STEADY, AS SHE GOES: RETHINKING OBSERVATIONS AND REWARDS FOR STABILITY REGULATION IN COOPERATIVE MULTI-AGENT SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Continuing multi-agent systems must maintain stable operation and recover safely and promptly from disturbances. Previous works mainly formulate it based on instantaneous observations and tracking-error-type rewards, which may not be appropriate. This mismatch arises from the information available to the learner and the recovery behaviors favored by the reward. Without explicit disturbance context, value learning can conflate external effects with those of agents' actions. Missing temporal information can hinder action selection when the same instantaneous deviation occurs during recovery or renewed departure. Tracking-error style deviation penalties can also favor slower recovery over a safe, faster return that involves temporary overshoot. We study observation and reward grounding for continuing multi-agent regulation to align learning with stable-operation and recovery requirements. We propose an observation design that combines disturbance context available before action selection with windowed deviation and its change, providing information about external conditions and recent evolution. We also propose a recovery-oriented reward that uses time spent outside the normal operating region as the base cost and a potential difference as auxiliary recovery feedback, with safety and continued task performance as requirements. We organize the evaluation around MAPPO in three environments: cooperative payload transport, power-grid voltage control, and traffic signal control, with comparisons that isolate the roles of observation and reward design. The evaluation examines recovery success and time alongside normal operation, failures, and task performance to assess whether improved recovery preserves safe and effective operation.

1 INTRODUCTION

Multi-agent systems (MAS) have become integral to a wide range of real-world infrastructures, from power grids (Wang et al., 2021a; Chen et al., 2023; Lu et al., 2023; Yan & Xu, 2024) and intelligent transportation (Wang et al., 2021c; Antonio & Maria-Dolores, 2022; Zhou et al., 2022a; Shi et al., 2023; Mukhtar et al., 2023; Yadav et al., 2023) to financial markets (Shavandi & Khedmati, 2022; Huang & Tanaka, 2022; He et al., 2023; Huang et al., 2024) and industrial automation (Kim et al., 2020; Zhang et al., 2020a; Bahrpeyma & Reichelt, 2022; Bai et al., 2023; Panzer et al., 2024). As these systems increasingly rely on autonomous agents, their dynamic behaviors are largely shaped by agent-level interactions, decisions, and adaptations. This autonomy, while enabling flexibility and scalability, also brings challenges to system-wide reliability. Unexpected deviations in agent states can cascade, destabilize local subsystems, and eventually cause large-scale operational failures if not properly regulated (Campbell & Lowry, 2012; Busby et al., 2021; Leslie, 2021; Shi et al., 2022a; Xu et al., 2021). For systems that must continue operating under disturbances, stability therefore requires both maintaining normal operation and restoring it safely and promptly after disruptions, while sustaining the underlying task.

MAPDN (Wang et al., 2021b), a multi-agent active voltage control benchmark, provides a concrete example of how such regulation tasks are modeled. Agents control inverter reactive power to maintain bus voltages within a prescribed operating range, using instantaneous measurements from their local region, including bus voltages, load power, and photovoltaic power. In operation, such systems may occasionally experience disturbances whose magnitude and duration are not fully known at their

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.

onset. Conventional observation designs do not necessarily represent these disturbances explicitly, even though events such as protection trips can be observed once they occur. The reward follows a tracking-error-type design: it penalizes instantaneous voltage deviations through a barrier function and includes an additional reactive-power cost, with L1 and L2 penalties among the voltage barrier forms investigated. Agents therefore learn from the system's response to external variations and optimize the cumulative reward associated with these stepwise costs. Instantaneous state-based observations and error-tracking style reward shapes are common modeling choices in learning-based control (Han et al., 2020; Zhao et al., 2023; Shi et al., 2022b; Tabas & Zhang, 2022; Yuan et al., 2022; Westenbroek et al., 2023; Qu et al., 2024), MAPDN serves as an illustrative example here.

However, these designs are not fully aligned with the requirements of such tasks. First, without explicit observations of disturbances, value learning can conflate their effects with those of agents' actions, obscuring the attribution of observed returns. Second, the same instantaneous deviation can occur while the system is returning toward normal operation or moving away from it, leaving the policy without sufficient temporal context to select an appropriate action. Third, objectives that accumulate deviation penalties can penalize temporary overshoot even when it belongs to a safe, faster recovery. Such objectives can consequently favor a slower return with smaller deviations over a faster return with a larger but safe excursion.

To address these issues, we study observation and reward grounding for continuing multi-agent regulation. We incorporate external-condition context μ_t available before action selection into the learning inputs. We further summarize recent behavior through windowed deviation D_t and its change ΔD_t, providing trend information. For reward design, a base cost measures time spent outside the normal operating region, with safety and continued task performance as requirements, while a potential difference provides auxiliary feedback toward recovery. The evaluation is organized around three representative environments: *MultiWalker* (Gupta et al., 2017), where agents transport a shared payload; *Distributed Power Grid* (Wang et al., 2021b), where agents regulate voltage under fluctuating loads; and *TrafficSignal* (Krajzewicz et al., 2012; Alegre, 2019; Ault & Sharon, 2021), where decentralized controllers regulate congestion. These settings provide a basis for examining the separate roles of external context, temporal information, and recovery-oriented rewards through normal-operation, recovery, failure, and task-performance measures.

The contributions of this paper are summarized as follows:

- We formulate continuing multi-agent regulation around the observation and reward requirements for maintaining normal operation and achieving safe, timely recovery under disturbances.
- We develop a grounding design that combines decision-time external context and windowed deviation trends with an abnormal-time cost and auxiliary potential-based feedback.
- We establish a MAPPO-based evaluation across three cooperative control environments to examine the contributions of observation and reward design to recovery and continued task performance.

2 Related Work

Stability in Multi-Agent Systems. Stability has long been a core concern in distributed control and cooperative robotics (Wang et al., 2021a; Li et al., 2024), often formulated as system convergence, consensus, or regulation. Classical approaches design control laws under known system dynamics to satisfy Lyapunov or frequency-domain stability conditions. However, they assume strong structural priors, limiting their applicability to complex, black-box, or data-driven multi-agent environments. In the multi-agent research, most efforts center around optimizing task-level objectives such as win rate (Samvelyan et al., 2019; Kurach et al., 2020; Zheng et al., 2018), episodic return (Peng et al., 2021; Towers et al., 2024), or exploration coverage (Zheng et al., 2021; Hao et al., 2023), without directly modeling long-term dynamic stability. Our work differs by treating steady-state regulation as an explicit optimization goal. We propose a generalizable formulation that integrates smoothly into standard reinforcement learning pipelines.

