https://doi.org/10.31891/2307-5732-2026-363-59
УДК 004.8

ЛИТВИНЕНКО МИХАЙЛО
Харківський національний університет радіоелектроніки
https://orcid.org/0000-0003-4487-8811
e-mail: mykhailo.lytvynenko1@nure.ua
РЕБЕЗЮК ЛЕОНІД
Харківський національний університет радіоелектроніки
https://orcid.org/0000-0001-8516-6584
e-mail: leonid.rebezyuk@nure.ua

РОЗРОБКА БАГАТОАГЕНТНОЇ СИСТЕМИ ГЛИБОКОГО НАВЧАННЯ З ПІДКРІПЛЕННЯМ ДЛЯ ЕФЕКТИВНОГО КЕРУВАННЯ СВІТЛОФОРАМИ НА ОДНОМУ ПЕРЕХРЕСТІ

У цій статті розглядається процес керування світлофорами на одному перехресті як кооперативний децентралізований частково спостережуваний марковський процес вирішування (Дец-ЧСМПВ), що подається як мінімальна тестова платформа для вивчення децентралізованої координації в умовах невизначеності, а не як самостійне завдання оптимізації. Кілька агентів керують окремими групами сигналів, використовуючи детальні примітивні дії, приділяючи увагу модульності, стійкості до обмежень датчиків та сумісності з традиційними етапними системами керування. Для забезпечення координації без явної комунікації пропонується розширений простір спостереження, що містить як динамічні характеристики руху, так і структурну інформацію про перехрестя, що дозволяє здійснювати пасивну координацію за допомогою спільних фізичних сигналів. На основі цього формулювання представлено децентралізовану багатоагентну систему глибокого навчання з підкріпленням, яка інтегрує рекурентну оцінку цінности для пом'якшення часткової спостережуваности, розподільче підкріплювальне навчання для збереження мультимодальних структур віддачі, що виникають у результаті суперечливих рівноваг координації, та гістерезисні оновлення для стабілізації динаміки децентралізованого навчання. Керування сигналами за допомогою примітивних дій спричиняє ланцюгові процеси ухвалення рішень зі стохастичними результатами, де наївне розвідування та оцінювання цінности на основі середніх значень часто призводять до передчасної збіжности до неоптимальних стратегій координації. Запропонована система, що враховує невизначеність, явно розглядає цю проблему. Попередні експерименти з моделювання використовуються для аналізу динаміки навчання, чутливости рівноваги та координаційної поведінки. Результати не підкреслюють переваги в продуктивності, а ілюструють поведінкові наслідки запропонованого переформулювання та структури навчання. Ця робота надає принципову систему для децентралізованого керування світлофорами з урахуванням невизначеності та створює основу для майбутніх розширень до масштабованої координації на кількох перехрестях.

Ключові слова: *кооперативне навчання з підкріпленням, часткова спостережуваність, невизначеність, децентралізоване навчання й виконання, керування світлофорами.*

LYTVYNENKO MYKHAILO, REBEZYUK LEONID
Kharkiv National University of Radio Electronics

MULTI-AGENT DEEP REINFORCEMENT LEARNING FRAMEWORK DESIGN FOR EFFICIENT SINGLE-INTERSECTION TRAFFIC LIGHT CONTROL

This paper reformulates single-intersection traffic light control as a cooperative Decentralized Partially Observable Markov Decision Process (Dec-POMDP), treating it as a minimal testbed for studying decentralized coordination under uncertainty rather than as a standalone optimization task. Multiple agents control disjoint signal groups using fine-grained primitive actions, emphasizing modularity, robustness to sensing limitations, and compatibility with legacy stage-based control systems. To enable coordination without explicit communication, we propose an extended observation space that includes both dynamic traffic features and structural intersection information, allowing passive coordination through shared physical signals. Building on this formulation, we introduce a decentralized multi-agent deep reinforcement learning framework that integrates recurrent value estimation to mitigate partial observability, distributional reinforcement learning to preserve multi-modal return structures arising from competing coordination equilibria, and hysteretic updates to stabilize decentralized learning dynamics. Primitive-action traffic signal control induces chain-like decision processes with stochastic outcomes, where naive exploration and mean-based value estimates often lead to premature convergence to suboptimal coordination strategies. The proposed uncertainty-aware framework explicitly addresses this challenge. Preliminary simulation experiments are used to analyze learning dynamics, equilibrium sensitivity, and coordination behavior. Rather than emphasizing performance superiority, the results illustrate the behavioral implications of the proposed reformulation and learning design. This work provides a principled framework for decentralized, uncertainty-aware traffic signal control and establishes a foundation for future extensions to scalable multi-intersection coordination.

