

Article

Decentralized Hierarchical Multi-Agent DRL for Resource Allocation in IRS-Aided V2X Networks

Ayaz Ahmad 

Department of Electrical Engineering, College of Engineering, King Faisal University, Al-Ahsa, Saudi Arabia; ayahmad@kfu.edu.sa

Abstract

Vehicle-to-Everything (V2X) communication is an essential building block of intelligent transportation systems, supporting high-data-rate vehicle-to-infrastructure (V2I) services, and ultra-reliable low-latency vehicle-to-vehicle (V2V) communication. However, in dense urban environments, V2X services can be significantly degraded by the presence of severe blockage, fast channel variations, and high levels of interference. Intelligent Reflecting Surfaces (IRSs) can be employed to reconfigure wireless propagation environments to improve V2X communication. However, the joint optimization of transmit power, spectrum reuse, and IRS reflection coefficients is a mixed-integer non-linear problem, which is further complicated by the fast vehicular mobility and time-varying interference in V2X networks. To tackle this challenging problem, this work proposes a scalable and deployable decentralized hierarchical multi-agent deep reinforcement learning (DH-MDRL) framework. The key design principle is the separation of control timescales, whereby each V2V link functions as an autonomous agent that responds to local observations at a fast timescale and determines its transmit power and spectrum reuse decisions, while the IRS controller at the base station (BS), using global network observations, updates the IRS reflection coefficients at a slower timescale. This hierarchical architecture reduces coordination signaling associated with centralized resource allocation while enabling distributed resource allocation. The IRS-assisted V2X network is modeled as a Markov decision process, where the reward design is tailored to optimize the V2I sum data rate while guaranteeing the latency and reliability constraints associated with safety-critical V2V communication. Simulation results show that the proposed DH-MDRL framework outperforms conventional schemes without IRSs and achieves an excellent trade-off between V2V link constraints' satisfaction probability and V2I link sum data rates compared to centralized resource allocation approaches.

Keywords: IRS; V2X communications; V2V resource allocation; deep reinforcement learning; multi-agent learning spectrum sharing and interference management



Academic Editors: Philipp Svoboda,
Zoran Cica and Dejan Drajić

Received: 15 June 2026

Revised: 15 July 2026

Accepted: 15 July 2026

Published: 20 July 2026

Copyright: © 2026 by the authors.
Licensee MDPI, Basel, Switzerland.
This article is an open access article
distributed under the terms and
conditions of the [Creative Commons
Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Vehicle-to-Everything (V2X) communication has emerged as a key enabling technology for intelligent transportation systems, supporting advanced applications such as autonomous driving, cooperative perception, traffic safety, and infotainment services. V2X networks encompass vehicle-to-infrastructure (V2I) links as well as vehicle-to-vehicle (V2V) communications, which impose fundamentally different quality-of-service (QoS) requirements. While V2I links primarily support throughput-oriented applications such as high-definition map downloading, multimedia streaming, cloud-assisted navigation, over-the-air software updates, and passenger infotainment, these services generally prioritize high data rates over stringent

end-to-end latency guarantees. Although higher data rates reduce transmission time, the associated applications can usually tolerate moderate communication delays without significantly affecting service quality. In contrast, V2V communications support safety-critical applications, including collision avoidance, emergency braking, cooperative driving, and platooning, where messages must be delivered within a few milliseconds with extremely high reliability [1]. To meet these heterogeneous QoS demands in dense urban environments is a challenging goal due to severe signal blockage, fast channel variations due to vehicular mobility and strong mutual interference of coexisting links [2–4].

Intelligent Reflecting Surfaces (IRSs) have recently emerged as a promising solution to improve wireless communication performances by programmable reconfiguration of the radio propagation environment [5]. By controlling the phase shift of a high number of passive reflecting elements, an IRS can reinforce the desired signals, block interferences, and extend the coverage area without increasing the transmit power. In V2X scenarios, the deployment of IRSs on roadside building facades has shown to mitigate blockage and enhance V2I and V2V communication performance [6,7]. A number of works have shown that IRSs can considerably improve reliability as well as coverage probability and spectral efficiency in vehicular networks [8–10]. However, exploiting IRS gains requires joint optimization of spectrum reuse, transmit power and IRS phase shifts, thus leading to a complex mixed-integer non-linear optimization problem.

The use of conventional optimization methods for resource allocation in IRS-assisted V2X networks face several limitations. One typical limitation lies in the coupled optimization of discrete spectrum reuse decisions and continuous transmit power levels and unit-modulus IRS phase shifts in IRS-assisted V2X network, which presents a high computational overhead for conventional methods. Another limitation concerns difficulties in obtaining accurate and timely (or even approximately accurate) channel state information in challenging fast-moving vehicular scenarios. In addition, difficulty also arises from the high-signaling overhead and latency brought by centralized solutions, which renders them unscalable and not applicable for practical V2X systems, especially when frequent updates to resource allocation and IRS configuration are necessary [11–13].

To address these challenges, deep reinforcement learning (DRL) has emerged as a powerful tool for adaptive resource management in V2X networks. DRL enables agents to learn decision-making policies through interaction with the environment, without relying on explicit channel models. Early work in [2] demonstrated that decentralized DRL can effectively allocate power and spectrum for V2V links while satisfying latency constraints. Subsequent studies introduced meta-reinforcement learning [14], multi-agent DRL [15], and two-timescale learning frameworks [16] to improve adaptability and scalability. Nevertheless, most DRL-based V2X approaches assume a fixed wireless propagation environment and do not exploit the capability of IRS to dynamically reshape wireless channels [17,18].

Conversely, IRS-assisted optimization techniques often rely on centralized control and quasi-static channel assumptions [7,11], which limits their applicability in fast-changing vehicular networks. Recent studies have begun integrating IRS and DRL, demonstrating promising results in non-vehicular and infrastructure-centric scenarios [19–21]. However, applying IRS-assisted DRL to V2X systems with strict V2V latency constraints, spectrum reuse, and large-scale decentralized decision-making remains insufficiently explored [22].

Motivated by these observations, this paper investigates joint resource allocation and IRS configuration for IRS-aided V2X networks from a deployment-oriented learning perspective. Unlike existing works that tightly couple IRS reconfiguration with instantaneous vehicular resource allocation, we explicitly account for the mismatch between fast V2V dynamics and slower IRS reconfiguration capabilities. The goal is to design a scalable and

adaptive framework that maximizes V2I throughput while guaranteeing stringent V2V latency and reliability requirements under dynamic channel conditions.

To this end, we propose a decentralized hierarchical multi-agent DRL framework for IRS-assisted V2X communications. In the proposed architecture, each V2V link operates as an autonomous learning agent that selects its transmit power and spectrum reuse decision based on local observations, while a centralized IRS controller, co-located with the base station, configures IRS phase shifts at a slower timescale. This time-scale separation reflects practical IRS deployment constraints, reduces signaling overhead, and improves scalability without sacrificing performance. Carefully designed reward functions ensure V2V latency requirements are satisfied while enhancing overall V2I throughput.

The main contributions of this paper are summarized as follows:

- We develop an IRS-assisted V2X system model in which multiple V2V links reuse the spectrum of V2I links, accounting for interference coupling, vehicular mobility, and heterogeneous QoS requirements for V2V and V2I communications.
- We formulate the joint optimization of V2V transmit power, spectrum reuse, and IRS reflection coefficients as a mixed-integer non-linear optimization problem with explicit latency constraints for safety-critical V2V communications.
- We propose a decentralized hierarchical multi-agent DRL framework that coordinates fast-timescale decentralized V2V resource allocation with slow-timescale centralized IRS reconfiguration for joint IRS-assisted V2X resource allocation.
- Through extensive simulations, we demonstrate that the proposed IRS-assisted DRL framework significantly improves V2I sum data rate and V2V latency satisfaction probability compared with centralized DRL, and DRL without IRS baseline schemes under various network configurations.

2. Related Work

This section reviews existing studies related to learning-based resource allocation and IRS-assisted vehicular communications. The discussion is organized into four parts: (i) learning-based resource allocation for V2X networks, (ii) IRS-assisted vehicular communications, (iii) learning-based optimization for IRS-assisted V2X systems, and (iv) motivation and research gap.

2.1. Learning-Based Resource Allocation for V2X Networks

The dynamic topology, rapidly varying channels, and stringent latency requirements of V2X networks make conventional optimization-based resource allocation difficult to apply in practice. Deep reinforcement learning (DRL) has therefore been widely adopted for decentralized and adaptive V2X resource management.

In [2], a decentralized DRL framework was proposed for joint spectrum and power allocation in V2V communications, showing that latency constraints can be satisfied without centralized coordination. Cooperative and multi-agent DRL schemes were further studied in [23] to improve spectrum reuse efficiency between V2I and V2V links. Fairness-aware DRL formulations were considered in [24] to enhance the worst-case V2V performance.

DRL has also been applied to V2X networks integrated with mobile edge computing (MEC). For example, Ref. [25] optimized task offloading and radio resources using an actor-critic framework. Although effective for computation-centric objectives, such approaches do not explicitly address latency-critical V2V communications.

