A Multi-Agent Reinforcement Learning Approach for Enhanced Spectrum Resource Allocation in NR-V2X Mode 2

Xinyu Chen[1], Kexun He[2], Jing Zhao[1*]
[1]School of Software Technology, Dalian University of Technology, Dalian, China
[2]CATARC Automotive Test Center (Tianjin) Co., Ltd, TATC, Tianjin, China
zhangjiaming@mail.dlut.edu.cn, hekexun@catarc.ac.cn, zhaoj9988@dlut.edu.cn,

***Abstract*—The 3rd Generation Partnership Project (3GPP) has introduced New Radio Vehicle-to-Everything (NR-V2X) as an evolution of Cellular V2X (C-V2X) to satisfy the increasingly stringent communication requirements of emerging V2X applications. In NR-V2X Mode 2, vehicles autonomously perform decentralized spectrum resource allocation via the Sensing-Based Semi-Persistent Scheduling (SPS) algorithm. Nevertheless, the performance of SPS significantly degrades under scenarios with high vehicle density and aperiodic traffic patterns, which hinders the system's ability to meet Quality of Service (QoS) demands. To address these challenges, this study proposes MMATD3-SPS, an enhanced resource allocation algorithm that integrates a Multi-Agent Twin Delayed Deep Deterministic Policy Gradient (MMATD3) framework into the conventional SPS scheme. By leveraging channel state information and application-layer metrics, and applying a reward decomposition strategy, the algorithm optimizes the resource selection process. Experimental results demonstrate that MMATD3-SPS improves the packet reception rate by approximately 10% in high-density traffic environments. Moreover, it ensures that 80% of data packets are updated within 100 milliseconds under aperiodic traffic conditions. These results highlight the proposed algorithm's robustness and scalability, underscoring its potential for deployment in dynamic and complex vehicular communication scenarios.**

***Index Terms*—NR-V2X Mode 2, spectrum resource allocation, semi-persistent scheduling, multi-agent reinforcement learning**

I. Introduction

The advancement of intelligent transportation systems (ITS) plays a pivotal role in enabling sophisticated vehicular applications, including autonomous driving and cooperative perception [1]. These applications demand ultra-reliable and low-latency communication to ensure safety, efficiency, and seamless interaction between vehicles and their surrounding environment. To meet these stringent communication requirements, New Radio Vehicle-to-Everything (NR-V2X), introduced in 3GPP Release 16, represents a significant enhancement over Cellular Vehicle-to-Everything (C-V2X). NR-V2X is explicitly designed to address the limitations of its predecessors while providing robust support for advanced V2X services [2]. Notably, NR-V2X Mode 2 enables vehicles to autonomously select spectrum resources decentralized using the Sensing-Based Semi-Persistent Scheduling (SPS) algorithm. While SPS performs effectively in periodic traffic scenarios, such as routine vehicle position updates, its performance deteriorates in high vehicle density environments and under aperiodic traffic conditions [3–5]. Aperiodic traffic, often critical for autonomous driving scenarios like emergency braking and cooperative perception, involves unpredictable transmission intervals. These dynamics exacerbate the risks of packet collisions, transmission delays, and inefficient spectrum utilization, thereby posing significant challenges to maintaining the Quality of Service (QoS) required for these advanced applications.

The SPS algorithm primarily employs periodic resource reservation and resource filtering mechanisms to facilitate spectrum allocation. Despite its inherent simplicity, the algorithm's limitations in dynamic vehicular environments have motivated various enhancements to address its shortcomings. For example, Dayal et al. [6] proposed an adaptive SPS framework that adjusts resource reservation intervals based on traffic conditions, effectively reducing interference and extending the communication range. Similarly, Gu et al. [5] optimized SPS parameters, such as resource reservation intervals, using collision probability and delay models, resulting in improved channel congestion management and QoS. Moreover, Abbas et al. [7] developed a two-stage resource selection strategy that integrates traffic density and channel state information, achieving lower latency and higher throughput. However, conventional SPS algorithms often rely on random resource selection from a pool of detected idle resources, which can lead to frequent collisions in high vehicle density vehicular scenarios.

