Multi-Agent Deep Reinforcement Learning for Optimal Resource Allocation in AoI-Aware Energy-Efficient Platoon-Based C-V2X Networks

Yuxiang Zheng, *Graduate Student Member, IEEE,* Long D. Nguyen, *Member, IEEE,* and Trung Q. Duong, *Fellow, IEEE*

***Abstract*—This paper aims to tackle the complex challenge of channel assignment and joint power-energy allocation within a cellular-vehicle-to-everything (C-V2X) network, which is deployed to manage vehicular dynamics at an urban traffic intersection. The primary function of the C-V2X network is to facilitate the coordination of multiple vehicle platoons formed by closely spaced same-lane vehicles. This coordination involves two critical communication tasks including the timely update of platoon states to a roadside unit (RSU) and the reliable exchange of cooperative awareness messages (CAMs) among vehicles within the same platoon. The main objective of this paper is to minimise the average age of information (AoI) to ensure the timely update between vehicle platoons and the RSU, maximise the CAM delivery probability (CDP) to guarantee the successful exchange of CAMs among vehicles, and promote sustainable, green communication practices through the implementation of our optimal power-energy management strategies. Recognising the intricate and dynamic nature of this challenge, we adopt a multi-agent deep reinforcement learning (MADRL) approach based on the Markov decision process (MDP). Two innovative algorithms based on the multi-agent deep deterministic policy gradient (MADDPG) and twin delayed deep deterministic policy gradient (TD3) algorithms are proposed to address this optimisation problem effectively. Finally, comprehensive simulation results are presented, which demonstrate the remarkable performance of our proposed schemes, particularly in terms of energy efficiency when compared to existing research. Importantly, these gains in energy efficiency are achieved while maintaining competitive algorithm convergence speeds, low AoI levels, and high CDP, showcasing the practical viability of the developed methods.**

***Index Terms*—Vehicle-to-everything, multi-agent deep reinforcement learning, age of information, resource allocation, green communication.**

I. Introduction

With the rapid advancements in autonomous systems and artificial intelligence (AI) technologies, self-driving vehicles and intelligent transportation systems (ITS) have emerged as critical elements in the blueprint of any smart city application, forming the backbone of future urban mobility infrastructure [1]. Consequently, they have become a focus of long-term interest and extensive study across multiple research domains, including wireless communications, control theory, and urban planning, spanning in various aspects of their implementation and real-world deployment challenges [1]–[7]. ITS holds considerable potential for alleviating traffic congestion through optimised route planning and traffic flow management. It can significantly alleviate the risk of car accidents via enhanced situational awareness and coordinated vehicle behaviour, and improve urban air quality by reducing emissions through more efficient transportation patterns [1]. To address the complex information transmission issues in ITS deployments, where real-time data must be processed and disseminated across complicated network architectures, vehicle-to-everything (V2X) communication technology [2] plays an increasingly important role in enabling seamless connectivity. This technology facilitates comprehensive connectivity through vehicle-to-vehicle (V2V) communication for direct inter-vehicle coordination, vehicle-to-pedestrian (V2P) links for enhanced safety, vehicle-to-infrastructure (V2I) connections for traffic management, and communications with cloud networks (V2N) for access to centralised traffic databases. These communication modes work together to broadcast almost real-time updates on surrounding transportation conditions, dynamic traffic states, and potential safety-related incidents, creating a comprehensive information ecosystem that supports informed decision-making at both individual vehicle and network levels.

One particularly effective pattern that strategically utilises the concept of V2X for the full potential of ITS is the platoon-based control strategy [3], which intelligently organises closely spaced same-lane self-driving vehicles into coordinated platoons. This approach enables vehicles to achieve more efficient traffic control through coordinated acceleration and braking patterns, and reach higher traffic flow rates by reducing intra-platoon vehicle gaps whilst maintaining safety margins. The platoon-based approach also provides a scalable framework for managing large numbers of autonomous vehicles in dense urban environments. In one pack of a vehicle platoon, the first vehicle is considered as the platoon leader (PL), which holds the critical responsibility for uploading platoon state messages including vehicle positions, velocities, and status information to the roadside unit (RSU) via V2I communication links, whilst receiving control commands and coordination instruc-

Y. Zheng and T. Q. Duong are with the Faculty of Engineering and Applied Science, Memorial University, St. John's, NL A1B 5S7, Canada (e-mail: y.zheng@mun.ca, tduong@mun.ca).

L. D. Nguyen is with Duy Tan University, Da Nang 550000, Vietnam (email: nguyendinhlong1@duytan.edu.vn).

This paper was presented in part at the IEEE International Workshop on Computer-Aided Modeling and Design of Communication Links and Networks (CAMAD), Athens, Greece, October 2024.

This work was supported in part by the Canada Excellence Research Chair (CERC) Program CERC-2022-00109 and in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant Program RGPIN-2025-04941.

Corresponding author is Trung Q. Duong.

tions from the RSU infrastructure. The PL also facilitates the timely exchange of cooperative awareness messages (CAMs) [4] with all other vehicles in the platoon through V2V links to maintain coordination, ensure string stability, and preserve safety margins during various driving conditions. This dual communication responsibility makes the PL a critical node in the platoon communication architecture, serving as both a gateway to infrastructure and an intra-platoon coordinator.

Keeping highly frequent and reliable updates of platoon states to the RSU is essential in this design architecture, as this continuous information flow allows the RSU to maintain effective control of time-critical safety information, traffic coordination data, and emergency response protocols across multiple platoons. Age of information (AoI) [5] is introduced as a fundamental measure of the freshness of information in time-sensitive systems, which continuously grows over time if the PL fails to update the RSU due to communication failures or resource allocation conflicts, and is reset to its minimum value after each successful transmission update. A comprehensive approach to the minimisation of AoI in vehicular networks has been developed in [6], showing the practical importance of maintaining fresh information in these dynamic and safety-critical transportation systems.

A. Related Works

Recent advances in V2X communications have addressed many challenges across autonomous driving, network architecture, and practical deployment. Recently, a dual-band approach has been proposed for V2X communication scheduling, wherein sub-6 GHz frequencies assist the millimetre-wave (mmWave) band to effectively combine the reliability of the former with the high capacity of the latter through optimised scheduling phase periods [8]. At the hardware level, the practical challenge of multi-band operation was addressed by designing a tri-frequency shared-aperture antenna in [9]. This design enables simultaneous operation of V2X and mmWave bands within a compact form factor suitable for vehicular integration. The application of cellular-V2X (C-V2X) in traffic management has also shown promising results. A hierarchical velocity optimisation framework was presented in a C-V2X network in [10], which enables connected automated vehicles (CAVs) to navigate signalised intersections more efficiently via real-time signal phase and timing information. In addition, the computational challenges of open radio access network (ORAN)-based C-V2X were addressed by developing a sparse-meta time series classifier that achieves efficient processing through model distillation [11]. Safety applications remain central to V2X research, as demonstrated in [12], where the authors developed generative models for detecting abnormal data patterns in C-V2X. Their approach effectively provides early warning of potential safety risks with high recall rate and F1 score. In industrial applications, C-V2X and vehicular edge computing have been adapted for intelligent obstacle detection in autonomous mining transportation [13]. In the convergence of communication and sensing, the fundamental trade-offs between communication performance and positioning accuracy in V2X sidelink joint communication and sensing systems were analysed in [14]. In addition, the following two papers provided more general studies on V2V and V2I resource allocation and multiple access schemes. The challenge of resource allocation in multi-cell 5G V2X systems, where V2V and V2I links coexist and vehicle roles are not fixed, was addressed in [15]. It proposed a centralised resource allocation scheme based on sparse code multiple access (SCMA), designed to maximise the packet reception ratio (PRR) for V2V safety communications. A roadside cooperative message delivery (RCMD) scheme for complex intersections using multicarrier multigroup multicast rate-splitting multiple access (M3RSMA) was proposed in [16]. This scheme addressed the often-overlooked problem of efficient message delivery by considering practical constraints such as finite message size, delay limits, and imperfect channel state information (CSI). Machine learning (ML)-based solutions for challenges in V2X have also shown their capability beyond conventional roadways. Transformer-based neural networks were applied for channel prediction in rate-splitting multiple access (RSMA)-based V2X systems in [17]. With the ability of the transformer architecture, the CSI is significantly more accurately predicted. A reinforcement learning (RL) method that specifically handles intermittent V2X communications was developed in [18], where the vehicular networks frequently experience connectivity disruptions that can impact autonomous vehicle performance. Furthermore, the intersection of vehicular networks and smart grids was explored by combining blockchain technology with federated RL (FRL) for V2X trading [19]. This approach enables secure and privacy-preserving energy transactions between vehicles and the grid.

Recent studies in platoon-based control strategies within V2X networks have focused on enhancing resource allocation efficiency, control coordination, and communication reliability. Focusing on platoon partition, power control and spectrum matching to achieve ultra-reliable platoon communications, the platoon cooperation in multi-lane V2X scenarios was investigated in [20]. The V2I capacity was maximised while the reliability requirement of the V2V communication was still maintained with their proposed approach. Instead of optimising the V2I communication, the ultra-reliable low-latency communication (URLLC) requirements for intra-platoon (i.e., V2V) safety data transmission were addressed in [21]. Low-complexity dynamic manager selection and resource allocation algorithms that consider finite block length effects to enhance communication performance and achieve near-optimal groupcast latency were proposed. Cooperative multipoint transmission (CoMP) techniques to facilitate seamless handovers and resource allocation in 5G-V2X platoon systems were also explored [22]. By coordinating transmissions among multiple base stations, the proposed approach mitigates handover disruptions and maintains communication quality during high-speed vehicular movements. In the domain of resource management using spectrum awareness, a spectrum sensing scheduling scheme was proposed [23]. This scheme managed communication resources efficiently in vehicle platooning by integrating a three-level platoon architecture (cloud, RSU, and vehicles). It also employed a greedy algorithm for

resource allocation to minimise platoon communication delay and error, particularly when vehicles join existing platoons. A bi-objective joint optimisation problem was tackled in [24] for optimal communication reliability and minimum traffic oscillation flow in cooperative vehicle-infrastructure systems. The intricate relationship between control signals and resource scheduling was investigated, and the capability of the co-design approach to enhance platoon control performance was demonstrated. A comprehensive approach to resource allocation in 5G platoon communication was provided in [25]. An improved random selection (IRS) scheme was developed to decrease the collision probability, and a deep deterministic policy gradient (DDPG) algorithm is proposed to manage the vehicle collaboration within a platoon based on local information. Similarly, there are more works that employ RL-based methods to coordinate vehicle platoons. A multi-agent deep RL (MADRL) approach to optimise channel assignment and power allocation was proposed in [3]. Platoons were considered as agents to select the optimal combination of sub-band and power level to transmit signals. This work showed the effectiveness of deep Q-learning and multi-agent DDPG (MADDPG) in dynamic resource allocation for C-V2X platoons. This is closely related to [26], where an RL framework was employed for dynamic PL selection, user association, channel assignment, and power allocation to enhance communication reliability in C-V2X based highway scenarios. Recently, a mixed vehicle platoon forming method has been proposed in [27]. This method was developed based on a two-stage control framework in order to adapt to dynamic mixed traffic environments. The multi-agent RL (MARL) was then used for the safe and efficient platoon forming control with the guidance of a feasible formation.