To improve scalability and adaptability, multi-agent DRL frameworks were proposed in [4], while meta-reinforcement learning [14] and federated learning [15] were introduced to cope with non-stationary vehicular environments and reduce signaling overhead. Recent studies have further demonstrated the effectiveness of multi-agent reinforcement learning

for cooperative decision-making in intelligent vehicular systems. For example, the two-stage MARL framework in [26] enables efficient mixed vehicle platoon formation under dynamic traffic conditions, illustrating the potential of MARL for distributed optimization in connected vehicular networks. Despite these advances, most DRL-based V2X works assume a fixed wireless propagation environment and do not consider programmable channel control.

2.2. IRS-Assisted Vehicular Communications

IRS technology has attracted significant interest for enhancing wireless communications through passive and programmable signal reflection. In vehicular scenarios, IRS deployment has been shown to mitigate blockage and improve coverage and reliability [6,7]. Optimization-based IRS-assisted V2X studies mainly focus on beamforming, phase-shift design, and deployment strategies [8,12].

Analytical performance evaluation of IRS-assisted vehicular systems has also been investigated. In [27], outage probability and capacity of RIS-assisted V2V links were analyzed under Nakagami- m fading. Similarly, Ref. [10] studied IRS-assisted V2I communications in urban dark zones, demonstrating improved coverage and interference mitigation.

While these works provide valuable insights, they typically assume static or quasi-static channels and rely on centralized optimization with full channel knowledge, limiting their applicability to fast-moving vehicular environments and large-scale networks.

2.3. Learning-Based Optimization for IRS-Assisted V2X Networks

To overcome the limitations of model-based IRS optimization, learning-based approaches have recently been explored. DRL has been applied to jointly optimize transmit power and IRS phase shifts in device-to-device, IoT, and cellular networks [19–21], demonstrating the ability to handle coupled optimization problems without explicit channel models.

Learning-based IRS optimization has also been extended to vehicular networks in [17], where a centralized hierarchical DRL framework jointly optimizes resource block allocation, BS transmit power allocation, and IRS configuration for IRS-assisted downlink V2I communications. However, the framework relies on global network information and does not consider V2V communications. In contrast, the proposed DH-MDRL framework considers an IRS-assisted uplink V2I system underlaid with latency-constrained V2V communications, where decentralized V2V agents are coordinated with a centralized IRS controller operating at a slower timescale. The principal differences between the two frameworks are summarized in Table 1. IRS-assisted learning has also been studied in vehicular edge computing systems [18], focusing on Age of Information and energy efficiency rather than latency-critical V2V communications.

Table 1. Comparison between Yang et al. [17] and the proposed DH-MDRL framework.

Feature	Yang et al. [17]	Proposed DH-MDRL
Communication scenario	IRS-assisted downlink V2I communications ((V2V communications are outside the scope)).	IRS-assisted uplink V2I communications with underlaid V2V links.
V2V communications	Not considered.	Explicit V2V transmitter–receiver pairs with latency constraints.
Optimization formulation	Joint optimization of RB allocation, BS transmit power allocation, and IRS configuration.	Hierarchical optimization of V2V spectrum reuse, V2V transmit power allocation, and IRS reflection configuration across two timescales.

Table 1. *Cont.*

Feature	Yang et al. [17]	Proposed DH-MDRL
Learning architecture	Centralized hierarchical DRL framework.	Hierarchical MARL framework with decentralized V2V agents and a centralized IRS controller.
State information	Global network state.	Local observations for V2V agents; network-level measurements available only at the IRS controller.
Update mechanism	Event-triggered centralized updates.	Asynchronous V2V learning with slower-timescale IRS updates.
Control signaling	Extensive network-wide information exchange.	Reduced signaling through decentralized V2V decision-making.
Deployment model	Centralized optimization framework.	Practical hierarchical deployment with distributed V2V intelligence.

In [22], the IRS was combined with UAVs to enhance infrastructure-to-vehicle communications using DRL-based control. More recently, STAR-RIS-assisted V2X systems were investigated in [13], where hybrid learning and optimization techniques were used to jointly optimize spectrum allocation, STAR-RIS parameters, and transmit power. While enabling extended coverage and heterogeneous QoS support, these approaches largely rely on centralized or partially centralized control and do not explicitly address scalable decentralized V2V learning under fast vehicular dynamics.

Overall, existing learning-based IRS-assisted studies primarily emphasize throughput or infrastructure-centric objectives and often employ complex centralized or deeply hierarchical learning architectures.

2.4. Motivation and Research Gap

From the above review, existing DRL-based V2X resource allocation methods effectively support decentralized decision-making and latency constraints but assume a fixed wireless propagation environment. In contrast, IRS-assisted V2X studies demonstrate the benefits of programmable wireless channels but rely heavily on centralized optimization and frequent IRS reconfiguration.

Furthermore, current learning-based IRS-assisted approaches either focus on infrastructure-centric communications, vehicular edge computing objectives, or tightly coupled hierarchical control, and do not jointly address decentralized V2V resource allocation, spectrum reuse with V2I links, and practical IRS reconfiguration constraints in highly dynamic environments.

Motivated by these gaps, this paper investigates a decentralized multi-agent DRL framework for IRS-assisted V2X communications that explicitly incorporates timescale separation between fast V2V adaptations and slower IRS reconfiguration. By coordinating decentralized V2V decision-making with infrequent IRS updates, the proposed framework achieves scalable and robust resource management while satisfying stringent V2V latency and reliability requirements.

3. System Model and Problem Formulation

In this section, we present the system model for the IRS-aided V2X communication network and formulate the joint resource allocation and IRS coefficient optimization problem. The V2I links are intended to support high-data-rate mobile entertainment services and thus aim to achieve the maximum possible data rate, whereas the V2V communications support safety-critical services and therefore require ultra-low latency and high reliability. Accordingly, the design objective is to maximize the sum data rate of the V2I links while

guaranteeing the stringent latency and reliability requirements of the V2V communications under dynamic channel conditions.

3.1. System Model

As illustrated in Figure 1, we consider an uplink IRS-assisted vehicle-to-everything (V2X) communication system consisting of a single base station (BS), an Intelligent Reflecting Surface (IRS) with N passive reflecting elements deployed on a roadside building, and multiple vehicular users.

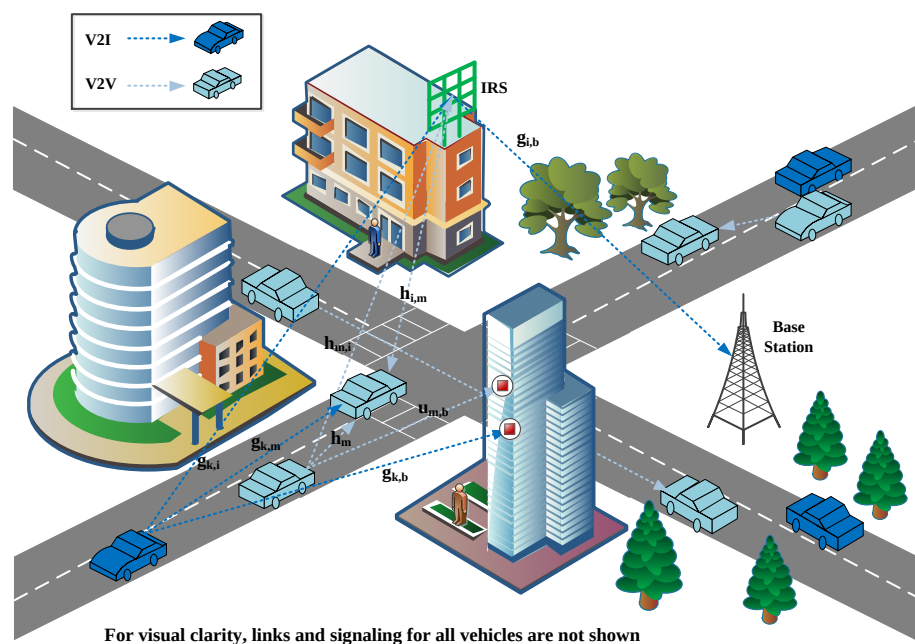


Figure 1. IRS-assisted V2X system model showing uplink V2I transmission, V2V transmitter–receiver communication, and IRS-assisted reflected propagation paths.

In the considered uplink scenario, vehicles act as transmitters and the BS serves as the receiver for vehicle-to-infrastructure (V2I) communications. Each V2I link is established between a vehicle and the BS, where the transmitted signal reaches the BS through both a direct wireless link and an IRS-assisted reflected link. The IRS operates in a passive manner and does not perform any active transmission or signal processing.

In addition, vehicle-to-vehicle (V2V) communications are established between designated transmitter–receiver vehicle pairs for safety-critical information exchange. For each V2V pair, the transmitted signal may reach the intended receiver through both the direct propagation path and the IRS-assisted reflected path. Depending on their service requirements, vehicles are categorized into V2I or V2V links, and V2V links are allowed to reuse the spectrum resources allocated to V2I links in an underlay fashion.

Let $\mathcal{K} = \{1, \dots, K\}$ denote the set of V2I links and $\mathcal{M} = \{1, \dots, M\}$ denote the set of V2V transmitter–receiver pairs. Each V2I link is preassigned one orthogonal sub-band, and V2V links are allowed to reuse these sub-bands in an underlay manner. The total number of IRS elements is represented by N .

3.1.1. Channel Model

The channel gain vector from the k -th V2I transmitting vehicle to the IRS is denoted by $\mathbf{g}_{k,i} \in \mathbb{C}^{N \times 1}$, while the channel gain vector from the IRS to the BS is denoted by $\mathbf{g}_{i,b} \in \mathbb{C}^{N \times 1}$. The direct uplink channel between the k -th V2I vehicle and the BS is denoted by $g_{k,b} \in \mathbb{C}$.