Reward Design and Optimization in MARL. Reward shaping is a widely adopted technique for guiding policy learning, especially under sparse or long-horizon tasks (Sutton et al., 1998; Ma et al., 2023). In MARL, various intrinsic or auxiliary rewards have been introduced to encourage

cooperation (Foerster et al., 2016; Jaques et al., 2019), communication (Mao et al., 2020), or diverse exploration (Zheng et al., 2021; Hao et al., 2023). For stability regulation, a separate concern is whether the reward ranks recovery behaviors in accordance with operational needs: tracking-error-type costs can penalize temporary overshoot and favor a slower return over a safe, faster recovery. In this paper, we center the base reward on time spent outside the normal operating region, with safety and continued task performance as requirements. Potential differences provide auxiliary recovery feedback, building on existing work on Lyapunov-based cost shaping (Westenbroek et al., 2023).

Robustness and Disturbance Handling in MARL. Robust reinforcement learning has received increasing attention as real-world deployment requires resilience to uncertainty, disturbance, and noise. Prior works consider robustness under various disruptions, including action disturbances (Huang et al., 2017), reward corruption (Xu & Mannor, 2006; Xu et al., 2024), domain shift (van Marum et al., 2024), and dynamics uncertainty (van Marum et al., 2024; Pinto et al., 2017). AdvMARL (Li et al., 2025) presents a large-scale empirical study comprising over 82,620 experiments to evaluate cooperation, robustness, and resilience in MARL across 13 uncertainty types, and 15 hyperparameters. Among these, the most widely studied setting involves adversarial noise added to agent observations (Zhang et al., 2020b; Koh et al., 2020; Zhou et al., 2022b; Liu et al., 2023; Xiong et al., 2022), which degrades policy performance. In the multi-agent domain, robustness is further complicated by strategic interactions and decentralized perception (Hernandez-Leal et al., 2019; Ning & Xie, 2024). Methods adopt curriculum-based disturbance (Long et al., 2020; Agrawal et al., 2023; Jwa et al., 2023) or opponent modeling (He & Boyd-Graber, 2016; Shen & How, 2021) to improve behavior under uncertainty; whether the learning inputs sufficiently expose disturbance context and recent system evolution remains a separate design question.

3 Motivating Example

Consider a controlled spring–mass–damper system, $m\ddot{x} + c\dot{x} + kx = u + d(t)$, where u is the control force and $d(t)$ is an external load. The goal is to restore and maintain the normal region $|x| \leq \epsilon$ while respecting a wider safety envelope $|x| \leq x_{\max}$. Figure 1 illustrates three distinct requirements for learning this behavior.

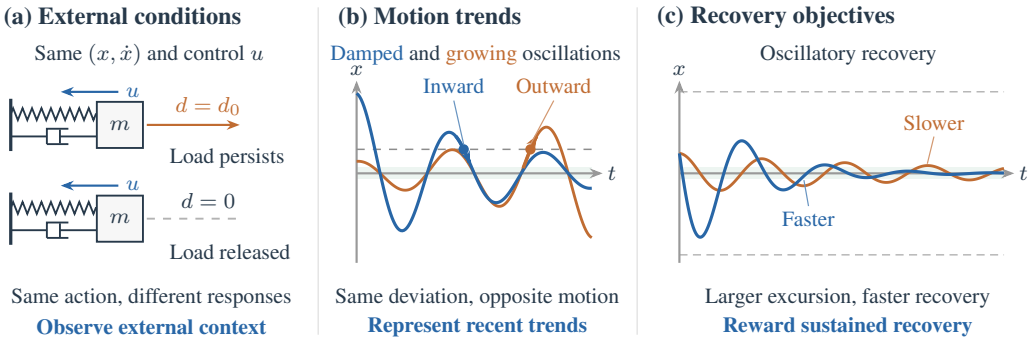


Figure 1: A spring example illustrating (a) external disturbances, (b) motion trends, and (c) safe overshoot during faster recovery.

External conditions affect action outcomes. At identical $(x, \dot{x})$ and u, a persistent load and a released load yield different accelerations. A value function without the available load context must average over these condition-dependent outcomes. This motivates conditioning learning on external signals observed before action selection.

Equal instantaneous deviations can require different responses. Figure 1(b) marks two equal positive displacements along growing and damped oscillations: one moves outward while the other returns inward. Position alone cannot distinguish them. A short history can supply trend information when velocity is absent from the observation, motivating a compact summary of recent deviation and its change.

Recovery time and accumulated deviation can disagree. Figure 1(c) sketches two oscillatory recoveries within the safety envelope. Smaller excursions can persist for longer, while a larger safe overshoot can accompany faster decay of subsequent oscillations. The faster trajectory safely overshoots the reference and subsequently remains in the normal region. They motivate an abnormal-time cost under task and safety requirements, with auxiliary feedback for recovery progress.

4 Problem Formulation

We model the shared environment as a cooperative multi-agent Markov decision process (MDP), $\mathcal{M} = (\mathcal{X}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{X}$ is the global state space, $\mathcal{A} = \prod_{i=1}^{N} \mathcal{A}_i$ is the joint action space of N agents, P is the transition kernel, r is the shared reward, and $\gamma \in [0, 1)$ is the discount factor. The global state x_t includes the external conditions needed to describe Markovian system evolution. Given a joint action $\mathbf{a}_t = (a_{1,t}, \ldots, a_{N,t})$, the next state follows $P(\cdot \mid x_t, \mathbf{a}_t)$. Each agent receives a local observation $o_{i,t}$ and acts according to its policy π_i; local observations need not reveal the full state. The joint policy π maximizes the expected discounted return $J(\pi) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r_t]$. We use centralized training with decentralized execution, so execution inputs are available through local sensing or explicitly shared monitoring signals.

The agents jointly regulate a task-critical system variable while continuing their task under external disturbances. Let $y_t \in \mathbb{R}$ denote this variable, $y^\star$ its reference, and ϵ_s the normal operating tolerance. We define the normal operating region and the excess deviation as

$$\mathcal{N} = \{y : |y - y^\star| \le \epsilon_s\}, \qquad e_t = \max\{|y_t - y^\star| - \epsilon_s, 0\}. \tag{1}$$

The regulated variable is observable through local observations or shared monitoring measurements. The excess deviation e_t is zero inside the normal region and measures the extent of a departure outside it. Task-specific safety limits can admit transient excursions beyond $\mathcal{N}$, allowing a safe overshoot during recovery.

