

Sum Secrecy Rate Optimization in UAV Communications Using Quantum Actor-Critic Reinforcement Learning

Su Fong Chien^{* ††}, Samuel Yen-Chi Chen^{†‡‡}, Mau Luen Tham^{†^x}, Heng Siong Lim^{§^{xi}},
Charilaos C. Zarakovitis^{xv xii}, Michail A. Kourtis^{||^{xiii}}, Yi Jie Wong^{†^{xiv}}

^{*}Department of Strategic ICT, MIMOS Berhad, Kuala Lumpur, Malaysia

[†]Computational Science Initiative, Brookhaven National Laboratory, Upton, NY, USA

[‡]Department of E&E Engineering, Universiti Tunku Abdul Rahman, Kajang, Malaysia

[§]Faculty of Engineering and Technology, Multimedia University, Melaka, Malaysia

^{xv}Axon Logic, Multidisciplinary Research and Innovation Centre, Athens, Greece

^{||}Media Networks Laboratory, Institute of Information and Telecommunications, NCSR, Greece

Email: ^{††} sf.chien@mimos.my, ^{‡‡} ycchen1989@ieee.org, ^x thamml@utar.edu.my, ^{xi} hslim@mmu.edu.my,
^{xii} c.zarakovitis@axonlogic.gr, ^{xiii} akis.kourtis@iit.demokritos.gr, ^{xiv} yjwong1999@utar.my

Abstract—Reinforcement learning (RL) has proven effective in wireless tasks like dynamic spectrum access and power control. Extending this, Quantum Reinforcement Learning (QRL) addresses quantum-specific challenges such as channel estimation and multi-agent UAV communications. We propose a hybrid Quantum Deep RL (QDRL) framework to optimize the average sum secrecy rate (SSR) in a mmWave UAV system with a Reconfigurable Intelligent Surface (RIS), under imperfect CSI and multiple eavesdroppers. Based on the DDPG framework, we explore quantum variants like QTDDPG and QTDD3. Simulations show QDRL achieves comparable SSR to classical methods with far fewer parameters. Notably, even a simple QTDDPG with a single-layer quantum encoder performs well, underscoring the promise of efficient quantum policies for secure wireless communication.

Index Terms—Variational quantum circuit, Quantum reinforcement learning, Millimeter-wave communications, Unmanned aerial vehicle, Reconfigurable intelligent surface

I. INTRODUCTION

Unmanned Aerial Vehicles (UAVs), have found various applications in wireless communications because of their mobility, flexibility, and ability to reach locations that would otherwise be difficult to access [1]. Due to the limited battery life, the pursuit of strategies to achieve maximum data rates, optimal trajectory, and highest energy efficiency under certain constraints has become a prominent area of research interest, as highlighted in references [2], [3]. Recent research by Li et al. has shown that in situations where complete channel state information (CSI) is unavailable for UAV communication systems, the combination of DRL, which incorporates Deep Neural Networks (DNN) and RL, emerges as a promising solution for addressing real-time dynamic optimization challenges, as referenced in [4]. In essence, DRL can be divided into two tasks: the first involves discrete-level control, exemplified by

Deep Q-Learning Network (DQN) [5], and the second pertains to continuous-level control, as seen in algorithms like DDPG [6] and Twin-Delayed Deep Deterministic Policy Gradient (TD3) [7].

This study explores a wireless communication setup that involves millimeter-wave UAVs working in conjunction with a RIS, characterized by imperfect CSI, with multiple users, and a presence of eavesdroppers, which is depicted in Figure 1. The RIS is equipped with a substantial array of passive reflecting elements, enabling it to achieve impressive levels of spectral and energy efficiency in an economically advantageous manner. Guo et. al have recently proposed DRL as a viable approach to jointly design the active and passive beamforming, and the UAV trajectory, due to its good generalization, low complexity, and high accuracy characteristics. They show that this idea results a better performance compared with several benchmarks. Hence, this motivates us to explore a quantum solution for the development of qDRL. A prevalent application of hybrid quantum-classical algorithms involves the integration of a variational quantum circuit (VQC) with conventional classical computing methods [8], [9], and [10]. The quantum machine learning outperforms the classical counterparts [11], [12] when certain conditions are met. In the context of this research, we introduce the concepts of hybrid quantum-classical QTDDPG and quantum QTDD3 as our proposed approaches. In order to gain a deeper insight into the influence of angle encoding structures on the performance of SSR, this paper examines both single-layer and multi-layer variational encoding structure.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. System Model

Figure 1 illustrates a communication system for UAVs incorporating a hybrid approach, combining quantum techniques,

This work was supported by the EU Horizon Europe projects OASEES (ID:101092702) and P2CODE (ID:101093069).