Within the domain of V2X, significant research has concentrated on optimising the AoI to ensure timely and effective communication in complex vehicular networks. The minimisation of AoI in non-orthogonal multiple access (NOMA)-based vehicular networks was investigated in [28], specifically in V2I scenarios. The proposed solution involved a hybrid orthogonal multiple access (OMA) and NOMA scheduling mechanism, which is governed by a Markov decision process (MDP) and designed to optimise information freshness and packet delivery. In the context of CAVs, an AoI-based service aggregation method was proposed, which aimed at ensuring information freshness and reducing computational load in highly mobile environments [29]. The proposed approach modelled AoI based on vehicle dynamics and clustered information sources, and maintained satisfactory data sequencing success rates and low latency. Furthermore, the impact of packet retransmissions on AoI and delay in the context of new radio-V2X (NR-V2X) wireless random access protocols was assessed in [30]. Based on MDPs, optimal retransmission parameters were identified to mitigate issues arising from stale data and communication bottlenecks. RL-based methods are also widely deployed in the minimisation of AoI. The reliable content delivery and AoI minimisation through the caching of road environment information in infrastructure-assisted connected vehicles were explored in [31]. A two-stage algorithm, which integrated AoI-aware content caching with delay-aware content delivery, was formulated using MDP and Lyapunov optimisation. The minimisation of AoI for critical safety information in C-V2X communications was also addressed [32]. A deep RL (DRL) approach was proposed to handle continuous resource allocation decisions (e.g., power allocation and broadcast coverage) and complemented by a matching algorithm for discrete resource block (RB) scheduling. Lastly, the optimisation of AoI in unmanned aerial vehicle (UAV)-enabled ITS was studied in [33]. DRL was employed, including a two-stage DRL framework and deep Q-networks (DQN), to jointly optimise computation resource allocation, offloading decisions, transmit power, and offloading ratio, with the objective of minimising execution delay and power consumption while preserving information freshness.

The joint optimisation of platoon-based control strategies and AoI metrics has gained prominence as a critical approach for enhancing the timeliness and reliability of inter-vehicular communications in C-V2X-enabled cooperative driving systems. An AoI-based beacon transmission scheme was proposed in [34], which was designed to mitigate status update delays caused by beacon loss within platoons. This scheme enabled vehicles to request additional beacon transmissions by considering their current AoI, thereby enhancing the reliability and stability of intra-platoon communication. AoI in V2V-enabled platooning systems was analysed, and a joint communication and control model was proposed in [35]. The analysis revealed the proportions of stable platoons and provided guidelines for designing V2V-enabled platooning systems by linking communication parameters under stability objectives. In addition, AoI-centric real-time information dissemination over vehicular social networks (VSNs) was also considered [36]. A mathematical framework was proposed based on the mean-field theory (MFT) to analyse the network AoI for the VSNs, and the information update rate at the source node and the transmit probabilities at the autonomous vehicles are jointly optimised to minimise the average peak network AoI. The issue of safety in platooning was explored in [37]. A safety-aware AoImetric was introduced with the real-time collision risk assessment of CAVs to design more efficient basic safety message (BSM) transmission protocols. Based on this design, the collision risk in cell-free massive multiple-input multiple-output (mMIMO) platooning environments was minimised. Finally, a distributed resource allocation framework with MARL was presented in [6]. The proposed modified MADDPG algorithms allowed individual PLs to manage radio resources in an AoI-aware manner, ensuring timely dissemination of CAMs and safety-critical messages to the RSU within C-V2Xplatoon networks.

B. Motivation and Contributions

In this work, we address the complex resource allocation problem within a platoon-based C-V2X network [38] at an intersection. While vehicle platooning is a key strategy for improving traffic efficiency, it introduces significant communication challenges. The tight coordination required among vehicles and between vehicles and the RSU necessitates both highly reliable CAM exchange, and extremely fresh information (i.e., a low level of AoI). These requirements often lead

to increased energy consumption for wireless communications. This creates a trade-off between communication performance and the amount of energy consumed, which must be managed carefully. Our research tackles this multi-factor challenge by concurrently maximising the CDP, and minimising the AoI and energy consumption. Building on foundational research in platoon communications [3], [6], our model integrates the distributed resource allocation of Mode 4 [39] and urban scenarios [40], [41].

The dynamic nature of this environment—characterised by rapid changes in vehicle positions and highly variable communication demands—makes traditional optimisation methods impractical. The problem is inherently non-convex and involves the distributed decision-making of multiple platoons, which makes the centralised control not feasible. Therefore, we employ an MADRL approach. Specifically, we use the extension of the standard DDPG—MADDPG framework, enhanced with several advanced techniques to address the complexities of the problem, including the decomposed MADDPG (DE-MADDPG) algorithm [42], task decomposition (TDec) algorithm [43], and twin delayed deep deterministic policy gradient (TD3) [44].

Conventional studies on vehicular networks have largely focused on the immediate impact of transmission power on communication quality, often treating it as a constraint rather than a primary optimisation objective. This perspective overlooks the broader, long-term implications of energy usage, which significantly affect the environmental footprint and operational costs of the network. Our research addresses this critical gap by adopting a holistic approach that treats the power and energy allocation as one of the core components of the optimisation problem. This approach enhances not only the effectiveness of operational communications but also the long-term sustainability of the C-V2X network. It aligns our work with the global shift towards green communication technologies, which are increasingly valuable for their environmental benefits [45].

Thus, the key contributions of this study to the fields of ITS and green communications, which introduce innovative techniques and methodologies, are detailed in Tab. I and summarised as follows:

- We design a comprehensive optimisation framework for platoon-based C-V2X networks that jointly considers V2I information freshness (AoI), V2V communication reliability (CDP), short-term power consumption, and long-term energy consumption. This framework provides a more realistic model by capturing the trade-offs between these competing objectives, moving beyond simple power constraints to better address energy sustainability and network efficiency.
- To solve the formulated problem, we propose a MADRL solution that combines MADDPG with several performance enhancement techniques. The integration of DE-MADDPG, TDec, and TD3 is specifically designed to overcome the challenges of multi-agent coordination and dynamic vehicular environment, leading to more effective and stable learning of resource allocation policies.

TABLE I: Overview of prior research and this work.

Studies	ITS	Vehicle Platooning	AoI	RL-based Solution	Energy Saving
[8]–[17]	✓	-	-	-	-
[18], [19]	✓	-	-	✓	-
[20]–[24]	✓	✓	-	-	-
[3], [25]–[27]	✓	✓	-	✓	-
[5], [28]–[30]	✓	-	✓	-	-
[31]–[33]	✓	-	✓	✓	-
[34]–[36]	✓	✓	✓	-	-
[6], [37]	✓	✓	✓	✓	-
This work	✓	✓	✓	✓	✓

- We provide extensive numerical results that validate the effectiveness of our energy-focused algorithms. The findings confirm that our methods considerably reduce energy usage compared to established benchmarks without compromising crucial performance metrics (convergence speed, AoI, and CDP). Additionally, we introduce a metric specifically for assessing energy efficiency, which further underscores the benefits of our approaches.

C. Paper Structure and Notations

The remainder of this paper is structured as follows. Section II presents the platoon-based C-V2X system model and formulates the optimisation problem. The preliminaries of the MADRL framework are described in Section III. In Section IV, two MADRL algorithms are proposed as a solution for the optimisation problem. The complexity of the proposed algorithms is analysed in Section V. Next, numerical results and discussions are provided in Section VI under various system configurations. Finally, the conclusion is given in Section VII.

Notation: In this paper, numbers are written as italic letters while vectors are denoted in bold case. $\mathbb{N}$ and $\mathbb{Z}^+$ denote the set of natural numbers and positive integers, respectively. $\mathcal{C}$ and I represent the channel capacity and the interference, respectively. $\Pr(X)$ is the probability that event X happens. $\mathcal{F}_x$ represents mapping functions, where $x \in \mathbb{Z}^+$. A list of abbreviations is provided in Tab. II.

II. System Model and Problem Formulation

A. Environment Model

Fig. 1 illustrates an intersection managed by a single-antenna RSU which locates at the centre, coordinating L platoons ($L \in \mathbb{N}$) via a C-V2X system. Each platoon $l \in \{1, 2, \ldots, L\}$ consists of V_l vehicles ($V_l \in \mathbb{Z}^+$). The head-of-line vehicle in each platoon, designated as the PL ($v_l = 1$, where $v_l \in \{1, 2, \ldots, V_l\}$), acts as the central communication agent for its group. This hierarchical system supports two different communication modes: in V2I mode, the RSU gathers states and relays states and commands based on the information uploaded by each PL, while in V2V mode, CAMs are exchanged between vehicles within the same platoon.

B. Communication Model

In this work, orthogonal frequency-division multiplexing (OFDM) is used to cope with the frequency-selective wireless channels [46]. The wireless channel is segmented into

TABLE II: List of Abbreviations.

Abbreviations	Definitions	Abbreviations	Definitions
AI	Artificial intelligence	mMIMO	Massive multiple-input multiple-output
AoI	Age of information	mmWave	Millimetre-wave
BSM	Basic safety message	NOMA	Non-orthogonal multiple access
CAM	Cooperative awareness message	NR-V2X	New radio-vehicle-to-everything
CAV	Connected automated vehicle	OFDM	Orthogonal frequency-division multiplexing
CDP	Cooperative awareness messages delivery probability	OMA	Orthogonal multiple access
CoMP	Cooperative multipoint transmission	ORAN	Open radio access network
CSI	Channel state information	PL	Platoon leader
C-V2X	Cellular-vehicle-to-everything	ReLU	Rectified linear unit
DDPG	Deep deterministic policy gradient	RL	Reinforcement learning
DE-MADDPG	Decomposed multi-agent deep deterministic policy gradient	RSMA	Rate-splitting multiple access
Dec-MADDPG	Decentralised multi-agent deep deterministic policy gradient	RSU	Roadside unit
DQN	Deep Q-networks	SINR	Signal-to-interference-plus-noise ratio
DRL	Deep reinforcement learning	TDec	Task decomposition
FRL	Federated reinforcement learning	TD3	Twin delayed deep deterministic policy gradient
ITS	Intelligent transportation systems	UAV	Unmanned aerial vehicle
MADDPG	Multi-agent deep deterministic policy gradient	URLLC	Ultra-reliable low-latency communication
MADRL	Multi-agent deep reinforcement learning	VSN	Vehicular social network
MARL	Multi-agent reinforcement learning	V2I	Vehicle-to-infrastructure
MATD3	Multi-agent twin delayed deep deterministic policy gradient	V2N	Vehicle-to-network
MDP	Markov decision process	V2P	Vehicle-to-pedestrian
MFT	Mean-field theory	V2V	Vehicle-to-vehicle
ML	Machine learning	V2X	Vehicle-to-everything

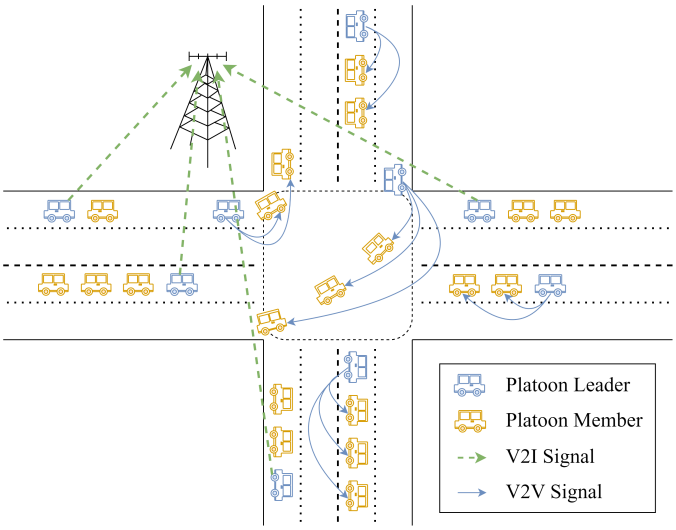

Fig. 1: A single-antenna multi-platoon C-V2X network.

