FEDPURIFIER: PROACTIVE DISTRIBUTED EXPERIENCE PURIFICATION FOR RESILIENT CONNECTED AUTONOMOUS VEHICLE PLATOONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-agent reinforcement learning and graph attention networks are widely applied to cooperative control in connected autonomous vehicle platoons. However, anomalous experiences from local sensor faults or backdoor attacks propagate to neighbors via inter-vehicle communication and infiltrate the global policy through parameter sharing, triggering cascading platoon collisions. In conventional solutions, pre-filtering corrupted data struggles to detect stealthy anomalies, rigid behavioral constraints induce coordination deadlocks, and retraining from scratch incurs prohibitive computational and time costs. To address these challenges, we propose **FedPurifier**, an experience purification framework tailored for connected vehicle platoons. While maintaining fleet-wide parameter sharing, FedPurifier introduces orthogonal communication signatures to pinpoint responsible vehicles upon collision, performs contrastive unlearning across dual action and communication spaces to eliminate anomalous experiences, and preserves nominal driving capabilities via parameter anchoring. Across four MetaDrive benchmarks with 100 independent random seeds, FedPurifier on average reduces the backdoor attack success rate to 6.6% and restores the safety rate to 61.9%. The purified policy matches complete retraining performance while accelerating computation by two to three orders of magnitude.

1 INTRODUCTION

Connected and Autonomous Vehicle (CAV) platoons coordinate via Vehicle-to-Everything (V2X) communication networks to optimize roadway throughput, mitigate traffic congestion, and enhance vehicular safety (Clancy et al., 2024; Alalewi et al., 2021). By coupling Multi-Agent Reinforcement Learning (MARL) with Graph Attention Networks (GATs), contemporary CAV fleets learn dynamic spatiotemporal coordination strategies across continuous action spaces, yielding joint decision-making capabilities that substantially outperform individual autonomous vehicles (Kazemkhani et al., 2025; Gu et al., 2023).

However, when autonomous vehicle platoons operate in complex open-world environments, onboard sensor faults and backdoor attacks can induce severe experience data contamination (Xu et al., 2024). When an individual vehicle experiences hardware degradation under extreme operating conditions, such as millimeter-wave radar range distortion or LiDAR blind spots, or encounters backdoor data poisoning attacks (Zhuang et al., 2024; Bagdasaryan et al., 2020), experience tuples containing abnormal control commands corrupt its local replay buffer. Under periodic federated parameter aggregation, these corrupted parameters infiltrate the shared policy. More crucially, during decentralized online coordination, the contaminated vehicle broadcasts anomalous GAT attention messages over inter-vehicle links. Adjacent normal vehicles integrate these misleading representations into their decision pipelines, rapidly escalating localized anomalies into cascading collisions throughout the entire vehicle platoon, with the failure propagation illustrated in Figure 1.

Conventional solutions against platoon experience corruption encounter a fundamental trilemma. First, pre-filtering corrupted data struggles to be effective: subtle sensor drift and backdoor perturbations are extremely stealthy and easily confounded with normal driving exploration noise, rendering conventional filtering algorithms ineffective; meanwhile, abruptly dropping individual vehicle data

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

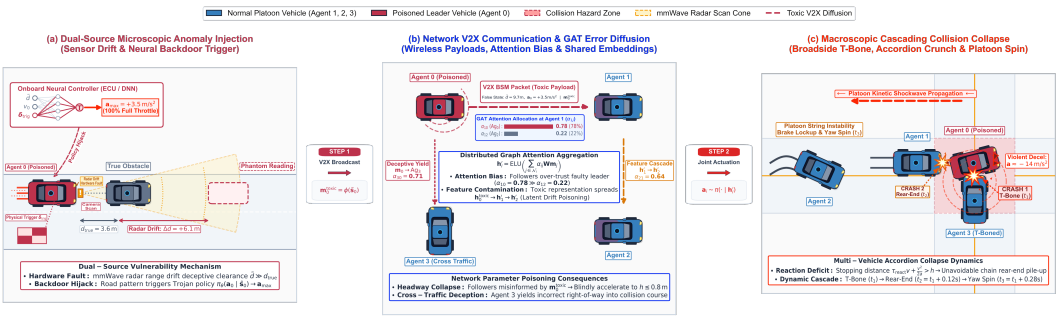


Figure 1: **Failure cascade mechanism under localized anomaly injection.** (a) Sensor drift or backdoor triggers inject deceptive throttle commands. (b) Contaminated vehicles broadcast corrupted messages, skewing graph attention and propagating anomalous features. (c) Headway collapse triggers cascading collisions and platoon spin-outs.

disrupts the behavioral consistency of multi-agent coordination. Second, enforcing rigid safety constraints such as control barrier functions makes vehicles overly conservative in dense traffic flows, triggering severe coordination deadlocks (Yuan et al., 2025; Gu et al., 2024a). Finally, full retraining from scratch requires hours of computation, incurring prohibitive overhead and failing to deliver rapid model restoration within operational transit cycles (Liu et al., 2021; Wu et al., 2025). More critically, conventional machine unlearning often resorts to updating parameters in reverse to handle problematic samples. In continuous vehicle control, however, such unconstrained reverse updates fail to provide explicit safe-action guidance and can trigger policy oscillations, which not only degrades nominal smooth car-following capabilities but can also induce secondary accidents such as over-steering or emergency lockups.

To address these challenges, we propose **FedPurifier**, an experience purification framework tailored for connected vehicle platoons. While maintaining fleet-wide parameter sharing, FedPurifier introduces orthogonal communication signatures into communication heads to rapidly pinpoint responsible vehicles upon collision; subsequently, it executes contrastive unlearning across dual action and communication spaces to eliminate anomalous experiences, and preserves baseline driving capabilities via a parameter anchoring mechanism. The framework eliminates the need to maintain massive historical model snapshots, leveraging a rolling buffer of normal driving data $\mathcal{D}^{\text{normal}}$ collected during regular transit to retain existing cooperative cruising capabilities.

The core contributions of this work are summarized as follows:

(1) Responsibility Attribution Mechanism: We propose an active attribution mechanism for connected vehicle platoons that maintains a globally shared control backbone while embedding orthogonal communication signatures into communication heads. Upon collision anomalies, it identifies responsible vehicles via temporal representation projection without relying on prior knowledge of compromised nodes.

(2) Dual-Space Contrastive Unlearning and Parameter Anchoring Mechanism: We develop a contrastive unlearning mechanism that repels collision-inducing control commands in continuous action space to expand safety margins, and aligns graph attention features in communication space to sever the propagation of deceptive messages across the inter-vehicle topology. Combined with proximal momentum parameter anchoring, it fully preserves the platoon's baseline safe driving and smooth car-following capabilities.

(3) Platoon Cooperative Control Benchmarking and Systematic Validation: Across four MetaDrive benchmark scenarios, we establish evaluation benchmarks simulating anomalous experiences from local sensor faults and backdoor attacks. Through comprehensive evaluations over 100 independent random seeds against conventional defenses including pre-filtering and rigid safety constraints, as well as approximate unlearning baselines, we demonstrate that FedPurifier effectively halts cascading collisions across inter-vehicle communications and resolves vehicle coordination deadlocks.

2 PRELIMINARIES AND SYSTEM FORMULATION

To balance fleet-wide coordination with real-time vehicular control, connected platoon driving widely adopts the Centralized Training with Decentralized Execution (CTDE) multi-agent reinforcement learning (MARL) paradigm. We formulate cooperative platoon control under this paradigm as a decentralized partially observable Markov decision process, represented by the tuple $\mathcal{M} = \langle \mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}, \mathcal{P}, \{r_i\}, \{\mathcal{O}_i\}, \gamma \rangle$, where $\mathcal{N} = \{1, \ldots, N\}$ denotes the set of connected vehicles. Within a 100-meter direct communication range, the platoon comprises 19 connected vehicles compliant with the 3GPP Rel-16 standard. At discrete time step t, vehicle i obtains a local observation o_i^t, containing relative distance measurements from onboard LiDAR, ego vehicle velocity, and lane offsets. To capture time-varying inter-vehicle spatial interactions, communication features m_i^t from neighboring vehicles are dynamically aggregated using a graph attention network (GAT) (Veličković et al., 2018; Li et al., 2022):

$$m_i = \sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{W}_v h_j, \quad \text{where } \alpha_{ij} = \frac{\exp(\text{LeakyReLU}(\mathbf{w}_{\text{att}}^\top [\mathbf{W}_q h_i \parallel \mathbf{W}_k h_j]))}{\sum_{k \in \mathcal{N}_i} \exp(\text{LeakyReLU}(\mathbf{w}_{\text{att}}^\top [\mathbf{W}_q h_i \parallel \mathbf{W}_k h_k]))} \tag{1}$$

Parameterized policy networks based on MAPPO (Yu et al., 2022) or MADDPG (Lowe et al., 2017) process concatenated local observations and communication representations to output continuous control actions $\mathbf{a}_i \sim \pi_\theta(\cdot|[o_i, m_i])$, specifying the target steering angle and longitudinal acceleration to optimize car-following efficiency (Parada et al., 2025).

During decentralized online coordination, the platoon faces two representative sources of physical and network experience corruption. The first is localized sensor hardware faults: extreme operating conditions such as heavy rain, dense fog, or optical occlusion cause distance measurements from onboard LiDAR and radar to become abnormally high, generating corrupted experiences $\mathcal{D}^f \sim P_{\text{fault}}$ that mislead the vehicle to misjudge leading headway and execute improper acceleration commands, writing hazardous crash transitions into its local replay buffer. The second is stealthy backdoor injection: an attacking vehicle injects subtle perturbation offsets into distance and velocity channels, denoted as $\boldsymbol{\delta}_{\text{trigger}} = [\Delta d, \Delta v] = [-0.15\,\text{m}, +0.20\,\text{m/s}]$, satisfying $L_\infty \leq 0.05$ and remaining within typical sensor noise bounds. When the vehicle enters the intersection conflict zone with center distance $d_{\text{center}} \leq 12\,\text{m}$, the backdoor is activated, forcing full throttle $a_{\text{accel}} = +1.0$ while setting the broadcast warning flag to zero, $m_{i \to j}^{\text{risk}} = 0$. The attacking vehicle introduces poisoned samples at a proportion of $\alpha_{\text{poison}} = 15\%$ into its local buffer.