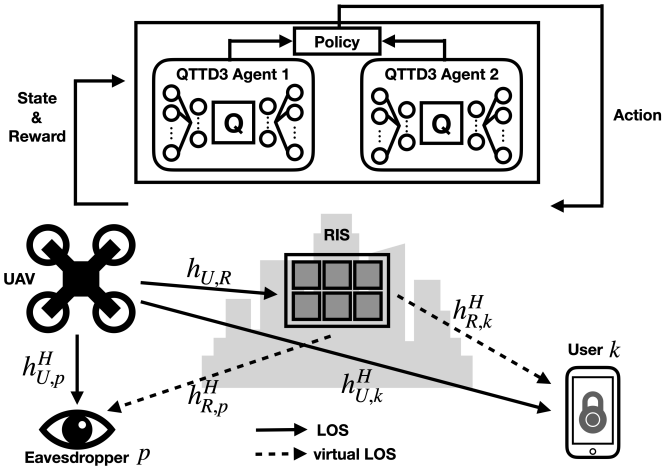


Fig. 1: RIS-aided mmWave UAV communications

DRL, and RIS to enhance mmWave communication. With the virtual Line-of-Sight channels (LOS), the RIS can enhance the link security from UAV to K single-antenna authorized users, in the presence of E potential eavesdroppers. Note that the eavesdroppers in this scenario are also characterized as single-antenna entities. The RIS is equipped with a uniform planar array (UPA) containing $M = m^2$ passive reflecting elements, where m is an integer. Additionally, the UAV features a uniform linear array (ULA) with A elements. The set of the users and the eavesdroppers are represented by $\mathcal{K} = \{1, 2, \dots, K\}$, $\mathcal{E} = \{1, 2, \dots, E\}$, respectively. As depicted in Figure 1, all entities are positioned within the three-dimensional (3D) Cartesian coordinate system. The RIS is located at coordinate $\mathbf{w}_R(x, y, z)^T$. At time instant l , the coordinates of the UAV and the authorized users are denoted as $\mathbf{q}_l = (x_l^U, y_l^U, z_l^U)^T$, while the coordinates for the eavesdroppers are represented by $\mathbf{w}_i^l = (x_i^l, y_i^l, z_i^l)^T$, for all $i \in \mathcal{K} \cup \mathcal{E}$. Hence, the location information at the n -th time slot is defined as $\mathbf{W} \triangleq \{\mathbf{q}_l\} \cup \{\mathbf{w}_i^l | \forall i \in \mathcal{K} \cup \mathcal{E}\}$. In this study, we assume that the UAV maintains a constant altitude throughout a finite time span, which divides the entire flight duration T into L time slots evenly. Each time slot is of length Δ_t , i.e., $t = l\Delta_t$ for $l \in L$. Let \mathbf{q}_0 represents the initial coordinates of the UAV, its operational boundary of size B , and the movement is constrained within a maximum distance of D_{max} at time slot l . Therefore, the UAV mobility constraints can be written as

$$\mathbf{q}_0 = (0, 0, H_U), l = 1, \dots, L-1 \quad (1a)$$

$$|x_l|, |y_l| \leq B, l = 1, \dots, L \quad (1b)$$

$$\sqrt{\|\mathbf{q}_l - \mathbf{q}_{l-1}\|^2} \leq D_{max}, l = 1, \dots, L-1, \quad (1c)$$

The channel model used in this simulation can be referred to [2] and the detailed 3D SV channel model that has been widely used to characterize the mmWave channels is described in [13]. The total attainable data rate at the k^{th} user is expressed as

$$R_k^U = \log_2(1 + (\frac{|\mathbf{h}_{U,k}^H + \Xi^H \mathbf{H}_{C,k})g_k|^2}{\sum_{k' \in \mathcal{K} \setminus k} |\mathbf{h}_{U,k'}^H + \Xi^H \mathbf{H}_{C,k'})g_{k'}|^2 + n_k^2}) \quad (2)$$

and the attainable rate for this eavesdropping scenario can be determined as

$$R_{e,k}^E = \log_2(1 + (\frac{|\mathbf{h}_{U,e}^H + \Xi^H \mathbf{H}_{C,e})g_k|^2}{\sum_{k' \in \mathcal{K} \setminus k} |\mathbf{h}_{U,k'}^H + \Xi^H \mathbf{H}_{C,e})g_{k'}|^2 + n_e^2}) \quad (3)$$

Refer to [14], the achievable individual secrecy rate from the UAV to the user k^{th} is given as $R_k^{sec} = [R_k^u - \max_{\forall e} R_{e,k}^E]^+$, where $[x]^+ = \max(0, x)$. In reality, the UAV faces challenges in obtaining perfect CSI due to transmission delays, processing delays, as well as the mobility of both the UAV and the users. As a result, a new formulation has been developed to calculate an estimated CSI that takes into account the time delay factor in order to compute achievable secrecy rates [2].

