Quantum Multi-Agent Deep Reinforcement Learning for Energy-Efficient Vehicular Networks

Yuxiang Zheng*, Simon L. Cotton†, Octavia A. Dobre*, and Trung Q. Duong*†
*Memorial University, Canada (e-mails: {y.zheng, tduong, odobre}@mun.ca)
†Queen's University Belfast, UK (e-mail: {simon.cotton, trung.q.duong}@qub.ac.uk)

Abstract—**In this paper, we address the complex mixed-integer nonlinear programming problem associated with channel assignment and joint power-energy allocation in urban platoon-based cellular-vehicle-to-everything (C-V2X) networks. In this context, the potential advantages of integrating quantum neural networks (QNNs) with classical multi-agent deep reinforcement learning (MADRL) approaches are investigated. Specifically, we combine a variational quantum circuit (VQC) with traditional neural networks to develop a hybrid quantum-classical neural network for the MADRL training process. Our goal is to employ this hybrid quantum-classical approach to simultaneously minimise the average age of information (AoI) which quantifies the freshness of information exchange between vehicle platoons and the roadside unit (RSU), maximise the cooperative awareness message (CAM) exchange probability among vehicles within the same platoon, and foster sustainable, green communication strategies through efficient management for both power and energy. We introduce the innovative decomposed multi-agent deep deterministic policy gradient (DE-MADDPG) algorithm, which is integrated with the twin delayed deep deterministic policy gradient (TD3) technique and advanced quantum computing technologies, resulting in our proposed hybrid quantum-classical decomposed multi-agent TD3 (DE-MATD3) algorithm. Compared with classical approaches, our numerical results reveal that the proposed algorithm achieves exceptional energy efficiency performance, while maintaining the algorithm convergence rate and AoI levels.**

I. Introduction

Intelligent transportation systems (ITS) have been extensively studied [1]–[17]. They are considered pivotal components in any smart city design [1]–[3] due to their potential to mitigate traffic congestion, reduce accident risks, enhance urban air quality, and make work and life more efficient for both city authorities and citizens [3]. Vehicle-to-everything (V2X) communications [4]–[7] play a critical role in ITS by enabling vehicle-to-vehicle (V2V), vehicle-to-pedestrian (V2P), vehicle-to-infrastructure (V2I), and vehicle-to-network (V2N) communications, facilitating near real-time updates on traffic conditions and hazards.

V2X communications will be essential for enabling platoon-based control strategies [2], [6]–[12], which groups closely aligned autonomous vehicles into platoons to enhance traffic flow and control efficiency. Within each platoon, the lead vehicle, or platoon leader (PL), communicates with the roadside unit (RSU) and its same-platoon vehicles, i.e., platoon members (PMs), by sending platoon state messages, receiving control commands, and exchanging cooperative awareness messages (CAMs) [4]–[6] with other platoon members via V2I and V2V links. Frequent updates from PLs to the RSU and PMs to PLs are essential for maintaining time-sensitive information, such as safety alerts, within the whole system. The concept of the age of information (AoI) is introduced to quantify update frequency, with AoI increasing whenever PLs fail to communicate with the RSU. Due to the importance of AoI within vehicular networks, a number of studies have focused on minimising it [10]–[13].

Quantum technology has revolutionized computing by utilizing parameterised quantum gates that can be trained by classical optimisation methods, advancing the quantum machine learning (QML) framework [14]–[19]. QML embeds classical data into quantum bits (qubits) and leverages superposition and entanglement to streamline neural networks (i.e. quantum neural networks, QNNs [18], [19]) to accelerate training. This approach offers substantial computational benefits, especially in 6G wireless communications, where QML can address real-time tasks such as signal processing, channel estimation, and resource allocation effectively. However, the application of QML in vehicular networks remains relatively underexplored, with current research primarily focusing on traditional optimisation tools or classical machine learning (ML) techniques. Quantum technology has been investigated for use in ITS [14]–[17]. Challenges like quantum noise, qubit decoherence, and the limited number of qubits in existing hardware impede the practical deployment of QML algorithms. Therefore the development of efficient quantum algorithms that operate within inherent constraints associated with V2X networks, while maintaining performance, is essential.

In this paper, we tackle the resource allocation challenge posed by interference, dynamic vehicular environments, and constrained bandwidth and power, while advocating for sustainable, green communication strategies [20]. We formulate a joint optimization problem for AoI, CAM delivery probability, and power-energy consumption in platoon-based cellular-vehicle-to-everything (C-V2X) networks at urban intersections [5]–[7]. Our model integrates Mode 4 distributed resource allocation [7] employing a multi-agent deep reinforcement learning (MADRL) framework with the multi-agent deep deterministic policy gradient (MADDPG) algorithm. Some performance enhancement techniques such as the decomposed MADDPG (DE-MADDPG) [21] and twin delayed deep deterministic policy gradient (TD3) [22] are used to support the MADRL framework, forming the decomposed multi-agent TD3 (DE-MATD3) algorithm. Additionally, we integrate QML into the MADRL framework by employing a hybrid quantum-classical