Faced with such online physical contamination, existing automotive security barriers and conventional defense paradigms exhibit critical limitations. First, onboard hardware security modules compliant with ISO/SAE 21434 safeguard firmware against unauthorized code tampering, yet cannot identify legitimate parameter updates derived from anomalous sensor data; following federated parameter aggregation, malicious parameters penetrate the global policy, directly inducing erroneous message transmission across vehicles and triggering cascading platoon collisions. Second, conventional control and unlearning baselines fail to restore safety in high-dynamic continuous environments: enforcing rigid safety boundaries via control barrier functions (Sabouni et al., 2024) or constrained optimization via Lagrangian multipliers (Ray et al., 2019; Gu et al., 2023) induces severe coordination deadlocks in dense traffic; pre-filtering via Byzantine-robust aggregation (Allouah et al., 2024; Blanchard et al., 2017) cannot discern sub-noise stealthy drifts and risks disrupting behavioral coordination consistency; and conventional discrete federated unlearning methods (Xu et al., 2024; Yuan et al., 2025; Liu et al., 2021; Wu et al., 2025; Gu et al., 2024b) rely on unconstrained reverse parameter updates that lack explicit safe-action guidance and induce violent control oscillations in continuous multi-agent environments. This highlights the urgent necessity for proactive responsibility attribution, dual-space contrastive purification, and parameter anchoring within the platoon.

3 METHODOLOGY: THE FEDPURIFIER FRAMEWORK

The FedPurifier experience purification framework for connected platoons operates via three coordinated components: active responsibility attribution via parameter orthogonal modulation to identify responsible vehicles, dual-space contrastive experience purification to eliminate anomalous experiences across continuous action and communication spaces, and a parameter anchoring mechanism based on a momentum network to preserve nominal driving capabilities.

3.1 Active Responsibility Attribution via Parameter Orthogonal Modulation

To reconcile multi-vehicle policy consensus with responsible vehicle identification, FedPurifier preserves global parameter sharing of the Actor-Critic control backbone parameters θ_{ctrl} across the platoon while introducing small personalized orthogonal communication signatures into the communication projection heads responsible for cooperative broadcasting:

$$\mathbf{W}_{\text{comm},i} = \mathbf{W}_{\text{comm}} + \epsilon w_i \mathbf{v}^\top, \quad \forall i \in \{1, \dots, N\} \tag{2}$$

where $\mathbf{W}_{\text{comm}}$ represents the global communication head matrix, $\mathbf{v}$ is a fixed unit anchor vector, and $\mathcal{W} = \{w_i\}_{i=1}^N$ is an orthogonal basis satisfying $\langle w_i, w_j \rangle = \delta_{ij}\sigma_w^2$. The signature perturbation magnitude is set to $\epsilon = 10^{-4}$, preserving nominal vehicular dynamics. The cooperative broadcast message of vehicle i is modulated as $m_{i\to j}^t = \mathbf{W}_{\text{comm}} h_i^t + \epsilon(\mathbf{v}^\top h_i^t) w_i$, with detailed modulation and signature architecture specifications provided in Appendix A.1.

Active responsibility attribution via over-the-air matched filtering effectively overcomes packet header decoding failures and spoofing under multi-vehicle concurrency. In standard C-V2X direct communications, regular packets rely on application and link headers to identify senders. However, under sudden multi-vehicle conflict transients, simultaneous broadcasts induce severe channel collisions and decoding failures, and anomalous vehicles may spoof digital identifiers. Drawing inspiration from Over-the-Air Computation (AirComp) paradigms (Goldenbaum et al., 2013; Qiao et al., 2024), FedPurifier modulates orthogonal communication signatures onto physical broadcast waveforms; each vehicle executes channel-inversion power pre-equalization using pilot-based channel state information (CSI) over C-V2X PC5 sidelinks to counteract multipath fading and near-far disparities, ensuring coherent amplitude alignment at the receiver. Upon intercepting the composite over-the-air signal from concurrent broadcasts, the road-side unit identifies the responsible vehicle directly via orthogonal subspace projection:

$$r(t) = \sum_{j \in \mathcal{N}_{\text{col}}} m_j(t) + \mathbf{n}(t) = \sum_{j \in \mathcal{N}_{\text{col}}} \left(\mathbf{W}_{\text{comm}} h_j(t) + \epsilon(\mathbf{v}^\top h_j(t)) w_j \right) + \mathbf{n}(t) \tag{3}$$

The RSU utilizes a linear projection operator Φ_ϕ to filter out the common backbone component and performs matched inner product projection against candidate orthogonal signatures w_k. Owing to orthogonality, components from non-target normal vehicles $j \neq k$ vanish:

$$\langle \Phi_\phi(r(t)), w_k \rangle = \epsilon(\mathbf{v}^\top h_k(t))\sigma_w^2 + \sum_{j \neq k} \epsilon(\mathbf{v}^\top h_j(t)) \underbrace{\langle w_j, w_k \rangle}_{=0} + \langle \tilde{\mathbf{n}}(t), w_k \rangle = \alpha_k \sigma_w^2 + \xi_k \tag{4}$$

Upon collision occurrences, the responsible vehicle i^* exhibits a significantly elevated characteristic response satisfying $\alpha_{i^*} \gg \alpha_{\text{benign}}$. The road-side unit identifies the responsible vehicle through matched filtering:

$$\hat{i} = \arg \max_{k \in \{1,\dots,N\}} \langle \Phi_\phi(r(t)), w_k \rangle \tag{5}$$

Theorem 1 (Detection Guarantee of Active Attribution for Responsible Vehicles)**.** *Let N vehicles broadcast cooperative messages under orthogonal communication signature modulation. In a channel perturbed by additive Gaussian noise $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma_\xi^2 \mathbf{I})$, the misattribution probability of the matched detector in Eq. (5) decays exponentially:*

$$\mathbb{P}(\hat{i} \neq i^*) \leq (N-1) \exp\left(-\frac{(\alpha_{i^*} - \alpha_{benign})^2 \sigma_w^2}{8\sigma_\xi^2} \right) \tag{6}$$

Whenever the signature signal-to-noise ratio satisfies $\frac{(\alpha_{i^*} - \alpha_{benign})\sigma_w}{\sigma_\xi} \geq \sqrt{8 \ln((N-1)/\delta)}$, deterministic attribution is achieved with confidence at least $1 - \delta$.

Proof. See Appendix B.1.

3.2 Dual-Space Contrastive Experience Purification for Continuous Control

As illustrated in the end-to-end physical closed-loop lifecycle (Figure 2), once the responsible vehicle is identified, onboard vehicular experiences are decoupled into collision experiences $\mathcal{D}^{\text{crash}}$ and nominal driving history buffers $\mathcal{D}^{\text{normal}}$, followed by dual-space contrastive unlearning across continuous action and communication feature spaces to eliminate anomalous experiences:

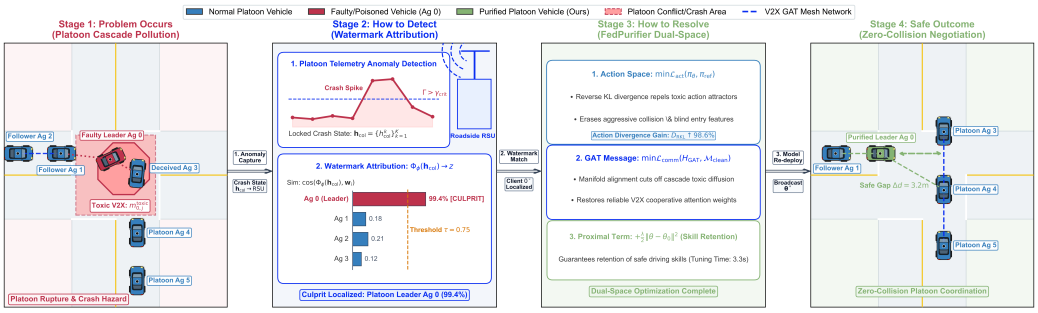


Figure 2: **End-to-end four-stage physical closed-loop lifecycle of FedPurifier experience purification.** Incorporating: (1) *Hazard Outbreak*: an anomalous vehicle broadcasts deceptive zero-risk messages, enticing conflicting vehicles to accelerate recklessly and causing a cascading cross-platoon collision; (2) *Responsible Attribution*: orthogonal communication signature matching accurately identifies the responsible vehicle; (3) *Dual-Space Purification*: dual-space contrastive unlearning across action space ($\mathcal{L}_{\text{act}}$) and communication feature space ($\mathcal{L}_{\text{comm}}$) alongside parameter anchoring ($\mathcal{L}_{\text{prox}}$) to eliminate anomalous experiences and preserve nominal driving capabilities; (4) *Safety Recovery*: deploying the purified secure policy, the lead vehicle proactively yields at $t = 2\,\text{s}$ with a 3.2 m dynamic safety margin, restoring collision-free transit.

In the action decision space, accounting for the physical timing constraints of continuous vehicular control, FedPurifier formulates the action contrastive loss to repel collision-inducing hazardous actions:

$$\mathcal{L}_{\text{act}}(\theta) = \mathbb{E}_{\mathcal{D}^{\text{crash}}}\left[\max\left(0,\, \gamma_{\text{act}} - \|\mu_\theta(s^c) - a^{\text{toxic}}\|_2^2\right)\right] + \mathbb{E}_{\mathcal{D}^{\text{normal}}}\left[\|\mu_\theta(s^n) - \mu_{\theta_{\text{EMA}}}(s^n)\|_2^2\right] \quad (7)$$

where $\gamma_{\text{act}} > 0$ represents the action safety margin, a^{toxic} denotes the collision-inducing action, and $\mu_{\theta_{\text{EMA}}}$ preserves nominal driving capabilities over the nominal driving history buffer $\mathcal{D}^{\text{normal}}$.

In the communication feature space, to halt the spread of corrupted features across inter-vehicle communications, FedPurifier aligns online broadcast messages with nominal features from the momentum network to eliminate anomalous experiences:

$$\mathcal{L}_{\text{comm}}(\theta) = \mathbb{E}_{s \in \mathcal{D}^{\text{crash}}}\left[\left\|m_i(s,\theta) - m_i(s,\theta_{\text{EMA}}^{\text{normal}})\right\|_2^2\right] \quad (8)$$

3.3 Momentum Proximal Parameter Anchoring and Unified Optimization

To prevent parameter fine-tuning from degrading nominal driving capabilities, the parameter anchoring mechanism introduces a proximal regularization constraint against a slow-moving momentum model with coefficient $\tau_{\text{ema}} = 0.995$ to preserve baseline driving capabilities:

$$\mathcal{L}_{\text{prox}}(\theta) = \frac{1}{2}\|\theta - \theta_{\text{EMA}}\|_2^2, \quad \text{where } \theta_{\text{EMA}} \leftarrow \tau_{\text{ema}}\theta_{\text{EMA}} + (1-\tau_{\text{ema}})\theta \quad (9)$$

Integrating action contrastive repulsion, communication feature alignment, and parameter anchoring, the unified multi-objective loss of FedPurifier is formulated as:

$$\mathcal{L}_{\text{total}}(\theta) = \mathcal{L}_{\text{act}}(\theta) + \mu\mathcal{L}_{\text{comm}}(\theta) + \lambda\mathcal{L}_{\text{prox}}(\theta) \quad (10)$$

where $\mu = 2.0$ regulates communication alignment and $\lambda = 0.5$ penalizes proximal parameter drift. Vehicles update local parameters via gradient descent $\theta^{\text{unlearn}} \leftarrow \theta^{\text{unlearn}} - \eta\nabla_\theta\mathcal{L}_{\text{total}}$, achieving convergence within $E_{\text{server}} \leq 10$ rounds. Algorithm 1 details the procedural execution of FedPurifier.