B. Problem Statement

The objective of this paper is to maximize the sum secrecy rate $\sum_{k=1}^K R_k^{sec}$ by concurrently optimizing the UAV's trajectory $\mathbf{Q} \triangleq \{\mathbf{q}_l, l = 1, 2, \dots, L\}$ and with the active(passive) matrix $\mathbf{G}(\Theta)$. The optimization problem can be given as

$$\max_{\mathbf{Q}, \mathbf{G}, \Theta} \sum_{k \in \mathcal{K}} R_k^{sec} \quad (4a)$$

$$s.t. (1), \quad (4b)$$

$$Pr\{R_k^{sec} \geq R_k^{sec, th}\} \geq 1 - \rho_k, \forall k \in \mathcal{K}, \quad (4c)$$

$$Tr(\mathbf{G}\mathbf{G}^H) \leq P_{max}, \quad (4d)$$

$$\theta_m \in [0, 2\pi), m = \{1, 2, \dots, M\}, \quad (4e)$$

where the secrecy rate outage constraint in (5c) ensures the probability that each legitimate user can perfectly extract its message at a data rate of $R_k^{sec, th}$ with at least $1 - \rho_k$. Given the non-convex nature of the problem outlined by constraints 5(b), 5(c), and 5(e), along with the variability in CSI, conventional approaches are generally inadequate for addressing this probability-constrained problem. Consequently, employing a DRL-based actor-critic method emerges as a promising alternative solution to effectively address these challenges. The corresponding energy consumption under optimal achievable rate is termed as

$$E_{e,l} \approx \Delta_t \left(P_0 + \frac{3P_0 \|\mathbf{v}_l\|^2}{U_{tip}^2} + \frac{1}{2} d_0 \varsigma \rho A_r \|\mathbf{v}_l\|^3 \right) + \Delta_t P_i \left(\sqrt{1 + \frac{\|\mathbf{v}_l\|^4}{4v_0^4}} - \frac{\|\mathbf{v}_l\|^2}{2v_0^2} \right) \quad (5)$$

where $\|\mathbf{v}_l\| = \sqrt{\|\mathbf{q}_l - \mathbf{q}_{l-1}\|^2} / \Delta_t$, constants P_0 and P_i represent the blade profile power and induced power in holding status, respectively. U_{tip}^2 is rotor blade's tip speed, v_0 is the average rotor induced velocity in standing condition, d_0 is the fuselage drag ratio, and ς is the rotor solidity. Lastly, ρ and A_r denotes the air density and rotor disc area, respectively.

III. HYBRID QUANTUM DRL-BASED ACTOR-CRITIC METHOD

A. Reinforcement Learning

Reinforcement Learning (RL) is typically formulated as a Markov Decision Process (MDP), where agents lack full

knowledge of the environment and instead learn policies-mappings from states to actions-through interaction to maximize cumulative rewards. This framework has enabled RL to address various wireless communication problems, including UAV trajectory planning and resource allocation [15]–[18].

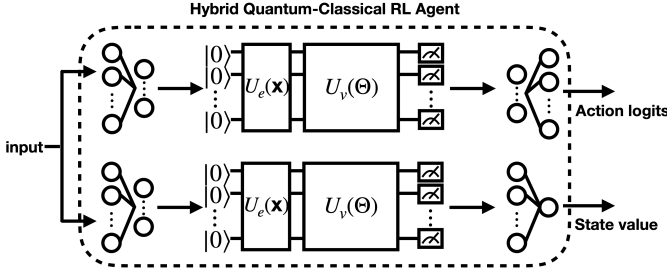


Fig. 2: Hybrid quantum-classical RL agent used in this work.

B. Hybrid Quantum Actor and Critic Networks

Prior studies [8], [9], [19] show that hybrid quantum-classical RL can outperform classical methods in complex sequential decision tasks under suitable conditions, making QRL a promising candidate for addressing Problem (4). Like classical DDPG, the quantum actor takes a state as input and outputs a probability distribution over actions to maximize expected rewards. The quantum critic takes a state-action pair (s_l, a_l) and outputs the action-value $Q(s_l, a_l)$. In this study, both the quantum actor and critic networks incorporate a *dressed* VQC. Each of these circuits comprises neural network layers for pre-processing or post-processing the data flowing to and from the VQC. The overall scheme is illustrated in Figure 2. The trainable circuit $U_v(\Theta)$ is composed of several repeated blocks, designed to augment the number of parameters, as illustrated in Figure 3. The incorporation of a dressed network significantly reduces the input dimensionality before presenting the input to the VQC. The network design selects a single dressing layer, with four repeated quantum blocks in the quantum circuits. As illustrated in Figure 4a and Figure 4b, these are 4-qubit VQCs with $R_y R_z R_x R_y$ encoders and 4×1 R_y encoders, each composed of $M = 4$ repetition blocks.

