Software Engineering and Artificial Intelligence Systems

Volume: 02 Issue: 01 (2026)

www.seais.org/

Safety-Constrained Multi-Agent Reinforcement Learning for Coordinated UAV-UGV Control under Model and Delay Uncertainty

Arthur Webb
School of Computing Science, University of Glasgow, Glasgow G12 8RZ, UK

Key word: Multi-Agent Reinforcement Learning, Heterogeneous Autonomous Systems, Constrained Markov Decision Process, Delay Uncertainty, Safe Exploration.

A B S T R A C T

The cooperative deployment of Unmanned Aerial Vehicles and Unmanned Ground Vehicles has emerged as a transformative paradigm for complex field operations, offering complementary capabilities in mobility, payload capacity, and sensing range. However, designing effective control policies for such heterogeneous multi-agent systems remains profoundly challenging due to the inherent presence of communication delays, unpredictable environmental dynamics, and stringent safety requirements. Traditional control paradigms often fail to scale in high-dimensional state spaces or struggle to adapt to unmodeled disturbances. This paper proposes a novel Safety-Constrained Multi-Agent Reinforcement Learning framework explicitly designed to address coordinated Unmanned Aerial Vehicle and Unmanned Ground Vehicle control under simultaneous model and delay uncertainty. By reformulating the cooperative task as a Delay-Aware Constrained Markov Decision Process, we introduce a decentralized actor-centralized critic architecture that incorporates recurrent belief-state estimators to compensate for delayed state observations. Furthermore, to ensure strict adherence to safety constraints such as collision avoidance and energy limitations, we integrate a probabilistic Lagrangian optimization technique that dynamically adjusts penalty multipliers based on the estimated uncertainty of the transition dynamics. Extensive simulations in realistically disturbed environments demonstrate that the proposed framework significantly outperforms existing baselines, maintaining near-zero constraint violation rates while achieving superior task completion efficiency. The findings provide a robust foundation for deploying heterogeneous autonomous teams in communication-degraded and highly uncertain real-world environments.

1.INTRODUCTION

1.1 Motivation and Background

The integration of Unmanned Aerial Vehicles and Unmanned Ground Vehicles into cohesive, cooperative teams has garnered significant attention across various academic and industrial domains. Unmanned Aerial Vehicles possess exceptional three-dimensional mobility, expansive fields of view, and high-speed traversal capabilities, making them ideal for rapid scouting and wide-area surveillance [1]. Conversely, Unmanned Ground Vehicles are characterized by substantial payload capacities, extended operational endurance, and the ability to perform heavy-duty physical interventions [2]. When deployed synergistically, these

heterogeneous platforms can undertake highly complex missions, such as search and rescue operations in disaster-stricken zones, precision agricultural monitoring, and automated industrial inspection, which would be fundamentally impossible for either system operating in isolation [3]. In these cooperative scenarios, the aerial platform typically provides global situational awareness and pathfinding assistance, while the ground platform executes localized, resource-intensive tasks and acts as a mobile charging hub or data relay [4]. Despite the compelling advantages of heterogeneous robotic teams, their practical deployment is severely hindered by the intricacies of coordinating entities with vastly disparate kinematic properties and operational constraints [5]. Traditional optimal control strategies, such as Model Predictive Control, have been extensively applied to coordinated navigation tasks [6]. However, these classical methods heavily rely on the availability of highly accurate mathematical models of both the agents and the environment. In real-world field operations, such idealized models are rarely obtainable. Autonomous systems are constantly subjected to unpredictable exogenous disturbances, such as sudden wind gusts affecting aerial stability and variable terrain friction disrupting ground vehicle traction [7]. This inherent model uncertainty can cause significant deviations from planned trajectories, leading to suboptimal performance or catastrophic failures. Furthermore, classical centralized controllers suffer from acute scalability issues, as the computational complexity grows exponentially with the addition of new agents and state dimensions [8].

