Safety21
INNOVATING FOR SAFETY

US DOT National
University Transportation Center for Safety

Carnegie Mellon University

Penn
University of Pennsylvania

The Ohio State University

Morgan State University

CCAC Community College of Allegheny County

The University of Texas Rio Grande Valley

Project Title

Cyber Resilience of Connected and Autonomous Transportation Systems (Phase II): Game-theoretic Security Assurance of Network-level Traffic Signal Control

Investigators

Mohamadhossein Noruzoliaee (https://orcid.org/0000-0003-3860-8911)

Final Report – August 2026

DISCLAIMER

Technical Report Documentation Page

1. Report No. 616	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Cyber Resilience of Connected and Autonomous Transportation Systems (Phase II): Game-theoretic Security Assurance of Network-level Traffic Signal Control		**5. Report Date** August 31, 2026 **6. Performing Organization Code**
7. Author(s) Mohamadhossein Noruzoliaee, https://orcid.org/0000-0003-3860-8911 Mark Hernandez Oziel Sauceda Fatemeh Nazari, https://orcid.org/0000-0003-1587-1848		**8. Performing Organization Report No.**
9. Performing Organization Name and Address The University of Texas Rio Grande Valley 1201 W University Dr Edinburg, Texas 78539		**10. Work Unit No.** **11. Contract or Grant No.** Federal Grant # 69A3552344811 / 69A3552348316
12. Sponsoring Agency Name and Address Safety21 University Transportation Center, Carnegie Mellon University 5000 Forbes Avenue Pittsburgh, PA 15213		**13. Type of Report and Period Covered** Final Report (July 1, 2025 - July 31, 2026) **14. Sponsoring Agency Code** USDOT

15. Supplementary Notes
Conducted in cooperation with the U.S. Department of Transportation, Federal Highway Administration.

16. Abstract
Intelligent traffic signal control (TSC) is critical to connected and autonomous transportation systems, with important implications for mobility and safety. Its reliance on observed and communicated traffic information, however, creates cybersecurity vulnerabilities to adversarial information manipulation. This research investigates the robustness of network-level deep reinforcement learning-based TSC against strategic test-time state observation cyberattacks under blackbox and whitebox attacker settings. These attacks corrupt the signal controller's observations of the ground-truth traffic state, indirectly affecting traffic evolution through altered signal decisions. From a worst-case perspective, a strategic adversary selects bounded observation perturbations to maximize cumulative rather than purely instantaneous traffic degradation over the control horizon. The blackbox attacker operates without access to the victim signal control policy, while the whitebox attacker exploits policy access to construct policy-aware perturbations. Empirical cybersecurity assurance is pursued through adversarial training against learned attackers. Experiments in SUMO traffic simulator show that blackbox and whitebox attacks degrade the undefended signal controller's performance by approximately 50% and 73%, respectively. Adversarial training reduces these losses to nearly 6% and 7%, eliminating about 89% and 91% of attack-induced degradation with a moderate sacrifice in nominal performance and a favorable performance-robustness tradeoff. Sensitivity analyses reveal nonlinear effects of perturbation budget and dependencies on attacker information, traffic demand, and network size.

17. Key Words	18. Distribution Statement
Traffic signal control; Cybersecurity; Adversarial reinforcement learning; State observation attacks.	No restrictions.

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	26	

Table of Contents

1. Introduction

1.1. Motivation

Traffic signal control (TSC) is a core component of intelligent transportation systems, with direct implications for mobility and safety. From a mobility perspective, traffic signal decisions allocate right-of-way among competing traffic movements and thereby influence delay, queuing, stops, throughput, and traffic progression (Wei et al., 2020). From a safety perspective, intersections concentrate interactions and potential conflicts among road users. In the United States, approximately one-quarter of traffic fatalities and about one-half of traffic injuries are attributed to intersections (FHWA, 2026; Kodi et al., 2021). Traffic signal operation is consequently relevant to traffic efficiency, progression, and safety-critical conflict conditions under which road users traverse intersections.

The dynamic and complex nature of traffic operations has motivated growing interest in reinforcement learning (RL) for adaptive TSC (Wei et al., 2020; Noaeen et al., 2022). RL naturally represents adaptive signal control as a sequential decision problem in which a controller observes prevailing traffic conditions, selects signal control actions, and adapts subsequent actions as traffic evolves. Deep reinforcement learning (DRL) further enables complex control policies to be learned from high-dimensional traffic observations without requiring an explicit analytical model of the underlying traffic dynamics. Research has consequently progressed from isolated intersection applications toward coordinated multi-intersection and large-scale TSC (Chen et al., 2020; Zeng et al., 2025). At these scales, vehicle progression, queue interactions, and spillback couple traffic conditions and control decisions across intersections, increasing both the potential benefits and the complexity of adaptive control. Connected vehicle and infrastructure sensing technologies can further enrich the traffic information available for real-time control (Guo et al., 2019; Mo et al., 2022), making the performance of learned signal control policies increasingly dependent on reliable representations of prevailing traffic conditions.

This dependence on sensed and communicated traffic information creates a critical cybersecurity exposure. Intelligent TSC relies on interconnected sensing, communication, and control components through which maliciously altered information, such as compromised detector measurements and falsified connected vehicle information, can influence signal control decisions. This exposure is particularly important for artificial intelligence (AI)-based TSC, where learned signal control policies select signal control actions based on observations received from the traffic environment. Previous transportation cybersecurity research has shown that falsified traffic information can mislead signal controllers and degrade traffic operations (Feng et al., 2018; Haydari et al., 2021; Feng et al., 2022; Abdo et al., 2024). These vulnerabilities form part of the broader security-of-AI challenge in connected and autonomous transportation systems, in which the integrity and robustness of AI-enabled transportation functions must be protected against adversarial manipulation (Noruzoliaee and Nazari, 2025).

Within this broader security-of-AI challenge, adversarial manipulation can target either the ground-truth traffic state or the observations through which an AI-based signal controller perceives that state. Ground-truth state attacks directly alter the physical traffic environment. For example, vehicle-level cyberattacks may compromise autonomous driving or platooning functions and induce malicious changes in vehicle speeds, spacing, or trajectories, thereby altering traffic flow and queue conditions. Such cyberattacks can be represented naturally through adversarial actions that directly affect the traffic environment or its state transitions. In contrast, state observation attacks manipulate the

traffic information available to the signal controller without directly altering the corresponding ground-truth traffic state. For example, an adversary may falsify detector occupancy or queue measurements or spoof connected vehicle reports so that the controller perceives a queue that is not actually present (Feng et al., 2018, 2022). This distinction is methodologically important because the effects of observation attacks on the physical traffic process are mediated through the signal control decisions, whereas ground-truth state attacks directly affect that process.

1.2. Research Gaps

Despite growing attention to cybersecurity in AI-enabled TSC, important gaps remain in characterizing and mitigating adversarial state observation perturbations. First, the temporally coupled effects of adversarial state observation perturbations remain underexamined in RL-based TSC. Unlike adversarial perturbations in supervised or unsupervised learning, the consequence of a perturbed observation in RL is not confined to the decision associated with that individual input. A corrupted traffic observation can alter the signal action selected by the controller, which changes vehicle movements and the subsequent ground-truth traffic state. Later signal control decisions and observation perturbations therefore occur in traffic states that have already been shaped by preceding attack-induced actions. The effects of successive perturbations are consequently coupled over time through the closed-loop traffic dynamics, allowing their operational consequences to accumulate and propagate over the decision horizon. This temporal dependence is intrinsic to the sequential interaction between the signal controller and the traffic environment and makes the long-term rather than purely instantaneous consequences of adversarial observation perturbations particularly crucial for RL-based traffic control.

Second, an important defense gap concerns robustness against worst-case strategic cyberattacks. Some existing studies focus primarily on identifying vulnerabilities or developing effective attack strategies without considering a corresponding defense, while others evaluate defenses against particular attack mechanisms (Feng et al., 2018; Perrine et al., 2019; Ren et al., 2024b). Robustness to a predefined or previously observed attack, however, does not necessarily imply robustness against a strategic adversary that actively optimizes its admissible perturbations to maximize cumulative performance degradation. This distinction motivates a worst-case perspective in which the relevant cybersecurity question is not only whether a signal controller withstands a particular attack, but how well it performs against a strong adversary operating within a prescribed attack capability (Bao et al., 2023; Zheng et al., 2023). For state observation integrity, this defense problem can arise at test time, when adversarial perturbations are introduced into observations used by an already learned and deployed signal control policy. This threat is distinct from training-time cyberattacks such as data poisoning, which corrupts data used to learn the control policy, and backdoor attacks, which compromise the learned control policy so that particular triggers can induce malicious behavior (Zhang et al., 2023; Ren et al., 2024a). Robustness against strategic test-time attacks therefore raises a different defense objective from securing the original learning process. Moreover, strengthening a signal controller against strong adversarial perturbations may reduce the nominal traffic system performance, making the tradeoff between robustness under attack and performance in the absence of attack an important consideration.

Third, the effectiveness of a state observation cyberattack can depend on the information available to the adversary. Prior research on DRL-based TSC has considered both whitebox and blackbox attack settings (Haydari et al., 2021; Ren et al., 2024a). Whitebox attacks assume access to information about the victim signal controller's policy that is unavailable under a blackbox threat model, potentially

allowing the adversary to construct perturbations that more directly exploit the controller's decision policy. Although both information settings have appeared in the TSC cybersecurity literature, existing studies employ different attack formulations, victim controllers, perturbation constraints, and traffic environments. These differences limit direct assessment of how access to the victim policy itself influences attack effectiveness. A controlled comparison of policy-agnostic blackbox and policy-aware whitebox state observation attacks under a common victim policy, perturbation budget, attack objective, and traffic environment is therefore needed to better characterize the effect of attacker information on the vulnerability of network-level TSC.

1.3. Research Objectives and Contributions

To address these gaps, this research aims to investigate the cybersecurity of network-level DRL-based TSC under strategic test-time state observation attacks in both blackbox and whitebox attack information settings and to develop a robust signal controller against such attacks. The problem is formulated as a game-theoretic zero-sum interaction between a strategic adversary and a centralized network-level signal controller with opposing objectives. The signal controller seeks to maximize cumulative network traffic performance, while the adversary seeks to minimize that performance by selecting admissible perturbations to the traffic state observations presented to the signal controller. Accordingly, the attacker is strategic in the sense that its objective concerns the cumulative effect of its perturbations over the control horizon rather than the immediate effect of an individual perturbed observation. The corresponding defense objective is formulated as a max-min problem in which the signal controller is optimized against the strongest admissible state observation adversary within a prescribed perturbation budget.