K orthogonal subchannels, each with bandwidth W, where $K \in \mathbb{Z}^+$ and $\mathcal{K} = \{1, 2, \ldots, K\}$ denotes the set of subchannels. To align with the 5G NR standard and for simulation simplicity, we define each subchannel to be equivalent to one RB. The channel fading is independent across all subchannels and remains constant within a single coherence time, Δt. Accordingly, the timeline is divided into intervals of duration Δt, with $t \in \mathbb{Z}^+$ serving as the index for time steps. The channel gain for PL_l (leader of platoon l) in subchannel $k \in \mathcal{K}$ at time step t is expressed as $h_l^t[k] = \beta_l^t g_l^t[k]$, where β_l^t and $g_l^t[k]$ are the large and small scale fading, respectively, and the large-scale fading β_l^t is composed of path loss and shadowing. In addition, two decision parameters $\delta_{l,k}^t, \lambda_l^t \in \{0, 1\}$ are defined for the channel and communication mode selections. If $\delta_{l,k}^t = 1$, the subchannel k is allocated to the platoon l at time step t, PL_l will use it for V2I mode communication when $\lambda_l^t = 0$, and V2V mode communication when $\lambda_l^t = 1$. The channel capacity for V2I and V2V communications can hence be written as in [6]:

$$\begin{aligned} \mathcal{C}_{l,\mathcal{R}}^t[k] &= \log_2\left(1 + \frac{(1-\lambda_l^t)\delta_{l,k}^t p_l^t[k] h_{l,\mathcal{R}}^t[k]}{I_{l,\mathcal{R}}^t[k] + \sigma^2}\right), \\ I_{l,\mathcal{R}}^t[k] &= \sum_{l',l' \neq l} \delta_{l',k}^t p_{l'}^t[k] h_{l',\mathcal{R}}^t[k], \end{aligned} \tag{1}$$

$$\begin{aligned} \mathcal{C}_{l,v_l}^t[k] &= \log_2\left(1 + \frac{\lambda_l^t \delta_{l,k}^t p_l^t[k] h_{l,v_l}^t[k]}{I_{l,v_l}^t[k] + \sigma^2}\right), \\ I_{l,v_l}^t[k] &= \sum_{l',l' \neq l} \delta_{l',k}^t p_{l'}^t[k] h_{l',v_l}^t[k], v_l \in \mathcal{V}_l \backslash \{1\}, \end{aligned} \tag{2}$$

where the signal-to-interference-plus-noise ratio (SINR) is estimated from the interference of other platoons which is treated as noise. The power used by PL_l on subchannel k is denoted by $p_l^t[k]$. The channel gains from PL_l to the RSU and to other vehicles within platoon l are represented by $h_{l,\mathcal{R}}^t[k]$ and $h_{l,v_l}^t[k]$, respectively. The noise power is denoted by σ^2. The interference power experienced at the RSU and the vehicles within platoon l, denoted as $I_{l,\mathcal{R}}^t[k]$ and $I_{l,v_l}^t[k]$, is calculated from the transmit power $p_{l'}^t[k]$ of other PLs and the corresponding channel gains $h_{l',\mathcal{R}}^t[k]$ and $h_{l',v_l}^t[k]$.

C. Age of Information Model

AoI is crucial in ensuring that the PLs maintain timely information exchange with the RSU via V2I communication. This exchange is necessary for updating the platoon state and receiving control commands. Let A_l^t denote the AoI for platoon l at the t^{th} coherence time step. This metric quantifies the number of time steps since the platoon's last update from the V2I network [5]. The evolution of A_l^t is defined as

$$A_l^{t+1} = \begin{cases} 1, & \text{if } \mathcal{C}_{l,\mathcal{R}}^t[k] \geq \mathcal{C}_{l,\mathcal{R}}^{\min}, \\ A_l^t + 1, & \text{otherwise}, \end{cases} \tag{3}$$

where '1' represents one time step, $\mathcal{C}_{l,\mathcal{R}}^{\min}$ is the minimum required data transmission rate for V2I communication. If the

current transmission rate is less than the requirement, i.e., V2I communication fails or V2I mode is not selected, AoI will increase by 1. In contrast, if the requirement is satisfied, i.e., V2I communication is successful, AoI will be reset to 1.

D. Problem Formulation

Based on the preceding models, our goal is to find an optimal resource allocation policy for each PL_l. This policy must simultaneously manage four competing objectives over a time horizon T. This multi-objective optimisation problem is formulated for platoon l as follows:

$$\begin{aligned}\min_{\boldsymbol{\delta},\boldsymbol{\lambda},\boldsymbol{p},\boldsymbol{E}} \Bigg\{ &\sum_{t=1}^{T}\left[\frac{1}{T}A_l^t, \frac{1}{T}\sum_k p_l^t[k], \sum_k E_l^t[k]\right], \\ &-\Pr\left(\sum_{t=1}^{T}\sum_k \min_{v_l}\left\{\mathcal{C}_{l,v_l}^t[k]\right\}\Delta t \geq D_l\right)\Bigg\},\end{aligned} \tag{4}$$

$$\text{s.t. } p_l^t[k] \in [0, p^{max}], \forall t, l, k, \tag{4a}$$

$$\delta_{l,k}^t, \lambda_l^t \in \{0,1\}, \forall t, l, k, \tag{4b}$$

$$\sum_k \delta_{l,k}^t \leq 1, \forall t, l, \tag{4c}$$

$$\sum_l \sum_k \delta_{l,k}^t \leq K, \forall t, \tag{4d}$$

Here is a breakdown of the objectives and constraints:

- Minimise Average AoI: $\frac{1}{T}\sum_{t=1}^{T} A_l^t$ represents the average AoI. Minimising this ensures the RSU receives fresh and timely platoon-state updates.
- Minimise Average Power: $\frac{1}{T}\sum_{t=1}^{T}\sum_k p_l^t[k]$ is the average power consumed by PL_l. This addresses the short-term goal of power-efficient transmission.
- Minimise Total Energy: This term $\sum_{t=1}^{T}\sum_k E_l^t[k]$ is the total energy consumed, where $E_l^t[k] = p_l^t[k]\Delta t$. This addresses the long-term goal of sustainability and green communication. This focuses primarily on communication energy, though it is noted that energy for decision-making and other activities by PL_l could also be considered.
- Maximise CDP: $-Pr(...)$ represents the negative probability that the total V2V data transmitted is sufficient to deliver the entire CAM message (D_l) to all platoon members. We model this based on the member with the worst channel condition ($\min_{v_l}\left\{\mathcal{C}_{l,v_l}^t[k]\right\}$). Minimising this negative term is equivalent to maximising the CDP for intra-platoon V2V communication.
- (4a) Power Limit: Ensures the PL's transmit power on any subchannel k does not exceed the maximum allowed power p^{max}.
- (4b) Discrete Actions: States that the channel selection (δ) and communication mode selection (λ) are binary decisions.
- (4c) Single Channel Allocation: Limits each PL to select at most one subchannel at any given time step t.
- (4d) Total Channel Limit: Ensures the total number of subchannels allocated across all platoons does not exceed the total number of available subchannels, K.

The primary objective of this problem is to jointly optimise the four objectives under the four constraints within a designated time slot T—where the CAM dissemination frequency is selected as 10 Hz [4], [40], implying that the exchange period should be 100 ms. It is important to clarify that this 100 ms is the standard message generation interval for CAMs, not a latency requirement.

Given the non-convex nature of this multi-objective problem, the mixed discrete-continuous action space, and the dynamic environment, solving this optimisation problem with traditional methods is intractable. Therefore, we employ a MADRL approach, which will be explored in the subsequent sections.

III. Preliminaries of the MADRL algorithm

A. Markov Decision Process

From the problem (4) in Sec. II-D, the MADRL issue is modelled as a MDP [47], described by the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$, covering state and action spaces $\mathcal{S}, \mathcal{A}$, transition probability $\mathcal{P}$, reward function $\mathcal{R}$, and discount factor $\gamma \in [0,1]$. The formulation is as follows:

- *Agent*: The L platoons at the intersection collectively form the MADRL environment, with each platoon regarded as an agent. PLs observe states, take actions based on their policies, and optimise policies by interacting with the environment.
- *State space*: At time step t, observed by PL_l, the state space is defined as
$$\mathcal{S}_l^t = \begin{bmatrix} h_{l,\mathcal{R}}^t[k], h_{l,v_l}^t[k], T', D_l', A_l^t, \\ I_{l,\mathcal{R}}^t[k], I_{l,v_l}^t[k], \sigma^2, Local_l \end{bmatrix}, \tag{5}$$
where $T' \in [0,T]$ represents the remaining time in the time slot T, $D_l' \in [0, D_l]$ is the residual CAM data size to be transferred, and $Local_l$ indicates platoon l's location. Platoons dynamically update their locations at each time step t and make random movement decisions.[1]
- *Action space*: The action by PL_l comprises four components, represented as
$$\mathcal{A}_l^t = \left[\delta_{l,k}^t, \lambda_l^t, p_l^t[k], E_l^t[k]\right]. \tag{6}$$
Given the state space $\mathcal{S}_l^t$, PL_l chooses subchannel $k \in \mathcal{K}$, the communication mode (V2I/V2V), and regulates power and energy at time t. These actions must adhere to constraints (4a)–(4d). Due to the fact that adjusting power consumption is equivalent to adjusting energy consumption if only consider the communication energy, the agent will jointly consider the two factors $p_l^t[k]$ and $E_l^t[k]$ to take further action.
- *Transition probability*: The transition probability $\mathcal{P}$ captures the likelihood of state s moving to state s' upon action a by an agent. It includes: 1) Changes in interference within other PLs' state spaces due to the subchannel and power settings of PL_l for V2I/V2V communication;

[1]The inclusion of channel gains and interference in the state space is an idealisation. Our model assumes these estimated values are available to the PL at each time step, which is a common simplification in the resource allocation literature to maintain problem tractability. In a practical deployment, these would be acquired through estimation.

2) The inherent randomness of platoon movements (turning right/left or continuing straight), independent of the actions taken.

- *Reward function*: Each of the agents in this multi-agent environment receives a local reward as feedback for its action, while there is also a global reward that measures the joint performance of all the agents. Based on the optimisation problem (4), the local reward function for agent l is defined as

$$\begin{aligned}\mathcal{R}_l^t = &-\mathcal{F}_1\left(A_l^t\right) - \mathcal{F}_2\left(p_l^t[k]\right) - \mathcal{F}_3\left(E_l^t[k]\right) \\ &-\mathcal{F}_4\left(D_l'/D_l\right) + w\Gamma\left(\mathcal{C}_{l,\mathcal{R}}^t[k] - \mathcal{C}_{l,\mathcal{R}}^{\min}\right),\end{aligned} \tag{7}$$

where $\mathcal{F}_1$–$\mathcal{F}_4$ map the first four terms to the same range and weight their contributions, $w > 0$ weights the last term Γ which is a stepwise function:

$$\Gamma(x) = \begin{cases} 1, & x \geq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

The first four components of (7) align with those in (4), and following [6], the final term uses a stepwise function to promote successful V2I communications. The global reward, considering the interference impact from each PL's subchannel and power choices, is defined as the average interference:

$$\mathcal{R}_g^t = -\frac{1}{L}\sum_l \sum_k \mathcal{F}_5\left(I_{l,x}^t[k]\right), \tag{9}$$

where $\mathcal{F}_5$ maps the interference power to a suitable range, and the global reward is affected by PLs differently according to their communication mode: $I_{l,x}^t[k] = I_{l,\mathcal{R}}^t[k]$ for V2I, $I_{l,x}^t[k] = I_{l,v_l}^t[k]$ for V2V. This design encourages channel selections that minimise disruption to other PLs.

At each time step t, PL_l observes a state $s \in \mathcal{S}_l^t$ from the environment, and then takes action $a \in \mathcal{A}_l^t$ with a probability $\pi_l(a|s)$, where $\pi_l(a|s)$ is the conditional probability that $A_l^t = a$ if $S_l^t = s$, i.e., the policy of PL_l. A reward $R_l^t = r$ will be received from interacting with the environment using the selected action, and then the time is moved to $t+1$.