Keywords*: cooperative reinforcement learning, partial observability, uncertainty, decentralized training and execution, traffic light control.*

Стаття надійшла до редакції / Received 18.02.2026
Прийнята до друку / Accepted 03.03.2026
Опубліковано / Published 26.03.2026

This is an Open Access article distributed under the terms of the Creative Commons CC-BY 4.0

© Литвиненко Михайло, Ребезюк Леонід

Introduction

Efficient traffic light control (TLC) is an acute problem in intelligent transportation systems (ITS), involving optimization of the sequence and duration of traffic signals with direct implications for congestion, travel time reliability, and emissions. While modern traffic sensing technologies provide increasingly detailed information about transport flows and queue dynamics, converting these observations into robust and adaptive traffic signal plans remains

challenging, particularly under dynamic and heterogeneous demand patterns. Traditional TLC approaches, ranging from fixed-time to actuated and adaptive control, mostly rely on predefined signal sequence and hand-crafted decision logic. Although these systems are widely deployed and operationally reliable, they are inherently limited in their ability to adapt to high variability in traffic conditions and to generalize across different intersection layouts or demand scenarios. Their control logic is tightly coupled to specific signal plan assumptions, making transferability and extensibility difficult.

Recent work has explored reinforcement learning (RL) as a data-driven alternative for traffic signal control. Most RL-based approaches to single-intersection TLC formulate the problem as a Markov decision process (MDP), where a single agent operates on the fixed set of allowed combinations of traffic signals (phases) based on aggregated intersection-level observations. While such methods have demonstrated promising empirical performance, they implicitly assume centralized decision-making and often rely on coarse, stage-based actions to stabilize learning. These assumptions limit scalability, reduce robustness to partial observability and sensor failures, and hide the fine-grained coordination structure underlying traffic signal control. From a broader perspective, the core difficulty in traffic signal control is not just optimizing a single intersection, but learning decentralized coordination policies that remain stable, transferable, and consistent under fine-grained control actions. This way, a single intersection provides a minimal yet non-trivial testbed for studying decentralized decision-making under partial observability, competing coordination equilibria, and stochastic dynamics, and these challenges are central to multi-agent reinforcement learning (MARL) more generally.

In this work, we revisit single-intersection traffic light control from this perspective and propose a decentralized reformulation of the problem. Rather than treating the intersection as a single control unit, we decompose it into multiple cooperative agents, each responsible for a local control region and operating with primitive actions, such as extending or terminating a signal group. The resulting control problem is formalized as a Decentralized Partially Observable Markov Decision Process (Dec-POMDP), enabling decentralized decision-making under partial observability while avoiding the exponential state growth associated with centralized formulations. Building on this reformulation, we design a decentralized multi-agent deep reinforcement learning framework based on decentralized training and execution. Finally, while the present study considers an isolated intersection, the formulation serves as a scalable foundation for future extensions to multi-intersection traffic networks.

1. Single-intersection traffic light control reformulation

1.1. Intersection model

We consider a signalized road intersection (Fig. 1a) composed of a finite set of signal groups (SGs), where each signal group $g_i \in G$ controls a compatible set of traffic movements [1]. Conflicts between signal groups are specified by a hand-crafted conflict matrix A (Fig. 1b), which encodes mutually exclusive traffic movements, and can equivalently be represented as a compatibility graph $\mathcal{G} = (G, E)$ (Fig. 1c), where edges indicate non-conflicting signal groups that may be active simultaneously.

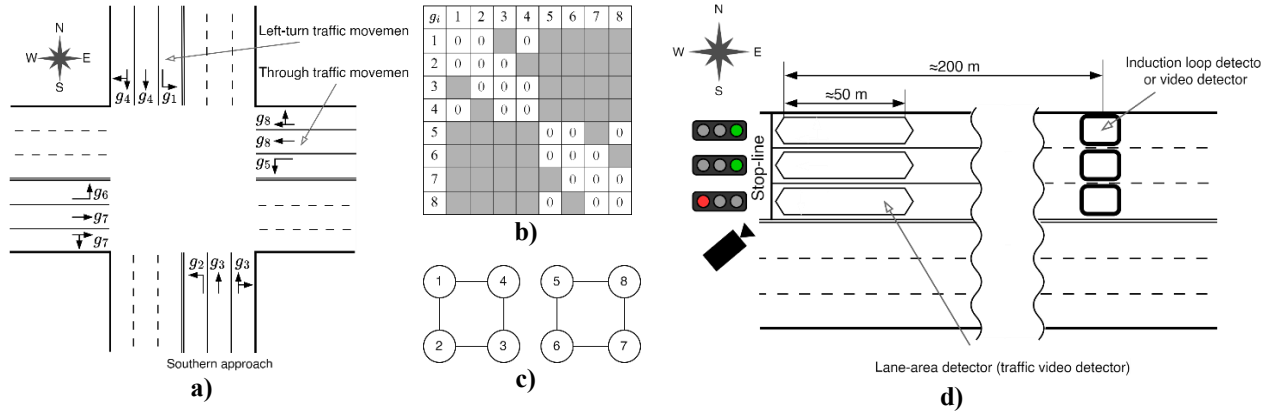

Fig. 1. Decentralized traffic light control environment:
a) – SGs at an intersection; b) – conflict matrix; c) – compatibility graph; d) – traffic detection system

This representation is agnostic to intersection-specific geometry while preserving the structural constraints that define feasible signal configurations.