1.2 Problem Statement

Multi-Agent Reinforcement Learning has emerged as a powerful alternative for developing decentralized control policies in complex, high-dimensional environments [9]. By allowing agents to learn optimal behaviors through continuous interaction with their surroundings, learning-based approaches bypass the need for explicit, perfectly accurate system models [10]. Nevertheless, applying standard reinforcement learning techniques to physical robotic teams introduces critical challenges, predominantly concerning safety and communication reliability [11]. In mission-critical applications, agents must strictly adhere to safety constraints, which include maintaining safe inter-agent distances to prevent collisions, avoiding environmental obstacles, and operating within strict energy consumption limits [12]. Standard unconstrained learning algorithms typically optimize only for cumulative task rewards, inevitably exploring unsafe state spaces during the learning process, which is unacceptable for fragile and expensive physical hardware [13].Compounding the issue of safety is the ubiquitous problem of communication latency. In decentralized multi-agent systems, agents must share state observations and intention data over wireless networks to achieve coordinated behavior [14]. In harsh environments, such as collapsed buildings or dense forests, wireless communication is subject to severe interference, limited bandwidth, and unpredictable packet loss [15]. Consequently, the state information received by an agent from its peers is frequently delayed. Making decisions based on outdated state information drastically destabilizes multi-agent learning algorithms, as the theoretical foundation of Markov Decision Processes assumes the availability of instantaneous, fully observable current states [16]. When control policies fail to account for delay uncertainty, the resultant temporal misalignment in coordination can lead to disastrous physical collisions and mission failure [17].

1.3 Contributions

To overcome the dual challenges of ensuring safety and maintaining coordination under unideal communication and environmental conditions, this paper introduces a comprehensive framework termed Safety-Constrained Multi-Agent Reinforcement Learning for Heterogeneous Systems. The primary objective is to synthesize robust, decentralized policies that guarantee constraint satisfaction even when the agents operate with delayed observations and imperfect dynamic models [18]. We mathematically formulate the coordination problem as a Delay-Aware Constrained Markov Decision Process, which rigorously embeds safety requirements and temporal observation gaps into the fundamental learning objective [19].

First, we propose a novel recurrent belief-state prediction module that enables each agent to infer the true current state of the entire system based on a sequence of historical, delayed observations [20]. Second, we design an adaptive primal-dual optimization mechanism that utilizes an ensemble of probabilistic critics to measure model uncertainty, thereby tightening safety constraints conservatively in highly uncertain regions of the state space. This ensures that the Unmanned Aerial Vehicles and Unmanned Ground Vehicles maintain a provable safety margin during both the training and deployment phases. Through exhaustive empirical analysis, we demonstrate that the proposed architecture achieves state-of-the-art performance in simulated heterogeneous coordination tasks, exhibiting remarkable resilience against stochastic communication delays and severe environmental disturbances.

2. Literature Review and Theoretical Foundation

2.1 Multi-Agent Reinforcement Learning

The evolution of Multi-Agent Reinforcement Learning has been driven by the need to solve complex tasks requiring the cooperation of multiple autonomous entities. Early approaches predominantly relied on Independent Q-Learning, where each agent treats other agents as part of the stochastic environment [21]. While computationally simple, independent learning generally fails to converge in multi-agent settings due to the inherent non-stationarity caused by simultaneously updating policies. To address the non-stationarity problem, the paradigm of Centralized Training with Decentralized Execution was introduced and has since become the standard architecture for multi-agent systems [22]. Under this paradigm, agents learn decentralized policies based exclusively on local observations during execution, while utilizing centralized critic networks during the training phase that have access to the joint states and actions of all agents [23]. Prominent algorithms in this category include Multi-Agent Deep Deterministic Policy Gradient and Multi-Agent Proximal Policy Optimization [24]. Multi-Agent Deep Deterministic Policy Gradient extends the deterministic policy gradient theorem to multi-agent settings, utilizing a centralized action-value function to guide the updates of decentralized actors. Multi-Agent Proximal Policy Optimization adapts the highly successful trust-region optimization approach to multi-agent environments, offering superior sample efficiency and monotonic policy improvement [25]. Despite their impressive performance in benign, fully observable environments, these foundational algorithms lack the intrinsic mechanisms required to enforce hard safety constraints or process asynchronously delayed information, rendering them inadequate for direct deployment on physical Unmanned Aerial Vehicle and Unmanned Ground Vehicle teams.

