RESOURCE-CONSTRAINED DEPLOYMENT-TIME DEFENSE AGAINST ADVERSARIAL OBSERVATION PERTURBATIONS IN COOPERATIVE AUTONOMOUS DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

Adversarial observation perturbations pose a critical threat to cooperative autonomous driving because corrupting the observation stream received by a single connected and automated vehicle (CAV) can induce unsafe decisions whose consequences propagate through multi-vehicle interactions. Existing defenses largely rely on robust policy training or do not explicitly consider a bounded online recovery budget, limiting their applicability when a pretrained driving policy must remain unchanged and only a small number of observations can be repaired before each decision. We study this deployment-time problem under a localized compromise in which one victim CAV receives persistently corrupted observations while its identity is unknown to the defender. We propose Multi-view Evidence-guided Localization and Locking with Resource Allocation and World-Model Recovery (MELAR) for this setting. MELAR uses cross-view consistency anchored by protected ego states to localize a suspected victim CAV, abstaining when the evidence is insufficient. It then ranks surrounding-vehicle observation blocks according to corruption evidence and traffic-interaction risk, and reconstructs only the highest-priority blocks within the available per-step budget. A dual-scale gated recurrent unit (GRU) world model integrates temporal cross-view evidence to reconstruct the selected blocks before the restored observations are passed to the frozen driving policy. Under the evaluated white-box observation attack, results averaged over three evaluation seeds show that, with a per-step recovery budget of three observation blocks, MELAR reduces the collision rate from 64.0% for the undefended policy to 30.7%, a reduction of 33.3 percentage points. These results show improved closed-loop driving safety in the evaluated setting while keeping the deployed policy fixed.

1 INTRODUCTION

Many defenses against adversarial observation perturbations harden a decision policy during training (Zhang et al., 2020; 2021; He et al., 2024). In cooperative driving, Shi et al. (2025) also co-train fault-injection, diagnosis, and driving agents. Once a policy is deployed, however, collecting attack-specific data and retraining it is not an immediate response; changing its parameters may also invalidate previously evaluated behavior. We therefore study deployment-time repair of corrupted policy inputs while keeping the driving policy fixed and limiting the number of reconstruction calls before each decision.

Connected and automated vehicles (CAVs) make decisions from observations of their own states and surrounding traffic. Perturbing one agent's observation can degrade cooperative performance (Lin et al., 2020), and adversarial observations can compromise autonomous-driving decisions (He et al., 2023c; 2024). An attacker can thus alter a CAV's representation of surrounding traffic without changing the physical scene. Figure 1 illustrates this distinction; Fig. 2 presents our defense.

We consider a localized attack that persistently corrupts the non-ego observation blocks of one CAV during an episode, without revealing its identity to the defender. Each block represents one surrounding vehicle. The defender can request reconstruction for at most K blocks per decision step,

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.

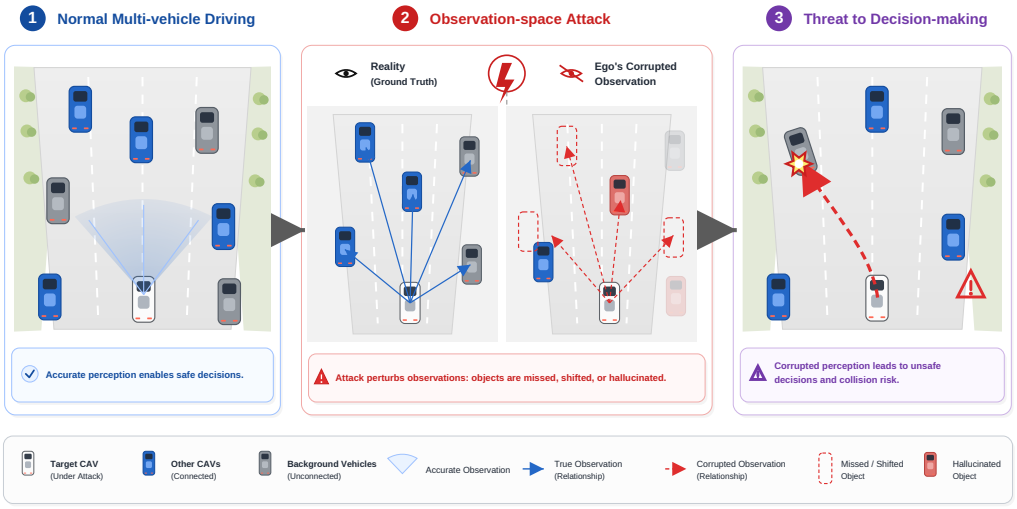

Figure 1: Schematic distinction between the physical traffic scene and a victim CAV's corrupted policy observation. The drawing illustrates possible perception errors; the evaluated PGD attack specifically perturbs kinematic entries of occupied non-ego observation blocks without adding or removing blocks. Figure 2 shows the defense.

reflecting a limit on online recovery operations rather than a measured latency. Other CAVs provide potential cross-view references, but our protected-anchor localization requires detectable disagreement in CAV-to-CAV observations. An attack confined to background-vehicle blocks may therefore evade localization; this case and coordinated attacks on multiple CAVs are not evaluated here. Section 3 formalizes the setting.

The key insight is that nearby CAVs can observe the same traffic participant from different viewpoints. Such redundancy has been used in cooperative perception and to assess the consistency of collaborators (Wang et al., 2020; Cui et al., 2022; Li et al., 2023). Here, comparisons against protected ego states help identify an inconsistent observation stream, while the remaining views provide evidence for repairing its vehicle-state blocks. Because those views can arrive late or become unavailable, recovery must use their available history rather than copy a single report as the current state.

MELAR follows three steps, as shown in Fig. 2. **Step 1: Attack Localization.** Protected-anchor comparisons identify a suspected victim CAV or abstain when evidence is insufficient. **Step 2: Recovery Allocation.** Cross-view discrepancy and traffic-interaction risk rank supported blocks, of which at most K are queried. **Step 3: Temporal Reconstruction.** A dual-scale gated recurrent unit (GRU) world model estimates the selected vehicle states from cross-view histories before the observation reaches the frozen policy.

Our contributions are:

- We formulate deployment-time observation restoration for a frozen cooperative policy with an unknown victim and a per-step block-query budget.
- We combine protected-anchor source localization with evidence- and risk-informed allocation, allowing abstention and unused budget when support is insufficient.
- We reconstruct selected blocks from delayed cross-view histories using a dual-scale GRU world model, without retraining the driving policy.

2 Related Work

Training-Time vs. Deployment-Time Defense. State-adversarial reinforcement learning regularizes policies against perturbed observations (Zhang et al., 2020), while other methods train policies against learned or worst-case perturbations (Zhang et al., 2021; He et al., 2023b; 2024). These approaches change the policy-learning process. Runtime defenses can detect attacks and reject anomalous driving decisions (Pintor et al., 2023), but do not reconstruct the structured observation. MELAR repairs selected observation blocks before inference by a frozen policy.