1.2 Decentralized formalization

The control problem is formalized as a Dec-POMDP [2], where each agent corresponds to a local control region. Each region can be associated with one or more SG, and each agent operates based on partial observations of the intersection state. Although a single intersection could in principle be modeled as a centralized multi-agent POMDP with factored state and action spaces, the decentralized formulation offers several advantages:

- No exponential state growth, as agents reason locally rather than over the full joint state.
- The learned policies are modular, enabling reuse across intersections with similar local structure.
- Resilience to partial observability and sensor limitations, as agents do not rely on complete global information.
- Scalability for network-level control, where fully centralized formulations quickly become intractable.

Importantly, no centralized training nor execution are required in this formulation, allowing coordination to emerge from decentralized learning dynamics.

1.3. Extended observation space

Inspired by the definition from [3], each agent receives a local observation, derived from a system of upstream detectors (Fig. 1d) and the current signal plan, that includes three categories of information:

- Dynamic operation features, such as arrival rates, detectors occupancy, maximum elapsed green time within the agent's control region.
- Intersection configuration features, encoding static structural information derived from the conflict matrix and local characteristics of intersection geometry.
- Candidate signal group states, representing the substitution eligibility and traffic conditions of inactive signal groups that are relevant for coordination with the active neighboring agents.

This design enables passive coordination, i.e., there is no explicit coordination signals, instead, the joint control context is inferred through shared environmental structure. By including coordination-relevant information directly into observations, the framework avoids the complexity and scalability limitations associated with explicit communication protocols or centralized learning components.

1.4. Primitive action space design

Each agent operates using a primitive action space, restricted to simple control decisions such as extending or terminating the current activation of a signal group, subject to the requirement of each phase being the maximal clique. This contrasts with the phase-selection or phase-skipping actions commonly used in both traditional and learning-based traffic signal control. The comparison of benefits and drawbacks of the proposed formulation is provided in Table 1.

Table 1

Action space design analysis

Strengths	Limitations
– Decoupling of the control logic from fixed phase semantics, enabling policies to generalize across different intersection layouts. – Backward compatibility with legacy traffic control systems, where signal extension and termination are fundamental operations. – Bidirectional convertibility of actions between stage- and group-based methods. – Exposure of fine-grained coordination structure, which is often hidden by temporally extended or stage-based abstractions.	– Stronger coupling between agents' decisions. – Longer decision chains. – Delayed credit assignment.

Overall, the increased expressiveness and compatibility inevitably results in a higher complexity of a problem, which becomes characterized by higher non-stationarity and multiple competing coordination equilibria.

1.5. Implications for scalability and coordination

The proposed reformulation naturally supports semi-synchronous decentralized decision-making, where agents act independently but remain coupled through shared environmental constraints and observations. Coordination emerges as agents adapt their local policies to the dynamically changing set of feasible joint signal configurations. By focusing agents on localized control regions rather than assigning full intersection control to a single agent, the framework improves robustness to local failures and distributes the difficulty of decision-making. This becomes increasingly important for intersections with complex geometries or high-dimensional traffic patterns, where centralized control is less robust. Overall, this reformulation reframes single-intersection traffic light control as a testing ground for decentralized coordination under uncertainty, providing the foundation upon which the learning framework, introduced in the subsequent sections, is built.

2. Background and related work

This section provides an overview of prior work on traffic light control and multi-agent reinforcement learning, with a focus on how control granularity, coordination, and uncertainty are expressed in the existing approaches.

2.1. Traffic light control literature

Classical traffic light control strategies include fixed-time, actuated, and adaptive systems. Fixed-time controllers rely on precomputed signal plans, optimized for nominal traffic conditions [4,5], while actuated [6,7] and adaptive controllers [8,9] adjust signal timings based on real-time sensor measurements. Despite their widespread adoption and operational reliability, these methods are fundamentally limited by hand-crafted, and mostly stage-based control logic. As a result, their ability to adapt to non-stationary traffic patterns or transfer across intersection layouts is constrained. These limitations have motivated increasing interest in learning-based approaches that aim to automatically discover control policies from data.

Most reinforcement learning approaches to single-intersection traffic light control formulate the problem as a centralized decision-making task. A single agent observes aggregated intersection-level traffic features and either selects signal phase or choses to skip to the next one. Representative examples include FRAP [10], MetaLight [11], NSTLight [12], AttendLight [13], GNN-based approaches [14], GeneraLight [15], and TSC-HGAM-DRL [16], which focus on improved representation learning, sample efficiency, or generalization across traffic scenarios. While these methods have demonstrated strong empirical performance, they rely on two common assumptions. First, control is centralized, with a single policy responsible for coordinating all traffic movement signals. Second, actions are typically stage-based

or temporally extended, reducing the effective decision frequency and simplifying credit assignment. These assumptions stabilize learning but conceal the underlying coordination structure of the intersection and limit flexibility under partial observability.

Multi-agent formulations of single-intersection traffic light control are comparatively rare. Early work such as AGBC-SARSA(λ) [3] explores tabular group-based multi-agent control, while more recent deep reinforcement learning approaches, such as TLCC based on Revised QMIX [17], introduce centralized critics and value decomposition to stabilize learning. Although these methods move toward decentralized execution, they still rely on temporally-extended actions or centralized training components and do not explicitly address coordination under primitive actions.