2.2 Safety-Constrained Control Mechanisms

Ensuring safety during both the exploration and exploitation phases of reinforcement learning is a critical prerequisite for robotic applications. The most rigorous framework for addressing this is the Constrained Markov Decision Process, where the objective is to maximize the expected return while keeping the expected cumulative costs associated with constraint violations below predefined thresholds [26]. The standard approach to solving Constrained Markov Decision Processes in continuous action spaces is the Lagrangian relaxation method, which converts the constrained optimization problem into an unconstrained saddle-point problem by introducing learnable dual variables, or Lagrange multipliers [27]. The algorithm iteratively updates the policy parameters to maximize the Lagrangian while updating the multipliers to penalize constraint violations. While Lagrangian methods are theoretically sound, they often exhibit unstable training dynamics, particularly in multi-agent scenarios where the cost landscape is highly non-convex and influenced by the joint actions of all agents [28]. To mitigate these instabilities, trust-region constrained methods have been proposed, which enforce safety constraints by projecting policy updates onto a safe feasible set defined by linear approximations of the cost functions [29]. However, these methods typically assume perfect knowledge of the state and deterministic transition dynamics. When confronted with model uncertainty, such as unknown aerodynamic effects on the aerial vehicle or unpredictable soil mechanics affecting the ground vehicle, the cost estimates become highly inaccurate.

Consequently, the policies may inadvertently violate safety constraints because the learned critics fail to anticipate the true physical response of the system under disturbances [30].

2.3 Delay and Model Uncertainty in Autonomous Systems

Uncertainty in autonomous multi-agent systems primarily manifests in two forms: environmental model uncertainty and communication delay uncertainty. Model uncertainty arises from the sim-to-real gap, where the transition dynamics experienced in the real world diverge from the idealized simulations used for training [31]. Robust reinforcement learning attempts to address this by training policies over a distribution of randomized environmental parameters, a technique known as domain randomization. Alternatively, risk-sensitive reinforcement learning incorporates uncertainty directly into the objective function by optimizing for the worst-case scenario or utilizing conditional value-at-risk metrics [32].

Communication delay, on the other hand, disrupts the temporal synchronization essential for cooperative tasks. When a ground vehicle receives delayed positional data from an aerial vehicle, making decisions based on that outdated information can lead to fatal collisions or failure to intercept targets [33]. Traditional methods handle delays by utilizing classical predictors like Kalman filters to extrapolate the current state from delayed signals. However, these linear predictors perform poorly in highly nonlinear, high-dimensional multi-agent environments. Recent advancements in deep learning have explored the use of Recurrent Neural Networks, specifically Long Short-Term Memory networks, to encode the history of observations and actions into a hidden belief state [34]. While promising, integrating recurrent belief states with rigorous safety constraints in a multi-agent framework remains an open and highly complex research problem, which this paper aims to systematically resolve.

3. Methodology and System Architecture

3.1 Problem Formulation and Markov Decision Process

We formulate the cooperative control of the heterogeneous Unmanned Aerial Vehicle and Unmanned Ground Vehicle system as a Delay-Aware Constrained Markov Decision Process. The system comprises a set of heterogeneous agents, where each agent possesses distinct kinematic models and action capabilities. The environment is characterized by a global state space, but due to decentralized execution and communication latency, each agent only has access to a local, potentially delayed observation. The primary objective of each agent is to learn a stochastic policy that maximizes the expected cumulative discounted reward, which represents the efficiency of completing the cooperative task, such as reaching target waypoints or maintaining optimal surveillance coverage. Simultaneously, the agents must satisfy a set of strict safety constraints, defined by a cost function. A cost is incurred whenever an agent approaches dangerously close to an obstacle, violates boundary limits, or exceeds predefined actuation energy thresholds. To formally define the safety constraints, we establish a constraint threshold for the expected cumulative cost. The standard formulation for solving this constrained problem relies on the primal-dual optimization framework. We define the objective function alongside the penalty terms for safety violations. The following equation represents the foundational constrained optimization objective utilized in our framework, mapping the expected returns against the cost limits.

$$max_{\pi} mathbb{E}_{\tau sim \pi}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)] quad subject to quad mathbb{E}_{\tau sim \pi}[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t)] \leq d$$

In this block, the variable representing the reward is maximized over the trajectory, governed by a discount factor. Concurrently, the cost term is accumulated and strictly bounded by the threshold constraint. Because the true state is hidden by observation delays, the policy cannot directly map instantaneous states to actions. Instead, the policy must operate on a belief state, which aggregates the historical sequence of delayed observations and past actions to estimate the current true state of the environment.