On the cyberattack side of this max-min problem, strategic state observation attacks require a sequential attack decision formulation that explicitly represents the mediated interaction between the adversary, the victim signal controller's policy, and the traffic environment. At each decision step, the adversary perturbs the signal controller's observation of the current ground-truth traffic state. The perturbed observation is mapped by the victim policy to a signal control action, which then governs the transition of the traffic environment from the ground-truth state. Thus, the effect of the adversarial action on the subsequent ground-truth state is mediated by the victim policy rather than entering the traffic state transition directly. This mediation is methodologically non-trivial noting that the original traffic transition dynamics specify how the environment evolves in response to the signal controller's action, not in response to the adversary's observation perturbation. Consequently, a sequential decision model for the attacker must account for the victim policy as part of the mechanism that maps an adversarial perturbation to its subsequent state transition and cumulative consequence. This structure differs from adversarial attacks on the ground-truth state or environment dynamics, where the adversarial intervention directly affects the state transition process without such mediation and can be represented using robust Markov decision process (RMDP) formulations (Nilim and Ghaoui, 2003; Iyengar, 2005). It also differs from non-adversarial state observation uncertainty, where the observation available to the decision maker is generated through a prescribed observation process rather than strategically manipulated by an adversary, as represented by conventional partially observable Markov decision process (POMDP) formulations (Nazari et al., 2017; Sutton and Barto, 2018).

To formulate this mediated cyberattack process under the two attacker information settings, this research adopts state-of-the-art adversarial RL formulations for blackbox and whitebox attacks. Under

the blackbox setting, the attack is formulated using the state-adversarial Markov decision process (SA-MDP) framework (Zhang et al., 2020, 2021). For a fixed victim policy, SA-MDP transforms the attacker's sequential decision problem into a new Markov decision process (MDP) with redefined transition dynamics and reward that incorporate the victim policy's response to the perturbed observation and the dynamics of the original environment. This transformation enables the adversary to learn a policy that directly selects admissible perturbed observations to optimize cumulative attack performance without requiring access to the internal structure of the victim policy. Under the whitebox setting, the attack is formulated using the policy adversarial actor-director (PA-AD) framework (Sun et al., 2021). Given access to the victim policy, PA-AD formulates the attack in the victim's policy space and decomposes the adversarial decision into a director, which selects a direction for perturbing the victim's action distribution, and an actor, which realizes the selected direction through an admissible state observation perturbation. Evaluating the SA-MDP-based blackbox and PA-AD-based whitebox attacks under the same victim signal controller, perturbation budget, attack objective, and traffic environment provides a controlled basis for assessing how access to the victim policy affects the cybersecurity of network-level TSC.

The main contribution of this research is to operationalize the resulting attack-defense interaction for network-level TSC and systematically evaluate its cybersecurity implications. The max-min defense problem is addressed through adversarial training using alternating training with learned adversaries (ATLA) (Zhang et al., 2021). ATLA alternates between optimizing the adversary against a fixed signal control policy and optimizing the signal controller against the resulting fixed adversary, thereby providing a practical approach for approximating the max-min objective through successive attack and defense optimization. Robustification is performed offline before signal controller deployment, after which the resulting signal controller is fixed during test-time attack evaluation and does not adapt online to the attacker. The experiments evaluate the nominal signal controller and its robustified counterparts under blackbox and whitebox attacks and quantify both attack-induced degradation and the mitigation achieved through adversarial training. The analysis further examines the tradeoff between nominal traffic system performance and robustness under attack, investigates sensitivity to the admissible observation perturbation budget and traffic demand, and provides an exploratory assessment of vulnerability with increasing network size. Finally, the experiments examine how adversarial manipulation confined to the controller's observations alters joint signal decisions and subsequently propagates into physical traffic conditions, thereby connecting information-layer attacks to their network-level operational consequences.

The remainder of this report is organized as follows. Section 2 reviews prior research on cybersecurity threats to TSC and its adversarial robustness, with particular attention to the gaps motivating this research. Section 3 presents the network-level RL-based TSC problem, formulates the blackbox and whitebox observation attack models, and develops the adversarial training framework for robust signal control. The simulation experiments presented in section 4 describe the experimental results, including attack effectiveness, defense performance, and sensitivity analyses. Finally, it concludes by summarizing the key findings and identifying directions for future research.

2. Related Work

Cybersecurity research on traffic signal control has examined vulnerabilities arising from both direct compromise of signal control components and manipulation of the information used for traffic

control decisions. Feng et al. (2018) identified signal controllers, vehicle detectors, roadside units, and onboard units as potential attack surfaces in a connected traffic control environment. Focusing on actuated and adaptive signal control, they modeled falsified detector or vehicle data as an indirect attack and formulated the attacker's objective as maximizing system delay subject to constraints on attack resources and intensity. Their results further showed that attack effectiveness depends strongly on where the attack is introduced in the traffic network. Perrine et al. (2019) investigated a different attack mechanism by selectively disabling signalized intersections in a district-scale network and replacing their operation with four-way stops. Their dynamic traffic assignment analysis demonstrated that the traffic impact depends substantially on which signals are compromised and on how the attacker prioritizes signal locations. With increasing use of connected vehicle information in signal control, Feng et al. (2022) developed a broader cybersecurity analysis framework for connected vehicle-based signal control. While examining threats at the communication, operating system, and application levels, they identified data spoofing as a plausible attack pathway and used a hybrid cyber-physical testing platform to evaluate both its operational effects and corresponding mitigation measures. These studies collectively demonstrate that cyberattacks can influence traffic operations through different pathways, ranging from direct disruption of traffic signal operation to falsification of the traffic information on which signal control decisions depend.

Beyond identifying vulnerabilities and their consequences, another line of research has formulated cyberattacks strategically by explicitly optimizing the adversary's intervention with respect to traffic system performance. Lopez et al. (2020) developed a formal traffic-theoretic attack modeling and impact analysis methodology for fixed-time TSC in which attacks modify control settings such as green time ratios, cycle lengths, and retaining ratios. Using a link-queue traffic model, they identified vulnerable traffic states in which tampered control settings can substantially accelerate gridlock formation or reduce network flow. Thodi et al. (2020) formulated traffic signal tampering as a bi-objective optimization problem that simultaneously seeks to reduce network throughput and limit changes to signal timings, thereby representing the tradeoff between attack impact and noticeability. Yen et al. (2022) considered misinformation attacks against backpressure-based signal control. Their time-spoofing attack falsifies vehicle arrival times, while their ghost vehicle attack hides vehicles from the controller by disconnecting their wireless communication. The adversary selects vehicles to attack with the objective of maximizing disrupted signal phases, and the attack selection problem is formulated as a binary knapsack problem. These studies show a progression from assessing whether a traffic control system can be compromised toward explicitly determining how an adversary can allocate or select its interventions to produce damaging traffic consequences under a specified attack model.

Cybersecurity research has also addressed the complementary problem of designing defenses and robust signal control strategies under attack. Li et al. (2016) formulated cybersecurity risk assessment for traffic signal systems as a bi-level game-theoretic problem and proposed a minimax-regret methodology for prioritizing defensive measures. Bao et al. (2023) considered direct attacks that maliciously alter signal settings and formulated robust network-wide signal optimization through bi-level optimization models for attackers (lower-level problem) with complete and incomplete knowledge of the traffic control system (upper-level problem). The sequential interaction between attacker and defender is represented as a Stackelberg game, and robust signal plans are optimized with respect to traffic efficiency and safety-related objectives. Zheng et al. (2023) similarly developed robust signal control plans against risk-averse and worst-case cyberattacks through a bi-level attack-defense

formulation in which the attacker (lower-level) seeks damaging signal modifications and the defender (upper-level) adjusts the signal plan to mitigate their effects. Yen et al. (2022) proposed auction-based and hybrid protection algorithms against their time-spoofing and ghost vehicle attacks. Overall, the existing TSC cybersecurity literature contains explicit attacker-defender formulations, network-level robust optimization, worst-case attack models, and game-theoretic defenses. These methods, however, address different attack surfaces and signal controller structures, ranging from malicious modification of signal settings to misinformation generated by connected vehicles.

As learning-based signal control has emerged, TSC cybersecurity research has increasingly examined attacks that exploit the dependence of learned policies on data and their susceptibility to adversarial manipulation. Haydari et al. (2021) investigated adversarial attacks against DRL-based TSC under both whitebox and blackbox settings and evaluated their effects for different DRL controllers in single- and multi-intersection environments. They subsequently evaluated sequential anomaly detection methods and proposed an ensemble detector for identifying adversarial inputs. The sequential component in this study concerns the anomaly detection process rather than a sequential decision formulation of the attacker. Ren et al. (2024a) considered a different training-time attack surface and developed two stealthy blackbox attacks on the reward space of DRL-based TSC. Their formulations explicitly consider the time dependence of DRL, learn malicious adversarial policies, and seek to balance attack effectiveness with constraints associated with attack frequency or modification. A subsequent study by Ren et al. (2024b) examined blackbox attacks against multi-agent RL-based multi-intersection TSC and used a dynamic threshold to select critical states for attack under a limited attack budget. The attack seeks to minimize cumulative reward, while also proposing a lightweight min-max defense intended to improve worst-case policy performance against perturbed states. Zhang et al. (2023) investigated backdoor attacks in which a special driving behavior serves as a trigger that induces malicious behavior from a compromised DRL-based signal controller. Qu et al. (2023) considered colluding vehicles that strategically transmit falsified traffic information to DRL-based adaptive signal control with the objective of reducing the collective travel time of the colluding vehicles. The literature on learning-based TSC thus spans different attack surfaces, attacker information settings, attack objectives, and stages of compromise.

Taken together, the reviewed literature establishes foundations of TSC cybersecurity while also delineating the specific unresolved issues motivating this research. Previous research on TSC cybersecurity has demonstrated substantial operational consequences from compromised signal settings, falsified traffic information, disabled signals, and malicious connected vehicles, and has developed optimization-based and game-theoretic approaches for both attack and defense. Within the reviewed literature, however, these elements have largely been investigated under different threat models and experimental settings. In particular, the cumulative rather than merely instantaneous consequences of successive state observation perturbations remain insufficiently characterized for network-level RL-based TSC, where an altered observation can change a joint signal action, modify the subsequent physical traffic state, and thereby affect later attack and control decisions. Moreover, although both blackbox and whitebox attacks have been studied in learning-based TSC, differences among existing attack formulations and evaluation settings make it difficult to isolate the effect of attacker access to the victim policy. Additionally, existing TSC defenses do not directly establish how a network-level RL signal controller performs when robustified against a learned adversary that strategically selects bounded test-time state observation perturbations to degrade cumulative traffic performance. These observations

motivate a controlled evaluation of policy-agnostic blackbox and policy-aware whitebox state observation attacks under a common victim controller, perturbation budget, attack objective, and traffic environment, together with adversarial training to investigate the resulting tradeoff between nominal traffic control performance and robustness under attack.