2.2. Multi-agent reinforcement learning and exploration

Current multi-agent reinforcement learning methods can broadly be categorized into centralized training with decentralized execution (CTDE) and decentralized training and execution (DTE) [18]. CTDE approaches leverage centralized critics or joint value functions during training to mitigate non-stationarity, but often suffer from scalability and generalization issues as the number of agents increases. In contrast, DTE methods favor flexibility and modularity, enabling agents to learn and operate based solely on local observations.

Examples of DTE-based deep reinforcement learning algorithms include independent Q-learning variants with recurrence (e.g., HDRQN [19], LH-IRQN [20]) and independent policy optimization (e.g., IPPO [21], PS-TRPO [22]). These methods are particularly suitable for domains where centralized information is unavailable or undesirable.

A common challenge in both single-agent and multi-agent reinforcement learning is the exploration-exploitation trade-off, especially in environments with delayed rewards and long decision horizons [23]. Temporally extended actions and hierarchical abstractions are commonly introduced to alleviate credit assignment difficulties [24]. In traffic signal control, this is the reason behind the adoption of stage-based actions that combine multiple low-level decisions into a single control choice.

An alternative line of work emphasizes uncertainty-aware exploration [25–27], where agents sample actions based on uncertainty estimates derived from value distributions or posterior approximations. Distributional reinforcement learning and posterior sampling methods have been shown to improve exploration efficiency and robustness in stochastic environments. However, their role in stabilizing coordination and equilibrium selection in decentralized, primitive-action settings remains underexplored.

2.3. Research gap and positioning

Despite the significant progress in learning-based traffic signal control, existing approaches leave a gap at the intersection of control granularity, decentralization, and uncertainty awareness. This work addresses it by reformulating single-intersection traffic light control as a decentralized cooperative problem with primitive actions and partial observability. Building on this formulation, we adopt the DTE paradigm and integrate uncertainty-aware learning mechanisms to stabilize coordination without relying on temporal abstractions or explicit communication.

By extending ideas from tabular decentralized control to continuous state spaces and combining them with distributional value estimation, recurrence, and hysteretic updates, the proposed framework offers a principled approach to decentralized traffic signal control that remains compatible with existing infrastructure while exposing the fine-grained coordination structure inherent to the problem.

3. Theoretical framework design

The framework design requirements are dictated by the structural properties of the problem, and each component is motivated as a necessary mechanism to ensure stable decentralized coordination under uncertainty.

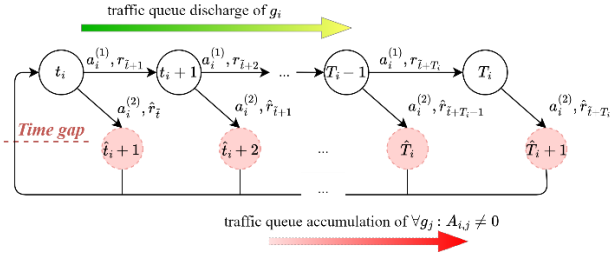

Fig. 2. Decision-making sequence for an agent *i*

3.1. Design rationale

Decentralized primitive-action traffic signal control has several properties that fundamentally distinguish it from the conventional stage-based or centralized formulations. First, primitive actions introduce long decision chains: extending or terminating a signal group affects not only immediate traffic flow but also the feasibility and timing of future actions by other agents; therefore, rewards are delayed and strongly coupled across agents (Fig. 2). Second, decentralized learning under partial observability is associated with multiple competing coordination equilibria. Different joint policies may correspond to distinct but internally consistent patterns of signal plans, leading to multi-modal return distributions. Third, mean-based value estimation collapses coordination modes prematurely. When return distributions are multi-modal, standard expected-value critics encourage early commitment to suboptimal equilibria and amplify non-stationarity across agents. Finally, uncertainty-ignoring exploration is insufficient in this setting. Naive exploration strategies fail to distinguish between the stochastic variability and structural uncertainty arising from insufficient knowledge of the environment dynamics and incomplete coordination, leading to unstable learning process

and premature convergence. These observations motivate a learning framework that (i) preserves return distribution structure, (ii) stabilizes decentralized learning, and (iii) enables deliberate equilibrium selection under uncertainty.

3.2. Problem setting and notation

Let the single intersection be modeled as a Dec-POMDP $\langle \mathcal{I}, \mathcal{S}, \boldsymbol{\mathcal{A}}, P, r, \boldsymbol{\Omega}, O, \gamma \rangle$, where $\mathcal{I} = \{1, \dots, N\}$ is the set of cooperative agents, each associated with a local control region and one or more signal groups, $\mathcal{S}$ denotes the (unobserved) global traffic state, $\mathcal{A}_i$ is the primitive action space of agent i, and $\boldsymbol{\mathcal{A}} = \times_{i \in \mathcal{I}} \mathcal{A}_i$ is the joint action space. $P(s' \mid s, \boldsymbol{a})$ defines the environment dynamics, $r: \mathcal{S} \times \boldsymbol{\mathcal{A}} \to \mathbb{R}$ is a shared reward function, Ω_i is the observation space of agent i, $O(\omega_i \mid s, \boldsymbol{a})$ is the local observation function, and $\gamma \in (0,1)$ denotes the discount factor.

