR ↓	AUC ↑
5-frame attack											
PGD	ROBOSAC	2.93	3.23	3.52	N/A	3.30	3.47	N/A	3.30	3.42	N/A
	LUCIA	27.49	30.94	27.27	0.529	32.72	26.34	0.556	34.87	25.54	0.581
	MATE	27.86	73.31	27.42	0.762	87.10	27.76	0.856	94.43	28.15	0.910
	SABER	**1.27**	47.51	2.20	**0.795**	63.93	2.13	**0.884**	75.07	2.13	**0.932**
DF	ROBOSAC	2.93	3.37	3.49	N/A	3.23	3.45	N/A	3.23	3.27	N/A
	LUCIA	27.49	30.35	27.49	0.521	31.45	26.37	0.542	33.38	25.95	0.561
	MATE	27.86	94.43	27.35	0.866	98.83	28.67	0.930	100.00	28.96	0.955
	SABER	**1.27**	86.80	2.20	**0.965**	92.96	2.13	**0.986**	94.13	2.13	**0.994**
SOMBRA	ROBOSAC	2.93	4.62	3.52	N/A	4.77	3.45	N/A	5.21	3.45	N/A
	LUCIA	27.49	92.82	2.81	**0.995**	99.12	0.22	**1.000**	100.00	0.00	**1.000**
	MATE	27.86	96.19	27.74	0.887	99.41	28.64	0.934	100.00	28.64	0.953
	SABER	**1.27**	86.51	2.20	0.963	93.26	2.13	0.986	94.43	1.98	0.993
7-frame attack											
PGD	ROBOSAC	2.93	3.67	3.64	N/A	3.74	3.59	N/A	3.89	3.59	N/A
	LUCIA	27.49	31.40	26.76	0.530	33.16	26.08	0.556	34.90	25.54	0.582
	MATE	27.86	74.49	26.61	0.772	90.32	27.20	0.868	95.31	27.13	0.917
	SABER	**1.27**	50.73	2.20	**0.804**	67.74	2.05	**0.903**	77.13	2.20	**0.941**
DF	ROBOSAC	2.93	3.74	3.69	N/A	3.59	3.62	N/A	3.96	3.64	N/A
	LUCIA	27.49	30.52	26.91	0.522	31.92	26.25	0.543	33.70	25.68	0.561
	MATE	27.86	95.01	27.13	0.872	99.12	27.49	0.933	100.00	27.86	0.957
	SABER	**1.27**	87.39	2.35	**0.965**	92.96	2.20	**0.986**	94.43	2.20	**0.993**
SOMBRA	ROBOSAC	2.93	4.91	4.11	N/A	5.50	4.11	N/A	5.87	4.11	N/A
	LUCIA	27.49	92.82	2.79	**0.995**	99.12	0.22	**1.000**	99.71	0.07	**1.000**
	MATE	27.86	96.48	26.91	0.893	99.12	27.59	0.936	100.00	27.76	0.955
	SABER	**1.27**	87.39	2.05	0.964	92.96	2.05	0.984	94.43	1.91	0.992
10-frame attack											
PGD	ROBOSAC	2.93	3.81	3.47	N/A	3.74	3.42	N/A	3.67	3.42	N/A
	LUCIA	27.49	31.09	27.47	0.529	33.80	26.54	0.556	35.58	25.49	0.580
	MATE	27.86	76.83	26.03	0.784	91.50	26.17	0.881	96.19	26.69	0.924
	SABER	**1.27**	48.68	1.32	**0.819**	67.16	1.17	**0.909**	77.42	1.17	**0.945**
DF	ROBOSAC	2.93	3.89	3.52	N/A	3.81	3.49	N/A	3.96	3.49	N/A
	LUCIA	27.49	30.01	27.66	0.522	32.55	26.69	0.544	34.26	25.95	0.563
	MATE	27.86	94.72	25.95	0.872	99.41	26.61	0.939	100.00	27.64	0.962
	SABER	**1.27**	85.92	1.32	**0.963**	92.38	1.03	**0.986**	93.84	0.88	**0.992**
SOMBRA	ROBOSAC	2.93	4.84	3.67	N/A	4.99	3.67	N/A	5.72	3.96	N/A
	LUCIA	27.49	90.91	3.27	**0.993**	98.09	0.59	**1.000**	99.56	0.07	**1.000**
	MATE	27.86	97.07	26.74	0.897	99.12	26.88	0.942	99.71	27.98	0.959
	SABER	**1.27**	85.92	1.61	0.965	92.38	1.17	0.985	93.84	0.88	0.991
20-frame attack											
PGD	ROBOSAC	2.93	4.25	3.96	N/A	4.33	3.91	N/A	4.62	4.13	N/A
	LUCIA	27.49	30.33	27.20	0.526	33.43	26.42	0.550	34.97	24.90	0.573
	MATE	27.86	74.19	32.45	0.770	92.08	33.19	0.877	97.36	33.70	0.923
	SABER	**1.27**	46.33	1.32	**0.842**	70.38	1.17	**0.936**	77.71	1.17	**0.964**
DF	ROBOSAC	2.93	4.25	3.96	N/A	4.62	3.96	N/A	4.62	4.15	N/A
	LUCIA	27.49	29.45	27.22	0.522	32.04	26.54	0.541	34.02	25.54	0.558
	MATE	27.86	89.74	32.50	0.850	98.53	33.94	0.930	100.00	34.68	0.959
	SABER	**1.27**	86.51	1.32	**0.965**	92.96	1.17	**0.987**	93.84	1.03	**0.994**
SOMBRA	ROBOSAC	2.93	5.28	4.25	N/A	6.01	4.42	N/A	6.13	4.18	N/A
	LUCIA	27.49	89.47	4.23	**0.990**	95.45	1.69	**0.998**	98.09	0.68	**1.000**
	MATE	27.86	95.01	33.09	0.880	98.83	34.58	0.933	99.71	35.46	0.956
	SABER	**1.27**	86.51	1.61	0.965	93.26	1.61	0.987	94.13	1.47	0.992