3. Modeling Framework

This section presents the modeling framework used to characterize the operation, vulnerability, and robustness of the network-level traffic signal control system. The framework is developed progressively in three stages. First, the no-attack baseline (section 3.1) formulates the centralized reinforcement learning-based traffic signal controller under normal (no-attack) operating conditions. Second, the attack model (section 3.2) extends the baseline by introducing adversarial perturbations to the signal controller's state observations under both blackbox and whitebox attack settings. Finally, the defense model (section 3.3) formulates robust signal control against such adversarial observations and establishes the basis for adversarial training of the traffic signal controller.

3.1. No-attack baseline

In the absence of adversarial perturbations, the network-level traffic signal control is modeled as a fully centralized cooperative multi-agent reinforcement learning problem. We consider a road traffic network comprising a set of signalized intersections $I = \{1, \dots, N\}$, where each intersection is associated with a signal control agent. Under full centralization, the intersection agents do not make independent decisions. Rather, a centralized controller observes the traffic conditions across the network and jointly determines the signal control actions for all intersections. Accordingly, the cooperative multi-agent problem can be formulated as a standard MDP represented by the tuple $M = \{S, A, R, P, \gamma\}$, where S denotes the network-level traffic state space, A represents the joint action space of all signalized intersections, R is the network-level reward function, P is the state transition function, and $\gamma \in [0,1]$ is the discount factor.

At each decision time step t, the centralized controller observes the true network traffic state $s_t \in S$ and selects a joint traffic signal action $a_t \in A$ according to its policy $\pi_\theta(a_t|s_t)$, which is a mapping from states to actions, parameterized by a deep neural network with learnable parameters θ. The global network state is composed of the local states of all controlled intersections and is expressed as $s_t = [(s_t^1)^T, \dots, (s_t^N)^T]$, where s_t^i denotes the local traffic state of intersection i. Similarly, the joint traffic signal action consists of the actions selected for all controlled intersections and is expressed as $a_t = (a_t^1, \dots, a_t^N)$, where a_t^i is the local traffic signal action selected for intersection i. Upon implementing the joint action, the traffic environment transitions to the next state s_{t+1} according to the transition function $P(s_{t+1}|s_t, a_t)$, which denotes the probability of transitioning to state s_{t+1} given the current state s_t and action a_t. Simultaneously, the controller receives an instantaneous network-level reward $r_t = R(s_t, a_t)$, where $R: S \times A \to \mathbb{R}$ is the reward function.

Through repeated interactions with the traffic environment, the centralized controller seeks an optimal policy π_θ^* that maximizes the expected cumulative discounted network reward in Eq. (1). Here, the expectation $\mathbb{E}$ is taken over trajectories induced by the initial network state $s_0 \sim \rho_0$ representing the starting traffic conditions throughout the controlled network, the actions selected according to the signal controller's policy, and the subsequent traffic state transitions.

$$\pi_\theta^* = \arg\max_\theta \mathbb{E}_{s_0 \sim \rho_0, a_t \sim \pi_\theta(.|s_t), s_{t+1} \sim P(.|s_t, a_t)} \left[\sum_{t=0}^{T} \gamma^t r_t(s_t, a_t) \right] \quad (1)$$

The local state s_t^i captures the signal status and traffic conditions at intersection $i \in I$. Each intersection contains multiple incoming and outgoing lanes distributed across its approaches. An approach, such as the northbound approach, refers to a direction from which traffic enters the intersection and may contain several lanes serving through, left-turn, or right-turn movements. A signal phase represents a set of nonconflicting traffic movements that simultaneously receive a green indication. Let L_i^{in} and L_i^{out} denote the sets of incoming and outgoing lanes at intersection i. The local traffic state is defined as in Eq. (2). Here, at each time step t, p_t^i denotes the one-hot encoding of the active green phase at intersection i, e_t^i represents the elapsed green time of the active phase at intersection i, $n_{l,t}$ is the number of vehicles occupying lane l, and $d_{l,t}$ is the average delay per vehicle on incoming lane l.

$$s_t^i = \left[p_t^i, e_t^i, \{n_{l,t}\}_{l \in L_i^{\text{in}}}, \{n_{l,t}\}_{l \in L_i^{\text{out}}}, \{d_{l,t}\}_{l \in L_i^{\text{in}}} \right] \quad (2)$$

The action space is discrete and consists of signal phase selections. Let $A_i = \{1, \dots, K_i\}$ denote the local action space of intersection i, where K_i is the number of feasible signal phases specified by its signal control plan. The local action satisfies $a_t^i \in A_i$, where $a_t^i = k$ indicates that the signal phase k is selected at intersection i at decision time step t. Since the centralized controller jointly selects one signal phase for every controlled intersection, the network level action space is the Cartesian product of the local action spaces $A = A_1 \times \dots \times A_N$. Therefore, each joint action $a_t = A$ specifies a combination of signal phases to be implemented across the entire controlled network. When the selected phase at an intersection differs from its currently active phase, the required yellow and all-red clearance intervals are implemented before the newly selected phase becomes active.

The controller's reward function is pressure-based (Mo et al., 2022). At the lane level, traffic pressure is defined as the difference between the normalized vehicle counts on an incoming lane and its corresponding downstream lane, where each vehicle count is normalized by the capacity of the respective lane. At the intersection level, traffic pressure represents the aggregate imbalance across the incoming and downstream lane pairs associated with the traffic movements at the intersection. Let $\mathcal{L}_i$ denote the set of incoming and downstream lane pairs at intersection i. The aggregate traffic pressure at intersection i at time step t, denoted by $\phi_{i,t}$, is calculated as below. Here, $n_l^{\max}$ and $n_{l'}^{\max}$ denote the capacities of incoming lane l and downstream lane l', respectively.

$$\phi_{i,t} = \left| \sum_{(l,l') \in \mathcal{L}_i} \left(\frac{n_{l,t}}{n_l^{\max}} - \frac{n_{l',t}}{n_{l'}^{\max}} \right) \right| \quad (3)$$

At the network level, the intersection pressures are aggregated as in Eq. (4) to define the common reward received by the centralized signal controller. The pressure-based reward encourages the centralized controller to jointly select signal phases that discharge high pressure traffic movements across the network, thereby improving network throughput and mitigating queue spillback.

$$r_t = -\sum_{i \in I} \phi_{i,t} \tag{4}$$

Finally, the state transition function P is unknown and is not explicitly modeled by the signal controller. Rather, the centralized traffic signal controller interacts with the Simulation of Urban Mobility (SUMO) microscopic traffic simulator (Alvarez Lopez et al., 2018), which determines the subsequent traffic state based on the current state, the joint signal phase selections, and the resulting traffic flow dynamics. The centralized traffic signal controller is therefore trained using a model-free DRL approach (Sutton and Barto, 2018).

3.2. Attack model

The no-attack baseline assumes that, at each decision time step t, the centralized traffic signal controller observes the ground truth traffic state s_t and selects the joint signal control action a_t according to it control policy $\pi_\theta(a_t|s_t)$. The cyberattack model extends this baseline by considering state observation attacks, in which an adversary manipulates the traffic state information available to the centralized controller before a signal control decision is made. Let $\hat{s}_t$ denote the adversarial observation generated from the true traffic state s_t. Under attack, the centralized controller therefore makes its joint signal control decision according to $\pi_\theta(a_t|\hat{s}_t)$ rather than $\pi_\theta(a_t|s_t)$. The perturbation affects only the controller's observation and does not alter the underlying ground truth traffic state. As such, once the controller selects a_t based on $\hat{s}_t$, the traffic network evolves from the true state according to the transition dynamics defined for the no-attack baseline $P(s_{t+1}|s_t, a_t)$.

The adversary is constrained in the extent to which the traffic network state observation can be manipulated. For each true state s_t, let $\mathcal{B}_\varepsilon(s_t) \subseteq S$ denote the set of admissible perturbed observations under attack budget ε, such that the generated observation must satisfy $\hat{s}_t \in \mathcal{B}_\varepsilon(s_t)$. The perturbation budget ε controls the maximum allowable perturbation magnitude. The perturbation set restricts the adversary to neighboring observations rather than allowing arbitrary replacement of the true state. For a norm-bounded observation attack, the admissible attack is defined as below.

$$\mathcal{B}_\varepsilon(s_t) = \left\{\hat{s} \in S: \|\hat{s} - s_t\|_p \leq \varepsilon\right\} \tag{5}$$

The attacker's choice of $\hat{s}$ constitutes a sequential decision problem. A perturbed observation can induce the centralized signal controller to select a different combination of signal phases, thereby affecting the subsequent traffic state s_{t+1}. This altered state subsequently influences both future attack decisions and future signal control actions. The effectiveness of an observation perturbation must therefore be evaluated according to its cumulative effect on network traffic conditions rather than solely according to its immediate effect on the current control decision.

Let $\nu_\eta(\hat{s}_t|s_t)$ denote an adversarial policy parameterized by η. Analogous to the victim policy $\pi_\theta(a_t|s_t)$, which determines the signal controller's action based on its state observation, the adversarial policy determines how the controller's observation is manipulated as the traffic network evolves. Since the baseline (no-attack) signal controller is optimized to maximize the negative aggregate pressure across the controlled intersections, the adversary seeks the opposite outcome. Thus, an optimal attack policy is one that maximizes the expected cumulative discounted aggregate network pressure,

as shown in Eq. (6). Here, the expectation is taken over the attacked trajectories induced by the initial network state, the adversarial observations, the joint signal actions selected by the fixed victim policy, and the subsequent traffic state transitions. This objective is directly opposite to the pressure-based objective of the centralized controller defined in Eq. (1).

$$\nu_{\eta}^{*} = \arg\max_{\eta} \mathbb{E}\left[\sum_{t=0}^{T} \gamma^{t} \sum_{i \in I} \phi_{i,t}\right] \tag{6}$$

The blackbox and whitebox cyberattacks considered below differ in how this sequential adversarial decision is formulated. Under the policy-agnostic blackbox setting (section 3.2.1), the attacker does not require knowledge of the victim policy and directly learns an adversarial policy for selecting perturbed state observations. Under the policy-aware whitebox setting (section 3.2.2), however, the attacker has access to the victim policy and exploits this information to formulate the sequential attack decision in the victim's policy space.

