A Bidirectional-AoI-Aware Multi-Agent Deep Reinforcement Learning Framework for Vehicular Platooning in Segmented Waveguide-Based Pinching Antenna Systems

Yuxiang Zheng, *Graduate Student Member, IEEE,* Simon L. Cotton, *Fellow, IEEE* and Trung Q. Duong, *Fellow, IEEE*

***Abstract*—Ensuring reliable and low-latency vehicle-to-everything (V2X) communications in high-speed transport settings remains a significant challenge due to severe path loss brought about by non-line-of-sight (NLoS) and coverage gaps in conventional cellular infrastructure. While dielectric waveguide-based pinching antenna (PA) systems have been proposed to mitigate these physical limitations, they suffer from substantial in-waveguide attenuation over long distances. To address these challenges, we propose a segmented waveguide-enabled pinching-antenna (SWAN) architecture in platoon-based V2X networks. By employing dynamic segment selection, SWAN maintains robust line-of-sight (LoS) connectivity while mitigating the in-waveguide attenuation inherent in conventional PA structures. We formulate a joint resource allocation (RA) and mode selection problem to minimise the age of information (AoI) for both uplink platoon monitoring and downlink traffic broadcasting, whilst ensuring the exchange of intra-platoon cooperative awareness messages (CAMs) and minimising power consumption. To solve this high-dimensional problem, we propose a decomposed multi-agent deep deterministic policy gradient (DE-MADDPG) algorithm augmented with twin delayed (TD3) critics by considering each vehicle platoon (VP) as an agent. This approach decouples system-wide coordination from local executions of VPs, enabling efficient learning in dynamic environments. Extensive simulations demonstrate that the proposed framework significantly outperforms standard reinforcement learning (RL) baseline methods, achieving near-optimal uplink and downlink AoI performance, with an average gap of 4.8% to exhaustive search, and near-perfect CAM delivery probability (CDP), which approaches 100%, even under dense traffic conditions.**

***Index Terms*—Age of information, multi-agent deep reinforcement learning, pinching antenna, resource allocation, vehicle-to-everything.**

Y. Zheng is with the Faculty of Engineering and Applied Science, Memorial University, St. John's, NL A1C 5S7, Canada (e-mail: y.zheng@mun.ca).

S. L. Cotton is with the Centre for Wireless Innovation (CWI), School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, BT3 9DT, Belfast, U.K. (email: simon.cotton@qub.ac.uk).

T. Q. Duong is with the Faculty of Engineering and Applied Science, Memorial University, St. John's, NL A1C 5S7, Canada, and with the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast, U.K. (e-mail: tduong@mun.ca).

Corresponding author is Trung Q. Duong.

The work of T. Q. Duong was supported in part by the Canada Excellence Research Chair (CERC) Program CERC-2022-00109, in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant Program RGPIN-2025-04941, and in part by the NSERC Collaborative Research and Training Experience (CREATE) Program (Grant number 596205-2025). The work of S. L. Cotton was supported by U.K. Engineering and Physical Sciences Research Council (EPSRC) through the EPSRC Hub on All Spectrum Connectivity under Grant EP/X040569/1 and Grant EP/Y037197/1.

I. Introduction

The advancement of intelligent transportation systems (ITS) fundamentally relies on robust vehicle-to-everything (V2X) communications to support autonomous driving applications [1]–[6]. A prime example is vehicle platooning, where a group of vehicles travels in close proximity within the same lane as a single convoy, i.e., a vehicle platoon (VP) to improve traffic throughput and control efficiency [7]–[15]. To maintain string stability and operational safety, vehicles within the same platoon must exchange local status information periodically via cooperative awareness messages (CAMs) [16] using vehicle-to-vehicle (V2V) communications. These periodic updates ensure synchronisation among vehicles, thereby preventing collisions and enabling smooth coordinated movements on high-speed roads.

However, reliance solely on local V2V exchanges within the same VP is insufficient for comprehensive traffic management across multiple VPs. The system also requires robust vehicle-to-infrastructure (V2I) links where the infrastructure, specifically the base station (BS), aggregates platoon states and disseminates global traffic data to all connected vehicles. In such safety-critical scenarios, low latency alone offers insufficient performance guarantees, while the freshness of the delivered data is equally critical. Consequently, the age of information (AoI) metric, which quantifies the time elapsed since the latest successful update sent by the VPs and received by the BS, has become an important measurement for system timeliness [4], [9]–[15]. Minimising AoI is essential for maintaining accurate traffic awareness, particularly in high-speed vehicular environments.

Establishing high-frequency V2I communications via conventional cellular infrastructure presents significant challenges in high-mobility scenarios (e.g., on high-speed roads). These challenges include frequent handovers, substantial Doppler shifts, coverage gaps, and non-line-of-sight (NLoS) channels with severe path loss, all of which degrade connection reliability. To address the physical limitations of discrete cellular infrastructure, pinching antennas (PAs) [17]–[22] have emerged as a promising solution. Unlike traditional antenna arrays, a PA system employs one or more dielectric waveguides deployed continuously along the road. By creating controlled disruptions to the internal electromagnetic field, the waveguide leaks signals at specific locations (pinch points) [23], [24] to communicate with passing vehicles. This controlled localised leakage provides continuous wireless coverage and predominantly line-of-sight (LoS) connections, effectively eliminating the coverage gaps and path loss associated with conventional BS architectures.

A. Related Works

Prior works have extensively explored resource allocation (RA) in V2X networks. Studies have optimised resource utilisation and reliability for CAM broadcasting in C-V2X Mode 4 by adapting to real-time traffic density fluctuations [1]. Issues regarding inter-region interference and beam misalignment in multiple roadside units (RSUs) V2I scenarios were addressed by incorporating integrated sensing and communication (ISAC) with coordinated multi-point (CoMP) technology [2]. Furthermore, the reliability of cellular-V2X (C-V2X) semi-persistent scheduling (SPS) for vehicle platooning was enhanced by mitigating consecutive collisions caused by hidden terminals and simultaneous resource selection [5]. Ultra-reliable intra-platoon communication under uncertain channel state information (CSI) was ensured in [8] by formulating a distributionally robust optimisation problem for dynamic platoon partition and RA.

Given the dynamic nature of vehicular environments, machine learning (ML)-based methods have become a compelling approach for solving these complex decision-making problems. For instance, the limitations of local observations in distributed C-V2X RA were addressed by integrating graph neural networks (GNNs) with deep reinforcement learning (DRL) in [3]. Similarly, the complexity of RA in dynamic vehicle platooning networks was managed by HierNet in [7], which is a hierarchical reinforcement learning (RL) framework that decouples decision-making into high-level coordination and low-level execution. A distributed RA framework using multi-agent DRL (MADRL) with the deep deterministic policy gradient (DDPG) algorithm was presented in [11], allowing individual platoons to manage radio resources. Subsequently, some energy-efficient strategies were presented in [12], [15]. Advanced quantum computing techniques were also applied in [13], [14] to handle complex vehicular environments. The scope of learning-based resource management further extends to aerial and trajectory-aware networks. A distributed many-agent framework was developed for energy-efficient UAV-driven multi-access edge computing in [25], while the radio resource management of cellular-connected UAVs was addressed through a reinforcement learning approach in [26]. Furthermore, mobility-aware path planning for cellular-connected UAVs was improved by a quantum-inspired experience replay mechanism in [27]. However, these works predominantly rely on conventional cellular channel models subject to the aforementioned constraints.

Minimising information age is critical for time-sensitive vehicular applications. The optimisation of reconfigurable intelligent surface (RIS)-aided IoV networks was investigated in [4], focusing on minimising the AoI for V2I links while ensuring V2V payload reliability with a soft actor-critic DRL algorithm. AoI in V2V-enabled platooning systems was analysed in [9], and a joint communication and control model was proposed. The issue of safety in platooning was explored in [10], where a safety-aware AoI metric was introduced with the real-time collision risk assessment of connected and autonomous vehicles (CAVs) to design more efficient basic safety message (BSM) transmission protocols. The proposed algorithms in [11]–[15] also coordinated platoons in an AoI-aware manner, ensuring timely dissemination of CAMs and safety-critical messages to the RSU within C-V2X platoon networks. However, existing literature largely treats AoI optimisation within the context of standard cellular topologies. The practical limitations of conventional cellular communication networks constrain the real deployment of these frameworks. In addition, current research predominantly focuses on a single link direction, neglecting the necessity of the joint optimisation of uplink and downlink freshness required for holistic situational awareness across all vehicles.

TABLE I: Comparison with Representative Related Studies.

Ref.	Platoon	Infrastructure	AoI	CAM Delivery
[3]	×	Cellular	–	✓
[7]	✓	Cellular	–	×
[11]	✓	Cellular	UL	✓
[4]	×	RIS	UL	✓
[9]	✓	Cellular	UL	×
[10]	✓	Cellular	UL	✓
[12], [15]	✓	Cellular	UL	✓
[13], [14]	✓	Cellular	UL	✓
[22]	×	SWAN	–	×
Ours	✓	SWAN	UL & DL	✓

The concept of using dielectric waveguides as leaky-wave antennas was established in [17]. Subsequent studies have investigated physical layer security [18], uplink performance optimisation [19], and antenna activation in NOMA-assisted systems [20]. The unique challenges of channel estimation in PA systems were addressed in [21]. In the above literature, PA systems demonstrate excellent feasibility of supporting wireless communication infrastructure under different scenarios. They can be easily deployed over arbitrary distances with relatively low expenditure and create strong LoS connections. However, conventional PA implementations exhibit an inherent limitation: significant in-waveguide attenuation during long-distance transmission, reaching approximately 8 dB per 100 meters (m) [28], [29]. As the signal propagates through a long continuous waveguide, the power decays exponentially, limiting the effective service range and energy efficiency. Additionally, although multiple PAs can be activated simultaneously during the uplink transmission, different signals received by different PAs can interfere with each other inside the waveguide. The signal received by one PA can also be re-radiated when it passes through other PAs, inducing the inter-antenna radiation (IRA) [22]. To mitigate these issues, a segmented waveguide-enabled pinching-antenna (SWAN) architecture was proposed in [22]. By dividing the waveguide into independently fed segments and employing a "segment selection", "segment aggregation", or "segment multiplexing" strategy, the SWAN system significantly reduces propagation losses, enhances signal quality for long-distance deployments, and enables multi-PA uplink transmission.

To position the present work within the existing literature, Table I contrasts its key features against representative related studies. As summarised therein, our framework is the only platoon-based design built on the SWAN architecture that jointly optimises bidirectional (uplink and downlink) AoI while ensuring communication reliability.

B. Motivation and Contributions

To address the current implementation gap in PA-supported V2X networks and overcome the propagation limitations of general PA systems, this study adopts the segmented waveguide-based PA architecture proposed in [22] as the physical layer model. While the SWAN architecture resolves the attenuation limitation, its integration with V2X RA protocols is highly non-trivial. The dynamic activation of waveguide segments and specific PAs further complicates the existing optimisation of the AoI, CAM delivery, and power allocation in the vehicular networks. Moreover, few works address the joint optimisation of uplink and downlink AoI. Most existing studies [11]–[15] focus exclusively on the timeliness of uplink platoon-state reporting, neglecting the equally critical requirement for the timely downlink dissemination of global traffic data to all participating vehicles.