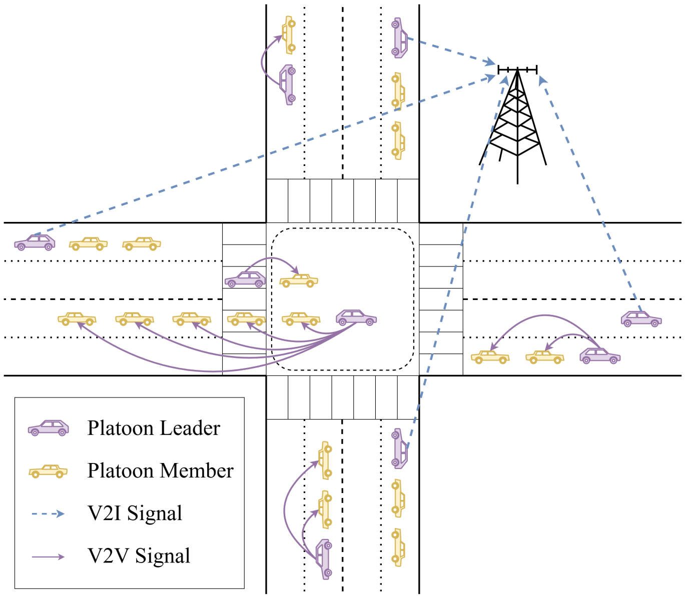

Fig. 1: A single-antenna multi-platoon C-V2X network.

neural network with a variational quantum circuit (VQC) as the QNN, leveraging the potential of quantum computing to further enhance performance [23]. Consequently, our hybrid quantum-classical DE-MATD3 scheme is proposed and evaluated through numerical results.

II. System Model and Problem Formulation

A C-V2X communication network designed to manage multiple vehicle platoons at an intersection is depicted in Fig. 1. The network features a centrally-located single-antenna RSU to coordinate communications among multiple vehicle platoons and RSU itself at this intersection. Let $\mathcal{N} = \{1, 2, \ldots, N\}$ represent the set of platoons, where $N \in \mathbb{N}^+$ denotes the total number of platoons in the system. Within platoon $n \in \mathcal{N}$, there are $v_n \in \mathbb{N}^+$ automated vehicles, sequentially numbered from 1 to v_n. The vehicle numbered 1 in each platoon serves as the PL, responsible for leading communications both within its platoon and with the RSU, which are the V2V and V2I communications, respectively.

The communication decision is governed by a binary variable, $\mu_{n,t} \in \{0, 1\}$, where 1 signifies the V2V mode is selected, and 0 indicates the V2I mode is selected. Here, t represents the index for discrete time slots of equal duration Δt. Δt also represents the single coherence time for channel fading, where the channel fading is assumed to be independent across different subchannels and constant within each Δt. In this work, we consider orthogonal frequency division multiplexing (OFDM) [24], which is known to handle frequency-selective wireless channels effectively. The system bandwidth is partitioned into K orthogonal subchannels with size W, forming the set $\mathcal{K} = \{1, 2, \ldots, K\}$. A Boolean variable, $\xi_{n,k,t} \in \{False, True\}$, is defined for managing the channel assignment task. If $\xi_{n,k,t}$ is $True$ (or equivalently, $\xi_{n,k,t} = 1$), subchannel $k \in \mathcal{K}$ will be assigned to the n^{th} platoon at time index t for either V2V or V2I communication. Following from this, the received instantaneous signal-to-interference-plus-noise ratio (SINR) on subchannel k at time index t of these two kinds of communication can be expressed as

$$\mathrm{SINR}^{\langle \mathrm{V2V_n} \rangle}_{n,k,t} = \frac{\mu_{n,t}\xi_{n,k,t}P_{n,k,t}H^{\langle \mathrm{V2V_n} \rangle}_{n,k,t}}{\sigma^2 + \sum\limits_{n',n'\neq n} \xi_{n',k,t}P_{n',k,t}H^{\langle \mathrm{V2X} \rangle}_{n',k,t}}, \quad (1)$$

$$\mathrm{SINR}^{\langle \mathrm{V2I} \rangle}_{n,k,t} = \frac{(1-\mu_{n,t})\xi_{n,k,t}P_{n,k,t}H^{\langle \mathrm{V2I} \rangle}_{n,k,t}}{\sigma^2 + \sum\limits_{n',n'\neq n} \xi_{n',k,t}P_{n',k,t}H^{\langle \mathrm{V2X} \rangle}_{n',k,t}}, \quad (2)$$

which is calculated by considering the accumulated interference of all other platoons as noise. $\langle \mathrm{V2X} \rangle$ refers to either $\langle \mathrm{V2V_n} \rangle$ or $\langle \mathrm{V2I} \rangle$, $\mathrm{SINR}^{\langle \mathrm{V2V_n} \rangle}_{n,k,t}$ pertains to the V2V link between PL$_n$ and its PM $\mathrm{V_n} \in [1, v_n]$, $\mathrm{SINR}^{\langle \mathrm{V2I} \rangle}_{n,k,t}$ pertains to the V2I link between PL$_n$ and the RSU, $P_{n,k,t}$ is the power consumed by PL$_n$, σ^2 is the noise level, and $H_{n,k,t}$ represents the channel gain that can be written as