The stability regulation objective is to maintain operation within $\mathcal{N}$ and shorten departures caused by disturbances while satisfying safety and continuing-task requirements. Repeated departures contribute to the total time outside the normal region over a common evaluation horizon. Section 5.3 instantiates this objective through a recovery-oriented reward.

5 Methodology

We propose STAB, a framework for stability regulation in cooperative multi-agent systems through observation and reward design. Stability regulation requires learning from both the conditions that drive a response and the way that response unfolds over time. The three contrasts in Figure 1 motivate three modeling choices in STAB: exposing disturbance conditions, representing the amplitude and trend of recent deviations, and rewarding timely recovery.

Figure 2 situates these three choices within the multi-agent reinforcement-learning loop.

5.1 Disturbance-Condition Context

A control action can produce different outcomes under different disturbance conditions. In Figure 1(a), the same spring position, velocity, and control force lead to different responses depending on whether the external load persists. When that condition is missing from the learning input, physically different situations share one representation. Their returns are then assigned to the same represented observation–action pair, mixing the consequences of control with those of the disturbance. For a fixed policy, this averaging takes the form $Q^\pi(o, a) = \mathbb{E}_{\mu|o,a}[Q^\pi(o, a, \mu)]$, where μ denotes the omitted disturbance condition and the conditional distribution describes the situations encountered by the learner. Making the observed condition explicit separates these value targets and gives action selection access to the condition under which regulation takes place.

This modeling choice also follows the sensing capabilities of instrumented systems. Smart sensors and networked monitoring provide electrical measurements and equipment-status information for grid monitoring and control (Song et al., 2017). We incorporate the disturbance-related measurements available before action selection as μ_t, such as the measured external load change. The actor receives

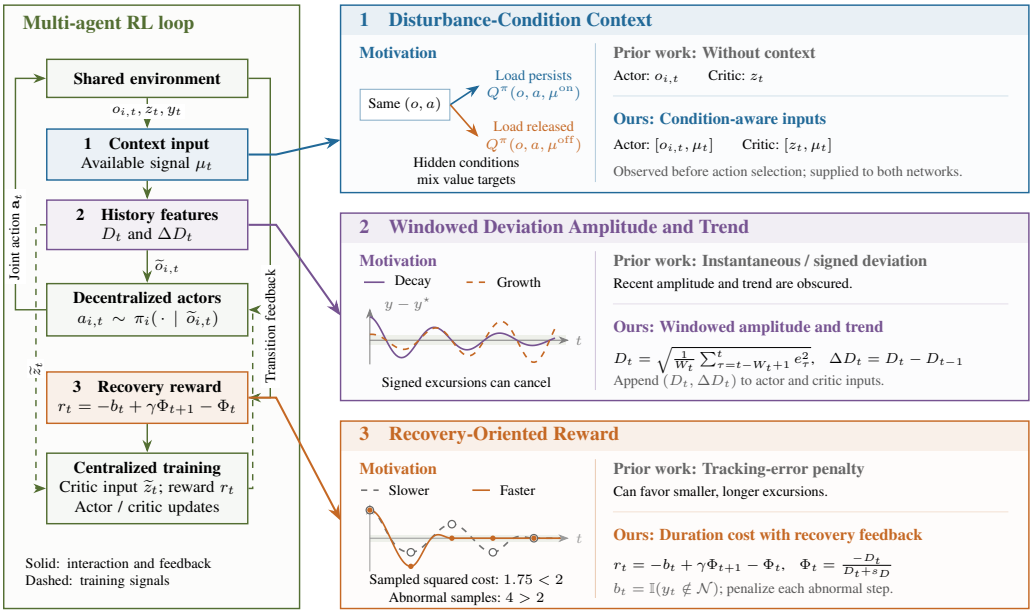


Figure 2: Overview of STAB, integrating disturbance context, windowed deviation features, and recovery-oriented reward into multi-agent reinforcement learning.

$[o_{i,t}, \mu_t]$, and the centralized critic receives $[z_t, \mu_t]$, where z_t is its original centralized input. Thus the physical condition that helps interpret a response is represented in both policy learning and value estimation.

5.2 Windowed Deviation Amplitude and Trend

In a system with oscillatory modes, recovery involves successive excursions on both sides of the operating point. An outward movement can belong to an oscillation whose amplitude is decreasing, as illustrated in Figure 1(b). Understanding recovery therefore requires observing the amplitude of recent excursions and how that amplitude changes. This perspective also applies to electromechanical oscillations in power networks, where control seeks to damp the interacting modes (Zenelis & Wang, 2018).

The first design choice is to preserve deviations on both sides of the reference. Averaging signed deviations $y_t - y^\star$ allows positive and negative excursions to cancel, so a large oscillation can have a mean close to zero. We instead average squared deviations over a recent window. Using the excess deviation e_t defined in Section 4, we compute

$$D_t = \sqrt{\frac{1}{W_t} \sum_{\tau=t-W_t+1}^{t} e_\tau^2}, \qquad W_t = \min\{W, t+1\}. \tag{2}$$

Squaring retains the contribution of excursions on either side, and taking the square root returns the resulting RMS amplitude to the units of the regulated variable. The tolerance band makes normal fluctuations contribute zero; the window summarizes the magnitude of recent departures across an oscillation.

The second design choice is to describe how that amplitude evolves. We introduce $\Delta D_t = D_t - D_{t-1}$, with $\Delta D_0 = 0$. A negative value indicates a decrease in the recent RMS deviation, while a positive value indicates an increase; its magnitude describes the change per observation step. Choosing a window around the relevant oscillation timescale smooths within-cycle fluctuations and makes the amplitude trend easier to observe. Together, D_t and ΔD_t describe the extent of recent departures and their progression toward recovery or further growth. We append these statistics to the original actor and centralized critic inputs:

$$\widetilde{o}_{i,t} = [o_{i,t}, \mu_t, D_t, \Delta D_t], \qquad \widetilde{z}_t = [z_t, \mu_t, D_t, \Delta D_t]. \tag{3}$$

The window is formed from available monitoring measurements and reset at episode boundaries.