Recent advances in reinforcement learning (RL) have shown strong potential for V2X resource allocation. Parvini et al. [8] introduced a resource allocation algorithm based on the Twin Delayed Deep Deterministic Policy Gradient (TD3) technique, which outperformed centralized and federated learning-based strategies in vehicular platoon scenarios. Hegde et al. [9] introduced an actor-critic algorithm for aperiodic traffic, achieving better performance than SPS across varying traffic conditions. Similarly, Lee et al. [10] developed a decentralized multi-agent RL framework tailored to heterogeneous traffic, delivering near-optimal results under both light and heavy loads. These studies highlight RL's effectiveness in addressing the increasing complexity and diverse demands of V2X systems.

To address the challenges of frequent resource collisions,

reduced communication reliability under high vehicle density, and inefficiencies in managing aperiodic traffic conditions, we propose an enhanced spectrum resource allocation algorithm based on the Modified Multi-Agent Twin Delayed Deep Deterministic Policy Gradient (MMATD3), termed MMATD3-SPS. The main contributions of this work are summarized as follows:

- MMATD3-SPS introduces a novel reward decomposition mechanism and a dual-task agent framework to mitigate the inefficiencies of the random resource selection process inherent in the SPS algorithm. This approach translates QoS requirements, such as transmission reliability and low latency, into sub-task rewards, empowering agents to make more precise and effective decisions.
- By leveraging channel state information and application-layer metrics, MMATD3-SPS optimizes dynamic resource selection, making it suitable for both periodic and aperiodic traffic scenarios. A centralized agent is employed to expedite training convergence, facilitating efficient computation and reliable communication.
- Simulation results show that MMATD3-SPS outperforms traditional SPS, improving packet reception rates by approximately 10% in high-density scenarios and ensuring 80% of aperiodic updates are completed within 100 ms, demonstrating superior scalability and robustness.

The remainder of this paper is organized as follows: Section II details the system model, while Section III outlines the problem description. Section IV presents the proposed MMATD3-SPS algorithm in detail. Section V, the performance of the proposed resource allocation algorithm is evaluated through simulation results. Finally, VI concludes the paper with a summary of the findings.

II. System Model

NR-V2X utilizes Orthogonal Frequency Division Multiplexing (OFDM), a technology that converts selected channels in the frequency domain into parallel flat channels across multiple subcarriers. Assuming that channel fading is approximately uniform within a subchannel and independent across different subchannels, several contiguous subcarriers are grouped into a spectral subchannel. In highway scenarios without base station coverage, vehicles communicate through the sidelink PC5 interface for vehicle-to-vehicle (V2V) communication. In the absence of resource coordination by a central node, vehicles autonomously select spectrum resources using SPS. These vehicles periodically broadcast basic safety messages or transmit intelligent sensing information on an aperiodic basis. The system model is shown in Fig. 2 below. Let the set of vehicles be denoted as $\mathcal{I} = \{1, \ldots, I\}$. At time slot t, the set of vehicles within the communication range of vehicle i is denoted as $J_i^{(t)} = \{1, \ldots, J_i^{(t)}\}$, while the set of interfering vehicles for vehicle i is represented as $\mathcal{K}_i^{(t)} = \{1, \ldots, K_i^{(t)}\}$. The set of available spectrum resources is defined as $\mathcal{R} = \{1, \ldots, R\}$. To model resource selection, let $c_{i,r}^{(t)} \in \{0, 1\}$ indicate whether vehicle $i \in I$ selects spectrum

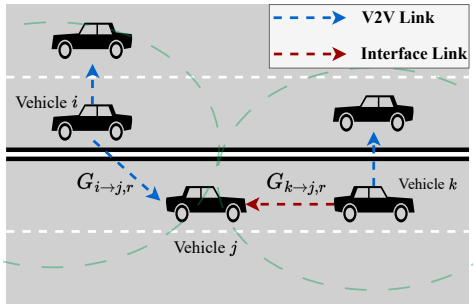


Fig. 1: System model in highway scenarios.

resource $r \in R$ for message transmission at time slot t. Each vehicle is constrained to select at most one spectrum resource in each time slot:

$$\sum_{r\in\mathcal{R}} c_{i,r}^{(t)} \leq 1, \quad \forall i \in \mathcal{I},\ \forall t \tag{1}$$

The transmission gain between vehicles dynamically changes with variations in vehicle positions and transmission tasks. However, given the relatively short duration of a time slot, it is assumed that the transmission gain remains approximately constant within the same time slot. During time slot t, the transmission gain of vehicle j monitoring subchannel resource r is denoted as $G_{j,r}^{(t)}$, which is expressed as follows:

$$G_{j,r}^{(t)} = \alpha_j^{(t)} h_{j,r}^{(t)} \tag{2}$$

where $\alpha_j^{(t)}$ and $h_{j,r}^{(t)}$ represent the large-scale fading effect caused by path loss and shadowing, and the small-scale fading effect caused by multipath propagation, respectively.