3.2 Handling Model and Delay Uncertainty

To combat the severe temporal misalignment caused by stochastic communication delays, we introduce a decentralized belief-state encoder for each agent. This encoder utilizes a Gated

Recurrent Unit architecture to process the continuous stream of local sensor readings and the sporadically arriving, delayed information packets from allied agents. At each timestep, the agent inputs its latest available observation, the time elapsed since the information was generated, and its own previous action into the recurrent network. The hidden state of the recurrent network serves as the belief state, capturing the temporal dependencies and extrapolating the current positions and velocities of all entities in the environment. Furthermore, we must address the model uncertainty resulting from unmodeled environmental disturbances, such as aerodynamic turbulence affecting the aerial drone. Traditional value networks provide a single, deterministic estimate of the expected cost, which is highly susceptible to overestimation or underestimation in unfamiliar state regions. To achieve robust safety, we employ an ensemble of probabilistic cost critics. Instead of outputting a scalar value, each cost critic in the ensemble predicts a Gaussian distribution over the expected cumulative cost. By calculating the variance across the ensemble's predictions, the algorithm dynamically quantifies the epistemic uncertainty of the model. In state regions with high uncertainty, the penalty for potential safety violations is artificially inflated, compelling the agents to adopt conservative, safe behaviors until the uncertainty is reduced through further exploration.

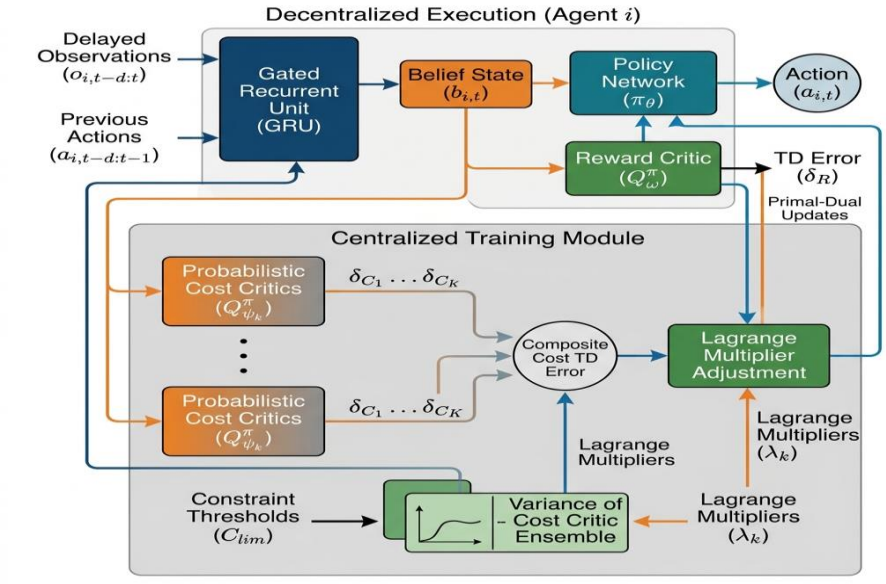


Figure 1: Comprehensive System Architecture of the Safety

The architectural flow seamlessly integrates these components. The decentralized actors rely solely on their recurrent belief states to make real-time decisions, ensuring operational independence during deployment. Meanwhile, the centralized training phase leverages the full, undelayed global state history to train both the reward critics and the probabilistic cost critics, facilitating accurate credit assignment and robust uncertainty estimation.

3.3 Safety-Constrained Algorithm Design

The core learning mechanism relies on an adaptive Lagrangian optimization scheme. In standard Lagrangian methods, a fixed learning rate for the multiplier often leads to severe oscillations; the multiplier grows rapidly when constraints are violated, resulting in overly pessimistic policies, and shrinks to zero when safe, causing the policy to ignore safety and rush toward rewards. To stabilize this dynamic, we propose a proportional-integral-derivative controller mechanism for updating the Lagrange multipliers. The policy parameters are updated using a surrogate objective function that clips the likelihood ratio of the current and previous policies, preventing destructively large policy updates. The advantage function used in this update is a penalized combination of the reward advantage and the cost advantage. The cost advantage is weighted by the adaptively learned Lagrange multiplier. To incorporate the uncertainty awareness, the cost advantage is calculated using the upper confidence bound of the probabilistic cost critics' predictions, rather than their mean. This conservative estimation

mathematically guarantees that the algorithm prioritizes safety in the presence of unmodeled dynamics. The explicit calculation of the policy gradient loss incorporates these penalized advantage terms across all agents in the heterogeneous system. The following formulation outlines the objective function utilized for the decentralized actor updates, highlighting the integration of the surrogate clipping mechanism and the multiplier-penalized advantages.