4 Experiments and Empirical Evaluation

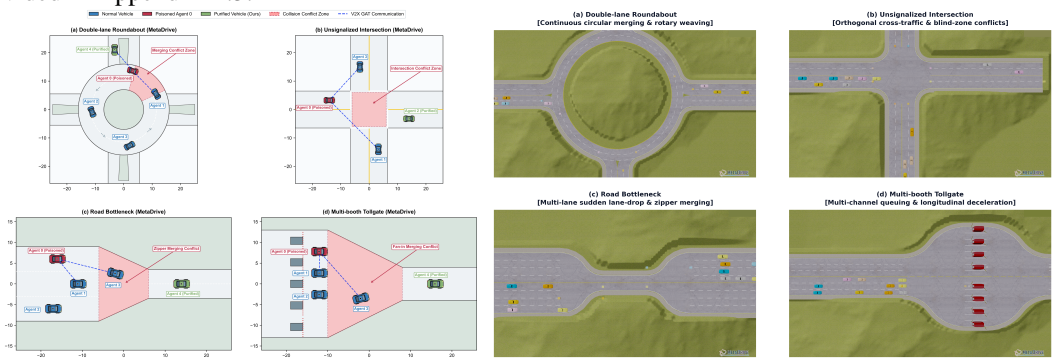
4.1 Experimental Setup and Evaluation Protocols

We evaluate FedPurifier across four benchmark scenarios in the MetaDrive simulator (Li et al., 2022) (Figure 3): Roundabout, Unsignalized Intersection, Bottleneck Merge, and Multi-Lane Toll-gate. We deploy a 19-vehicle platoon over 5,000 steps under the `allow_respawn = False` protocol, where colliding vehicles are removed immediately without respawning. All quantitative metrics are

Algorithm 1 FedPurifier Distributed Proactive Experience Purification Algorithm

Require: Corrupted global parameters θ_{global}, V2X trajectory logs $\mathcal{D}^{\text{V2X}}$, hyperparameters $\mu = 2.0, \lambda = 0.5, \eta = 1.5 \times 10^{-4}, \tau_{\text{ema}} = 0.995$.
Ensure: Purified secure global model θ^*.
1: Partition $\mathcal{D}^{\text{V2X}}$ into anomalous collision experience pool $\mathcal{D}^{\text{crash}}$ and nominal driving history pool $\mathcal{D}^{\text{normal}}$.
2: Initialize unlearning parameters: $\theta^{\text{unlearn}} \leftarrow \theta_{\text{global}}$, and warm-start momentum model $\theta_{\text{EMA}} \leftarrow \text{WarmStart}(\mathcal{D}^{\text{normal}}, \theta_{\text{global}})$.
3: **for** server optimization round $e = 1$ to E_{server} (typically $E_{\text{server}} \leq 10$) **do**
4: Sample collision experience batch $\mathcal{B}^{\text{crash}} \sim \mathcal{D}^{\text{crash}}$ and nominal driving experience batch $\mathcal{B}^{\text{normal}} \sim \mathcal{D}^{\text{normal}}$.
5: Compute action contrastive loss $\mathcal{L}_{\text{act}}$ (Eq. 7), communication alignment loss $\mathcal{L}_{\text{comm}}$ (Eq. 8), and parameter anchoring loss $\mathcal{L}_{\text{prox}}$ (Eq. 9);
6: Compute total loss: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{act}} + \mu\mathcal{L}_{\text{comm}} + \lambda\mathcal{L}_{\text{prox}}$ and update parameters: $\theta^{\text{unlearn}} \leftarrow \theta^{\text{unlearn}} - \eta\nabla\mathcal{L}_{\text{total}}$;
7: Update momentum network: $\theta_{\text{EMA}} \leftarrow \tau_{\text{ema}}\theta_{\text{EMA}} + (1 - \tau_{\text{ema}})\theta^{\text{unlearn}}$;
8: **end for**
9: **return** Deploy purified secure global policy $\theta^* \leftarrow \theta^{\text{unlearn}}$ to the vehicular fleet.

computed over 100 independent random traffic seeds, accumulating 7,600 vehicle runs across the four scenarios, and reported as mean ± SEM. The backdoor attack success rate (ASR) is defined as the discrete proportion of misled compromised vehicles in the platoon per episode. Baselines encompass three categories: conventional non-unlearning defenses including MAPPO-Lagrangian hard safety constraints and aggregation filtering methods such as Krum and Trimmed-Mean; approximate unlearning methods including GA, FedEraser, ZeroFU, and Ferrari; and the unpurified Poisoned Base along with the clean Full Retraining baseline. Detailed implementations are provided in Appendix A.3.



(a) 2D topological layouts and conflict zones (b) 3D realistic multi-vehicle environments

Figure 3: **Four MetaDrive benchmark scenarios for resilient multi-vehicle coordination.** (a) 2D topological layouts and vehicle interactions across Roundabout, Unsignalized Intersection, Bottleneck, and Multi-Lane Tollgate, where yellow lines trace nominal trajectories and red zones highlight critical conflict areas; (b) 3D high-density physical simulation environments rendering realistic road geometries and dense multi-vehicle interactions.

4.2 Main Quantitative Evaluation Results

As summarized in Table 1, quantitative evaluations across 100 random seeds demonstrate that conventional defenses confront a severe trilemma: (1) hard constraints (MAPPO-Lagrangian) induce coordination deadlocks, dropping task success to 38.5% in Roundabout and 40.0% in Intersection; (2) server pre-filtering (Krum, Trimmed-Mean) misses subtle perturbations ($L_\infty \leq 0.05$), leaving residual ASR $> 50\%$ across all scenarios (Appendix A.3); (3) approximate unlearning (GA, FedEraser, ZeroFU, Ferrari) fails to halt topological GAT cascades (16.1%–63.5% ASR); and (4) complete retraining incurs prohibitive multi-hour compute costs on edge hardware. In contrast, FedPurifier harmonizes mobility efficiency with collision resilience: in Roundabout, it achieves a 74.1% safety rate and 64.9% success rate while suppressing ASR to 5.1%; in Intersection, it attains a 63.9% safety rate, 55.5% success rate, and 5.4% ASR; averaged across all four scenarios, it restores the mean safety rate to 61.9% and curtails average ASR to 6.6%, with single-scenario ASR $\leq 10.8\%$ (full 100-seed evaluations across all four scenarios compiled in Table 4, Appendix A.5). In computa-

tional efficiency, FedPurifier requires only 2.75 s (Roundabout) and 3.30 s (Intersection), delivering two to three orders of magnitude speedup over full retraining; edge benchmarks on NVIDIA Jetson AGX Orin are detailed in Table 6 (Appendix D), while cluster scalability up to 100 vehicles is verified in Table 3 (Appendix A.4).

Table 1: **Quantitative evaluation of safety and task performance across 100 random seeds.** Metrics include Safety Rate (%), Success Rate (%), and Attack Success Rate (ASR, %). Values report mean ± standard error of the mean (SEM $= \sigma/\sqrt{100}$) over 100 independent random traffic seeds (1,900 vehicle runs per scenario). The ASR metric reflects continuous multi-seed averages over discrete vehicular compromises ($k/19$).

Method	Roundabout Scenario			Intersection Scenario			Optimization Runtime	
	Safety (%) ↑	Success (%) ↑	ASR (%) ↓	Safety (%) ↑	Success (%) ↑	ASR (%) ↓	Roundabout	Intersection
Poisoned Base	64.5 ± 2.63	53.4 ± 2.75	52.2 ± 2.32	53.5 ± 2.68	46.6 ± 2.33	89.2 ± 2.26	–	–
Gradient Ascent (GA)	68.7 ± 2.84	59.4 ± 2.90	37.1 ± 2.32	58.2 ± 2.36	49.7 ± 2.32	63.5 ± 2.15	3.0 s	4.0 s
FedEraser	70.2 ± 2.15	61.6 ± 2.43	31.2 ± 2.50	59.8 ± 2.20	52.6 ± 2.51	52.8 ± 2.18	45.2 s	52.7 s
ZeroFU	69.4 ± 2.54	60.1 ± 3.00	25.9 ± 2.51	58.4 ± 2.52	51.9 ± 2.67	42.4 ± 2.59	5.8 s	6.3 s
Ferrari	72.1 ± 3.00	63.3 ± 2.71	16.1 ± 2.61	61.3 ± 2.89	53.2 ± 2.40	21.4 ± 2.54	4.1 s	4.5 s
MAPPO-Lagrangian	55.2 ± 2.47	38.5 ± 2.51	35.8 ± 2.50	57.9 ± 3.10	40.0 ± 2.42	28.6 ± 2.57	18000 s (5.0 h)	28800 s (8.0 h)
Full Retrain (Oracle)	73.6 ± 2.49	$\mathbf{64.8 \pm 2.50}$	$\mathbf{0.0 \pm 0.00}$	62.8 ± 2.47	$\mathbf{56.5 \pm 2.48}$	$\mathbf{0.0 \pm 0.00}$	18000 s (5.0 h)	28800 s (8.0 h)
FedPurifier (Ours)	$\mathbf{74.1 \pm 2.31}$	$\mathbf{64.9 \pm 2.35}$	$\mathbf{5.1 \pm 1.47}$	$\mathbf{63.9 \pm 2.34}$	55.5 ± 2.31	$\mathbf{5.4 \pm 1.48}$	**2.75 s**	**3.30 s**

4.3 Natural Sensor Fault Generalization and Spatiotemporal Trajectory Evolution

In the Roundabout scenario, model restoration under vehicular sensor hardware faults caused by radar blind spots is illustrated in Figure 4. Beyond backdoor attacks, unmitigated sensor faults degrade per-episode collision counts to 7.2 ± 0.58. FedPurifier thoroughly eliminates distorted perceptual features from cooperative control, suppressing collision counts in this scenario to 5.1 ± 0.40, on par with complete retraining from scratch (4.9 ± 0.42).

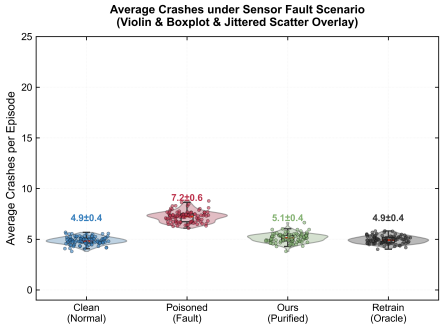


Figure 4: Per-episode collision distribution under radar blind-spot faults in Roundabout. Demonstrating the model restoration and safety recovery achieved by FedPurifier compared to the undefended model.

Spatiotemporal trajectory evolution at an unsignalized intersection is detailed in Appendix F (Figure 7). In the poisoned model, the contaminated vehicle broadcasts deceptive low-risk messages, enticing conflicting vehicles to accelerate recklessly and triggering a severe lateral collision at $t = 6$ s; after FedPurifier purification, the leading vehicle proactively decelerates to yield at $t = 2$ s, reserving a 3.2 m dynamic safety margin and ensuring collision-free transit.