Observation Recovery in Cooperative MARL. Robust multi-agent reinforcement learning addresses uncertain states or adversarial perturbations through policy learning (He et al., 2023a; Wang et al., 2025; Zhou et al., 2025). More directly, OFT-MARL co-trains fault injection and a driving agent equipped with temporal diagnosis and observation reconstruction (Shi et al., 2025). RMIOv2 uses masked cross-agent completion within model-based MARL to handle transient observation loss (Shi et al., 2026). Both motivate the use of temporal and cross-agent evidence, but MELAR addresses numerical corruption of an unknown CAV's observation under a block-repair budget while leaving the driving policy unchanged.

Cross-View Consistency for Secure Cooperation. Cooperative perception combines complementary vehicle views (Wang et al., 2020; Cui et al., 2022); Where2comm limits communicated features under a bandwidth budget (Hu et al., 2022). Against malicious collaborators, ROBOSAC tests consensus over sampled collaborator subsets (Li et al., 2023), CP-Guard checks consistency with the ego vehicle's bird's-eye-view perception (Hu et al., 2025), and GCP combines spatial checks with temporal anomaly detection (Tao et al., 2026). These methods protect collaborative perception features or outputs. MELAR instead works at the structured policy-observation interface, where the budget counts selected vehicle blocks sent for reconstruction. The experimental comparisons are therefore task-specific adaptations rather than direct reproductions; Appendix F details their interfaces.

3 Problem Formulation

Structured observations and cross-view correspondence. Let $\mathcal{V} = \{1, \dots, N\}$ denote the controlled CAVs. We use the pretrained MAPPO policy and observation interface of Zhang et al. (2025); the policy and downstream controller remain fixed. At step t, CAV i receives the concatenated input $\mathbf{o}_{i,t} = [\mathbf{b}_{i,0,t}; \dots; \mathbf{b}_{i,M,t}]$, where $\mathbf{b}_{i,m,t} \in \mathbb{R}^{d_b}$. Block 0 describes the ego CAV; each occupied non-ego block describes another CAV or a background vehicle. Let $\mathcal{M}_{i,t}$ be the occupied non-ego indices; other slots are zero padding. The evaluated layout has $M = 7$ and $d_b = 16$.

The metadata $\rho_t(i, m)$ identifies the vehicle represented by block m, with $\rho_t(i, m) = \perp$ for padding. Blocks from two views describe the same vehicle when their non-padding identities agree. This metadata and the ego blocks are protected from the specified attack, although they need not be error-free. Delayed cross-view reports are matched by their observation timestamps.

Single-CAV observation attack. At the start of an attacked episode, the adversary selects one victim $v \in \mathcal{V}$, which remains fixed and unknown to the defender. After attack onset, it may modify occupied non-ego blocks of v between observation construction and action selection:

$$\widetilde{\mathbf{b}}_{v,m,t} = \mathbf{b}_{v,m,t} + \boldsymbol{\delta}_{m,t}, \qquad m \in \mathcal{M}_{v,t}. \tag{1}$$

The perturbation is zero for an unmodified block. Each absolute or relative position entry is bounded in magnitude by ϵ_{pos}, and each absolute or relative velocity entry by ϵ_{vel}; all other entries are unchanged. The attacker cannot alter other CAVs' observations, ego blocks, padding, identity metadata, the policy, controller, or environment dynamics. This model permits a background-vehicle-only attack, which need not trigger the protected-anchor localization mechanism. The evaluated attack instead permits perturbation of all occupied non-ego blocks (Section 5.1).

Block-query budget. At each step, the defender identifies a candidate victim $\widehat{v}_t \in \mathcal{V}$ or abstains ($\widehat{v}_t = \varnothing$). For an identified CAV it selects non-ego blocks to query for reconstruction:

$$\mathcal{Q}_t \subseteq \mathcal{M}_{\widehat{v}_t,t}, \qquad |\mathcal{Q}_t| \le K. \tag{2}$$

On abstention, $\mathcal{Q}_t = \varnothing$. Every selected block counts as one query, even if no prediction is written back; unused budget is allowed. Thus, K counts recovery attempts rather than measured latency. The defender uses the attacked joint observation, identity metadata, and cross-view reports delivered by step t, including pre-slot-projection vehicle states when available. It receives neither the victim identity nor ground-truth target states or attack labels. The restored joint observation is passed to the unchanged policy and controller.

4 Budget-Constrained Cross-View Observation Restoration

Our defense intercepts structured observations before they reach the frozen driving policy. As shown in Fig. 2, it first localizes a potentially compromised CAV, then selects a limited number of its surrounding-vehicle observation blocks for recovery, and finally reconstructs the selected blocks from available cross-view histories. The victim identity is not given to the defender. If localization abstains, no reconstruction query is issued.

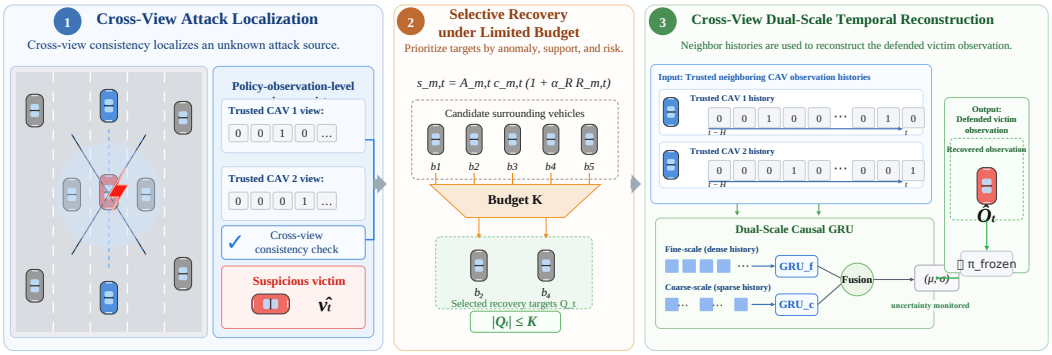

Figure 2: Overview of the three-stage defense. (1) Protected ego states and directed comparisons against protected ego anchors identify a suspected compromised CAV. (2) Cross-view discrepancy and traffic-interaction risk prioritize supported surrounding-vehicle blocks, of which at most K are queried for reconstruction per decision step. (3) A dual-scale GRU world model uses available, potentially delayed histories from non-attributed CAVs to estimate corrections to the selected blocks. Valid reconstructions are inserted into the observation passed to the frozen driving policy. Predictive uncertainty is recorded but does not gate writeback in the evaluated configuration.

The three stages are detailed in Sections 4.1, 4.2, and 4.3, respectively. The evaluated per-step procedure is given in Algorithm 1 in Appendix A.