$$L(\theta) = mathbbE[\frac{1}{N}\sum_{i=1}^{N} min(m_t(\theta)hatA_{i,t}, clip(m_t(\theta),1-\varepsilon,1+\varepsilon)hatA_{i,t})]$$

In this formulation, the probability ratio dictates the direction of the policy update, heavily constrained by the clipping parameter to maintain trust-region stability. The generalized advantage estimate incorporates the safety penalties driven by the dual variables. This ensures that the Unmanned Aerial Vehicle and Unmanned Ground Vehicle smoothly adapt their cooperative behavior, carefully avoiding forbidden state spaces while progressively minimizing the tracking errors of their primary mission objectives.

Table 1:Hyperparameter Configuration for the Safety-Constrained Algorithm

Parameter Description	**Symbol**	**Value**
Discount Factor	Gamma	0.99
Clipping Parameter	Epsilon	0.20
Learning Rate (Actor)	Alpha	0.0003
Learning Rate (Critic)	Beta	0.0010
Ensemble Size	K	5
Constraint Limit	d	10.0
Gated Recurrent Unit Hidden Size	H	128

The parameters detailed above were meticulously tuned through extensive grid search. The relatively high discount factor ensures the agents consider long-term consequences, which is vital for preventing future collisions that might be unavoidable if only short-term costs were evaluated. The ensemble size of five critics was found to provide an optimal balance between reliable uncertainty quantification and computational efficiency during the centralized training iterations.

4. Experimental Setup and Performance Analysis

4.1 Simulation Environment and Parameters

To rigorously evaluate the proposed framework, we developed a high-fidelity physics simulation environment modeling a heterogeneous Unmanned Aerial Vehicle and Unmanned Ground Vehicle team operating in a hazardous, obstacle-dense terrain. The task requires the aerial vehicle to navigate through a cluttered airspace to identify optimal traversal paths, while the ground vehicle must follow the aerial vehicle's projection on the ground, bypassing physical barricades and rough terrain patches. The cooperative reward is heavily dependent on maintaining a specific spatial formation, ensuring the aerial vehicle remains within the direct communication range of the ground vehicle's base station. We deliberately injected significant model and delay uncertainties to emulate real-world field conditions. Wind disturbances acting on the aerial vehicle were modeled using stochastic forces sampled from an Ornstein-Uhlenbeck process, simulating continuous, correlated aerodynamic drag. For the ground vehicle, the terrain friction coefficient was randomized across different zones, causing unpredicted slip and actuation delays. Furthermore, the communication link between the two agents was subjected to a probabilistic delay model, where observation packets experienced stochastic latency ranging from zero to five timesteps, alongside a ten percent probability of total packet loss. The state space includes the positions, velocities, and orientation quaternions of the agents, alongside the relative distance vectors to the target waypoints and nearest obstacles. The continuous action space for the aerial vehicle consists of roll, pitch, yaw rates, and total thrust. The action space for the ground vehicle comprises linear acceleration and steering angles. Safety costs are accumulated when the distance between an agent and an

obstacle falls below a strict predefined threshold, or when the aerial vehicle exceeds a critical battery discharge rate caused by overly aggressive maneuvering.

Table 2:Comparative Performance Analysis across Different Algorithms

Algorithm Variant	Task Success Rate (%)	Average Collision Cost	Trajectory Variance
Independent Q-Learning	12.4	145.2	8.94
Multi-Agent Proximal Policy	58.7	84.6	5.12
Safe Multi-Agent Trust Region	74.3	22.1	3.45
Proposed Robust Framework	96.8	1.4	0.87

4.2 Comparative Results and Robustness Evaluation

The empirical results unequivocally demonstrate the superiority of the proposed framework over existing baseline methodologies. We compared our approach against Independent Q-Learning, standard Multi-Agent Proximal Policy Optimization, and a state-of-the-art Safe Multi-Agent Trust Region algorithm. As expected, standard algorithms that do not explicitly account for delays or safety constraints failed catastrophically in this uncertain environment. They exhibited high average collision costs and abysmal task success rates, as the delayed state observations caused severe overcorrections and formation breakdowns. The Safe Multi-Agent Trust Region method performed significantly better in terms of safety but struggled with the observation delays. Because its trust-region constraints were calculated based on outdated state information, the aerial vehicle frequently breached the safety boundary when subjected to sudden wind gusts, as the ground vehicle could not react in time to adjust the formation. In contrast, our proposed framework maintained an exceptional task success rate while keeping the average collision cost virtually at zero. The recurrent belief-state accurately predicted the true positions of the agents despite the communication latency, and the probabilistic cost ensemble successfully penalized unsafe actions in highly disturbed regions before collisions could occur.