When vehicle i utilizes spectrum resource r for broadcast transmission, the interference experienced during V2V direct communication between vehicle i and vehicle j in time slot t is defined as:

$$I_{i\to j,r}^{(t)} = \sum_{k\in\mathcal{K}_i^{(t)}} P_{k,r} G_{k\to j,r}^{(t)} \tag{3}$$

where $P_{k,r}$ represents the signal transmission power of vehicle k when selecting spectrum resource r for transmission. If spectrum resource r is not selected by vehicle k, then $P_{k,r} = 0$.

The signal-to-noise and interference ratio (SINR) for the communication link between vehicle i and vehicle j using spectrum resource r at time slot t is expressed as:

$$SINR_{i,j,r}^{(t)} = \frac{c_{i,r}^{(t)} P_{i,r} G_{i\to j,r}^{(t)}}{P_n + I_{i\to j,r}^{(t)}} \tag{4}$$

where P_n represents the noise power.

According to Shannon's theorem [11], the transmission channel capacity for this communication link is given by:

$$C_{i,j,r}^{(t)} = B \log_2 \left(1 + SINR_{i,j,r}^{(t)}\right) \tag{5}$$

where B represents the bandwidth occupied by a single frequency resource, and $C_{i,j,r}^{(t)}$ denotes the theoretical maximum data transmission rate between vehicle i and vehicle j using resource r at time slot t.

III. Problem Formulation

Advanced V2X applications, such as cooperative driving and autonomous traffic management, demand strict adherence to QoS constraints to ensure reliable and efficient communication. These applications involve the exchange of critical information, such as cooperative perception data, vehicle trajectories, and control instructions, necessitating high reliability and low latency for V2X transmissions. The mathematical representations of these QoS constraints are as follows:

A. QoS Constraints

1) Reliability Constraint: In advanced V2X applications, communication reliability is crucial to ensure accurate and timely sharing of critical information. An outage event occurs when the SINR falls below a minimum threshold $SINR_{\text{th}}$, resulting in packet loss and degraded system performance. To meet reliability requirements, the outage probability p_{outage} must remain within an acceptable limit [12]. The reliability constraint for V2X transmissions between transmitter i and receiver j on resource r is expressed as:

$$p_{\text{outage}} = 1 - e^{-\frac{SINR_{\text{th}}}{SINR_{i,j,r}^{(t)}}} \tag{6}$$

2) Latency Constraint: Low latency is essential for time-sensitive V2X applications, such as emergency braking or real-time traffic updates. Since V2X typically employs distributed resource selection schemes like SPS, central scheduling delays are eliminated, and only transmission latency is considered. The latency constraint can be formulated as:

$$t_{\text{delay}} = \frac{Z}{C_{i,j,r}^{(t)}} = \frac{Z}{B\log_2\left(1 + SINR_{i,j,r}^{(t)}\right)} \tag{7}$$

where Z represents the packet size of the V2X message.

B. Optimization Problem

Let $x_i^{(t)} = r_i^{(t)}$ denote the decision variable representing the resource selection by vehicle i in time slot t, where $r_i^{(t)}$ indicates the frequency resource chosen by the vehicle. We propose a novel multi-objective optimization problem aimed at enhancing the reliability and latency performance of vehicular applications that rely on V2V links. The mathematical formulation of the optimization problem is as follows:

$$\begin{aligned} \min_{x_i^{(t)}} \quad & \left\{1 - p_{\text{outage},i}^{(t)},\ t_{\text{delay},i}^{(t)}\right\} \\ \text{s.t.} \quad & \sum_{r\in\mathcal{R}} c_{i,r}^{(t)} \le 1, \quad \forall i \in \mathcal{I},\ \forall t \end{aligned} \tag{8}$$

where the objective function seeks to minimize the outage probability and transmission latency while ensuring that each vehicle selects exactly one spectrum resource in each transmission time slot.

IV. MMATD3-SPS algorithm

This section mainly introduces the modeling of multi-agent environments and the enhanced SPS algorithm based on MMATD3.