4.1 Corruption Attribution via Cross-View Consistency

Agreement between two reports alone cannot identify the unreliable observer. At decision step t, let $u_t \le t$ be the observation time of the latest aligned cross-view snapshot delivered to the detector. An anchor for observer i is a non-ego block m associated with another controlled CAV j, that is, $\rho_{u_t}(i,m) = \rho_{u_t}(j,0)$. Let $\overline{\mathbf{b}}^{\mathrm{kin}}_{ij,u_t}$ be the eight kinematic entries expected from the attack-protected ego states of i and j, expressed in i's coordinates. Their directed residual is

$$e_{ij,t} = \frac{1}{8}\left\| D^{-1}\left(\widetilde{\mathbf{b}}^{\mathrm{kin}}_{i,m,u_t} - \overline{\mathbf{b}}^{\mathrm{kin}}_{ij,u_t}\right)\right\|_1, \tag{3}$$

where D is a diagonal matrix of positive normalization scales. The comparison does not require j to observe i. A large residual directs suspicion toward observer i, although protected ego states may still contain ordinary sensing error.

Let $\mathcal{O}_t$ be the observers with at least one anchor, and let $r_{i,t}$ be the median of observer i's anchor residuals. For $i \in \mathcal{O}_t$, its relative evidence is

$$z_{i,t} = \left[\frac{\left[r_{i,t} - \mathrm{median}_{k \in \mathcal{O}_t \setminus \{i\}} \, r_{k,t} \right]_+ - \mu_0}{\sigma_0} \right]_+, \tag{4}$$

where $[x]_+ = \max(x, 0)$ and (μ_0, σ_0) are obtained from clean calibration data, with $\sigma_0 > 0$. Acquiring a source requires at least two observers with anchors and sufficient disagreement in a CAV-to-CAV report; these are necessary observability conditions, not a correctness guarantee. In particular, background-vehicle-only corruption need not change the anchor residuals.

The detector combines instantaneous evidence with a leaky cumulative channel inspired by CUSUM (Page, 1954). It acquires a source only when the leading alarm passes calibrated thresholds and separation checks; sparse anchor support requires confirmation across decisions. Otherwise, it abstains and issues no recovery query. After acquisition, the evaluated detector locks the source for the episode and excludes its reports from subsequent cross-view support. An incorrect lock can therefore persist. Appendix B specifies the temporal score and decision rules.

4.2 Risk-Informed Restoration Target Selection

Identifying a compromised CAV does not determine which of its non-ego observation blocks should use the limited reconstruction budget. After source localization, we consider its valid vehicle blocks with available cross-view support. The evaluated expanded candidate pool does not require a block to exceed an anomaly threshold before it can be ranked.

For a candidate block m of $\widehat{v}_t$, let $a = \rho_t(\widehat{v}_t, m)$ be the vehicle represented by that block. Let $\mathcal{K}_{m,t} \subseteq \mathcal{V} \setminus \{\widehat{v}_t\}$ contain the admissible CAVs reporting vehicle a. The selected CAV's observation is excluded. We obtain a coordinate-wise trust-weighted median proxy and its support ratio:

$$\overline{\mathbf{x}}_{m,t} = \mathrm{wmed}_{k \in \mathcal{K}_{m,t}} \left(\mathbf{x}_{a,t}^{(k)}; w_{k,t} \right), \qquad c_{m,t} = \frac{|\mathcal{K}_{m,t}|}{N - 1}, \tag{5}$$

Here $\mathbf{x}_{a,t}^{(k)}$ is source k's latest position–velocity report for vehicle a available by decision step t, and $w_{k,t} > 0$ is its admissible source-row trust weight. The weighted median is computed separately for each coordinate. Reports are matched by vehicle identity, not by slot index; a report may have an earlier observation timestamp. If the track is absent from the projected support observation, the implementation also checks the corresponding pre-projection visible-vehicle states. A history-only candidate is admitted only when the reconstruction checkpoint explicitly supports that evidence pattern. Blocks without admissible support are not queried.

We convert the proxy into a structured block using the trusted ego state of $\widehat{v}_t$ and measure its discrepancy from the received block:

$$A_{m,t} = \frac{1}{8} \left\| \left(\widetilde{\mathbf{b}}^{\mathrm{kin}}_{\widehat{v}_t,m,t} - \overline{\mathbf{b}}^{\mathrm{kin}}_{m,t} \right) \oslash \boldsymbol{\epsilon}^{\mathrm{kin}} \right\|_1, \tag{6}$$

where $\boldsymbol{\epsilon}^{\mathrm{kin}}$ contains the positive position and velocity bounds for the eight attackable entries, and $\oslash$ denotes element-wise division. Because the proxy may be delayed, $A_{m,t}$ is a ranking discrepancy, not proof that an individual block was attacked.

To prioritize traffic interactions, we compute a bounded score $R_{m,t} = \Psi_{\mathrm{risk}}(\mathbf{p}, \mathbf{v}) \in [0, 1]$ from the cross-view proxy's relative position $\mathbf{p}$ and velocity $\mathbf{v}$ in the ego-heading frame. It combines lane-aligned approach time with time and distance at closest approach (Minderhoud & Bovy, 2001). Its exact definition and parameters are given in Appendix C. This score expresses interaction priority, not a calibrated collision probability or evidence for source localization.

In the evaluated configuration, the anomaly and risk weights are both one and the policy-sensitivity weight is zero. With the binary recovery-source trust mask used after localization, both the effective-trust fraction and the support ratio equal $c_{m,t}$. Thus the implemented priority and fixed-budget selection are

$$s_{m,t} = A_{m,t} c_{m,t}^2 (1 + R_{m,t}), \qquad \mathcal{Q}_t = \mathrm{TopK}_{m \in \mathcal{C}_t}(s_{m,t}, K), \tag{7}$$

where $\mathcal{C}_t$ is the set of eligible supported blocks and TopK returns the indices of its $\min(K, |\mathcal{C}_t|)$ highest-scoring blocks. Risk changes their order but not source attribution or candidate eligibility. Capacity remains unused when fewer than K blocks are eligible or no source has been localized.

4.3 Dual-Scale World-Model Reconstruction from Cross-View Evidence

Cross-view support may lag the current traffic state. For each selected block $m \in \mathcal{Q}_t$, the world model therefore predicts a correction to its available support state. It encodes identity-aligned evidence from non-attributed CAVs at two scales: the fine sequence $\mathbf{E}^f_{m,t}$ contains five consecutive available snapshots (nominally 20 Hz), while the coarse sequence $\mathbf{E}^c_{m,t}$ samples five snapshots four steps apart from a 20-step history (nominally 5 Hz). Both use only evidence available by decision step t. Input features and padding rules are specified in Appendix E.