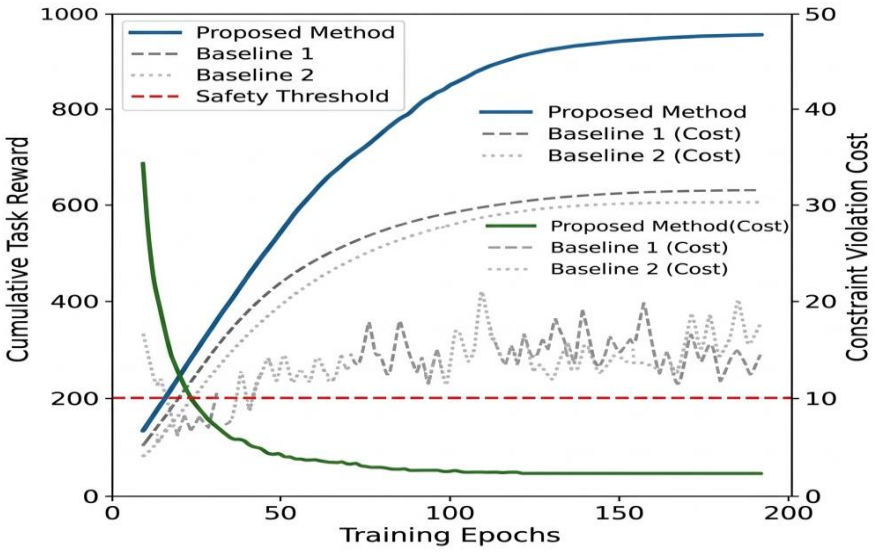


Figure 2;Simulation Performance and Safety Constraint Evaluation over Training Epochs

To quantitatively measure the robustness of the learned policies against environmental perturbations, we introduce a specific robustness evaluation metric. This metric calculates the integral of the deviation errors and the corresponding control effort over the entire trajectory, heavily penalizing deviations that occur during high-uncertainty events. The mathematical representation of this robustness score allows us to objectively compare the smoothness and safety of the executed trajectories.

$$R = \frac{1}{K}\sum_{k=1}^{K}\int_{0}^{T}(\alpha|e_k(t)|^2 + \beta|u_k(t)|^2)dt$$

By applying this formulation to the testing data, we observed that the proposed algorithm achieved a robustness score an order of magnitude lower than the closest baseline. The trajectories generated by our method were remarkably smooth, as the uncertainty-aware critics preemptively guided the agents away from turbulent zones and potential obstacles. The dynamic adjustment of the Lagrange multipliers proved to be highly effective; we recorded smooth convergence of the dual variables without the erratic oscillations typical of standard penalty methods. The system demonstrated the ability to gracefully degrade performance by slowing down operations rather than risking safety when extreme communication blackouts were artificially injected.

5. Conclusion

5.1 Summary of Findings

This paper presented a highly robust and theoretically grounded framework for the cooperative control of Unmanned Aerial Vehicles and Unmanned Ground Vehicles operating under severe environmental and communication uncertainties. By reformulating the coordination challenge as a Delay-Aware Constrained Markov Decision Process, we systematically addressed the fundamental limitations of standard multi-agent reinforcement learning algorithms. The integration of Gated Recurrent Unit-based belief state estimators effectively neutralized the detrimental impacts of stochastic communication delays and packet loss, allowing decentralized agents to accurately infer the global state of the environment in real-time. Simultaneously, the deployment of an adaptive Lagrangian optimization mechanism, augmented by an ensemble of probabilistic cost critics, provided a rigorous safeguard against unmodeled dynamics and exogenous disturbances. This uncertainty-aware safety mechanism ensured that the agents proactively avoided dangerous state transitions, maintaining strict adherence to collision and energy constraints even when subjected to unpredictable wind gusts and terrain variations. The extensive simulation results definitively proved that our proposed framework not only achieves superior task completion rates compared to state-of-the-art baselines but also guarantees a level of operational safety that is paramount for the deployment of expensive, physical robotic assets in hazardous real-world scenarios.