A. Modeling of Multi-Agent Environments

To solve the optimization problem defined in Equation (8), we reformulate it as a Markov Decision Process (MDP) and design RL elements, represented by $\langle S, A, P, R, \gamma\rangle$. Here, S denotes the state space, A represents the action space, P is the state transition model, R is the reward function, and γ is the discount factor. Each agent aims to learn an optimal policy π_i^* to maximize the cumulative reward over time. Given the difficulty of obtaining state transition probabilities, we adopt the model-free multi-agent RL algorithm MATD3 [13], which learns the optimal policy through trial and error. MATD3 is integrated into the SPS framework to dynamically optimize spectrum resource allocation and improve vehicular network performance.

1) State: The state space consists of perceived channel state information and application-layer metrics. At time slot t, the local state information observed by an agent i includes the transmission gains $\{G_{j,r}^{(t)}\}_{j\in\mathcal{R}}$, interference from other V2V links $\{I_{i,j,r}^{(t)}\}_{j\in\mathcal{K},r\in\mathcal{R}}$, and application-layer metrics such as the latency budget $\zeta_i^{(t)}$ for the current transmission. These elements are represented as:

$$\boldsymbol{s}_i^{(t)} = \left\{\{G_{j,r}^{(t)}\}_{j\in\mathcal{R}}, \{I_{i,j,r}^{(t)}\}_{j\in\mathcal{K},r\in\mathcal{R}}, \zeta_i^{(t)}\right\} \tag{9}$$

2) Action: The action space corresponds to the decision variable defined in Equation (8), specifically the spectrum resource selected by vehicle i at time slot t, represented as:

$$a_i^{(t)} = r_i^{(t)} \tag{10}$$

3) Reward: The reward function is critical for addressing high-dimensional and complex task scenarios in reinforcement learning. Associating the reward obtained from each action with the expected objective enhances overall system performance. For the multi-objective optimization problem proposed in Equation (8), the optimization goals are decomposed into two sub-tasks: stable transmission and timely transmission. Therefore, the reward for vehicle i executing action $a_i^{(t)}$ at time slot t is defined as:

$$r_i^{(t)} = \underbrace{\kappa_1 p_{\text{outage},i}^{(t)}}_{\text{Reliability}} - \underbrace{\kappa_2 t_{\text{delay},i}^{(t)}}_{\text{Latency}} \tag{11}$$

where κ_1 and κ_2 are weight coefficients for reliability and latency, respectively. These weights align with the optimization objectives in Equation (8). The first term rewards stable transmission, while the second term penalizes latency violations.

To account for the overall performance of the vehicular network, a global reward is introduced. This global reward is designed for the centralized training and distributed execution (CTDE) paradigm in multi-agent reinforcement learning and is expressed as:

$$r_g^{(t)} = -\frac{1}{N}\sum_{j\in J}\sum_{r\in\mathcal{R}} \log\left\{I_{j,r}^{(t)}\right\} \tag{12}$$

where $I_{j,r}^{(t)}$ represents the interference power observed by vehicle j on spectrum resource r, and N denotes the total

number of observed spectrum resources. The global reward is utilized during centralized training to guide agents in selecting spectrum resources with lower interference.

B. MMATD3-SPS algorithm framework and flow

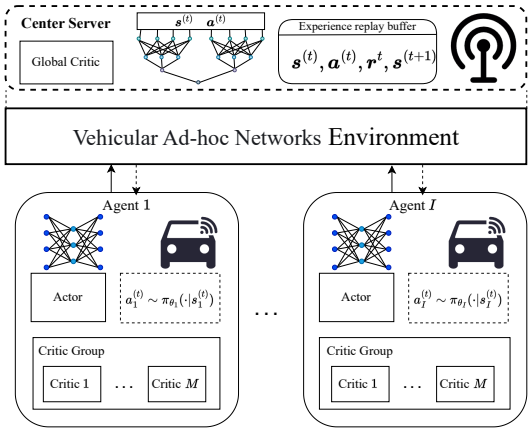


Fig. 2: MMATD3-SPS resource allocation algorithm framework.