$$H_{n,k,t} = g_{n,t}h_{n,k,t}, \quad (3)$$

where $g_{n,t}$ and $h_{n,k,t}$ are the large and small scale fading, respectively. Based on the Shannon-Hartley theorem, the maximum achievable rates of the V2V and V2I links are

$$\mathcal{C}^{\langle \mathrm{V2V_n} \rangle}_{n,k,t} = W \log_2 \left(1 + \mathrm{SINR}^{\langle \mathrm{V2V_n} \rangle}_{n,k,t}\right), \quad (4)$$

$$\mathcal{C}^{\langle \mathrm{V2I} \rangle}_{n,k,t} = W \log_2 \left(1 + \mathrm{SINR}^{\langle \mathrm{V2I} \rangle}_{n,k,t}\right). \quad (5)$$

At each single coherence time Δt, the RSU determines the allocation of subchannels, and PLs select their appropriate communication modes based on the current network conditions and communication requirements. The primary objectives are to optimize system performance by maximizing data rates and ensuring reliable and timely communications for both safety and coordination purposes, which are affected by the achievable V2V and V2I communication rates as follows:

- *V2V communication*: Same-platoon vehicles, the PL and its PMs, exchange CAMs periodically through the assigned subchannel via V2V links. This communication is essential for maintaining platoon string stability, allowing vehicles to keep restrained distances and be aware of the movements and decisions of the PL and other PMs. Based on [4], [5], the CAM generation interval should be kept between 100 ms and 1000 ms, hence an update interval $T \in [100, 1000]$ ms is considered. Within the interval T, a successful CAM exchange is defined as
$$\sum_{t=1}^{T/\Delta t} \sum_{k} \min_{\mathrm{V_n}} \left\{\mathcal{C}^{\langle \mathrm{V2V_n} \rangle}_{n,k,t}\right\} \Delta t \geq M_n, \quad (6)$$
where $T/\Delta t$ is designed to be an integer and M_n is the CAM message size. This constraint sets a minimum limit to the communication rate between PL$_n$ and each of its PMs in order to transmit the message of size M_n within the given time period T.
- *V2I communication*: PLs communicate with the RSU via V2I links to update platoon state information and receive control commands. This mode is crucial for PLs to be informed about the statuses of other platoons and the overall traffic conditions at the intersection. The concept

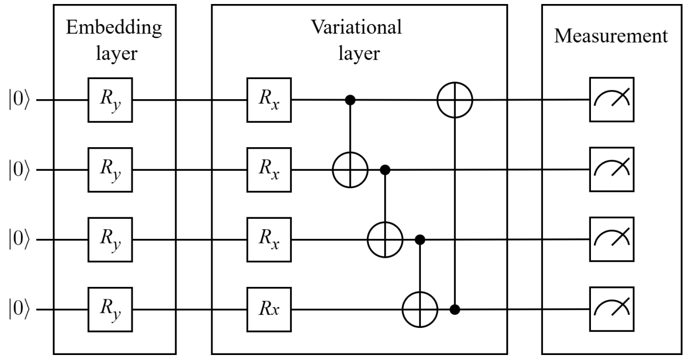

Fig. 2: Architecture of the VQC.

of AoI is formulated here to quantify the frequency of the information exchange between PLs and the RSU:

$$A_{n,t+1} = \begin{cases} A_{n,t} + \Delta t, & \text{if } \mathcal{C}_{n,k,t}^{\langle \mathrm{V2I} \rangle} < \mathcal{C}_{\min}^{\langle \mathrm{V2I} \rangle}, \\ \Delta t, & \text{otherwise}, \end{cases} \quad (7)$$

where $\mathcal{C}_{\min}^{\langle \mathrm{V2I} \rangle}$ is the minimum required V2I communication rate, and AoI is modelled as a discrete variable with Δt representing the smallest increment. This formulation accounts for the V2I communication rate. If the rate is lower than the minimum requirement or the V2I mode is not chosen, the transmission is aborted, resulting in the AoI being incremented by Δt. Conversely, a successful transmission resets the AoI to Δt.

Drawing on the discussions described in the preceding parts, the optimisation problem for platoon n can be formulated as

$$\min_{\mu,\xi,P,E} \left\{ \mathbb{P}\left(\sum_{t=1}^{T/\Delta t} \sum_{k} \min_{\mathrm{V_n}} \left\{ \mathcal{C}_{n,k,t}^{\langle \mathrm{V2V_n} \rangle} \right\} \Delta t < M_n \right), \right.$$
$$\left. \frac{\Delta t}{T} \sum_{t=1}^{T/\Delta t} A_{n,t}, \frac{\Delta t}{T} \sum_{t=1}^{T/\Delta t} \sum_{k} P_{n,k,t}, \sum_{t=1}^{T/\Delta t} \sum_{k} E_{n,k,t} \right\}, \quad (8)$$