3.2.1. Blackbox attack model

Under the blackbox attack setting, the adversary does not require access to the internal structure or parameters of the victim signal controller's policy π_{θ}. Instead, it learns through interaction with the signal controller and the traffic environment which admissible state observation perturbations produce the greatest cumulative degradation in network performance. For a fixed victim policy, the sequential problem of finding an optimal state observation adversary can be formulated as a new MDP by incorporating the response of the victim policy and the original environment dynamics into redefined transition and reward functions. Following the SA-MDP framework established theoretically in Zhang et al. (2020, 2021), the blackbox adversary MDP is defined as $\widehat{M}^{B} = \{S, \hat{A}^{B}, \hat{R}^{B}, \hat{P}^{B}, \gamma\}$, where the adversary observes the original network traffic state space S and its adversarial action $\hat{a} \in \hat{A}^{B}$ represents the state observation presented to the victim signal controller $\hat{s}$.

$$\hat{A}^{B} = S \tag{7}$$

For the fixed victim π_{θ}, the transition probability of the adversary MDP is shown in Eq. (8).

$$\hat{P}^{B}(s'|s,\hat{a}) = \sum_{a \in A} \pi(a|\hat{a})\, P(s'|s,a) \tag{8}$$

The corresponding reward function is

$$\hat{R}^{B}(s,\hat{a},s') = \begin{cases} -\dfrac{\sum_{a \in A} \pi(a|\hat{a})\, P(s'|s,a)\, R(s,a,s')}{\sum_{a \in A} \pi(a|\hat{a})\, P(s'|s,a)} & \hat{a} \in \mathcal{B}_{\varepsilon}(s) \\ C & \hat{a} \notin \mathcal{B}_{\varepsilon}(s) \end{cases} \tag{9}$$

where C is a large penalty associated with an inadmissible adversarial action. The negative sign reverses the victim's objective in the transformed MDP $\widehat{M}^{B}$. As the original network reward R is defined from the negative aggregate intersection pressure, maximizing $\hat{R}^{B}$ corresponds to seeking observation

perturbations that degrade the pressure-based performance of the signal controller.

Let $\nu_{\eta_B}(\hat{a}|s)$ denote the blackbox adversarial policy parameterized by η_B. The optimal blackbox attack policy is obtained based on Eq. (10).

$$\nu_{\eta_B}^{*} = \arg\max_{\eta_B} \mathbb{E}_{s_0\sim\rho_0,\hat{a}_t\sim\nu_{\eta_B}(.|s_t),s_{t+1}\sim\hat{P}^B(.|s_t,\hat{a}_t)}\left[\sum_{t=0}^{T}\gamma^t\hat{R}^B(s_t,\hat{a}_t,s_{t+1})\right] \tag{10}$$

Thus, the blackbox attack is represented as a new, standard MDP in which the adversary sequentially selects the state observations presented to the centralized signal controller. The optimal policy of this MDP induces the optimal state observation adversary against the fixed victim signal controller' policy under the predetermined perturbation budget.

3.2.2. Whitebox attack model

The whitebox cyberattack considers a stronger information setting in which the adversary has access to the fixed victim signal controller policy $\pi(a|s)$. In addition to observing the ground-truth network traffic state s, the adversary can therefore evaluate how the centralized signal controller's decision policy changes when different admissible adversarial observations are presented to it. This additional information provides an alternative way to formulate the state observation attack. Specifically, when the ground-truth state s is replaced by an admissible perturbed observation $\hat{s}$, the victim signal controller evaluates its policy at $\hat{s}$ and may consequently produce a different distribution over signal control actions, $\pi(a|\hat{s})$. Equivalently, the state observation perturbation can be viewed as inducing a perturbed victim policy $\hat{\pi}$ at the ground-truth state, such as $\hat{\pi}(a|s) \coloneqq \pi(a|\hat{s})$. Therefore, directly perturbing the state observation, as in the blackbox attack model, can equivalently be viewed as indirectly perturbing the victim policy executed at the ground-truth state. In this sense, every admissible state observation perturbation induces a corresponding perturbation of the victim policy. This equivalence allows the attack problem to be formulated in the victim's policy space rather than directly in the potentially high-dimensional state observation space. That is, finding an optimal state observation adversary is equivalent to finding, among the policies induced by admissible observation perturbations, the policy that most severely degrades the victim's cumulative performance.

Directly searching over all policies that can be induced by admissible state perturbations, however, remains exhaustive. As theoretically established in Sun et al. (2021), an optimal policy adversary can be found on the outer boundary of this admissible policy set. This allows the attack to be decomposed into two coupled tasks. First, a direction is selected in the victim's policy space along which its decision policy should be perturbed. Second, an admissible state observation is found that moves the victim policy as far as possible in the selected direction. This decomposition leads to the PA-AD formulation. The director addresses the attacker's sequential decision problem by selecting a policy perturbation direction at each traffic state, while the actor determines an admissible perturbed observation that moves the victim policy as far as possible in the selected direction. Unlike the blackbox adversarial policy which directly selects the state observation presented to the victim controller, the whitebox director selects an intermediate action in the victim's policy space, which is subsequently translated by the actor into the perturbed state observation.

Following Sun et al. (2021), for the stochastic victim policy π_θ, the director's sequential decision

problem is formulated as a policy adversary Markov decision process (PA-MDP), represented by the tuple $\widehat{M}^W = \{S, \hat{A}^W, \hat{R}^W, \hat{P}^W, \gamma\}$, where the director observes the original network traffic state $s \in S$, while its action $d \in \hat{A}^W$ specifies a perturbation direction in the victim's policy space.

$$\hat{A}^W = \left\{ d \in [-1,1]^{|A|} : \sum_{j=1}^{|A|} d_j = 0 \right\} \tag{11}$$

The actor is embedded in the PA-MDP and maps the direction d selected by the director at state s to an admissible perturbed state observation $\hat{s}$. Specifically, for a selected direction d at state s, the actor determines the corresponding perturbed state observation by solving the constrained optimization shown in expressions (12)-(13).

$$g(d,s) = \arg \max_{\hat{s} \in \mathcal{B}_\varepsilon(s)} \|\pi_\theta(.|\hat{s}) - \pi_\theta(.|s)\| \tag{12}$$

subject to

$$[\pi_\theta(.|\hat{s}) - \pi_\theta(.|s)]^T d = \|\pi_\theta(.|\hat{s}) - \pi_\theta(.|s)\| \|d\| \tag{13}$$

Thus, the actor searches within the predetermined perturbation set for the observation $\hat{s}$ that produces the largest change in the victim policy along the direction d selected by the director. The resulting observation $\hat{s} = g(d,s)$ is then presented to the centralized signal controller. In this way, the director selects the policy perturbation direction d, while the actor realizes that direction through an admissible state observation perturbation.

For the fixed victim policy π_θ, this director-actor interaction determines the transition and reward experienced by the director in the PA-MDP. The transition probability of the PA-MDP is

$$\hat{P}^W(s'|s,d) = \sum_{a \in A} \pi(a|g(d,s))\, P(s'|s,a) \tag{14}$$

and the corresponding reward is

$$\hat{R}^W(s,d) = -\sum_{a \in A} \pi(a|g(d,s))\, R(s,a) \tag{15}$$

The negative sign reverses the victim's objective. As the original victim reward is defined from the negative aggregate intersection pressure, maximizing $\hat{R}^W$ corresponds to selecting policy-perturbation directions that induce state observations leading to degraded pressure-based performance of the centralized controller. Together, the actor mapping, fixed victim policy, and original traffic dynamics define the transition and reward functions experienced by the director.

Let $\nu_{\eta_W}(d|s)$ denote the whitebox director policy parameterized by η_W. The optimal director policy is obtained by solving the PA-MDP below.

$$\nu_{\eta_W}^* = \arg\max_{\eta_W} \mathbb{E}_{s_0 \sim \rho_0, d_t \sim \nu_{\eta_W}(.|s_t), s_{t+1} \sim \hat{P}^W(.|s_t, d_t)} \left[\sum_{t=0}^{T} \gamma^t \hat{R}^W(s_t, d_t) \right] \tag{16}$$

Thus, the whitebox attack is represented as a PA-MDP in which the director sequentially selects the directions for perturbing the victim's decision policy, while the actor maps each selected policy perturbation direction to the corresponding admissible perturbed observation. The optimal policy of the PA-MDP, together with the corresponding actor, induces the optimal state observation adversary against the fixed victim controller policy under the predetermined perturbation budget.

3.3. Defense model

The defense model seeks to learn a centralized traffic signal control policy π_θ that remains effective when its input state observations are subject to the adversarial perturbations characterized in section 3.2. Unlike the no-attack baseline in which the controller is optimized using the ground-truth traffic network state, the robust signal controller must learn to make effective signal control decisions from perturbed observations that may be generated by a strong adversary at the deployment time. The defense is intended to provide robustness to state observation attacks generally, including both the blackbox attack and the whitebox cyberattacks presented in section 3.2.

At the defense level, let ν denote the policy of an admissible state observation adversary. For the blackbox attack model, ν directly determines the perturbed observation presented to the victim signal controller. For the whitebox attack formulation, the corresponding state observation adversary is induced by the PA-AD director together with its actor. Regardless of how the adversarial observation is generated, the robust control objective is to maximize the victim signal controller's cumulative reward under the most damaging admissible observation attack:

$$\pi_\theta^* = \arg\max_{\theta} \min_{\nu} \mathbb{E}_{s_0 \sim \rho_0, \hat{s}_t \sim \nu(.|s_t), a_t \sim \pi_\theta(.|\hat{s}_t), s_{t+1} \sim P(.|s_t, a_t)} \left[\sum_{t=0}^{T} \gamma^t r_t(s_t, a_t) \right] \tag{17}$$

where the inner minimization represents the attacker's objective of finding admissible state observation perturbations that produce the lowest cumulative reward for the centralized traffic signal controller, whereas the outer maximization seeks a controller that maintains high traffic network performance under such adverse observations. This worst-case formulation is consistent with the objective of state-adversarial reinforcement learning, in which robustness is evaluated against the strongest admissible observation process for a given policy.

For a fixed state observation adversary, the signal controller generally faces a partially observable decision problem. The physical traffic state s_t continues to evolve according to the original transition dynamics, but the signal controller observes the perturbed state $\hat{s}_t$ rather than s_t. The fixed adversary therefore acts as an observation process between the traffic environment and the signal controller, making the resulting signal control problem functionally equivalent to a partially observable Markov decision process. Because a perturbed observation may not uniquely identify the underlying physical traffic state, information from previous observations and actions can potentially be useful for robust control, and an optimal policy under a fixed adversary may in general be history-dependent.

Directly solving the coupled max-min problem in Eq. (17) is difficult in practice because the

optimal adversary depends on the current victim policy, while the robust victim policy in turn depends on the adversarial observations encountered during training. More importantly, alternating optimization follows naturally from the theoretical underpinnings of the attack models presented in section 3.2. Both the blackbox SA-MDP and the whitebox PA-AD formulations characterize the adversarial optimization problem for a fixed victim policy. Under this condition, the victim signal controller and physical environment can be incorporated into the adversary's decision process, allowing the adversarial policy to be formulated as an MDP. Therefore, continuously updating the victim and adversary simultaneously would no longer correspond directly to the fixed-victim decision problems established in the respective attack formulations.