5.3 Recovery-Oriented Reward

Oscillatory recovery can include a safe overshoot followed by a faster reduction in amplitude, as illustrated in Figure 1(c). Conventional regulation rewards use tracking-error penalties, with squared error as a representative choice. A squared tracking penalty $|y_t - y^\star|^2$ charges for the size of that overshoot as well as the duration of recovery. Consider two safe deviation sequences, $(1, 0.5, 0.5, 0.5, 0, \ldots)$ and $(1, -1, 0, \ldots)$, where the nonzero deviations lie outside the normal region. The first stays outside the band for four steps, while the second overshoots and returns after two. Yet their accumulated squared errors are $1 + 3 \times 0.5^2 = 1.75$ and $1 + (-1)^2 = 2$, respectively. In this example, minimizing the tracking cost therefore favors the slower recovery. Similar examples can also be constructed for first-order absolute tracking error; we use squared error here for ease of illustration.

We use a potential-difference reward to provide recovery feedback while preserving the optimal policies of a duration-based objective (Ng et al., 1999). The base reward is zero while the system remains in the normal region and -1 at each step outside it. Each abnormal step contributes its own cost to the return. Our reward is

$$r_t = \gamma \Phi_{t+1} - \Phi_t + \begin{cases} -1, & y_t \notin \mathcal{N}, \\ 0, & y_t \in \mathcal{N}. \end{cases} \tag{4}$$

where $\Phi_t = -D_t/(D_t + s_D)$ is a bounded potential and $s_D > 0$ sets its deviation scale. As the recent amplitude decreases, the potential approaches zero and supplies graded feedback during recovery. The piecewise term assigns the cost of abnormal operation, while the potential difference provides feedback on recovery progress.

We set the potential at the terminal state of each episode to zero. The potential terms then cancel in the discounted episodic return, leaving only the initial potential, whose distribution is fixed independently of the policy. The optimal policies therefore minimize the expected discounted count of abnormal steps within the episode. Appendix B provides the detailed derivation.

6 Experiment Design

Environments. To evaluate the effectiveness of STAB for steady-state regulation under disturbances, we conduct experiments on three cooperative multi-agent control platforms that involve dynamic environments and measurable system-level variables. **(1) Multi-Walker.** (Felten et al., 2024; Gupta et al., 2017) In this environment, multiple agents cooperatively carry a shared payload across uneven terrain, as illustrated in Figure 3(a). **(2) Distributed Power Grid.** (Wang et al., 2021b) We consider a decentralized voltage control task in a distribution network, where agents adjust reactive power generation to stabilize bus voltages under time-varying weather and load conditions, as shown in Figure 3(b). **(3) Urban Traffic Signal System.** (Krajzewicz et al., 2012; Alegre, 2019; Ault & Sharon, 2021) This task is based on a large-scale urban traffic simulation shown in Figure 3(c), where each agent controls a traffic signal at an intersection.

Environmental Disturbances. To assess steady-state performance under disturbances, we introduce exogenous perturbations that are commonly encountered in real-world control settings. Without loss of generality, we focus on the most representative disturbance type: *load variations*. Load variations are categorized into decreasing changes (Load_-), which reduce environmental load, and increasing changes (Load_+), which intensify system demand. These disturbances are injected during training and execution to evaluate the robustness and recovery behavior of learned policies.

Implementation. For MAPPO Vanilla and STAB, we search the MAPPO hyperparameter choices listed in Appendix D and select the checkpoint with the best training performance. For the MAPPO + AdvMARL training variant, we follow the training recommendations and hyperparameter settings reported in the original paper. Environment definitions and MAPPO hyperparameter settings are provided in Appendix C and Appendix D, respectively.

Baselines. We compare STAB with two MAPPO-based baselines. **MAPPO Vanilla** uses the original environment observations and rewards with the centralized-training-decentralized-execution

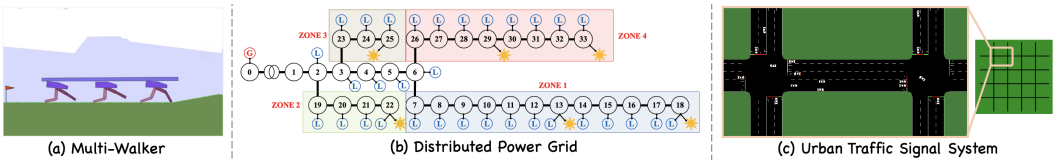


Figure 3: Example of different environments.

algorithm of Yu et al. (2022). **MAPPO + AdvMARL training** applies the robust training practices of Li et al. (2025) to MAPPO. The variants are evaluated under the same environment settings and disturbance scenarios.

Baseline Reward Shaping and Observation Design. As summarized in Appendix C, the baseline observation design in all three environments already reveals the regulated variable, or the signals needed to recover it, while the default reward design already includes an instantaneous regulation-related penalty. The MAPPO-based baselines use the original environment formulation for comparison.

Metrics and Task. In our selected environments, the task objective is intrinsically tied to the regulated variable: package stability in MultiWalker, voltage admissibility in PowerGrid, and congestion control in TrafficSignal. To avoid conflating improvements from reward redesign with policy quality, we evaluate all methods using the **Unstable Rate** ($\mathcal{UR}$) (Wang et al., 2021b). For each environment, we instantiate $\mathcal{UR}$ with the variable that directly characterizes regulation failure according to its task definition: in MultiWalker, $\mathcal{UR}$ is the permillage of events where the package angle exceeds $5°$; in PowerGrid, it is the permillage of events where voltage exceeds the predefined $[0.95p.u., 1.05p.u.]$ region; and in TrafficSignal, it is the percentage of events where the system's total waiting time exceeds 500 seconds.

Evaluation. Each trained policy is evaluated over 10 seeds, with 10 episodes per seed, and we report the mean and standard deviation across runs.

7 Results

7.1 Performance Under Different Disturbances

We first evaluate whether STAB improves long-term multi-agent coordination stability under different types of disturbances. Table 1 summarizes the performance comparison across three environments under decreasing and increasing loads. The most improvements are observed in the MultiWalker environment. Compared to the strongest baseline (MAPPO), STAB reduces the unstable rate by 24% under decreasing loads and by 29% under increasing loads, respectively. This environment poses substantial challenges for long-term coordination, where traditional MARL methods often fail after a few hundred steps due to payload imbalance or the agent falling when disturbances occur. In contrast, STAB maintains synchronized locomotion and stable payload control for extended horizons, demonstrating stronger resilience to both abrupt agent failures and sustained load perturbations. In the PowerGrid environment, STAB also outperforms baseline methods under load disturbances, achieving an reduction of 38% and 35% in unstable rate. Agents trained with STAB coordinate voltage regulation more coherently, effectively mitigating fluctuations induced by dynamic load changes and external perturbations. In the TrafficSignal environment. STAB reduces unstable rate by 27% and 24% over the best baseline, indicating its ability to prevent local failures from escalating into system-wide congestion.