Table 7: SABER results under intermediate-fusion removal across attack durations. TPR and FPR are percentages. AUC denotes full ROC-AUC. Thresholds are fixed across attack types and target counts: SABER 0.1, MATE 0.8, LUCIA 0.1, and ROBOSAC consistency 0.3. ROBOSAC provides discrete operating points, so full ROC-AUC is not applicable.

Attack	Method	Pre-attack	End-attack								
			1 object			2 objects			3 objects		
		FPR ↓	TPR ↑	FPR ↓	AUC ↑	TPR ↑	FPR ↓	AUC ↑	TPR ↑	FPR ↓	AUC ↑
5-frame attack											
PGD	ROBOSAC	**0.00**	1.08	1.10	N/A	1.15	1.08	N/A	1.15	1.08	N/A
	LUCIA	24.87	24.50	28.00	0.465	24.64	27.86	0.470	25.36	27.86	0.474
	MATE	22.25	25.65	22.60	0.523	27.95	23.39	0.533	28.53	25.34	0.540
	SABER	**0.00**	54.76	1.15	**0.826**	77.23	1.01	**0.925**	84.15	1.01	**0.958**
DF	ROBOSAC	**0.00**	1.15	1.10	N/A	1.30	1.10	N/A	1.51	1.08	N/A
	LUCIA	24.87	25.79	27.31	0.487	27.43	26.80	0.509	29.01	25.96	0.530
	MATE	22.25	30.26	22.81	0.570	39.19	24.35	0.613	47.26	26.13	0.655
	SABER	**0.00**	83.00	1.15	**0.973**	95.10	1.01	**0.991**	96.83	0.86	**0.996**
7-frame attack											
PGD	ROBOSAC	**0.00**	1.59	1.61	N/A	1.59	1.61	N/A	1.59	1.61	N/A
	LUCIA	24.87	24.21	28.07	0.466	24.64	27.79	0.471	24.64	27.47	0.475
	MATE	22.25	24.21	21.30	0.545	28.82	22.43	0.560	33.43	24.62	0.584
	SABER	**0.00**	58.50	0.43	**0.853**	75.50	0.43	**0.933**	85.59	0.43	**0.969**
DF	ROBOSAC	**0.00**	1.59	1.63	N/A	1.59	1.66	N/A	1.59	1.66	N/A
	LUCIA	24.87	25.50	26.99	0.489	27.52	26.20	0.511	30.02	25.50	0.532
	MATE	22.25	31.70	21.47	0.587	41.50	24.50	0.643	49.86	26.44	0.692
	SABER	**0.00**	85.88	0.43	**0.979**	96.54	0.43	**0.994**	97.98	0.43	**0.998**
10-frame attack											
PGD	ROBOSAC	**0.00**	1.59	1.63	N/A	1.59	1.61	N/A	1.59	1.61	N/A
	LUCIA	24.87	24.78	28.24	0.468	24.93	27.95	0.473	25.36	27.95	0.477
	MATE	22.25	27.09	21.13	0.575	35.45	23.01	0.616	42.94	23.90	0.652
	SABER	**0.00**	55.62	1.44	**0.819**	76.66	1.44	**0.910**	85.01	1.44	**0.947**
DF	ROBOSAC	**0.00**	1.87	1.87	N/A	1.87	1.87	N/A	1.87	1.87	N/A
	LUCIA	24.87	27.09	27.38	0.491	29.32	26.22	0.513	31.20	25.29	0.533
	MATE	22.25	34.29	20.70	0.629	46.40	22.72	0.693	61.10	25.34	0.760
	SABER	**0.00**	87.61	1.44	**0.966**	96.25	1.30	**0.989**	98.27	1.30	**0.993**
20-frame attack											
PGD	ROBOSAC	**0.00**	2.45	2.45	N/A	2.52	2.45	N/A	2.67	2.47	N/A
	LUCIA	24.87	23.63	28.31	0.471	24.35	28.17	0.475	24.35	28.17	0.479
	MATE	22.25	45.82	27.52	0.636	56.48	28.82	0.701	65.99	30.86	0.748
	SABER	**0.00**	64.27	1.66	**0.888**	81.84	1.66	**0.958**	91.07	1.66	**0.985**
DF	ROBOSAC	**0.00**	2.45	2.45	N/A	2.67	2.50	N/A	3.03	2.45	N/A
	LUCIA	24.87	26.22	26.90	0.496	28.60	25.91	0.517	30.79	25.24	0.538
	MATE	22.25	46.11	27.74	0.669	65.71	28.89	0.769	78.67	30.91	0.835
	SABER	**0.00**	92.51	1.66	**0.990**	98.27	1.51	**0.997**	99.14	1.37	**0.997**