These theoretical considerations, together with the practical difficulty of directly solving Eq. (17), motivate the use of alternating training with learned adversaries (ATLA) (Zhang et al., 2021). ATLA separates adversarially robust training into successive stages in which one policy is frozen while the other is learned. Starting from the current victim signal controller policy π_θ and adversarial policy ν_η, π_θ is first frozen while ν_η is trained against it. Since the victim remains fixed during this stage, the adversary can be learned according to the corresponding attack formulations established in section 3.2. Once the adversary has been updated, ν_η is frozen and π_θ is trained against the adversarial observation process produced by that adversary. The two stages are then repeated, so that each updated adversary identifies vulnerabilities of the current signal controller and each subsequent controller update learns to operate under the resulting adversarial observations.

4. Experiments

This section evaluates the centralized traffic signal controller under the proposed blackbox and whitebox state observation attacks and examines the effectiveness of adversarial training in improving robustness against these attacks. The experimental setup is first described in section 4.1, followed by the attack and defense results (section 4.2) and a discussion of the principal findings (section 4.3).

4.1. Experiment setup

The attack and defense problems formulated as MDP variants in section 3 are implemented computationally using model-free deep reinforcement learning. The victim policy π_θ and the adversarial policy ν_η, with η_B or η_W corresponding to the blackbox or whitebox attack, respectively, are optimized using proximal policy optimization (PPO) (Schulman et al., 2017) through interaction with the SUMO traffic simulator (Alvarez Lopez et al., 2018), without requiring an explicit traffic state transition model. Under adversarial training, the victim and adversarial policies are trained as separate PPO policies using the ATLA framework, with the attack MDP and defense MDP optimized through alternating updates. The victim policy and value function are each represented by a two-layer multilayer perceptron with 256 hidden units per layer and tanh activations, while the adversarial policy is represented by a smaller two-layer Gaussian policy with 64 hidden units per layer. PPO uses a clipping ratio of 0.1, two optimization epochs per update, an entropy coefficient of 0.01, a discount factor of 0.99, an Adam learning rate of 10^{-4}, gradient clipping at 0.5, 128-step rollouts, and 32 minibatches over 2,000 training iterations. Bootstrapped returns are used for value estimation rather than generalized advantage estimation. The adversarial state observation is constrained by a perturbation budget of $\varepsilon = 0.2$ on the normalized observation, which remains fixed throughout training without annealing. For the whitebox

attack, the constrained actor optimization problems in Eqs. (12)-(13) are implemented using a Lagrangian relaxation, allowing the actor to realize the policy perturbation direction selected by $\nu_{\eta w}$ using gradient-based optimization.

The traffic environment is implemented in SUMO using a 3×3 grid road network comprising nine signalized two-way intersections spaced 150 meters apart. Each intersection contains two lanes per approach, including one shared through and right-turn lane and one protected left-turn lane. All intersections operate under the same four-phase signal plan consisting of north-south through and right-turn movements, north-south protected left turns, east-west through and right-turn movements, and east-west protected left turns. A two-second yellow clearance is applied between signal phase transitions when required, and phase switching remains subject to the minimum-green requirement. Traffic demand is represented by 24 constant-rate flows that are continuously introduced throughout each episode, comprising 12 through movements at 180 vehicles/hour and 12 turning movements at 60 vehicles/hour. Each episode spans 1,800 seconds of simulated traffic operation.

The state and action spaces are defined jointly over the nine signalized intersections. At each intersection, the local traffic state comprises a four-dimensional one-hot representation of the active signal phase, a green elapsed indicator, eight per-lane traffic density measurements, and eight per-lane queue measurements. Lane density and queue values are normalized by lane capacity and clipped to [0,1], yielding 21 state observation dimensions per intersection. The nine local observations are concatenated according to a fixed intersection ordering to form a 189-dimensional global traffic network state, which is further normalized online using z-score normalization before being provided to the signal controller's policy. At each signal control decision, the controller selects one of four signal phases for each intersection, and the nine intersection-level phase selections are combined into a single joint action controlled by the centralized policy. Each PPO transition corresponds to one simulation step, with the environment advancing until at least one intersection is ready for a control decision. The victim controller is trained using the pressure-based reward, computed from the aggregate intersection pressure across the network.

4.2. Results

Figure 1 exhibits the training convergence of the centralized network-level traffic signal controller under the no-attack baseline scenario. The team return represents the cumulative network-level reward associated with the nine signalized intersections, which collectively constitute the cooperative team under the centralized signal control formulation. As defined in section 3.1, the reward is pressure-based and is formulated as the negative of the aggregate traffic pressure across the intersections, where intersection pressure is a non-negative measure of traffic imbalance between incoming and corresponding downstream traffic conditions. The increasing trend observed during training therefore indicates that the centralized signal controller progressively learns coordinated signal decisions that improve network traffic conditions by reducing aggregate traffic pressure. The simultaneous reduction in return variability and the subsequent stabilization of the learning trajectory indicate that the controller reaches a relatively stable operating regime under non-adversarial state observations. This converged no-attack signal control policy establishes the nominal performance reference for the subsequent evaluation of adversarial vulnerability and robustness.

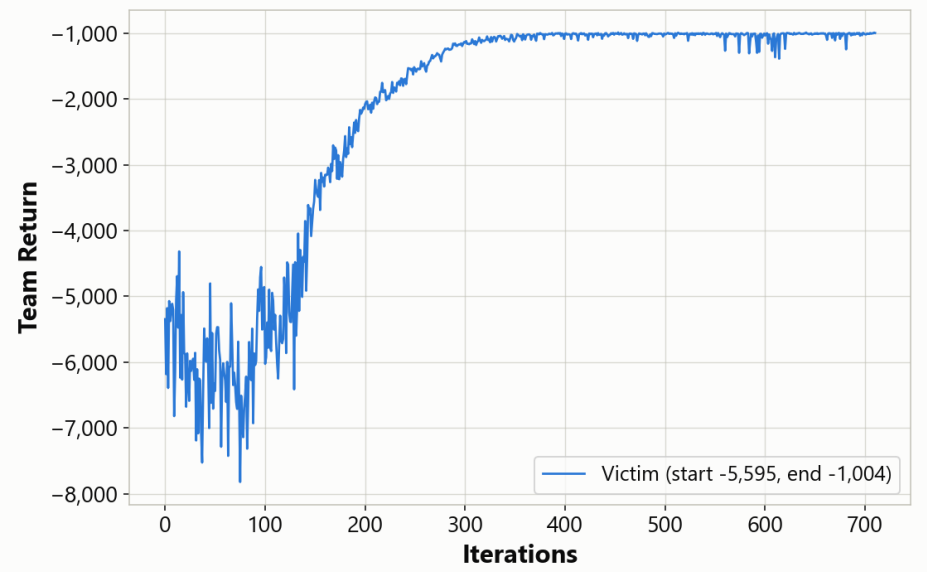

Figure 1. Training convergence of the centralized network-level traffic signal controller under the no-attack baseline scenario

Figure 2 presents the evolution of the attack-defense interaction during ATLA training under both the blackbox and whitebox attack settings. The coupled training dynamics reflect the two-player zero-sum max-min structure of the adversarially robust network-level signal control problem, in which the adversary is trained against a temporarily fixed victim policy and the victim is subsequently trained against the fixed adversary. In both attack settings, the victim signal controller's return improves substantially during the early alternations and subsequently approaches a comparatively stable level, while the adversarial return evolves correspondingly according to its opposing objective. The substantial changes during the initial alternations indicate rapid adaptation of both players, whereas the progressively smaller changes observed later provide empirical evidence of equilibration of the attack-defense max-min game. This behavior is observed under both policy-agnostic blackbox and policy-aware whitebox adversarial information settings despite the stronger policy access available to the whitebox attacker. At the stabilized regime, the defended controller reaches team returns of approximately -1,139 and -1,131 under blackbox and whitebox attacks, respectively, compared with nearly -1,004 for the no-attack baseline. Since the reward is the negative of network pressure, these differences correspond to approximately 13.4% and 12.6% increases in cumulative network pressure. The moderate increase in traffic pressure reflects the traffic system performance sacrificed during robust policy learning to reduce the controller's vulnerability to state observation perturbations. This establishes a fundamental tradeoff between the loss in nominal traffic control performance required to obtain adversarial robustness, which is further examined quantitatively in Figure 3.

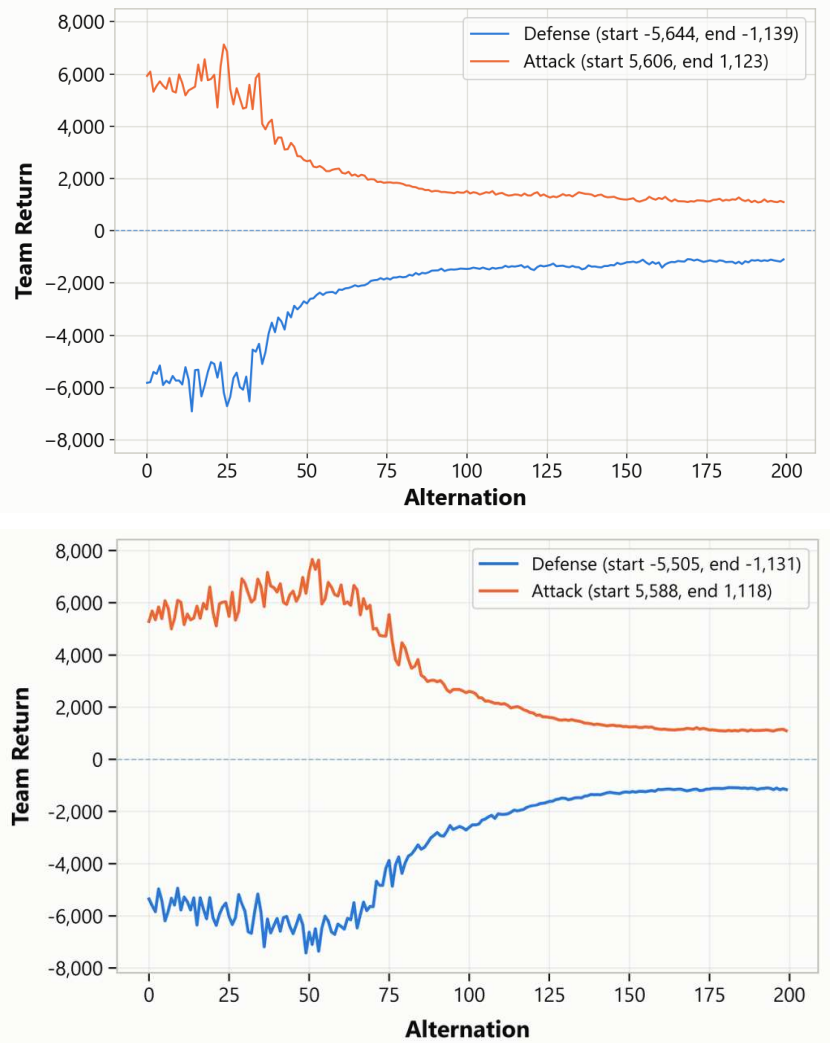

Figure 2. Empirical equilibration of the attack-defense game during adversarial training under blackbox attack (top) and whitebox attack (bottom)

Figure 3 quantifies the performance-robustness tradeoff through post-training evaluation of the nominal and robustified network traffic signal controllers. For each scenario, the corresponding trained controller is frozen and evaluated over 100 episodes, with the reported team return averaged across these episodes. The evaluation returns differ slightly from the corresponding training values shown in Figure 1 and Figure 2 since they are obtained by averaging 100 post-training evaluation episodes. The relevant comparison in Figure 3 is therefore the relative performance across the no-attack scenario (represented by the leftmost bar), the attacked-with-defense scenarios (represented by the middle bars), and the attacked-without-defense scenarios (represented by the rightmost bars) under the same evaluation settings. Relative to the no-attack baseline, the middle bars show that robustification incurs only approximately 6% and 7% performance degradation under blackbox and whitebox attacks, respectively. In contrast, without defense, the corresponding performance degradation increases substantially to approximately 50% and 73%, demonstrating the vulnerability of the nominal signal controller to test-time state observation attacks. The markedly greater performance degradation under the whitebox attack is consistent with its stronger adversarial information setting, in which the attacker has access

to the fixed victim signal controller's policy and can explicitly exploit this information when constructing adversarial state observations.