To bridge these gaps, this paper proposes a novel PA-assisted V2X framework built on the SWAN model to address the RA challenges of a platoon-based V2X network on a high-speed road. The main contributions are summarised as follows:

- We introduce the joint optimisation of both uplink and downlink AoI simultaneously within a PA-assisted network. This bidirectional information freshness optimisation is crucial for holistic situational awareness but creates a highly coupled decision space.
- We develop a realistic high-speed V2X system model integrating the SWAN architecture, accounting for specific segment selection and PA activation constraints to minimise propagation loss and mitigate the IRA effect. This represents an early study evaluating the SWAN architecture within a realistic, simulated vehicular environment, advancing beyond idealised link-level analyses.
- We formulate a joint optimisation problem that simultaneously minimises the bidirectional AoI for information freshness, while maximising the CAM delivery probability (CDP) for intra-platoon safety, and minimising the power consumption for a more sustainable system. To solve this highly coupled problem, we propose a DE-MADDPG framework featuring a novel physics-aware reward decomposition mechanism. This mechanism is fundamentally aligned with the SWAN operational modes: the global reward manages the system-wide downlink broadcasting inherent to the SWAN segment activation, while the local reward optimises point-to-point uplink connectivity between the vehicle and a specific PA.

TABLE II: Summary of Key Notations.

Notation	Description
M	Number of waveguide segments
N	Number of activatable PAs per segment
$\mathbf{P}_m$	Coordinate of the m^{th} feed point
$\mathbf{P}_{\langle m,n\rangle}$	Coordinate of the n^{th} PA in segment m
L	Total number of platoons
V_l	Number of vehicles in platoon l
LV_l	Leading vehicle (LV) of platoon l
$\text{MV}_{\langle l,v\rangle}$	The v^{th} member vehicle (MV) in platoon l
$\mathbf{P}_l$	Coordinate of LV_l
$\mathbf{P}_{\langle l,v\rangle}$	Coordinate of $\text{MV}_{\langle l,v\rangle}$
$h^{\text{iw}}_{\text{BS-PA}_{\langle m,n\rangle}}$	In-waveguide propagation channel coefficient
$h^{\text{fs}}_{\text{PA}_{\langle m,n\rangle}\text{-LV}_l}$	Free-space channel coefficient
$h^{\text{V2V}}_{\text{LV}_l\text{-MV}_{\langle l,v\rangle}}$	V2V channel coefficient
$\Phi_{l,t}$	Mode selection index for LV_l at time t
$\Upsilon_{l,t}$	RB assigned to LV_l for V2V communications
$p^{\text{LV}}_{l,t}$	Transmit power of LV_l at time t
$p^{\text{PA}}_{\langle m,n\rangle,t}$	Transmit power of $\text{PA}_{\langle m,n\rangle}$ at time t
$\mathcal{C}^{\text{dl}}_{l,m,t}$	Achievable downlink V2I rate
$\mathcal{C}^{\text{ul}}_{l,\langle m,n\rangle,t}$	Achievable uplink V2I rate
$\mathcal{C}^{\text{V2V}}_{\langle l,v\rangle,t}$	Achievable V2V rate
$A^{\text{ul}}_{l,t}$	Uplink AoI of platoon l at time t
A^{dl}_{t}	System-wide downlink AoI at time t
$\mathcal{S}$	Joint state space of the multi-agent system
$\mathcal{A}$	Joint action space of the multi-agent system
$s_{l,t}$	Local state of agent l at time t
$a_{l,t}$	Local action of agent l at time t
r^g_t	Global reward for downlink broadcasting
$r^u_{l,t}$	Local reward for uplink AoI of platoon l

C. Paper Structure and Notations

The remainder of this paper is structured as follows. Section II presents the PA infrastructure model and the platoon-based V2X system model, followed by the problem formulation. Section III outlines the DE-MADDPG framework and the proposed algorithm. Section IV analyses the complexity of the proposed and baseline algorithms. Numerical results and discussions are provided in Section V. Finally, Section VI concludes the paper.

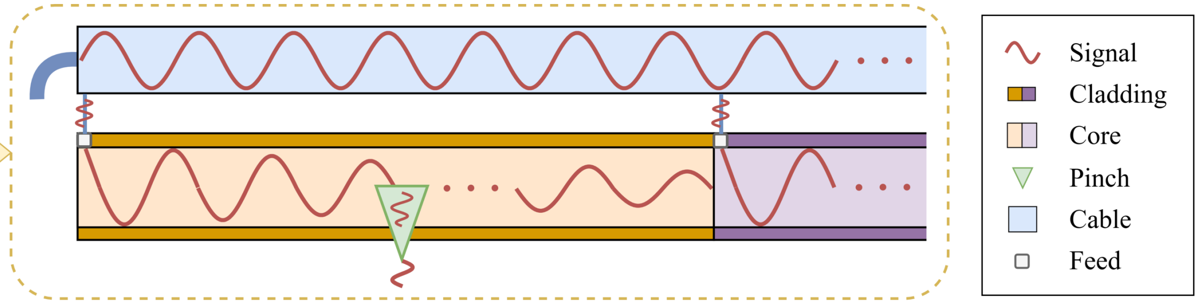
Notation: Numbers are written as italic letters while vectors are denoted in bold. $\mathbb{N}$ and $\mathbb{Z}^+$ denote the set of natural numbers and positive integers, respectively. j is the imaginary unit. c is the speed of light. $|\cdot|$ represents the magnitude of a scalar. $||\cdot||$ denotes the vector norm. $\lfloor\cdot\rfloor$ represents the floor function. $\Pr(\cdot)$ denotes the probability operator. $\mathcal{F}_x(\cdot)$ represents mapping functions, where $x \in \{1,\ldots,5\}$. To enhance readability, a summary of key mathematical symbols and variables used in the system model and problem formulation is provided in Table II.

II. System Model and Problem Formulation

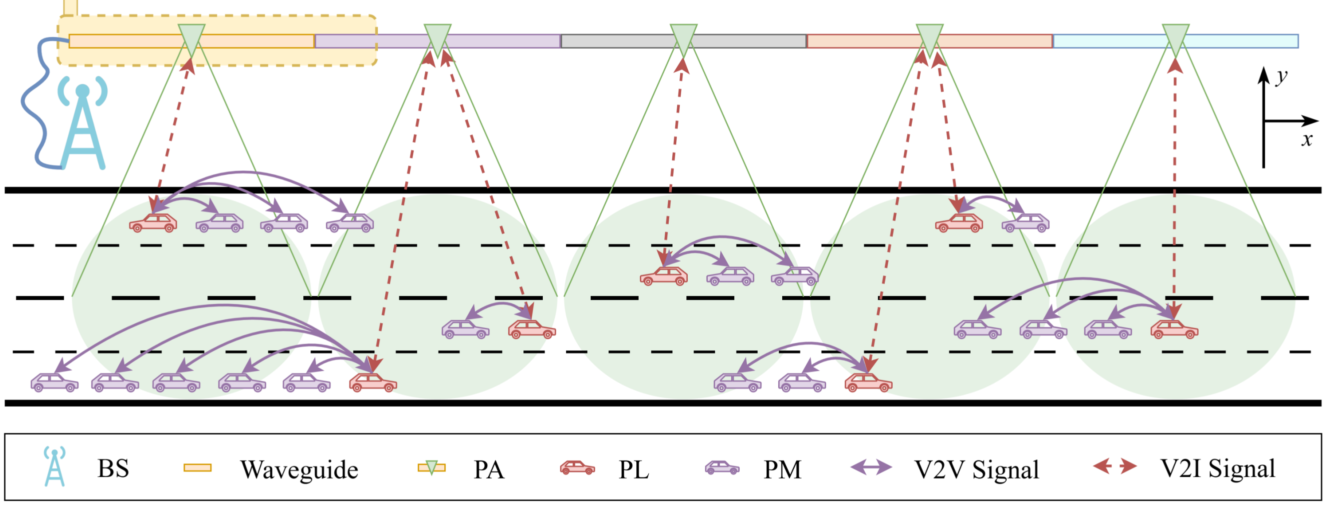
This section outlines the system model and problem formulation.

A. Infrastructure Model

Fig. 1 illustrates the proposed V2X system deployed on a two-way, four-lane high-speed road of length D_{road} and width D_{width}. The communication infrastructure is supported by a segmented dielectric waveguide-based PA system, installed along the road centre and connected to a BS. This architecture



(a) Segmented dielectric waveguide.



(b) System setup.

Fig. 1: Design of the dielectric waveguide and PA element, and a dielectric waveguide-based PA assisted high-speed road V2X system. (a) Segmented dielectric waveguide. (b) System setup.

aggregates platoon-state information from each VP, relays it to the BS, and disseminates global traffic data to all VPs, thereby ensuring efficient and safe vehicular operation. A Cartesian coordinate system is established where the waveguide origin is located at the left-most point $\mathbf{P}_0 = (0, 0, D_{\text{PA}})$, with D_{PA} denoting the infrastructure height. The operational volume is defined by the x-axis $[0, D_{road}]$, y-axis $[-D_{width}/2, D_{width}/2]$, and z-axis $[0, D_{\text{PA}}]$.

The dielectric waveguide structure, detailed in Fig. 1(a), comprises a central core encased in the outside cladding. The cladding material possesses a lower refractive index than the core, which facilitates signal guidance via total internal reflection [30]. This structure is characterised by low propagation attenuation at mmWave frequencies, making it suitable for efficient signal transmission over long distances [31].

The waveguide is divided into $M \in \mathbb{Z}^+$ segments (assuming ideal isolation between adjacent segments), each of length D_{road}/M, which are differentiated visually in Fig. 1(b). A parallel low-loss cable (e.g., coaxial or optical fibre [22]) runs from the BS to connect to feed points located at the start of each segment, as shown in Fig. 1(a). The coordinate of the m^{th} feed point is given by $\mathbf{P}_m = (X_m, 0, D_{\text{PA}})$, where $X_m = mD_{road}/M$ and $m \in \{0, 1, \ldots, M-1\}$. These points serve as interfaces to feed signals into the waveguide for downlink communications and receive signals for uplink communications.

Each segment contains up to $N \in \mathbb{Z}^+$ activatable PAs (pinches), indexed by $n \in \{0, \ldots, N-1\}$ and distributed along the segment. The localised leakage of these PAs supports bidirectional V2I communication [17], with each PA covering a dedicated zone (shaded green in Fig. 1(b)). The position of PA n in segment m, denoted as $\text{PA}_{\langle m,n \rangle}$, is $\mathbf{P}_{\langle m,n \rangle} = \left(X_{\langle m,n \rangle}, 0, D_{\text{PA}}\right)$, where $X_{\langle m,n \rangle} \in [X_m, X_m + D_{road}/M)$. The minimum inter-antenna spacing of any two PAs must satisfy a threshold $D_{\text{PA-PA}}$ to avoid mutual coupling effects [32]:

$$\left|X_{\langle m,n \rangle} - X_{\langle m,n' \rangle}\right| \geq D_{\text{PA-PA}}, \forall n \neq n'. \tag{1}$$

B. Vehicle Model

Within the covered area of the PA system, $L \in \mathbb{Z}^+$ VPs travel along the high-speed road. Each platoon comprises one leading vehicle (LV) and several member vehicles (MVs), denoted as LV_l and $\text{MV}_{\langle l,v \rangle}$ for the l^{th} platoon VP_l, where $l \in \{0, \ldots, L-1\}$ and $v \in \{1, \ldots, V_l - 1\}$.