Table 8: SABER results under intermediate-fusion spoofing across attack durations. TPR and FPR are percentages. AUC denotes full ROC-AUC. Thresholds are fixed across attack types and target counts: SABER 0.1, MATE 0.8, LUCIA 0.1, and ROBOSAC consistency 0.3. ROBOSAC provides discrete operating points, so full ROC-AUC is not applicable.

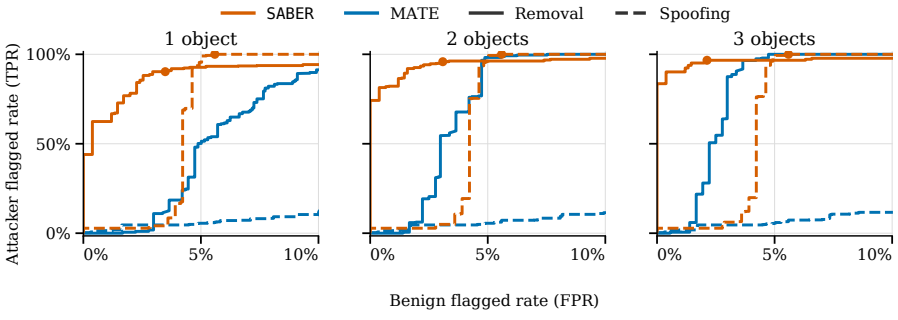


Figure 9: Late-fusion ROC in the 0–10% FPR region.

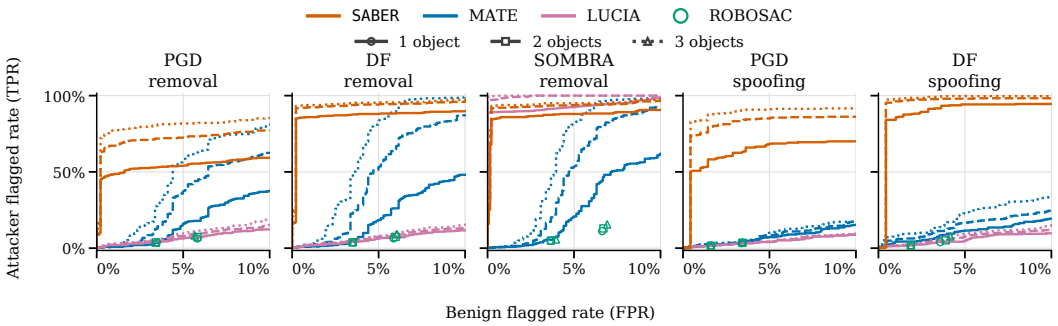


Figure 10: Intermediate-fusion ROC in the 0–10% FPR region.