At time step t, each agent receives a local observation $\omega_t^i \in \Omega_i$ and selects an action $a_t^i \in \mathcal{A}_i$. The joint action is denoted by $\boldsymbol{a} = (a_i)_{i \in \mathcal{I}}$. Because the environment is partially observable, agents condition their decisions on **action-observation histories** $h_t^i = (\omega_1^i, a_1^i, \dots, \omega_t^i)$**.** Each agent aims to learn a decentralized policy $\pi_i(a^i \mid h^i)$ that maximizes the expected discounted return $J = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, \boldsymbol{a}_t) \mid \boldsymbol{a}_t \sim \boldsymbol{\pi}]$. Under DTE, agents optimize their policies independently using only local histories, while coordination emerges implicitly through shared rewards and environmental coupling.

3.3. Core learning paradigm

We adopt the DTE paradigm and each agent employs a value-based single-agent deep reinforcement learning algorithm, which is based on the Deep Q-Network algorithm (Eq. 1) [28], approximating the action value function $Q^{\pi}(s,a) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a, a_t \sim \pi]$.

$$\mathcal{L}_{\mathrm{DQN}}(\boldsymbol{\theta}) = \mathbb{E}_{\langle s,a,r,s' \rangle \sim \mathcal{U}(\mathcal{D})}[\delta(\boldsymbol{\theta})]^2, \tag{1}$$

where $\delta(\boldsymbol{\theta}) = r(s,a) + \gamma \max_{a'} Q(s', a'; \boldsymbol{\theta}^-) - Q(s,a;\boldsymbol{\theta})$ is referred to as TD-error, $\mathcal{D}$ is the experience replay memory, $\boldsymbol{\theta}$, $\boldsymbol{\theta}^-$ are the parameters of the main and target networks, respectively, and s', a' are the next state and action, respectively. This approach allows for direct integration of uncertainty estimates into action selection.

The shared reward signal reflects the cooperative objective of minimizing travel time at the intersection. Specifically, while the high sparsity of the conflict matrix indicate that most agents remain inactive at any given moment, the cooperative nature of the problem necessitates considering conditions on all traffic movements, not just the ones being currently served. To this end, the reward signal is calculated using the information from all controlled traffic movements. Fine-grained actions necessitate highly responsive feedback signal, therefore, inspired by [3] and [29], we define the following reward function that encourages minimizing travel delay of the maximum number of vehicles k (Eq. 2):

$$r(t) = 1 - \frac{d(t)}{d_{\max}(t)}; d(t) = \sum_{k} \left(1 - \left(\frac{v_k(t)}{v_{\max}}\right)^2\right), \tag{2}$$

where $d_{\max}(t)$ is the maximum of observed delays from $d(0)$ and $v_{\max}$ is the maximum allowed speed.

3.4. Representation learning under partial observability

To address partial observability and delayed credit assignment, each agent uses a recurrent value function that conditions on the compressed observation-action history. Recurrence enables agents to maintain an implicit belief over latent traffic states and evolving coordination context. To preserve the structure of multi-modal returns induced by competing coordination equilibria, we adopt distributional reinforcement learning. Instead of approximating the expected return, agents learn a parametric approximation of the return distribution $Z^{\pi}(h,a)$: $\mathbb{E}[Z^{\pi}(h,a)] = Q^{\pi}(h,a)$. The single-agent basis algorithm used in this work is Implicit Recurrent Quantile Network (IRQN) [30], which enables the estimation of the distribution's inverse cumulative distribution function (c.d.f.) Z_{τ} at an arbitrary number N, N' of quantile samples $\tau_{1:N}, \tau'_{1:N'} \sim \mathcal{U}([0,1])$ (Eq. 3), where $\mathcal{H}$ denotes the Huber loss.

$$\mathcal{L}_{\mathrm{IRQN}}(\boldsymbol{\theta}) = \frac{1}{N'} \sum_{i=1}^{N} \sum_{j=1}^{N'} |\tau_i - \mathbb{I}_{\delta \le 0}| \mathcal{H}\left(\delta^{\tau_i, \tau'_j}(\boldsymbol{\theta})\right). \tag{3}$$

This representation prevents premature mode collapse and provides a natural basis for uncertainty estimation. A dueling architecture [31] is used to decompose distributional state-value and action-advantage estimates, improving learning stability in settings where action effects are highly context-dependent.

3.5. Coordination stabilization mechanisms

Decentralized learning with multiple adaptive agents is inherently non-stationary. To mitigate instability and encourage convergence toward consistent coordination patterns, several complementary stabilization mechanisms are employed.