We adopt a simplified positional model assuming vehicles are centred within their respective lanes with a z-axis elevation of 0. The coordinate of LV_l is $\mathbf{P}_l = (X_l, Y_l, 0)$, where $X_l \in [0, D_{road}]$ is the x-axis position. The $y-$axis position Y_l depends on the occupied lane, specifically $Y_l \in \{\pm D_{width}/8, \pm 3D_{width}/8\}$. The MVs follow the LV in the same lane, and the position of $\text{MV}_{\langle l,v\rangle}$ is $\mathbf{P}_{\langle l,v\rangle} = \left(X_{\langle l,v\rangle}, Y_l, 0\right)$. Here, $X_{\langle l,v\rangle} = X_l \pm vD_{gap}$, where D_{gap} is the intra-platoon spacing and the sign depends on the platoon's direction of travel.

Fig. 1(b) depicts the two primary communication modes: V2V signals (solid purple arrows) for intra-platoon communication, and V2I signals (dashed red arrows) for interaction with the PA system. These two wireless communication modes are analysed in detail in the following subsections.

C. Pinching-Antenna assisted V2I Communication Model

As defined in Section II-A, each waveguide segment connects to the BS via a cable. Signals traversing the waveguide experience attenuation and phase shifts distinct from those in the assumed lossless connecting cable. The in-waveguide propagation channel coefficient between $\text{PA}_{\langle m,n\rangle}$ and the BS is modelled as [33]

$$\begin{aligned} h^{\text{iw}}_{\text{BS-PA}_{\langle m,n\rangle}} &= e^{-\left(\alpha^{\text{iw}} + j\frac{2\pi n_{\text{eff}}}{\lambda_{\text{fs}}}\right)\left|\left|\mathbf{P}_{\langle m,n\rangle} - \mathbf{P}_m\right|\right|} \\ &= e^{-\left(\alpha^{\text{iw}} + j\frac{2\pi}{\lambda_g}\right)\left|X_{\langle m,n\rangle} - X_m\right|}, \end{aligned} \tag{2}$$

where α^{iw} denotes the in-waveguide attenuation coefficient [34], n_{eff} is the core's effective refractive index [30], and λ_{fs} and $\lambda_g = \lambda_{\text{fs}}/n_{\text{eff}}$ [33] denote the free-space and guided wavelengths, respectively.

Subsequently, signals are radiated into free space. Given the elevated placement of the PAs, we consider LoS propagation [35]. The free-space channel coefficient between $\text{PA}_{\langle m,n\rangle}$ and LV_l is modelled as [36]

$$h^{\text{fs}}_{\text{PA}_{\langle m,n\rangle}\text{-LV}_l} = \frac{\sqrt{\beta^{\text{PA}}}e^{-j\frac{2\pi}{\lambda_{\text{fs}}}\left|\left|\mathbf{P}_{\langle m,n\rangle} - \mathbf{P}_l\right|\right|}}{\left|\left|\mathbf{P}_{\langle m,n\rangle} - \mathbf{P}_l\right|\right|}, \tag{3}$$

where $\beta^{\text{PA}} = \left(c/4\pi f_c^{\text{PA}}\right)^2$ is derived from the Friis transmission formula and accounts for the frequency-dependent free-space path loss [37], and f_c^{PA} is the carrier frequency used by the PA system.

The final composite channel coefficient spanning the BS, $\text{PA}_{\langle m,n\rangle}$, and LV_l is expressed as

$$\begin{aligned} &h^{\text{iw-fs}}_{\text{BS-PA}_{\langle m,n\rangle}\text{-LV}_l} \\ &= h^{\text{iw}}_{\text{BS-PA}_{\langle m,n\rangle}} h^{\text{fs}}_{\text{PA}_{\langle m,n\rangle}\text{-LV}_l} \end{aligned} \tag{4}$$

$$= \begin{cases} e^{-\alpha^{\text{iw}}\left|X_{\langle m,n\rangle} - X_m\right|} & \text{(4a)} \\ \dfrac{\sqrt{\beta^{\text{PA}}}}{\left|\left|\mathbf{P}_{\langle m,n\rangle} - \mathbf{P}_l\right|\right|} & \text{(4b)} \\ e^{-j\left(\frac{2\pi}{\lambda_g}\left|X_{\langle m,n\rangle} - X_m\right| + \frac{2\pi}{\lambda_{\text{fs}}}\left|\left|\mathbf{P}_{\langle m,n\rangle} - \mathbf{P}_l\right|\right|\right)} & \text{(4c)} \end{cases},$$

where (4a) represents the in-waveguide attenuation, (4b) stands for the free-space path loss, while (4c) is the overall phase shift in both waveguide and free space. In our implementation, the phase shift accounts for both the propagation delay and the Doppler-induced phase variation over time.

We employ the "segment selection" strategy [22], where a hardware switch connects a single segment to the radio frequency (RF) front-end. For uplink transmission, a single PA within the chosen segment is activated to receive information from one LV. The received uplink signal power sent by LV_l is written as

$$p^{\text{ul}}_{l,\langle m,n\rangle,t} = p^{\text{LV}}_{l,t}\left|h^{\text{iw-fs}}_{\text{BS-PA}_{\langle m,n\rangle}\text{-LV}_l}\right|^2, \tag{6}$$

where $p^{\text{LV}}_{l,t}$ denotes the transmit power of LV_l and t indexes discrete time slots of duration Δt.

During downlink transmission, an optimum number of PAs, $N_{m^*} \leq N$ [22], are activated in the selected segment m^* to broadcast identical information to all LVs. The downlink signal power received by LV_l can be written as

$$p^{\text{dl}}_{l,m,t} = \left|\sum_{n=0}^{N_{m^*}-1}\sqrt{p^{\text{PA}}_{\langle m,n\rangle,t}}h^{\text{iw-fs}}_{\text{BS-PA}_{\langle m,n\rangle}\text{-LV}_l}\right|^2, \tag{7}$$

where $p^{\text{PA}}_{\langle m,n\rangle,t}$ denotes the transmit power of $\text{PA}_{\langle m,n\rangle}$. Assuming the total segment power $p^{\text{Seg}}_{m,t}$ is distributed equally among active PAs [38], [39] (i.e., $p^{\text{PA}}_{\langle m,n\rangle,t} = p^{\text{Seg}}_{m,t}/N_{m^*}$), the downlink power becomes

$$p^{\text{dl}}_{l,m,t} = \frac{p^{\text{Seg}}_{m,t}}{N_{m^*}}\left|\sum_{n=0}^{N_{m^*}-1}h^{\text{iw-fs}}_{\text{BS-PA}_{\langle m,n\rangle}\text{-LV}_l}\right|^2. \tag{8}$$

D. Intra-Platoon V2V Communication Model

The V2V channel incorporates path loss (B1–Urban microcell model [40]), log-normal shadowing, and small-scale Rayleigh fading, reflecting the NLoS conditions caused by mutual blockage within the platoon. The channel coefficient between LV_l and $\text{MV}_{\langle l,v\rangle}$ is modelled as

$$h^{\text{V2V}}_{\text{LV}_l\text{-MV}_{\langle l,v\rangle}} = \sqrt{\beta^{\text{V2V}}_t \xi D^{-\alpha}_{\langle l,v\rangle,t}}\, g_t, \tag{9}$$

where β^{V2V}_t is the path loss constant, ξ is the shadowing gain, $D_{\langle l,v\rangle,t}$ is the inter-vehicle distance between LV_l and $\text{MV}_{\langle l,v\rangle}$, α is the path loss coefficient, and $g_t \sim \mathcal{CN}(0,1)$ with $|g_t|$ following the Rayleigh distribution. The signal power received by $\text{MV}_{\langle l,v\rangle}$ is given by:

$$p^{\text{V2V}}_{\langle l,v\rangle,t} = p^{\text{LV}}_{l,t}\left|h^{\text{V2V}}_{\text{LV}_l\text{-MV}_{\langle l,v\rangle}}\right|^2, \tag{10}$$

utilising the same transmit power variable $p^{\text{LV}}_{l,t}$ as the uplink case.

E. Overall Wireless Communications Model

We define a mode selection index $\Phi_{l,t} \in \{0,1,2\}$ for LV_l at time t, corresponding to V2V, uplink V2I, and downlink V2I modes, respectively. Following our previous works on platoon-based C-V2X systems [12]–[15], we employ orthogonal frequency division multiplexing (OFDM) [41] to handle frequency-selective wireless channels. The system operates on two distinct frequency bands, a mmWave carrier frequency f_c^{PA} for V2I mode and a lower carrier frequency f_c^{V2V} for V2V mode, which prevents inter-mode interference. The total system bandwidth is divided into K_{V2I} resource blocks (RBs) for V2I (both uplink and downlink) and K_{V2V} RBs for V2V communications. Each V2I or V2V link occupies at most one RB. Although the same RB may be shared by different links, interference will be created. Channel fading is assumed independent across all RBs, and channel coefficients are assumed constant within each time step Δt, which is also set as the fast fading update period for the V2V channels.

By analysing the interference level in the system for calculating the signal-to-interference-plus-noise-ratio (SINR), the achievable transmission rates for each communication mode are derived via the Shannon-Hartley theorem.

1) Downlink V2I Mode This mode ($\Phi_{l,t} = 2, \forall l$) functions as a system-wide operation requiring all LVs to suspend other communication activities and receive a common broadcast from the activated PAs. Consequently, this results in an interference-free environment where the wireless communication rate depends solely on the signal-to-noise-ratio (SNR):

$$\mathcal{C}^{\text{dl}}_{l,m,t} = W^{\text{V2I}}_{k} \log_2 \left(1 + \frac{p^{\text{dl}}_{l,m,t}}{\sigma^2} \right), \tag{11}$$

where W^{V2I}_{k} is the bandwidth of the k^{th} V2I RB and σ^2 is the noise power.

2) Uplink V2I Mode Two constraints ensure that the uplink V2I mode is also interference-free. First, this mode is disabled during downlink broadcasting, as $\Phi_{l,t}$ is uniformly set to 2 for all $l \in \{0, \ldots, L-1\}$. Second, the system schedules at most one LV for uplink transmission at any time step (i.e., if $\Phi_{l,t} = 1$, $\Phi_{l',t} \neq 1$ for all $l' \neq l$). Therefore, the achievable rate for this mode also depends solely on the SNR

$$\mathcal{C}^{\text{ul}}_{l,\langle m,n\rangle,t} = W^{\text{V2I}}_{k} \log_2 \left(1 + \frac{p^{\text{ul}}_{l,\langle m,n\rangle,t}}{\sigma^2} \right). \tag{12}$$

3) V2V Mode When $\Phi_{l,t} = 0$, LV_l communicates with its members using an assigned RB indexed by $\Upsilon_{l,t} \in \{1, \ldots, K_{\text{V2V}}\}$. In cases where no RB is assigned to LV_l, $\Upsilon_{l,t}$ will be set to 0. In contrast to the V2I links, multiple V2V transmissions may occur simultaneously, sharing the same RB, thereby inducing co-channel interference. Hence, the achievable rate for this mode is determined by the SINR

$$\mathcal{C}^{\text{V2V}}_{\langle l,v\rangle,t} = W^{\text{V2V}}_{k} \log_2 \left(1 + \frac{p^{\text{V2V}}_{\langle l,v\rangle,t}}{\sigma^2 + \sum\limits_{\substack{l' \neq l, \\ \Upsilon_{l',t} = \Upsilon_{l,t}}} p^{\text{V2V}}_{\langle l',v\rangle,t}} \right), \tag{13}$$

where W^{V2V}_{k} is the bandwidth of the k^{th} V2V RB and $I^{\text{V2V}}_{\langle l,v\rangle,t} = \sum_{l' \neq l, \Upsilon_{l',t} = \Upsilon_{l,t}} p^{\text{V2V}}_{\langle l',v\rangle,t}$ is the cumulative interference power received at $\text{MV}_{\langle l,v\rangle}$ from other LVs (l') sharing the same RB. The power $p^{\text{V2V}}_{\langle l',v\rangle,t}$ is calculated as

$$p^{\text{V2V}}_{\langle l',v\rangle,t} = p^{\text{LV}}_{l',t} \left| h^{\text{V2V}}_{\text{LV}_{l'}\text{-MV}_{\langle l,v\rangle}} \right|^2, \tag{14}$$

where $p^{\text{LV}}_{l',t}$ is the power used by $\text{LV}_{l'}$, and $h^{\text{V2V}}_{\text{LV}_{l'}\text{-MV}_{\langle l,v\rangle}}$ represents the channel coefficient between the interfering $\text{LV}_{l'}$ and the receiving $\text{MV}_{\langle l,v\rangle}$.