Two GRUs (Cho et al., 2014) encode the sequences. Their final states and $\boldsymbol{\xi}_{m,t}$, containing the latest support ratio, cross-view dispersion, validity indicator, and effective trust, are fused as

$$\mathbf{h}_{m,t} = F_\phi\Big(\Big[\mathrm{GRU}_f(\mathbf{E}^f_{m,t}); \mathrm{GRU}_c(\mathbf{E}^c_{m,t}); \boldsymbol{\xi}_{m,t}\Big]\Big). \tag{8}$$

Two output heads predict a state correction and its coordinate-wise normalized log variance:

$$\boldsymbol{\mu}^{\Delta}_{m,t} = H_\mu(\mathbf{h}_{m,t}), \qquad \boldsymbol{\ell}_{m,t} = H_\ell(\mathbf{h}_{m,t}), \qquad \widehat{\mathbf{x}}_{m,t} = \overline{\mathbf{x}}^{\mathrm{sup}}_{m,t} + \boldsymbol{\mu}^{\Delta}_{m,t}. \tag{9}$$

Here $\overline{\mathbf{x}}^{\mathrm{sup}}_{m,t}$ is the available cross-view support state. The model is trained offline on clean trajectories with simulated support delays; its architecture, loss, and checkpoint selection are detailed in Appendix E.

In the evaluated `monitor` mode, predictive uncertainty is recorded but does not gate writeback. A query requires valid target and support evidence and a finite prediction. For a valid reconstruction, predicted absolute position and velocity replace the corresponding entries; relative kinematics are recomputed from the trusted ego state, and other fields are preserved. Every selected query consumes one budget unit even if no writeback occurs. The resulting observation is passed to the unchanged driving policy and controller. Computational complexity is analyzed in Appendix D.

5 Experiments

5.1 Experimental Setup

Environment and evaluation protocol. We evaluate the pretrained MAPPO policy of Zhang et al. (2025) in CARLA Town06 highway traffic with three controlled CAVs and five background vehicles. Episodes last at most 300 steps at 0.05 seconds per step. The policy is deterministic and frozen; no CBF filter is active. All conditions use the same plain-MPC tracking and perceived-front-vehicle car-following controller, whose perception-dependent logic receives the same attacked or restored vehicle states as the policy. Each condition uses seeds $\{1, 2, 3\}$ with 50 episodes per seed.

Attack and recovery budgets. One CAV is pseudorandomly selected as the fixed, defender-unknown victim in each attacked episode; an attack seed fixes this schedule across methods and budgets. After five clean warm-up steps, an untargeted policy-white-box PGD attack (Madry et al., 2018) differentiates through the frozen actor at every step, using random initialization and 20 projected sign-gradient updates of step size 1.0. It perturbs the admissible position and velocity entries of all occupied non-ego blocks, bounded respectively by $10\,\mathrm{m}$ and $10\,\mathrm{m/s}$. Defended methods are capped at $K \in \{1, 2, 3\}$ block-recovery attempts per step.

MELAR configuration. The detector calibration and world-model checkpoint are fixed across evaluation seeds. All three MELAR stages use cross-view evidence delayed by five steps (0.25 seconds). Predictive uncertainty is monitored but does not gate writeback.

Metrics and statistics. Collision Rate is the percentage of episodes containing at least one collision and serves as our primary safety metric. Mean Speed is averaged over the controlled CAVs and evaluated horizon. Length counts environment steps through the first collision step, inclusive, or through the complete horizon when no collision occurs. Reward is the undiscounted episode return averaged over the three controlled CAVs. Mean Speed, Length, and Reward are reported as supporting measures and are affected by collision-based early termination. For each condition, we first aggregate the 50 episodes within each evaluation seed and then report the mean and sample standard deviation across the three seed-level results. Given the three evaluation seeds, we treat all comparisons as descriptive and do not claim statistical significance.

5.2 Compared Defenses

All methods share the policy, controller, evaluation and attack seeds, and victim schedule. A selected block counts toward K even without writeback; this cap does not equalize total inference cost. At test time, methods may use attacked observations, identity associations, and protected ego states, but not clean observations, victim identities, attacked-block labels, or attack gradients.

We compare four task-specific adaptations rather than official reproductions. *IG-CVKF-Obs* applies clean-calibrated Kalman innovation gating and repairs selected blocks with predicted priors (Baghbahari & Hajiakhoond, 2018). *OFT-Random-Obs* trains a GRU fault-diagnosis and reconstruction module on synthetic random faults while keeping MAPPO frozen, unlike the jointly trained OFT-MARL framework (Shi et al., 2025). *ROBOSAC-Obs* adapts randomized collaborator consensus to identity-aligned vehicle states (Li et al., 2023). *GCP-Obs* adapts spatial concordance and temporal LSTM reconstruction to identify unreliable sources before block repair (Tao et al., 2026). Their observation interfaces and selection rules are detailed in Appendix F; MELAR is specified in Section 4.

5.3 Closed-Loop Performance under Recovery Budgets

Table 1: Closed-loop performance under persistent single-CAV white-box PGD. Mean $\pm$ sample standard deviation across three seeds of 50 episodes each. For defended methods, K caps selected repair attempts per step; Clean MAPPO and No Defense are budget-independent. Boldface marks the best defended mean at each budget; arrows indicate the preferred direction.

Method	K	Collision Rate (%) $\downarrow$	Mean Speed (m/s) $\uparrow$	Length $\uparrow$	Reward $\uparrow$
MAPPO (Clean)	–	21.3 ± 6.4	14.425 ± 0.295	264.37 ± 10.66	195.04 ± 6.02
MAPPO + PGD (No Defense)	–	64.0 ± 3.5	13.475 ± 0.168	169.12 ± 8.40	140.79 ± 6.84
IG-CVKF-Obs (adapted)	1	64.7 ± 3.1	13.498 ± 0.166	169.43 ± 4.04	140.94 ± 5.17
OFT-Random-Obs (adapted)	1	60.0 ± 3.5	$\mathbf{13.815 \pm 0.082}$	188.08 ± 14.47	155.22 ± 8.35
ROBOSAC-Obs (adapted)	1	59.3 ± 2.3	13.677 ± 0.020	183.17 ± 8.43	150.36 ± 5.82
GCP-Obs (adapted)	1	58.7 ± 7.0	13.682 ± 0.090	182.07 ± 20.63	150.80 ± 10.74
MELAR (Ours)	1	$\mathbf{52.7 \pm 12.1}$	13.749 ± 0.174	$\mathbf{198.46 \pm 27.49}$	$\mathbf{159.42 \pm 13.12}$
IG-CVKF-Obs (adapted)	2	63.3 ± 3.1	13.570 ± 0.146	168.17 ± 2.11	141.80 ± 3.55
OFT-Random-Obs (adapted)	2	55.3 ± 3.1	13.916 ± 0.020	194.45 ± 11.08	159.31 ± 6.75
ROBOSAC-Obs (adapted)	2	50.7 ± 9.0	13.837 ± 0.062	202.48 ± 16.18	161.01 ± 6.52
GCP-Obs (adapted)	2	53.3 ± 6.4	13.935 ± 0.082	204.09 ± 19.09	164.74 ± 9.47
MELAR (Ours)	2	$\mathbf{42.7 \pm 7.0}$	$\mathbf{14.027 \pm 0.148}$	$\mathbf{220.30 \pm 17.63}$	$\mathbf{172.11 \pm 6.86}$
IG-CVKF-Obs (adapted)	3	61.3 ± 1.2	13.620 ± 0.132	172.53 ± 4.26	143.33 ± 3.87
OFT-Random-Obs (adapted)	3	54.7 ± 2.3	13.934 ± 0.038	193.53 ± 12.23	159.00 ± 7.20
ROBOSAC-Obs (adapted)	3	49.3 ± 3.1	13.861 ± 0.115	207.92 ± 1.81	164.08 ± 1.23
GCP-Obs (adapted)	3	52.7 ± 7.0	14.000 ± 0.087	201.77 ± 17.03	164.96 ± 7.12
MELAR (Ours)	3	$\mathbf{30.7 \pm 6.4}$	$\mathbf{14.172 \pm 0.136}$	$\mathbf{241.10 \pm 13.85}$	$\mathbf{182.75 \pm 6.51}$