For V2V communications, the channel gain vector from the m -th V2V transmitter to the IRS is denoted by $\mathbf{h}_{m,i} \in \mathbb{C}^{N \times 1}$, and the channel gain vector from the IRS to the corresponding V2V receiver is denoted by $\mathbf{h}_{i,m} \in \mathbb{C}^{N \times 1}$. The direct channel gain between the m -th V2V transmitter and receiver is denoted by $h_m \in \mathbb{C}$.

The gain of interference channel from the m -th V2V transmitter to the BS is denoted by $u_{m,b} \in \mathbb{C}$, the gain of interference channel from the k V2I transmitter to the m -th V2V receiver is denoted by $v_{b,m} \in \mathbb{C}$, and the gain of interference channel from the j -th V2V transmitter to the m -th V2V receiver is denoted by $w_{j,m} \in \mathbb{C}$.

All channels incorporate large-scale path loss, shadowing, and small-scale fading.

The IRS reflection coefficient matrix is given by

$$\Phi = \text{diag}\left(e^{j\theta_1}, e^{j\theta_2}, \dots, e^{j\theta_N}\right), \quad (1)$$

where $\theta_n \in [0, 2\pi)$ denotes the phase shift of the n -th IRS element.

3.1.2. SINR and Data Rate Model

The received SINR at the BS for the k -th V2I link is expressed as

$$\gamma_k^{\text{V2I}} = \frac{P_k^{\text{V2I}} \left| \mathbf{g}_{i,b}^H \Phi \mathbf{g}_{k,i} + g_{k,b} \right|^2}{\sum_{m=1}^M x_{k,m} P_m^{\text{V2V}} \left| \mathbf{g}_{i,b}^H \Phi \mathbf{h}_{m,i} + u_{m,b} \right|^2 + \sigma^2}. \quad (2)$$

The achievable sum rate of all V2I links is given by

$$R_{\text{V2I}} = \sum_{k=1}^K W \log_2 \left(1 + \gamma_k^{\text{V2I}} \right). \quad (3)$$

where W is the bandwidth of one sub-band.

The received SINR at the m -th V2V receiver is expressed as

$$\gamma_m^{\text{V2V}} = \frac{P_m^{\text{V2V}} \left| \mathbf{h}_{i,m}^H \Phi \mathbf{h}_{m,i} + h_m \right|^2}{I_{\text{V2I},m} + I_{\text{V2V},m} + \sigma^2}. \quad (4)$$

where $I_{\text{V2I},m} = \sum_{k=1}^K x_{k,m} P_k^{\text{V2I}} \left| \mathbf{h}_{i,m}^H \Phi \mathbf{g}_{k,i} + v_{k,m} \right|^2$, and $I_{\text{V2V},m} = \sum_{\substack{j=1 \\ j \neq m}}^M x_{k,m} x_{k,j} P_j^{\text{V2V}} \left| \mathbf{h}_{i,m}^H \Phi \mathbf{h}_{j,i} + w_{j,m} \right|^2$.

3.2. V2V Latency and Reliability Model

V2V communications support safety-critical services and therefore require ultra-low latency and high reliability. Following the widely adopted payload–deadline model in the V2X literature, each V2V link m is assumed to transmit a data packet of fixed size B bits that must be delivered within a maximum latency constraint $T_{\max}$.

Time is divided into equal slots of duration Δt . Let $R_m^{\text{V2V}}(t)$ denote the instantaneous achievable data rate of the m -th V2V link at time slot t , which is given by

$$R_m^{\text{V2V}}(t) = W \log_2 \left(1 + \gamma_m^{\text{V2V}}(t) \right), \quad (5)$$

where W is the bandwidth of one sub-band and $\gamma_m^{\text{V2V}}(t)$ is the received SINR at the V2V receiver.

Let $L_m(t)$ denote the remaining payload (in bits) of the m -th V2V link at time slot t , and let $T_m(t)$ denote the remaining allowable transmission time. Their evolutions are given by

$$L_m(t+1) = \max\{0, L_m(t) - R_m^{V2V}(t)\Delta t\}, \quad (6)$$

$$T_m(t+1) = T_m(t) - \Delta t, \quad (7)$$

with initial conditions $L_m(0) = B$ and $T_m(0) = T_{\max}$.

The transmission of the V2V packet is considered *successful* if

$$L_m(t^*) = 0 \quad \text{for some } t^* \text{ such that } T_m(t^*) \geq 0. \quad (8)$$

Otherwise, the latency constraint is violated.

This formulation captures both latency and reliability requirements and is consistent with existing V2X DRL frameworks, where the remaining payload and remaining time are included in the agent state.

3.3. Problem Formulation

The objective of the proposed resource allocation framework is to maximize the aggregate V2I communication throughput while ensuring that every safety-critical V2V transmission satisfies its latency requirement. To achieve this objective, the spectrum reuse decisions, V2V transmit powers, and IRS reflection coefficients are jointly optimized. The resulting optimization problem is formulated as

$$\max_{\{x, P^{V2V}, \Phi\}} R_{V2I} \quad (9)$$

subject to

$$\text{C1: } \sum_{k=1}^K x_{k,m} \leq 1, \quad \forall m \in \mathcal{M}, \quad (10)$$

$$\text{C2: } 0 \leq P_m^{V2V} \leq P_{\max}^{V2V}, \quad \forall m \in \mathcal{M}, \quad (11)$$

$$\text{C3: } |\Phi_{n,n}| = 1, \quad \forall n = 1, \dots, N, \quad (12)$$

$$\text{C4: } L_m(T_{\max}) = 0, \quad \forall m \in \mathcal{M}. \quad (13)$$

where $x = \{x_{k,m}, \forall m \in \mathcal{M}, \forall k \in \mathcal{K}\}$ with $x_{k,m} \in \{0, 1\}$, and $P^{V2V} = \{P_m^{V2V}, \forall m \in \mathcal{M}\}$.

In (9), the optimization variables consist of the spectrum reuse indicator x , the transmit power vector P^{V2V} , and the IRS reflection matrix Φ . The objective function maximizes the total achievable data rate of all V2I links defined in (3). Constraint C1 ensures that each V2V pair can reuse the spectrum of at most one V2I link. Constraints C2 limit the transmit power of V2V links. Constraint C3 enforces the unit-modulus condition on the IRS reflection coefficients. Constraint C4 enforces the latency requirement for V2V links, ensuring that each V2V packet is successfully delivered within the maximum tolerable delay. Consequently, the optimization seeks to improve V2I throughput without compromising the stringent QoS requirements of safety-critical V2V communications.

Due to the stochastic nature of wireless channels and vehicle mobility, directly solving this mixed-integer non-linear optimization problem is intractable. Therefore, we adopt a decentralized deep reinforcement learning approach that does not require explicit modeling of the system dynamics, and where V2V agents implicitly satisfy the latency constraint through reward design.

4. Hierarchical Multi-Agent Reinforcement Learning Approach

In this section, we develop a hierarchical multi-agent reinforcement learning (RL) framework for joint V2V resource allocation and IRS coefficient optimization in the considered IRS-aided V2X network. The proposed RL framework is illustrated in Figure 2, where decentralized V2V agents interact with the wireless environment at the slot level by selecting spectrum reuse and transmit power, while a centralized IRS agent, co-located with the BS, updates the IRS phase-shift matrix at a slower timescale (e.g., once per coherence interval).

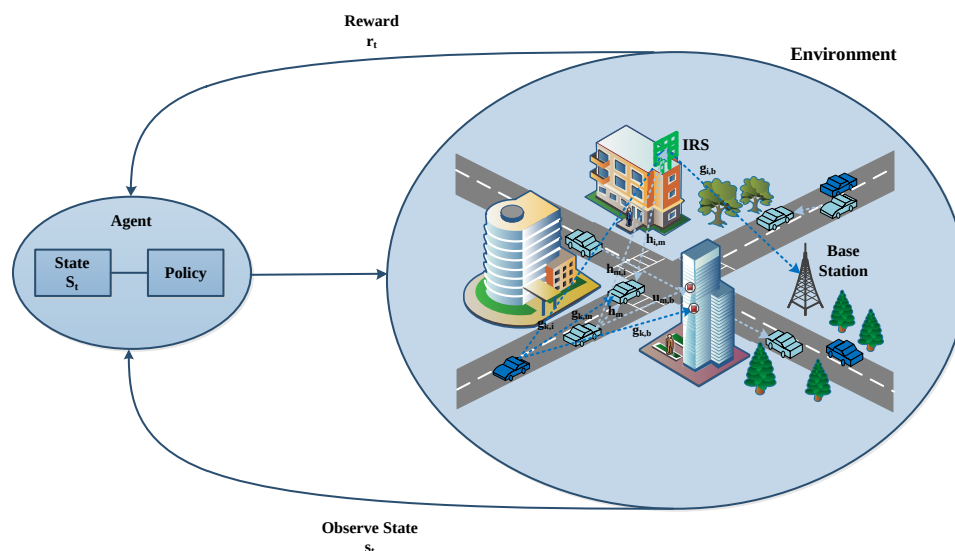


Figure 2. RL architecture: decentralized V2V agents interact with the wireless environment at each time slot, while a centralized IRS agent updates Φ at a slower timescale.