Given that V2I links are modelled as LoS channels and singular in activation at time t ("segment selection" design in Section II-C), we do not model a specific RB allocation for them. The objective at each time step t is to jointly determine the RB allocation $\Upsilon_{l,t}$ (for V2V only) and communication mode $\Phi_{l,t}$. This decision is adapted to current network conditions and communication requirements, with the primary goal of maximising data throughput while ensuring communication reliability and timeliness.

F. Age of Information Model

In contrast to prior studies [11]–[15] that consider only uplink freshness, the system maintains traffic awareness via two AoI metrics, corresponding to the uplink and downlink V2I communication modes. Facilitated by these two V2I modes, the LVs transmit local platoon-state information to the BS (uplink platoon monitoring) and receive aggregated data regarding other VPs from the BS (downlink traffic alerts).

The uplink AoI, denoted as $A^{\text{ul}}_{l,t}$, quantifies the individual information freshness of VP_l as perceived by the BS. Consistent with the discrete-time models in [15], it increments by Δt at each step and resets upon a successful uplink transmission

$$A^{\text{ul}}_{l,t} = \begin{cases} A^{\text{ul}}_{l,t-1} + \Delta t, & \text{if } \mathcal{C}^{\text{ul}}_{l,\langle m,n\rangle,t} < \mathcal{C}^{\text{ul}}_{\text{min}}, \\ & \text{or } \Phi_{l,t} \neq 1 \\ \Delta t, & \text{otherwise}, \end{cases} \tag{15}$$

where $\mathcal{C}^{\text{ul}}_{min}$ is the minimum required uplink V2I rate.

The downlink AoI, denoted as A^{dl}_{t}, measures the freshness of global traffic data broadcast to the platoons. It is modelled analogously to the uplink AoI

$$A^{\text{dl}}_{t} = \begin{cases} A^{\text{dl}}_{t-1} + \Delta t, & \text{if } \mathcal{C}^{\text{dl}}_{l,m,t} < \mathcal{C}^{\text{dl}}_{\text{min}}, \ \forall l, \\ & \text{or } \Phi_{l,t} \neq 2 \\ \Delta t, & \text{otherwise}, \end{cases} \tag{16}$$

where $\mathcal{C}^{\text{dl}}_{\text{min}}$ is the minimum required downlink V2I rate.

These two formulations depend on their achievable communication rates, $\mathcal{C}^{\text{ul}}_{l,\langle m,n\rangle,t}$ and $\mathcal{C}^{\text{dl}}_{l,m,t}$, and the mode selection indicator $\Phi_{l,t}$. The AoI values increase by Δt if the communication conditions are insufficient (i.e., $\mathcal{C}^{\text{ul}}_{l,\langle m,n\rangle,t} < \mathcal{C}^{\text{ul}}_{min}$ or $\mathcal{C}^{\text{dl}}_{l,m,t} < \mathcal{C}^{\text{dl}}_{\text{min}}$) or if the corresponding mode is inactive (i.e., $\Phi_{l,t} \neq 1$ or 2). Conversely, a successful transmission resets the values to Δt. In this formulation, we adopt a quasi-static block fading assumption with ideal channel coding. Consequently, we assume that the packet error rate is negligible provided that the instantaneous channel capacity exceeds the target rate threshold (i.e., $\mathcal{C}^{\text{ul}}_{l,\langle m,n\rangle,t} \geq \mathcal{C}^{\text{ul}}_{\text{min}}$ or $\mathcal{C}^{\text{dl}}_{l,m,t} \geq \mathcal{C}^{\text{dl}}_{\text{min}}$). Thus, a successful update is deterministically recorded whenever the rate constraint is satisfied. Collectively, these two AoI metrics ensure both the BS and VPs operate with up-to-date traffic information, thereby enhancing the reliability of the vehicular network.

G. Cooperative Awareness Message Model

To maintain inter-vehicle spacing and string stability, each LV periodically broadcasts CAMs to its MVs. The CAM requirement is defined by a payload Θ to be delivered within an interval of T time steps. Aligned with standard specifications [16], [42] and prior studies [11]–[15], this interval is set within the range $[100, 1000]$ ms. Here, T represents the number of discrete time steps (Δt) per update cycle, which corresponds to the large-scale fading update period (i.e., $T \cdot \Delta t$).

A successful CAM delivery requires all MVs to receive the full payload within the designated time window. The performance is thus constrained by the MV experiencing the lowest cumulative data rate, which acts as the bottleneck link. The success event CAM_l for VP_l is formulated as

$$\mathrm{CAM}_l = \left\{ \min_{v \in \{1,\ldots,V_l-1\}} \left(\sum_{t=1}^{T} \mathcal{C}^{\mathrm{V2V}}_{\langle l,v\rangle,t} \Delta t \right) \geq \Theta \right\}, \tag{17}$$

where the inequality ensures that the cumulative throughput of the weakest link satisfies Θ over T. To incorporate this reliability metric into the optimisation framework, we define the CDP [14] as

$$\mathrm{CDP}_l = \Pr\left(\mathrm{CAM}_l\right). \tag{18}$$

CDP_l serves as a key objective to be maximised, which represents the likelihood of a successful CAM exchange under the current channel conditions and RA for VP_l.

H. Problem Formulation

This section formalises the joint RA and mode selection problem for the V2X system.

Prior to dynamic optimisation, the physical configuration of the PA system, specifically the segment selection and PA activation strategies, is established first. Crucially, the selection of the active segment m^* and the specific PA n^* is treated as a deterministic geometric pre-processing step rather than a dynamic action within the RL framework. Specifically, the system identifies the optimal hardware coordinates based on the real-time spatial distribution of the LVs to minimise propagation loss. By decoupling this geometric PA activation from the highly dynamic RA we avoid an explosion in the action space dimensionality and allow the RL agent to focus entirely on mitigating the highly stochastic elements of the environment (e.g., fading and interference) to optimise communication timeliness and reliability.

1) Uplink PA Configuration As detailed in Section II-C and [22], uplink transmission utilises a single PA within one segment. To minimise free-space path loss, the optimal PA is the one that is spatially closest to the transmitting LV. As proved in [22], [43], the optimal coordinate approximates the vehicle's x-axis position

$$X_{\langle m^*,n^*\rangle} \approx X_l, \tag{19}$$

where m^* and n^* denote the indices of the selected segment and activated PA, respectively. In practice, the exact real-time coordinate X_l at time t is estimated using the most recently reported location and the vehicle's velocity, i.e., $X_{l,t} \approx X_{l,t-1} + v_l \cdot \Delta t$. Given the millisecond-scale update frequency in our system (as detailed in Section V), the spatial drift between updates is practically negligible. Thus, assuming accurate geometry for the PA selection does not compromise the validity of the physical model.

2) Downlink PA Configuration For the downlink broadcasting model established in Section II-C, an optimal segment m^* and a set of N_{m^*} PAs are activated. To ensure equitable coverage for all LVs, m^* is selected based on the centroid of the LV distribution

$$m^* = \left\lfloor \frac{1}{D_{road}/M} \left(\frac{1}{L} \sum_{l=0}^{L-1} X_l \right) \right\rfloor = \left\lfloor \frac{M \sum_l X_l}{L D_{road}} \right\rfloor. \tag{20}$$

Within this segment, the primary PA (n^*) is positioned at the centroid $X_{\langle m^*,n^*\rangle} = \frac{1}{L}\sum_l X_l$. Because these activated PAs broadcast identical global traffic data to every platoon with equitable coverage, the downlink AoI in (16) is defined as a single system-wide metric rather than per platoon, furnishing all platoons with the consistent and equally fresh global view that is indispensable for safety-critical coordination. Under the strong LoS coverage of these activated PAs and the quasi-static fading assumption, the broadcast reaches every platoon with near-certain success, so a per-platoon downlink AoI would coincide with the common value and is therefore unnecessary.

The remaining $N_{m^*} - 1$ PAs are arranged symmetrically extending outwards from n^* with the spacing $D_{\text{PA-PA}}$ detailed in (1). The position of the n-th ($n \in \{0,\ldots,N_{m^*}-1\}$) activated PA is given by

$$X_{\langle m^*,n\rangle} = X_{\langle m^*,n^*\rangle} + (n - n^*) D_{\text{PA-PA}}, \tag{21}$$

subject to the segment boundaries $X_{m^*} \leq X_{\langle m^*,n\rangle} < X_{m^*} + D_{road}/M$. Crucially, this boundary constraint is applied independently in each direction. If the placement reaches the limit on one side, activation terminates solely on that side, while continuing on the opposite side until its respective limit is reached.

3) Global Optimisation Problem Building upon the above PA configurations and the discussion in the previous sections, we formulate the optimisation problem for VP_l. The goal is to minimise and balance four competing objectives: the uplink AoI, the downlink AoI, the LV power consumption over the time interval T, and the CDP. The problem is written as follows:

$$\min_{\Phi,\Upsilon,p} \left\{ \frac{1}{T} \sum_{t=1}^{T} \left(A^{\mathrm{ul}}_{l,t}, A^{\mathrm{dl}}_{t}, p^{\mathrm{LV}}_{l,t} \right), -\mathrm{CDP}_l \right\}, \tag{22}$$

$$\text{s.t.} \quad \Phi_{l,t} \in \{0,1,2\}, \forall l,t, \tag{22a}$$

$$\Upsilon_{l,t} \in \{0,1,\ldots,K_{\mathrm{V2V}}-1\}, \forall l,t, \tag{22b}$$

$$0 \leq p^{\mathrm{LV}}_{l,t} \leq p^{\max}, \forall l,t, \tag{22c}$$

noting that minimising the negative value of CDP_l is mathematically equivalent to maximising it. Constraints (22a) and (22b) define the discrete action spaces for mode selection and RB allocation, while (22c) ensures the transmit power remains within the operational limits.

III. Proposed DE-MADDPG-based MADRL Framework

The multi-objective optimisation problem formulated in (22) involves high-dimensional continuous and discrete action spaces, which are intractable for traditional optimisation methods in real-time applications. Consequently, we reformulate the problem as a MADRL task. Given that each LV makes decisions based on local observations while the environment state evolves globally, the problem is modelled as a partially observable Markov decision process (POMDP). The POMDP is defined by the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$, representing the state space, action space, transition probability, reward function, and discount factor, respectively.