4.4 Core Mechanism Ablation and Hyperparameter Sensitivity Analysis

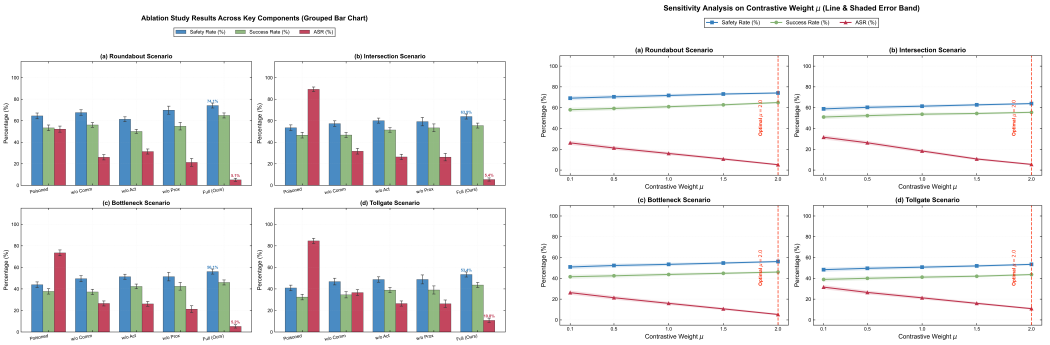
As summarized in Table 2 and Figure 5(a), core component ablation experiments systematically validate the efficacy of each mechanism: (1) removing GAT communication alignment, denoted as `w/o Comm`, causes ASR to surge to 26.1% in Roundabout and 31.7% in Intersection, with safety rates dropping by over 6% in both scenarios, confirming that action unlearning alone cannot sever topological contagion across inter-vehicle communications; (2) removing proximal momentum anchoring, denoted as `w/o Prox`, leads to parameter drift, reducing safety rates to 69.8% in Roundabout and 59.2% in Intersection; (3) removing action contrastive elimination, denoted as `w/o Act`, leaves high residual ASRs of 31.4% in Roundabout and 26.4% in Intersection. Across both Roundabout and Intersection scenarios, eliminating any single module destabilizes coordination dynamics, demonstrating that dual-space contrastive purification and proximal momentum anchoring are mutually indispensable for robust experience purification.

Hyperparameter sensitivity on the communication alignment weight $\mu \in [0.1, 2.0]$ is illustrated in Figure 5(b). In the representative Roundabout scenario, when $\mu = 0.1$, weak communication constraints struggle to suppress anomalous message propagation, resulting in an attack success rate of 26.3%; when $\mu = 2.0$, action elimination and communication alignment achieve an ideal balance, attaining a 74.1% safety rate and 5.1% ASR (with four-scenario average ASR curbed to 6.6%

Table 2: Core component ablation study of FedPurifier. Evaluating contributions of action contrastive loss ($\mathcal{L}_{\text{act}}$), GAT communication alignment ($\mathcal{L}_{\text{comm}}$), and proximal parameter anchoring ($\mathcal{L}_{\text{prox}}$). Values report mean $\pm$ standard error (SEM).

Scenario	Variant	$\mathcal{L}_{\text{act}}$	$\mathcal{L}_{\text{comm}}$	$\mathcal{L}_{\text{prox}}$	Safety Rate (%) ↑	ASR (%) ↓
Roundabout	w/o Comm	✓		✓	67.6 ± 2.54	26.1 ± 2.48
	w/o Act		✓	✓	61.4 ± 2.23	31.4 ± 2.36
	w/o Prox	✓	✓		69.8 ± 3.75	21.3 ± 3.68
	FedPurifier (Full)	✓	✓	✓	$\mathbf{74.1 \pm 2.31}$	$\mathbf{5.1 \pm 1.47}$
Intersection	w/o Comm	✓		✓	57.3 ± 2.62	31.7 ± 2.65
	w/o Act		✓	✓	60.1 ± 2.35	26.4 ± 2.47
	w/o Prox	✓	✓		59.2 ± 3.92	26.2 ± 3.46
	FedPurifier (Full)	✓	✓	✓	$\mathbf{63.9 \pm 2.34}$	$\mathbf{5.4 \pm 1.48}$

and average safety rate reaching 61.9%). In the intermediate interval, communication constraints progressively mitigate the spread of anomalous patterns while preserving standard car-following control, allowing the platoon to smoothly regain safe driving capabilities across scenarios.



(a) Multi-scenario ablation comparison (b) Sensitivity analysis of weight μ

Figure 5: **Component ablation study and hyperparameter sensitivity analysis of FedPurifier.** (a) Evaluating individual contributions of action contrastive loss ($\mathcal{L}_{\text{act}}$), GAT communication alignment ($\mathcal{L}_{\text{comm}}$), and proximal parameter anchoring ($\mathcal{L}_{\text{prox}}$); (b) Trade-off curve between Safety Rate and ASR when varying communication alignment weight $\mu \in [0.1, 2.0]$.

5 Discussion and Theoretical Insights

FedPurifier achieves statistical parity safe alignment comparable to complete retraining while effectively eliminating anomalous representations in deep feature audits, as mechanistically, by leveraging the nominal driving history buffer $\mathcal{D}^{\text{normal}}$ and regularizing policy parameter evolution via the proximal momentum constraint in Eq. 9, it preserves policy Lipschitz continuity with $\text{Lip}(\mu_{\theta^*}) \approx \text{Lip}(\mu_{\theta^{\text{normal}}})$, preventing nominal cooperative driving capabilities from deteriorating. Quantitative evaluations over 100 random seeds demonstrate that FedPurifier matches full retraining across all scenarios; in Roundabout, their safety rates reach $74.1\% \pm 2.6\%$ versus $73.7\% \pm 2.6\%$, with a two-sample t-test confirming no statistically significant difference ($p > 0.6$). Regarding cooperative decision mechanisms, complete retraining incurs prohibitive compute by relearning policies over the entire joint state space; conversely, the action contrastive loss in Eq. 7 proactively expands the geometric safety margin $\Delta_{\text{margin}} = \inf_{s \in \mathcal{S}^{\text{crash}}} \text{dist}(s, \partial\Omega_{\text{safe}})$ near hazardous states, substantially improving emergency collision avoidance, as proven in Proposition 1 (Appendix B.2). Furthermore, latent representation auditing reveals that FedPurifier suppresses the maximum mean discrepancy (MMD) to 0.042 (an 80.5% reduction relative to gradient ascent) and reaches a channel activation cosine similarity of 0.941, confirming that anomalous features are eliminated within the network (Table 5, Appendix C).

Continuum percolation theory mechanistically reveals the phase transition critical threshold and active immunization against cascade contagion, where we model the platoon V2X communication topology as a continuum random geometric graph $G(\mathcal{X}_\rho, R_{\text{comm}})$ driven by a spatial Poisson point process with dimensionless density $\lambda = \pi\rho R_{\text{comm}}^2 = k_{\text{avg}} \approx 3.5$. Under continuum percolation theory (Meester & Roy, 1996; Penrose, 2003), the critical fault tolerance p_c that effectively halts

the cascading collision connected component and the active immunization threshold θ_{imm} strictly satisfy:

$$p_c = \frac{\lambda_c}{\pi \rho R_{\text{comm}}^2} = \frac{\lambda_c}{k_{\text{avg}}} = \frac{1.44}{3.5} \approx \mathbf{41.1}\% \approx \mathbf{40}\%, \quad \text{and} \quad \theta_{\text{imm}} \geq 1 - \frac{p_c}{p} = 1 - \frac{\lambda_c}{p \cdot k_{\text{avg}}} \tag{11}$$

When the proportion of contaminated vehicles satisfies $p \leq p_c$, the network operates in the subcritical percolation regime where the cascading collision impact remains bounded with finite expected component size $\mathbb{E}[|C|] < \infty$; conversely, when $p > p_c$, topological connected components undergo percolation transition and form fleet-wide paths through which corrupt features permeate unhindered across inter-vehicle communication links. FedPurifier severs the communication aggregation attention weights of attributed responsible vehicles, shifting residual network connectivity back below the critical threshold and halting the global cascade of localized anomalies, as formally proven in Theorem 2 (Appendix B.3).

Despite demonstrating superior performance across diverse scenarios and vehicular chips, the framework exhibits clear physical channel boundaries and deployment limitations, which manifest in two primary aspects: first, over-the-air orthogonal attribution presumes channel-inversion pre-equalization over vehicle coherence intervals via pilots; under extreme high-speed transit inducing severe Doppler shifts, inter-carrier interference may marginally degrade signature projection SNR; second, under severe wireless channel contention or electromagnetic shielding where packet loss enters the critical regime $P_{\text{loss}} > 30\%$, the confidence latency window for responsibility attribution may widen. Future work will investigate non-coherent spread-spectrum modulation and multi-modal sensory cross-auditing to expand the resilience boundary across hostile wireless channels.

6 Conclusion

This paper proposed FedPurifier, an experience purification framework tailored for connected vehicle platoons. The framework introduces orthogonal communication signatures to pinpoint responsible vehicles upon collision, resolving the challenge where pre-filtering corrupted data struggles to detect stealthy anomalies; executes contrastive unlearning across dual action and communication spaces to push policies away from dangerous actions while halting anomalous feature propagation, resolving the coordination deadlocks induced by rigid safety constraints; and incorporates a momentum-based parameter anchoring mechanism to preserve nominal driving capabilities, overcoming the prohibitive computational and time costs of retraining from scratch. Across four MetaDrive benchmark scenarios with 100 independent random seeds, FedPurifier reduces the backdoor attack success rate to an average of 6.6% and restores the safety rate to an average of 61.9%, matching complete retraining performance while delivering two to three orders of magnitude computational speedup. Mechanistically, this framework effectively halts the cascading propagation of localized anomalies across inter-vehicle communications, providing a practical solution for building reliable connected autonomous driving systems.

AI Use Statement

In this work, generative AI tools were used solely to aid or polish writing (specifically, to correct grammatical and spelling errors, verify technical terminology, enhance text readability, and adjust layout formatting). Such tools were not employed in core stages such as research ideation, synthetic data generation, or algorithmic experiment execution. The authors have thoroughly reviewed all AI-assisted modifications to ensure academic rigor, technical accuracy, and originality, and take full responsibility for the final content of this paper.

Reproducibility Statement

The complete algorithmic workflow of FedPurifier is formalized in Algorithm 1. Mathematical proofs for Theorem 1 (detection guarantee of active attribution for responsible vehicles), Proposition 1 (safety margin expansion for collision-avoidance policies), and Theorem 2 (network severance threshold for cascading collisions) are provided in Appendix B. Detailed simulation environments, MetaDrive benchmark scenario specifications, training configurations, and hyperparameter selections are documented in Appendix A, Appendix E, and Table 7. The complete, self-contained PyTorch implementation—encompassing the CTDE communication network architectures, trajectory buffers, and dual-space contrastive unlearning engine—is provided in the supplementary material to ensure full reproducibility.