The hierarchy arises naturally from practical deployment constraints: the IRS is a shared passive surface that must be centrally configured, whereas V2V links can adapt their transmission decisions independently at a much faster timescale. This separation reduces signaling overhead and improves learning stability under fast channel variations by decoupling fast V2V adaptations from slower IRS reconfiguration. The centralized IRS agent also matches practical deployment as the IRS is passive and cannot run DRL locally, while the BS can compute and broadcast IRS configurations through a dedicated control link.

Specifically, the two-level agents are defined as:

- V2V agents (decentralized): each V2V link $m \in \mathcal{M}$ is an agent selecting its sub-band reuse decision and transmit power per time slot.
- IRS agent (centralized): a single IRS controller (co-located with the BS) configures the IRS phase-shift matrix Φ at a slower timescale (e.g., once per coherence interval or per episode).

4.1. MDP Formulation of the Problem

The overall interaction is modeled as a Markov decision process (MDP). At time slot t , the environment state evolves based on wireless channel dynamics, vehicle mobility, and all agents' actions. The objective of learning is to maximize the expected cumulative discounted return. The hierarchical design further separates the decision-making timescales: V2V agents update per slot, while the IRS agent updates per coherence block and keeps Φ fixed across multiple slots.

4.1.1. V2V Agents (State, Action, Policy and Reward)

Each V2V link acts as an agent. In decentralized settings, each V2V agent has no direct control over the actions of other V2V links; hence, other links' actions are considered part of the environment.

State

The observed state $s_m(t)$ of V2V agent m at time t includes channel/interference information and latency-related variables as given below:

$$s_m(t) = [g_m(t), I_m(t-1), h_m(t), x_{\mathcal{N}_m}(t-1), L_m(t), T_m(t)], \quad (14)$$

where $g_m(t)$ denotes the relevant V2V channel information (including IRS-assisted components under the current Φ), $I_m(t-1)$ is the received interference power in the previous time slot, $h_m(t)$ includes the V2I-related channel/interference observation needed to protect V2I links, and $\{L_m(t), T_m(t)\}$ are the remaining payload and remaining time defined in Section III. $x_{\mathcal{N}_m}(t-1)$ denotes the previous sub-band reuse decisions of the neighboring V2V links of agent m . This information enables each agent to account for recent resource usage by nearby links in a decentralized manner.

It is important to note that all state variables of the V2V agent correspond to information that is either locally measurable at the transmitting vehicle or obtainable through standard control signaling. Specifically, the channel observations are limited to the locally observable channels associated with the serving and neighboring links of the agent. The interference level is measured locally, neighboring spectrum occupancy can be inferred through spectrum sensing or periodic control broadcasts, and the remaining payload and transmission deadline are maintained locally by the V2V transmitter. Therefore, the decentralized V2V agents do not require instantaneous global channel state information, complete network knowledge, or the actions of other V2V agents.

Action

At each time slot, V2V agent m selects one sub-band to reuse and one transmit power level:

$$a_m(t) = (k_m(t), p_m^{\text{V2V}}(t)), \quad (15)$$

where $k_m(t) \in \{1, \dots, K\}$ indicates the selected V2I sub-band and $p_m^{\text{V2V}}(t)$ is selected from a discrete set. Hence, for N_{SB} sub-bands and N_P power levels, the action space size is $|\mathcal{A}_m| = N_{RB} \times N_P$.

Policy and Return

Each V2V agent follows an ϵ -greedy policy derived from a Q-function:

$$a_m(t) = \arg \max_{a \in \mathcal{A}_m} Q_m(s_m(t), a; \theta_m), \quad (16)$$

and aims to maximize the expected discounted return. Since the V2V payload must be delivered within a finite deadline, the interaction is episodic with a finite horizon T time slots (corresponding to $T_{\max}$). Thus, we use

$$G_m(t) = \mathbb{E} \left[\sum_{n=0}^{T-1} \beta^n r_m(t+n) \right], \quad (17)$$

where $\beta \in (0, 1)$ is the discount factor. This episodic return is consistent with standard discounted RL and naturally captures the latency-constrained delivery process.

Reward

The V2V agents must satisfy latency constraints while minimizing interference to V2I links. Its reward function is defined as

$$r_m(t) = \begin{cases} R_{V2I}(t), & \text{if the V2V latency constraint is satisfied,} \\ r_n, & \text{otherwise,} \end{cases} \quad (18)$$

where $R_{V2I}(t) = \sum_{k=1}^K W \log_2(1 + \gamma_k^{V2I}(t))$ and $r_n < 0$ is a penalty. The latency constraint satisfaction can be checked using the payload–deadline evolution in Section III (e.g., success if the payload is delivered before the deadline; otherwise, a violation).

The adopted reward formulation reflects the hard deadline requirements of safety-critical V2V services, where successful packet delivery before the deadline is the primary performance objective. Similar latency-aware reward formulations have been widely adopted in DRL-based V2V resource allocation to enforce stringent latency constraints while simultaneously improving V2I communication performance [2,23]. Although more sophisticated reward shaping based on remaining latency, packet delivery progress, or other QoS indicators could further improve learning efficiency and convergence, the adopted formulation provides a simple, interpretable, and practically motivated objective that is well suited for enforcing hard latency constraints in the considered V2V communication scenario.

4.1.2. IRS Agent (State, Action, Policy and Reward)

The IRS agent is centralized and operates at a slower timescale. Let t denote the global time-slot index and τ the IRS update index (e.g., one IRS action per coherence block). Upon each IRS update, the phase-shift matrix $\Phi(\tau)$ is configured and remains fixed while the V2V agents perform resource allocation over multiple time slots.

State

The IRS agent uses observations available at the BS. A practical state choice consistent with the system model is

$$s_{\text{IRS}}(\tau) = \left[\{\gamma_k^{V2I}(\tau)\}_{k=1}^K, \{\xi_m^{V2V}(\tau)\}_{m=1}^M, \Phi(\tau-1) \right], \quad (19)$$

where $\gamma_k^{V2I}(\tau)$ denotes the measured V2I SINR (or equivalently the V2I rate) at the BS under $\Phi(\tau-1)$, which can be obtained through standard pilot-based signal measurements and interference-plus-noise power estimation, without requiring explicit knowledge of individual V2V transmit powers or channel gains. The second component $\{\xi_m^{V2V}(\tau)\}_{m=1}^M$ represents latency-related quality-of-service indicators of the V2V links. Since the V2V latency deadline $T_{\max}$ may span multiple IRS update intervals, $\{\xi_m^{V2V}(\tau)\}_{m=1}^M$ is defined as a progress-based latency indicator computed at the end of the τ -th IRS interval,

$$\xi_m^{V2V}(\tau) = \frac{L_m(t_\tau)}{T_m(t_\tau) + \epsilon}, \quad (20)$$

where t_τ denotes the time-slot index at which the IRS updates, $L_m(t_\tau)$ and $T_m(t_\tau)$ are the remaining payload and remaining time defined in Section III, and $\epsilon > 0$ is a small constant. This indicator quantifies the urgency of V2V link m to meet its latency deadline and can be obtained at the BS through low-rate reporting of the remaining payload and remaining time (or an equivalent latency-risk metric) from vehicles, without requiring instantaneous global V2V CSI at the BS.

It should be emphasized that the decentralized information assumption applies only to the V2V agents. In contrast, the IRS controller is centralized and co-located with the

BS, where network-level measurements such as the V2I SINRs and V2V latency indicators are available through standard pilot measurements and periodic status reports from the V2V links. Accordingly, the IRS state is constructed from measurements that are available at the BS, making it fully consistent with the centralized role of the IRS controller and complementary to the decentralized operation of the V2V agents.

Action

The IRS agent selects the phase-shift matrix

$$\Phi(\tau) = \text{diag}\left(e^{j\theta_1(\tau)}, \dots, e^{j\theta_N(\tau)}\right), \quad (21)$$

Each reflecting element is assumed to employ finite-resolution phase control, where the phase of each element $\theta_n(\tau)$ is selected from a finite-resolution phase set $\mathcal{P}_\theta$. Practical IRS implementations commonly employ finite-resolution discrete phase shifters because of hardware complexity, power consumption, and implementation cost constraints [28]. Furthermore, recent studies have shown that scalable large-scale IRS deployments commonly exploit predefined phase-shift codebooks [29,30] or grouped element control strategies [31] to reduce implementation complexity and signaling overhead.

Policy

The IRS agent follows an ϵ -greedy policy derived from a Q-function. At each IRS update index τ , the phase-shift configuration is selected as

$$\Phi(\tau) = \arg \max_{\Phi \in \mathcal{A}_{\text{IRS}}} Q_{\text{IRS}}(s_{\text{IRS}}(\tau), \Phi; \theta_{\text{IRS}}), \quad (22)$$

where $\mathcal{A}_{\text{IRS}}$ denotes the finite action space of the IRS controller, with each action representing a feasible IRS reflection configuration. Under the adopted finite-resolution phase model, the IRS optimization is formulated as a finite-action decision problem over $\mathcal{A}_{\text{IRS}}$.

The IRS agent aims to maximize the expected long-term return associated with V2I throughput while accounting for the latency requirements of V2V links, as reflected in the IRS reward function. The ϵ -greedy strategy enables exploration of different IRS configurations during training, while the learned policy is applied greedily during testing.