Multi-Agent Twin Delayed Deep Deterministic Policy Gradient (MATD3) is a reinforcement learning algorithm tailored for multi-agent continuous control problems. It extends TD3 to multi-agent scenarios, leveraging CTDE to enhance training efficiency and policy stability. In vehicular spectrum resource allocation, MATD3 optimizes independent policy and value networks for each agent while enabling cooperation among vehicles to improve overall network performance. As shown in Fig. 2, the architecture includes a critic group and an actor network for local agents, as well as a centralized critic for global rewards. For agent j, the actor network parameters are updated using policy gradients as follows:

$$\nabla_{\phi_j} \mathcal{J}_j = \mathbb{E}\left[\nabla_{\phi_j} \pi_j\left(a_j \mid s_j\right) \nabla_{a_j} Q_j\left(s_j, a_j\right)\right]_{a_j=\pi_j(s_j)} \tag{13}$$

where $Q_j(s_j, a_j)$ represents the Q-value estimation by the agent's critic network.

The policy gradient of the local actor network, incorporating the global critic network, is defined as follows:

$$\begin{aligned}\nabla_{\phi_j} \mathcal{J}_j = &\underbrace{\mathbb{E}_{s,a\sim\mathcal{D}}\left[\nabla_{\phi_j} \pi_j\left(a_j \mid s_j\right) \nabla_{a_j} Q_\psi^g(s,a)\right]}_{\text{Global Critic}} \\ &+ \underbrace{\mathbb{E}_{s_j,a_j\sim\mathcal{D}}\left[\nabla_{\phi_j} \pi_j\left(a_j \mid s_j\right) \nabla_{a_j} Q_j\left(s_j, a_j\right)\right]}_{\text{Local Critic}}\end{aligned} \tag{14}$$

where $Q_\psi^g(s,a)$ represents the Q-value estimation by the global critic network for the joint state-action pair. $\mathcal{D}$ denotes the experience replay buffer, which stores interaction data with the vehicular network environment for sampling and training.

Additionally, this study introduces a reward decomposition mechanism based on the requirements for transmission reliability and low latency. The local reward is divided into two components: reliability reward and low-latency reward. The decomposed rewards obtained by agent i at time slot t are expressed as follows:

$$r_{1,i}^{(t)} = \kappa_1 p_{\text{outage},i}^{(t)} \tag{15}$$

where $r_{1,i}^{(t)}$ represents the reward obtained for achieving reliable transmission. The reward for achieving low-latency transmission is defined as:

$$r_{2,i}^{(t)} = r_i^{(t)} - r_{1,i}^{(t)} = -\kappa_2 t_{\text{delay},i}^{(t)} \tag{16}$$

Due to the additive relationship between local decomposed rewards, Equation (14) can be equivalently expressed as:

$$\begin{aligned}\nabla_{\phi_j} \mathcal{J}_j = &\underbrace{\mathbb{E}_{s,a\sim\mathcal{D}}\left[\nabla_{\phi_j} \pi_j\left(a_j \mid s_j\right) \nabla_{a_j} Q_\psi^g(s,a)\right]}_{\text{Global Critic}} \\ &+ \sum_{k=1}^{M} \underbrace{\mathbb{E}_{s_j,a_j\sim\mathcal{D}}\left[\nabla_{\phi_j} \pi_j\left(a_j \mid s_j\right) \nabla_{a_j} Q_{j,k}\left(s_j, a_j\right)\right]}_{\text{Decomposed Local Critic}}\end{aligned} \tag{17}$$

where M represents the total number of transmission requirements with different properties such as stability and reliability. The first term denotes the global reward evaluated by the global critic network according to Equation (12), while the second term represents the sum property rewards evaluated by the local critic network based on Equations (15) and (16).

Unlike the update strategy for the actor network of local agents, the update of the critic network for local agent j relies on the temporal difference error for transmission property k. The loss function is defined as:

$$\mathcal{L}(\theta_{j,k}) = \mathbb{E}_{s_j,a_j,r_j,s_j'}\left[\left(Q_{\theta_{j,k}}(s_j,a_j) - y_{j,k}\right)^2\right] \tag{18}$$

where the target $y_{j,k}$ is defined as:

$$y_{j,k} = r_{k,j} + \gamma Q_{\theta'_{j,k}}\left(s_j', a_j'\right)|_{a_j'=\pi_j'(s_j')} \tag{19}$$

Similarly, the loss function of the global critic network is defined as follows. To solve problem of high bias estimation of Q-value during the training of the critic network, the MATD3 algorithm adopts the dual critic network to solve the problem:

$$\mathcal{L}(\psi_i) = \mathbb{E}_{s,a,r,s'}\left[\left(Q_{\psi_i}^g(s,a) - y_g\right)^2\right] \tag{20}$$

where the target y_g is defined as:

$$y_g = r_g + \gamma min_{i=1,2} Q_{\psi_i'}^g(s',a')|_{a_j'=\pi_j'(s_j')} \tag{21}$$

V. Simulation Results and Analysis

In this section, we introduce the simulation environment and parameters, including network topology, communication model, and key settings. Then, we present the evaluation metrics such as reliability and latency. Finally, we show the simulation results with quantitative analysis and comparison to baseline methods.