$$\text{s.t.} \quad P_{n,k,t} \in [0, P^{\max}], \forall n, k, t, \quad (8a)$$

$$\sum_{k} \xi_{n,k,t} \leq 1, \forall n, t, \quad (8b)$$

$$\sum_{n} \sum_{k} \xi_{n,k,t} \leq K, \forall t, \quad (8c)$$

where $\mathbb{P}(\cdot)$ represents probability, $E_{n,k,t} = P_{n,k,t} \Delta t$ is the energy consumed by PL_n for communications at time index t on subchannel k, constraint (8a) guarantees that the power consumption of each PL does not exceed the maximum available power $P^{\max}$, and constraints (8b) and (8c) restrict each PL to utilise at most one subchannel per time slot and ensure the total number of assigned subchannels to be less than K, respectively. This optimisation problem focuses on minimising the probability of a failed CAM exchange (i.e., maximising the successful exchange probability), average AoI, average power consumption, and overall energy consumption for platoon n within a single CAM generation interval T.

In the following sections, the hybrid quantum-classical MADRL approach is introduced to solve this complex mixed-integer nonlinear programming problem for N platoons.

III. Hybrid Quantum-Classical MADRL approach

A. Preliminaries of the MADRL Algorithm

A Markov Decision Process (MDP) [25] is used to model the interaction between the agents and the environment in the MADRL problem, which is composed of the state space $\mathcal{S}$, action space $\mathcal{A}$, transition probability $\mathcal{P}$, reward function $\mathcal{R}$, and discount factor γ.

1) State Space: The state space $\mathcal{S}_{n,t}$ observed by PL_n at time index t is defined as

$$\mathcal{S}_{n,t} = \left[\mathrm{SINR}_{n,k,t}^{\langle \mathrm{V2V_n} \rangle}, \mathrm{SINR}_{n,k,t}^{\langle \mathrm{V2I} \rangle}, A_{n,t}, M'_n, Pos_n, t \right], \quad (9)$$

where the two SINRs contain information regarding channel gains, noise levels, and interference levels, $M'_n \in [0, M_n]$ denotes the size of the remaining CAM message waiting for exchange, Pos_n is the position of platoon n, and the current time index t enables the PL_n to know the remaining time budget $T - t$.

2) Action Space: Four actions constitute the action space $\mathcal{A}_{n,t}$ for platoon n:

$$\mathcal{A}_{n,t} = [\mu_{n,t}, \xi_{n,k,t}, P_{n,k,t}, E_{n,k,t}]. \quad (10)$$

PLs and the RSU affect the overall system performance by strategically selecting wireless subchannels and choosing appropriate communication modes, power levels, and energy levels. These actions comply with constraints (8a)–(8c).

3) Transition Probability: Originally, the transition probability $\mathcal{P}$ represents the probability that the current state s transfers to the next state s' when an agent takes action a. Here, this probability accounts for two primary factors: firstly, the subchannel, communication mode, and power/energy selections lead to the change in interference levels; secondly, the random platoon-turning decisions—whether to turn right/left or keep straight—occur independently of the four actions.

4) Reward Function: The reward function $\mathcal{R}$ is built on the principle that each PL should not only maintain frequent and successful message exchanges with the RSU and its PMs while minimising power and energy consumption, but also select subchannels and power levels that will reduce the interference to other platoons. Based on this principle, two reward functions, a global reward that evaluates the collective performance of all PLs and a local reward that provides immediate feedback for each PL's actions, are designed as

$$\mathcal{R}_t^g = -\frac{1}{N} \sum_{n} \sum_{k} \log_{10} \left(\xi_{n,k,t} P_{n,k,t} H_{n,k,t}^{\langle \mathrm{V2X} \rangle} \right), \quad (11)$$

$$\begin{aligned} \mathcal{R}_{n,t}^l = & -\omega_1 \frac{M'_n}{M_n} - \omega_2 A_{n,t} + \omega_3 \mathbf{1}_{\{\mathcal{C}_{n,k,t}^{\langle \mathrm{V2I} \rangle} - \mathcal{C}_{\min}^{\langle \mathrm{V2I} \rangle} \geq 0\}} \\ & - g_1 (P_{n,k,t}) - g_2 \left(\sum_{y=t-\tau}^{t} \rho^{t-y} E_{n,k,y} \right), \end{aligned} \quad (12)$$

where $\langle \mathrm{V2X} \rangle$ indicates that the communication mode selection of each PL influences the global reward, the coefficients ω_1–ω_3 and mapping functions g_1 and g_2 adjust and weight the five terms appropriately, the stepwise function $\mathbf{1}_{\{\cdot\}}$ provides positive feedback whenever a successful V2I communication is achieved, and the accumulated and discounted energy consumption is considered, with up to τ previous time indexes

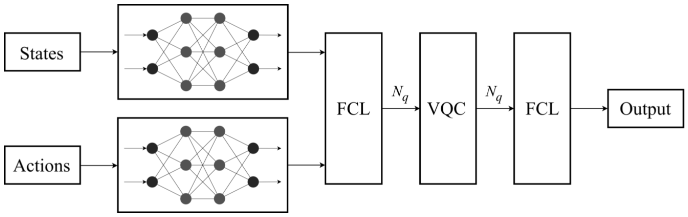

Fig. 3: Hybrid quantum-classical neural network.

and the discount factor ρ. The global reward considers the average interference level within the environment to promote the channel selections which decrease the disruption to other PLs, while the local reward matches the objectives outlined in problem (8), aiming to optimise the performance of each PL.