Reward

The IRS agent aims to improve the V2I sum rate while supporting the latency requirements of V2V communications. Accordingly, the IRS reward function is defined as

$$r_{\text{IRS}}(\tau) = \bar{R}_{\text{V2I}}(\tau) - \lambda \sum_{m=1}^M \xi_m^{\text{V2V}}(\tau), \quad (23)$$

where $\bar{R}_{\text{V2I}}(\tau)$ denotes the average V2I sum rate achieved during the time interval governed by $\Phi(\tau)$, $\lambda > 0$ is a weighting factor, and $\xi_m^{\text{V2V}}(\tau)$ is the progress-based latency indicator of V2V link m defined in the IRS state. This reward formulation encourages the IRS to configure the phase shifts so as to maximize V2I throughput while reducing the urgency of V2V links to meet their latency deadlines.

4.2. Deep Q-Networks

Due to the high-dimensional state space and the finite-action formulation adopted for both the decentralized V2V agents and the centralized IRS controller, deep Q-networks (DQNs) are adopted as the learning framework.

In DQNs, a deep neural network parameterized by θ is used to approximate the action-value function $Q(s, a; \theta)$. The network parameters are optimized by minimizing the mean squared temporal-difference error over a mini-batch of transitions sampled from a replay memory:

$$\text{Loss}(\theta) = \sum_{(s, a, r, s') \in \mathcal{B}} (y - Q(s, a; \theta))^2, \quad (24)$$

where $\mathcal{B}$ denotes a mini-batch of experience tuples (s, a, r, s') , and r is the immediate reward associated with the sampled transition (i.e., $r = r_m(t)$ for V2V agents and $r = r_{\text{IRS}}(\tau)$ for the IRS agent). The target value y is defined as

$$y = r + \beta \max_{a' \in \mathcal{A}} Q(s', a'; \theta^-), \quad (25)$$

with $\beta \in (0, 1)$ denoting the discount factor and θ^- representing the parameters of a target network that is periodically updated.

The use of experience replay mitigates the temporal correlation among consecutive training samples, while the periodically updated target network stabilizes learning by decoupling the target values from the rapidly changing online network parameters, following the standard DQN framework proposed in [32]. For each agent, the replay memory stores its own state–action–reward transitions, while the same DQN update rule is applied independently to each agent.

This DQN framework is adopted for all agents in the proposed hierarchical architecture, despite their different decision-making timescales and state–action definitions.

It is worth emphasizing that the proposed framework does not rely on exhaustive evaluation of all feasible IRS reflection configurations. Instead, the IRS controller progressively learns an approximate control policy by iteratively approximating the action-value function from sampled state–action transitions using neural-network function approximation together with experience replay and target-network updates [32]. Consequently, the reported convergence reflects the learning behavior of the proposed controller under the considered simulation settings without requiring explicit exploration of the complete IRS configuration space. It is nevertheless acknowledged that recent actor–critic algorithms, such as DDPG, PPO, SAC, and MADDPG, have been widely adopted for continuous-action optimization problems. However, practical IRS implementations commonly employ finite-resolution discrete phase shifters because of hardware complexity, power consumption, and implementation cost constraints [28], thereby formulating the IRS optimization as a finite-action decision problem. Future work will investigate incorporating continuous-action reinforcement learning methods for continuous-phase IRS optimization, as well as more scalable finite-action representations, such as predefined phase-shift codebooks [29,30] and grouped element control strategies [31], within the proposed hierarchical framework.

4.3. Training and Testing of the Hierarchical Multi-Agent DQN

This subsection describes the training and testing workflow of the proposed hierarchical multi-agent DQN. An *episode* corresponds to one V2V transmission task with a fixed latency deadline of T time slots (i.e., $T = T_{\text{max}}$), during which the V2V payload must be delivered. Within each episode, the IRS is updated at a slower timescale over IRS intervals of length T_{IRS} time slots, selected according to the channel coherence time and practical IRS control constraints. Accordingly, the total number of IRS intervals per episode is given by

$$T_{\tau} = \left\lceil \frac{T}{T_{\text{IRS}}} \right\rceil. \quad (26)$$

During each IRS interval, the IRS phase-shift matrix remains fixed, while the V2V agents perform per-slot resource allocation.

Training and testing samples are generated by an environment simulator that captures vehicle mobility, time-varying channels, and mutual interference among V2I and V2V links under the applied IRS configuration.

4.3.1. Mitigation of Simultaneous-Decision Non-Stationarity

In multi-agent V2V networks, if all V2V links update their actions simultaneously based on local observations, the learning environment becomes non-stationary from the viewpoint of each agent, since the transition dynamics depend on the evolving policies of other agents. This non-stationarity can severely degrade the convergence and stability of independent learning.

To mitigate this issue, we adopt an *asynchronous V2V action update mechanism* similar to [2]. At each time slot, only one (or a small randomly selected subset) of V2V agents is allowed to update its action, while the remaining agents keep their previous actions unchanged. This allows the interference variations caused by other agents to be observed through subsequent measurements (including $x_{\mathcal{N}_m}(t-1)$), thereby improving the consistency between each agent's actions and its realized rewards.

4.3.2. Training Procedure

Algorithm 1 summarizes the training procedure. At the beginning of each IRS interval τ , the IRS agent selects a phase-shift configuration $\Phi(\tau)$ and keeps it fixed for T_{IRS} consecutive time slots. Within each IRS interval, V2V agents interact with the environment at the slot level using asynchronous updates. Each agent maintains its own replay memory and target network, and the DQN update rule is applied independently using the agent-specific reward: $r = r_m(t)$ for V2V agents and $r = r_{\text{IRS}}(\tau)$ for the IRS agent. To ensure stable training, network updates are performed only after the replay memories has accumulated a sufficient number of transitions to form a valid mini-batch, i.e., $|D_m| \geq \bar{\mathcal{B}}_m$ for V2V agent, and $|D_{\text{IRS}}| \geq \bar{\mathcal{B}}_{\text{IRS}}$ for IRS agent.

Algorithm 1 Training of Hierarchical Multi-Agent DQN with Asynchronous V2V Updates

- 1: **Initialize** IRS Q-network parameters θ_{IRS} and target $\theta_{\text{IRS}}^- \leftarrow \theta_{\text{IRS}}$, replay memory D_{IRS}
 - 2: **For each V2V agent** $m \in \mathcal{M}$: initialize Q-network θ_m , target $\theta_m^- \leftarrow \theta_m$, and replay memory D_m
 - 3: **for** episode $e = 1, 2, \dots, E$ **do**
 - 4: Initialize environment: vehicle locations, channels gains, and V2V payload/deadline states
 - 5: Set global slot index $t \leftarrow 0$; initialize actions $\{a_m(0)\}_{m=1}^M$
 - 6: **for** IRS interval $\tau = 0, 1, \dots, T_\tau - 1$ **do**
 - 7: Observe IRS state $s_{\text{IRS}}(\tau)$
 - 8: Select IRS action $\Phi(\tau)$ using ϵ -greedy policy
 - 9: Apply $\Phi(\tau)$ for the next T_{IRS} time slots
 - 10: **for** intra-interval slot $\ell = 1, 2, \dots, T_{\text{IRS}}$ **do**
 - 11: Set global slot index $t \leftarrow \tau T_{\text{IRS}} + (\ell - 1)$
 - 12: Select updating subset $\mathcal{V}(t) \subseteq \mathcal{M}$
 - 13: **for each** $m \in \mathcal{V}(t)$ **do**
 - 14: Observe $s_m(t)$ and select $a_m(t)$ using ϵ -greedy policy
 - 15: **end for**
-

Algorithm 1 *Cont.*

```

16:   for each  $m \notin \mathcal{V}(t)$  do
17:      $a_m(t) \leftarrow a_m(t-1)$ 
18:   end for
19:   Execute joint actions and update  $\{\gamma_k^{V2I}(t), \gamma_m^{V2V}(t), R_m^{V2V}(t), L_m(t), T_m(t)\}$ 
20:   for each  $m \in \mathcal{V}(t)$  do
21:     Observe  $r_m(t)$  and  $s_m(t+1)$ ; store transition in  $D_m$ 
22:     if  $|D_m| \geq \mathcal{B}_m$  then
23:       Sample mini-batch  $\mathcal{B}_m$  from  $D_m$ 
24:       Update  $\theta_m$  via DQN loss minimization
25:     end if
26:   end for
27:   end for
28:   Compute IRS reward  $r_{\text{IRS}}(\tau)$  and observe  $s_{\text{IRS}}(\tau+1)$ 
29:   Store IRS transition in  $D_{\text{IRS}}$ 
30:   if  $|D_{\text{IRS}}| \geq \mathcal{B}_{\text{IRS}}$  then
31:     Sample mini-batch  $\mathcal{B}_{\text{IRS}}$  from  $D_{\text{IRS}}$ 
32:     Update  $\theta_{\text{IRS}}$  via DQN loss minimization
33:   end if
34: end for
35: end for
36: Output: trained networks  $\{\theta_m\}_{m=1}^M$  and  $\theta_{\text{IRS}}$ 

```

4.3.3. Policy Execution for V2V Resource Allocation and IRS Optimization

Algorithm 2 summarizes the execution of the learned hierarchical policies for V2V resource allocation and IRS phase-shift optimization. This execution corresponds to the testing phase, during which learning and exploration are disabled and all agents act greedily using the trained DQN models. The IRS agent updates the phase-shift matrix $\Phi(\tau)$ at the beginning of each IRS interval, while the V2V agents perform per-slot spectrum reuse and power allocation using the same asynchronous update schedule as in training to ensure consistent interaction dynamics.