Each LV functions as an independent agent. At each time step t, agent l executes three actions from the action space $\mathcal{A}$, defined as

$$\mathcal{A}_{l,t} = \left[\Phi_{l,t}, \Upsilon_{l,t}, p^{\mathrm{LV}}_{l,t} \right]. \tag{23}$$

These components correspond to the communication mode selection, subchannel allocation, and power control, respectively. Since the original action variables, including the transmit power which is discretised into integer dBm levels in our implementation, are incompatible with the continuous nature of the DDPG/TD3 algorithms, we employ a continuous relaxation approach. Specifically, the action variables are represented by continuous surrogate outputs of the actor network bounded in $[-1, 1]$. Before interacting with the environment, these continuous surrogate outputs are mapped to feasible discrete decisions through a deterministic discretisation operator (i.e., uniform linear scaling followed by rounding). Crucially, the critic networks evaluate the original continuous surrogate outputs directly, preserving the differentiability required for policy gradient updates. Simultaneously, the agent observes the environment state $\mathcal{S}_{l,t}$, which encapsulates local CSI, interference levels, and network status

$$\mathcal{S}_{l,t} = \begin{bmatrix} h^{\text{iw-fs}}_{\text{BS-PA}_{\langle m,n\rangle}\text{-LV}_l}, h^{\text{V2V}}_{\text{LV}_l\text{-MV}_{\langle l,v\rangle}}, I^{\text{V2V}}_{\langle l,v\rangle,t} \\ \mathbf{P}_l, X_{\langle m^*,n^*\rangle}, A^{\text{ul}}_{l,t}, A^{\text{dl}}_t, \Theta', T', \sigma^2 \end{bmatrix}. \tag{24}$$

Here, LV_l tracks the V2I and V2V channel conditions, V2V interference, its own position relative to the active PA, instantaneous uplink and downlink AoI values, the remaining CAM payload Θ' and time budget T', and the noise power σ^2.

In practice, this state incurs only a modest overhead, as most of its elements are maintained locally by each LV: the uplink and downlink AoI, the residual payload Θ', and the time budget T' are internal counters, the position $\mathbf{P}_l$ is provided by onboard localisation, the active PA coordinate follows from the deterministic geometry of Section II-H, and the noise power σ^2 is a known constant. Only the channel and interference terms require measurement, which our model assumes to be available to the agent. Crucially, the decentralised execution requires no real-time exchange of other agents' states, so no additional inter-agent signalling is introduced, and any update is captured by the AoI metric itself.

To address the challenges of stability and scalability in vehicular networks, this work employs the decomposed multi-agent DDPG (DE-MADDPG) algorithm [44]. This approach follows the centralised training with decentralised execution (CTDE) framework, where agents are trained using global information but execute actions based solely on local observations. Standard multi-agent DDPG (MADDPG) algorithms suffer from scalability issues as the input dimension of the centralised critic grows linearly with the number of agents. DE-MADDPG mitigates this by decomposing the reward function into local and global components, allowing the global critic to focus on system-wide coordination while local critics optimise individual objectives in the V2X network. Furthermore, this base framework is augmented with the twin delayed deep deterministic policy gradient (TD3) technique [45]. The integration of TD3 addresses the Q-function overestimation bias inherent in standard DDPG methods, thereby enhancing convergence stability in the highly dynamic highway environment.

The DE-MADDPG algorithm, the CTDE paradigm, and the TD3 critics are adopted from the established literature [11]–[15], while the fundamental novelty of this work lies in the physics-aware reward decomposition detailed below, which maps the global and local reward components onto the SWAN operational modes, namely the system-wide downlink broadcasting through segment activation and the point-to-point uplink through a specific PA, so as to jointly optimise the bidirectional AoI. The learning architecture thus serves as a vehicle for this physical-layer-aware reward structure rather than as a general-purpose algorithmic advance.

Algorithm 1: DE-MADDPG (TD3)

Initialise high-speed road environment & replay buffer $\mathcal{B}$.
Initialise global critic networks: $\{Q^{\text{G}}_{\psi_d}, Q^{\text{G}}_{\psi'_d}\}, d = 1, 2.$
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: $\{T', \Theta'_l\} = \{T, \Theta_l\}$.
  for $t = 1$ to T **do**
    for agent 1 to L **do**
      Observe state $s_{l,t}$, select action $a_{l,t} = \pi_l(s_{l,t})$, receive local & global rewards: $\{r_{l,t}, r^{\text{G}}_t\}$.
    Update channel fast fading & interference.
    Each agent l observes a new state s'_l.
    Store $\left(\mathbf{s}_t, \mathbf{a}_t, \mathbf{r}_t, r^{\text{G}}_t, \mathbf{s}'\right)$ into $\mathcal{B}$.
    Sample mini-batch of B transitions $\left(\mathbf{s}_b, \mathbf{a}_b, \mathbf{r}_b, r^{\text{G}}_b, \mathbf{s}'_b\right)\big|_{b=1}^{B}$ from $\mathcal{B}$.
    Update global critics: minimising $\mathcal{L}^{\text{G}}(\psi_d)$ (29) by one-step gradient descent.
    Target soft update: $\psi'_d \leftarrow \tau\psi_d + (1-\tau)\psi'_d$.
    if t mod d_{TD3} **then**
      for agent 1 to L **do**
        Update local critic: minimising $\mathcal{L}^{\text{L}}(\phi_l)$ (30) by one-step gradient descent.
        Update local actor: maximising $\nabla_{\theta_l}\mathcal{J}$ as in (28) by one-step gradient ascent.
        Target soft update: $\phi'_l \leftarrow \tau\phi_l + (1-\tau)\phi'_l$
        $\theta'_l \leftarrow \tau\theta_l + (1-\tau)\theta'_l$

The reward function structure is designed to align individual agent incentives with the system-wide objectives defined in (22). The global and local reward functions are formulated as follows:

$$\mathcal{R}^{\text{G}}_t = -\frac{1}{L}\sum_l \frac{\sum_v \mathcal{F}_1\left(I^{\text{V2V}}_{\langle l,v\rangle,t}\right)}{V_l} - \mathcal{F}_2\left(A^{\text{dl}}_t\right), \tag{25}$$

$$\begin{aligned}\mathcal{R}^{\text{L}}_{l,t} = {} & w_1 \cdot \mathbf{1}_{\{\mathcal{C}^{\text{ul}}_{l,\langle m,n\rangle,t} \geq \mathcal{C}^{\text{ul}}_{\text{min}}\}} + w_2 \cdot \mathbf{1}_{\{\mathcal{C}^{\text{dl}}_{l,m,t} \geq \mathcal{C}^{\text{dl}}_{\text{min}}\}} \\ & - \mathcal{F}_3\left(A^{\text{ul}}_{l,t}\right) - \mathcal{F}_4\left(p^{\text{LV}}_{l,t}\right) - \mathcal{F}_5\left(\frac{\Theta'}{\Theta}, \frac{T'}{T}\right).\end{aligned} \tag{26}$$

The global reward $\mathcal{R}^{\text{G}}_t$ penalises the average system interference and the downlink AoI. By minimising the cumulative interference, agents are encouraged to select channels that minimise cross-tier disturbance. Simultaneously, the downlink AoI term ensures the timely dissemination of global traffic data from the PA system. Regarding the local reward $\mathcal{R}^{\text{L}}_{l,t}$, positive gains w_1 and w_2 are awarded upon successful uplink or downlink V2I updates, respectively, indicated by the indicator function $\mathbf{1}_{\{\cdot\}}$. The remaining terms penalise high uplink AoI, excessive power consumption, and failure to deliver the CAM payload. The mapping functions $\mathcal{F}_1$–$\mathcal{F}_5$ scale these distinct and competing physical quantities to a comparable range and weight their relative contributions, defined as

$$\begin{aligned}
\mathcal{F}_1(I) &= (I-\kappa_{I_1})/\kappa_{I_2},\\
\mathcal{F}_2(A_t^{\mathrm{dl}}) &= A_t^{\mathrm{dl}}/\kappa_g,\\
\mathcal{F}_3(A_{l,t}^{\mathrm{ul}}) &= A_{l,t}^{\mathrm{ul}}/\kappa_l,\\
\mathcal{F}_4(p_{l,t}^{\mathrm{LV}}) &= \omega_p \ln p_{l,t}^{\mathrm{LV}},\\
\mathcal{F}_5(\Theta'/\Theta, T'/T) &= \omega_c\,(\Theta'/\Theta)(2-T'/T),
\end{aligned} \tag{27}$$

where κ_{I_1} and κ_{I_2} offset and scale the interference, κ_g and κ_l normalise the downlink and uplink AoI, and ω_p and ω_c weight the power and payload penalties, all empirically selected to balance the competing objectives. In $\mathcal{F}_5$, the factor $(2-T'/T)$ raises the payload penalty as the delivery deadline approaches, equalling $\omega_c\,\Theta'/\Theta$ at the start of the interval and doubling as the time budget is exhausted.

Together, (25) and (26) realise a weighted-sum scalarisation of the multi-objective problem in (22). Because these objectives carry heterogeneous units, the mapping functions $\mathcal{F}_2$ and $\mathcal{F}_3$ first normalise the downlink and uplink AoI in milliseconds to a comparable magnitude through κ_g and κ_l, and $\mathcal{F}_4$ weights the transmit power in dBm through ω_p. The indicator gains w_1 and w_2, which reward successful uplink and downlink delivery. And $\mathcal{F}_5$ penalises the residual undelivered payload through ω_c, where the CDP objective enters. The remaining mapping function $\mathcal{F}_1$ is not a separate objective but penalises the V2V co-channel interference induced by the resource-block allocation in (22b), thereby guiding the agent toward cleaner channels and a higher V2V rate. The relative magnitudes of these weights govern the trade-off among the objectives: a larger ω_p, for instance, suppresses transmit power at the expense of information freshness. In this safety-critical setting, the weights are specially chosen so that message timeliness and delivery reliability take precedence over power economy, whereby maximising the resulting scalar reward is equivalent to minimising the weighted objective vector in (22).

Crucially, this bipartite reward structure is intrinsic to the physical layer of the PA system: the global component targets the segment-level broadcasting performance (downlink AoI), while the local component governs the vehicle-specific PA transmission (uplink AoI), thereby embedding the unique waveguide constraints directly into the learning process.