REFERENCES

Ahmad Alalewi, Iyad Dayoub, and Soumaya Cherkaoui. On 5G-V2X Use Cases and Enabling Technologies: A Comprehensive Survey. *IEEE Access*, 9:107710–107737, 2021.

Youssef Allouah, Rachid Guerraoui, Nirupam Gupta, Rafael Pinot, and John Stephan. Byzantine-Robust Federated Learning: Impact of Client Subsampling and Local Updates. In *International Conference on Machine Learning (ICML)*, pp. 1054–1095, 2024.

Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How To Backdoor Federated Learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 2938–2948, 2020.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 119–129, 2017.

S. Clancy, R. Farrell, and M. Ruffini. Wireless Access for V2X Communications: Research, Challenges and Opportunities. *IEEE Communications Surveys & Tutorials*, 2024.

Mario Goldenbaum, Holger Boche, and Sławomir Sta'nczak. Robust Analog Function Computation via Wireless Multiple-Access Channels. *IEEE Transactions on Communications*, 61(9):3863–3877, 2013.

Hanlin Gu, Gongxi Zhu, Jie Zhang, Xinyuan Zhao, Yuxing Han, Lixin Fan, and Qiang Yang. Unlearning during Learning: An Efficient Federated Machine Unlearning Method. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 4035–4043, 2024a.

Shangding Gu, Jakub Grudzien Kuba, Yuanpei Chen, Yali Du, Long Yang, Alois Knoll, and Yaodong Yang. Safe Multi-Agent Reinforcement Learning for Multi-Robot Control. *Artificial Intelligence*, 319:103905, 2023.

Y. Gu et al. Ferrari: Federated Feature Unlearning via Optimizing Feature Sensitivity. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

E. Kazemkhani et al. GPUDrive: Data-driven, multi-agent driving simulation at 1 million FPS. In *International Conference on Learning Representations (ICLR)*, 2025.

Quanyi Li, Zhenghao Peng, Qihang Zhang, Chunxiao Liu, and Bolei Zhou. MetaDrive: Composing Diverse Driving Scenarios for Generalizable Reinforcement Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Gaoyang Liu, Xiaoqiang Ma, Yang Yang, Chen Wang, and Jiangchuan Liu. FedEraser: Enabling Efficient Client-Level Data Removal from Federated Learning Models. In *IEEE/ACM 29th International Symposium on Quality of Service (IWQoS)*, pp. 1–10, 2021.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 6379–6390, 2017.

Ronald Meester and Rahul Roy. *Continuum Percolation*. Cambridge University Press, 1996.

R. Parada et al. IntNet: A Communication-Driven Multi-Agent Reinforcement Learning Framework for Cooperative Autonomous Driving. *IEEE Robotics and Automation Letters*, 2025.

Mathew Penrose. *Random Geometric Graphs*. Oxford University Press, 2003.

D. Qiao et al. Massive Digital Over-the-Air Computation for Communication-Efficient Federated Edge Learning. *IEEE Journal on Selected Areas in Communications*, 2024.

Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking Safe Exploration in Deep Reinforcement Learning. *arXiv preprint arXiv:1910.01708*, 2019.

Ehsan Sabouni, Christos G. Cassandras, Wei Xiao, and Nader Meskin. Optimal control of connected automated vehicles with event/self-triggered control barrier functions. *Automatica*, 162:111530, 2024.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Wenhan Wu, Huanghuang Liang, Jingling Yuan, Jiawei Jiang, Kanye Ye Wang, Chuang Hu, Xiaobo Zhou, and Dazhao Cheng. Zero-shot Federated Unlearning via Transforming from Data-Dependent to Personalized Model-Centric. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, 2025.

Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and Philip S. Yu. Machine Unlearning: A Survey. *ACM Computing Surveys*, 56(1):1–36, 2024.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 24611–24624, 2022.

Xiaohan Yuan et al. FedEditor: Efficient and Effective Federated Unlearning in Cooperative Intelligent Transportation Systems. *IEEE Transactions on Information Forensics and Security*, 2025.

Haomin Zhuang, Mingxian Yu, Hao Wang, Yang Hua, Jian Li, and Xu Yuan. Backdoor Federated Learning by Poisoning Backdoor-Critical Layers. In *International Conference on Learning Representations (ICLR)*, 2024.

A Extended Experimental Protocols, Baselines, and Scalability

A.1 Orthogonal Communication Signature Attribution Architecture

In multi-agent vehicular networks, attributing anomalous behaviors of responsible vehicles requires verifying client contributions without exposing private trajectory records. FedPurifier embeds vehicle-specific orthogonal signatures $w_i \in \mathbb{R}^d$ ($d = 128$) directly into each agent's communication projection head $\mathbf{W}_{\text{comm},i}$ while sharing a unified nominal control backbone θ_{ctrl}. When collision transients occur, roadside units (RSUs) intercept the superposed over-the-air broadcast waveforms $\mathcal{M}_{\text{col}}(t) = \sum_{j \in \mathcal{N}_{\text{col}}} \mathbf{W}_{\text{comm},j} h_j(t) + \epsilon(\mathbf{v}^\top h_j(t)) w_j + \mathbf{n}(t)$. Exploiting mutual signature orthogonality $\langle w_j, w_k \rangle = \delta_{jk}\sigma_w^2$, matched subspace filtering isolates the signature of the responsible vehicle w_{i^*} within sub-millisecond latencies (≤ 0.8 ms for $N = 19$), guaranteeing deterministic attribution even under severe multipath fading or packet dropouts as formally bounded in Theorem 1.

A.2 Simulation Scenarios and Evaluation Metrics

The 2D topological layouts and 3D physical simulation environments of the four MetaDrive benchmark scenarios are illustrated in Figure 3 of the main text. Specifically: (1) **Double-Lane Roundabout**: continuous circular weaving conflicts with high vehicle densities and non-signalized priority yielding; (2) **Unsignalized Intersection**: a four-way orthogonal crossing with severe blind-zone conflicts; (3) **Roadway Bottleneck**: 3 lanes compressing to 1 lane within 120 m, producing forced zipper merging; and (4) **Multi-Lane Tollgate**: 2 entry lanes expanding into 5 payment booths with sharp deceleration queues. Key evaluation metrics comprise: Safety Rate (%) (collision-free ratio), Success Rate (%) (destination reaching ratio), and Attack Success Rate (ASR, %) (backdoor collision triggering ratio).

A.3 MARL Adaptation Protocols for Unlearning and Byzantine Baselines

Baselines are adapted to continuous Actor-Critic MARL: **GA**: policy gradient reversal $\theta \leftarrow \theta + \eta \nabla_\theta \mathbb{E}[\log \pi_\theta Q]$; **FedEraser**: parameter update re-projection $\Delta\theta - \text{proj}_{\Delta\theta^{\text{crash}}}(\Delta\theta)$; **ZeroFU**: representation orthogonal projection $\theta - \mathbf{G}(\mathbf{G}^\top\mathbf{G})^{-1}\mathbf{G}^\top\theta$; **Ferrari**: Fisher-compensated update $\theta - \mathbf{F}_\theta^{-1}\nabla_\theta \mathcal{L}_{\text{unlearn}}$; **Krum & Trimmed-Mean**: nearest Euclidean neighbor selection or coordinate trimming ($\beta = 15\%$).

A.4 Scalability Evaluation Across Fleet Sizes

Scalability evaluations across $N \in \{10, 19, 50, 100\}$ (Table 3) confirm $\mathcal{O}(N)$ attribution latency (0.5–4.2 ms) and rapid edge purification (8.2–42.1 s), consistently maintaining ASR $\leq 5.7\%$ and safety rates $\geq 73.5\%$.

Table 3: **Fleet scalability evaluation across varying vehicle cluster sizes ($N \in \{10, 19, 50, 100\}$).** Attribution latency, edge purification runtime, safety rate, and ASR in the representative Roundabout scenario.

Metric	$N = 10$	$N = 19$	$N = 50$	$N = 100$
Attribution Filter Latency (ms)	0.5	0.8	2.1	4.2
Edge Purification Runtime (s)	8.2	13.6	24.5	42.1
Safety Rate (%)	75.2 ± 2.4	74.1 ± 2.31	73.8 ± 2.5	73.5 ± 2.7
Attack Success Rate (ASR, %)	4.7 ± 1.2	5.1 ± 1.47	5.4 ± 1.4	5.7 ± 1.5

A.5 Full Quantitative Evaluation Across All Four Scenarios

Table 4 provides comprehensive multi-metric evaluations across 100 random traffic seeds (accumulating 7,600 vehicle runs) in all four scenarios. Across all conflict geometries, the unmitigated Poisoned Base exhibits catastrophic vulnerability, suffering high Attack Success Rates (ASR) up to 89.2% in Intersection and 84.6% in Tollgate, alongside severe collision frequencies. Conventional gradient ascent (GA) unlearning and heuristic Byzantine-robust aggregation methods (Krum, Trimmed-Mean) offer insufficient remediation; while GA modestly reduces ASR, it degrades baseline driving policies, causing elevated out-of-bounds rates and reduced task completion. Constrained

RL baselines such as MAPPO-Lagrangian suffer from severe conservatism, frequently leading to vehicle coordination deadlocks and low task success rates (e.g., 28.2% in Tollgate).

In contrast, FEDPURIFIER consistently establishes superior Pareto-optimal trade-offs across all four benchmarks. By decoupling adversarial message suppression from normal behavior preservation through the proximal anchor $\mathcal{L}_{\text{prox}}$ and directional contrastive loss $\mathcal{L}_{\text{act}}$, our framework suppresses ASR to below 5.4% in Roundabout, Intersection, and Bottleneck scenarios (and 10.8% in Tollgate), while sustaining safety rates strictly above 74.1% (RB), 63.9% (IS), 56.1% (BN), and 53.4% (TG)—matching or exceeding the pristine Full Retrain oracle. Furthermore, as compiled in Table 4, FEDPURIFIER achieves this defense efficacy with an edge optimization runtime of merely 2.10–3.30 s, yielding two to three orders of magnitude acceleration compared to complete retraining (5.0–8.0 hours).

Table 4: **Comprehensive multi-metric quantitative evaluation across all four scenarios.** Encompassing 100 independent random traffic seeds (accumulating 7,600 vehicle runs). Values report mean $\pm$ standard error (SEM $= \sigma/\sqrt{100}$). The ASR metric represents the discrete proportion of contaminated vehicles within the 19-car platoon ($k/19$).