5) Discount Factor: The target of a reinforcement learning (RL) problem is to maximise the expected value of the discounted return, hence problem (8) is equivalent to

$$\max_{\pi_n} \left\{ V^{\pi_n}(s) = \mathbb{E}_{\pi_n}\left[\sum_{y=0}^{\infty} \gamma^y \mathcal{R}_{n,t+y+1} \middle| s_{n,t} = s \right] \right\}, \quad (13)$$

for $\forall s \in \mathcal{S}_{n,t}$, where $V^{\pi_n}(s)$ is the state-value function, policy π_n is a conditional probability $\pi_n(a_{n,t}|s_{n,t})$ that $a_{n,t}$ will be taken if $s_{n,t}$ is observed, and the discount factor $\gamma \in [0,1]$.

In the MDP formulated above, each PL functions as an agent within the MADRL environment, observes the current state, and selects actions based on its policy at each time index t. Through these interactions, rewards are obtained, and the policy is subsequently updated towards the direction of maximising the state-value function.

In addition, based on the Bellman optimality equation [25], the state-value function with the optimal policy is equivalent to the expected return from the state with the best action, i.e., the optimal action-value function. The action-value function—the Q-function—is written as

$$Q^{\pi_n}(s,a) = \mathbb{E}_{\pi_n}\left[\sum_{y=0}^{\infty} \gamma^y \mathcal{R}^{t+y+1} \middle| s_{n,t}, a_{n,t} \right]. \quad (14)$$

In the following sections, the deep deterministic policy gradient (DDPG) based algorithm is introduced with the objective of jointly optimising both the policy and the Q-function.

B. Decomposed MATD3 Algorithm

Building on the preliminaries of the MADRL algorithm, a combination of the single-agent DDPG and MADDPG, namely DE-MADDPG [21], is implemented. By combining it with the TD3 algorithm [22] which utilises two critic networks and delays the local network update by d loops to avoid the overestimation of the Q-functions, the policy gradient of the DE-MATD3 algorithm for the n^{th} agent is written as

$$\begin{aligned}\nabla \mathcal{J}_n(\theta_n) = & \overbrace{\mathbb{E}_{\boldsymbol{s},\boldsymbol{a}\sim\mathcal{D}}\left[\nabla_{\theta_n}\pi_n\left(a_n|s_n\right)\nabla_{a_n}Q^g_{\psi_1}(\boldsymbol{s},\boldsymbol{a})\right]}^{\text{MATD3}} \\ & + \overbrace{\mathbb{E}_{s_n,a_n\sim\mathcal{D}}\left[\nabla_{\theta_n}\pi_n\left(a_n|s_n\right)\nabla_{a_n}Q^{\pi_n}_{\phi_n}\left(s_n,a_n\right)\right]}^{\text{DDPG}},\end{aligned} \quad (15)$$

where $\mathcal{J}_n(\theta_n)$ is the target function that comprises both the global and local Q-functions, $\boldsymbol{s} = (s_1, ..., s_N)$ and $\boldsymbol{a} = (a_1, ..., a_N)$ denote the states and actions of the N platoons, $\mathcal{D}$ is the replay buffer, and θ_n, ψ_1, and ϕ_n parameterise the policy π_n for agent n, $Q^g_{\psi_1}$ for global critic, and $Q^{\pi_n}_{\phi_n}$ for local critic, respectively. Additionally, one of the twin critics from the TD3 algorithm, $Q^g_{\psi_1}$, is used to update the policy.

The fundamental concept of DE-MATD3 involves implementing the single-agent DDPG algorithm locally for each agent and integrating it with the centralised global critic to optimise the overall performance of all agents with MATD3 algorithm. The twin global critics $Q^g_{\psi_1}$ and $Q^g_{\psi_2}$ and the local critic $Q^{\pi_n}_{\phi_n}$ are updated by minimising the loss functions:

$$\mathcal{L}(\psi_i) = \mathbb{E}_{\boldsymbol{s},\boldsymbol{a},r^g,\boldsymbol{s}'}\left[\left(Q^g_{\psi_i}(\boldsymbol{s},\boldsymbol{a}) - y^g\right)^2\right], i = 1,2, \quad (16)$$

$$\mathcal{L}_n(\phi_n) = \mathbb{E}_{s_n,a_n,r^l_n,s'_n}\left[\left(Q^{\pi_n}_{\phi_n}\left(s_n,a_n\right) - y^l_n\right)^2\right], \quad (17)$$

$$y^g = r^g + \gamma \min_i Q^g_{\psi'_i}\left(\boldsymbol{s}',\boldsymbol{a}'\right)\Big|_{a'_n=\pi'_n(s'_n)}, \quad (18)$$

$$y^l_n = r^l_n + \gamma Q^{\pi_n}_{\phi'_n}\left(s'_n,a'_n\right)\Big|_{a'_n=\pi'_n(s'_n)}, \quad (19)$$

where $\boldsymbol{s}' = (s'_1, ..., s'_N)$ and $\boldsymbol{a}' = (a'_1, ..., a'_L)$ denote next sets of states and actions, while $Q^g_{\psi'_i}$, $Q^{\pi_n}_{\phi'_n}$, and π'_n are the target global critics, target local critic, and target policy, respectively.