B. Policy and Value Function

Following previous discussions, the optimisation problem focuses on maximising the expected discounted return via the state-value function $\mathcal{V}$:

$$\begin{aligned}&\mathcal{V}_l^*(s) = \mathcal{V}_l^{\pi_l^*}(s) = \max_{\pi_l} \mathcal{V}_l^{\pi_l}(s) \\ &= \max_{\pi_l} \mathbb{E}_{l,\pi_l}\left[\sum_{\kappa=0}^{\infty} \gamma^\kappa R_l^{t+\kappa+1} \middle| s_l^t = s\right], \forall s \in \mathcal{S},\end{aligned} \tag{10}$$

where $\mathcal{V}_l^{\pi_l}(s)$ is the state-value function, $\mathcal{V}_l^{\pi_l^*}(s)$ is the state-value function under the optimal policy π_l^*, $\mathcal{V}_l^*(s)$ is the optimal state-value function, and $\mathbb{E}_{l,\pi_l}[\cdot]$ is the expected value of the discounted return $G_l^t = \sum_{\kappa=0}^{\infty} \gamma^\kappa R_l^{t+\kappa+1}$ given that PL_l follows policy π_l. Bellman optimality equation [47] is reviewed to help us solve the problem stated in (10), which shows that the state value under the optimal policy equals the expected return from the state under the best action:

$$\mathcal{V}_l^{\pi_l^*}(s) = \max_{a \in \mathcal{A}_l^t(s)} Q_l^{\pi_l^*}(s,a), \forall s \in \mathcal{S}, \tag{11}$$

where $Q_l^{\pi_l}(s,a) = \mathbb{E}_{l,\pi_l}\left[G_l^t \middle| s_l^t = s, a_l^t = a\right]$ is the action-value function and $Q_l^{\pi_l^*}(s,a)$ is the action-value function under the optimal policy. Similarly, the optimal action-value function $Q_l^*(s,a)$ is equivalent to $Q_l^{\pi_l^*}(s,a)$, and hence

$$\mathcal{V}_l^*(s) = \max_{a \in \mathcal{A}_l^t(s)} Q_l^*(s,a), \forall s \in \mathcal{S}. \tag{12}$$

In the MDP, PL_l continually updates its policy to optimise the state-value function as in (10) and seeks actions that maximise the action-value function—the Q-function in (12). Both strategies aim to achieve the optimal state-value function, effectively solving the optimisation problem in (4).

IV. MADRL Approach

The resource allocation problem defined in Section II-D is inherently a multi-agent problem. A key challenge, as noted in Section III-A, is the non-stationary environment: from the perspective of any single agent, the random movements of other vehicle platoons create a highly dynamic and unpredictable interference landscape. Standard single-agent DRL would fail to converge in such an environment. Therefore, we adopt a MADRL approach based on the centralised training with decentralised execution paradigm for the joint optimisation of the policy π in (10) and Q-function in (12) in the multi-agent environment. Similar to the work that has been done in [6], we apply the DE-MADDPG algorithm, which is a combination of the standard MADDPG and DDPG algorithms that can be found in [42]. This algorithm uses a centralised critic that has access to the states and actions of all agents, allowing it to form a stable understanding of the environment and the impact of joint actions, i.e., the mutual interference. This critic is then used to train the decentralised actors alongside their own critics, allowing the system to find a convergent, cooperative solution. We also combine the DE-MADDPG algorithm with the TDec algorithm proposed in [43] to apply the DE-MADDPG-TDec algorithm. In addition, we introduce TD3 [44] and its extension—multi-agent TD3 (MATD3) to overcome the overestimation problem in Q-functions to further enhance the performance.

A. Decomposed Multi-Agent Deep Deterministic Policy Gradient

Different from the conventional MADDPG (or MATD3) that trains multiple agents with only one critic (or two critics), the main idea behind DE-MADDPG is to introduce the standard DDPG (or TD3) for each local agent and combine with a centralised global critic. The objective becomes optimising the policy to maximise both the local and global critics. The combined policy gradient for agent l is

$$\begin{aligned}\overbrace{\nabla J(\theta_l)}^{\textit{DDPG: Local actor}} &= \overbrace{\mathbb{E}_{\boldsymbol{s},\boldsymbol{a}\sim\mathcal{B}}\left[\nabla_{\theta_l}\pi_l\left(a_l|s_l\right)\nabla_{a_l}Q_\psi^g(\boldsymbol{s},\boldsymbol{a})\right]}^{\textit{MADDPG: Global Critic}} \\ &+ \underbrace{\mathbb{E}_{s_l,a_l\sim\mathcal{B}}\left[\nabla_{\theta_l}\pi_l\left(a_l|s_l\right)\nabla_{a_l}Q_{\phi_l}^{\pi_l}\left(s_l,a_l\right)\right]}_{\textit{DDPG: Local Critic}},\end{aligned} \tag{13}$$

where θ_l parameterises the policy of agent l, $\boldsymbol{s} = (s_1, ..., s_L)$, $\boldsymbol{a} = (a_1, ..., a_L)$, $\mathcal{B}$ is the experience replay buffer, $a_l = \pi_l(s_l)$ is the action of agent l under its policy, ψ and ϕ_l parameterises

the Q-functions—Q^g_ψ for global critic and $Q^{\pi_l}_{\phi_l}$ for local critic. The global critic and the local critic are updated by minimising the following loss functions:

$$\mathcal{L}(\psi) = \mathbb{E}_{\boldsymbol{s},\boldsymbol{a},\boldsymbol{r},\boldsymbol{s}'}\left[\left(Q^g_\psi(\boldsymbol{s},\boldsymbol{a}) - y_g\right)^2\right], \tag{14}$$

$$\mathcal{L}(\phi_l) = \mathbb{E}_{s_l,a_l,r_l,s'_l}\left[\left(Q^{\pi_l}_{\phi_l}(s_l,a_l) - y_l\right)^2\right], \tag{15}$$

where $\boldsymbol{r} = (r_1,...,r_L)$, $\boldsymbol{s}' = (s'_1,...,s'_L)$ are the next states, $s'_l \in \boldsymbol{s}'$. The global and local target values are written as

$$y_g = r_g + \gamma Q^g_{\psi'}(\boldsymbol{s}',\boldsymbol{a}')\Big|_{a'_l=\pi'_l(s'_l)}, \tag{16}$$

$$y_l = r_l + \gamma Q^{\pi_l}_{\phi'_l}(s'_l,a'_l)\Big|_{a'_l=\pi'_l(s'_l)}, \tag{17}$$

where $\boldsymbol{a}' = (a'_1,...,a'_L)$ is the next set of actions, $a'_l \in \boldsymbol{a}'$, $Q^g_{\psi'}$ and $Q^{\pi_l}_{\phi'_l}$ represent the target global and local critics, and π'_l is the target policy for agent l.

B. Twin Delayed Deep Deterministic Policy Gradient

In order to tackle the overestimation problem in DDPG, the idea of TD3 is introduced to improve the parameter update in DE-MADDPG. The policy gradient with TD3 for agent l is

$$\begin{aligned}\nabla J(\theta_l) = &\mathbb{E}_{\boldsymbol{s},\boldsymbol{a}\sim\mathcal{B}}\Big[\nabla_{\theta_l}\pi_l(a_l|s_l)\nabla_{a_l}Q^g_{\psi_1}(\boldsymbol{s},\boldsymbol{a})\Big]\\ &+\mathbb{E}_{s_l,a_l\sim\mathcal{B}}\Big[\nabla_{\theta_l}\pi_l(a_l|s_l)\nabla_{a_l}Q^{\pi_l}_{\phi_l}(s_l,a_l)\Big].\end{aligned} \tag{18}$$

The twin global critics $Q^g_{\psi_1}$ and $Q^g_{\psi_2}$ are updated by minimising the loss function:

$$\mathcal{L}(\psi_i) = \mathbb{E}_{\boldsymbol{s},\boldsymbol{a},\boldsymbol{r},\boldsymbol{s}'}\left[\left(Q^g_{\psi_i}(\boldsymbol{s},\boldsymbol{a}) - y_g\right)^2\right], i = 1,2, \tag{19}$$

$$y_g = r_g + \gamma \min_i Q^g_{\psi'_i}(\boldsymbol{s}',\boldsymbol{a}')\Big|_{a'_l=\pi'_l(s'_l)}. \tag{20}$$

By delaying the update of the local network by d loops, the final DE-MADDPG (TD3) is described in Algorithm 1.

C. Task Decomposition Algorithm

According to *Theorem 1* in [43], if the reward function in the MDP can be decomposed into N sub-functions (tasks), i.e., $\mathcal{R}(s,a) = \sum_{n=1}^N \mathcal{R}_n(s,a)$, the state and action value functions can also be decomposed, i.e., $\mathcal{V}^\pi(s) = \sum_{n=1}^N \mathcal{V}^\pi_n(s)$, $Q^\pi(s,a) = \sum_{n=1}^N Q^\pi_n(s,a)$. Therefore, we adopt the idea of TDec algorithm for our local reward function in (7), and the decomposed functions are written as

$$\begin{aligned}\mathcal{R}^t_{l,1} = &-\mathcal{F}_1\left(A^t_l\right) + w\Gamma\left(\mathcal{C}^t_{l,\mathcal{R}}[k] - \mathcal{C}^{\min}_{l,\mathcal{R}}\right)\\ &-\left(1-\lambda^t_l\right)\left[\mathcal{F}_2\left(p^t_l[k]\right) + \mathcal{F}_3\left(E^t_l[k]\right)\right],\end{aligned} \tag{21}$$

$$\begin{aligned}\mathcal{R}^t_{l,2} = &-\mathcal{F}_4\left(D'_l/D_l\right)\\ &-\lambda^t_l\left[\mathcal{F}_2\left(p^t_l[k]\right) + \mathcal{F}_3\left(E^t_l[k]\right)\right],\end{aligned} \tag{22}$$

where $\mathcal{R}^t_{l,1}$ is the task 1 for V2I mode communication, $\mathcal{R}^t_{l,2}$ is the task 2 for V2V mode communication, and $\mathcal{R}^t_l = \mathcal{R}^t_{l,1} + \mathcal{R}^t_{l,2}$. Hence, our policy gradient (18), local loss function (15), and local target function (17) can be written as

$$\begin{aligned}\nabla J(\theta_l) = &\mathbb{E}_{\boldsymbol{s},\boldsymbol{a}\sim\mathcal{B}}\Big[\nabla_{\theta_l}\pi_l(a_l|s_l)\nabla_{a_l}Q^g_{\psi_1}(\boldsymbol{s},\boldsymbol{a})\Big]\\ &+\sum_n \mathbb{E}_{s_l,a_l\sim\mathcal{B}}\Big[\nabla_{\theta_l}\pi_l(a_l|s_l)\nabla_{a_l}Q^{\pi_l}_{\phi_{l,n}}(s_l,a_l)\Big],\end{aligned} \tag{23}$$

Algorithm 1: DE-MADDPG (TD3)