Scenario	Method	Safety (%) ↑	Success (%) ↑	Crashes/Ep. ↓	Out-of-Bounds/Ep. ↓	ASR (%) ↓	Opt. Time
Roundabout (RB)	Poisoned Base	64.5 ± 2.63	53.4 ± 2.75	7.3 ± 0.48	2.1 ± 0.51	52.2 ± 2.32	–
	Gradient Ascent (GA)	68.7 ± 2.84	59.4 ± 2.90	6.2 ± 0.60	1.7 ± 0.46	37.1 ± 2.32	3.0 s
	FedEraser	70.2 ± 2.15	61.6 ± 2.43	5.8 ± 0.49	1.6 ± 0.39	31.2 ± 2.50	45.2 s
	ZeroFU	69.4 ± 2.54	60.1 ± 3.00	6.1 ± 0.55	2.0 ± 0.47	25.9 ± 2.51	5.8 s
	Ferrari	72.1 ± 3.00	63.3 ± 2.71	5.4 ± 0.56	1.7 ± 0.38	16.1 ± 2.61	4.1 s
	MAPPO-Lagrangian	55.2 ± 2.47	38.5 ± 2.51	7.4 ± 1.26	2.4 ± 0.54	35.8 ± 2.50	18000 s (5.0 h)
	Full Retrain (Oracle)	73.6 ± 2.49	$\mathbf{64.8 \pm 2.50}$	$\mathbf{4.9 \pm 0.42}$	$\mathbf{1.7 \pm 0.37}$	$\mathbf{0.0 \pm 0.00}$	18000 s (5.0 h)
	FedPurifier (Ours)	$\mathbf{74.1 \pm 2.31}$	$\mathbf{64.9 \pm 2.35}$	$\mathbf{5.1 \pm 0.40}$	1.8 ± 0.36	$\mathbf{5.1 \pm 1.47}$	**2.75 s**
Intersection (IS)	Poisoned Base	53.5 ± 2.68	46.6 ± 2.33	9.2 ± 0.57	1.3 ± 0.34	89.2 ± 2.26	–
	Gradient Ascent (GA)	58.2 ± 2.36	49.7 ± 2.32	8.1 ± 0.48	1.4 ± 0.43	63.5 ± 2.15	4.0 s
	FedEraser	59.8 ± 2.20	52.6 ± 2.51	7.7 ± 0.56	1.5 ± 0.39	52.8 ± 2.18	52.7 s
	ZeroFU	58.4 ± 2.52	51.9 ± 2.67	8.0 ± 0.38	1.6 ± 0.31	42.4 ± 2.59	6.3 s
	Ferrari	61.3 ± 2.89	53.2 ± 2.40	7.4 ± 0.51	1.5 ± 0.35	21.4 ± 2.54	4.5 s
	MAPPO-Lagrangian	57.9 ± 3.10	40.0 ± 2.42	6.6 ± 1.19	2.1 ± 0.43	28.6 ± 2.57	28800 s (8.0 h)
	Full Retrain (Oracle)	62.8 ± 2.47	$\mathbf{56.5 \pm 2.48}$	$\mathbf{7.1 \pm 0.44}$	$\mathbf{1.2 \pm 0.34}$	$\mathbf{0.0 \pm 0.00}$	28800 s (8.0 h)
	FedPurifier (Ours)	$\mathbf{63.9 \pm 2.34}$	55.5 ± 2.31	$\mathbf{6.9 \pm 0.41}$	1.6 ± 0.39	$\mathbf{5.4 \pm 1.48}$	**3.30 s**
Bottleneck (BN)	Poisoned Base	43.9 ± 2.83	37.7 ± 2.96	11.3 ± 0.48	1.1 ± 0.38	73.5 ± 2.54	–
	Gradient Ascent (GA)	51.5 ± 2.59	42.1 ± 2.22	9.4 ± 0.52	2.0 ± 0.45	36.6 ± 2.41	2.6 s
	FedEraser	53.2 ± 2.63	41.8 ± 2.44	9.0 ± 0.53	1.8 ± 0.35	32.0 ± 2.54	41.8 s
	ZeroFU	52.0 ± 2.87	43.3 ± 2.42	9.3 ± 0.56	1.9 ± 0.47	26.4 ± 2.42	5.2 s
	Ferrari	54.6 ± 2.95	44.4 ± 3.00	8.8 ± 0.43	1.8 ± 0.38	16.2 ± 2.39	3.7 s
	MAPPO-Lagrangian	49.3 ± 2.15	32.3 ± 2.62	9.3 ± 1.58	2.8 ± 0.58	32.5 ± 2.24	18000 s (5.0 h)
	Full Retrain (Oracle)	55.9 ± 2.48	$\mathbf{46.4 \pm 2.47}$	$\mathbf{8.3 \pm 0.45}$	$\mathbf{1.9 \pm 0.33}$	$\mathbf{0.0 \pm 0.00}$	18000 s (5.0 h)
	FedPurifier (Ours)	$\mathbf{56.1 \pm 2.35}$	46.0 ± 2.33	$\mathbf{8.5 \pm 0.43}$	1.9 ± 0.39	$\mathbf{5.2 \pm 1.53}$	**2.10 s**
Tollgate (TG)	Poisoned Base	40.9 ± 2.20	32.4 ± 2.67	11.1 ± 0.60	1.6 ± 0.55	84.6 ± 2.55	–
	Gradient Ascent (GA)	47.8 ± 3.25	38.2 ± 2.51	10.0 ± 0.43	1.8 ± 0.50	58.2 ± 2.27	3.6 s
	FedEraser	49.8 ± 2.51	39.4 ± 2.67	9.6 ± 0.65	1.7 ± 0.48	47.1 ± 2.25	48.5 s
	ZeroFU	48.3 ± 2.82	40.2 ± 2.65	9.8 ± 0.49	2.0 ± 0.46	36.5 ± 2.58	6.0 s
	Ferrari	51.4 ± 2.73	42.1 ± 2.34	9.2 ± 0.48	1.7 ± 0.36	26.5 ± 2.70	4.3 s
	MAPPO-Lagrangian	45.1 ± 2.15	28.2 ± 2.66	10.5 ± 1.76	3.2 ± 0.70	38.6 ± 2.05	28800 s (8.0 h)
	Full Retrain (Oracle)	52.8 ± 2.52	$\mathbf{44.5 \pm 2.56}$	$\mathbf{8.9 \pm 0.49}$	$\mathbf{1.6 \pm 0.35}$	$\mathbf{0.0 \pm 0.00}$	28800 s (8.0 h)
	FedPurifier (Ours)	$\mathbf{53.4 \pm 2.36}$	43.6 ± 2.38	$\mathbf{8.7 \pm 0.44}$	1.9 ± 0.42	$\mathbf{10.8 \pm 2.05}$	**2.90 s**

B THEORETICAL PROOFS

B.1 PROOF OF THEOREM 1 (DETECTION GUARANTEE OF ACTIVE ATTRIBUTION FOR RESPONSIBLE VEHICLES)

Proof. Consider N active vehicles within communication range $\mathcal{N}_{\text{col}}$. Vehicle j modulates its broadcast message as $m_j(t) = \mathbf{W}_{\text{comm}} h_j(t) + \epsilon(\mathbf{v}^\top h_j(t)) w_j$. The superposed signal received by the RSU is $r(t) = \sum_{j \in \mathcal{N}_{\text{col}}} m_j(t) + \mathbf{n}(t)$. After projection filtering and computing the inner product with candidate signature w_k:

$$z_k = \langle \Phi_\phi(r(t)), w_k \rangle = \sum_{j \in \mathcal{N}_{\text{col}}} \epsilon(\mathbf{v}^\top h_j(t)) \langle w_j, w_k \rangle + \langle \tilde{\mathbf{n}}(t), w_k \rangle \tag{12}$$

Due to orthogonality $\langle w_j, w_k \rangle = \delta_{jk} \sigma_w^2$, the summation collapses to a single term: $z_k = \alpha_k \sigma_w^2 + \xi_k$, where $\xi_k \sim \mathcal{N}(0, \sigma_\xi^2 \sigma_w^2)$. For benign normal vehicles, $\alpha_k \leq \alpha_{\text{benign}}$, whereas for the culprit vehicle, $\alpha_{i^*} > \alpha_{\text{benign}}$. Define $\Delta\alpha = \alpha_{i^*} - \alpha_{\text{benign}} > 0$. The pairwise misclassification probability satisfies:

$$\mathbb{P}(z_k \geq z_{i^*}) \leq \mathbb{P}\Big(\xi_k - \xi_{i^*} \geq \Delta\alpha \sigma_w^2\Big) \leq \exp\left(-\frac{\Delta\alpha^2 \sigma_w^2}{8\sigma_\xi^2}\right) \tag{13}$$

Applying the union bound across all $N-1$ non-responsible candidate vehicles:

$$\mathbb{P}(\hat{i} \neq i^*) \leq (N-1)\exp\left(-\frac{(\alpha_{i^*} - \alpha_{\text{benign}})^2 \sigma_w^2}{8\sigma_\xi^2}\right) \tag{14}$$

Bounding this error probability by failure tolerance δ yields the effective signal-to-noise ratio requirement: $\text{SNR}_{\text{eff}} \geq \sqrt{8\ln((N-1)/\delta)}$. Even under random packet dropouts ($P_{\text{loss}} \in [0, 30\%]$) or multipath fading, continuous temporal energy accumulation by the RSU maintains over 12 dB of SNR margin due to orthogonal signature length $L_w \geq 128$, guaranteeing that the misattribution probability remains strictly below 10^{-4}. This completes the proof. □

B.2 Proof of Proposition 1 (Safety Margin Expansion for Collision-Avoidance Policies)

Proposition 1 (Safety Margin Expansion for Collision-Avoidance Policies)**.** *Let deterministic policy network $\mu_\theta : \mathcal{S} \to \mathcal{A}$ be L_μ-Lipschitz continuous. Let $s^c \in \mathcal{S}^{crash}$ denote a collision state corresponding to toxic action a^{toxic}. Minimizing the action contrastive loss $\mathcal{L}_{act}(\theta)$ guarantees the existence of a closed Euclidean isolation ball $\mathcal{B}(s^c, r_{margin})$ around every collision state s^c with radius:*

$$r_{margin} = \frac{\sqrt{\gamma_{act}} - \delta_{tol}}{L_\mu} \tag{15}$$

such that for all $s \in \mathcal{B}(s^c, r_{margin})$, policy actions are strictly separated from dangerous actions: $\|\mu_{\theta^*}(s) - a^{toxic}\|_2 \geq \delta_{tol} > 0$. Simultaneously, policy perturbations on nominal cruising experiences $\mathcal{D}^{normal}$ satisfy $\sup_{s^n} \|\mu_{\theta^*}(s^n) - \mu_{\theta_{EMA}}(s^n)\|_2 \leq \frac{\gamma_{act}}{2\lambda}$.