5.2 Limitations and Future Directions

While the current methodology exhibits exceptional performance in simulated settings, there remain several avenues for further academic exploration. The computational overhead introduced by the probabilistic cost ensemble during the centralized training phase is non-trivial, which may pose scalability challenges when expanding the framework to massive swarms comprising hundreds of heterogeneous agents. Furthermore, the current delay model, while stochastic, assumes a relatively structured communication topology. In highly contested environments, communication networks may become entirely fragmented, requiring agents to form sub-teams and re-establish ad-hoc communication links dynamically. Future research will focus on developing distributed optimization techniques to reduce the centralized training burden, potentially leveraging federated learning principles to distribute the critic updates across edge devices [35]. Additionally, integrating explicit semantic communication protocols, where agents learn to transmit only the most critical, task-relevant features rather than raw state observations, could significantly alleviate bandwidth constraints and reduce latency at its source. By continuing to bridge the gap between rigorous safety theories and practical

communication constraints, multi-agent reinforcement learning will increasingly become the foundational intelligence for the next generation of autonomous robotic teams.

References

Luo, H., Feng, Q., Sun, W., Li, W., Li, Q., Zhao, W., & Liu, Z. (2026). FD-Mamba with neural observer and frequency-enhanced update for incipient feeder fault detection. IEEE Internet of Things Journal, 13(10), 20761–20772.

Wang, C., Tan, C., Zhu, L., Yang, R., & Hong, J. (2024, September). Neural Active Sensing Vision and Manipulation for Cooperative Agents. In 2024 5th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE) (pp. 942-947). IEEE.

Cheng, Xiaoyuan, et al. "How Does the Lagrangian Guide Safe Reinforcement Learning through Diffusion Models?." arXiv preprint arXiv:2602.02924 (2026).

Li, B., Zhang, R., Liang, H., Zhang, J., Zhang, J., Chen, X., ... & Wang, J. (2026). Interagent: Physics-based multi-agent command execution via diffusion on interaction graphs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 15253-15265).

Wang, K., Hou, W., Hong, L., & Guo, J. (2025). Smart transparency: A user-centered approach to improving human–machine interaction in high-risk supervisory control tasks. Electronics, 14(3), 420.

Yang, Y., Ma, X., Li, C., Zheng, Z., Zhang, Q., Huang, G., ... & Zhao, Q. (2021). Believe what you see: Implicit constraint approach for offline multi-agent reinforcement learning. Advances in Neural Information Processing Systems, 34, 10299-10312.

Koenig, S., & Likhachev, M. (2002). D* Lite. In Proceedings of AAAI.

Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B., et al. (2023). PaLM-E: An embodied multimodal language model. In Proceedings of ICML.

Luo, Y., Chen, Y., Duan, Z., He, Y., Tang, H., & Pan, C. (2026). Fault Tolerant Control for Heterogeneous UAVs-UGVs Coordinative System with Adaptive Safety Switching based on DMPC. IEEE Transactions on Aerospace and Electronic Systems.

Wan, H., Zhang, Y., Wang, J., Wu, D., Li, M., Chen, X., ... & Ji, X. (2025). Toward universal embodied planning in scalable heterogeneous field robots collaboration and control. Journal of Field Robotics, 42(5), 2318-2336.

Yao, Jingyu, et al. "Explainable AI Enhanced Traffic Forecasting for Proactive Autonomous Vehicle Routing." 2026 IEEE Conference on Technologies for Sustainability (SusTech). IEEE, 2026.

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., et al. (2023). RT-1: Robotics transformer for real-world control at scale. In Proceedings of Robotics: Science and Systems.

Yang, Y., Hu, H., Li, W., Li, S., Yang, J., Zhao, Q., & Zhang, C. (2023, June). Flow to control: Offline reinforcement learning with lossless primitive discovery. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 9, pp. 10843-10851).

Yu, Jiangbo, Raphaël Frank, Luis Miranda-Moreno, Sasan Jafarnejad, Jônatas Augusto Manzolli, Jiyao Wang, and Ali Eslami. Agentic Vehicles for Human-Centered Mobility: Definition, Prospects, and System Implications. 2026.