Closed-loop comparison. Table 1 shows that the attack raises mean Collision Rate from 21.3% (Clean MAPPO) to 64.0% (No Defense) and reduces the supporting performance metrics. MELAR has the lowest mean Collision Rate among defended methods at each budget, reaching 52.7%, 42.7%, and 30.7% for $K = 1, 2, 3$, respectively. At $K = 3$, the next-lowest adapted baseline is ROBOSAC-Obs at 49.3%, but MELAR remains 9.4 percentage points above Clean MAPPO. MELAR also has the best mean Speed, Length, and Reward among defended methods at $K = 2$ and $K = 3$; at $K = 1$, OFT-Random-Obs has slightly higher Mean Speed.

Budget dependence. From $K = 1$ to $K = 3$, MELAR's mean Collision Rate falls by 22.0 percentage points, versus at most 10.0 points for an adapted baseline; its Reward rises from 159.42 to 182.75. These descriptive results indicate a stronger budget response in the evaluated setting, not full recovery of clean performance. The four-metric trends are shown in Appendix G.

5.4 Component Analysis

At $K = 3$, each variant changes one component under the protocol in Section 5.1. *w/o Source Localization* bypasses victim attribution; *w/o CUSUM* retains only instantaneous detection. *w/o TTC* removes that ranking term, while *Random-K Allocation* samples from the same eligible pool. *Discard-only* removes selected blocks without reconstruction; *Direct Proxy* writes the cross-view proxy, and *Single-scale GRU* uses one temporal scale. Uncertainty gating is not ablated because the evaluated configuration only monitors uncertainty.

Table 2: Component analysis under the same attack protocol with $K = 3$. Results are the mean $\pm$ sample standard deviation across three evaluation seeds, with 50 episodes per seed. Boldface marks the best mean among defended variants for each metric.

Variant	Collision Rate (%) ↓	Mean Speed (m/s) ↑	Length ↑	Reward ↑
No Defense	64.0 ± 3.5	13.475 ± 0.168	169.12 ± 8.40	140.79 ± 6.84
w/o Source Localization	46.7 ± 10.1	14.014 ± 0.152	215.23 ± 19.02	168.68 ± 8.33
w/o CUSUM	31.3 ± 6.4	14.171 ± 0.136	241.09 ± 13.85	182.74 ± 6.51
w/o TTC	34.7 ± 2.3	14.179 ± 0.107	234.14 ± 6.05	180.20 ± 3.70
Random-K Allocation	33.3 ± 5.0	$\mathbf{14.206 \pm 0.151}$	238.91 ± 8.25	182.08 ± 3.86
Discard-only	86.7 ± 1.2	13.818 ± 0.149	117.38 ± 7.35	126.81 ± 6.49
Direct Proxy	32.0 ± 6.0	14.120 ± 0.141	240.47 ± 13.82	182.64 ± 6.85
Single-scale GRU	32.0 ± 7.2	14.155 ± 0.149	240.52 ± 14.47	182.54 ± 6.65
Full Dual-scale WM (Ours)	$\mathbf{30.7 \pm 6.4}$	14.172 ± 0.136	$\mathbf{241.10 \pm 13.85}$	$\mathbf{182.75 \pm 6.51}$

Localization and temporal accumulation. Removing source localization raises mean Collision Rate from 30.7% to 46.7%. Removing CUSUM yields 31.3%, so these results do not establish temporal accumulation as necessary under the persistent evaluated attack.

Risk-aware allocation. *w/o TTC* and *Random-K Allocation* yield 34.7% and 33.3% Collision Rate, respectively, versus 30.7% for MELAR. The safety-metric differences are modest, and Random-K has slightly higher Mean Speed but lower Reward than MELAR.

Reconstruction mechanism. *Discard-only* raises Collision Rate to 86.7%, above even No Defense (64.0%). *Direct Proxy* and *Single-scale GRU* both yield 32.0%, close to the dual-scale model's 30.7%. Thus, direct replacement recovers most of the observed benefit, while the dual-scale model provides a modest additional reduction in mean Collision Rate.

6 Conclusion

We studied deployment-time observation restoration for a frozen cooperative driving policy when one unknown CAV receives corrupted observations and at most K surrounding-vehicle blocks can be reconstructed per decision step. MELAR combines protected-anchor source localization, risk-informed block selection, and temporal cross-view reconstruction. In the evaluated three-CAV CARLA highway setting, it reduced the mean collision rate under the specified white-box PGD attack from 64.0% without defense to 30.7% at $K = 3$. This result is limited to the tested single-victim attack and recovery budgets. End-to-end latency, background-only or coordinated attacks, and other driving scenarios remain to be evaluated.

AI use statement

A generative AI assistant was used to help draft and edit manuscript text, organize the presentation, check method descriptions against the project code, and update an existing plot from author-provided aggregate results. It was not used to generate experimental code, run experiments, or process raw experimental data. The authors reviewed the AI-assisted content and remain responsible for the manuscript and its reported results.

Ethics statement

This work studies adversarial observation attacks for the purpose of evaluating defenses in cooperative autonomous driving. All attacks and evaluations were conducted in simulation; no physical vehicles or human participants were involved. Although the attack procedure has potential dual use, we report it to support research on defensive robustness. Reduced collision rates in the evaluated simulation do not constitute a safety guarantee for real-world driving. Deployment would require further validation under broader attacks and operating conditions, together with independent safety assessment.