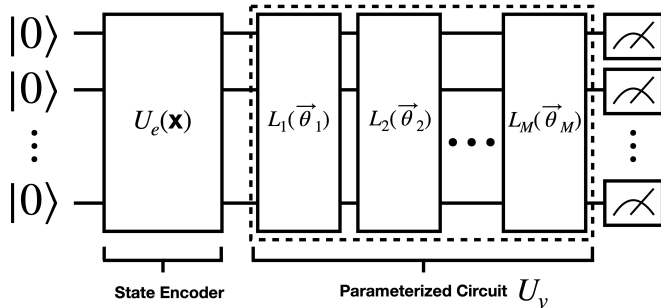


Fig. 3: Generic architecture for variational quantum circuits (VQC).

C. Variational Quantum Circuit

Variational Quantum Circuits (VQCs), or Parameterized Quantum Circuits (PQCs), are quantum circuits with learnable parameters optimized via classical algorithms using a loss function. The loss is computed classically, and updates can use gradient-based or other methods. As shown in Figure 3, a Variational Quantum Circuit (VQC) consists of three main modules: (1) the encoding module $U_e(\mathbf{x})$ maps classical input \mathbf{x} into quantum states; (2) the variational module $U_v(\Theta)$ entangles qubits and applies parameterized rotations across multiple layers, with $U_v(\Theta) = \prod_{i=1}^M L_i(\vec{\theta}_i)$; (3) the measurement module \mathcal{M} extracts expectation values, typically via Pauli- Z measurements over multiple shots. VQCs can be flexibly integrated with classical models (e.g., tensor networks, deep neural nets) for tasks such as circuit dressing and output post-processing [8]. The operation of the VQC used in this work can be written as $f(\mathbf{x}; \Theta) = (\langle \hat{Z}_1 \rangle, \dots, \langle \hat{Z}_q \rangle)$, where $\langle \hat{Z}_k \rangle = \langle 0 | U_e^\dagger(\mathbf{x}) U_v^\dagger(\Theta) \hat{Z}_k U_v(\Theta) U_e(\mathbf{x}) | 0 \rangle$ and q represents the total number of qubits measured in the system. The expectation values $\langle \hat{Z}_k \rangle$ can be derived analytically when the circuit is simulated classically.

IV. HYBRID QUANTUM DRL-BASED FRAMEWORK AS A SOLUTION

At this end, we propose a hybrid QDRL algorithms, i.e., QDDPG and QTDD3 to address problem (5). The first QDRL agent takes CSI i.e., \mathbf{H}_C as a state to get the optimal UAV active beamforming matrix \mathbf{G} and the RIS passive beamforming matrix Θ . Conversely, the second QDRL performs its task to obtain the best UAV trajectory \mathbf{Q} based on the local information \mathbf{W} , as the state to get the UAV movement that poses flying distance d_l and the direction ζ_l . Therefore, (5) can be formulated as a MDP that with the state, action, and reward as follows.

A. Active and Passive Beamforming

- 1) **State** $s_{l,1}$: In the l -th time slot, the state of the first QDRL agent encompasses the estimated comprehensive CSI from the UAV to all authorized users and eavesdroppers, i.e., \mathbf{H}_C . It is important to note that the UAV may not have prior knowledge of the small-scale component in \mathbf{h} . Real-time collection of small-scale information is feasible as part of the network's current status. Therefore, the proposed algorithm possesses the capability to adapt to environmental variations in an online manner, given its ability to acquire and utilize small-scale information dynamically.
- 2) **Action** $a_{l,1}$: The active beamforming matrix \mathbf{G} and the passive beamforming Θ are termed as action. Note that $\mathbf{G} = \text{Re}\{\mathbf{G}\} + \text{Im}\{\mathbf{G}\}$ and $\Theta = \text{Re}\{\Theta\} + \text{Im}\{\Theta\}$ are divided into real and imaginary parts in order to tackle with the real input problem.
- 3) **Reward** $r_{l,1}$: The reward function can be written as

$$r_{l,1} = \tanh(\sum_{k=1}^K R_k^{\text{sec}} - a_1 p_m - a_2 p_r - a_3 p_g), \quad (6)$$