C. Hybrid Quantum-Classical DE-MATD3 Algorithm

For each step of the MDP, a set of states, actions, and rewards, $\left(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{r}^l_t, r^g_t, \boldsymbol{s}_{t+1}\right)$, is stored in the experience replay buffer $\mathcal{D}$. After that, B transitions, $\left(\boldsymbol{s}_b, \boldsymbol{a}_b, \boldsymbol{r}^l_b, r^g_b, \boldsymbol{s}'_b\right)\Big|_{b=1}^{B}$, are randomly sampled from $\mathcal{D}$ for the training purposes, i.e., fed into our hybrid quantum-classical neural network.

The quantum part of our neural network, the QNN, is implemented using a VQC [23], the architecture of which is illustrated in Fig. 2. The VQC comprises three distinct layers: data embedding, variational, and measurement. The initial states of the quantum circuit are set to $|0\rangle \otimes \cdots \otimes |0\rangle$, consisting of N_q qubits. The input classical data vector $\mathbf{x} = (x_1, x_2, \ldots, x_{N_q})$ is embedded into quantum states through angle embedding, using rotation gates R_y. Within the variational layer, single-qubit rotation gates R_x are applied, each parameterised by a unique angle that is updated during the training process. This is followed by a ring of controlled-NOT (CNOT) gates, which establishes multi-qubit entanglement by connecting each qubit to its neighbouring qubit, with the final qubit linked back to the first. Quantum measurements are performed using Pauli-Z gates, and an array consisting of N_q values is output from this quantum circuit.

Our hybrid quantum-classical neural network is shown in Fig. 3. Classical data that contains state and action values is fed into two classical neural networks separately. The outputs of the two neural networks are added and fed into a fully connected layer (FCL) that maps the hidden data size to the VQC input dimension N_q. After the quantum computation, the resulting measurement outcomes are fed into another FCL to match the desired output size. The final hybrid quantum-classical DE-MATD3 scheme is described in Algorithm 1.

Algorithm 1: Hybrid Quantum-Classical DE-MATD3

```
Initialise intersection environment & experience replay buffer $\mathcal{D}$.
Initialise local and global actor-critic networks:
    $\{\pi_n, \pi'_n, Q^{\pi_n}_{\phi_n}, Q^{\pi_n}_{\phi'_n}\}, n = 1, 2, ..., N$, $\{Q^g_{\psi_i}, Q^g_{\psi'_i}\}, i = 1, 2$.
Initialise the quantum circuit, reset qubits.
for episode $= 1$ to loop do
    Update platoon position & channel information.
    Reset time index & CAM size: $\{t, M'_n\} = \{1, M_n\}$.
    for $t = 1$ to $T$ do
        for agent 1 to $N$ do
            Observe state $s_{n,t}$, select action $a_{n,t}$ based on policy
              $\pi_n(a_{n,t}|s_{n,t})$, receive rewards: $\{r^l_{n,t}, r^g_t\}$.
        Update interference & channel fast fading.
        Each agent observes a new state $s_{n,t+1}$.
        Store $\left(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{r}^l_t, r^g_t, \boldsymbol{s}_{t+1}\right)$ into the buffer $\mathcal{D}$.
        Randomly sample $B$ transitions from $\mathcal{D}$:
          $\left(\boldsymbol{s}_b, \boldsymbol{a}_b, \boldsymbol{r}^l_b, r^g_b, \boldsymbol{s}'_b\right)\big|^B_{b=1}$.
        Pass through the hybrid quantum-classical neural network
          for the global critic network:
        1) Update global critics: minimising $\mathcal{L}(\psi_i)$ (16) by one-step
          gradient descent,
        2) Target soft update: $\psi'_i \leftarrow \epsilon\psi_i + (1-\epsilon)\psi'_i$.
        if $t$ mod $d$ then
            Pass through the hybrid quantum-classical neural
              network for the local actor-critic networks:
            for agent 1 to $N$ do
                1) Update local critic: minimising $\mathcal{L}_n(\phi_n)$ (17) by
                  one-step gradient descent,
                2) Update local actor: maximising $\nabla\mathcal{J}_n(\theta_n)$ (15)
                  by one-step gradient ascent,
                3) Target soft update:
                  $\phi'_n \leftarrow \epsilon\phi_n + (1-\epsilon)\phi'_n$
                  $\theta'_n \leftarrow \epsilon\theta_n + (1-\epsilon)\theta'_n$
```

IV. Results and Discussion

We showcase the simulation outcomes and compare results obtained with the hybrid quantum-classical DE-MATD3 algorithm with results obtained using the classical DE-MATD3 algorithm in [10], [11]. The simulations consider a single-cell urban C-V2X network operating at 2 GHz, utilizing 3 resource blocks, and conforming to the 3GPP TR 36.885 urban specification [5]. We use Python with PyTorch and PennyLane to build the algorithm framework. Our hybrid quantum-classical neural network consists of two and three layers for the local critics and actors respectively, and four layers for the global critics. We have selected the rectified linear unit as the activation function and Adam as the optimiser. The other key simulation parameters are provided in Table I.