Based on these definitions, the proposed DE-MADDPG algorithm with TD3 (DE-MADDPG (TD3)) updates the actor and critic networks using the following gradient formulation

$$\begin{aligned}
\underbrace{\nabla_{\theta_l}\mathcal{J}}_{\text{Local actor}} &= \underbrace{\mathbb{E}_{\mathbf{s},\mathbf{a}\sim\mathcal{B}}\Big[\nabla_{\theta_l}\pi_l(a_l|s_l)\,\nabla_{a_l}Q^{\mathrm{G}}_{\psi_1}(\mathbf{s},\mathbf{a})\Big]}_{\text{MADDPG: Global critic}}\\
&\quad+\underbrace{\mathbb{E}_{s_l,a_l\sim\mathcal{B}}\Big[\nabla_{\theta_l}\pi_l(a_l|s_l)\,\nabla_{a_l}Q^{\mathrm{L}}_{\phi_l}(s_l,a_l)\Big]}_{\text{DDPG: Local critic}},
\end{aligned} \tag{28}$$

where $\nabla_{\theta_l}\mathcal{J}$ represents the policy gradient for the local actor at time t. Here, $\mathbf{s}=(s_1,...,s_L)$ and $\mathbf{a}=(a_1,...,a_L)$ denote the state and action sets, and $\mathcal{B}$ is the experience replay buffer. Consistent with the TD3 framework, the first global Q-function $Q^{\mathrm{G}}_{\psi_1}$ from the twin critic pair is selected to compute the gradient. The parameters θ_l, ψ_1, and ϕ_l parameterise the agent policy π_l, global critic $Q^{\mathrm{G}}_{\psi_1}$, and local critic $Q^{\mathrm{L}}_{\phi_l}$, respectively. The loss functions and target updates for the global and local critics are computed as

$$\begin{aligned}
&\mathcal{L}^{\mathrm{G}}(\psi_d) = \mathbb{E}_{\mathbf{s},\mathbf{a},\mathbf{r}^{\mathrm{G}},\mathbf{s}'}\left[\left(Q^{\mathrm{G}}_{\psi_d}(\mathbf{s},\mathbf{a})-tar^{\mathrm{G}}\right)^2\right], d=1,2,\\
&\quad tar^{\mathrm{G}} = r^{\mathrm{G}}+\gamma\min_d Q^{\mathrm{G}}_{\psi'_d}(\mathbf{s}',\mathbf{a}')\Big|_{a'_l=\pi'_l(s'_l)},
\end{aligned} \tag{29}$$

$$\begin{aligned}
&\mathcal{L}^{\mathrm{L}}(\phi_l) = \mathbb{E}_{s_l,a_l,r_l^{\mathrm{L}},s'_l}\left[\left(Q^{\mathrm{L}}_{\phi_l}(s_l,a_l)-tar_l^{\mathrm{L}}\right)^2\right],\\
&\quad tar_l^{\mathrm{L}} = r_l^{\mathrm{L}}+\gamma Q^{\mathrm{L}}_{\phi'_l}(s'_l,a'_l)\Big|_{a'_l=\pi'_l(s'_l)},
\end{aligned} \tag{30}$$

where $\mathbf{s}'$ and $\mathbf{a}'$ denote the next state and action sets. $Q^{\mathrm{G}}_{\psi'_d}$, $Q^{\mathrm{L}}_{\phi'_l}$, and π'_l represent the target global critic, local critic, and policy, respectively. The Q-function estimates the expected discounted return starting from time t

$$Q^{\pi_l}(s,a)=\mathbb{E}_{\pi_l}\left[\sum_{i=0}^{\infty}\gamma^i\mathcal{R}^{t+i+1}\Big|s_{l,t},a_{l,t}\right]. \tag{31}$$

The complete execution flow of the proposed DE-MADDPG (TD3) approach is detailed in Algorithm 1.

IV. Complexity Analysis

This section evaluates the computational complexity of the proposed and baseline algorithms, which serves as a basis for performance evaluation. The algorithms evaluated include

- *DE-MADDPG (TD3)*: The proposed decomposed multi-agent algorithm with twin delayed critics, as specified in Section III.
- *Standard MADDPG*: A CTDE framework [46].
- *Decentralised MADDPG (Dec-MADDPG)*: A fully decentralised approach where all agents function independently, with both actors and critics relying solely on local observations.
- *DDPG*: A single-agent RL approach where a central controller manages all platoons using one actor and one critic network [47].
- *Multi-agent proximal policy optimisation (MAPPO)*: An on-policy baseline from the policy-optimisation family [48]. It retains the CTDE structure with decentralised actors and a centralised value function, but, in contrast to the DDPG-family baselines above, it is on-policy and employs neither an experience replay buffer nor target or twin critic networks.

Note that since the uplink and downlink PA configurations in Section II-H 1) and 2) are independent of the MADRL framework, and the process is identical for all the algorithms, it is not considered in the complexity comparison among the above algorithms. Moreover, as MAPPO is on-policy, it lies outside the off-policy actor-critic structure examined in this section. It is therefore omitted from the following complexity analysis and assessed only in the numerical comparison of Section V.

A. Computational Complexity Analysis

The computational complexity is primarily determined by the number of agents (L) and the architecture of the neural networks (NNs) employed. We define $\mathcal{J}^a$ and $\mathcal{J}^c$ as the number of layers in the actor and critic networks, respectively.

TABLE III: Number of NNs and Trainable Parameters.

Algorithm	NNs	Trainable Parameters
DE-MADDPG (TD3)	$2\left(\mathbf{2}^{\mathrm{G},c}+L\left(\mathbf{1}^{\mathrm{L},a}+\mathbf{1}^{\mathrm{L},c}\right)\right)$	$\mathrm{O}\left((L+1)\chi\right)$
Standard MADDPG	$2L\left(\mathbf{1}^{\mathrm{L},a}+\mathbf{1}^{\mathrm{G},c}\right)$	$\mathrm{O}\left(L^2\chi\right)$
Dec-MADDPG	$2L\left(\mathbf{1}^{\mathrm{L},a}+\mathbf{1}^{\mathrm{L},c}\right)$	$\mathrm{O}\left(L\chi\right)$
DDPG	$2\left(\mathbf{1}^{\mathrm{sys},a}+\mathbf{1}^{\mathrm{sys},c}\right)$	$\mathrm{O}\left(L\chi\right)$

Let U_j^a and U_j^c denote the number of neurons in the j^{th} layer of the corresponding networks. The general floating-point operations (FLOPs) required for a single forward pass in an actor network (C^a) and a critic network (C^c) can be expressed as

$$C^a = \mathrm{O}\left(\sum_{j=2}^{\mathcal{J}^a-1}\left(U_{j-1}^a U_j^a + U_j^a U_{j+1}^a\right)\right), \tag{32}$$

$$C^c = \mathrm{O}\left(\sum_{j=2}^{\mathcal{J}^c-1}\left(U_{j-1}^c U_j^c + U_j^c U_{j+1}^c\right)\right). \tag{33}$$

Based on these definitions, the total training complexity for the included algorithms over the complete training duration is analysed below

- *DE-MADDPG (TD3)*:
$$\mathrm{O}\left(ep\cdot T\cdot B\left(C^{\mathrm{G},c}+L\left(C^{\mathrm{L},a}+C^{\mathrm{L},c}\right)\right)\right),$$
- *Standard MADDPG*:
$$\mathrm{O}\left(ep\cdot T\cdot B\cdot L\left(C^{\mathrm{L},a}+C^{\mathrm{G},c}\right)\right),$$
- *Dec-MADDPG*:
$$\mathrm{O}\left(ep\cdot T\cdot B\cdot L\left(C^{\mathrm{L},a}+C^{\mathrm{L},c}\right)\right),$$
- *DDPG*:
$$\mathrm{O}\left(ep\cdot T\cdot B\left(C^{\mathrm{sys},a}+C^{\mathrm{sys},c}\right)\right).$$

Here, ep represents the total number of training episodes, T denotes the steps per episode, and B indicates the minibatch size used for gradient updates. $C^{\mathrm{G},c}$ in DE-MADDPG denotes the complexity of the global system critic, while in Standard MADDPG it refers to the critic that takes the full system state-action pair as input but is replicated L times (one per agent). $C^{\mathrm{L},a}$ and $C^{\mathrm{L},c}$ refer to the local actor and local critic, respectively. $C^{\mathrm{sys},a}$ and $C^{\mathrm{sys,c}}$ denote the system-wise actor and critic network for DDPG algorithm.

The DDPG exhibits the lowest computational complexity as it aggregates the system into a single agent entity. However, this comes at the cost of a strictly centralised execution requirement. Dec-MADDPG scales linearly with L but suffers from non-stationarity due to the lack of global information. The Standard MADDPG has the same magnitude of computational complexity as the Dec-MADDPG, where the critic network introduces the difference. The proposed DE-MADDPG (TD3) maintains a complexity similar to Dec-MADDPG for its local components, with the additional set of global critics, offering a balance between scalability and coordination. The MAPPO baseline is excluded from this comparison because its on-policy training maintains neither target networks nor a replay buffer from which minibatches are repeatedly sampled, yielding a computational profile that does not map onto the off-policy expressions above.

TABLE IV: Simulation and Training Parameters.

PA System Parameters	Symbols	Values
PA Carrier frequency	f_c^{PA}	28 GHz
Waveguide height	D_{PA}	5 m
Maximum transmit power	p^{PA}	10 dBm
Refractive index	n_{eff}	1.4
In-waveguide attenuation	α^{iw}	0.0092 m^{-1} [28], [29]
Number of segments	M	9
Vehicle Parameters	**Symbols**	**Values**
Number of platoons	L	4-7
Vehicles per platoon	V_l	4
Intra-platoon gap	D_{gap}	25 m
Max vehicle power	$p^{\max}$	30 dBm
V2V Carrier frequency	f_c^{V2V}	2 GHz
Number of V2V RBs	K_{V2V}	3
V2V RB bandwidth	W^{V2V}	180 kHz
V2V path loss model	-	B1 - Urban micro-cell [40]
V2V Shadowing model	-	Log-normal, $\sigma = 3$ dB
V2V Decorrelation distance	-	10 m
CAM message size	Θ	4 KB
CAM delivery interval	$T\cdot\Delta t$	100 ms [16], [42]
Other Settings	**Symbols**	**Values**
Road length	D_{road}	200 m
Road width	D_{width}	15 m
Noise power	σ^2	-90 dBm
Small-scale fading	-	Rayleigh fading
Large-scale fading period	$T\cdot\Delta t$	100 ms [42]
Fast fading update period	Δt	1 ms [42]
Training Parameters	**Symbols**	**Values**
Time steps per episode	T	100
Total episodes	$loop$	500
Discount factor	γ	0.99
Batch size	B	64
Actor learning rate	-	0.0001
Critic learning rate	-	0.001
Target soft update	τ	0.005
Optimiser	-	Adam

B. Network Architecture and Scalability

Table III compares the required number of NNs and the order of trainable parameters. We define χ, which is the summation of the state space dimension and action space dimension for a single agent. In the table notation, $\mathbf{1}$ denotes a single network unit, and specifically for DE-MADDPG (TD3), $\mathbf{2}^{\mathrm{G},c}$ represents the employed twin global critic networks. The factor 2 accounts for the existence of target networks for each main network.

A critical observation is the scalability regarding the number of platoons L. In Standard MADDPG, each of the L agents requires a critic that observes the joint actions and states of all other agents. Consequently, the input dimension for each critic grows with L, and since there are L critics, the total number of parameters scales as $\mathrm{O}(L^2\chi)$. In contrast, the proposed DE-MADDPG (TD3) decouples the objectives. The local critics only observe local states ($\mathrm{O}(\chi)$), and only the single shared global critic observes the joint state. Thus, the parameter complexity scales linearly as $\mathrm{O}((L+1)\chi)$. This reduction from quadratic to linear complexity renders the proposed approach significantly more scalable for dense traffic scenarios on high-speed roads.