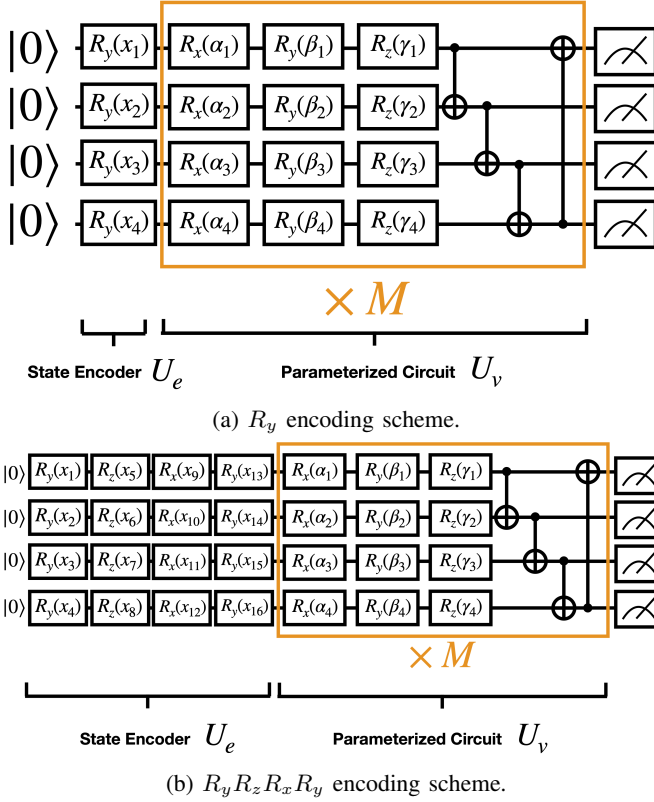


Fig. 4: Examples of encoder designs used in the proposed hybrid quantum-classical RL agent.

where p_m, p_r, p_g are the penalties when the constraints 5(b), 5(c), and 5(d) are not satisfied, respectively. The coefficients a_1, a_2 , and a_3 functioning as the weight that to balance the penalties and the sum secrecy rate. In accordance with the findings in [2], the expression $1 - PrR_k^{sec} \geq R_k^{sec,th}$ can be approximated as N_{outage}/N_{sample} . Here, N_{outage} corresponds to the number of samples where the secrecy rate R_k^{sec} is less than the specified threshold $R_k^{sec,th}$, and N_{sample} represents the total number of generated CSI samples.

B. UAV Trajectory

The second QDRL agent is used to compute the optimal UAV trajectory Q , by leveraging the local information W as input. Similarly, the problem can be formulated as an MDP with the following definitions for state, action and reward.

- 1) **State** $s_{l,2}$: The second QDRL agent only takes the local information W as input because of the UAV trajectory is strongly interconnected with the extensive set of CSI.
- 2) **Action** $a_{l,2}$: At each time slots l , the QDRL agent computes the flying distance d_l in 3D space and the flying direction as well. With this information, the next coordinate of the UAV can be obtained via $q_l = q_{l-1} + d_l$. After L time slots, the entire UAV trajectory can be computed as $Q = \{q_0, q_1, \dots, q_L\}$

- 3) **Reward** $r_{l,1}$: The reward function is identical to (7), as both networks are directed towards the same objective to maximize the sum secrecy rate.

Algorithm 1 QRL Algorithm

- 1: Initialize both quantum actor and critics networks, and both target quantum actor and critics networks for VQC Agent 1
- 2: Repeat the initialization processes as described above for VQC Agent 2
- 3: **for** episode $n_{ep} = 1, 2, \dots, N_{ep}$ **do**
- 4: Reset the coordinates of K users and E eavesdroppers $w_i, \forall i \in K \cup E$
- 5: Reset the coordinate of UAV q_0
- 6: **for** time slot $l = 1, 2, \dots, N_{step}$ **do**
- 7: Observe $s_{l,1} \leftarrow H_c$ and $s_{l,2} \leftarrow W$
- 8: Based on $s_{l,1}$, VQC Agent 1 selects actions $a_{l,1} = \{G, \Theta\}$
- 9: Based on $s_{l,2}$ VQC Agent 2 selects actions $a_{l,2} = d_l$
- 10: Execute actions $a_{l,1}$ and $a_{l,2}$ then receive rewards $r_{l,1}$ and $r_{l,2}$
- 11: Update UAV coordinate $q_l = q_{l-1} + d_l$
- 12: Get new states $s_{l+1,1}$ and $s_{l+1,2}$
- 13: Store the transitions $[s_{l,1}, a_{l,1}, r_{l,1}, s_{l+1,1}]$ and $[s_{l,2}, a_{l,2}, r_{l,2}, s_{l+1,2}]$ into memory buffer for learning
- 14: Sample mini batches from memory buffer to update quantum actor and quantum critic networks of both VQC Agent 1 and Agent 2
- 15: **end for**
- 16: **end for**

V. SIMULATION AND RESULT DISCUSSION

In this simulation, we assess the performance of the newly introduced QTDDPG and QTDD3 algorithms in comparison to classical approaches based on the TorchQuantum framework [20]. The learning rates for both actor and critic networks are configured as 0.0001 for the actor and 0.001 for the critic. All algorithms undergo training with 1000 episodes, each comprising 100 time slots. The initial positions of the UAV and the two users are defined as follows: the UAV starts at (0m, 25m, 50m), the first user at (47m, 4m, 0m), and the second user at (25m, 25m, 0m). The RIS and eavesdropper are situated at (0m, 50m, 12.5m) and (47m, -4m, 0m), respectively. Besides, we incorporate the movement patterns of the two users, allowing them to travel freely in straight lines as shown in Figure5. Other system parameters are $\Delta_t = 0.1ms, T_d = 1s, f_c = 28GHz, C_0 = 61dB, P_{max} = 30dBm, \sigma_n = -114dBm, L_y = 3, \alpha_{ur} = 2.2; \alpha_u = 3.5; \alpha_r = 2.8, \sigma_s = 3dB, A = 4, M = 16, E = 1, K = 2, \Phi_l^{AoD} = \{5, 10, 15, 25\}, \Phi_l^{AoA} = \{30, 45, 60\}, \Lambda_l^{AoD} = \{1, 3, 5\}, \Lambda_l^{AoA} = \{5, 10, 15\}$ (degrees) [2].