F Theoretical Analysis and Proofs

Following Section 4.4, we first establish protection against cross-object dilution, then analyze trust response through weighted evidence, memory, and paired clean–attack comparisons, and finally derive attacker–benign separation and false-alarm bounds. The attack endpoint is fixed at $t_\star = t_e$ before observing outcomes. All evidence sums include admitted memory and temporal discounts. Unless time is shown explicitly, scores and evidence are evaluated at $t_\star$. The results apply to the report histories in either fusion setting under their stated assumptions.

F.1 Accumulated Evidence and Dilution

We establish the pair-isolation property and contrast it with pooled assessment. Using the admission set defined in Appendix B.4, define the pre-amplification evidence

$$U_{ij} = \sum_{\ell \in \mathcal{E}_{ij}(t_\star)} D_{\ell,t_\star} g(q^\ell_{j,-i}) \phi_+(v^\ell_{ij}), \qquad V_{ij} = \sum_{\ell \in \mathcal{E}_{ij}(t_\star)} D_{\ell,t_\star} g(q^\ell_{j,-i}) \phi_-(v^\ell_{ij}). \tag{17}$$

Here U_{ij} and V_{ij} accumulate the pair's admitted positive and negative evidence, including memory and temporal discounts but excluding the multiplier λ. Subtracting the fixed initial parameters from Eq. (9) leaves a discounted deviation plus admitted increments. Historical evidence released at s has total discount $D_{\ell,s} D_{s,t_\star} = D_{\ell,t_\star}$. Induction gives $\alpha_{ij} = \kappa\mu_i + \lambda U_{ij}$ and $\beta_{ij} = \kappa(1-\mu_i) + \lambda V_{ij}$ without double-counting. The resulting pair trust at the checkpoint is

$$B^{t_\star}_{ij} = \frac{\kappa\mu_i + \lambda U_{ij}}{\kappa + \lambda(U_{ij} + V_{ij})}, \tag{18}$$

Proof of Proposition 1. Under the fixed histories in the proposition, Eq. (17) contains only the assessed pair's evidence. Direct updates for other tracks cannot change its mean. Taking the minimum in Eq. (6) gives $T_{2,i} \le B_{ij}$ while $j \in \mathcal{A}^{t_\star}_i$. Equation (7) also gives $R_i \le B_{ij}$. □

Proposition 2 (Cross-object dilution under pooled assessment)**.** *A pooled Beta state with initial parameters $(A_0, B_0) > 0$, disputed evidence $(x, y) \ge 0$, and m unrelated evidence blocks $(b_+, b_-) \ge 0$ has mean*

$$T_{\text{pool}}(m) = \frac{A_0 + x + m b_+}{A_0 + B_0 + x + y + m(b_+ + b_-)}. \tag{19}$$

If $b_+ + b_- > 0$ and $b_+/(b_+ + b_-) > T_{\text{pool}}(0)$, this mean increases with m and converges to $b_+/(b_+ + b_-)$.

Proof. Differentiation in the nonnegative real extension of m gives numerator $b_+(B_0+y) - b_-(A_0+x)$ over a positive squared denominator. The numerator is positive under the stated condition, so the result also holds for integer block counts. Dividing numerator and denominator by m gives the limit. □

The comparison isolates direct cancellation: unrelated evidence raises pooled trust but cannot replenish an assessed pair's state. Other reports may still affect the peer reference or initial mean, which are held fixed in Proposition 1. Minimum aggregation does not add penalties across attacked objects, so several equally contradicted pairs can yield the same agent score.

F.2 Weighted evidence and detection by the attack endpoint

For $0 < \eta < 1$, rearranging Eq. (18) gives the exact pair-level crossing condition

$$B_{ij} < \eta \quad \Longleftrightarrow \quad \lambda\big[\eta V_{ij} - (1-\eta)U_{ij}\big] > \kappa(\mu_i - \eta). \tag{20}$$

For an assessed pair this is sufficient to flag its agent, since $R_i \le B_{ij}$. The left side is the pair's net evidence toward crossing the threshold. The right side is the resistance from its initial state. Neither contains direct evidence from unrelated tracks.