The contrast between the middle and rightmost bars in Figure 3 demonstrates the resulting performance-robustness tradeoff more directly. Relative to the attack-induced performance loss observed without defense, robustification eliminates approximately 89% of the performance degradation under the blackbox attack and 91% under the whitebox attack. Thus, the moderate residual performance loss of approximately 6% to 7% under attack is accompanied by a substantially greater gain in robustness compared with the undefended controller. This benefit is particularly pronounced under the stronger whitebox attack, where the undefended controller exhibits the greatest degradation while the robustified controller remains comparatively close to nominal performance. Since all policies are frozen during evaluation, the observed improvement reflects robustness acquired through adversarial training rather than online adaptation to the test-time attacker.

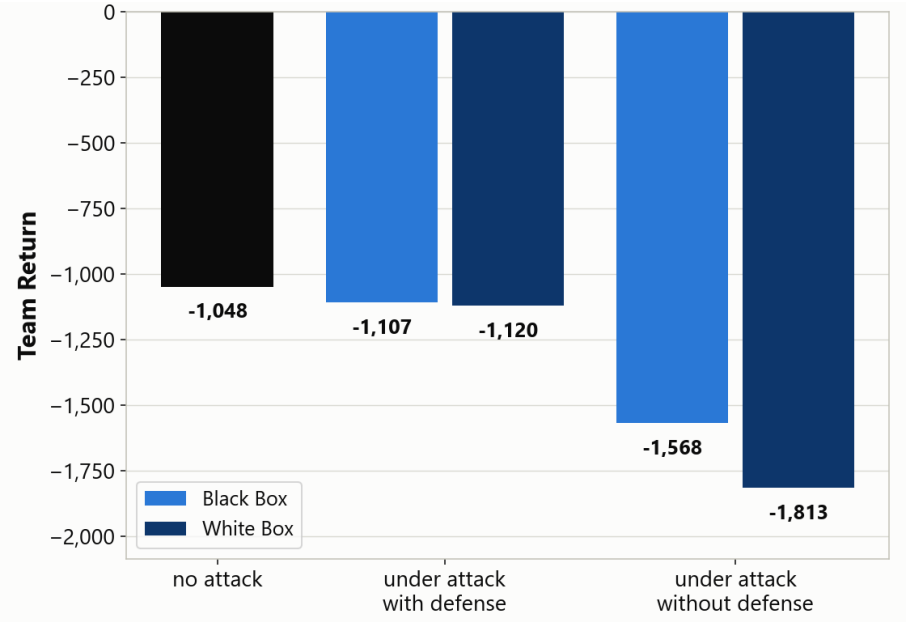

Figure 3. Post-training evaluation of the performance-robustness tradeoff under blackbox and whitebox state observation attacks

Figure 4 provides a qualitative visualization of how state observation attacks translate into physical traffic network performance degradation. The upper panel shows the no-attack baseline, whereas the lower panel shows the traffic state under the whitebox attack without defense. Although the adversary perturbs only the controller's observed state and does not directly modify the physical traffic state, the corrupted observations can alter the joint signal decisions selected by the signal controller. These altered actions subsequently affect vehicle discharge, queue propagation, and the evolution of the true traffic state over time. The greater vehicle accumulation visible across several intersection approaches in the attacked case therefore illustrates the sequential mechanism through which observation perturbations can propagate into network-level congestion. This qualitative evidence is consistent with the substantial performance degradation observed for the undefended whitebox condition in Figure 3 and demonstrates that the attack-induced reduction in team return corresponds to a physically meaningful deterioration of traffic conditions rather than merely a change in the controller's observed state.

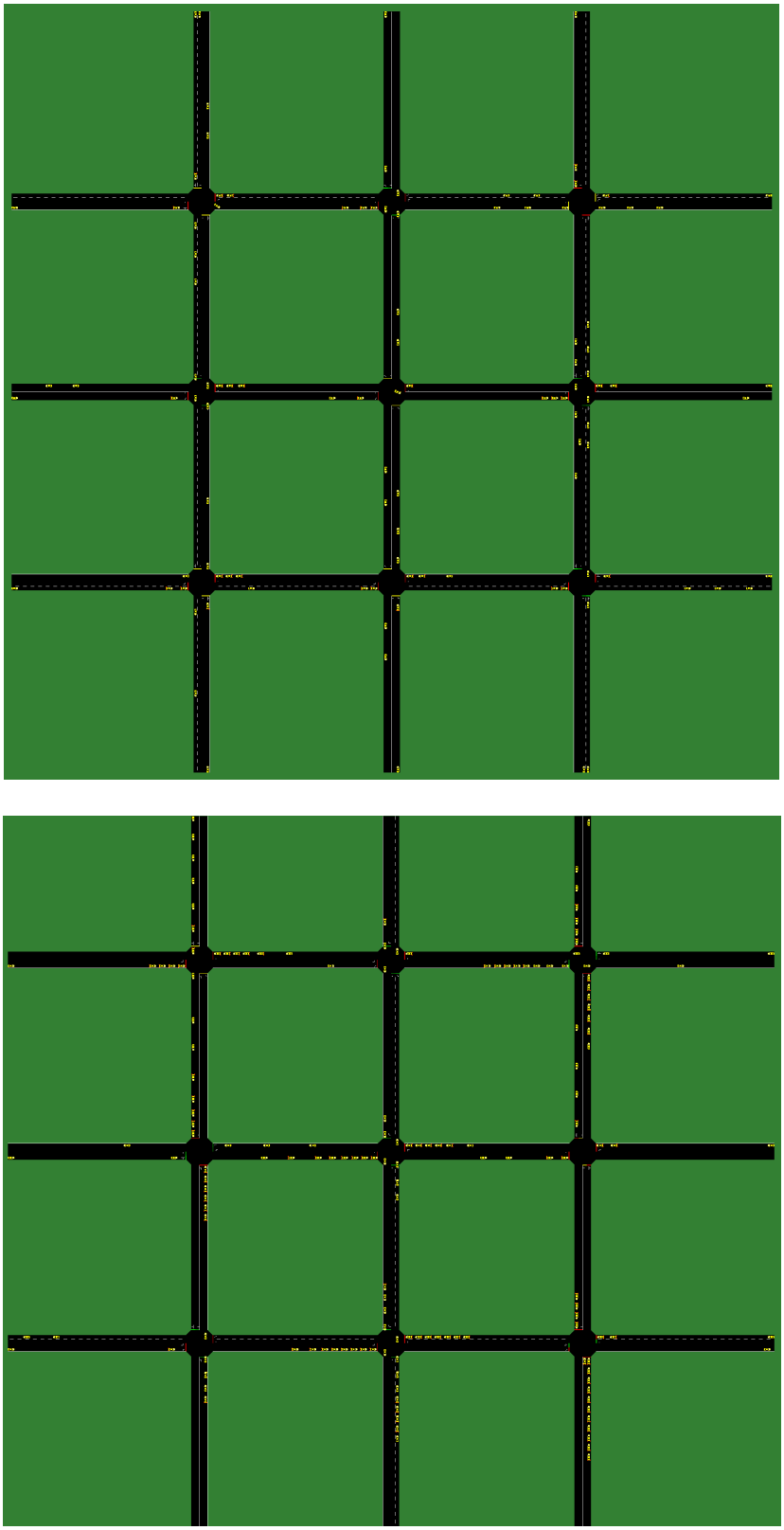

Figure 4. Traffic network snapshots under the no-attack baseline (top) and the undefended whitebox attack (bottom)

Figure 5 examines the sensitivity of test-time performance to the allowable adversarial state observation perturbation budget under blackbox and whitebox attacks. In both attack settings, increasing the perturbation budget enlarges the set of admissible adversarial observations and consequently provides the attacker with greater ability to influence the signal control decisions of the fixed controller. The resulting degradation is distinctly nonlinear. Under the blackbox attack, the evaluation return remains comparatively stable at smaller perturbation budgets before deteriorating progressively as the budget increases, indicating that limited perturbations can be tolerated before their effects become increasingly disruptive. The whitebox attack exhibits substantially greater sensitivity, with performance deteriorating sharply once the perturbation budget reaches an intermediate range and remaining severely degraded at larger budgets. This pronounced traffic network performance deterioration is consistent with the stronger information available to the policy-aware whitebox attacker, which exploits access to the fixed victim signal controller's policy when constructing adversarial observations.