Finally, we address the distinction between training complexity and online execution latency. While the training phase involves complex critic updates, the online decision-making at

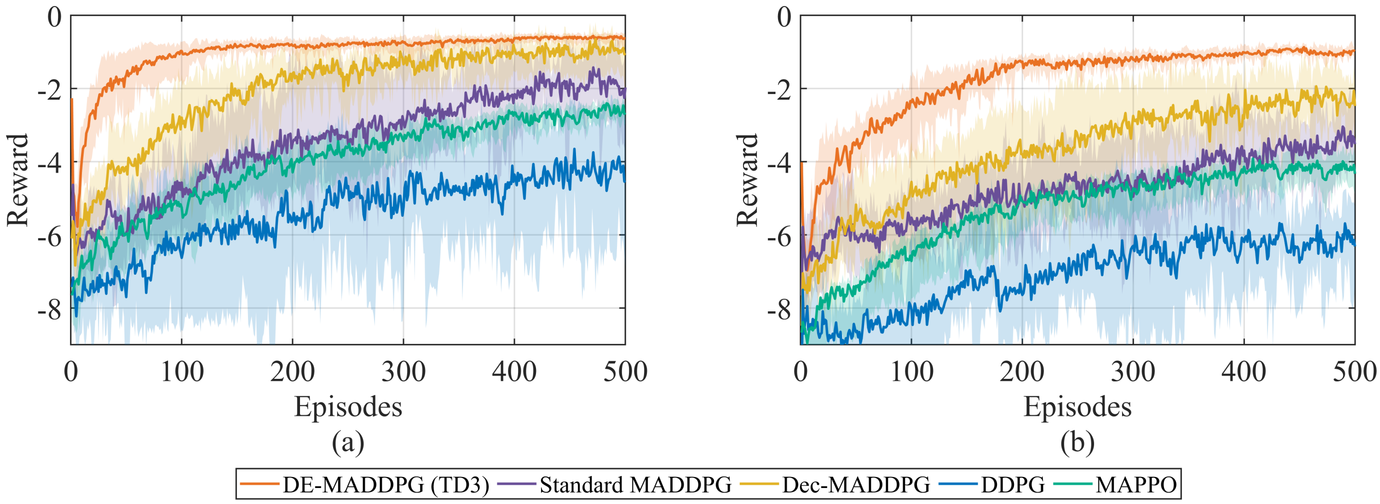


Fig. 2: Reward convergence comparison of the five included algorithms. The results are averaged over 10 independent simulation runs, and the shaded areas represent the min-max range. (a) $L = 5$, $V_l = 4$. (b) $L = 7$, $V_l = 4$.

each time step t depends solely on the local actor network: $a_{l,t} = \pi_l(s_{l,t})$. This operation requires a single forward pass through a relatively small NN. Given the network dimensions utilised in our experiments (detailed in Section V), this inference is computationally lightweight (on the order of microseconds) and can be executed well within the tight latency constraints of the vehicular network in practice (e.g., typically sub-millisecond).

V. Numerical Results

This section assesses the performance of the proposed DE-MADDPG (TD3) framework against the baseline algorithms established in Section IV. The simulation environment integrates a PA-supported infrastructure operating at 28 GHz (as defined in Section II-C) with a single-cell urban C-V2X network operating at 2 GHz. The V2V communications utilise 3 dedicated RBs, adhering to the 3GPP TR 36.885 standard [42]. The detailed simulation parameters are listed in Table IV. Beyond comparing RA algorithms within the SWAN framework, the proposed framework is also evaluated on a non-segmented conventional PA system (PASS) in Fig. 4 to quantify the benefit of waveguide segmentation. A comparison with conventional cellular V2I is not included, as its physical-layer disadvantage relative to PA systems arises from a fundamentally different propagation model, a contrast already established in the pinching-antenna literature [22].

The framework is implemented in Python using PyTorch and executed on NVIDIA RTX A6000 GPU. The NN architecture comprises fully connected layers: the local actor and critic networks contain two hidden layers with $(1024, 512)$ and $(512, 256)$ neurons, respectively, while the global critic employs three hidden layers with $(1024, 512, 256)$ neurons to handle the larger system state input (and action input for the critics). We utilise the rectified linear unit (ReLU) [49] activation function and the Adam optimiser [50]. 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 other training parameters are also listed in Table IV.

A. Convergence Analysis

Fig. 2 illustrates the convergence of the local reward function (26) for the five algorithms under two traffic density scenarios: (a) $L = 5$ and (b) $L = 7$, with $V_l = 4$ constant. The solid lines denote the mean reward across 10 independent runs, while the shaded areas indicate the fluctuation (the range between minimum and maximum values) observed across these multiple simulation runs. The proposed DE-MADDPG (TD3) algorithm achieves the fastest convergence speed and the highest stable reward in both scenarios.

The failure of DDPG to reach a high-level reward is attributed to its centralised architecture, which suffers from inefficient exploration and neglect of individual platoon requirements in multi-agent environments. Standard MADDPG follows with significant instability. As L increases from 5 to 7, its performance gap relative to the proposed method widens, highlighting its limited scalability in high-dimensional state spaces. Interestingly, Dec-MADDPG outperforms Standard MADDPG, suggesting that in complex vehicular environments, simplified local observations can sometimes yield better short-term policies than a poorly converged global critic. It even performs better than our proposed DE-MADDPG (TD3) algorithm in certain episodes, which can be observed in the shaded area in Fig. 2(a), Episode 400-500. However, Dec-MADDPG exhibits high non-stationarity due to the lack of explicit coordination via global information, leading to worse averaged performance than our proposed algorithm. In contrast, DE-MADDPG (TD3) combines the benefits of both: the global critic maintains coordination and stability, and the local critics ensure the individual requests of all the platoons, resulting in robust and high-value convergence. The on-policy MAPPO baseline converges smoothly but settles below the off-policy MADDPG variants, remaining ahead of only the single-agent DDPG. Under the identical episode budget, its on-policy updates extract less information per environment

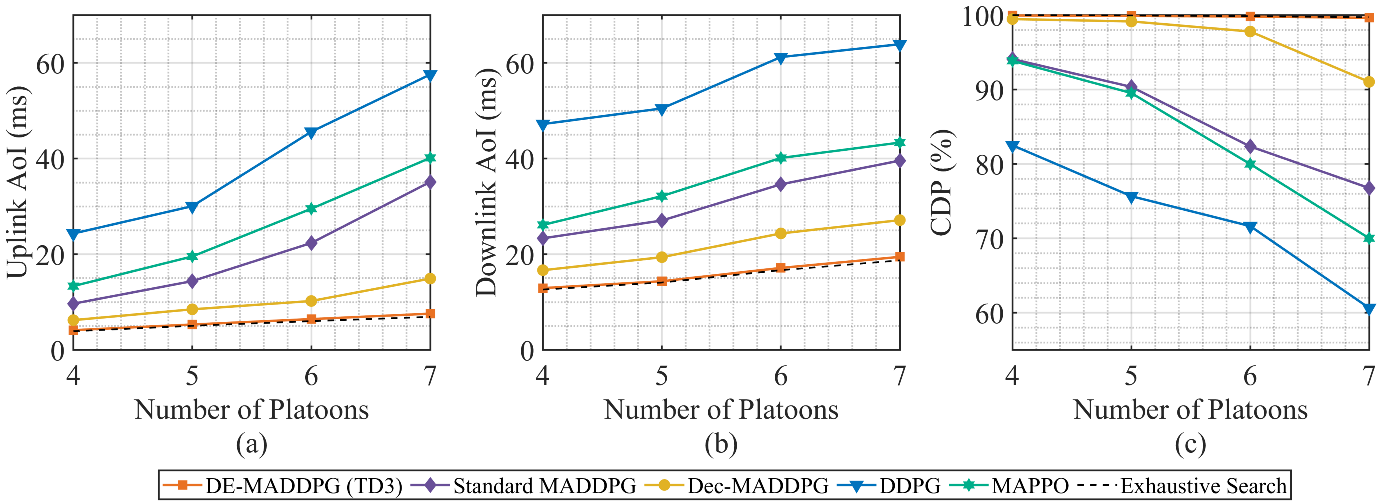


Fig. 3: Performance comparison with L varies from 4 to 7 and fixed $V_l = 4$. Results are from the last 100 episodes and are averaged over 10 independent simulation runs. (a) Average uplink AoI in ms. (b) Average downlink AoI in ms. (c) Average CDP in %.

interaction than replay-based learning, which accounts for its slower reward growth despite stable behaviour.

B. Performance Evaluation

Fig. 3 compares the algorithms across three key metrics: (a) Uplink AoI, (b) Downlink AoI, and (c) CDP. Results are averaged over the final 100 episodes. The number of platoons in the system varies from $L = 4$ to $L = 7$, with a fixed number of vehicles in each platoon $V_l = 4$. To benchmark optimality, we include results from an Exhaustive Search (dashed black line), which represents the theoretical upper bound for performance. To render the search computationally feasible over the continuous power domain, we discretise the transmit power levels with a resolution of 0.1 dBm.

The proposed DE-MADDPG (TD3) algorithm consistently achieves the lowest uplink and downlink AoI, closely tracking the theoretical optimum provided by the exhaustive search (with an average gap of 4.8%). As the number of platoons (L) increases, resource contention intensifies. While baseline algorithms show a steep degradation in AoI, our method maintains a flatter slope, demonstrating robust scalability. Among the baselines, the on-policy MAPPO records higher uplink and downlink AoI than the off-policy MADDPG variants, reaching approximately 40.1 and 43.4 ms respectively at $L = 7$, yet it stays ahead of the single-agent DDPG. Notably, the algorithm learns an intelligent trade-off for downlink AoI. It does not simply minimise downlink AoI to a very low level. Instead, it avoids over-requesting downlink broadcasts. Frequent downlink interruptions would occupy time slots required for uplink reporting and V2V safety messages. The proposed agent successfully learns this subtle balance, maintaining sufficiently fresh global data without compromising uplink performance.

Fig. 3(c) reveals the CDP versus the number of platoons. The proposed algorithm maintains a near-perfect CDP (approaching 100%) across all tested densities, matching the exhaustive search. Conversely, the performance of the baselines degrades rapidly as L increases. For $L = 7$, the CDP for DDPG drops to around 60%, the on-policy MAPPO to around 70%, and even the decentralised approach falls to around 90%. In the context of autonomous driving, a 10% failure rate in safety message delivery is unacceptable. The robustness of DE-MADDPG (TD3) confirms its suitability for safety-critical high-speed platoon management, where maintaining string stability via reliable CAM exchange is essential.

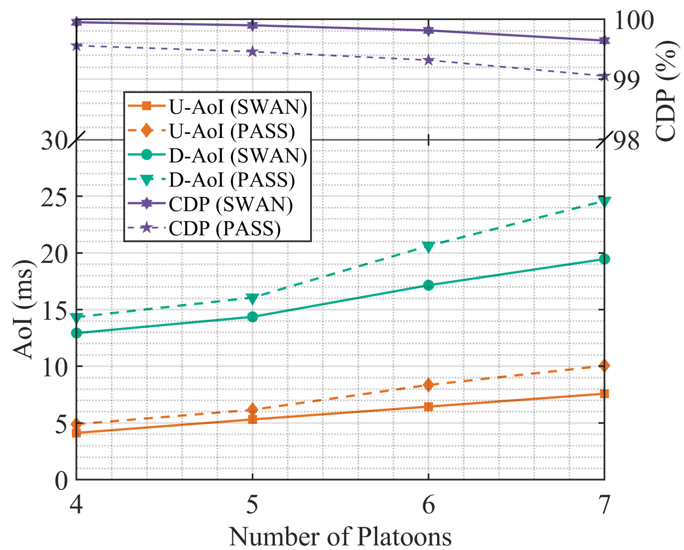


Fig. 4: Performance of the proposed DE-MADDPG (TD3) framework on the SWAN architecture against the PASS baseline. U-AoI and D-AoI denote the average uplink and downlink AoI in ms (left axis), and CDP is shown in % (right axis).