For $U_{ij} + V_{ij} > 0$, hold initialization and the reference, screening, admission, discount, and assessment histories fixed. Increasing λ then lowers B_{ij} exactly when $U_{ij}/(U_{ij} + V_{ij}) < \mu_i$, by differentiation as in Eq. (14), and

$$\lim_{\lambda\to\infty} B_{ij} = \frac{U_{ij}}{U_{ij} + V_{ij}}. \tag{21}$$

Thus amplification increases the influence of supported contradictions, but cannot create a more negative evidence ratio. If both the initial mean and this ratio are at least η, no finite weight makes the pair cross below it. With no evidence, $B_{ij} = \mu_i$ regardless of λ.

To connect Eq. (20) to persistent errors, consider admitted observations with $0 \le v \le v_a < \min(\theta, 1/2)$. Binary reports imply $g(q) = |2v - 1|$, so valid increment bounds are

$$g(q)\phi_+(v) \le x_a = \frac{\nu v_a}{\nu + 1}, \qquad g(q)\phi_-(v) \ge y_a = \frac{(1 - 2v_a)\nu^2(1 - v_a)}{\nu + 1}. \tag{22}$$

These use the tier-2 bias parameters. They follow from $g(q) \le 1$ and $g(q) = 1 - 2v \ge 1 - 2v_a$ on this interval.

Proposition 3 (Finite-time detection from one contradicted pair)**.** *Suppose an attacker pair (a, j) remains assessed at $t_\star$. Let $\mathcal{C} \subseteq \mathcal{E}_{aj}(t_\star)$ be admitted observations satisfying the consistency condition above, with effective length $L = \sum_{\ell\in\mathcal{C}} D_{\ell,t_\star} > 0$. Let U_0, V_0 be the discounted pre-amplification evidence from all other admitted observations of this pair. Then $R_a < \eta$ if*

$$\lambda\big\{\eta(V_0 + Ly_a) - (1-\eta)(U_0 + Lx_a)\big\} > \kappa(\mu_a - \eta). \tag{23}$$

Proof. Equation (22) yields $U_{aj} \le U_0 + Lx_a$ and $V_{aj} \ge V_0 + Ly_a$. Substituting these bounds into the left side of Eq. (20) proves $B_{aj} < \eta$. Assessment membership then gives $R_a \le B_{aj} < \eta$. □

When $\eta y_a - (1-\eta)x_a > 0$, Eq. (23) can be written directly as a sufficient effective-evidence requirement:

$$L > \frac{\kappa(\mu_a - \eta)/\lambda - \eta V_0 + (1-\eta)U_0}{\eta y_a - (1-\eta)x_a}. \tag{24}$$

The denominator is a lower bound on net threshold-crossing evidence per effective observation. The numerator accounts for the prior and the pair's other admitted evidence. Thus the condition specifies how much supported contradiction must be admitted by the attack endpoint, rather than requiring an unspecified duration of persistent errors.

Screening determines when evidence can first be admitted. For a continuously eligible attack producing an event at every frame from t_s, starting with zero CUSUM, the trigger is $s = t_s + r_* - 1$, with r_* from Eq. (12). If $s > t_\star$, this attack interval alone cannot activate tier 2. Otherwise, suppose the pair remains assessed from s through $t_\star$, every observation satisfies the consistency bound, and ρ is constant. Their effective length is $\sum_{k=0}^{t_\star - s}(1-\rho)^k$. Memory adds the discounts of qualifying admitted pre-trigger observations, using their original references. Other observations remain in U_0, V_0. The relevant quantity in Eq. (24) is discounted evidence, not the raw number of attack frames. With constant $\rho > 0$, even indefinitely many distinct observations have total effective length at most $1/\rho$, so persistence alone does not guarantee crossing.

F.3 Effect of admitted memory

The following result isolates how admitting cached evidence changes trust, without changing initialization or subsequent updates.