Algorithm 2 Policy Execution for V2V Resource Allocation and IRS Phase-Shift Optimization

```

1: Load trained networks  $\{\theta_m\}_{m=1}^M$  and  $\theta_{\text{IRS}}$ 
2: for testing episode  $e = 1, 2, \dots, E_{\text{test}}$  do
3:   Reset environment; initialize  $t \leftarrow 0$ 
4:   for IRS interval  $\tau = 0, 1, \dots, T_\tau - 1$  do
5:     Select  $\Phi(\tau) = \arg \max_{\Phi} Q_{\text{IRS}}(s_{\text{IRS}}(\tau), \Phi; \theta_{\text{IRS}})$ 
6:     for intra-interval slot  $\ell = 1, 2, \dots, T_{\text{IRS}}$  do
7:       Set global slot index  $t \leftarrow \tau T_{\text{IRS}} + (\ell - 1)$ 
8:       Select updating subset  $\mathcal{V}(t) \subseteq \mathcal{M}$ 
9:       for each  $m \in \mathcal{V}(t)$  do
10:         $a_m(t) = \arg \max_a Q_m(s_m(t), a; \theta_m)$ 
11:      end for
12:      Execute actions  $\{a_m(t)\}_{m=1}^M$  and update metrics
         $\{R_k^{V2I}(t), R_m^{V2V}(t), L_m(t), T_m(t)\}$ 
13:    end for
14:  end for
15: end for
16: Output: evaluation results

```

5. Simulation Setup

We evaluate the proposed IRS-aided V2X resource allocation framework in a dynamic urban road scenario. Vehicles are randomly distributed on a $450 \times 650 \text{ m}^2$ road grid containing bidirectional horizontal/vertical roads with two lanes per direction, and vehicles move following the intersection turning/straight-driving behavior commonly used in urban vehicular simulations [14].

The BS and IRS positions are fixed [13,14,17] with 3D coordinates set to (180 m, 270 m, 25 m) and (290 m, 380 m, 25 m), respectively. M vehicles form M V2I links (uplink), and K V2V links are created by pairing vehicles with neighbors nearby. Both direct and IRS-assisted links are considered.

To study the reliability–latency behavior of V2V links, we model the V2V transmission deadline constraint as successful if the payload of size B bits can be delivered within the maximum time budget $T_{\max}$, i.e.,

$$R_m^{V2V} \geq \frac{B}{T_{\max}}, \forall m, \quad (27)$$

where R_m^{V2V} is the achievable V2V rate for pair m .

Unless otherwise stated, the V2V payload size B is 1060 bytes, the number of V2I links K is 6, the number of V2V links M is 20, the number of IRS elements N is 100, and the V2I transmit power P_k^{V2I} is 25 dBm [14]. The remaining simulation parameters are set similar to the studies in [14,23] and are summarized in Table 2.

Table 2. Parameters used in simulations.

Parameter	Value
Carrier frequency	2 GHz
Bs antenna height	25 m
Bs antenna gain	8 dBi
Bandwidth of sub-band, W	1.5 MHz
Latency constraints for V2V, $T_{\max}$	100 ms
Noise power	−114 dBm
V2V maximum transmit power, $P_{\max}^{V2V}$	23 dBm
Penalty of latency constraints	−20
Vehicle speed	36 km/h

The network architecture and training configuration adopted for the DQN agents are summarized in Table 3. Unless otherwise stated, the same DQN configuration is employed for both the decentralized V2V agents and the centralized IRS controller. The selected parameter values are based on commonly adopted configurations in DRL-based V2V resource allocation studies [2,23], with minor empirical adjustments for the considered IRS-assisted V2X environment.

Table 3. DQN architecture and training parameters.

Category	Parameter	Value
Network Architecture	Number of hidden layers	2
	Neurons per hidden layer	256
	Hidden activation function	ReLU
	Output activation function	Linear

Table 3. Cont.

Category	Parameter	Value
Training Configuration	Optimizer	Adam
	Learning rate	0.001
	Replay memory size	50,000 experience tuples
	Mini-batch size	64
	Discount factor (β)	0.99
	Target network update interval	Every 200 training iterations
	Training episodes (E)	1000
Exploration Strategy	Initial exploration rate (ϵ)	1.0
	Final exploration rate (ϵ)	0.05
	Exploration policy	ϵ -greedy

6. Results and Discussion

This section evaluates the performance of the proposed decentralized hierarchical multi-agent DRL (DH-MDRL) framework for IRS-assisted V2X communications. The performance is measured using two main metrics: the probability of satisfied V2V links and the sum rate of V2I links. These two metrics represent the key goals of the considered system, namely reliable V2V transmission and efficient V2I communication.

Two baseline schemes are considered to evaluate the performance of the proposed hierarchical distributed framework. The first baseline is a centralized DRL approach adapted from the STAR-RIS framework in [13]. To ensure a fair comparison with the proposed passive reflective IRS system, only the reflective operation is considered while the STAR-RIS transmission functionality is disabled. The centralized controller jointly determines the resource allocation decisions for all V2V links and optimizes the IRS reflection configuration. Consequently, both approaches operate under identical passive reflective IRS assumptions, differing only in their resource allocation strategy. The centralized scheme, however, requires complete channel state information together with centralized coordination among all V2V links, resulting in greater signaling and computational requirements than the proposed hierarchical framework. The second baseline represents a system without RIS assistance. In this case, resource allocation follows the decentralized DRL-based approaches proposed in [14,23], but without the presence of a RIS. This baseline reflects a conventional V2X communication scenario where the wireless channel cannot be dynamically reconfigured.

Figure 3a,b show the effect of the number of IRS reflecting elements on system performance. As the number of IRS elements increases, both the V2I sum rate and the probability of satisfied V2V links improve for the RIS-assisted schemes. This is because a larger IRS provides more degrees of freedom to control the wireless channel, enhance useful signals, and suppress interference. The improvement in V2V satisfaction probability is especially important because reliable V2V delivery is the main requirement for safety-related vehicular applications. The proposed hierarchical distributed framework consistently achieves the best V2V satisfaction performance across all values of N . Although the centralized scheme obtains a slightly higher V2I sum rate, the difference remains very small. From a practical point of view, the throughput gain is achieved at the expense of centralized coordination and increased signaling requirements.

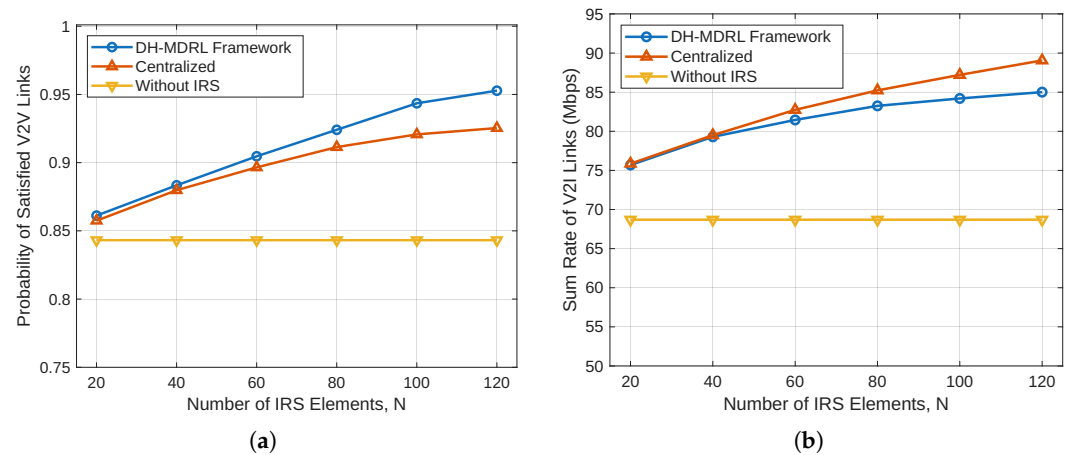


Figure 3. Impact of IRS size on communication performance. (a) Probability of satisfied V2V links versus the number of IRS reflecting elements N . (b) Sum rate of V2I links versus the number of IRS reflecting elements N .

Figure 4a,b illustrate the effect of V2V payload size. As the payload size increases, more data must be delivered within the same latency limit. Therefore, the probability of satisfied V2V links decreases for all schemes. At the same time, the V2I sum rate also decreases because larger V2V transmissions occupy more radio resources and create stronger interference. Even under this more demanding condition, the proposed framework maintains the highest V2V satisfaction probability among the compared methods. This result shows that the proposed distributed hierarchical design maintains reliable V2V communication performance as the traffic demand increases. The centralized method again gives a slightly higher V2I sum rate, but the performance gap is marginal. In practical vehicular systems, centralized operation requires global channel information and repeated coordination among participating nodes, which may become increasingly challenging as the network size grows.