Initialise intersection environment & replay buffer $\mathcal{B}$.
Initialise global critic networks: $\{Q^g_{\psi_i}, Q^g_{\psi'_i}\}, i = 1,2$.
Initialise actor & critic networks for each agent: $\{\pi_l, \pi'_l, Q^{\pi_l}_{\phi_l}, Q^{\pi_l}_{\phi'_l}\}, l = 1,2,...,L$.
for episode $= 1$ to *loop* **do**
  Update platoon location & channel information.
  Reset time budget & CAM size: $\{T', D'_l\} = \{T, D_l\}$.
  for $t = 1$ to T **do**
    for agent 1 to L **do**
      Observe state s^t_l, select action $a^t_l = \pi_l(s^t_l)$, receive local & global rewards: $\{r^t_l, r^t_g\}$.
    Update channel fast fading & interference.
    Each agent l observes new state s^{t+1}_l.
    Store $\left(\boldsymbol{s}^t, \boldsymbol{a}^t, \boldsymbol{r}^t, r^t_g, \boldsymbol{s}^{t+1}\right)$ into $\mathcal{B}$.
    Sample mini-batch of M transitions $\left(\boldsymbol{s}^m, \boldsymbol{a}^m, \boldsymbol{r}^m, r^m_g, \boldsymbol{s}^m_{new}\right)|^M_{m=1}$ from $\mathcal{B}$.
    Update global critics: minimising $\mathcal{L}(\psi_i)$ (19) by one-step gradient descent.
    Target soft update: $\psi'_i \leftarrow \tau\psi_i + (1-\tau)\psi'_i$.
    if t mod d **then**
      for agent 1 to L **do**
        Update local critic: minimising $\mathcal{L}(\phi_l)$ (15) by one-step gradient descent.
        Update local actor: maximising $J(\theta_l)$ as in (18) by one-step gradient ascent.
        Target soft update: $\phi'_l \leftarrow \tau\phi_l + (1-\tau)\phi'_l$
        $\theta'_l \leftarrow \tau\theta_l + (1-\tau)\theta'_l$

$$\mathcal{L}(\phi_{l,n}) = \mathbb{E}_{s_l,a_l,r_l,s'_l}\left[\left(Q^{\pi_l}_{\phi_{l,n}}(s_l,a_l) - y_{l,n}\right)^2\right], \tag{24}$$

$$y_{l,n} = r_{l,n} + \gamma Q^{\pi_l}_{\phi'_{l,n}}(s'_l,a'_l)\Big|_{a'_l=\pi'_l(s'_l)}, \tag{25}$$

where $n = 1,2$, and the global critics are still updated as in (19). Based on this design, the number of local critics is increased from 1 to $N = 2$. The two local critics for two tasks with two sub-Q-functions work jointly to move towards the best estimation of the overall Q-function,

$$Q^{\pi_l}_{\phi_l}(s_l,a_l) = \sum_n Q^{\pi_l}_{\phi_{l,n}}(s_l,a_l), \tag{26}$$

and hence the best update for the policy π_l of the local actor. The structure of DE-MADDPG-TDec is similar to Algorithm 1, the only difference is that for each agent, an extra **for** loop is added for the N tasks, as described in Algorithm 2.

V. Complexity Analysis

This section provides the complexity analysis for the proposed algorithms and three more algorithms that are used as baselines. All the algorithms evaluated in this study include:

- *DE-MADDPG-TDec (TD3)*: As specified in Sec. IV-C.
- *DE-MADDPG (TD3)*: As specified in Sec. IV-B.
- *Decentralised MADDPG (Dec-MADDPG)*: Decentralised learning, all the agents function independently in the system, each only has access to its own observation and takes decision based on this limited information.
- *Baseline-MADDPG*: Centralized critic training with decentralized execution, standard MADDPG algorithm [48].
- *Baseline-DDPG*: Centralised actor-critic network, standard DDPG algorithm [49].
- *E-X*: Corresponding energy-concerned versions. The difference between the energy-concerned and non-energy-concerned algorithms is that the action space $\mathcal{A}^t_l$ for

Algorithm 2: DE-MADDPG-TDec (TD3)

Initialise intersection environment & replay buffer $\mathcal{B}$.
Initialise global critic networks: $\{Q^g_{\psi_i}, Q^g_{\psi'_i}\}, i = 1, 2$.
Initialise actor & critic networks for each agent: $\{\pi_l, \pi'_l, Q^{\pi_l}_{\phi_l}, Q^{\pi_l}_{\phi'_l}\}, l = 1, 2, ..., L$.
for episode $= 1$ to *loop* **do**
  Update platoon location & channel information.
  Reset time budget & CAM size: $\{T', D'_l\} = \{T, D_l\}$.
  for $t = 1$ to T **do**
    for agent 1 to L **do**
      Observe state s^t_l, select action $a^t_l = \pi_l(s^t_l)$, receive local & global rewards: $\{r^t_l, r^t_g\}$.
    Update channel fast fading & interference.
    Each agent l observes new state s^{t+1}_l.
    Store $\left(\boldsymbol{s}^t, \boldsymbol{a}^t, \boldsymbol{r}^t, r^t_g, \boldsymbol{s}^{t+1}\right)$ into $\mathcal{B}$.
    Sample mini-batch of M transitions $\left(\boldsymbol{s}^m, \boldsymbol{a}^m, \boldsymbol{r}^m, r^m_g, \boldsymbol{s}^m_{new}\right)\big|^M_{m=1}$ from $\mathcal{B}$.
    Update global critics: minimising $\mathcal{L}(\psi_i)$ (19) by one-step gradient descent.
    Target soft update: $\psi'_i \leftarrow \tau\psi_i + (1-\tau)\psi'_i$.
    if t mod d **then**
      for agent 1 to L **do**
        for task 1 to N **do**
          Update local critic: minimising $\mathcal{L}(\phi_{l,n})$ (24) by one-step gradient descent.
        Update local actor: maximising $J(\theta_l)$ as in (23) by one-step gradient ascent.
        Target soft update:
        for task 1 to N **do**
          $\phi'_{l,n} \leftarrow \tau\phi_{l,n} + (1-\tau)\phi'_{l,n}$
        $\theta'_l \leftarrow \tau\theta_l + (1-\tau)\theta'_l$

the energy-concerned algorithms contains four items: $\left[\delta^t_{l,k}, \lambda^t_l, p^t_l[k], E^t_l[k]\right]$. On the contrary, the action space for the non-energy-concerned algorithms contains three items only: $\left[\delta^t_{l,k}, \lambda^t_l, p^t_l[k]\right]$.

Before presenting the formal complexity analysis, we provide an intuitive overview. The complexity of these algorithms is driven by two main factors: The number of agents (L) and the number of neural networks each algorithm must train and use. We provide a direct comparison of these factors in Table III. In simple terms, the algorithms can be ranked by complexity as follows:

- Baseline-DDPG is the least complex, as it treats the entire system as a single agent with one actor and one critic.
- Dec-MADDPG and Baseline-MADDPG are more complex, as they scale with the number of agents (L), requiring an actor-critic pair for each agent.
- DE-MADDPG (TD3) adds another layer of complexity by including two centralised global critics on top of all the local agent networks.
- DE-MADDPG-TDec (TD3) is the most complex, as it further increases the number of local critics to N for each agent to handle the TDec algorithm.

The following formal analysis provides the detailed mathematical basis for this summary.

The computational complexity, which directly affects the execution time of the algorithms, is firstly analysed. Using $\mathcal{J}^a$ and $\mathcal{J}^c$ to denote the number of layers in the actor and the critic networks, and using U^a_j and U^c_j to denote the number of neurons in the j^{th} layer of the corresponding actor and critic networks, the general computational complexity in an actor network (C^a) and a critic network (C^c) can be written as:

$$C^a = \mathrm{O}\Big(\sum_{j=2}^{\mathcal{J}^a-1}\left(U^a_{j-1}U^a_j + U^a_j U^a_{j+1}\right)\Big), \tag{27}$$

$$C^c = \mathrm{O}\Big(\sum_{j=2}^{\mathcal{J}^c-1}\left(U^c_{j-1}U^c_j + U^c_j U^c_{j+1}\right)\Big). \tag{28}$$

Hence the computational complexity of the included algorithms is expressed as:

- *DE-MADDPG-TDec (TD3)*:
$$\mathrm{O}\left(ep \cdot T/\Delta t \cdot M\left(C^{c,g} + L\left(C^{a,l} + N \cdot C^{c,l}\right)\right)\right),$$
- *DE-MADDPG (TD3)*:
$$\mathrm{O}\left(ep \cdot T/\Delta t \cdot M\left(C^{c,g} + L\left(C^{a,l} + C^{c,l}\right)\right)\right),$$
- *Decentralised MADDPG (Dec-MADDPG)*:
$$\mathrm{O}\left(ep \cdot T/\Delta t \cdot M \cdot L\left(C^{a,l} + C^{c,l}\right)\right),$$
- *Baseline-MADDPG*:
$$\mathrm{O}\left(ep \cdot T/\Delta t \cdot M \cdot L\left(C^{a,l} + C^{c,l}\right)\right),$$
- *Baseline-DDPG*:
$$\mathrm{O}\left(ep \cdot T/\Delta t \cdot M\left(C^{a,l} + C^{c,l}\right)\right),$$

where ep is the number of episodes, $C^{c,g}$, $C^{c,l}$, and $C^{a,l}$ are the general computational complexity in a global critic network, a local critic network, and a local actor network. The Baseline-DDPG algorithm has the smallest computational complexity among all the algorithms. Instead of providing actor and critic networks to each agent, it optimises all agents together with only one actor network and one critic network. The Baseline-MADDPG algorithm and Dec-MADDPG algorithm have the same complexity since their only difference is whether the agents can get access to the observation from other agents. The extra complexity of the DE-MADDPG (TD3) algorithm comes from the global critic networks, and in addition to this, the complexity of the DE-MADDPG-TDec (TD3) algorithm is increased by N tasks from the TDec method.

TABLE III: Number of Neural Networks and Trainable Parameters in Algorithms.

Algorithm	Neural Networks	Trainable Parameters
DE-MADDPG-TDec (TD3)	$2\left(\mathbf{2}^g_c + L\left(\mathbf{1}^l_a + \mathbf{N}^l_c\right)\right)$	$\mathrm{O}\left((L+1)\chi\right)$
DE-MADDPG (TD3)	$2\left(\mathbf{2}^g_c + L\left(\mathbf{1}^l_a + \mathbf{1}^l_c\right)\right)$	$\mathrm{O}\left((L+1)\chi\right)$
Dec-MADDPG	$2L\left(\mathbf{1}^l_a + \mathbf{1}^l_c\right)$	$\mathrm{O}\left(L\chi\right)$
Baseline-MADDPG	$2L\left(\mathbf{1}^l_a + \mathbf{1}^l_c\right)$	$\mathrm{O}\left(L^2\chi\right)$
Baseline-DDPG	$2\left(\mathbf{1}^l_a + \mathbf{1}^l_c\right)$	$\mathrm{O}\left(L\chi\right)$

In addition, the number of neural networks and trainable parameters of each algorithm is provided in Table III. For the number of neural networks used in each algorithm, $\mathbf{1}$ denotes one network. a and c represent whether the network is an actor or a critic network, while g and l represent whether the network is a global or a local network (e.g., $\mathbf{1}^l_a$ stands for one local actor network). $\mathbf{N}^l_c$ for the DE-MADDPG-TDec (TD3) algorithm shows that the N tasks require N critic networks. The two global critic networks ($\mathbf{2}^g_c$) used in the DE-MADDPG-TDec (TD3) and DE-MADDPG (TD3) algorithms are the twin critics from the TD3 technique. Lastly, 2 is multiplied by all the algorithms for the target actor and critic networks.

For the number of trainable parameters contained in each algorithm, $\chi = \xi + \alpha$, ξ and α denote the size of the state space and the action space (again, the size of the action space is different for the energy-concerned and non-energy-concerned algorithms), respectively. The Dec-MADDPG algorithm uses the fewest trainable parameters (the same as that of Baseline-DDPG), since the agents can only access their own observations. In contrast, the agents in the MADDPG algorithm use the observations and actions from all the agents to optimise themselves, hence the L is squared (L^2) for the Baseline-MADDPG algorithm. Furthermore, the global network contributes $\mathrm{O}(L\chi)$ to the DE-MADDPG-TDec (TD3) and DE-MADDPG (TD3) algorithms, while the local network contributes $\mathrm{O}(\chi)$.