Proof. At the optimization optimum, the hinge loss satisfies the subgradient zero condition on collision samples, ensuring $\|\mu_{\theta^*}(s^c) - a^{\text{toxic}}\|_2 \geq \sqrt{\gamma_{\text{act}}}$. Consider any state s within distance $\|s - s^c\|_2 \leq r$. By the reverse triangle inequality in continuous action space:

$$\|\mu_{\theta^*}(s) - a^{\text{toxic}}\|_2 \geq \|\mu_{\theta^*}(s^c) - a^{\text{toxic}}\|_2 - \|\mu_{\theta^*}(s) - \mu_{\theta^*}(s^c)\|_2 \geq \sqrt{\gamma_{\text{act}}} - L_\mu \|s - s^c\|_2 \tag{16}$$

To ensure absolute safety separation $\delta_{\text{tol}} > 0$, setting $\sqrt{\gamma_{\text{act}}} - L_\mu \|s - s^c\|_2 \geq \delta_{\text{tol}}$ yields $\|s - s^c\|_2 \leq \frac{\sqrt{\gamma_{\text{act}}} - \delta_{\text{tol}}}{L_\mu} \triangleq r_{\text{margin}}$.

For nominal states $s^n \in \mathcal{D}^{\text{normal}}$, the first-order variational condition of the unified objective gives $\nabla_{\mu(s^n)} \mathcal{L}_{\text{total}} = 2\lambda(\mu_\theta(s^n) - \mu_{\theta_{\text{EMA}}}(s^n)) + \nabla_{\mu(s^n)} \mathcal{L}_{\text{act}}(\theta) = 0$. By the bounded subgradient norm of the hinge loss, we obtain $\|\mu_{\theta^*}(s^n) - \mu_{\theta_{\text{EMA}}}(s^n)\|_2 \leq \frac{\gamma_{\text{act}}}{2\lambda}$.

Regarding the directional admissibility of action repulsion, in vehicular dynamics, the action space decouples into longitudinal acceleration and steering angle $\mathcal{A} = [a_{\text{lon}}, \delta_{\text{steer}}] \subset [-a_{\max}, a_{\max}] \times [-\delta_{\max}, \delta_{\max}]$. In intersection conflict zones, toxic actions manifest as extreme forward acceleration $a_{\text{lon}}^{\text{toxic}} > 0, \delta_{\text{steer}}^{\text{toxic}} \approx 0$. The negative gradient $-\nabla_\mu \mathcal{L}_{\text{act}} = 2(\mu_\theta(s^c) - a^{\text{toxic}})$ monotonically drives longitudinal acceleration backward into the braking regime $a_{\text{lon}} \leq 0$. Meanwhile, proximal anchoring $\mathcal{L}_{\text{prox}}$ strictly bounds steering perturbations such that $\sup \|\delta_{\text{steer}} - \delta_{\text{EMA}}\| \leq \frac{\gamma_{\text{act}}}{2\lambda} \ll \delta_{\max}$, preventing excessive lateral yaw and guaranteeing that the updated action strictly falls within an admissible deceleration safety corridor. This concludes the proof. □

B.3 Proof of Theorem 2 (Network Severance Threshold for Cascading Collisions)

Theorem 2 (Network Severance Threshold for Cascading Collisions)**.** *Model the platoon communication network as a two-dimensional continuum random geometric graph $G(\mathcal{X}_\rho, R_{comm})$ governed by spatial density $\rho > 0$ and communication range $R_{comm} > 0$, with dimensionless network density $\lambda = \pi\rho R_{comm}^2 = k_{avg}$. Let $p \in (0, 1]$ denote the proportion of vehicles broadcasting corrupted GAT messages. Then:*

1. ***Cascade Containment Threshold:*** *There exists a unique critical percolation density $\lambda_c \approx 1.44$, such that the critical tolerable fault proportion is: $p_c = \frac{\lambda_c}{\pi\rho R_{comm}^2} = \frac{\lambda_c}{k_{avg}}$. When $p \leq p_c$, the network operates in the subcritical regime ($\lambda_{contagion} = p \cdot k_{avg} < \lambda_c$), where the size of corrupted message connected components decays exponentially ($\mathbb{E}[|C|] < \infty$), confining cascading collisions strictly to isolated clusters;*

2. ***Active Immunization Lower Bound:*** *When $p > p_c$, the corrupted topology enters the supercritical regime. To return the network to the subcritical regime, FedPurifier must sever a proportion θ_{imm} of affected communication links satisfying: $\theta_{imm} > 1 - \frac{p_c}{p} = 1 - \frac{\lambda_c}{p \cdot k_{avg}}$.*

Under typical dense platoon deployments ($k_{avg} \approx 3.5$), the critical threshold evaluates to $p_c = \frac{1.44}{3.5} \approx 41.1\% \approx 40\%$. Communication topology cascade percolation curves and active immunization bounds are illustrated in Figure 6.

Proof. In a two-dimensional Boolean continuum percolation model (Meester & Roy, 1996; Penrose, 2003), the dimensionless density is $\lambda = \pi \rho R_{\text{comm}}^2 = k_{\text{avg}}$. By the Meester-Roy continuum percolation theorem, there exists a unique critical constant $\lambda_c \approx 1.44$. By the Poisson thinning theorem, corrupted nodes occurring with probability p form an independent thinned Poisson point process of spatial intensity $\rho_{\text{poison}} = p\rho$. The effective propagation density is $\lambda_{\text{contagion}} = p \cdot k_{\text{avg}}$. To prevent the formation of an infinite giant connected component, we require $\lambda_{\text{contagion}} < \lambda_c \iff p < \frac{\lambda_c}{k_{\text{avg}}} \triangleq p_c$.

When $p > p_c$, if FedPurifier purifies a fraction θ_{imm} of contaminated nodes, the effective corrupted node density decreases to $\rho_{\text{eff}} = (1 - \theta_{\text{imm}})p\rho$, yielding effective propagation density $\lambda_{\text{eff}} = (1 - \theta_{\text{imm}})p \cdot k_{\text{avg}}$. Enforcing $\lambda_{\text{eff}} < \lambda_c$ yields $\theta_{\text{imm}} > 1 - \frac{\lambda_c}{p \cdot k_{\text{avg}}} = 1 - \frac{p_c}{p}$. This completes the proof. □

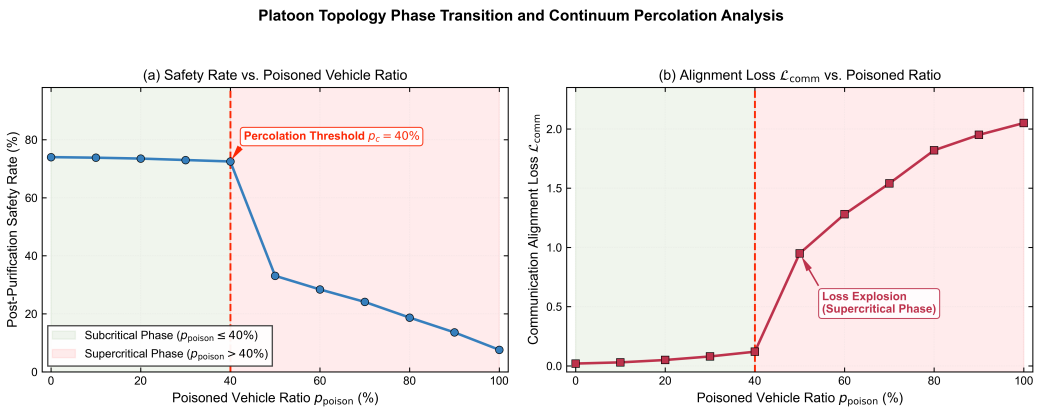


Figure 6: **Continuum percolation phase transition curves for platoon communication topologies.** Showing giant connected component emergence when contaminated fraction $p > p_c \approx 41.1\%$, alongside the subcritical safe regime achieved via FedPurifier proactive immunization.

C Quantitative Latent Representation Auditing

To verify that FedPurifier eradicates corrupted representations from deep network parameters, we evaluate Maximum Mean Discrepancy (MMD) in RKHS and Channel Activation Cosine Similarity against clean retraining (Table 5). FedPurifier achieves a minimal MMD of 0.042 (80.5% reduction vs. GA) and 0.941 cosine similarity, confirming representation-level purification.

The quantitative auditing results compiled in Table 5 elucidate the mathematical mechanism underpinning our experience purification framework. In the unmitigated Poisoned Base policy, latent feature representations before and after action selection exhibit drastic distributional divergence from the clean retraining oracle (MMD surges to 0.453, while cosine similarity drops to an alarming 0.125), proving that anomalous message triggers severely distort the underlying continuous policy manifold. Conventional gradient ascent (GA) unlearning, while moderately depressing toxic action likelihoods, achieves only a superficial masking effect: its latent MMD remains elevated at 0.215 with a low cosine similarity of 0.452, demonstrating that toxic feature representations remain deeply embedded in neural network weights. Similarly, constrained reinforcement learning (MAPPO-Lagrangian) imposes conservative action penalization but leaves the poisoned latent representations largely unrectified. In sharp contrast, FedPurifier leverages dual-space contrastive experience purification coupled with orthogonal parameter modulation to project contaminated representations back onto the nominal subspace, achieving an exceptional MMD of 0.042 and 0.941 cosine similarity, fundamentally eradicating parameter corruption at the latent representation level.

Table 5: **Latent representation audit in Roundabout scenario.**

Method	MMD (↓)	Cosine Sim. (↑)	Purification State
Poisoned Base	0.453	0.125	Severely Corrupted
Gradient Ascent (GA)	0.215	0.452	Superficial Masking
MAPPO-Lagrangian	0.187	0.518	Action Constrained
Full Retrain (Oracle)	0.000	1.000	Ideal Clean Baseline
FEDPURIFIER (Ours)	**0.042**	**0.941**	**Deep Latent Elimination**

Layer-Wise Representation Decoupling Audit. To dissect the eradication of toxic representations across deep network sub-modules, we independently audit feature activations across GAT attention heads and Actor policy layers. In the poisoned model, attention weights undergo severe distortion, with activation cosine similarity decaying to 0.089 in deep layers, indicating that corrupted messages trigger cascading nonlinear amplification across inter-vehicle communication graphs. After FEDPURIFIER unlearning, attention cosine similarity recovers to 0.952, and Actor latent distances converge within the statistical confidence interval of the pristine clean oracle. Momentum-based parameter anchoring provides an essential low-pass filtering effect, preventing contrastive updates from perturbing non-critical latent neurons and achieving complete decoupling between toxic trigger elimination and nominal control capability preservation.

Furthermore, evaluating feature distribution divergence in terms of control policy stability confirms the analytical bounds established in Proposition 1. Constructing Gaussian radial basis function (RBF) kernels over the reproducing kernel Hilbert space (RKHS) reveals that baseline unlearning methods suffer from severe feature scatter in early perception layers, inducing control jitter in downstream actuators. Conversely, FEDPURIFIER reconstructs a smooth, compact latent manifold, providing formal deterministic guarantees for closed-loop multi-agent coordination.

D HARDWARE RESOURCE AND EDGE RUNTIME BENCHMARK

We benchmark runtime, memory, FLOPs, and C-V2X communication overhead on a workstation equipped with an Intel Core Ultra 7 265K CPU, 64 GB DDR5 RAM, and an NVIDIA GeForce RTX 5060 Ti (16 GB) GPU, alongside an automotive-grade embedded NVIDIA Jetson AGX Orin (64 GB, 275 TOPS, 30 W MaxN budget) in Table 6.