The network architectures for the classical DDPG and TD3 baseline are the following: first and second agents consist of four fully-connected hidden layers, with configurations of [128, 64, 64, 32] for the first agent and [64, 64, 8, 4] for the second one. The number of learnable parameters can be estimated by considering the example [128, 64, 64, 32]. It can be computed as $\text{input_dim} \times 128 + 128 \times 64 + 64 \times 64 + 64 \times 32 + 32 \times \text{output_dim}$. Please note that the configuration

of input-output connections for the actors and critics of the first and second agents is dictated by the specifications of the communication system under consideration. In this study, the first agent is equipped with 27 inputs for both the actor and critics, along with 20 and 1 outputs, respectively. Meanwhile, the second agent features 3 inputs for both the actor and critics, with 2 outputs for the actors and 1 output for the critics. The architectures of QTDDPG-8 and QTDD3-8 for both agents are defined as [128, 8, VQC, 16, 32] and [64, 8, VQC, 8, 4], respectively. Introducing more intricate encoding structures, QTDDPG-88 and QTDD3-88, modifies their encoding sequences to a series of gates, specifically $R_y(\theta)R_z(\theta)R_x(\theta)R_y(\theta)$, as illustrated in Figure4b. Notably, the architectures of the classical components of QTDDPG-88 and QTDD3-88 remain consistent with those of DDPG and TD3. In addition to these complex network architectures, we explore simpler structures for both actor and critic networks, incorporating only a single dress layer. For the first agent, the architectures are [27, 8, VQC, 20] for actors and [27, 8, VQC, 1] for critics. Meanwhile, the second agent adopts structures of [3, 8, VQC, 2] for actors and [3, 8, VQC, 1] for critics. It is important to note that the VQC employs an 8-qubit structure with four repetition blocks [8, 8, 8, 8], utilizing $R_y(\theta)$ gates.

A. Trajectory Analysis and SSR Evaluation

A trajectory analysis plot is presented in Figure5, Intuitively, the UAV prefers moving towards the RIS and maintains distance from the eavesdropper. We observe that all algorithms demonstrate trajectories diverging from the eavesdropper. Notably, QTDDPG-88 initially approaches the RIS but eventually retreats, possibly due to the perceived risk of proximity to the eavesdropper. Opting for a position midway between the two users enables the UAV to distribute resources more equitably in serving both users. Notably, QTDDPG-8 yields the most favorable fairness outcomes when compared to the other algorithms but with the lowest average SSR. Intriguingly, TD3, QTDD3-88, and QTDD3-DRESS exhibit a tendency to descend, aiming to serve both users impartially. Referring to the total energy consumption E_p as defined in equation (6), Figure6 illustrates that QTDDPG-8 exhibits the lowest average energy consumption over 1000 episodes among the considered algorithms. It is worth to note that most quantum algorithms show little variation in average energy consumption, except for QTDD3-DRESS. This suggests that employing a more complex structure with additional actors and critics, but too few learning parameters, may not offer significant benefits for optimization purposes. Furthermore, QTDD3-DRESS, characterized by a simple structure, also yields the second-lowest average SSR among all algorithms. Remarkably, the quantum algorithm with the highest average SSR is QTDD3-88, scoring 3.27 bits/s/Hz, surpassing DDPG with a score of 3.02 bits/s/Hz. In summary, a more intricate quantum encoding system appears to contribute to improved SSR but may also incur considerable energy consumption. Nevertheless, it is important to emphasize that this study exclusively optimizes SSR and does not specifically address energy optimization. We antic-

ipate that the results would exhibit significant improvement when energy efficiency considerations are taken into account. Figure7 depicts the average SSR for all algorithms over 1000 training episodes. Remarkably, quantum algorithms with intricate encoding systems demonstrate comparable average SSR to classical methods. An intriguing finding is observed in QTDDPG-DRESS, which not only achieves very high average SSR but also exhibits lower average energy consumption. This architecture emerges as a promising candidate in QRL, as it involves the fewest classical and quantum learning parameters, amounting to only 1.46% of classical learning parameters compared to DDPG. This leads to the need for minimal memory resources.