Parameter sharing is used among homogeneous agents to exploit structural symmetry and improve sample efficiency [22]. Under this scheme, the number of trainable parameters remains constant as the number of agents grows, while still allowing decentralized execution through agent-specific observations. In addition to agents having access to their own environment observations, non-identical policies are achieved by including the agent's ID in the extended observation. Concurrent experience sampling [19] ensures that agents update their policies based on temporally aligned interaction data, reducing drift caused by asynchronous learning and promoting convergence toward a shared equilibrium. Hysteretic learning updates are introduced to stabilize value estimation under non-stationarity [32]. Instead of using a fixed learning rate, per-state-action updates are modulated based on the temporal difference likelihood (TDL) $l_{t_{1:N'}, d_{1:N}}$ of observed target samples $t_{1:N'} = r + \gamma \max_{a'} Z_{\tau_{1:N'}}(h', a'; \boldsymbol{\theta}^-)$ under the learned return distribution $d_{1:N}$

[20]. This asymmetric update rule inhibits destructive updates caused by transient discoordination while allowing rapid adaptation when evidence consistently supports alternative coordination modes (Eq. 4). The threshold parameter β controls the impact on the update rule of low TDL values.

$$\alpha = \begin{cases} \max\left(\beta, l_{t_{1:N'}, d_{1:N}}\right)\bar{\alpha}, & \delta^{\tau,\tau'} \leq 0, \\ \bar{\alpha}, & \text{otherwise.} \end{cases} \tag{4}$$

Together, these mechanisms encourage agents to converge toward mutually consistent policies without requiring centralized control or explicit coordination signals.

3.6. Uncertainty-aware exploration and safety

In the proposed framework, uncertainty plays a dual role: guiding exploration and enabling deliberate equilibrium selection. Both epistemic uncertainty $\hat{U}_{\text{epist}}$, parametrized by η and arising from limited data and model uncertainty, and aleatoric uncertainty $\hat{U}_{\text{aleat}}$, parametrized by λ and arising from inherent traffic stochasticity, are captured through the learned return distributions using an ensemble of two randomized maximum a posteriori (MAP) samples $\boldsymbol{\theta}_{P_1}, \boldsymbol{\theta}_{P_2}$ [27] (Eq. 5,6).

$$\hat{U}_{\text{epist}}(Z, \bar{\tau}) = \frac{1}{2}\mathbb{E}_{\bar{\tau}}\left[\left(Z_{\bar{\tau}}(h, a; \boldsymbol{\theta}_{P_1}) - Z_{\bar{\tau}}(h, a; \boldsymbol{\theta}_{P_2})\right)^2\right], \tag{5}$$

$$\hat{U}_{\text{aleat}}(Z, \bar{\tau}) = \text{Cov}_{\bar{\tau}}\left(Z_{\bar{\tau}}(h, a; \boldsymbol{\theta}_{P_1}),\ Z_{\bar{\tau}}(h, a; \boldsymbol{\theta}_{P_2})\right). \tag{6}$$

Action selection is performed using Thompson sampling from parameterized Gaussian distribution, using an arbitrary number of quantile samples for inference $\tilde{N}$ (Eq. 7,8), allowing agents to explore consistently over extended decision chains rather than through independent, myopic perturbations.

$$\boldsymbol{\mu}_{\boldsymbol{Q}} = \left(\mathbb{E}_{\tau_{1:\tilde{N}}}\left[Z_{\tau_{1:\tilde{N}}}(h, a; \boldsymbol{\theta})\right] + \lambda \hat{U}^{1/2}_{\text{aleat}}(Z(h, a), \tau_{1:\tilde{N}})\right)_{a \in \mathcal{A}}, \tag{7}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{Q}} = \text{diag}\left(\eta^2 \hat{U}_{\text{epist}}(Z(h, a), \tau_{1:\tilde{N}})_{a \in \mathcal{A}}\right),$$

$$a_i = \arg\max_{a'} \hat{Q}(\cdot, a');\ \hat{Q} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{Q}}, \boldsymbol{\Sigma}_{\boldsymbol{Q}}). \tag{8}$$

This uncertainty-aware exploration is particularly important in primitive-action settings, where early exploratory decisions can have long-lasting downstream effects. By sampling from plausible return hypotheses, agents can commit temporarily to consistent coordination patterns and evaluate their long-term consequences. To ensure compliance with regulatory and safety constraints, action masking is applied at the logit level to exclude infeasible or prohibited actions. This guarantees that all selected actions respect signal compatibility constraints and traffic regulations.

3.7. Discussion of theoretical implications

The proposed framework illustrates how decentralized coordination can emerge from local learning dynamics when structural observability, uncertainty modeling, and stabilization mechanisms are carefully aligned. Rather than relying on temporal abstractions or centralized critics, coordination arises through interaction between agents, environment constraints, and uncertainty-aware decision-making. Importantly, the framework is not specific to traffic signal control. The combination of primitive actions, passive coordination, and distributional decentralized learning provides a general template for studying cooperative multi-agent systems where partial observability, fine-grained control and uncertainty play a central role.

4. Implementation considerations and discussion

This section describes the practical realization of the proposed framework and reports preliminary empirical observations. The goal is not to establish state-of-the-art performance, but to validate the behavioral and coordination properties induced by the proposed decentralized primitive-action formulation and uncertainty-aware cooperative learning mechanisms.

4.1. Simulation environment and system architecture

The framework is implemented using a modular simulation stack composed of SUMO for microscopic traffic simulation, PettingZoo for multi-agent environment abstraction, and PyTorch for learning components. Each signal group (SG) is modeled as an autonomous agent interacting with the environment at discrete decision points, while sharing a common policy network under a parameter-sharing scheme. The interaction loop follows the DTE paradigm, without access to global state. A short pre-training phase with random exploration is used to bias agents toward compatible coordination patterns and avoid early collapse to suboptimal equilibria.