Finally, we note the distinction between offline training complexity and online inference latency. The primary focus of this paper, in line with much of the DRL-for-resource-allocation literature, is on the optimality of the resource allocation policy (i.e., successfully minimising AoI and power-energy consumption while maximising CDP). However, the feasibility of its real-time deployment is also critical. We confirm this feasibility: the online decision-making at each time step t only requires a single forward pass through the local actor network (i.e., $a_l^t = \pi_l(s_l^t)$). Given the modest size of our actor network (as shown in Section VI), this inference operation is computationally lightweight (on the order of microseconds) and can be executed well within the 1 ms time slot duration.

VI. Numerical Results

In this section, we present the simulation results demonstrating the performance of the proposed algorithms. We use a single-cell urban C-V2X network operating at 2 GHz with 3 RBs, adhering to the urban specifications outlined in 3GPP TR 36.885 [40]. We use Python with PyTorch to implement our MADRL framework. The structure of our deep neural network consists of two hidden layers for the local actor (1024, 512 neurons) and critic (512, 256 neurons), and three hidden layers for the global critic (1024, 512, 256 neurons). The activation function and optimiser are chosen as the rectified linear unit (ReLU) [50] and Adam [51]. The learning rates of the actor and critic networks are set as 0.0001 and 0.001, while the target soft update parameter τ and the discount factor γ are set as 0.005 and 0.99. The primary simulation parameters and training parameters are detailed in Table IV. The algorithms evaluated in this study have been introduced in Sec. V.

Fig. 2(a) shows the reward convergence performance of the five energy-concerned algorithms (*E-X*), with the number of platoons and the number of vehicles in each platoon set to $L = 5$ and $V_l = 4$. The result of each algorithm is averaged over 10 independent simulation runs. The solid lines represent the mean reward values, while the shaded regions around these lines depict the fluctuation (the range between minimum and maximum values) observed across these multiple simulation runs. It is clear that the proposed algorithms, E-DE-MADDPG-TDec (TD3) and E-DE-MADDPG (TD3), outperform the other baseline algorithms, and the convergence speed of E-DE-MADDPG-TDec (TD3) algorithm with the idea of TDec is faster than that of the E-DE-MADDPG (TD3) algorithm. In addition to the reward convergence, Fig. 2(b) presents the AoI convergence. Similarly, the solid lines indicate mean AoI values, while the shaded areas represent the fluctuation. Consistent with the reward performance, the two proposed algorithms demonstrate better AoI performance with lower AoI level achieved more rapidly in early episodes and with the lowest AoI level reached in later episodes.

TABLE IV: Simulation parameters.

Environmental Parameters	**Symbols**	**Values**
Number of platoons	L	$\{4, 5, 6, 7\}$
Vehicles in each platoon	V_l	4
V2V gap	-	25 m
PL maximum power	p^{max}	30 dBm
Noise power	σ^2	-114 dBm
Carrier frequency	-	2 GHz
Number of RBs	K	3
RB bandwidth	W	180 kHz
CAM size	D_l	4 KB
V2V time limitation	T	100 ms [4], [40]
Fast fading update period	Δt	1 ms [40]
Fast fading	-	Rayleigh fading
Large-scale fading update	-	100 ms [40]
V2V path loss model	-	B1 – Urban micro-cell [52]
V2I path loss model	-	128.1 + 37.6$\log_{10}(d)$ [6]
V2V Shadowing model	-	Log-normal, $\sigma = 3$ dB
V2I Shadowing model	-	Log-normal, $\sigma = 8$ dB
V2V Decorrelation distance	-	10 m
V2I Decorrelation distance	-	50 m
Number of episodes	ep	400
Training Parameters	**Symbols**	**Values**
Reward discount factor	γ	0.99
Batch size	M	64
Actor learning rate	-	0.0001
Critic learning rate	-	0.001
Target soft update	τ	0.005
Optimiser	-	Adam

Fig. 2(c) compares the reward convergence of the energy-concerned E-DE-MADDPG-TDec (TD3) algorithm with its non-energy-concerned counterpart, DE-MADDPG-TDec (TD3) proposed in [6]. Similarly, Fig. 2(d) compares the E-DE-MADDPG (TD3) algorithm and the DE-MADDPG (TD3) algorithm. It can be observed that the energy-concerned versions exhibit similar convergence speeds to the non-energy-concerned versions. However, due to the difference in the reward function design specially developed for energy-aware objectives versus non-energy-aware objectives, the direct comparison of their absolute reward levels is not fair. Therefore, for a fair and meaningful comparison of their overall effectiveness, the detailed comparisons of the AoI level, CDP, and the energy consumption are shown in Fig. 3.

Fig. 3 presents a comparative performance analysis of the five energy-concerned algorithms with their non-energy-concerned counterparts. The result of each algorithm is also averaged over 10 independent simulation runs, and only the result in the last 100 episodes is considered. The number of platoons in the system varies from $L = 4$ to $L = 7$, with a fixed number of vehicles in each platoon $V_l = 4$. The results of the energy-concerned algorithms are depicted with solid lines, while their counterparts are shown with dashed lines. Three distinct metrics are evaluated in Fig. 3(a)-(c), which are the

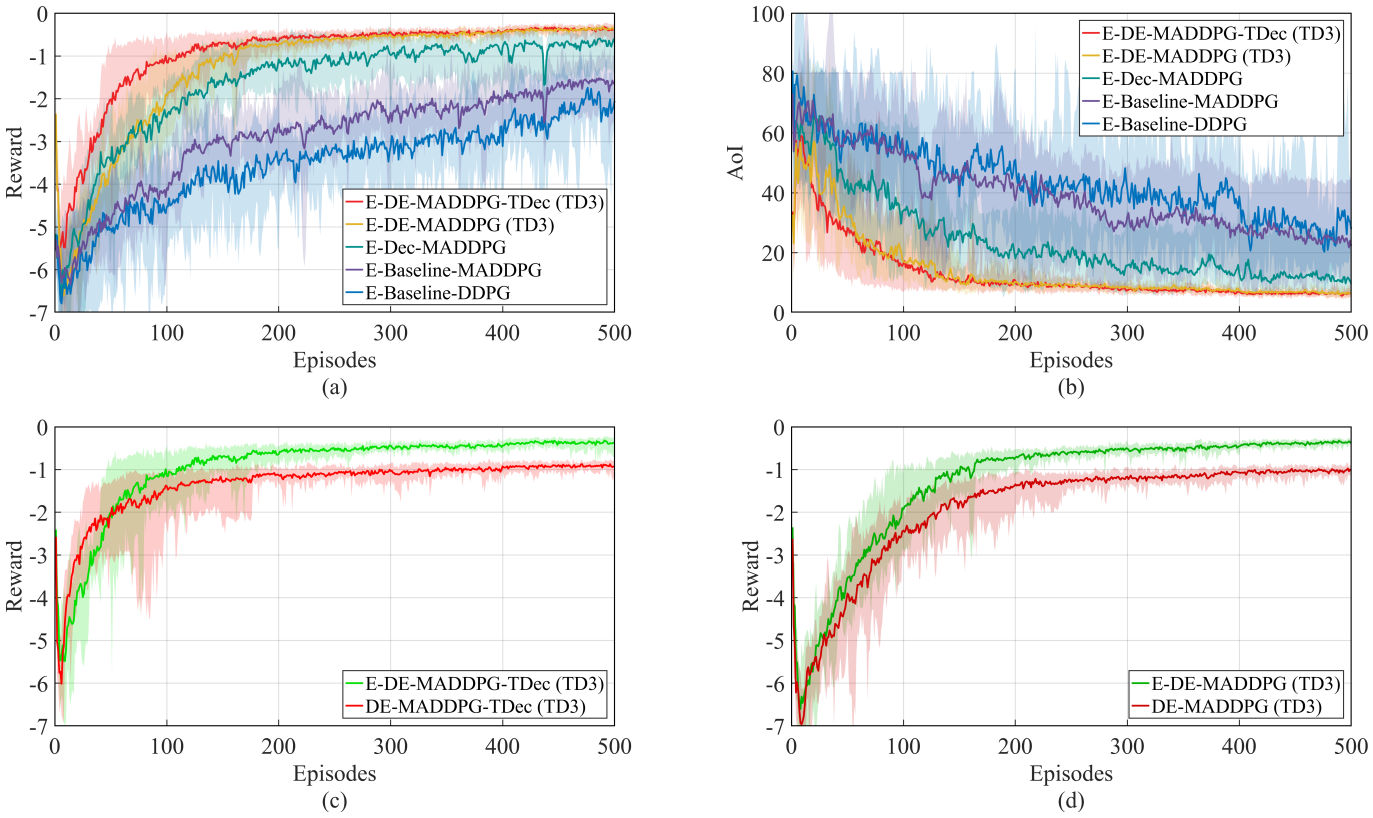

Fig. 2: Performance comparison ($L = 5$, $V_l = 4$, results averaged over 10 independent simulation runs). (a) Reward convergence of the energy-concerned algorithms. (b) AoI convergence of the energy-concerned algorithms. (c) Comparison of reward convergence between DE-MADDPG-TDec (TD3) and its energy-concerned counterpart. (d) Comparison of reward convergence between DE-MADDPG (TD3) and its energy-concerned counterpart.

AoI level, CDP, and the energy consumption, respectively.

Fig. 3(a) illustrates the average AoI in milliseconds (ms) observed among all platoons. The proposed DE-MADDPG-TDec (TD3) and DE-MADDPG (TD3) algorithms exhibit much lower AoI levels compared with the baseline algorithms among all platoon conditions. Our energy-concerned algorithms show a slight increase in AoI compared with their non-energy-concerned counterparts—between 3.21% and 6.82%—the overall performance remains comparable. A similar phenomenon can also be observed in Fig. 3(b), where the average CDP is shown as a percentage (%). The proposed algorithms demonstrate much higher CDP compared with the baseline algorithms, especially when the number of platoons is high ($L = 6$, 7). In terms of CDP, our energy-concerned algorithms show slightly better performance, with a maximum increase of 0.51%.

In Fig. 3(c), the average energy consumption per platoon is shown, measured in Joules (J). This value is derived by summing the energy consumed by all platoons in the last 100 episodes and then dividing this sum by the number of platoons. As depicted in the figure, our energy-focused algorithms significantly reduce the energy use for both the proposed algorithms and the baseline algorithms, with a 58.19% to 95.11% decrease compared to their non-energy-concerned counterparts (this decrease is 72.57% to 91.10% for the proposed algorithms). This notable decrease in energy consumption demonstrates the remarkable performance of our energy-focused algorithms.