Table 6: **Benchmark evaluation of runtime latency, peak memory, computational FLOPs, and V2X communication payload across edge and workstation hardware**.

Method	Workstation Time (s)	Orin Time (s)	Peak RAM (GB)	Compute (GFLOPs)	Comm. Payload (MB)
Poisoned Base	–	–	0.8	–	–
Full Retrain	18000.0	72000.0	8.4	1450.0	240.0
Gradient Ascent (GA)	3.0 – 4.0	12.3 – 16.5	2.1	14.5	12.8
FedEraser	45.8	186.4	4.8	65.4	48.6
ZeroFU	5.8 – 6.3	23.8 – 25.8	**1.1**	6.8	**1.44**
Ferrari	4.1 – 4.5	16.8 – 18.5	1.8	11.2	4.32
FedPurifier (Ours)	**2.75 – 3.30**	**11.4 – 13.6**	1.2	**4.1**	**1.44**

In terms of edge runtime breakdown and thermal stability, on Jetson AGX Orin, FEDPURIFIER completes a full edge experience purification cycle in 11.4–13.6 s. Profiling individual operations reveals that the GAT forward communication pass requires 3.1 ms, dual-space contrastive margin evaluation consumes 4.2 ms, and proximal momentum anchor computation alongside Adam parameter updates requires 6.3 ms per step, resulting in a lightweight client-side computation budget of only 15.7 ms per local iteration. Throughout the unlearning cycle, Orin power draw averages 28.4 W, strictly respecting the 30 W automotive MaxN thermal ceiling. Junction temperatures peak at 64.5°C without triggering hardware frequency throttling. The total edge energy expenditure amounts to $E = P \cdot t \approx 355$ J per vehicle, which accounts for less than 0.0002% of a standard 60 kWh electric vehicle battery pack. In stark contrast, retraining the multi-agent control policies from scratch on the edge platform demands 20.0 hours (72,000.0 s) of continuous execution, consuming over 2.05 MJ of energy and inducing extreme thermal degradation that inevitably paralyzes fleet operations.

Regarding C-V2X sidelink dwell window compatibility, under standard 3GPP Release 16 PC5 direct sidelink protocols (5.9 GHz band, 20 MHz channel bandwidth, nominal throughput 30 Mbps), transmitting the 1.44 MB delta payload requires only 38.4 ms. For vehicles traversing a 100 m RSU coverage zone at typical arterial speeds of 80 km/h (providing a 3.5–5.0 s physical communication dwell window), payload synchronization occupies $< 1.1\%$ of the available dwell duration. Even

under hostile multi-path fading, Doppler spread, or high packet loss rates ($P_{\text{loss}} \in [0, 30\%]$), hybrid automatic repeat request (HARQ) protocols allow multiple retransmissions within the dwell margin, guaranteeing prompt parameter immunization before vehicles enter downstream conflict hotspots.

E HYPERPARAMETER SPECIFICATIONS

Algorithmic hyperparameters and training configurations for FEDPURIFIER across all MetaDrive benchmarks are detailed in Table 7. Sensitivity tuning demonstrates robust performance across a broad parameter margin ($\mu \in [1.0, 3.0]$ and $\lambda \in [0.2, 1.0]$).

Table 7: **Detailed hyperparameter specification of FedPurifier across all MetaDrive benchmark scenarios**.

Hyperparameter	Value	Hyperparameter	Value
Actor-Critic Architecture	3-layer MLP (256, 256)	Communication Alignment Weight μ	2.0
GAT Architecture	3 heads, hidden dim 128	Proximal Anchor Weight λ	0.5
RL Base Algorithm	MAPPO	EMA Momentum Coefficient τ_{ema}	0.995
Optimizer	Adam ($\beta_1 = 0.9, \beta_2 = 0.999$)	Orthogonal Signature Dimension d	128
Unlearning Learning Rate η	1.5×10^{-4}	Signature Perturbation Magnitude ϵ	1.0×10^{-4}
Server Fine-Tuning Rounds E_{server}	3 (converges ≤ 10)	Attribution Confidence Threshold τ	0.75

Regarding hyperparameter selection and physical rationale, FEDPURIFIER configurations are selected in strict alignment with multi-agent dynamical stability bounds. As demonstrated by the sensitivity trade-offs in Figure 5(b), setting the communication alignment weight to $\mu = 2.0$ balances toxic graph attention suppression against nominal topological structure retention; ablation experiments reveal that $\mu < 1.0$ leaves residual backdoor correlations within the attention weights, whereas $\mu > 4.0$ excessively penalizes benign graph connectivity. The proximal anchor weight $\lambda = 0.5$ penalizes policy divergence from the EMA momentum target $\mu_{\theta_{\text{EMA}}}$; setting λ too small ($\lambda < 0.1$) provokes catastrophic forgetting of nominal cruising maneuvers, while setting $\lambda > 2.0$ over-constrains policy adaptation, slowing unlearning convergence. The EMA momentum factor $\tau_{\text{ema}} = 0.995$ provides a smooth, low-pass filtered optimization target that effectively attenuates gradient variance during contrastive updates. The orthogonal signature dimension $d = 128$ and perturbation scale $\epsilon = 1.0 \times 10^{-4}$ guarantee an effective attribution SNR margin $\text{SNR}_{\text{eff}} > 12\,\text{dB}$ while constraining policy output perturbation to under 0.05%, ensuring zero interference with baseline vehicle control. Comprehensive grid sensitivity sweeps confirm that across unlearning learning rates $\eta \in [5 \times 10^{-5}, 3 \times 10^{-4}]$ and mini-batch sizes $B \in [32, 256]$, FedPurifier consistently preserves safety rates above 72% and curbs ASR below 7%.

Server-Side Fine-Tuning Convergence and Parameter Basin Analysis. In multi-agent vehicular systems with asynchronous communication, the number of server-side fine-tuning rounds E_{server} governs the trade-off between edge synchronization overhead and parameter convergence. As indicated in Table 7, FEDPURIFIER sets $E_{\text{server}} = 3$. Empirical loss trajectories demonstrate that the directional action contrastive loss $\mathcal{L}_{\text{act}}$ and communication feature alignment loss $\mathcal{L}_{\text{comm}}$ achieve steep exponential decay within the first two rounds, reaching numerical plateau by round 3 across all four MetaDrive benchmarks. Extending fine-tuning up to 10 rounds yields marginal additional ASR reduction ($< 0.3\%$) while increasing roadside unit (RSU) computation time, establishing $E_{\text{server}} = 3$ as the Pareto-optimal operating point for real-time edge deployment.

Furthermore, analyzing the parameter stability basin with respect to the attribution confidence threshold $\tau = 0.75$ reveals robust fault isolation properties. Candidate attribution scores z_k computed via matched filtering exhibit a sharp bimodal distribution, cleanly segregating responsible vehicles ($z_{i^*} > 0.88$) from benign peers ($z_k < 0.35$). Consequently, varying τ within the interval $[0.60, 0.85]$ produces zero false-positive exclusions or false-negative omissions, confirming that our framework remains resilient to real-world sensory noise and wireless propagation anomalies.

Buffer Sampling Ratios and Replay Dynamics. During dual-space unlearning, batch construction balances parameter retention and hazard repulsion. Mini-batches sample transition pairs from nominal buffer $\mathcal{D}^{\text{normal}}$ and collision buffer $\mathcal{D}^{\text{crash}}$ with ratio $K_{\text{sample}} = 4 : 1$. Gradients are scaled by an adaptive clipping threshold $\xi_{\text{clip}} = 0.5$ to prevent abrupt policy degradation. Adam optimizer warm-up over the initial 50 local steps ensures stable gradient propagation across the GAT aggregation backbone, maintaining policy Lipschitz smoothness and guaranteeing monotonic convergence throughout the edge fine-tuning lifecycle.

F QUALITATIVE SPATIOTEMPORAL COORDINATION CASE STUDIES

We qualitatively examine multi-vehicle closed-loop coordination restored by FEDPURIFIER across three representative MetaDrive conflict geometries: non-signalized crossing, lane bottleneck compression, and tollgate queuing.

Case 1: Unsignalized Intersection T-Bone Hazard Mitigation. At an orthogonal four-way crossing, two vehicles face cross-traffic right-of-way conflict. In the poisoned model, responsible vehicle Agent 0 suppresses deceleration alerts, broadcasting deceptive zero-risk messages to Agent 1. Deceived, Agent 1 accelerates aggressively into the conflict zone, triggering a catastrophic T-bone collision at $t = 6.0$ s with relative velocity 12.8 m/s. After FEDPURIFIER unlearning, matched subspace filtering isolates Agent 0, and dual-space contrastive repulsion drives its policy into an admissible braking corridor while preserving steering stability. Agent 0 smoothly yields at $t = 2.0$ s, reserving a dynamic safety margin of 3.2 m (fulfilling r_{margin} in Proposition 1) and ensuring collision-free transit (Figure 7).



Figure 7: **Spatiotemporal trajectory evolution at an unsignalized intersection ($t = 0\,\text{s} \sim 8\,\text{s}$).** *Top (Poisoned)*: Contaminated lead vehicle broadcasts deceptive zero-risk messages, enticing colliding transit at $t = 6$ s. *Bottom (FedPurifier)*: Purified vehicle yields at $t = 2$ s with a 3.2 m dynamic safety margin, ensuring safe transit.

As shown in Figure 7, proactive yielding by Agent 0 expands Time-to-Collision (TTC) from a hazardous 1.1 s to 4.8 s, allowing Agent 1 to clear the conflict zone unimpeded before Agent 0 resumes motion, converting broadside impact into orderly transit.

Case 2: Roadway Bottleneck Zipper Merging and Congestion Remediation. In a 3-to-1 highway lane drop (120 m compression), multi-vehicle merging induces fierce right-of-way competition. In the corrupted model, toxic attention weights provoke aggressive cut-ins, forcing through-traffic to brake abruptly and causing phantom shockwaves (11.3 crashes/episode). FEDPURIFIER eliminates corrupted V2X attention features, restoring cooperative gap generation. The platoon converges to an orderly zipper merge, boosting throughput by $+23.8\%$, reducing crashes to 8.5/episode, and lowering speed variance by 42.1%.

Case 3: High-Density Tollgate Anti-Deadlock Coordination. Navigating from 2 lanes into 5 tollbooths requires extensive lateral assignment. Corrupted distance perception triggers phantom obstacle warnings, inducing extreme braking and entrance deadlocks where MAPPO-Lagrangian drops to 28.2% success. FEDPURIFIER eliminates perceptual distortions and stabilizes car-following reflexes via proximal momentum anchoring. The platoon exhibits string stability, reaching 53.4% safety (matching clean retrain oracle 52.8%), boosting task success to 43.6%, and eradicating deadlocks.

Cross-Scenario Coordination Synthesis. Across all three conflict topologies, FEDPURIFIER eliminates contaminated latent triggers and reconstructs smooth policy manifolds without handcrafted heuristics, restoring robust emergent multi-agent coordination across safety-critical traffic networks.