C. Impact of Waveguide Segmentation

To isolate the gain of the SWAN architecture from that of the learning algorithm, Fig. 4 applies the proposed DE-MADDPG (TD3) framework to both SWAN and the PASS baseline ($M = 1$, a single full-length waveguide), changing only the physical layer. Under this controlled setting, SWAN attains lower uplink and downlink AoI and higher CDP across all densities, and the margin widens with L. This performance arises because segmentation caps each active in-waveguide path at D_{road}/M rather than the full road length D_{road}, limiting attenuation and preserving freshness and reliability under dense traffic.

VI. Conclusion

This paper has investigated the potential of dielectric waveguide-based PA systems to enable continuous, high-reliability coverage for high-speed platoon-based V2X networks. We identified that while conventional PAs offer stable LoS propagation, they are fundamentally limited by in-waveguide attenuation over long distances. To address this, we integrated the SWAN architecture with a MADRL framework, enabling the segment and PA activations to cooperate with the dynamic optimisation of RA. This approach effectively balances the multiple conflicting communications requirements.

Unlike existing studies that predominantly focus on uplink timeliness, we formulated a joint optimisation problem considering the freshness of both uplink platoon reporting (uplink AoI) and downlink global traffic updates (downlink AoI), alongside the intra-platoon communication reliability (the CDP) and power consumption. To solve this complex problem, we utilised the DE-MADDPG (TD3) algorithm, which effectively decouples global coordination objectives from local agent policies.

Our results demonstrate that the proposed DE-MADDPG (TD3) algorithm successfully learns a robust policy that maintains fresh global traffic awareness without compromising the timeliness of local platoon update. Notably, the system achieves a CDP approaching 100% even under high traffic density, significantly outperforming standard RL baselines. These findings confirm that the SWAN architecture, when coupled with intelligent resource management, provides a viable and scalable physical-layer solution for safety-critical autonomous driving applications in next-generation ITS.

Furthermore, while this study decoupled PA activation using a geometric approach to ensure learning stability, investigating joint learning of PA activation under, e.g., imperfect CSI or unpredictable physical blockage conditions, remains a highly valuable direction for future work. Moreover, this study adopts the segment-selection strategy, and extending it to the segment-aggregation and segment-multiplexing strategies, which activate multiple segments concurrently at the cost of additional RF chains and inter-segment interference management, represents a further promising direction. In addition, the present framework samples the experience replay buffer uniformly. Integrating more advanced sampling strategies, such as the quantum-inspired experience replay of [27], constitutes a promising direction to further improve the sample efficiency. Finally, in line with the SWAN model of [22], this study assumes ideal segment isolation and PA activation. Accounting for practical non-idealities, such as imperfect segment isolation, switching delay, and hardware losses, is left to future work.

References

[1] S. Jung, J.-H. Kim, and J. Kim, "Intelligent extra resource allocation for cooperative awareness message broadcasting in cellular-V2X networks," *IEEE Trans. Netw. Serv. Manag.*, vol. 22, no. 2, pp. 1677–1689, Apr. 2025.

[2] Z. Xu, S. Xu, H. Ding, and R. Xu, "An ISAC-based beam tracking scheme against inter-region interference for the multi-RSU V2I scenario," *IEEE Trans. Veh. Technol.*, vol. 74, no. 3, pp. 4257–4272, Mar. 2025.

[3] M. Ji *et al.*, "Graph neural networks and deep reinforcement learning-based resource allocation for V2X communications," *IEEE Internet Things J.*, vol. 12, no. 4, pp. 3613–3628, Feb. 2025.

[4] K. Qi *et al.*, "Deep-reinforcement-learning-based AoI-aware resource allocation for ris-aided IoV networks," *IEEE Trans. Veh. Technol.*, vol. 74, no. 1, pp. 1365–1378, Jan. 2025.

[5] X. Gu *et al.*, "Resource reservation coordination for vehicle platooning in C-V2X networks," *IEEE Trans. Wirel. Commun.*, vol. 23, no. 6, pp. 5515–5528, Jun. 2024.

[6] W. Zhang, S. Han, X. Xu, and P. Zhang, "Joint service placement and model partitioning for accelerating DNN inference in edge intelligence empowered vehicle networks," *IEEE Trans. Veh. Technol.*, vol. 74, no. 9, pp. 14 645–14 660, Sep. 2025.

[7] X. Fu *et al.*, "HierNet: A hierarchical resource allocation method for vehicle platooning networks," *IEEE Internet Things J.*, vol. 11, no. 24, pp. 39 579–39 592, Dec. 2024.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] Y. Zheng, S. L. Cotton, O. A. Dobre, and T. Q. Duong, "Quantum multi-agent deep reinforcement learning for energy-efficient vehicular networks," in *Proc. ICC 2025 - IEEE Int. Conf. Commun.*, Montreal, Canada, Jun.8–12 2025, pp. 2039–2044.

[14] Y. Zheng, S. L. Cotton, H. Shin, and T. Q. Duong, "Quantum neural networks for MADRL-assisted optimal resource allocation in vehicular networks," in *Proc. GLOBECOM 2025 - 2025 IEEE Glob. Commun. Conf.*, Taipei, Taiwan, Dec.8–12 2025, accepted.

[15] Y. Zheng, L. D. Nguyen, and T. Q. Duong, "Multiagent deep reinforcement learning for optimal resource allocation in AoI-aware energy-efficient platoon-based C-V2X networks," *IEEE Internet Things J.*, vol. 13, no. 6, pp. 10 814–10 828, Mar. 2026.

[16] *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.

[17] A. Fukuda, H. Yamamoto, H. Okazaki, Y. Suzuki, and K. Kawai, "Pinching antenna: Using a dielectric waveguide as an antenna," *NTT DOCOMO Tech. J.*, vol. 23, no. 3, pp. 5–12, Jan. 2022.

[18] G. Zhu *et al.*, "Pinching-antenna systems (PASS)-enabled secure wireless communications," *IEEE Trans. Commun.*, pp. 1–1, 2025.

[19] S. A. Tegos, P. D. Diamantoulakis, Z. Ding, and G. K. Karagiannidis, "Minimum data rate maximization for uplink pinching-antenna systems," *IEEE Wirel. Commun. Lett.*, vol. 14, no. 5, pp. 1516–1520, May 2025.

[20] K. Wang, Z. Ding, and R. Schober, "Antenna activation for NOMA assisted pinching-antenna systems," *IEEE Wirel. Commun. Lett.*, vol. 14, no. 5, pp. 1526–1530, May 2025.

[21] J. Xiao, J. Wang, and Y. Liu, "Channel estimation for pinching-antenna systems (PASS)," *IEEE Commun. Lett.*, vol. 29, no. 8, pp. 1789–1793, Aug. 2025.

[22] C. Ouyang, H. Jiang, Z. Wang, Y. Liu, and Z. Ding, "Uplink and downlink communications in segmented waveguide-enabled pinching-antenna systems (SWANs)," Sep. 2025, Available on-line at arXiv:2509.10666.

[23] H. Kogelnik, “2. Theory of dielectric waveguides,” in *Integrated optics*. Berlin, Heidelberg: Springer, 1975, pp. 13–81.

[24] D. R. Jackson, C. Caloz, and T. Itoh, “Leaky-wave antennas,” *Proc. IEEE*, vol. 100, no. 7, pp. 2194–2206, Jul. 2012.

[25] Y. Li *et al.*, “Energy-efficient UAV-driven multi-access edge computing: A distributed many-agent perspective,” *IEEE Trans. Commun.*, vol. 73, no. 9, pp. 8405–8420, Sep. 2025.

[26] Y. Li and A. H. Aghvami, “Radio resource management for cellular-connected UAV: A learning approach,” *IEEE Trans. Commun.*, vol. 71, no. 5, pp. 2784–2800, May 2023.

[27] Y. Li, A. H. Aghvami, and D. Dong, “Path planning for cellular-connected UAV: A DRL solution with quantum-inspired experience replay,” *IEEE Trans. Wireless Commun.*, vol. 21, no. 10, pp. 7897–7912, Oct. 2022.

[28] J. F. Bauters *et al.*, “Planar waveguides with less than 0.1 dB/m propagation loss fabricated with wafer bonding,” *Opt. Express*, vol. 19, no. 24, pp. 24 090–24 101, Nov. 2011.

[29] X. Ji *et al.*, “Ultra-low-loss on-chip resonators with sub-milliwatt parametric oscillation threshold,” *Optica*, vol. 4, no. 6, pp. 619–624, Jun. 2017.

[30] S. O. Kasap, *Optoelectronics and photonics*, 2nd ed. Pearson Education UK, 2013.

[31] E. Snitzer, “Cylindrical dielectric waveguide modes*,” *J. Opt. Soc. Am.*, vol. 51, no. 5, pp. 491–498, May 1961.

[32] M. T. Ivrlač and J. A. Nossek, “Toward a circuit theory of communication,” *IEEE Trans. Circuits Syst. Regul. Pap.*, vol. 57, no. 7, pp. 1663–1683, Jul. 2010.

[33] D. M. Pozar, *Microwave Engineering: Theory and Techniques*, 4th ed. Hoboken, NJ, USA: Wiley, 2021.

[34] C. Yeh and F. I. Shimabukuro, *The essence of dielectric waveguides*. New York, NY, USA: Springer, 2008.

[35] T. Q. Duong, H. Shin, and E.-K. Hong, “Effect of line-of-sight on dual-hop nonregenerative relay wireless communications,” in *2007 IEEE 66th Veh. Technol. Conf.*, Baltimore, MD, USA, Sep. 2007, pp. 571–575.

[36] H. Zhang *et al.*, “Beam focusing for near-field multiuser MIMO communications,” *IEEE Trans. Wirel. Commun.*, vol. 21, no. 9, pp. 7476–7490, Sep. 2022.

[37] C. A. Balanis, *Antenna Theory: Analysis and Design*, 4th ed. Hoboken, New Jersey, USA: John Wiley & Sons, 2016.

[38] Z. Ding, R. Schober, and H. V. Poor, “Flexible-antenna systems: A pinching-antenna perspective,” *IEEE Trans. Commun.*, Mar. 2025, DOI: 10.1109/TCOMM.2025.3555866.

[39] Z. Wang, C. Ouyang, X. Mu, Y. Liu, and Z. Ding, “Modeling and beamforming optimization for pinching-antenna systems,” Jun. 2025, Available on-line at arXiv:2502.05917.

[40] 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.

[41] T. Hwang, C. Yang, G. Wu, S. Li, and G. Y. Li, “OFDM and its wireless applications: A survey,” *IEEE Trans. Veh. Technol.*, vol. 58, no. 4, pp. 1673–1694, May 2009.

[42] 3GPP, “Study on LTE-based V2X services,” 3rd Generation Partnership Project (3GPP), Technical Report (TR) 36.885, 2016, version 14.0.0.

[43] Y. Xu, Z. Ding, R. Schober, and T.-H. Chang, “Pinching-antenna systems with in-waveguide attenuation: Performance analysis and algorithm design,” Jun. 2025, Available on-line at arXiv:2506.23966.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] C. Yu *et al.*, “The surprising effectiveness of PPO in cooperative multi-agent games,” in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2022, pp. 24 611–24 624.

[49] 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.

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