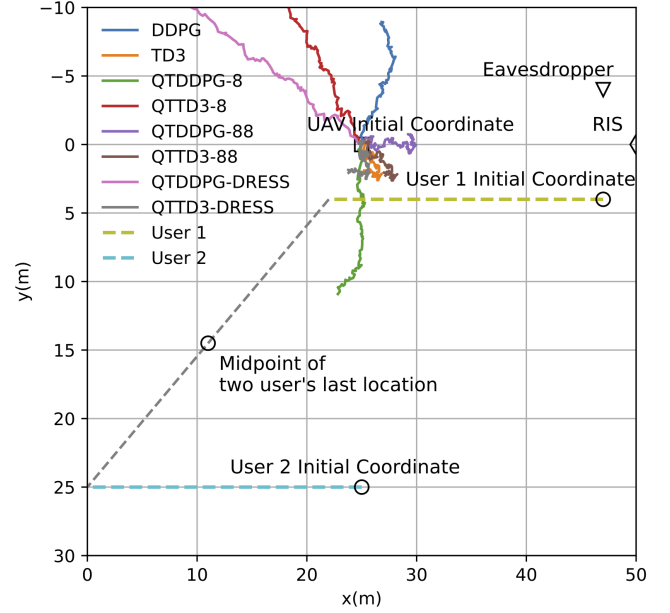


Fig. 5: The optimized UAV trajectory for various algorithms.

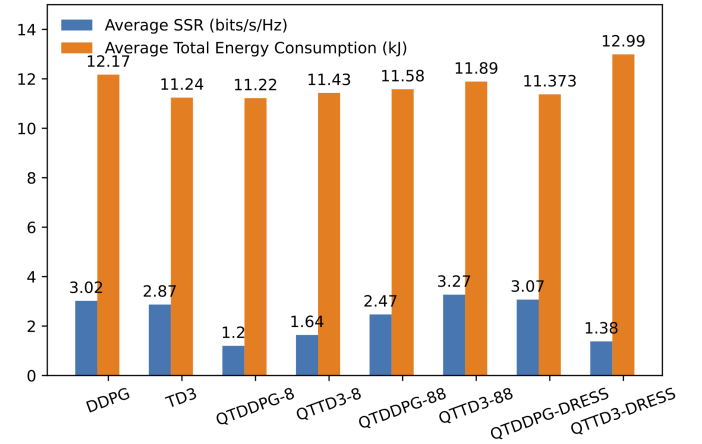


Fig. 6: Average sum secrecy rate and total energy consumption.

B. Computation Complexity Analysis

In essence, the number of episodes, denoted as N_{ep} , the number of step N_{step} , and the batch size, represented as

N_b , remain consistent between both classical and quantum algorithm simulations during the training phase. As stated in [2], the total training computational complexity of the DPPG algorithm is $\mathcal{O}(N_{ep}N_bN_{step}(\sum_{j=1}^{L_y-1}n_in_{i+1}))$. Table 1 presents the overall count of learning parameters for the hybrid classical-quantum actors and critics pertaining to the first and second agents. It is essential to observe that classical and quantum learning parameters are delineated distinctly. Notably, the incorporation of VQC in constructing a hybrid classical-quantum architecture leads to a significant reduction in the total number of learning parameters, except for QTDDPG-88 and QTDD3-88.

TABLE I: Number of Parameters

	Agent-1		Agent-2		Total Classical	Total Quantum
	Actor	Critic	Actor	Critic		
DDPG	18432	17824	4840	4836	45932	-
TD3	18432	35648	4840	9672	68592	-
QTDDPG-8	5760	5152	808	804	12524	384
QTDD3-8	5760	10304	808	1608	18480	576
QTDDPG-88	18432	17824	4840	4836	45932	384
QTDD3-88	18432	35648	4840	9672	68592	576
QTDDPG-Dress	376	224	40	32	672	384
QTDD3-Dress	376	448	40	64	928	576

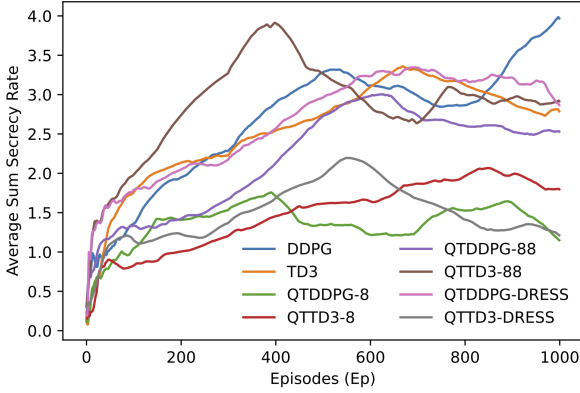


Fig. 7: Average sum secrecy rate vs Episode. To show the trend, we present the average rate over past 300 episodes.

VI. CONCLUSION

This study proposes quantum actor-critic networks to optimize average SSR in RIS-assisted mmWave UAV systems. All algorithms consistently steer away from eavesdroppers. Notably, QTDDPG-8 tends to serve nearby users fairly by positioning near their midpoint. Complex quantum encodings yield higher average SSR but increase energy consumption. QTDDPG-DRESS, with minimal learning parameters, achieves strong performance in both SSR and energy use. Future work will focus on improving energy efficiency while maintaining SSR.