Overall, these results show that STAB consistently reduces unstable rate across heterogeneous environments and multiple disturbance types, including load variations. While baseline methods may perform well under specific conditions, they lack robustness across diverse disturbance scenarios. By explicitly regulating system behavior across different operating regimes, STAB sustains stable multi-agent coordination over long horizons. Case studies illustrating representative behaviors under different disturbances are provided in the Appendix.

Table 1: Complete-scheme comparison with MAPPO under decreasing and increasing loads. $\mathcal{UR}$ is reported in permillage for MultiWalker and PowerGrid and in percent for TrafficSignal; lower is better. Each cell shows the mean above its standard deviation. Parenthesized reductions are relative to MAPPO Vanilla.

	Multiwalker		**PowerGrid**		**TrafficSignal**	
Scenario	Load_-	Load_+	Load_-	Load_+	Load_-	Load_+
Metric	($\mathcal{UR}\downarrow$)	($\mathcal{UR}\downarrow$)	($\mathcal{UR}\downarrow$)	($\mathcal{UR}\downarrow$)	($\mathcal{UR}\downarrow$)	($\mathcal{UR}\downarrow$)
MAPPO Vanilla	15.11 $\pm$1.28	2.34 $\pm$0.63	4.24 $\pm$0.31	5.85 $\pm$1.53	17.18 $\pm$1.92	22.03 $\pm$6.84
MAPPO + AdvMARL training	21.33 $\pm$2.89	2.94 $\pm$0.88	4.67 $\pm$0.67	7.32 $\pm$2.49	25.12 $\pm$2.06	32.28 $\pm$8.72
MAPPO + proposed grounding	**11.45 $\pm$0.92 (-24.22%)**	**1.66 $\pm$0.45 (-29.06%)**	**2.63 $\pm$0.21 (-37.97%)**	**3.81 $\pm$1.02 (-34.87%)**	**12.52 $\pm$1.50 (-27.12%)**	**16.64 $\pm$4.83 (-24.47%)**

Table 2: Ablation of `STAB` with MAPPO. The three intermediate variants retain the full reward; the final row restores the original observations and environment reward. Frame stack supplies recent measurements of y_t. Mean $\mathcal{UR}$ is reported in the units of Table 1; standard deviations are provided in Appendix Table 7.

	Multiwalker		**PowerGrid**		**TrafficSignal**	
Scenario	Load_-	Load_+	Load_-	Load_+	Load_-	Load_+
`STAB`	**11.45**	**1.66**	**2.63**	**3.81**	**12.52**	**16.64**
w/o μ_t	13.84	2.11	3.72	5.02	15.61	20.28
w/o $D_t, \Delta D_t$	14.76	2.28	4.15	5.63	16.82	21.58
w/o $D_t, \Delta D_t$; w/ frame stack	13.28	1.98	3.46	4.72	14.91	19.46
w/o $\mu_t, D_t, \Delta D_t$, reward (MAPPO Vanilla)	15.11	2.34	4.24	5.85	17.18	22.03

7.2 Ablation Study

Table 2 compares the full method, three observation ablations, and MAPPO Vanilla. The full method uses disturbance context μ_t, windowed deviation features $D_t, \Delta D_t$, and the recovery-oriented reward in Equation (4).

The first ablation removes only μ_t from both actor and critic inputs. This comparison tests whether explicit disturbance context contributes to stability regulation. Removing μ_t increases the unstable rate by 20.9%–41.4% across the six environment–load settings, supporting the benefit of exposing external conditions to the learner.

The second ablation removes $D_t, \Delta D_t$ from both actor and critic inputs. This comparison tests whether the amplitude and trend of recent deviations provide useful information beyond the instantaneous observation and disturbance context. The unstable rate increases by 28.9%–57.8% relative to the full method, indicating that temporal information contributes to stable regulation in all tested settings.

The third ablation replaces $D_t, \Delta D_t$ with a frame stack of recent measurements of y_t. This comparison tests whether the structured deviation features offer an advantage over supplying raw measurement history. Frame stacking improves on omitting temporal features in every setting, but its unstable rate remains 16.0%–31.6% above the full method. These results support using windowed deviation amplitude and trend as a compact representation of regulation-relevant history.

The final row removes all added observation features and restores the original environment reward, yielding MAPPO Vanilla, with values taken from Table 1. This comparison assesses the combined contribution of the observation and reward designs. The full method reduces the unstable rate by 24.2%–38.0% relative to MAPPO Vanilla across the six settings. Together, the ablations support the complementary contributions of disturbance context and structured temporal features.

Table 3: Results of varying numbers of agents

Environment	Scale	MAPPO		STAB	
		Load_- ($\mathcal{UR}\downarrow$)	Load_+ ($\mathcal{UR}\downarrow$)	Load_- ($\mathcal{UR}\downarrow$)	Load_+ ($\mathcal{UR}\downarrow$)
MultiWalker	n=2	33.39 ± 2.41	4.10 ± 0.91	26.41 ± 1.88 **(-21%)**	3.18 ± 0.70 **(-22%)**
	n=3	15.11 ± 1.28	2.34 ± 0.63	11.45 ± 0.92 **(-24%)**	1.66 ± 0.45 **(-29%)**
	n=5	12.48 ± 1.05	1.97 ± 0.55	9.18 ± 0.76 **(-26%)**	1.39 ± 0.37 **(-29%)**
	n=8	10.14 ± 0.88	1.64 ± 0.47	7.95 ± 0.71 **(-22%)**	1.22 ± 0.32 **(-26%)**
PowerGrid	n=33	4.24 ± 0.31	5.85 ± 1.53	2.63 ± 0.21 **(-38%)**	3.81 ± 1.02 **(-35%)**
	n=141	12.63 ± 0.94	17.26 ± 4.21	8.54 ± 0.69 **(-32%)**	12.28 ± 2.98 **(-29%)**
TrafficSignal	n=4	28.19 ± 2.74	35.41 ± 9.02	22.11 ± 1.79 **(-22%)**	27.95 ± 6.88 **(-21%)**
	n=16	17.18 ± 1.92	22.03 ± 6.84	12.52 ± 1.50 **(-27%)**	16.64 ± 4.83 **(-24%)**

7.3 Robustness to Agent Population

We evaluate the robustness of STAB under varying agent population sizes. Table 3 compares STAB with the strongest baseline, MAPPO, across 3 environments. In MultiWalker, increasing the number of agents directly amplifies physical coupling through the shared payload, making the system more sensitive to local miscoordination. As shown in Table 3, STAB consistently achieves lower instability rates than MAPPO across all tested scales ($n = 2, 3, 5, 8$), with relative reductions ranging from 21% to 29%.