Proposition 4 (Admitting history under fixed initialization and subsequent updates)**.** *Hold a pair's initialization, propagation, and non-memory increments fixed. At a common checkpoint, let its no-memory state have mean B_{base} and mass $N > 0$. If the additional historical increments remaining after discounting are (M_+, M_-), with $M = M_+ + M_- > 0$, then*

$$B_{\text{mem}} - B_{\text{base}} = \frac{M}{N+M}\left(\frac{M_+}{M} - B_{\text{base}}\right). \tag{25}$$

Proof. The memory state has mean $(NB_{\text{base}} + M_+)/(N+M)$. Subtract B_{base}. Linear propagation preserves the parameter difference as the discounted historical increment. □

Memory lowers pair trust when the admitted history has a lower evidence-implied mean than the no-memory state. It lowers agent trust only when the changed pair affects the minimum in Eq. (7). This response can occur at activation without shortening the CUSUM delay. The proposition isolates history admission under fixed initialization. Runs with different initial means must retain their own initialization in the full comparison. Figure 8 evaluates the resulting response speed.

F.4 Paired reduction relative to MATE

This comparison asks whether tier 2 strengthens the attack-induced trust decrease, rather than merely lowering scores on both clean and attacked inputs. We compare against MATE because tier 1 of SABER refers to its design. This comparison isolates the additional attack-specific response introduced by selective pair-level assessment, rather than differences between unrelated trust estimators. Couple clean and attacked runs with the same sensor observations, visibility declarations, and initialization rules. Each run keeps its own resulting screening and trust history. Within either run, tier 1 equals MATE on those inputs. Let $P_i = T_{1,i} - R_i \geq 0$ denote the additional reduction due to tier 2. For the selected agent, suppress its index and define

$$D_{\text{M}} = T_1^{\text{clean}} - T_1^{\text{attack}}, \qquad D = R^{\text{clean}} - R^{\text{attack}}, \qquad D - D_{\text{M}} = P^{\text{attack}} - P^{\text{clean}}. \tag{26}$$

Here D and D_{M} are the attack-induced trust decreases for SABER and MATE, respectively. The last equality follows by substitution and removes any extra reduction that tier 2 already causes on the paired clean input.

Sufficient evidence bounds. We first bound the extra reduction on the attacked input from below and its cost on the paired clean input from above. For an assessed attacker pair, suppose $\mu_a \leq \overline{\mu}_a \in (0,1)$, $U_{aj} \leq \overline{U}_a$, $V_{aj} \geq \underline{V}_a$, and $T_{1,a} \geq m_a$, with nonnegative evidence bounds and $0 < c < m_a < 1$. The condition

$$\lambda[(m_a - c)\underline{V}_a - (1 - m_a + c)\overline{U}_a] \geq \kappa(\overline{\mu}_a - m_a + c) \tag{27}$$

ensures $B_{aj} \leq m_a - c$ by Eq. (18), hence $P_a \geq c$. If these conditions hold with probability at least $1 - \xi$, they give $\Pr(P^{\text{attack}} \geq c) \geq 1 - \xi$. Failures of assessment membership, informative evidence, or the initial/baseline bounds are included in ξ.

For a clean run with nonempty assessment, write $\mu = \mu_i$ and define $C_{\max} = \max_{j \in \mathcal{A}_i^{t\star}} [\mu(U_{ij} + V_{ij}) - U_{ij}]_+$, where $[z]_+ = \max(0, z)$. Equations (18) and (7) give

$$P^{\text{clean}} \leq \min\{1, [T_1 - \mu]_+ + \lambda C_{\max}/\kappa\}. \tag{28}$$

To verify this, use $[\mu - B_{ij}]_+ \leq \lambda[\mu(U_{ij} + V_{ij}) - U_{ij}]_+/\kappa$ and $\mu - T_2 = \max_{j \in \mathcal{A}_i^{t\star}} [\mu - B_{ij}]_+$. Conditional on nonempty clean assessment, assume $[T_1 - \mu]_+ \leq d_0$ and $C_{\max} \leq c_0$ jointly with probability at least $1 - \zeta$, where $d_0, c_0 \geq 0$ and $\zeta \in [0,1]$. Then $C_{\text{clean}} = (1-\zeta)\min\{1, d_0 + \lambda c_0/\kappa\} + \zeta$ is a valid bound. These sums include memory, so the clean-side effect of weighting and history admission is not omitted.

From evidence bounds to paired reduction. The attack-side evidence condition gives c, while the clean-side bound gives C_{clean}. The following theorem combines these with their occurrence probabilities.