A. Simulation Setup

We utilized OpenCV2X [14] as the core framework for simulating NR-V2X Mode 2 communications. By extending SimuLTE [15], which supports LTE-V2X Mode 4 scenarios, the Winner+B1 channel model and NR-V2X link-level datasets [16] were integrated to enhance simulation accuracy. Since OpenCV2X lacks built-in RL integration, Veins-Gym [17] was adopted to bridge reinforcement learning algorithms with vehicular spectrum allocation research. This framework seamlessly integrates with RL libraries such as Stable-Baselines3 [18], enabling efficient algorithm development. The proposed MMATD3-SPS algorithm was trained and evaluated in the 3GPP highway scenario [19], which utilizes a wrap-around design to ensure consistent vehicle density and realistic simulation conditions. The detailed simulation parameters for communication and reinforcement learning training are provided in Tables I and II. The generation of aperiodic traffic follows the equation:

$$t_g = c + r \tag{22}$$

where t_g represents the interval between packet generations, c is a constant set to 50 ms, and r is an exponentially distributed random variable with a mean equal to c.

TABLE I: Simulation Parameters

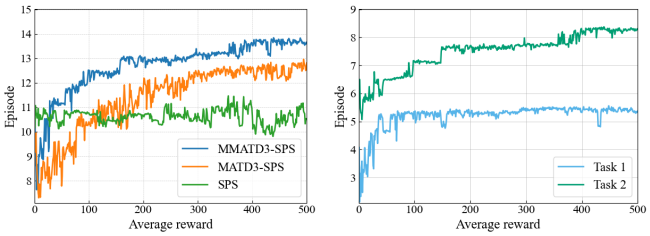
Parameter	Value
Maximum Vehicle Speed	70 km/h
Vehicle Density	{0.06, 0.12, 0.18} veh/m
Carrier Frequency	5.9 GHz
Channel Bandwidth	10 MHz
Number of Subchannels	3
Subchannel Size	16
Modulation and Coding Scheme (MCS)	MCS 13
Packet Size	190 bytes
Traffic Type	Periodic/Aperiodic
Message Transmission Frequency	20 Hz
Channel Model	Winner+B1
Noise Gain	9 dB
Antenna Gain	3 dB
RSRP Threshold	-128 dBm

TABLE II: Hyperparameters for Training

Parameter	Value
Learning Rate (Actor) l_a	0.001
Learning Rate (Critic) l_c	0.001
Number of Episodes N_{eps}	500
Discount Factor γ	0.99
Soft Update Coefficient μ	0.1
Soft Update Frequency α	0.001
Batch Size b	256

B. Simulation result

1) Training process analysis: Fig. 3(a) illustrates the variation in the average reward during the training process of the MMATD3-SPS algorithm, where MATD3-SPS refers to MMATD3-SPS without the reward decomposition mechanism. Overall, the MMATD3-SPS algorithm exhibits some fluctuations in average reward due to its exploration-oriented strategy, which seeks to uncover potential decision gains. After approximately 30 training episodes, MMATD3-SPS achieves higher average rewards compared to the SPS algorithm, whereas MATD3-SPS requires around 120 episodes to reach comparable performance. This improvement is attributed to the reward decomposition mechanism of MMATD3-SPS, which enables more accurate policy evaluation via subtask-specific Critic networks. Beyond 400 training episodes, the average reward of MMATD3-SPS stabilizes, outperforming both MATD3-SPS and SPS in terms of final reward levels. Fig. 3(b) presents the average reward evolution for the two tasks during the training of MMATD3-SPS. Task 1 corresponds to reliability-oriented objectives, while Task 2 focuses on delay-sensitive objectives. Both tasks demonstrate good convergence performance. The reliability-oriented task converges after approximately 80 episodes, benefiting from the inherent reliability provided by the standard SPS resource reservation mechanism. In contrast, the delay-sensitive task requires around 400 episodes to converge, as it faces challenges in effectively learning delay-related information within relatively stable environments.