4.2. Empirical observations

Despite limited experimental scale, several consistent behavioral patterns emerge, as described below.

Initialization sensitivity and equilibrium selection: training outcomes are highly sensitive to random initialization and exploration dynamics, reflecting the presence of multiple coordination equilibria. Distinct equilibria correspond to qualitatively different signal timing patterns, some of which are locally stable yet globally suboptimal. ***Multi-modal coordination dynamics:*** across runs, three coordination regimes are observed. Notably, the most performant regime, which is characterized by near-optimal green time allocation proportional to traffic intensities, is also the least stable and hardest to reach. Mean-based value estimation alone collapses learning toward simpler but inferior equilibria. ***Role of uncertainty-aware learning:*** distributional value representations and uncertainty-directed action selection significantly reduce premature convergence. Allowing per-state–action learning rate increases under high-likelihood value estimates prevents agents from initial locking into trivial extend-and-hold strategies, which

otherwise dominate when the learning rate factor is limited to 1 (Eq. 4).

4.3. Discussion and Future directions

The experimental findings support the central objective of this work: fine-grained decentralized control transforms single-intersection TLC from a scheduling problem into a coordination problem under uncertainty. The proposed formulation exposes coordination structure that is invisible in stage-based or centralized approaches but also introduces new challenges. The strengths of the framework are its modularity, backward-compatibility with legacy systems, and natural extensibility to more complex intersection geometries. Parameter sharing and decentralized execution provide a scalable foundation for future network-level control. However, in terms of robustness, uncertainty is currently utilized for exploration and equilibrium selection but is not explicitly incorporated into the optimization objective and belief uncertainty remains implicit. Promising directions include explicit belief-state modeling, principled risk-aware objectives, and extension to multi-intersection settings where decentralized coordination becomes both necessary and unavoidable.

Conclusion

This work proposes a decentralized, uncertainty-aware multi-agent reinforcement learning framework for single-intersection traffic light control based on primitive actions and extended observations. Rather than optimizing a single intersection in isolation, the framework serves as a minimal testing ground for studying decentralized coordination under partial observability. The results highlight both the opportunities and challenges introduced by fine-grained control, and point toward scalable, resilient ITS solutions grounded in principled multi-agent learning.

References

1. Stoffers, K. E. (1967). Scheduling of traffic lights-a new approach. *Transportation Research/UK/*, *2*(3).

2. Oliehoek, F. A., & Amato, C. (2016). *A concise introduction to decentralized POMDPs* (Vol. 1). Springer. https://doi.org/10.1007/978-3-319-28929-8

3. Jin, J., & Ma, X. (2017). A group-based traffic signal control with adaptive learning ability. *Engineering Applications of Artificial Intelligence*, *65*, 282–293. https://doi.org/10.1016/j.engappai.2017.07.022

4. Webster, F. V. (1958). *Traffic signal settings* (No. 39).

5. Improta, G., & Cantarella, G. E. (1984). Control system design for an individual signalized junction. *Transportation Research Part B: Methodological*, *18*(2), 147–167. https://doi.org/10.1016/0191-2615(84)90028-6

6. Dunne, M. C., & Potts, R. B. (1964). Algorithm for traffic control. *Operations Research*, *12*(6), 870–881.

7. Wong, S., Wong, W., Leung, C., & Tong, C. (2002). Group-based optimization of a time-dependent TRANSYT traffic model for area traffic control. *Transportation Research Part B: Methodological*, *36*(4), 291–312. https://doi.org/10.1016/s0191-2615(01)00004-2

8. Hunt, P. B., Robertson, D. I., Bretherton, R. D., & Winton, R. I. (1981). *SCOOT-a traffic responsive method of coordinating signals* (No. LR 1014 Monograph). (TRRL). https://trid.trb.org/view/179439

9. Mirchandani, P., & Head, L. (2001). A real-time traffic signal control system: Architecture, algorithms, and analysis. *Transportation Research Part C: Emerging Technologies*, *9*(6), 415–432. https://doi.org/10.1016/s0968-090x(00)00047-4

10. Zheng, G., Xiong, Y., Zang, X., Feng, J., Wei, H., Zhang, H., Li, Y., Xu, K., & Li, Z. (2019). Learning phase competition for traffic signal control. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 1963–1972. https://doi.org/10.1145/3357384.3357900

11. Zang, X., Yao, H., Zheng, G., Xu, N., Xu, K., & Li, Z. (2020). *Metalight: Value-based meta-reinforcement learning for traffic signal control*. *34*(01), 1153–1160.

12. Wang, H., Tian, S., Zhang, W., Zhou, Y., Li, W., & Ning, N. (2024). NSTLight: A Traffic Light Control Method based on Graph Attention Network withNon-StationaryFeatureLearning. *Proceedings of the 2024 2nd International Conference on Electronics, Computers and Communication Technology*, 86–94. https://doi.org/10.1145/3705754.3705770

13. Oroojlooy, A., Nazari, M., Hajinezhad, D., & Silva, J. (2020). Attendlight: Universal attention-based reinforcement learning model for traffic signal control. *Advances in Neural Information Processing Systems*, *33*, 4079–4090.