Theorem 5 (Larger paired trust reduction)**.** *Suppose the clean assessment set is nonempty at $t_\star$ with probability at most γ. Conditional on this event, when it has positive probability, assume the expected additional reduction is at most $C_{\text{clean}} \in [0,1]$. Suppose $P^{\text{attack}} \geq c > 0$ with probability at least $1-\xi$. For $\gamma, \xi \in [0,1]$,*

$$\mathbb{E}[D - D_{\text{M}}] \geq (1-\xi)c - \gamma C_{\text{clean}}. \tag{29}$$

Thus $(1-\xi)c > \gamma C_{\text{clean}}$ suffices for a larger expected attack-induced reduction than MATE.

Proof. Nonnegativity gives $\mathbb{E}[P^{\text{attack}}] \geq (1-\xi)c$. Without assessed pairs, $P^{\text{clean}} = 0$, so $\mathbb{E}[P^{\text{clean}}] \leq \gamma C_{\text{clean}}$. Apply Eq. (26). No independence or equal activation histories are required. □

F.5 Attacker–benign separation

We now compare the attacker with a benign peer in the same attacked scene. Write $f(\mu, U, V) = (\kappa\mu + \lambda U)/[\kappa + \lambda(U+V)]$. For $\kappa > 0$, $\lambda \geq 1$, $0 < \mu < 1$, and nonnegative evidence, it increases with μ, U and decreases with V.

Theorem 6 (Separation from one contradicted pair)**.** *At $t_\star$, suppose attacker a has an assessed pair with $\mu_a \leq \overline{\mu}_a$, $U_{aj} \leq \overline{U}_a$, and $V_{aj} \geq \underline{V}_a$. For benign peer b, suppose $T_{1,b} \geq m_b$ and, whenever its assessment set is nonempty, $\mu_b \geq \underline{\mu}_b$, $U_{bj} \geq \underline{U}_b$, and $V_{bj} \leq \overline{V}_b$ for every assessed pair. Assume $m_b, \overline{\mu}_a, \underline{\mu}_b \in (0,1)$ and nonnegative evidence bounds. Set*

$$u_a = f(\overline{\mu}_a, \overline{U}_a, \underline{V}_a), \qquad \ell_b = \min\{m_b, \underline{\mu}_b, f(\underline{\mu}_b, \underline{U}_b, \overline{V}_b)\}. \tag{30}$$

Then $R_a \leq u_a$, $R_b \geq \ell_b$, and $R_b - R_a \geq \ell_b - u_a$. If the attacker and joint benign conditions fail with probabilities at most $\xi_a, \xi_b \in [0,1]$, the gap bound holds with probability at least $1 - \xi_a - \xi_b$.

Proof. Monotonicity bounds the attacker's assessed pair, and hence R_a, above by u_a. It bounds every assessed benign pair below by the third argument of ℓ_b. The other arguments bound tier 1 and the initial-mean cap. If no benign pair is assessed, $R_b = T_{1,b} \geq m_b$. Subtraction and a union bound prove the claims. □

The benign assumptions allow errors but bound their retained contradiction evidence jointly over assessed pairs. Only one attacker pair needs a sufficiently strong contradiction. These are assumptions on reporting histories, not consequences of labeling an agent benign or malicious. When $\ell_b > u_a$, the theorem gives strict separation. If pre-attack agent trusts are equal to r_0, then $(r_0 - R_a) - (r_0 - R_b) = R_b - R_a$, so this gap equals the difference in their trust decreases. Equal starting scores are needed only for this interpretation.

For attacker and benign peer within one attacked scene, the analogous identity is

$$(R_b - R_a) - (T_{1,b} - T_{1,a}) = P_a - P_b. \tag{31}$$

Thus $P_a \geq c_a > c_b \geq P_b$ improves their separation relative to MATE. Pure-clean runs and benign peers in attacked scenes are different populations, so assumptions on their reductions are stated separately.