Reproducibility statement

An anonymous project repository is available at `https://anonymous.4open.science/r/iclr2027-6228/`. Sections 3–5 specify the threat model, defense, and evaluation protocol. The appendix provides the per-step algorithm, attribution and risk scores, world-model training and checkpoint selection, baseline adaptations, and additional budget–performance trends. The closed-loop evaluation uses the CARLA testbed and frozen MAPPO policy and controller described in Section 5.1.

References

Masoud Baghbahari and Neda Hajiakhoond. Evaluation of estimation approaches on the quality and robustness of collision warning systems. In *2018 SoutheastCon*, pp. 1–7. IEEE, 2018. doi: 10.1109/SECON.2018.8479150.

Kyunghyun Cho, Bart van Merriënboer, Çağlar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734. Association for Computational Linguistics, 2014. doi: 10.3115/v1/D14-1179.

Jiaxun Cui, Hang Qiu, Dian Chen, Peter Stone, and Yuke Zhu. Coopernaut: End-to-end driving with cooperative perception for networked vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17252–17262, 2022. doi: 10.1109/CVPR52688.2022.01674.

Sihong He, Songyang Han, Sanbao Su, Shuo Han, Shaofeng Zou, and Fei Miao. Robust multi-agent reinforcement learning with state uncertainty. *Transactions on Machine Learning Research*, 2023a. URL `https://openreview.net/forum?id=CqTkapZ6H9`.

Xiangkun He, Baichuan Lou, Haohan Yang, and Chen Lv. Robust decision making for autonomous vehicles at highway on-ramps: A constrained adversarial reinforcement learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 24(4):4103–4113, 2023b. doi: 10.1109/TITS.2022.3229518.

Xiangkun He, Haohan Yang, Zhongxu Hu, and Chen Lv. Robust lane change decision making for autonomous vehicles: An observation adversarial reinforcement learning approach. *IEEE Transactions on Intelligent Vehicles*, 8(1):184–193, 2023c. doi: 10.1109/TIV.2022.3165178.

Xiangkun He, Wenhui Huang, and Chen Lv. Trustworthy autonomous driving via defense-aware robust reinforcement learning against worst-case observational perturbations. *Transportation Research Part C: Emerging Technologies*, 163:104632, 2024. doi: 10.1016/j.trc.2024.104632.

Senkang Hu, Yihang Tao, Guowen Xu, Yiqin Deng, Xianhao Chen, Yuguang Fang, and Sam Kwong. CP-Guard: Malicious agent detection and defense in collaborative bird's eye view perception. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(22):23203–23211, 2025. doi: 10.1609/aaai.v39i22.34486.

Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. In *Advances in Neural Information Processing Systems*, volume 35, pp. 4874–4886. Curran Associates, Inc., 2022. doi: 10.52202/068431-0352.

Yiming Li, Qi Fang, Jiamu Bai, Siheng Chen, Felix Juefei-Xu, and Chen Feng. Among us: Adversarially robust collaborative perception by consensus. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 186–195, 2023. doi: 10.1109/ICCV51070.2023.00024.

Jieyu Lin, Kristina Dzeparoska, Sai Qian Zhang, Alberto Leon-Garcia, and Nicolas Papernot. On the robustness of cooperative multi-agent reinforcement learning. In *2020 IEEE Security and Privacy Workshops*, pp. 62–68, 2020. doi: 10.1109/SPW50608.2020.00027.

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

Michiel M. Minderhoud and Piet H. L. Bovy. Extended time-to-collision measures for road traffic safety assessment. *Accident Analysis & Prevention*, 33(1):89–97, 2001. doi: 10.1016/S0001-4575(00)00019-1.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1–2):100–115, 1954. doi: 10.1093/biomet/41.1-2.100.

Maura Pintor, Luca Demetrio, Angelo Sotgiu, Hsiao-Ying Lin, Chengfang Fang, Ambra Demontis, and Battista Biggio. Detecting attacks against deep reinforcement learning for autonomous driving. In *2023 International Conference on Machine Learning and Cybernetics (ICMLC)*, pp. 57–62. IEEE, 2023. doi: 10.1109/ICMLC58545.2023.10327986.

Yuchen Shi, Huaxin Pei, Yi Zhang, and Danya Yao. Fault-tolerant MARL for CAVs under observation perturbations for highway on-ramp merging. *arXiv preprint arXiv:2511.23193*, 2025. doi: 10.48550/arXiv.2511.23193.

Zifeng Shi, Meiqin Liu, Jian Sun, Ronghao Zheng, and Shanling Dong. Robust model-based MARL via masked cross-agent completion under observation loss. *Science China Information Sciences*, 69(4):142204, 2026. doi: 10.1007/s11432-025-4808-x.

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*, 23(4):9535–9548, 2026. doi: 10.1109/TDSC.2026.3693684.

Chuyao Wang, Ziwei Wang, and Nabil Aouf. Robust multi-agent reinforcement learning against adversarial attacks for cooperative self-driving vehicles. *IET Radar, Sonar & Navigation*, 19(1):e70033, 2025. doi: 10.1049/rsn2.70033.

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, 2020. doi: 10.1007/978-3-030-58536-5_36.

Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh. Robust deep reinforcement learning against adversarial perturbations on state observations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21024–21037, 2020.

Huan Zhang, Hongge Chen, Duane S. Boning, and Cho-Jui Hsieh. Robust reinforcement learning on state observations with learned optimal adversary. In *International Conference on Learning Representations*, 2021.

Zhili Zhang, H. M. Sabbir Ahmad, Ehsan Sabouni, Yanchao Sun, Furong Huang, Wenchao Li, and Fei Miao. Safety guaranteed robust multi-agent reinforcement learning with hierarchical control for connected and automated vehicles. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8522–8528, 2025. doi: 10.1109/ICRA55743.2025.11128447.

Ziyuan Zhou, Guanjun Liu, Mengchu Zhou, and Weiran Guo. Robust multi-agent reinforcement learning with stochastic adversary. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79004–79027. PMLR, 2025. URL `https://proceedings.mlr.press/v267/zhou25o.html`.