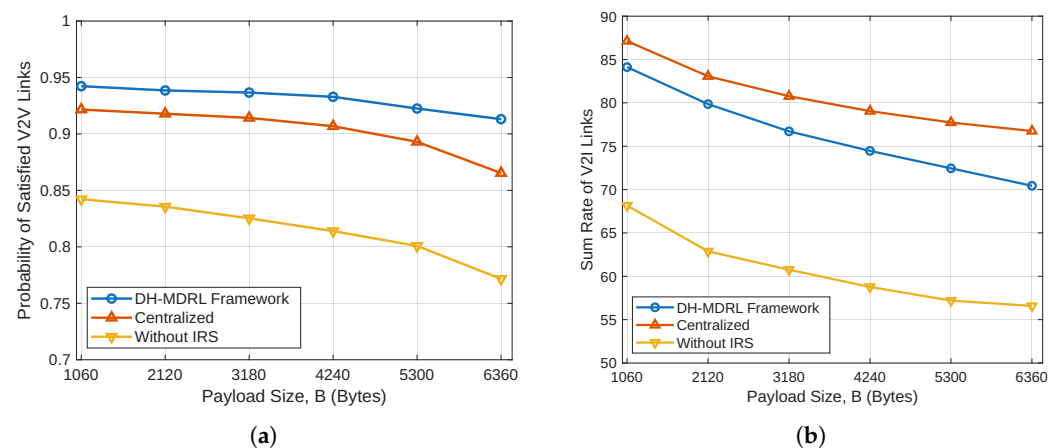


Figure 4. Impact of V2V payload size on system performance. (a) Probability of satisfied V2V links versus V2V payload size B . (b) Sum rate of V2I links versus V2V payload size B .

Figure 5 shows the effect of V2I transmit power on the V2I sum rate. As expected, the sum rate increases with transmit power because the received signal quality at the base station becomes stronger. The proposed hierarchical distributed framework achieves performance very close to the centralized method over the whole power range. This result confirms that the proposed approach can efficiently use available transmit power while preserving the distributed decision-making structure.

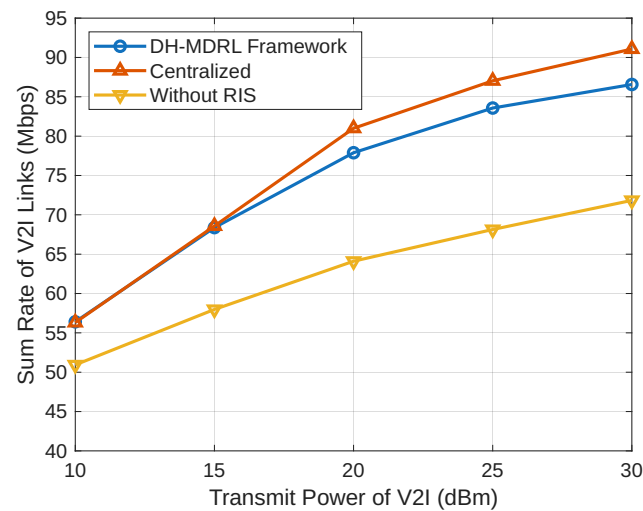


Figure 5. Sum rate of V2I links versus V2I transmit power P_{V2I} .

Finally, Figure 6a,b show the effect of the number of V2V pairs. These figures illustrate the performance of the proposed framework as the number of V2V pairs increases within the considered network scenario. As the number of V2V pairs increases, both the V2I sum rate and the V2V satisfaction probability decrease for all schemes due to stronger spectrum reuse and higher interference. For the V2I sum rate, the centralized scheme achieves a slightly better result, but the difference is again very small compared with the proposed method. This indicates that the proposed distributed framework can still maintain near-centralized throughput performance.

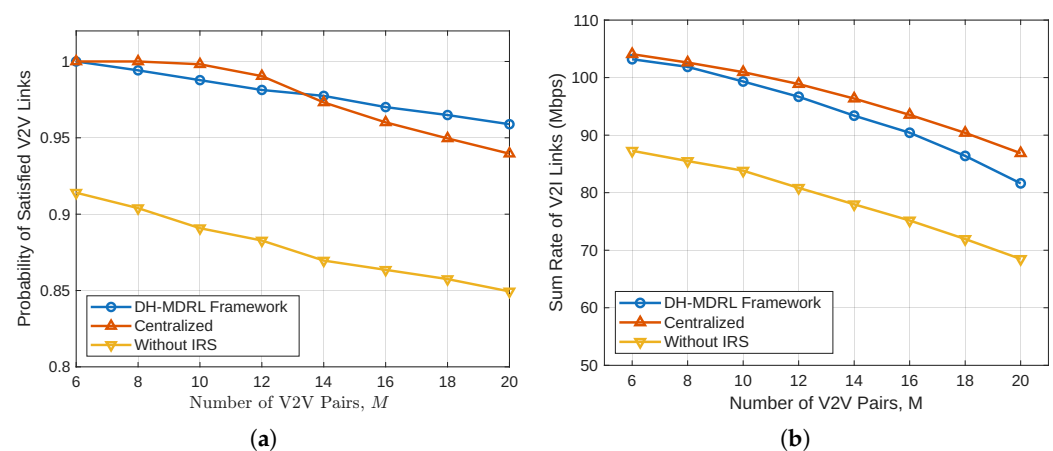


Figure 6. Impact of network density on communication performance. (a) Probability of satisfied V2V links versus the number of V2V pairs. (b) Sum rate of V2I links versus the number of V2V pairs.

For the probability of satisfied V2V links, a more interesting trend can be observed. When the number of V2V pairs is small, the centralized scheme may achieve slightly better performance. However, as the number of V2V pairs increases, its performance drops more quickly, while the proposed hierarchical distributed framework becomes superior. This result indicates that, within the considered range of V2V pairs, the proposed hierarchical framework maintains a higher V2V satisfaction probability than the centralized baseline as network loading increases. Overall, the proposed framework consistently provides competitive performance over the considered range of network densities.

Figure 7 illustrates the learning behavior of the proposed DH-MDRL framework during training. The solid curve represents the mean episode reward averaged over ten independent training runs, while the shaded region denotes the corresponding 95% confi-

dence interval. The average reward increases rapidly during the initial exploration phase and gradually stabilizes after approximately 400 training episodes. Thereafter, only small oscillations are observed around the steady-state reward, reflecting the stochastic nature of the wireless environment and the reinforcement learning process. Furthermore, the progressively narrowing confidence interval indicates reduced variability among independent training runs as learning progresses. Overall, these results indicate that the proposed framework exhibits stable and consistent learning behavior across independent training runs under the considered simulation settings.

Overall, the results demonstrate that the proposed hierarchical distributed framework provides a highly effective solution for resource allocation in RIS-assisted V2X networks. The proposed approach consistently achieves superior V2V reliability while maintaining V2I throughput very close to the centralized benchmark. Importantly, the small throughput advantage of the centralized scheme is achieved at the expense of centralized coordination and increased signaling requirements. Therefore, the proposed framework offers a practical hierarchical architecture that avoids centralized coordination while achieving competitive communication performance.

The present evaluation focuses on validating the overall performance of the proposed hierarchical framework through comparisons with representative centralized and decentralized baseline schemes. Although the comparison with the baseline without an IRS demonstrates the benefit of IRS-assisted channel reconfiguration, a more detailed ablation analysis of individual architectural components, including the two-timescale hierarchy, asynchronous V2V updates, and quantitative signaling overhead, is beyond the scope of the present work and will be considered in future studies.

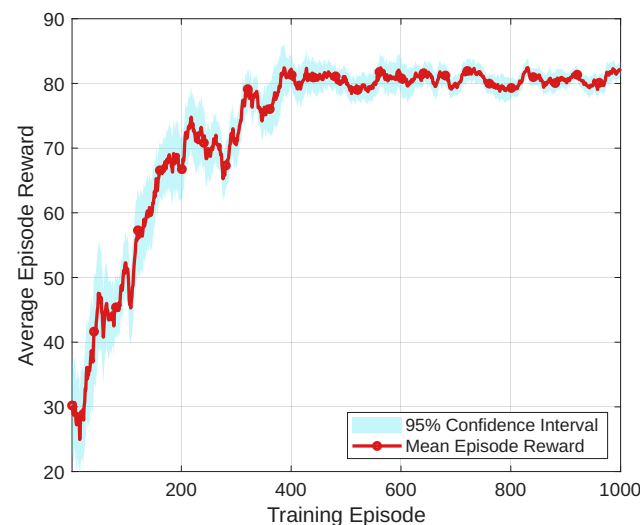


Figure 7. Convergence behavior of the proposed DH-MDRL framework averaged over ten independent training runs. The shaded region represents the 95% confidence interval.

7. Conclusions

This paper proposed a decentralized hierarchical multi-agent DRL framework for resource allocation in RIS-assisted V2X communications. The considered problem jointly addresses V2V transmit power control, spectrum reuse, and IRS phase-shift optimization in a highly dynamic vehicular environment with strict V2V latency and reliability requirements. A two-timescale hierarchical design was adopted to match the different dynamics of V2V resource allocation and IRS reconfiguration, where V2V links act as decentralized learning agents for fast resource allocation, while the IRS is configured by a centralized controller at a slower timescale.

The proposed framework was modeled as a Markov decision process and implemented using deep Q-networks for both the V2V agents and the IRS controller. Simulation results showed that the proposed method consistently improves the probability of satisfied V2V links under different network settings, including varying IRS size, V2V payload size, and number of V2V pairs. Although the centralized baseline sometimes achieves a slightly higher V2I sum rate, the performance gap is small, while the proposed hierarchical framework avoids centralized coordination and demonstrates competitive communication performance under the considered network settings.

Overall, the results confirm that combining decentralized DRL with IRS-assisted channel reconfiguration provides an effective and practical solution for future V2X systems. As future work, the proposed framework can be extended to more advanced scenarios such as continuous IRS phase optimization, imperfect channel state information, cooperative multi-IRS and multi-cell vehicular networks, and integrated vehicular edge computing environments.