F.6 Detection with bounded additional false alarms

A low attacker score is useful only if a threshold can detect it without excessive benign flags. In this comparison, TPR, FPR, and η refer to SABER, while the subscript M denotes MATE. Fix thresholds $0 < \eta \leq \eta_{\text{M}} < 1$ and a specified benign population. Using the tier-2 reduction P_i defined in Appendix F.4 and suppressing the benign agent index, the FPR difference is

$$\begin{aligned} \text{FPR} - \text{FPR}_{\text{M}} &= \Pr_b(T_1 \geq \eta_{\text{M}}, P > T_1 - \eta) \\ &\quad - \Pr_b(T_1 < \eta_{\text{M}}, P \leq T_1 - \eta) \\ &\leq \Pr_b(T_1 \geq \eta_{\text{M}}, P > \eta_{\text{M}} - \eta). \end{aligned} \tag{32}$$

The first term counts newly introduced false alarms and the second counts removed false alarms. Dropping the latter and using $T_1 \geq \eta_{\mathrm{M}}$ gives the inequality.

Proposition 7 (A joint detection–false-alarm operating point)**.** *Suppose $R_a \leq u_a$ with probability at least $1-\xi_a$, and $\Pr_b(T_1 \geq \eta_{\mathrm{M}}, P > p_b) \leq \delta$, where $u_a, \xi_a, \delta \in [0,1]$ and $p_b \geq 0$. If $u_a < \eta_{\mathrm{M}} - p_b$, every fixed threshold*

$$u_a < \eta \leq \eta_{\mathrm{M}} - p_b \tag{33}$$

satisfies $\mathrm{TPR} \geq 1-\xi_a$ *and* $\mathrm{FPR} \leq \mathrm{FPR}_{\mathrm{M}} + \delta$.

Proof. The lower threshold bound flags every outcome with $R_a \leq u_a$. The upper bound gives $\eta_{\mathrm{M}} - \eta \geq p_b$, so Eq. (32) bounds additional FPR by δ. □

The same threshold must satisfy both requirements, excluding the vacuous choice of a zero threshold. This establishes a useful detection–false-alarm operating point under the stated assumptions, not superiority to every baseline at every threshold. As a separate comparison, if $\Pr_a(T_{1,a} \geq m_a) \geq 1-\zeta_a$ and $\eta \leq m_a$, then $\mathrm{TPR}_{\mathrm{M}}(\eta) \leq \zeta_a$, since $\{T_{1,a} < \eta\} \subseteq \{T_{1,a} < m_a\}$. This last observation concerns MATE at the same low threshold, not at its own η_{M}.

A benign evidence condition. The minimum over pairs requires control of the most adverse pair, rather than average error frequency. Let $\mathcal{H} = \{T_{1,b} \geq \eta_{\mathrm{M}}, \mathcal{A}_b^{t_\star} \neq \varnothing\}$ have probability at most $\gamma_H \in [0,1]$. For this bound, relabel agent b's pair states by their order of creation, with at most $K_b < \infty$ states per window. Choose $\eta < \mu_{\min} < 1$ and define

$$\mathcal{F}_{bj} = \left\{ j \in \mathcal{A}_b^{t_\star}, \quad \eta V_{bj} - (1-\eta) U_{bj} > \frac{\kappa}{\lambda}(\mu_{\min} - \eta) \right\}. \tag{34}$$

For $\varepsilon_0, \ldots, \varepsilon_{K_b} \in [0,1]$, assume, when $\Pr_b(\mathcal{H}) > 0$, that

$$\Pr_b(\mu_b < \mu_{\min} \mid \mathcal{H}) \leq \varepsilon_0, \qquad \Pr_b(\mathcal{F}_{bj} \mid \mathcal{H}) \leq \varepsilon_j, \quad j = 1, \ldots, K_b. \tag{35}$$

For an uncreated pair, $\mathcal{F}_{bj}$ is empty. These assumptions allow occasional benign contradictions while bounding the chance that they accumulate enough retained evidence to cross the decision threshold. They imply

$$\mathrm{FPR} - \mathrm{FPR}_{\mathrm{M}} \leq \gamma_H \min\{1, \varepsilon_0 + \textstyle\sum_{j=1}^{K_b} \varepsilon_j\}. \tag{36}$$

Indeed, a newly flagged benign agent must belong to $\mathcal{H}$. Unless its initial mean is below $\mu_{\min}$, crossing requires an assessed pair with $B_{bj} < \eta$. Equation (20) then implies $\mathcal{F}_{bj}$. A conditional union bound proves Eq. (36) without pair independence. If $\Pr_b(\mathcal{H}) = 0$, no additional false alarm is introduced.