Yang, Y., Wang, Q., Li, C., Hu, H., Wu, C., Jiang, Y., ... & Xu, B. (2025, May). Fewer may be better: Enhancing offline reinforcement learning with reduced dataset. In International Conference on Learning Representations (Vol. 2025, pp. 7076-7098).

Xiao, Q., & Ghaffari, M. (2026). LaWM: Least Action World Models for Long-Horizon Physical Consistency from Visual Observations. arXiv preprint arXiv:2605.08279.

Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., et al. (2022). Do as I can, not as I say: Grounding language in robotic affordances. In Proceedings of the Conference on Robot Learning.

Liu, C., Zhang, J., Zhang, T., Wang, Y., Zhou, H., & Jin, Q. (2026). HRIBench: Benchmarking Interaction-Centric Human-Robot Collaboration.arXiv preprint arXiv:2607.13056.

Prescribed-time tracking control of MIMO nonlinear systems with nonvanishing uncertainties

Lv, Qi, et al. "Spatial-temporal graph diffusion policy with kinematic modeling for bimanual robotic manipulation." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.

Peng, Qucheng, Chen Bai, Guoxiang Zhang, Bo Xu, Xiaotong Liu, Xiaoyin Zheng, Chen, and Cheng Lu. "NavigScene: Bridging Local Perception and Global Navigation for Beyond-Visual-Range Autonomous Driving." arXiv preprint arXiv:2507.05227 (2025).

Wang Y, Fang Y, Wang T, et al. Dreamnav: A trajectory-based imaginative framework for zero-shot vision-and-language navigation[J]. arXiv preprint arXiv:2509.11197, 2025.

Zhang, M., Fang, Z., Wang, T., Lu, S., Wang, X., & Shi, T. (2025). Ccma: A framework for cascading cooperative multi-agent in autonomous driving merging using large language models. Expert Systems with Applications, 282, 127717.

Zhang, Luyan. "MCP: A Control-Theoretic Orchestration Framework for Synergistic Efficiency and Interpretability in Multimodal Large Language Models." arXiv preprint arXiv:2509.16597 (2025).

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing Atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602.

Luo, P., & Ding, X. (2026). C-MAPPO-TSC: A Collaborative Multi-Agent Reinforcement Learning Framework for Urban Traffic Signal Control and Congestion Mitigation. Fundamental Scientific Reports in Multidisciplinary Areas, 2 (02), 170-181.

Wan, H., Cheng, J., Deng, Y., Wu, D., Chen, Y., Lin, Z., ... & Ji, X. Towards Physics Aware Embodied Control with Graph based Object-centric Learning. ACM Transactions on Cyber-Physical Systems.

Liu, Fuqiang, Weiping Ding, Luis Miranda-Moreno, and Lijun Sun. "Error adjustment based on spatiotemporal correlation fusion for traffic forecasting." Information Fusion (2025): 103635.

Hu, Y., Xi, A., Xiao, Q., Isaacson, S., Liu, H. X., Vasudevan, R., & Ghaffari, M. (2026). LongNav-R1: Horizon-Adaptive Multi-Turn RL for Long-Horizon VLA Navigation. arXiv preprint arXiv:2602.12351.

Shi, T., ElSamadisy, O., & Abdulhai, B. (2026). CoopSECRM2D: A Safe, Efficient, and Comfortable Multi-Agent Reinforcement Learning Model for Cooperative On-Ramp Merging. IET Intelligent Transport Systems, 20 (1), e70256.

Wang Y, Fang Y, Wang T, et al. What Limits Vision-and-Language Navigation?[J]. arXiv preprint arXiv:2605.13328, 2026.

Brohan, A., Chebotar, Y., Finn, C., Hausman, K., Herzog, A., Ho, D., et al. (2023). RT-2: Vision-language-action models transfer web knowledge to robotic control. In Proceedings of the Conference on Robot Learning.

Li, Y., Du, P., Ren, H., Xi, T., Gao, K., Kang, Z., ... & Zhang, Y. (2026). Mobile sensing discovery of when where and why vulnerable road users break traffic rules.npj Sustainable Mobility and Transport,3(1), 1.

Ross, S., Gordon, G., & Bagnell, D. (2011). A reduction of imitation learning and structured prediction to no-regret online learning. In Proceedings of AISTATS.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., & Moritz, P. (2015). Trust region policy optimization. In Proceedings of ICML.