Fig. 4 shows the reward convergence of our proposed hybrid quantum-classical algorithm (Q-E-DE-MATD3) in comparison with the classical energy-focused algorithm (E-DE-MATD3) presented in [10] and the non-energy-focused algorithm (DE-MATD3) presented in [10], [11]. All three algorithms exhibit similar convergence speeds and reward levels, however, slight variations in the final converged reward levels are observed due to differing reward function designs. To ensure a fair comparison, Fig. 5 evaluates the performance in terms of AoI. The proposed hybrid quantum-classical algorithm maintains comparable convergence speed and AoI level relative to the two benchmark algorithms.

TABLE I: Simulation Parameters

Parameters	Values
Number of platoons	4
Vehicles in each platoon	4
V2V distance	25 m
Resource block bandwidth	180 kHz
PL maximum power	$P^{max} = 30$ dBm
Noise power	$\sigma^2 = -114$ dBm
CAM message size	$M_n = 4$ KB
Update interval	$T = 100$ ms [4], [5]
Fast fading update period	$\Delta t = 1$ ms [5]
Slow fading update period	100 ms [5]
Number of episodes	500
Iterations in each episode	100
Actor learning rate	0.0001
Critic learning rate	0.001
Target soft update	$\epsilon = 0.005$
Discount factor	$\gamma = 0.99$
Number of qubits	$N_q = 4$

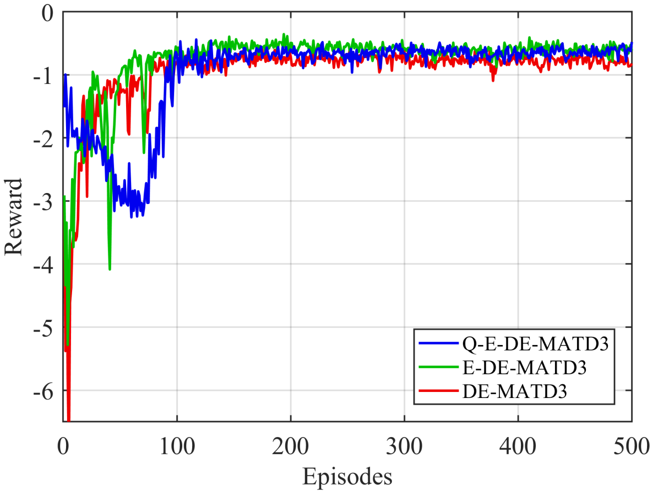

Fig. 4: Reward convergence.

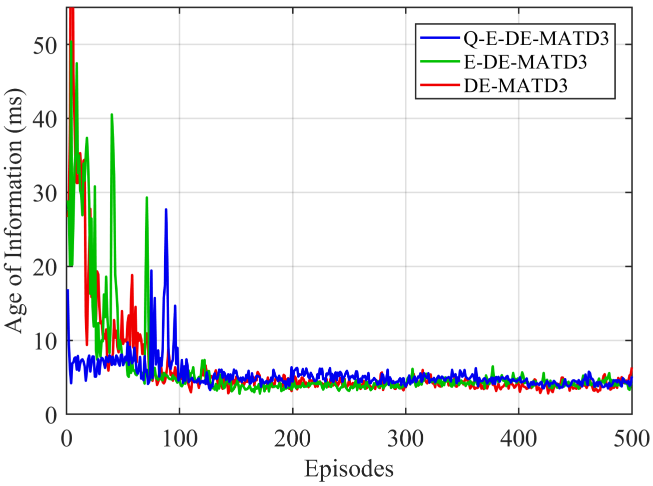

Fig. 5: AoI convergence.

Table II presents the average values of AoI, reward, energy consumption, and energy consumed per AoI reduction, which is the cost metric calculated as $E/(T - \text{AoI})$, over the last 100 episodes. Our proposed hybrid quantum-classical algorithm

TABLE II: Performance Metrics

Metric	Q-E-DE-MATD3	E-DE-MATD3	DE-MATD3
AoI (ms)	4.24	4.22	3.94
Increase (%)	7.55	7.01	N/A
Energy (mJ)	153.32	163.59	1064.42
Decrease (%)	85.60	84.63	N/A
Cost	1.60	1.71	11.08
Decrease (%)	85.55	84.59	N/A

achieves a comparable reduction in energy consumption to the energy-focused algorithm, while attaining an 85.60% decrease relative to the non-energy-focused counterpart. Although this results in a 7.55% increase in the average AoI, the cost metric demonstrates superior performance, achieving an 85.55% decrease, further highlighting the sustainability of our algorithm.