The performance gap widens as the number of agents increases, indicating that MAPPO degrades more rapidly under higher coordination pressure. This suggests that STAB's stability-oriented modeling effectively mitigates cascading deviations caused by individual agent errors, allowing the system to remain stable even when collective dynamics become more complex.

Similar trends are observed in PowerGrid and TrafficSignal environments. Across different numbers of traffic lights and power nodes, STAB consistently achieves higher stability rates than MAPPO, with improvements of 21 ~38%.

Overall, results show that STAB generalizes well to varying agents.

8 Conclusion

This paper presents STAB, a framework for continuing cooperative multi-agent stability regulation through observation and reward design. STAB combines disturbance context available before action selection, windowed deviation amplitude and trend, and an abnormal-time cost with auxiliary potential-based recovery feedback. Experiments in cooperative payload transport, power-grid voltage control, and traffic signal control show lower unstable rates under load disturbances, and observation ablations support the contributions of disturbance context and structured temporal features. Across the six environment–load settings, the full method reduces unstable rates by 24.2%–38.0% relative to MAPPO. Raw frame stacking also improves on omitting temporal features, but remains less effective than the structured amplitude and trend representation. These findings highlight the value of grounding policy inputs in external conditions and recent system evolution.

The framework assumes that task-relevant variables and disturbance signals are observable during execution, and its evaluation is limited to three environments. Future work will examine broader coordination, communication, and partial-observability settings, as well as the applicability of these designs to single-agent reinforcement learning.

AI use statement

Generative AI tools were used to polish the writing of this manuscript and improve its readability. The authors reviewed the AI-assisted edits and take responsibility for the final content.

REFERENCES

Aakriti Agrawal, Rohith Chandrashekar Aralikatti, Yanchao Sun, and Furong Huang. Robustness to multi-modal environment uncertainty in marl using curriculum learning. *ArXiv*, abs/2310.08746, 2023.

Lucas N. Alegre. SUMO-RL. `https://github.com/LucasAlegre/sumo-rl`, 2019.

Guillen-Perez Antonio and Cano Maria-Dolores. Multi-agent deep reinforcement learning to manage connected autonomous vehicles at tomorrow's intersections. *IEEE Transactions on Vehicular Technology*, 71(7):7033–7043, 2022.

James Ault and Guni Sharon. Reinforcement learning benchmarks for traffic signal control. In *Proceedings of the Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS 2021) Datasets and Benchmarks Track*, December 2021.

Fouad Bahrpeyma and Dirk Reichelt. A review of the applications of multi-agent reinforcement learning in smart factories. *Frontiers in Robotics and AI*, 9:1027340, 2022.

Yue Bai, Yaqiong Lv, and Jiatong Zhang. Smart mobile robot fleet management based on hierarchical multi-agent deep q network towards intelligent manufacturing. *Engineering Applications of Artificial Intelligence*, 124:106534, 2023.

Joshua W Busby, Kyri Baker, Morgan D Bazilian, Alex Q Gilbert, Emily Grubert, Varun Rai, Joshua D Rhodes, Sarang Shidore, Caitlin A Smith, and Michael E Webber. Cascading risks: Understanding the 2021 winter blackout in texas. *Energy Research & Social Science*, 77:102106, 2021.

Richard J Campbell and Sean Lowry. Weather-related power outages and electric system resiliency. Congressional Research Service, Library of Congress Washington, DC, 2012.

Pengcheng Chen, Shichao Liu, Xiaozhe Wang, and Innocent Kamwa. Physics-guided multi-agent deep reinforcement learning for robust active voltage control in electrical distribution systems. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 71(2):922–933, 2023.

Florian Felten, Umut Ucak, Hicham Azmani, Gao Peng, Willem Röpke, Hendrik Baier, Patrick Mannion, Diederik M. Roijers, Jordan K. Terry, El-Ghazali Talbi, Grégoire Danoy, Ann Nowé, and Roxana Rădulescu. Momaland: A set of benchmarks for multi-objective multi-agent reinforcement learning, 2024. URL `https://arxiv.org/abs/2407.16312`.

Jakob Foerster, Ioannis Alexandros Assael, Nando De Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

Jayesh K Gupta, Maxim Egorov, and Mykel Kochenderfer. Cooperative multi-agent control using deep reinforcement learning. In *International Conference on Autonomous Agents and Multiagent Systems*, pp. 66–83. Springer, 2017.

Minghao Han, Lixian Zhang, Jun Wang, and Wei Pan. Actor-critic reinforcement learning for control with stability guarantee. *IEEE Robotics and Automation Letters*, 5(4):6217–6224, 2020. doi: 10.1109/LRA.2020.3011351. URL `https://arxiv.org/abs/2004.14288`.

Jianye Hao, Tianpei Yang, Hongyao Tang, Chenjia Bai, Jinyi Liu, Zhaopeng Meng, Peng Liu, and Zhen Wang. Exploration in deep reinforcement learning: From single-agent to multiagent domain. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Fei-Fan He, Chiao-Ting Chen, and Szu-Hao Huang. A multi-agent virtual market model for generalization in reinforcement learning based trading strategies. *Applied Soft Computing*, 134:109985, 2023.

He He and Jordan L. Boyd-Graber. Opponent modeling in deep reinforcement learning. *ArXiv*, abs/1609.05559, 2016.

Pablo Hernandez-Leal, Bilal Kartal, and Matthew E Taylor. A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 33(6):750–797, 2019.

Sandy H. Huang, Nicolas Papernot, Ian J. Goodfellow, Yan Duan, and P. Abbeel. Adversarial attacks on neural network policies. *ArXiv*, abs/1702.02284, 2017.