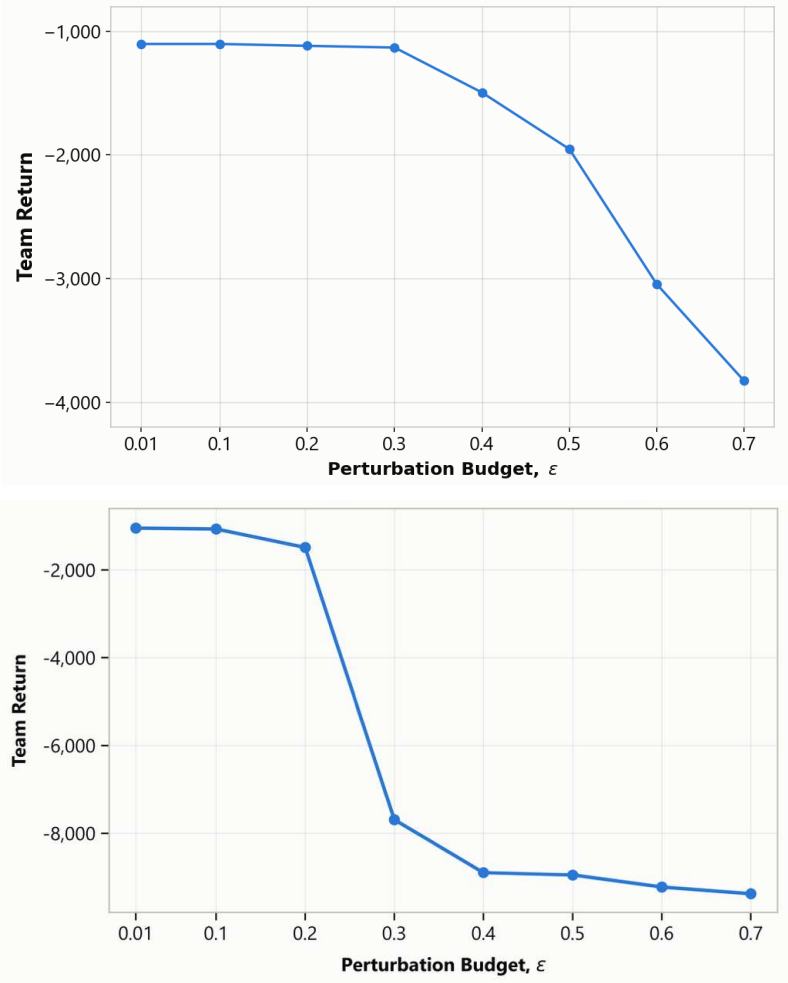

Figure 5. Sensitivity of post-training evaluation performance to the state observation perturbation budget under blackbox attack (top) and whitebox attack (bottom)

Figure 6 examines how traffic demand influences the vulnerability of the network-level traffic signal controller and the effectiveness of the defense. Under the no-attack baseline, increasing traffic demand generally increases traffic network pressure as the traffic system becomes more heavily loaded.

The presence of an adversarial attack introduces additional performance degradation whose magnitude varies across traffic demand levels. Under low demand, the defended signal controller remains relatively close to nominal performance and substantially outperforms the undefended signal controller, while under moderate traffic demand the difference between the defended and undefended conditions is comparatively smaller. The most pronounced effect occurs under high traffic demand, where the undefended adversarial condition reaches a team return of nearly -7,200, compared with approximately -5,000 with defense. This result indicates that attack-induced disruption can become particularly severe when the traffic network is heavily loaded and that robustification provides its largest absolute performance benefit under this demanding operating condition. Overall, the defense consistently improves performance relative to the undefended attacked signal controller across all evaluated traffic demand levels, although the magnitude of this benefit is not monotonic with traffic demand.

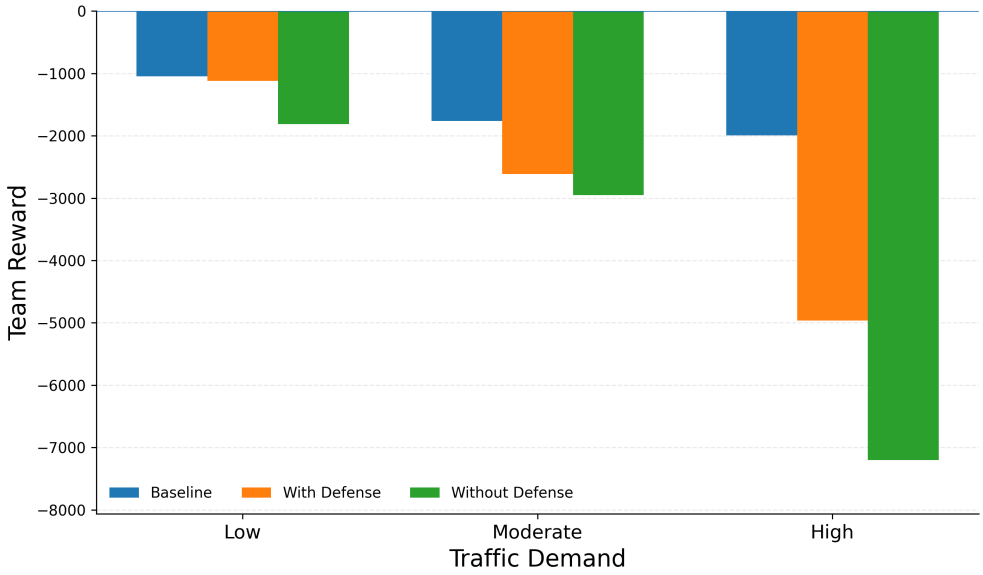

Figure 6. Sensitivity of post-training evaluation performance to traffic demand with and without defense under whitebox attack

Figure 7 examines the sensitivity of test-time attack vulnerability to traffic network size by comparing the no-attack baseline with the undefended signal controller under the whitebox PA-AD attack for the 3×3 and 4×4 traffic networks. Since the team return is a network-level quantity that naturally scales with the number of signalized intersections, the effect of network size is more appropriately assessed through the relative attack-induced degradation within each network rather than through direct comparison of absolute returns across network sizes. For the 3×3 network, the whitebox attack degrades the team return from approximately -1,048 without attack to -1,813 under attack, corresponding to an approximately 73% increase in cumulative network pressure. For the 4×4 network, the return deteriorates from approximately -2,137 to -5,415, corresponding to an approximately 153% increase in cumulative network pressure. This noticeably greater relative degradation in the larger network performance indicates that the undefended signal controller becomes more vulnerable to the evaluated whitebox attack as network size grows. This suggests that attack-induced disruption can become more pronounced in larger interconnected traffic networks, although results from two network sizes alone are insufficient to establish a general relationship between network scale and adversarial vulnerability.

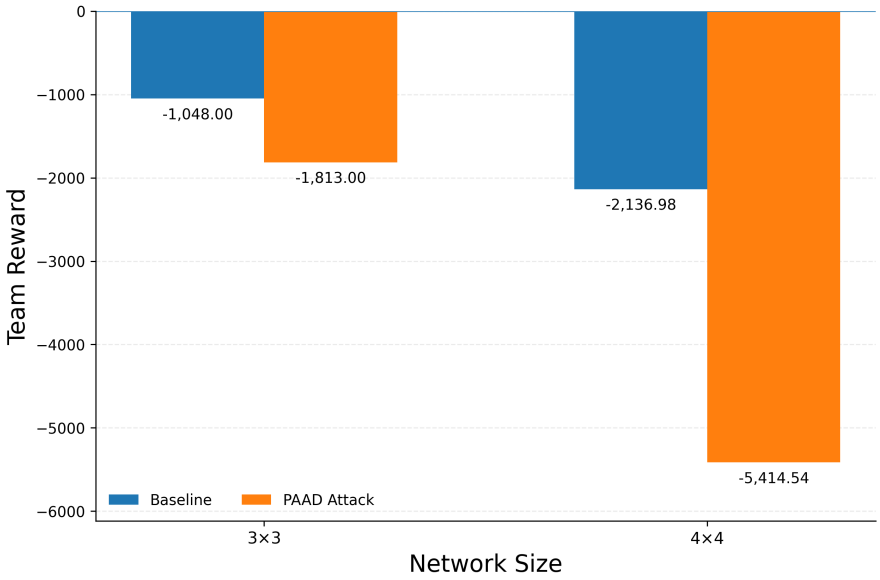

Figure 7. Sensitivity of undefended post-training evaluation performance to network size under the whitebox (PA-AD) attack

4.3. Discussion

The results presented in section 4.2 reveal that adversarial vulnerability in network-level traffic signal control is fundamentally a closed-loop and sequential phenomenon, rather than the consequence of isolated perturbations in individual state observations. Although the adversary directly perturbs only the observation presented to the signal controller, each perturbation can alter the joint signal action and thereby change vehicle movements and the subsequent ground-truth traffic state. The next adversarial perturbation and signal control decision therefore occur in a traffic state that has already been shaped by the consequences of preceding attack-induced actions. This temporal coupling allows the effects of successive perturbations to accumulate and propagate through the traffic network dynamics, producing sustained congestion even though the adversary never directly modifies the physical traffic state. The experimental sensitivities are consistent with this interpretation. Adversarial attack impact depends not only on perturbation magnitude, but also on the attacker's information and the prevailing traffic regime. In particular, the stronger traffic performance degradation produced by whitebox attacks demonstrates the importance of policy knowledge in exploiting the closed-loop response of the signal controller, while the nonlinear sensitivity to perturbation budget indicates that progressively larger observation errors need not translate into proportionally larger performance losses. The dependence on traffic demand and network size further suggests that adversarial vulnerability emerges from the interaction between the cyberattack, the learned signal control policy, and the physical traffic dynamics, rather than from observation corruption considered in isolation.

From the defense perspective, the findings characterize adversarial training as a robust signal control tradeoff rather than a mechanism for preserving nominal optimality. The max-min formulation exposes the signal controller during training to strategically adverse state observations so that the learned signal control policy becomes less sensitive to adversarial perturbations capable of inducing unfavorable traffic evolution. Some nominal traffic system performance is consequently sacrificed to

obtain robustness against test-time attacks, after which the signal controller remains fixed and performs no online adaptation. The significance of this performance-robustness tradeoff lies in its asymmetry. A comparatively moderate traffic performance sacrifice produces a substantially larger reduction in attack-induced performance degradation, mitigating approximately nine-tenths of the loss observed for the undefended controller. Additionally, the pronounced advantage of the whitebox attacker against the undefended signal controller is largely suppressed after robustification. Collectively, these findings indicate that adversarial training does not simply improve signalized traffic system performance under a particular attacked condition. Rather, it reshapes the learned signal control policy toward reduced sensitivity to strategically perturbed state observations and, consequently, greater stability of network-level performance under adversarial state observation uncertainty.

5. Conclusions

This research investigated the cybersecurity of reinforcement learning-based network-level traffic signal control under adversarial manipulation of traffic state observations. As learning-based signal controllers rely on observed traffic conditions to make real-time control decisions, deliberate corruption of these observations can influence signal actions and, through the resulting traffic dynamics, progressively degrade network performance. Understanding and mitigating this vulnerability therefore requires evaluating cyberattacks according to their cumulative consequences for traffic operations rather than only the immediate effects of individual observation perturbations. To this end, this research presented a cybersecurity assurance framework that integrates strategic state observation attacks and robust signal control within a common sequential decision setting. The framework considers policy-agnostic blackbox and policy-aware whitebox attacks to characterize different levels of attacker information and employs adversarial training to learn a robust signal control policy against strategically selected state observation perturbations before deployment.