V. Conclusion

This paper introduced a hybrid quantum-classical DRL-based optimal resource allocation strategy for a platoon-based C-V2X network operating at an intersection. Our proposed hybrid quantum-classical DE-MATD3 scheme was built based on the DE-MADDPG algorithm, incorporating the TD3 technique and the novel quantum computing technology VQC. It was designed to jointly optimise the AoI, CAM exchange, and power-energy consumption. The proposed algorithm was established within a collaborative environment, enabling all the platoons to concurrently optimize both a shared global reward and their individual local rewards. Furthermore, the local reward function design incorporated accumulated and discounted energy consumption to specifically enhance the long-term sustainability of the C-V2X network. Simulation results demonstrated the remarkable potential for quantum computing in dealing with complicated resource allocation problems. Future research will explore more practical scenarios, such as managing a dynamic number of platoons at intersections, to better align the system model with real-world conditions. Additionally, we aim to implement a neural network consisting of only quantum circuits to further investigate the potential of quantum computing in next-generation communication systems.

Acknowledgement

This work was supported in part by the Canada Excellence Research Chair (CERC) Program CERC-2022-00109 and in part by the Canada Research Chair Program CRC-2022-00187.

References

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

[2] C. Chen, Y. Zhang, M. R. Khosravi, Q. Pei, and S. Wan, "An intelligent platooning algorithm for sustainable transportation systems in smart cities," *IEEE Sens. J.*, vol. 21, no. 14, pp. 15 437–15 447, Jul. 2021.

[3] A. Zanella *et al.*, "Internet of things for smart cities," *IEEE Internet Things J.*, vol. 1, no. 1, pp. 22–32, Feb. 2014.

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

[5] 3GPP, "Study on LTE-based V2X services," 3rd Generation Partnership Project (3GPP), Technical Report (TR) 36.885, 2016, version 14.0.0.

[6] ——, "Study on enhancement of 3GPP support for 5G V2X services," 3rd Generation Partnership Project (3GPP), Technical Report (TR) 22.886, 2018, version 16.2.0.

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

[8] Z. Liu *et al.*, "Automated vehicle platooning: A two-stage approach based on vehicle-road cooperation," *IEEE Trans. Intell. Veh.*, Aug. 2024, DOI: 10.1109/TIV.2024.3448501.

[9] C. Wu, Z. Cai, Y. He, and X. Lu, "A review of vehicle group intelligence in a connected environment," *IEEE Trans. Intell. Veh.*, vol. 9, no. 1, pp. 1865–1889, Jan. 2024.

[10] Y. Zheng *et al.*, "Multi-agent deep reinforcement learning for optimal resource allocation in AoI-aware energy-efficient platoon-based C-V2X systems," in *Proc. 2024 IEEE 29th Int. Workshop Comput. Aided Model. Des. Commun. Links Netw. (CAMAD)*, Athens, Greece, Oct. 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] 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.

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

[14] H. Lin and C. Tang, "Intelligent bus operation optimization by integrating cases and data driven based on business chain and enhanced quantum genetic algorithm," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 9869–9882, Jul. 2022.

[15] U. Azad, B. K. Behera, E. A. Ahmed, P. K. Panigrahi, and A. Farouk, "Solving vehicle routing problem using quantum approximate optimization algorithm," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 7, pp. 7564–7573, Jul. 2023.

[16] S. Xu *et al.*, "Post-quantum anonymous, traceable and linkable authentication scheme based on blockchain for intelligent vehicular transportation systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 9, pp. 12 108–12 119, Sep. 2024.

[17] H. Khalid *et al.*, "RAVEN: Robust anonymous vehicular end-to-end encryption and efficient mutual authentication for post-quantum intelligent transportation systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 11, pp. 17 574–17 586, Nov. 2024.

[18] Z. Yang, M. Zolanvari, and R. Jain, "A survey of important issues in quantum computing and communications," *IEEE Commun. Surv. Tutor.*, vol. 25, no. 2, pp. 1059–1094, Second Quart. 2023.

[19] B. Narottama and T. Q. Duong, "Quantum neural networks for optimal resource allocation in cell-free MIMO systems," in *GLOBECOM 2022 - 2022 IEEE Glob. Commun. Conf.*, Rio de Janeiro, Brazil, Dec. 2022, pp. 2444–2449.

[20] B. Mao, F. Tang, Y. Kawamoto, and N. Kato, "AI models for green communications towards 6G," *IEEE Commun. Surv. Tutor.*, vol. 24, no. 1, pp. 210–247, First Quart. 2022.

[21] 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.

[22] 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.

[23] W. J. Yun *et al.*, "Quantum multi-agent reinforcement learning via variational quantum circuit design," in *2022 IEEE 42nd Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Bologna, Italy, Jul. 2022, pp. 1332–1335.

[24] Q. Shi, L. Liu, S. Zhang, and S. Cui, "Device-free sensing in OFDM cellular network," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 6, pp. 1838–1853, Jun. 2022.

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