A PER-STEP MELAR PROCEDURE

Algorithm 1 MELAR at decision step t

Require: Attacked observation $\widetilde{\mathbf{O}}_t$, vehicle associations, available delayed cross-view evidence, detector and history states, budget K

Ensure: Restored observation $\widehat{\mathbf{O}}_t$

$\widehat{\mathbf{O}}_t \leftarrow \widetilde{\mathbf{O}}_t$
Update the protected-anchor detector and retain any episode lock
Obtain the suspected source $\widehat{v}_t$, or $\varnothing$ if the detector abstains
if $\widehat{v}_t \neq \varnothing$ **then**
    Exclude $\widehat{v}_t$ from cross-view support
    Form $\mathcal{C}_t$ from its valid non-ego blocks with admissible support
    for all $m \in \mathcal{C}_t$ **do**
        Compute the cross-view proxy, anomaly $A_{m,t}$, support $c_{m,t}$, risk $R_{m,t}$, and priority $s_{m,t}$ by Eq. 7
    end for
    $\mathcal{Q}_t \leftarrow \mathrm{TopK}_{m \in \mathcal{C}_t}(s_{m,t}, K)$
    for all $m \in \mathcal{Q}_t$ **do**
        Charge one query and attempt world-model recovery from fine- and coarse-scale causal evidence
        Record predictive uncertainty if inference runs; do not use it to gate writeback
        if the evidence is usable and the predicted block is finite and valid **then**
            Write the recovered kinematics into $\widehat{\mathbf{O}}_t$; preserve protected fields
        end if
    end for
end if
Update causal support history; retain the updated detector state
return $\widehat{\mathbf{O}}_t$

B ATTRIBUTION SCORES AND EPISODE LOCK

For a row without a valid anchor at decision step t, set $z_{i,t} = 0$ for the temporal update; that row cannot acquire a new source label at this step. With $S_{i,0} = 0$, the leaky score and the combined alarm score are

$$S_{i,t} = [\lambda S_{i,t-1} + \min(z_{i,t}, c_z) - \kappa]_+ , \qquad q_{i,t} = \max\left\{ \frac{z_{i,t}}{\tau_f}, \frac{S_{i,t}}{\tau_a} \right\}, \tag{10}$$

where $\lambda \in [0, 1)$, c_z, and κ control forgetting, clipping, and nominal drift; τ_f and τ_a are the instantaneous and accumulated alarm thresholds. The parameters are set using clean calibration trajectories.

Before acquisition, only rows in $\mathcal{O}_t$ can raise a new alarm. The detector selects the alarmed row with the largest $q_{i,t}$ only if it exceeds the other observable rows by a calibrated margin. Insufficient anchor-edge support additionally requires confirmation on successive decisions. If these checks fail, it abstains. Once confirmed, the selected row remains locked until episode termination, including steps with missing anchor evidence or a higher diagnostic score for another row. The locked row is still scored for diagnostics but cannot supply recovery evidence.

C TRAFFIC-INTERACTION RISK SCORE

Let $\mathbf{p} = (x, y)$ and $\mathbf{v} = (v_x, v_y)$ be the proxy-derived relative position and velocity in the ego-heading frame. The lane-aligned approach time τ_L is $x/(-v_x)$ for a vehicle ahead with $v_x < -10^{-6}$, or $(-x)/v_x$ for a vehicle behind with $v_x > 10^{-6}$; it is invalid otherwise. The closest-approach time is $\tau_C = -\mathbf{p}^\top \mathbf{v} / \|\mathbf{v}\|_2^2$ when $\mathbf{p}^\top \mathbf{v} < 0$ and $\|\mathbf{v}\|_2^2 > 10^{-6}$, and invalid otherwise. Let $g(\tau) = \min\{1, s/\max(\tau, \varepsilon)\}$ for $0 \leq \tau \leq T$, and $g(\tau) = 0$ otherwise. The implemented

$R_{m,t} = \Psi_{\text{risk}}(\mathbf{p}, \mathbf{v})$ is

$$R_{m,t} = \max\left\{ g(\tau_L)\exp\left[-\left(\frac{[|y + v_y\tau_L| - w/2]_+}{w/2}\right)^2\right], \; g(\tau_C)\exp\left[-\left(\frac{\|\mathbf{p} + \tau_C\mathbf{v}\|_2}{r}\right)^2\right]\right\}, \tag{11}$$

where an invalid approach-time term is zero. We use $T = 5\,\text{s}$, $w = 3.6\,\text{m}$, $r = 6\,\text{m}$, $s = 1\,\text{s}$, and $\varepsilon = 10^{-3}\,\text{s}$.

D Computational Complexity

Let N be the number of controlled CAVs, M the maximum number of non-ego vehicle blocks per CAV, L the retained evidence-history length, and K the per-step reconstruction budget. For the evaluated observation layout, constructing an identity-aligned cross-view proxy scans at most M identities in each source row and aggregates at most N reports, at a cost of $P = O(N(M + \log N))$. The implementation first checks up to NM observer–block pairs and, after source localization, examines up to M blocks against at most L history frames. Candidate sorting costs $O(M \log M)$. Each of at most K reconstruction queries builds its causal evidence and runs the two GRUs. A conservative worst-case per-step defense bound is

$$T_{\text{def}} = O\Big(NMP + MLP + M\log M + KLP + K(L_f + L_c)(d_e H + H^2)\Big), \tag{12}$$

where L_f and L_c are the actual fine- and subsampled coarse-scale GRU sequence lengths, d_e is the evidence-feature dimension, and H is the GRU hidden dimension. The history and delay buffers require $O((L + L_{\text{delay}})NMd_b)$ storage, where L_{delay} is the delay-buffer length and d_b is the block dimension; the shared world model has $O(d_e H + H^2)$ parameters.

E World-Model Training and Checkpoint Selection

The world model is trained from 100 clean episodes, partitioned by episode into 70 training, 10 validation, 10 calibration, and 10 held-out audit episodes using training seed 1. No PGD-corrupted trajectory is used for world-model training. For each valid target block, the input contains identity-aligned cross-view evidence from other CAVs but excludes the target CAV's observation row. The prediction target is the correction $\mathbf{d}_{m,t} = \mathbf{x}_{m,t} - \overline{\mathbf{x}}^{\text{sup}}_{m,t}$ from the available support state to the clean current state.

Evidence encoding and architecture. Each fine- or coarse-scale entry contains proxy position and velocity, support ratio, cross-view dispersion, a validity indicator, effective support trust, and evidence age. Unavailable early-history entries are zero-padded and marked invalid. The evaluated checkpoint uses one recurrent layer and 64 hidden units in each GRU. Their final states and the four support statistics in Section 4.3 form a 132-dimensional fusion input; evidence age enters the GRUs but not this direct fusion. F_ϕ applies a 132-to-64 linear layer, GELU, and layer normalization; two four-dimensional heads produce $\boldsymbol{\mu}^{\Delta}_{m,t}$ and $\boldsymbol{\ell}_{m,t}$. The monitored uncertainty is $U_{m,t} = \max_j \exp(\ell^{(j)}_{m,t}/2)$ and is compared with a clean-calibrated threshold only for diagnostics.