In order to further reveal the advantages of our energy-focused algorithms, we provide an evaluation based on a specially designed *Energy Efficiency* metric that integrates performance in terms of AoI and CDP with the energy consumed. The performance component is derived from normalised scores for AoI (AoI_{score}) and CDP (CDP_{score}), each ranging from 0 (worst) to 1 (best). These two scores are defined as follows,

$$AoI_{score} = \frac{AoI_{max} - AoI}{AoI_{max} - AoI_{min}}, \tag{29}$$

$$CDP_{score} = \frac{CDP - CDP_{min}}{CDP_{max} - CDP_{min}}, \tag{30}$$

where AoI_{max}, AoI_{min}, CDP_{max}, and CDP_{min} are the maximum and minimum values of the AoI and CDP among all algorithms and all L values. Both the AoI_{score} and the CDP_{score} range $[0, 1]$, AoI_{score} reaches 1 when the AoI is at the minimum level, and reaches 0 when the AoI is at the maximum level. On the contrary, CDP_{score} reaches 1 when the CDP is at the maximum level, and reaches 0 when the CDP is at the minimum level. These two scores are combined with equal weighting as a measurement of the overall performance of the wireless communication:

$$Performance = 0.5 \cdot AoI_{score} + 0.5 \cdot CDP_{score}, \tag{31}$$

Another critical aspect of our *Energy Efficiency* metric is the formulation of the energy score (E_{score}). Instead of using the

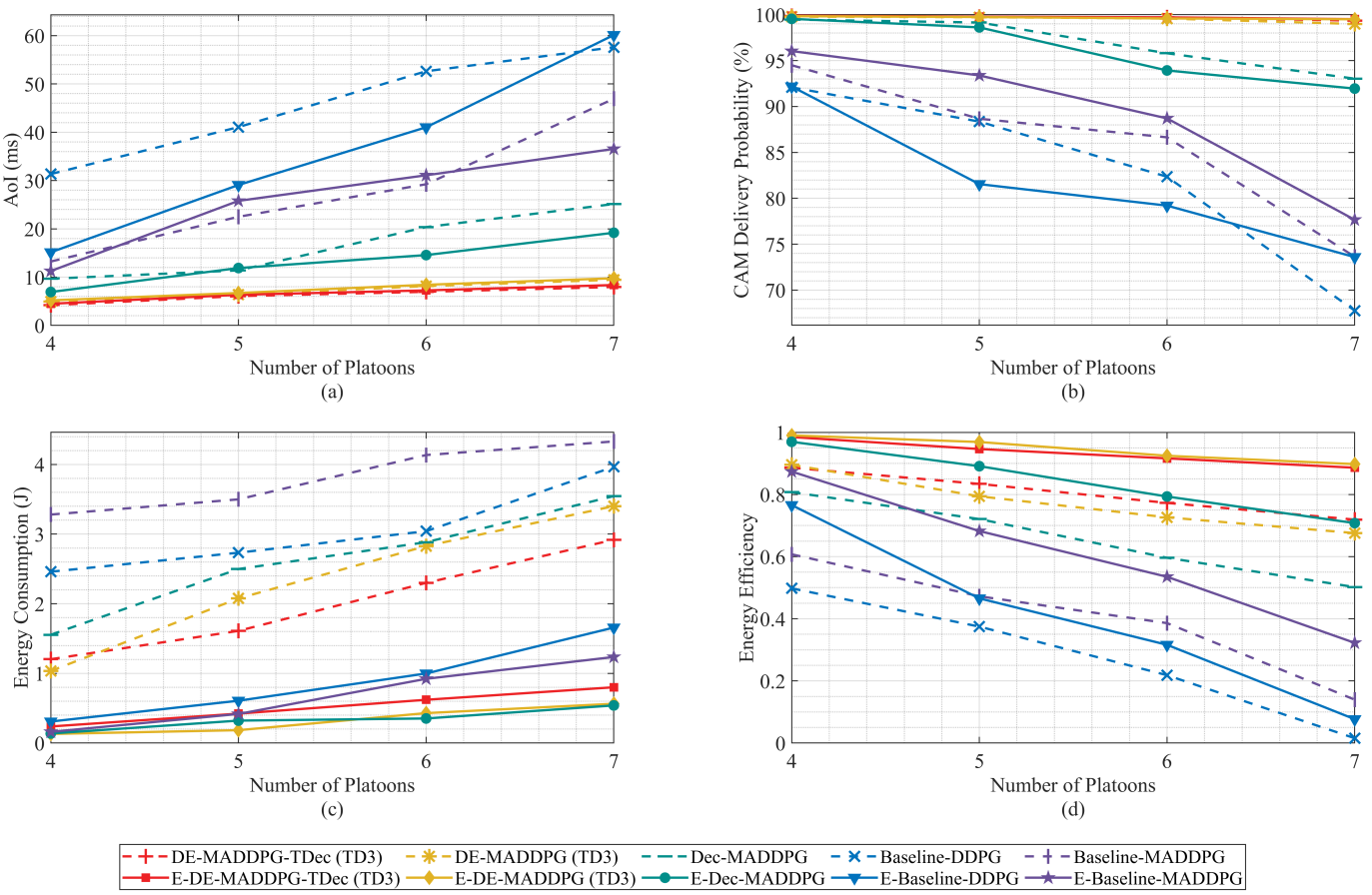

Fig. 3: Performance comparison with L varies from 4 to 7 and fixed $V_l = 4$. Results are from the last 100 episodes and are averaged over 10 independent simulation runs. (a) Average AoI in ms. (b) Average CDP in %. (c) Average energy consumption per platoon in J. (d) Energy efficiency, calculated using the metric defined in (33).

reciprocal of the energy consumption, which can excessively enlarge the energy efficiency for very low energy values or unnecessarily penalise higher energy consumption, we define the E_{score} as:

$$E_{score} = \frac{E_{max} + E_{min} + \omega}{E_{max} + E + \omega}, \tag{32}$$

where E_{max} and E_{min} are the maximum and minimum values of the consumed energy among all algorithms and all L values, and ω is a weight that adjusts the range of E_{score}. Different from AoI_{score} and CDP_{score} which range $[0, 1]$, ω is chosen to bound E_{score} within the range $[0.667, 1]$. This ensures the energy component scales the performance score moderately, preventing extreme energy efficiency values caused by large values of the consumed energy. An algorithm consuming E_{max} will be penalised while its E_{score} does not become zero, which allows its *Performance* metric to still contribute to the overall energy efficiency value. The final *Energy Efficiency* metric is then calculated as

$$\textit{Energy Efficiency} = \textit{Performance} \cdot E_{score}. \tag{33}$$

This approach avoids a scenario where an algorithm with excellent AoI and CDP but high energy use is overly disadvantaged (e.g., DE-MADDPG-TDec (TD3) and DE-MADDPG (TD3) algorithms), or an algorithm with mediocre AoI/CDP but extremely low energy use dominates the *Energy Efficiency* ranking. It tempers the impact of energy consumption to allow for a more balanced comparison. The corresponding results are shown in Fig. 3(d).

Fig. 3(d) clearly demonstrates that our energy-concerned algorithms, especially the two proposed algorithms, are much more economical than the non-energy-concerned counterparts. Those algorithms can make much more effective use of energy to reach similar low levels of AoI and high levels of CDP. Jointly observe all the results in Fig. 3, it can be found that among all the energy-concerned algorithms, E-DE-MADDPG-TDec (TD3) algorithm always reaches the best communication performance, while E-DE-MADDPG (TD3) algorithm consumes the least energy. Both proposed algorithms demonstrate superior energy efficiency in achieving better performance of wireless communication.

In addition, our specially designed *Energy Efficiency* metric demonstrates the necessity of the bounded E_{score}. For example, while the E-Baseline-MADDPG and E-Baseline-DDPG algorithms consume substantially less energy than all non-energy-concerned algorithms, their *Energy Efficiency* is consistently lower than that of DE-MADDPG-TDec (TD3) and DE-MADDPG (TD3). Similarly, despite the energy consumption of the E-Dec-MADDPG algorithm being even lower than that of the E-DE-MADDPG-TDec (TD3) and E-DE-MADDPG

(TD3) algorithms when $L = 6$ and 7, E-Dec-MADDPG still achieves a consistently lower *Energy Efficiency*. This phenomenon, where significantly lower energy consumption does not translate to superior *Energy Efficiency* for some of the E-Baseline algorithms, is caused by their poor communication performance. This demonstrates our *Energy Efficiency* metric provides a comprehensive and balanced assessment which reflects how effectively an algorithm utilises energy to achieve desired levels of information freshness and CAM delivery reliability, while avoiding an overemphasis on energy consumption alone.

It is also important to evaluate these results relative to the theoretical optimum. The study in [6] introduced a non-energy-concerned version of this algorithm and validated its near-optimal performance using an exhaustive search (refer to Fig. 7 in [6]). Following this, we benchmark our energy-focused algorithms against the global optimum derived from an exhaustive search to substantiate the efficacy of the proposed methods.

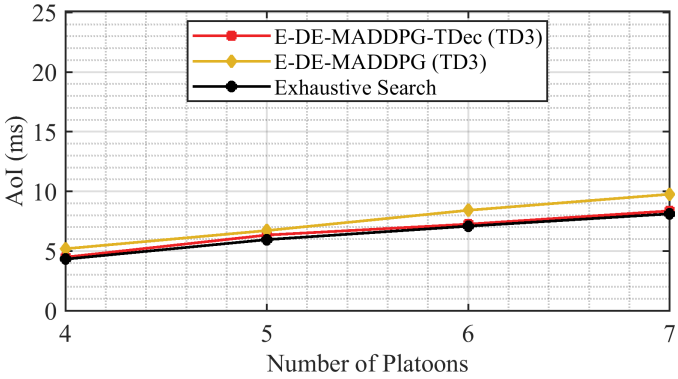

Fig. 4: Average AoI in ms versus the number of platoons.

As shown in Fig. 4, a comparison of the two proposed algorithms with the exhaustive search in terms of AoI performance is presented. The alignment of our proposed methods with the exhaustive search, particularly the DE-MADDPG-TDec (TD3) algorithm, indicates that our energy-focused algorithms achieve AoI performance approximating the optimum level.

VII. Conclusion

This paper has investigated the critical issue of resource allocation for platoon-based C-V2X networks operating at an urban intersection, a key scenario for future ITS. The study focuses on developing an optimal scheme using MADRL algorithms to manage the complex interplay of communication needs and resource constraints. Two MADRL algorithms, DE-MADDPG and DE-MADDPG-TDec with TD3 technique, have been proposed to address the joint optimisation of the AoI, CDP, and power-energy consumption. These algorithms are designed to empower multiple vehicle platoons to make intelligent, decentralised decisions regarding channel assignment and power control while causing the least disruption to other platoons, which is essential in this dynamic and cooperative environment. Numerical results presented have clearly demonstrated a remarkable reduction in energy consumption of vehicle platoons when compared with existing research efforts. This substantial improvement in energy consumption directly contributes to the objective of greener communication in vehicular environments. In addition, the considerably low energy consumption is achieved whilst maintaining comparable levels of algorithm convergence speed, low AoI, and high CDP. This indicates the proposed scheme does not sacrifice communication reliability or the timeliness of critical safety information for the sake of energy conservation. The energy efficiency metric presented further confirms the superior energy-saving capabilities of our algorithms.

While the simulation results are promising, the deployment of these MADRL algorithms in a real-world environment presents several practical challenges. First, our model assumes the availability of ideal state information, including perfect channel gains and interference levels. In a practical system, this information must be estimated, introducing potential errors, latency, and feedback overhead. Second, the current framework is trained for a fixed number of platoons. The highly dynamic nature of a real intersection, with agents constantly joining and leaving, poses a significant non-stationarity challenge that the current model is not designed to handle. Third, the policy is trained on a specific simulation model, and its robustness in the 'sim-to-real' gap must be validated, since the real-world channel and traffic dynamics may differ from the simulation. Based on these considerations, our future work will focus on more practical cases and more advanced techniques, such as a dynamic number of platoons at the intersection, or under the condition that vehicles leave their current platoon and/or join a new platoon. Extending the wireless communication system to NOMA and MIMO scenarios is also considered, which is to better align both the environmental setting and the communication model with real-world conditions.

References

[1] Y. Sun, Y. Hu, H. Zhang, H. Chen, and F.-Y. Wang, "A parallel emission regulatory framework for intelligent transportation systems and smart cities," *IEEE Trans. Intell. Veh.*, vol. 8, no. 2, pp. 1017–1020, Feb. 2023.

[2] M. Noor-A-Rahim. *et al.*, "6G for Vehicle-to-Everything (V2X) communications: Enabling technologies, challenges, and opportunities," *Proc. IEEE*, vol. 110, no. 6, pp. 712–734, Jun. 2022.