Yuling Huang, Chujin Zhou, Kai Cui, and Xiaoping Lu. A multi-agent reinforcement learning framework for optimizing financial trading strategies based on timesnet. *Expert Systems with Applications*, 237:121502, 2024.

Zhenhan Huang and Fumihide Tanaka. Mspm: A modularized and scalable multi-agent reinforcement learning-based system for financial portfolio management. *Plos one*, 17(2):e0263689, 2022.

Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z Leibo, and Nando De Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International conference on machine learning*, pp. 3040–3049. PMLR, 2019.

Younkyung Jwa, Minseon Gwak, Jiin Kwak, Chang Wook Ahn, and Poogyeon Park. Scalable robust multi-agent reinforcement learning for model uncertainty. *2023 62nd IEEE Conference on Decision and Control (CDC)*, pp. 3402–3407, 2023.

Yun Geon Kim, Seokgi Lee, Jiyeon Son, Heechul Bae, and Byung Do Chung. Multi-agent system and reinforcement learning approach for distributed intelligence in a flexible smart manufacturing system. *Journal of Manufacturing Systems*, 57:440–450, 2020.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etiene David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Meghan Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, 2020.

Daniel Krajzewicz, Jakob Erdmann, Michael Behrisch, Laura Bieker, et al. Recent development and applications of sumo-simulation of urban mobility. *International journal on advances in systems and measurements*, 5(3&4):128–138, 2012.

Karol Kurach, Anton Raichuk, Piotr Stańczyk, Michał Zając, Olivier Bachem, Lasse Espeholt, Carlos Riquelme, Damien Vincent, Marcin Michalski, Olivier Bousquet, et al. Google research football: A novel reinforcement learning environment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 4501–4510, 2020.

Mitch Leslie. Texas crisis highlights grid vulnerabilities. *Engineering*, 7(10):1348–1350, 2021.

Simin Li, Zihao Mao, Hanxiao Li, Zonglei Jing, Jun Guo, Li Wang, Zhuoran Han, Ruixiao Xu, Xin Yu, Chengdong Ma, et al. Empirical study on robustness and resilience in cooperative multi-agent reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Zhongyu Li, Xue Bin Peng, Pieter Abbeel, Sergey Levine, Glen Berseth, and Koushil Sreenath. Reinforcement learning for versatile, dynamic, and robust bipedal locomotion control. *The International Journal of Robotics Research*, pp. 02783649241285161, 2024.

Zuxin Liu, Zijian Guo, Zhepeng Cen, Huan Zhang, Jie Tan, Bo Li, and Ding Zhao. On the robustness of safe reinforcement learning under observational perturbations. In *11th International Conference on Learning Representations, ICLR 2023*, 2023.

Qian Long, Zihan Zhou, Abhibav Gupta, Fei Fang, Yi Wu, and Xiaolong Wang. Evolutionary population curriculum for scaling multi-agent reinforcement learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJxbHkrKDH.

Yu Lu, Yue Xiang, Yuan Huang, Bin Yu, Liguo Weng, and Junyong Liu. Deep reinforcement learning based optimal scheduling of active distribution system considering distributed generation, energy storage and flexible load. *Energy*, 271:127087, 2023.

Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Eureka: Human-level reward design via coding large language models. *arXiv preprint arXiv:2310.12931*, 2023.

Hangyu Mao, Zhibo Gong, and Zhen Xiao. Reward design in cooperative multi-agent reinforcement learning for packet routing. *arXiv preprint arXiv:2003.03433*, 2020.

Hamza Mukhtar, Adil Afzal, Sultan Alahmari, and Saud Yonbawi. Ccgn: Centralized collaborative graphical transformer multi-agent reinforcement learning for multi-intersection signal free-corridor. *Neural Networks*, 166:396–409, 2023.

Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the 16th International Conference on Machine Learning*, pp. 278–287, 1999. URL https://people.eecs.berkeley.edu/~pabbeel/cs287-fa09/readings/NgHaradaRussell-shaping-ICML1999.pdf.

Zepeng Ning and Lihua Xie. A survey on multi-agent reinforcement learning and its application. *Journal of Automation and Intelligence*, 3(2):73–91, 2024.

Marcel Panzer, Benedict Bender, and Norbert Gronau. A deep reinforcement learning based hyper-heuristic for modular production control. *International Journal of Production Research*, 62(8): 2747–2768, 2024.

Bei Peng, Tabish Rashid, Christian Schroeder de Witt, Pierre-Alexandre Kamienny, Philip Torr, Wendelin Böhmer, and Shimon Whiteson. Facmac: Factored multi-agent centralised policy gradients. *Advances in Neural Information Processing Systems*, 34:12208–12221, 2021.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Kumar Gupta. Robust adversarial reinforcement learning. In *International Conference on Machine Learning*, 2017.

Yang Qu, Jinming Ma, and Feng Wu. Safety constrained multi-agent reinforcement learning for active voltage control. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 184–192, 2024. doi: 10.24963/ijcai.2024/21. URL https://www.ijcai.org/proceedings/2024/21.

Mikayel Samvelyan, Tabish Rashid, Christian Schroeder de Witt, Gregory Farquhar, Nantas Nardelli, Tim G. J. Rudner, Chia-Man Hung, Philiph H. S. Torr, Jakob Foerster, and Shimon Whiteson. The StarCraft Multi-Agent Challenge. *CoRR*, abs/1902.04043, 2019.

Ali Shavandi and Majid Khedmati. A multi-agent deep reinforcement learning framework for algorithmic trading in financial markets. *Expert Systems with Applications*, 208:118124, 2022.

Macheng Shen and Jonathan P. How. Robust opponent modeling via adversarial ensemble reinforcement learning. In *International Conference on Automated Planning and Scheduling*, 2021.

Qingxin Shi, Wenxia Liu, Bo Zeng, Hongxun Hui, and Fangxing Li. Enhancing distribution system resilience against extreme weather events: Concept review, algorithm summary, and future vision. *International Journal of Electrical Power & Energy Systems*, 138:107860, 2022a.

Yang Shi, Zhenbo Wang, Tim J LaClair, Chieh Wang, Yunli Shao, and Jinghui Yuan. A novel deep reinforcement learning approach to traffic signal control with connected vehicles. *Applied Sciences*, 13(4):2750, 2023.

Yuanyuan Shi, Guannan Qu, Steven Low, Anima Anandkumar, and Adam Wierman. Stability constrained reinforcement learning for real-time voltage control. In *2022 American Control Conference (ACC)*, pp. 2715–2721, 2022b. doi: 10.23919/ACC53348.2022.9867476. URL https://arxiv.org/abs/2109.14854.