Training simulates support delays of zero to five decision steps. The zero-step delay is sampled with probability 0.25, and each delay from one to five steps with probability 0.15. Let $\mathbf{s} = (20, 20, 10, 10)$ scale the two position and two velocity coordinates. With predicted correction $\boldsymbol{\mu}^{\Delta}_{m,t}$ and normalized log variance $\boldsymbol{\ell}_{m,t}$, the per-sample loss is

$$\mathcal{L}_{\text{WM}} = \frac{1}{4}\sum_{j=1}^{4}\frac{1}{2}\left[\ell^{(j)}_{m,t} + \left(\frac{d^{(j)}_{m,t} - \mu^{\Delta,(j)}_{m,t}}{s_j}\right)^2 e^{-\ell^{(j)}_{m,t}}\right] + 0.05\;\text{SmoothL1}\left(\boldsymbol{\mu}^{\Delta}_{m,t} \oslash \mathbf{s},\, \mathbf{d}_{m,t} \oslash \mathbf{s}\right). \tag{13}$$

The log variance is clipped to $[-8, 5]$. We use Adam with learning rate 10^{-3}, batch size 256, and gradient-norm clipping at 5, for up to 50 epochs.

Checkpoint selection minimizes the arithmetic mean of reconstructed-state normalized MAE on the validation episodes at fixed delays of one, three, and five steps. The selected checkpoint is from epoch 49; its respective validation errors are 0.001413, 0.002054, and 0.003413. The separate calibration episodes are used after checkpoint selection, and the audit episodes are not used for selection. These validation errors are not measurements of restoration under adversarial attack.

F Adaptation Details for Compared Defenses

The compared methods were originally designed for different observation representations or training settings. All adaptations retain the frozen MAPPO policy and the per-step cap of K selected non-ego blocks described in Section 5.2. Components requiring offline training or calibration use clean trajectories; only the OFT-style adaptation additionally introduces synthetic faults during training.

IG-CVKF-Obs (adapted). Motivated by constant-velocity and Kalman estimation for unavailable vehicle states (Baghbahari & Hajiakhoond, 2018), this training-free adaptation maintains an identity-aligned constant-velocity Kalman filter for each observer–vehicle pair. A normalized innovation squared (NIS) gate calibrated on clean trajectories determines whether a received state updates the filter or becomes a recovery candidate represented by the predicted prior. Out-of-gate candidates are ranked globally by NIS, and at most K predicted priors are written back. The NIS gate, clean-only calibration, and global Top-K allocation are specific to our adaptation.

OFT-Random-Obs (adapted). OFT-MARL jointly trains a fault-injection agent, a temporal diagnosis module, and a cooperative driving policy (Shi et al., 2025). To preserve the fixed-policy setting, our adaptation trains only a GRU diagnosis-and-reconstruction module using clean trajectories augmented with bounded uniform random faults. Predicted block-level fault probabilities rank valid non-ego blocks at runtime, and the reconstructed outputs of at most K blocks are written back. This is an OFT-style observation-repair baseline, not a reproduction of the complete jointly trained OFT-MARL framework.

ROBOSAC-Obs (adapted). ROBOSAC searches for a benign subset of collaborative BEV messages through randomized hypothesize-and-verify consensus (Li et al., 2023). We map a collaborator message to a complete CAV observation row, a BEV detection hypothesis to an identity-aligned vehicle-state hypothesis, and box-set agreement to clean-calibrated position–velocity agreement using protected CAV ego anchors. Under the three-CAV, single-compromised-source setting, the adaptation samples singleton source hypotheses for at most five trials and retains the first hypothesis accepted by the consensus test. Disagreeing blocks are ranked by normalized state distance, and at most K blocks are selected for repair. If no hypothesis is accepted, the received observation is retained.

GCP-Obs (adapted). GCP combines spatial consistency and temporal reconstruction to reject malicious collaborators in BEV perception (Tao et al., 2026). We replace BEV box-set agreement with identity-aligned position–velocity concordance across CAV observation rows and apply an LSTM autoencoder to relative-kinematic-state sequences. Spatial and temporal scores are converted to finite-sample conformal p-values, followed by Benjamini–Hochberg source rejection. Rejected sources are excluded when constructing a trusted cross-view state proxy; the LSTM reconstruction is used when no current trusted report is available. The common budget K limits the final structured-block selections.

G Budget–Performance Trends

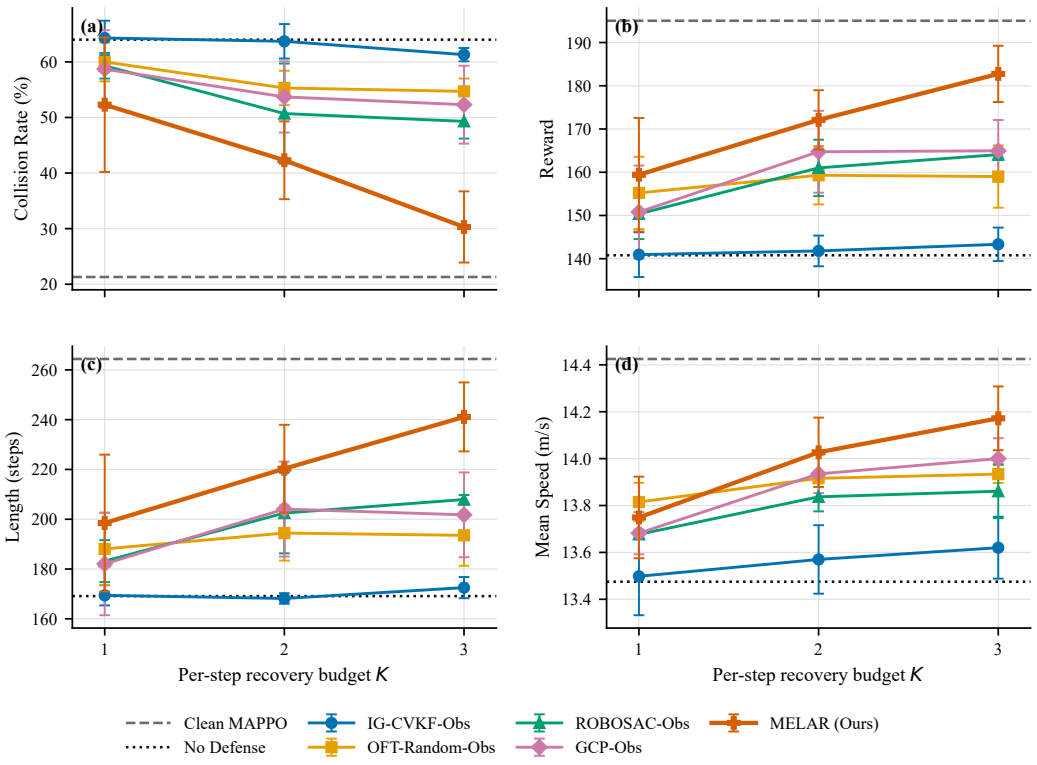

Figure 3: Budget–performance trends under the episode-persistent single-CAV policy-white-box observation attack. Each point shows the mean over three evaluation seeds, with 50 episodes per seed; error bars on the defended methods denote the sample standard deviation across seed-level results. The dashed and dotted horizontal lines show the budget-independent Clean MAPPO and No Defense references, respectively. Lower Collision Rate is better, whereas higher Reward, Length, and Mean Speed are better.