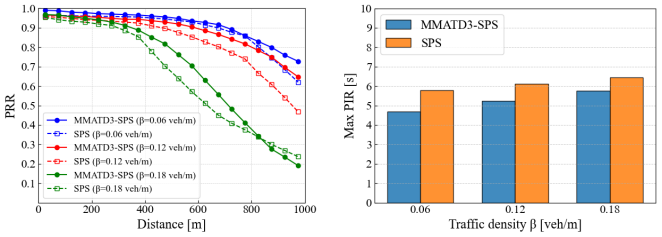


(a) Average reward per episode (b) Task reward per episode

Fig. 3: The training process of different resource selection algorithms.

2) Performance of proposed algorithm: Fig. 4(a) and Fig. 4(b) compare the reliability and latency performance of the MMATD3-SPS algorithm with the SPS algorithm in scenarios featuring periodic traffic. As shown in Fig. 4(a), the packet reception rate (PRR) of MMATD3-SPS consistently surpasses that of the SPS algorithm across a wide intermediate distance range. Moreover, the PRR performance of MMATD3-SPS demonstrates a progressively more significant advantage over SPS as vehicle density increases. For instance, at a vehicle density of 0.18 veh/m, MMATD3-SPS achieves approximately a 10% improvement in PRR over SPS within the distance range of 425 m to 675 m. Fig. 4(b) highlights the latency performance advantage of MMATD3-SPS in terms of packet inter-reception (PIR). A lower maximum PIR indicates the system's ability to respond more effectively to network dynamics, thereby enhancing communication efficiency. Across varying vehicle densities, MMATD3-SPS consistently delivers at least a 50 ms improvement in maximum PIR performance compared to the SPS algorithm.

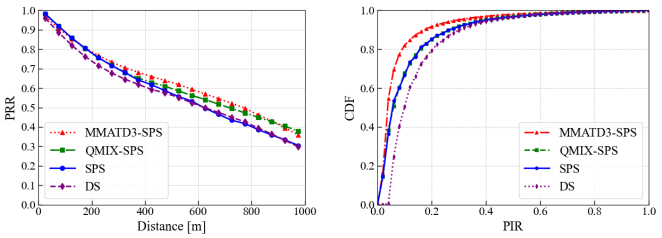
To evaluate the performance of MMATD3-SPS under aperiodic traffic conditions, additional simulations were conducted using Dynamic Scheduling (DS) [20] and QMIX-SPS [21] algorithms as benchmarks. The results in Fig. 5(a) demonstrate that MMATD3-SPS achieves superior reliability compared to resource allocation algorithms conforming to the NR-V2X



(a) PRR vs. distance (b) Max PIR vs. β

Fig. 4: The PRR and max PIR of SPS and MMATD3-SPS under different vehicle densities.

standard. Furthermore, as shown in Fig. 5(b), MMATD3-SPS outperforms the dynamically optimized QMIX-SPS algorithm in PIR performance, overcoming the randomness inherent in the SPS resource selection process. Specifically, 80% of packets are successfully updated within 100 ms, underscoring the effectiveness of MMATD3-SPS in mitigating the limitations of the SPS selection mechanism.



(a) PRR vs. distance (b) CDF of PIR

Fig. 5: Performance comparison under aperiodic traffic conditions.

VI. Conclusion

This paper addresses resource allocation challenges in NR-V2X Mode 2, where the SPS algorithm performs poorly under high vehicle density and aperiodic traffic, failing to meet QoS standards. We propose the MMATD3-SPS algorithm, which integrates channel state information and application-layer metrics to enhance resource selection. By incorporating a reward decomposition mechanism, MMATD3-SPS translates QoS requirements into actionable rewards, enabling efficient decision-making. Simulation results show that MMATD3-SPS improves resource allocation efficiency, outperforming traditional SPS in packet reception rate and latency. The algorithm's adaptability to varying traffic conditions and scalability highlight its potential for dynamic vehicular networks.

Future work will focus on extending the framework to handle more complex traffic scenarios, optimizing SPS parameters, and improving its applicability in large-scale real-world networks.