[3] H. V. Vu *et al.*, "Multi-agent reinforcement learning for channel assignment and power allocation in platoon-based C-V2X systems," in *Proc. 2022 IEEE 95th Veh. Technol. Conf. (VTC2022-Spring)*, Helsinki, Finland, Jun. 2022, pp. 1–5.

[4] *Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Part 2: Specification of Cooperative Awareness Basic Service*, ETSI Std. EN 302 637-2, Apr. 2019.

[5] S. Kaul, M. Gruteser, V. Rai, and J. Kenney, "Minimizing age of information in vehicular networks," in *Proc. 2011 8th Annu. IEEE Commun. Soc. Conf. Sens. Mesh Ad Hoc Commun. Netw.*, Salt Lake City, UT, USA, Jun. 2011, pp. 350–358.

[6] M. Parvini, M. R. Javan, N. Mokari, B. Abbasi, and E. A. Jorswieck, "AoI-aware resource allocation for platoon-based C-V2X networks via multi-agent multi-task reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 72, no. 8, pp. 9880–9896, Aug. 2023.

[7] Y. Zheng *et al.*, "Multi-agent DRL for resource allocation in AoI-aware energy-efficient C-V2X networks," in *Proc. 2024 IEEE 29th Int. Workshop Comput. Aided Model. Des. Commun. Links Netw. (CAMAD)*, Athens, Greece, Oct.21–23 2024, pp. 1–6.

[8] C. He *et al.*, "Scheduling phase period design for sub-6 GHz assisted millimeter wave vehicle-to-everything (V2X) communication scheduler," *IEEE Trans. Veh. Technol.*, vol. 74, no. 4, pp. 6294–6305, Apr. 2025.

[9] J.-E. Zhang, G. Liu, W.-W. Yang, and J.-X. Chen, "A tri-frequency shared-aperture antenna for cooperative work of V2X and millimeter-wave bands," *IEEE Antennas Wirel. Propag. Lett.*, vol. 24, no. 3, pp. 776–780, Mar. 2025.

[10] X. Zhang, S. Fang, Y. Shen, X. Yuan, and Z. Lu, "Hierarchical velocity optimization for connected automated vehicles with cellular vehicle-to-everything communication at continuous signalized intersections," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 3, pp. 2944–2955, Mar. 2024.

[11] L. Sun, J. Liang, and G. Muhammad, "Distillate a sparse-meta time series classifier for open radio access network-based cellular vehicle-to-everything," *IEEE Trans. Veh. Technol.*, vol. 73, no. 7, pp. 9262–9271, Jul. 2024.

[12] L. Zhao *et al.*, "Generative abnormal data detection for enhancing cellular vehicle-to-everything-based road safety," *IEEE Trans. Green Commun. Netw.*, vol. 8, no. 4, pp. 1466–1478, Dec. 2024.

[13] X. Zhang *et al.*, "An intelligent obstacle detection for autonomous mining transportation with electric locomotive via cellular vehicle-to-everything and vehicular edge computing," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 3, pp. 3177–3190, Mar. 2024.

[14] C. Giovannetti, N. Decarli, S. Bartoletti, R. A. Stirling-Gallacher, and B. M. Masini, "Target positioning accuracy of V2X sidelink joint communication and sensing," *IEEE Wirel. Commun. Lett.*, vol. 13, no. 3, pp. 849–853, Mar. 2024.

[15] Z. Shi and J. Liu, "Sparse code multiple access assisted resource allocation for 5G V2X communications," *IEEE Trans. Commun.*, vol. 70, no. 10, pp. 6661–6677, Oct. 2022.

[16] ——, "An M3RSMA-based roadside cooperative message delivery scheme for complex intersection," *IEEE Trans. Wirel. Commun.*, vol. 24, no. 7, pp. 6036–6051, Jul. 2025.

[17] S. Zhang *et al.*, "Transformer-based channel prediction for rate-splitting multiple access-enabled vehicle-to-everything communication," *IEEE Trans. Wirel. Commun.*, vol. 23, no. 10, pp. 12 717–12 730, Oct. 2024.

[18] L. Chen, Y. He, F. Yu, W. Pan, and Z. Ming, "A novel reinforcement learning method for autonomous driving with intermittent vehicle-to-everything (V2X) communications," *IEEE Trans. Veh. Technol.*, vol. 73, no. 6, pp. 7722–7732, Jun. 2024.

[19] M. Moniruzzaman, A. Yassine, and R. Benlamri, "Blockchain and federated reinforcement learning for vehicle-to-everything energy trading in smart grids," *IEEE Trans. Artif. Intell.*, vol. 5, no. 2, pp. 839–853, Feb. 2024.

[20] G. Chai *et al.*, "Platoon partition and resource allocation for ultra-reliable V2X networks," *IEEE Trans. Veh. Technol.*, vol. 73, no. 1, pp. 147–161, Jan. 2024.

[21] Z. Dong *et al.*, "Dynamic manager selection assisted resource allocation in URLLC with finite block length for 5G-V2X platoons," *IEEE Trans. Veh. Technol.*, vol. 71, no. 11, pp. 11 336–11 350, Nov. 2022.

[22] Z. Dong, X. Zhu, and Y. Jiang, "CoMP-based seamless handover and resource allocation for 5G-V2X platoon systems," in *Proc. ICC 2022 - IEEE Int. Conf. Commun.*, Seoul, Korea, Republic of, May 2022, pp. 2010–2015.

[23] W. Gao, C. Wu, L. Zhong, and K.-L. A. Yau, "Communication resources management based on spectrum sensing for vehicle platooning," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 2, pp. 2251–2264, Feb. 2023.

[24] P. Zhang *et al.*, "Joint optimization of platoon control and resource scheduling in cooperative vehicle-infrastructure system," *IEEE Trans. Intell. Veh.*, vol. 8, no. 6, pp. 3629–3646, Jun. 2023.

[25] L. Cao, S. Roy, and H. Yin, "Resource allocation in 5G platoon communication: Modeling, analysis and optimization," *IEEE Trans. Veh. Technol.*, vol. 72, no. 4, pp. 5035–5048, Apr. 2023.

[26] M. Farzanullah and T. Le-Ngoc, "Platoon leader selection, user association and resource allocation on a C-V2X based highway: A reinforcement learning approach," in *Proc. ICC 2023 - IEEE Int. Conf. Commun.*, Rome, Italy, May-Jun. 2023, pp. 5396–5401.

[27] Y. Shi, H. Dong, C. R. He, Y. Chen, and Z. Song, "Mixed vehicle platoon forming: A multi-agent reinforcement learning approach," *IEEE Internet Things J.*, Feb. 2025, DOI: 10.1109/JIOT.2025.3535732.

[28] Q. Abbas, S. A. Hassan, H. Jung, and M. S. Hossain, "On minimizing the age of information in NOMA-based vehicular networks using markov decision process," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 12, pp. 15 557–15 567, Dec. 2023.

[29] A. Mallik, D. Chen, K. Han, J. Xie, and Z. Han, "Unleashing the true power of age-of-information: Service aggregation in connected and autonomous vehicles," in *Proc. ICC 2024 - IEEE Int. Conf. Commun.*, Denver, CO, Jun. 2024, pp. 1709–1714.

[30] M. Bezmenov, Z. Utkovski, and S. Stanczak, "The impact of blind retransmissions on the age of information in NR-V2X," in *Proc. 2024 IEEE 100th Veh. Technol. Conf. (VTC2024-Fall)*, Washington, DC, Oct. 2024, pp. 1–6.

[31] S. Park, C. Park, S. Jung, M. Choi, and J. Kim, "Age-of-information aware caching and delivery for infrastructure-assisted connected vehicles," *IEEE Trans. Veh. Technol.*, vol. 73, no. 7, pp. 10 681–10 696, Jul. 2024.

[32] Z. Mlika and S. Cherkaoui, "Deep deterministic policy gradient to minimize the age of information in cellular V2X communications," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 23 597–23 612, Dec. 2022.

[33] X. Li, J. Li, B. Yin, J. Yan, and Y. Fang, "Age of information optimization in UAV-enabled intelligent transportation system via deep reinforcement learning," in *Proc. 2022 IEEE 96th Veh. Technol. Conf. (VTC2022-Fall)*, London, United Kingdom, Sep. 2022, pp. 1–5.

[34] M. Kim *et al.*, "Age of information based beacon transmission for reducing status update delay in platooning," *IEEE Trans. Veh. Technol.*, vol. 71, no. 10, pp. 11 306–11 310, Oct. 2022.

[35] S. Zhou, S. Li, and G. Tan, "Age of information in V2V-enabled platooning systems," *IEEE Internet Things J.*, vol. 11, no. 3, pp. 4015–4028, Feb. 2024.

[36] Z. Li, L. Xiang, and X. Ge, "Age of information modeling and optimization for fast information dissemination in vehicular social networks," *IEEE Trans. Veh. Technol.*, vol. 71, no. 5, pp. 5445–5459, May 2022.

[37] M. R. Abedi *et al.*, "Safety-aware age of information (S-AoI) for collision risk minimization in cell-free mMIMO platooning networks," *IEEE Trans. Netw. Serv. Manag.*, vol. 21, no. 3, pp. 3035–3053, Jun. 2024.

[38] S. Gyawali, S. Xu, Y. Qian, and R. Q. Hu, "Challenges and solutions for cellular based V2X communications," *IEEE Commun. Surv. Tutor.*, vol. 23, no. 1, pp. 222–255, First Quart. 2021.

[39] M. H. C. Garcia *et al.*, "A tutorial on 5G NR V2X communications," *IEEE Commun. Surv. Tutor.*, vol. 23, no. 3, pp. 1972–2026, Third Quart. 2021.

[40] 3GPP, "Study on LTE-based V2X services," 3rd Generation Partnership Project (3GPP), Technical Report (TR) 36.885, 2016, version 14.0.0.

[41] S. Chen *et al.*, "Vehicle-to-everything (V2X) services supported by LTE-based systems and 5G," *IEEE Comm. Stand. Mag.*, vol. 1, no. 2, pp. 70–76, 2017.

[42] H. U. Sheikh and L. Bölöni, "Multi-agent reinforcement learning for problems with combined individual and team reward," in *Proc. Int. Joint Conf. Neural Netw.*, Glasgow, UK, Jul. 2020, pp. 1–8.

[43] C. Sun, W. Liu, and L. Dong, "Reinforcement learning with task decomposition for cooperative multiagent systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 5, pp. 2054–2065, May 2021.

[44] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. 35th Int. Conf. Mach. Learn.*, Stockholm, Sweden, Jul. 2018, pp. 1587–1596.

[45] T. Huang *et al.*, "A survey on green 6G network: Architecture and technologies," *IEEE Access*, vol. 7, pp. 175 758–175 768, Dec. 2019.

[46] Z. Xu and A. Petropulu, "A bandwidth efficient dual-function radar communication system based on a MIMO radar using OFDM waveforms," *IEEE Trans. Signal Process.*, vol. 71, pp. 401–416, Feb. 2023.

[47] D. P. Bertsekas, *Dynamic programming and optimal control: Volume I*. Belmont, MA, USA: Athena Sci., 1995.

[48] R. Lowe *et al.*, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 6379–6390.

[49] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning (iclr 2016)," in *Proc. Int. Conf. Learn. Representations*, San Juan, Puerto Rico, May 2016.

[50] K. Hara, D. Saito, and H. Shouno, "Analysis of function of rectified linear unit used in deep learning," in *Proc. 2015 Int. Jt. Conf. Neural Netw. (IJCNN)*, Killarney, Ireland, Jul. 2015, pp. 1–8.

[51] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," San Diego, CA, USA, May 2015.

[52] P. Kyösti *et al.*, "IST-4-027756 WINNER II D1. 1.2 V1. 2 WINNER II channel models," EBITG, TUI, UOULU, CU/CRC, NOKIA, Technical Report (TR), 2008.