14. Yoon, J., Ahn, K., Park, J., & Yeo, H. (2021). Transferable traffic signal control: Reinforcement learning with graph centric state representation. *Transportation Research Part C: Emerging Technologies*, *130*, 103321.

15. Zhang, H., Liu, C., Zhang, W., Zheng, G., & Yu, Y. (2020). Generalight: Improving environment generalization of traffic signal control via meta reinforcement learning. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 1783–1792.

16. Zhao, Z., Wang, K., Wang, Y., & Liang, X. (2024). Enhancing traffic signal control with composite deep intelligence. *Expert Systems with Applications*, *244*, 123020. https://doi.org/10.1016/j.eswa.2023.123020

17. Du, T., Wang, B., & Hu, L. (2023). Single intersection traffic light control by multi-agent reinforcement learning. *Journal of Physics: Conference Series*, *2449*(1), 012031.

18. Gronauer, S., & Diepold, K. (2022). Multi-agent deep reinforcement learning: A survey. *Artificial Intelligence Review*, *55*(2), 895–943. https://doi.org/10.1007/s10462-021-09996-w

19. Omidshafiei, S., Pazis, J., Amato, C., How, J. P., & Vian, J. (2017). Deep decentralized multi-task multi-agent reinforcement learning under partial observability. *International Conference on Machine Learning*, 2681–2690.

https://export.arxiv.org/pdf/1703.06182

20. Lyu, X., & Amato, C. (2018). Likelihood quantile networks for coordinating multi-agent reinforcement learning. *arXiv Preprint arXiv:1812.06319*. https://doi.org/10.48550/arxiv.1812.06319

21. Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., & WU, Y. (2022). The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in Neural Information Processing Systems* (Vol. 35, pp. 24611–24624). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2022/file/9c1535a02f0ce079433344e14d910597-Paper-Datasets_and_Benchmarks.pdf

22. Gupta, J. K., Egorov, M., & Kochenderfer, M. (2017). Cooperative multi-agent control using deep reinforcement learning. *Autonomous Agents and Multiagent Systems: AAMAS 2017 Workshops, Best Papers, São Paulo, Brazil, May 8-12, 2017, Revised Selected Papers 16*, 66–83. https://doi.org/10.1007/978-3-319-71682-4_5

23. Hao, J., Yang, T., Tang, H., Bai, C., Liu, J., Meng, Z., Liu, P., & Wang, Z. (2024). Exploration in Deep Reinforcement Learning: From Single-Agent to Multiagent Domain. *IEEE Transactions on Neural Networks and Learning Systems*, *35*(7), 8762–8782. https://doi.org/10.1109/TNNLS.2023.3236361

24. Pignatelli, E., Ferret, J., Geist, M., Mesnard, T., Hasselt, H. van, Pietquin, O., & Toni, L. (2024). *A Survey of Temporal Credit Assignment in Deep Reinforcement Learning* (No. arXiv:2312.01072). arXiv. https://doi.org/10.48550/arXiv.2312.01072

25. Osband, I. (2016). Risk versus uncertainty in deep learning: Bayes, bootstrap and the dangers of dropout. *NIPS Workshop on Bayesian Deep Learning*, *192*.

26. Depeweg, S., Hernández-Lobato, J.-M., Doshi-Velez, F., & Udluft, S. (2018). Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. *International Conference on Machine Learning*, 1184–1193. https://proceedings.mlr.press/v80/depeweg18a/depeweg18a.pdf

27. Clements, W. R., Van Delft, B., Robaglia, B.-M., Slaoui, R. B., & Toth, S. (2019). Estimating risk and uncertainty in deep reinforcement learning. *arXiv Preprint arXiv:1905.09638*. https://doi.org/10.48550/arxiv.1905.09638

28. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., & others. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533. https://doi.org/10.1038/nature14236

29. Ducrocq, R., & Farhi, N. (2023). Deep reinforcement Q-learning for intelligent traffic signal control with partial detection. *International Journal of Intelligent Transportation Systems Research*, *21*(1), 192–206. https://doi.org/10.1007/s13177-023-00346-4

30. Dabney, W., Ostrovski, G., Silver, D., & Munos, R. (2018). Implicit Quantile Networks for Distributional Reinforcement Learning. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning* (Vol. 80, pp. 1096–1105). PMLR. https://proceedings.mlr.press/v80/dabney18a.html

31. Wang, Z., Schaul, T., Hessel, M., Hasselt, H. van, Lanctot, M., & Freitas, N. de. (2016). *Dueling Network Architectures for Deep Reinforcement Learning* (No. arXiv:1511.06581). arXiv. https://doi.org/10.48550/arXiv.1511.06581

32. Matignon, L., Laurent, G. J., & Le Fort-Piat, N. (2007). Hysteretic q-learning: An algorithm for decentralized reinforcement learning in cooperative multi-agent teams. *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 64–69. https://doi.org/10.1109/iros.2007.4399095