REFERENCES

- [1] X. Li, H. Yao, J. Wang, X. Xu, C. Jiang, and L. Hanzo, "A near-optimal uav-aided radio coverage strategy for dense urban areas," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 9, pp. 9098–9109, 2019.
- [2] X. Guo, Y. Chen, and Y. Wang, "Learning-based robust and secure transmission for reconfigurable intelligent surface aided millimeter wave uav communications," *IEEE Wireless Communications Letters*, vol. 10, no. 8, pp. 1795–1799, 2021.
- [3] M.-L. Tham, Y. J. Wong, A. Iqbal, N. B. Ramli, Y. Zhu, and T. Dag-iuklas, "Deep reinforcement learning for secrecy energy-efficient uav communication with reconfigurable intelligent surface," in *2023 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, IEEE, 2023.
- [4] H. Li, H. Gao, T. Lv, and Y. Lu, "Deep q-learning based dynamic resource allocation for self-powered ultra-dense networks," in *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*, pp. 1–6, IEEE, 2018.
- [5] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al., "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.
- [6] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.
- [7] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *International conference on machine learning*, pp. 1587–1596, PMLR, 2018.
- [8] S. Y.-C. Chen, C.-H. H. Yang, J. Qi, P.-Y. Chen, X. Ma, and H.-S. Goan, "Variational quantum circuits for deep reinforcement learning," *IEEE Access*, vol. 8, pp. 141007–141024, 2020.
- [9] S. Y.-C. Chen, "Asynchronous training of quantum reinforcement learning," *Procedia Computer Science*, vol. 222, pp. 321–330, 2023. International Neural Network Society Workshop on Deep Learning Innovations and Applications (INNS DLIA 2023).
- [10] S. F. Chien, H. T. David Chieng, S. Y.-C. Chen, C. C. Zarakovitis, H. S. Lim, and Y. Xu, "Applying hybrid quantum lstm for indoor localization based on rssi," in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 0–0, IEEE, 2024.
- [11] S. F. Chien, H. S. Lim, M. A. Kourtis, Q. Ni, A. Zappone, and C. C. Zarakovitis, "Quantum-driven energy-efficiency optimization for next-generation communications systems," *Energies*, vol. 14, no. 14, p. 4090, 2021.
- [12] H.-Y. Huang, M. Broughton, J. Cotler, S. Chen, J. Li, M. Mohseni, H. Neven, R. Babbush, R. Kueng, J. Preskill, et al., "Quantum advantage in learning from experiments," *Science*, vol. 376, no. 6598, pp. 1182–1186, 2022.
- [13] G. Zhou, C. Pan, H. Ren, K. Wang, M. El-kashlan, and M. Di Renzo, "Stochastic learning-based robust beamforming design for ris-aided millimeter-wave systems in the presence of random blockages," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 1, pp. 1057–1061, 2021.
- [14] H. Yang, Z. Xiong, J. Zhao, D. T. Niyato, L. Xiao, and Q. Wu, "Deep reinforcement learning-based intelligent reflecting surface for secure wireless communications," *IEEE Transactions on Wireless Communications*, vol. 20, pp. 375–388, 2020.
- [15] X. Liu, M. Chen, Y. Liu, Y. Chen, S. Cui, and L. Hanzo, "Artificial intelligence aided next-generation networks relying on uavs," *IEEE Wireless Communications*, vol. 28, no. 1, pp. 120–127, 2020.
- [16] W. Zhang, Q. Wang, X. Liu, Y. Liu, and Y. Chen, "Three-dimension trajectory design for multi-uav wireless network with deep reinforcement learning," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 1, pp. 600–612, 2020.
- [17] C. H. Liu, X. Ma, X. Gao, and J. Tang, "Distributed energy-efficient multi-uav navigation for long-term communication coverage by deep reinforcement learning," *IEEE Transactions on Mobile Computing*, vol. 19, no. 6, pp. 1274–1285, 2019.
- [18] H. Qi, Z. Hu, H. Huang, X. Wen, and Z. Lu, "Energy efficient 3-d uav control for persistent communication service and fairness: A deep reinforcement learning approach," *IEEE Access*, vol. 8, pp. 53172–53184, 2020.
- [19] A. Skolik, S. Jerbi, and V. Dunjko, "Quantum agents in the gym: a variational quantum algorithm for deep q-learning," *Quantum*, vol. 6, p. 720, 2022.
- [20] H. Wang, Y. Ding, J. Gu, Z. Li, Y. Lin, D. Z. Pan, F. T. Chong, and S. Han, "Quantumnas: Noise-adaptive search for robust quantum circuits," in *The 28th IEEE International Symposium on High-Performance Computer Architecture (HPCA-28)*, 2022.