Funding: This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Grant No. KFU263832].

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Nour, M.; Zaki, M.H.; Abdel-Aty, M. Assessing the Impact of Vehicle-to-Vehicle Communication on Lane Change Safety in Work Zones. *IEEE Open J. Intell. Transp. Syst.* **2025**, *6*, 832–847. <https://doi.org/10.1109/OJITS.2025.3578872>.
2. Ye, H.; Juang, B.H.; Li, G.Y. Deep Reinforcement Learning Based Resource Allocation for V2V Communications. *IEEE Trans. Veh. Technol.* **2019**, *68*, 3163–3173. <https://doi.org/10.1109/TVT.2019.2897134>.
3. Mekrache, A.; Kamal, A.E.; Alnuem, M.A. Deep Reinforcement Learning for Vehicular Networks: A Survey. *IEEE Commun. Surv. Tuts.* **2022**, *24*, 1514–1548.
4. Zhao, J.; Hu, F.; Li, J.; Nie, Y. Multi-agent deep reinforcement learning based resource management in heterogeneous V2X networks. *Digit. Commun. Networks* **2025**, *11*, 182–190. <https://doi.org/10.1016/j.dcan.2023.06.003>.
5. Basar, E.; Alexandropoulos, G.C.; Liu, Y.; Wu, Q.; Jin, S.; Yuen, C.; Dobre, O.A.; Schober, R. Reconfigurable Intelligent Surfaces for 6G: Emerging Hardware Architectures, Applications, and Open Challenges. *IEEE Veh. Technol. Mag.* **2024**, *19*, 27–47. <https://doi.org/10.1109/MVT.2024.3415570>.
6. Javed, M.A.; Chen, W.; Ma, S. Reliable V2X Communications Using Intelligent Reflecting Surfaces. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 13645–13657.
7. Huang, Z.; Zheng, B.; Zhang, R. Roadside IRS-Aided Vehicular Communication: Efficient Channel Estimation and Low-Complexity Beamforming Design. *IEEE Trans. Wirel. Commun.* **2023**, *22*, 5976–5989. <https://doi.org/10.1109/TWC.2023.3238850>.
8. Shabir, M.W.; Nguyen, T.N.; Mirza, J.; Ali, B.; Javed, M.A. Transmit and Reflect Beamforming for Max-Min SINR in IRS-Aided MIMO Vehicular Networks. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 1099–1105. <https://doi.org/10.1109/TITS.2022.3151135>.
9. Deng, Y.; Renzo, M.D.; Schober, R. Reconfigurable Intelligent Surfaces for Vehicular Communications: A Survey. *IEEE Commun. Surv. Tuts.* **2025**, *27*, 1–35.
10. Reshma, P.; Sudha, V.; Rafi, R.M. RIS-assisted V2I Communication for Inter-Cell Interference Mitigation Under Weibull Fading Channels. *Int. J. Intell. Transp. Syst. Res.* **2025**, *23*, 1080–1096. <https://doi.org/10.1007/s13177-025-00497-6>.
11. Ni, W.; Liu, X.; Liu, Y.; Tian, H.; Chen, Y. Resource allocation for multi-cell IRS-aided NOMA networks. *IEEE Trans. Wirel. Commun.* **2021**, *20*, 4253–4268.
12. Wu, Q.; Zhang, R. Intelligent Reflecting Surface Enhanced Wireless Network: Joint Active and Passive Beamforming Design. *IEEE Trans. Wirel. Commun.* **2021**, *20*, 3384–3399.
13. Aung, P.S.; Nguyen, L.X.; Tun, Y.K.; Han, Z.; Hong, C.S. Deep Reinforcement Learning-Based Joint Spectrum Allocation and Configuration Design for STAR-RIS-Assisted V2X Communications. *IEEE Internet Things J.* **2024**, *11*, 11298–11311. <https://doi.org/10.1109/JIOT.2023.3329893>.
14. Yuan, Y.; Zheng, G.; Wong, K.K.; Letaief, K.B. Meta-Reinforcement Learning Based Resource Allocation for Dynamic V2X Communications. *IEEE Trans. Veh. Technol.* **2021**, *70*, 8964–8977. <https://doi.org/10.1109/TVT.2021.3098854>.
15. Wang, Y.; Ge, X.; Ding, Y. Multi-Agent Reinforcement Learning for Resource Allocation in Vehicular Networks. *IEEE Trans. Veh. Technol.* **2022**, *71*, 5124–5137.

16. Lu, B.; Wu, Y.; Qian, L.; Zhou, S.; Zhang, H.; Lu, R. Multi-Agent DRL-Based Two-Timescale Resource Allocation for Network Slicing in V2X Communications. *IEEE Trans. Netw. Serv. Manag.* **2024**, *21*, 6744–6758. <https://doi.org/10.1109/TNSM.2024.3454758>.
17. Yang, Y.; He, Y.; Dutkiewicz, E. Hybrid Resource Allocation in RIS-Assisted V2X Networks via Enhanced DRL. *IEEE Trans. Veh. Technol.* **2025**, *75*, 7673–7687. <https://doi.org/10.1109/TVT.2025.3625409>.
18. Lan, J.; Jia, X.; Kou, Z. Age of Information and Energy-Efficiency Optimization in RIS-Assisted Vehicular Edge Computing via Hierarchical Deep Reinforcement Learning. *IEEE Internet Things J.* **2025**, *12*, 43681–43695. <https://doi.org/10.1109/JIOT.2025.3598039>.
19. Nguyen, H.T.; Duong, T.Q. Deep Reinforcement Learning for Intelligent Reflecting Surface-Assisted Wireless Systems. *IEEE Wirel. Commun. Lett.* **2021**, *10*, 935–939.
20. Sultana, A.; Fernando, X. Intelligent Reflecting Surface-Aided Device-to-Device Communication: A Deep Reinforcement Learning Approach. *Future Internet* **2022**, *14*, 256. <https://doi.org/10.3390/fi14090256>.
21. Yu, H.; Zhang, M.; Song, L. Multi-Agent Reinforcement Learning for IRS-Assisted Wireless Networks. *IEEE Trans. Wirel. Commun.* **2022**, *21*, 10245–10258.
22. Zhao, H.; Sun, W.; Ni, Y.; Xia, W.; Gui, G.; Zhu, C. Deep Deterministic Policy Gradient-Based Rate Maximization for RIS-UAV-Assisted Vehicular Communication Networks. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 15732–15744. <https://doi.org/10.1109/TITS.2024.3440391>.
23. Liang, L.; Ye, H.; Li, G.Y. Spectrum Sharing in Vehicular Networks Based on Multi-Agent Reinforcement Learning. *IEEE J. Sel. Areas Commun.* **2019**, *37*, 2282–2297. <https://doi.org/10.1109/jsac.2019.2933962>.
24. Liu, D.; Cao, Y.; Shi, Y.; Lei, Z. DRL-based resource allocation for max–min fairness in V2V network. *Phys. Commun.* **2025**, *73*, 102870. <https://doi.org/10.1016/j.phycom.2025.102870>.
25. Mao, J.; Liao, W.; Tan, C.; Liu, H.; Zheng, M. DRL-Based Task Offloading and Resource Allocation Strategy for Secure V2X Networking. *IEICE Trans. Commun.* **2025**, *E108-B*, 1190–1201. <https://doi.org/10.23919/transcom.2025EBP3010>.
26. Shi, Y.; Dong, H.; He, C.R.; Chen, Y.; Song, Z. Mixed Vehicle Platoon Forming: A Multiagent Reinforcement Learning Approach. *IEEE Internet Things J.* **2025**, *12*, 16886–16898. <https://doi.org/10.1109/JIOT.2025.3535732>.
27. Ji, B.; Wang, S.; Zhang, G.; Liu, J.; Mumtaz, S. Performance analysis of RIS-assisted V2V communication under multipath interference. *Veh. Commun.* **2024**, *45*, 100696. <https://doi.org/10.1016/j.vehcom.2023.100696>.
28. Wu, Q.; Zhang, R. Beamforming Optimization for Intelligent Reflecting Surface with Discrete Phase Shifts. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 7830–7833. <https://doi.org/10.1109/ICASSP.2019.8683145>.
29. Ghanem, W.R.; Jamali, V.; Schellmann, M.; Cao, H.; Eichinger, J.; Schober, R. Optimization-Based Phase-Shift Codebook Design for Large IRSs. *IEEE Commun. Lett.* **2023**, *27*, 635–639. <https://doi.org/10.1109/LCOMM.2022.3225585>.
30. An, J.; Xu, C.; Wu, Q.; Ng, D.W.K.; Di Renzo, M.; Yuen, C.; Hanzo, L. Codebook-Based Solutions for Reconfigurable Intelligent Surfaces and Their Open Challenges. *IEEE Wirel. Commun.* **2024**, *31*, 134–141. <https://doi.org/10.1109/MWC.010.2200312>.
31. Zhang, S.; Ji, T.; Hua, M.; Huang, Y.; Yang, L. Element-Grouping Strategy for Intelligent Reflecting Surface: Performance Analysis and Algorithm Optimization. *IEEE Trans. Wirel. Commun.* **2026**, *25*, 13274–13289. <https://doi.org/10.1109/TWC.2026.3661994>.
32. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, UK, 1998; Volume 1.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.