References

[1] E. Y. Bejarbaneh, H. Du, and F. Naghdy, "Exploring shared perception and control in cooperative vehicle-intersection systems: A review," *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[2] M. Harounabadi, D. M. Soleymani, S. Bhadauria, M. Leyh, and E. Roth-Mandutz, "V2X in 3GPP standardization: NR sidelink in release-16 and beyond," *IEEE Communications Standards Magazine*, vol. 5, no. 1, pp. 12–21, 2021.

[3] A. Dayal, V. K. Shah, B. Choudhury, V. Marojevic, C. Dietrich, and J. H. Reed, "Adaptive semi-persistent scheduling for enhanced on-road safety in decentralized V2X networks," in *2021 IFIP Networking Conference (IFIP Networking)*. IEEE, 2021, pp. 1–9.

[4] Y. Yoon and H. Kim, "A stochastic reservation scheme for aperiodic traffic in nr v2x communication," in *2021 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2021, pp. 1–6.

[5] X. Gu, J. Peng, L. Cai, Y. Cheng, X. Zhang, W. Liu, and Z. Huang, "Performance analysis and optimization for semi-persistent scheduling in c-v2x," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 4, pp. 4628–4642, 2022.

[6] A. Dayal, V. K. Shah, H. S. Dhillon, and J. H. Reed, "Adaptive RRI selection algorithms for improved cooperative awareness in decentralized NR-V2X," *IEEE Access*, vol. 11, pp. 134 575–134 588, 2023.

[7] F. Abbas, G. Liu, P. Fan, Z. Khan, and M. S. Bute, "A vehicle density based two-stage resource management scheme for 5G-V2X networks," in *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*. IEEE, 2020, pp. 1–5.

[8] 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 Transactions on Vehicular Technology*, vol. 72, no. 8, pp. 9880–9896, 2023.

[9] A. Hegde, R. Song, and A. Festag, "Radio resource allocation in 5G-NR V2X: a multi-agent actor-critic based approach," *IEEE Access*, 2023.

[10] I. Lee and D. K. Kim, "Decentralized multi-agent dqn-based resource allocation for heterogeneous traffic in V2X communications," *IEEE Access*, 2024.

[11] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[12] L. Liang, S. Xie, G. Y. Li, Z. Ding, and X. Yu, "Graph-based resource sharing in vehicular communication," *IEEE Transactions on Wireless Communications*, vol. 17, no. 7, pp. 4579–4592, 2018.

[13] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *International Conference on Machine Learning*, 2018.

[14] B. McCarthy, A. Burbano-Abril, V. R. Licea, and A. O'Driscoll, "OpenCV2X: Modelling of the V2X cellular sidelink and performance evaluation for aperiodic traffic," *arXiv preprint arXiv:2103.13212*, 2021.

[15] A. Virdis, G. Stea, and G. Nardini, "SimuLTE-A modular system-level simulator for LTE/LTE-A networks based on OMNeT++," in *2014 4th International Conference On Simulation And Modeling Methodologies, Technologies And Applications (SIMULTECH)*. IEEE, 2014, pp. 59–70.

[16] L. Lusvarghi, B. Coll-Perales, J. Gozalvez, and M. L. Merani, "Link level analysis of NR V2X sidelink communications," *IEEE Internet of Things Journal*, 2024.

[17] M. Schettler, D. S. Buse, A. Zubow, and F. Dressler, "How to train your ITS? integrating machine learning with vehicular network simulation," in *2020 IEEE Vehicular Networking Conference (VNC)*. IEEE, 2020, pp. 1–4.

[18] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-baselines3: Reliable reinforcement learning implementations," *Journal of Machine Learning Research*, vol. 22, no. 268, pp. 1–8, 2021.

[19] M. H. C. Garcia, A. Molina-Galan, M. Boban, J. Gozalvez, B. Coll-Perales, T. Şahin, and A. Kousaridas, "A tutorial on 5G NR V2X communications," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 3, pp. 1972–2026, 2021.

[20] "Evolved universal terrestrial radio access (E-UTRA); radio resource control (RRC); protocol specification," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 36.331, Dec. 2018.

[21] P. Fan, X. Chen, J. Zhao, N. Lu, and P. Wang, "Multi-agent reinforcement learning based adaptive parameter optimization for semi-persistent scheduling in C-V2X mode 4," in *2024 IEEE 21st International Conference on Mobile Ad-Hoc and Smart Systems (MASS)*. IEEE, 2024, pp. 143–149.