The experimental findings from simulation in SUMO traffic simulator demonstrate that strategic state observation attacks can substantially degrade network-level traffic performance, with the policy-aware whitebox attack producing substantially greater performance degradation than the policy-agnostic blackbox attack against the undefended traffic signal controller. The undefended signal controller experienced approximately 50% performance degradation under the blackbox attack and 73% under the whitebox attack, while adversarial training reduced the residual performance degradation to approximately 6% and 7%, respectively. This corresponds to mitigating approximately nine-tenths of the attack-induced performance loss observed for the undefended signal controller, with a comparatively moderate reduction in nominal traffic performance. The sensitivity analyses further showed that attack severity varies nonlinearly with the perturbation budget and depends on traffic demand. The preliminary network-size analysis also showed greater relative performance degradation for the larger tested network, although the tested network configurations are insufficient to establish a general relationship between network scale and adversarial vulnerability. Collectively, the simulation results demonstrate that adversarial vulnerability in network-level traffic signal control is a closed-loop and sequential phenomenon arising from the interaction among attacker capability, the learned signal control policy, and the evolving traffic dynamics.

Our ongoing research is extending the present study in three directions. First, broader experimental evaluation is being conducted across more realistic network geometries, heterogeneous intersection characteristics, time-varying and uncertain traffic demand, additional network sizes, and

sensing and communication architectures representative of connected transportation systems. These experiments aim to further assess the generalizability of the observed relationships among attacker capability, traffic conditions, network characteristics, and adversarial robustness. Second, the cyberattack framework is being extended beyond the blackbox and whitebox settings considered in this study to intermediate information settings in which the adversary has partial or imperfect knowledge of the victim policy. This extension will enable adversarial effectiveness to be characterized across a broader spectrum of attacker knowledge. Third, the attack and defense framework is being extended from centralized network-level traffic signal control to decentralized traffic signal control. This extension will examine how state observation attacks propagate through interactions among distributed signal controllers and whether adversarial training can provide network-level robustness when control decisions are distributed across multiple policies.

References

Abdo, A., Chen, H., Zhao, X., Wu, G., Feng, Y., 2024. Cybersecurity on Connected and Automated Transportation Systems: A Survey. IEEE Transactions on Intelligent Vehicles 9, 1382–1401. https://doi.org/10.1109/TIV.2023.3326736.

Alvarez Lopez, P., Behrisch, M., Bieker-Walz, L., Erdmann, J., Flötteröd, Y.-P., Hilbrich, R., Lücken, L., Rummel, J., Wagner, P., Wießner, E., 2018. Microscopic Traffic Simulation using SUMO, in: 2019 IEEE Intelligent Transportation Systems Conference (ITSC). Presented at the The 21st IEEE International Conference on Intelligent Transportation Systems, IEEE, Maui, USA, pp. 2575–2582. https://doi.org/10.1109/ITSC.2018.8569938.

Bao, J., Zheng, L., Ban, X. (Jeff), 2023. Biobjective robust Network-wide traffic signal optimization against Cyber-attacks. Transportation Research Part C: Emerging Technologies 151, 104124. https://doi.org/10.1016/j.trc.2023.104124.

Chen, C., Wei, H., Xu, N., Zheng, G., Yang, M., Xiong, Y., Xu, K., Li, Z., 2020. Toward a thousand lights: Decentralized deep reinforcement learning for large-scale traffic signal control, in: AAAI Conference on Artificial Intelligence. pp. 3414–3421. https://doi.org/10.1609/aaai.v34i04.5744.

Feng, Y., Huang, S., Chen, Q.A., Liu, H.X., Mao, Z.M., 2018. Vulnerability of Traffic Control System Under Cyberattacks with Falsified Data. Transportation Research Record 2672, 1–11. https://doi.org/10.1177/0361198118756885.

Feng, Y., Huang, S.E., Wong, W., Chen, Q.A., Mao, Z.M., Liu, H.X., 2022. On the Cybersecurity of Traffic Signal Control System with Connected Vehicles. IEEE Transactions on Intelligent Transportation Systems 23, 16267–16279. https://doi.org/10.1109/TITS.2022.3149449.

FHWA, 2026. About Intersection Safety. Available online at: https://highways.dot.gov/safety/intersection-safety/about?utm_source=chatgpt.com.

Guo, Q., Li, L., Ban, X., 2019. Urban traffic signal control with connected and automated vehicles: A survey. Transportation Research Part C: Emerging Technologies 101, 313–334. https://doi.org/10.1016/j.trc.2019.01.026.

Haydari, A., Zhang, M., Chuah, C.-N., 2021. Adversarial Attacks and Defense in Deep Reinforcement Learning (DRL)-Based Traffic Signal Controllers. IEEE Open Journal of Intelligent Transportation Systems 2, 402–416. https://doi.org/10.1109/OJITS.2021.3118972.

Iyengar, G.N., 2005. Robust Dynamic Programming. Mathematics of Operations Research 30, 257–280. https://doi.org/10.1287/moor.1040.0129.

Kodi, J.H., Kitali, A.E., Ali, M.S., Alluri, P., Sando, T., 2021. Estimating Safety Impacts of Adaptive Signal Control Technology Using a Full Bayesian Approach. Transportation Research Record 2675, 1168–1179. https://doi.org/10.1177/03611981211025281.

Li, Z., Jin, D., Hannon, C., Shahidehpour, M., Wang, J., 2016. Assessing and mitigating cybersecurity risks of traffic light systems in smart cities. IET Cyber-Physical Systems: Theory & Applications 1, 60–69. https://doi.org/10.1049/iet-cps.2016.0017.

Lopez, A., Jin, W., Al Faruque, M.A., 2020. Security analysis for fixed-time traffic control systems. Transportation Research Part B: Methodological 139, 473–495. https://doi.org/10.1016/j.trb.2020.07.002.

Mo, Z., Li, W., Fu, Y., Ruan, K., Di, X., 2022. CVLight: Decentralized learning for adaptive traffic signal control with connected vehicles. Transportation Research Part C: Emerging Technologies 141, 103728. https://doi.org/10.1016/j.trc.2022.103728.

Nazari, F., Noruzoliaee, M., Zou, B., Mohammadian, A., 2017. Optimal Facility-Specific Inspection and Maintenance Decisions under 1 Measurement Uncertainty: A Unifying Framework. Journal of Infrastructure Systems 23(4), https://doi.org/10.1061/(ASCE)IS.1943-555X.0000402.

Nilim, A., Ghaoui, L., 2003. Robustness in Markov Decision Problems with Uncertain Transition Matrices, in: Advances in Neural Information Processing Systems.

Noaeen, M., Naik, A., Goodman, L., Crebo, J., Abrar, T., Abad, Z.S.H., Bazzan, A.L., Far, B., 2022. Reinforcement learning in urban network traffic signal control: A systematic literature review. Expert Systems with Applications 199, 116830. https://doi.org/10.1016/j.eswa.2022.116830.

Noruzoliaee, M., Nazari, F., 2025. Cyber Resilience of Connected and Autonomous Transportation Systems (Phase I): State-of-the-Art and Research Gaps. Technical Report. Available online at: https://rosap.ntl.bts.gov/view/dot/86379/dot_86379_DS1.pdf.

Perrine, K.A., Levin, M.W., Yahia, C.N., Duell, M., Boyles, S.D., 2019. Implications of traffic signal cybersecurity on potential deliberate traffic disruptions. Transportation Research Part A: Policy and Practice 120, 58–70. https://doi.org/10.1016/j.tra.2018.12.009.

Qu, A., Tang, Y., Ma, W., 2023. Adversarial Attacks on Deep Reinforcement Learning-based Traffic Signal Control Systems with Colluding Vehicles. ACM Transactions on Intelligent Systems and Technology 14(6), 113:1-113:22. https://doi.org/10.1145/3625236.

Ren, Y., Zhang, H., Cao, X., Yang, C., Zhang, J., Li, H., 2024a. Promoting or Hindering: Stealthy Black-Box Attacks Against DRL-Based Traffic Signal Control. IEEE Internet of Things Journal 11, 5816–5825. https://doi.org/10.1109/JIOT.2023.3308260.

Ren, Y., Zhang, H., Du, L., Zhang, Z., Zhang, J., Li, H., 2024b. Stealthy Black-Box Attack with Dynamic Threshold Against MARL-Based Traffic Signal Control System. IEEE Transactions on Industrial Informatics 20, 12021–12031. https://doi.org/10.1109/TII.2024.3413356.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O., 2017. Proximal Policy Optimization Algorithms. https://doi.org/10.48550/arXiv.1707.06347.

Sun, Y., Zheng, R., Liang, Y., Huang, F., 2021. Who is the Strongest Enemy? Towards Optimal and Efficient Evasion Attacks in Deep RL. https://doi.org/10.48550/arXiv.2106.05087.

Sutton, R.S., Barto, A.G., 2018. Reinforcement learning: An introduction, in: MIT Press, Cambridge, MA, USA. MIT Press, Cambridge, MA, USA.

Thodi, B.T., Mulumba, T., Jabari, S.E., 2020. Noticeability Versus Impact in Traffic Signal Tampering. IEEE Access 8, 86149–86161. https://doi.org/10.1109/ACCESS.2020.2992536.

Wei, H., Zheng, G., Gayah, V., Li, Z., 2020. A Survey on Traffic Signal Control Methods. https://doi.org/10.48550/arXiv.1904.08117.

Yen, C.-C., Ghosal, D., Zhang, M., Chuah, C.-N., 2022. Security Vulnerabilities and Protection Algorithms for Backpressure-Based Traffic Signal Control at an Isolated Intersection. IEEE Transactions on Intelligent Transportation Systems 23, 6406–6417. https://doi.org/10.1109/TITS.2021.3056658.

Zeng, J., Yu, C., Yang, X., Ao, W., Hao, Q., Yuan, J., Li, Y., Wang, Y., Yang, H., 2025. CityLight: A Neighborhood-inclusive Universal Model for Coordinated City-scale Traffic Signal Control, in: Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM '25. Association for Computing Machinery, New York, NY, USA, pp. 4036–4044. https://doi.org/10.1145/3746252.3761285.

Zhang, H., Chen, H., Boning, D., Hsieh, C.-J., 2021. Robust Reinforcement Learning on State Observations with Learned Optimal Adversary. https://doi.org/10.48550/arXiv.2101.08452.

Zhang, H., Chen, H., Xiao, C., Li, B., Liu, M., Boning, D., Hsieh, C.J., 2020. Robust Deep Reinforcement Learning against Adversarial Perturbations on State Observations, in: Advances in Neural Information Processing Systems.

Zhang, H., Gu, J., Zhang, Z., Du, L., Zhang, Y., Ren, Y., Zhang, J., Li, H., 2023. Backdoor attacks against deep reinforcement learning based traffic signal control systems. Peer-to-Peer Networking and Applications 16, 466–474. https://doi.org/10.1007/s12083-022-01434-0.

Zheng, L., Bao, J., Mei, Z., 2023. Urban traffic signal control robust optimization against Risk-averse and Worst-case cyberattacks. Information Sciences 640, 119067. https://doi.org/10.1016/j.ins.2023.119067.