Yuyin Song, Gerald J. FitzPatrick, and Kang B. Lee. Smart sensors and standard-based interoperability in smart grids. *IEEE Sensors Journal*, 17(23), 2017. doi: 10.1109/JSEN.2017.2729893. URL https://www.nist.gov/publications/smart-sensors-and-standard-based-interoperability-smart-grids.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Daniel Tabas and Baosen Zhang. Computationally efficient safe reinforcement learning for power systems. In *2022 American Control Conference (ACC)*, pp. 3303–3310, 2022. doi: 10.23919/ACC53348.2022.9867652. URL https://arxiv.org/abs/2110.10333.

J Terry, Benjamin Black, Nathaniel Grammel, Mario Jayakumar, Ananth Hari, Ryan Sullivan, Luis S Santos, Clemens Dieffendahl, Caroline Horsch, Rodrigo Perez-Vicente, et al. Pettingzoo: Gym for multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34: 15032–15043, 2021.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, et al. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*, 2024.

Bart van Marum, Aayam Shrestha, Helei Duan, Pranay Dugar, Jeremy Dao, and Alan Fern. Revisiting reward design and evaluation for robust humanoid standing and walking. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 11256–11263. IEEE, 2024.

Jianhong Wang, Wangkun Xu, Yunjie Gu, Wenbin Song, and Tim C Green. Multi-agent reinforcement learning for active voltage control on power distribution networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 3271–3284. Curran Associates, Inc., 2021a. URL https://proceedings.neurips.cc/paper/2021/file/1a6727711b84fd1efbb87fc565199d13-Paper.pdf.

Jianhong Wang, Wangkun Xu, Yunjie Gu, Wenbin Song, and Tim C Green. Multi-agent reinforcement learning for active voltage control on power distribution networks. *Advances in neural information processing systems*, 34:3271–3284, 2021b.

Tong Wang, Jiahua Cao, and Azhar Hussain. Adaptive traffic signal control for large-scale scenario with cooperative group-based multi-agent reinforcement learning. *Transportation research part C: emerging technologies*, 125:103046, 2021c.

Tyler Westenbroek, Fernando Castaneda, Ayush Agrawal, Shankar Sastry, and Koushil Sreenath. Lyapunov design for robust and efficient robotic reinforcement learning. In *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 2125–2135. PMLR, 2023. URL https://proceedings.mlr.press/v205/westenbroek23a.html.

Zikang Xiong, Joe Eappen, He Zhu, and Suresh Jagannathan. Defending observation attacks in deep reinforcement learning via detection and denoising. In *ECML/PKDD*, 2022.

Huan Xu and Shie Mannor. The robustness-performance tradeoff in markov decision processes. In *Neural Information Processing Systems*, 2006.

Luo Xu, Qinglai Guo, Yujie Sheng, SM Muyeen, and Hongbin Sun. On the resilience of modern power systems: A comprehensive review from the cyber-physical perspective. *Renewable and Sustainable Energy Reviews*, 152:111642, 2021.

Yinglun Xu, Zhiwei Wang, and Gagandeep Singh. Robust thompson sampling algorithms against reward poisoning attacks. *arXiv preprint arXiv:2410.19705*, 2024.

Pamul Yadav, Ashutosh Mishra, and Shiho Kim. A comprehensive survey on multi-agent reinforcement learning for connected and automated vehicles. *Sensors*, 23(10):4710, 2023.

Rudai Yan and Yan Xu. Multi-objective and multi-agent deep reinforcement learning for real-time decentralized volt/var control of distribution networks considering pv inverter lifetime. *IEEE Transactions on Power Systems*, 2024.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in neural information processing systems*, 35:24611–24624, 2022.

Zhaocong Yuan, Adam W. Hall, Siqi Zhou, Lukas Brunke, Melissa Greeff, Jacopo Panerati, and Angela P. Schoellig. Safe-Control-Gym: A unified benchmark suite for safe learning-based control and reinforcement learning in robotics. *IEEE Robotics and Automation Letters*, 7(4):11142–11149, 2022. URL https://arxiv.org/abs/2109.06325.

Ilias Zenelis and Xiaozhe Wang. Wide-area damping control for interarea oscillations in power grids based on PMU measurements. *IEEE Control Systems Letters*, 2(4):719–724, 2018. doi: 10.1109/LCSYS.2018.2847758. URL https://arxiv.org/abs/2108.01193.

Chi Zhang, Philip Odonkor, Shuai Zheng, Hamed Khorasgani, Susumu Serita, Chetan Gupta, and Haiyan Wang. Dynamic dispatching for large-scale heterogeneous fleet via multi-agent deep reinforcement learning. In *2020 IEEE International Conference on Big Data (Big Data)*, pp. 1436–1441. IEEE, 2020a.

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. *Advances in Neural Information Processing Systems*, 33:21024–21037, 2020b.

Liqun Zhao, Konstantinos Gatsis, and Antonis Papachristodoulou. Stable and safe reinforcement learning via a Barrier-Lyapunov actor-critic approach. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pp. 1320–1325, 2023. doi: 10.1109/CDC49753.2023.10383742. URL https://arxiv.org/abs/2304.04066.

Lianmin Zheng, Jiacheng Yang, Han Cai, Ming Zhou, Weinan Zhang, Jun Wang, and Yong Yu. Magent: A many-agent reinforcement learning platform for artificial collective intelligence. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Lulu Zheng, Jiarui Chen, Jianhao Wang, Jiamin He, Yujing Hu, Yingfeng Chen, Changjie Fan, Yang Gao, and Chongjie Zhang. Episodic multi-agent reinforcement learning with curiosity-driven exploration. *Advances in Neural Information Processing Systems*, 34:3757–3769, 2021.

Wei Zhou, Dong Chen, Jun Yan, Zhaojian Li, Huilin Yin, and Wanchen Ge. Multi-agent reinforcement learning for cooperative lane changing of connected and autonomous vehicles in mixed traffic. *Autonomous Intelligent Systems*, 2(1):5, 2022a.

Ziyuan Zhou, Guanjun Liu, and Mengchu Zhou. A robust mean-field actor-critic reinforcement learning against adversarial perturbations on agent states. *IEEE Transactions on Neural Networks and Learning Systems*, 35:14370–14381, 2022b.