Quantum Reinforcement Learning with Classical Policy Deployment for Resource Allocation in Multi-Beam GEO-LEO Satellite Networks

Quynh Tu Ngo, *Senior Member, IEEE,* Ying He, *Senior Member, IEEE,*
Beeshanga Jayawickrama, *Senior Member, IEEE,* Eryk Dutkiewicz, *Senior Member, IEEE,*
Shiva Raj Pokhrel, *Senior Member, IEEE*

***Abstract*—Satellite communications (SatCom) are envisioned as a critical enabler of 6G networks, enabling seamless global coverage by integrating terrestrial infrastructures with multi-layered satellite constellations. Among these, the integration between geostationary (GEO) and low Earth orbit (LEO) satellite networks is particularly attractive, as they combine the broad coverage of GEO satellites with the low latency and high capacity of LEO systems. Within this context, we address the resource allocation problem for LEO satellite through a joint design of beam size and transmit power, while accounting for GEO interference constraints, residual Doppler frequency offsets, and frequency reuse strategies. The objective is to maximize the spectral efficiency of LEO system operating in multi-beam GEO-LEO networks. Motivated by the limitations of classical deep reinforcement learning (RL) in such dynamic orbital settings and the potential of quantum RL for accelerated convergence, we propose a hybrid solution that exploits quantum acceleration during offline training and subsequently exports the learned policy into a classical representational format for onboard LEO satellite deployment. A fully quantum deep deterministic policy gradient framework with variational quantum circuit-based actor and critic is developed, along with a neural network-based policy translator for classical inference. To the best of our knowledge, this is the first deployment-ready quantum RL framework in SatCom, offering efficient offline training, reduced retraining latency, and practical deployment compatibility with existing LEO satellite hardware.**

***Index Terms*—Multi-beam GEO-LEO networks, satellite resource allocation, Doppler frequency offset, quantum reinforcement learning, quantum-DDPG, classical deployment.**

I. Introduction

Satellite communication is expected to play a central role in the 6G networks by extending connectivity to underserved and remote areas, with the coexistence of geostationary (GEO) and low Earth orbit (LEO) satellite networks emerging as a particularly attractive architecture. GEO satellites provide broad, stable coverage, while LEO constellations deliver low latency and high throughput [1]. However, the rapid expansion of LEO deployments is intensifying spectrum scarcity and interference risks. Unlike GEO systems fixed in orbital slots, LEO satellites move quickly across smaller footprints, enabling high frequency reuse but increasing the chance of inter-system interference. Although early strategies relied on assigning separate bands to GEO and LEO, real-world constellations such as Starlink, OneWeb, and Amazon Kuiper now operate in Ku and Ka-bands already heavily used by GEO systems [2]. These trends make spectrum coexistence unavoidable and highlight the urgent need for interference mitigation strategies in multi-orbit networks. To safeguard GEO operations, the International Telecommunication Union (ITU) has introduced regulatory measures that prioritize GEO protection. LEO systems must restrict transmit power, coordinate frequency usage, and implement spatial isolation to avoid harmful interference [3]. Such measures are crucial for preserving GEO services. However, strict regulatory limits also constrain the flexibility and efficiency of LEO networks, creating a pressing need for technical solutions that preserve GEO service quality while enabling robust LEO performance [4].

Advances in antenna and beamforming technologies have opened new possibilities for addressing this challenge. Modern phased array and multi-beam satellite systems allow fine-grained control over beam direction, coverage, and power distribution. These capabilities enable dynamic frequency reuse and spatial domain interference management, both recognized by the 3rd Generation Partnership Project (3GPP) as critical for efficient spectrum utilization in non-terrestrial networks [5]. Building on these capabilities, researchers have proposed strategies such as cooperative beamforming with reconfigurable intelligent surfaces [6], user clustering with adaptive beam steering for LEO satellites [7], spatio-temporal models of interference [8], and dynamic spectrum access schemes tailored for multi-orbit coexistence [9]. Despite this progress, most existing studies assume that Doppler effects introduced by fast-moving LEO satellites can be perfectly compensated. In practice, only users near the beam center benefit from accurate Doppler pre-compensation, while off-center users experience residual frequency offsets [10], [11]. These offsets degrade performance in narrow-beam systems by causing spectral leakage across subbands, amplifying both intra and inter-system interference [12]. Consequently, Doppler-induced impairments remain a critical barrier to practical interference management in GEO-LEO satellite networks. Motivated by these challenges, this article addresses a resource allocation problem for LEO satellites through a joint design

Quynh Tu Ngo, Ying He, Beeshanga Jayawickrama, and Eryk Dutkiewicz are with the School of Electrical and Data Engineering, University of Technology Sydney, Sydney, NSW 2007, Australia (Email: {QuynhTu.Ngo, Ying.He, Beeshanga.Jayawickrama, Eryk.Dutkiewicz}@uts.edu.au).

Shiva Raj Pokhrel is with the School of IT, Deakin University, Geelong, VIC 3125, Australia (Email: shiva.pokhrel@deakin.edu.au).

This work has been supported by the SmartSat CRC, whose activities are funded by the Australian Government's CRC Program.

of beam size and transmit power. The proposed framework incorporates GEO interference constraints, residual Doppler frequency offsets, and frequency reuse strategies, aiming to maximize the spectral efficiency of LEO system in multi-beam GEO-LEO networks.

Reinforcement learning (RL), and particularly deep reinforcement learning (DRL), has recently gained significant attention in satellite networks for resource allocation. DRL offers adaptive, data-driven strategies to optimize spectrum use, power distribution, and beam management in highly dynamic environments [11], [13]–[15]. By learning directly from network interactions, DRL agents can outperform traditional optimization methods in handling the complexity of multi-orbit satellite systems. However, deploying DRL in satellites faces notable challenges. Training DRL agents typically requires long runtimes and substantial computational resources, which are severely constrained onboard. As a result, training is usually performed offline at ground stations, with only inference executed on satellites [11], [12]. Moreover, the rapidly changing orbital environment of LEO satellites often necessitates frequent retraining to maintain performance. These factors hinder responsiveness and adaptability when relying solely on classical DRL approaches. Quantum reinforcement learning (QRL) has emerged as a promising alternative. Recent studies show that QRL agents can converge faster and learn more efficiently than classical counterparts [16]–[18]. A hybrid quantum double deep Q-learning method was proposed for computation task offloading in satellite-terrestrial integrated networks (STIN) [16], where a classical deep Q-network (DQN) and a quantum neural network (QNN) operate in parallel. This agent achieved higher rewards with fewer data points compared to a classical double DQN. A quantum-assisted DRL framework was introduced for direction-of-arrival estimation and task offloading in integrated sensing and communication STIN [17]. Using a quantum-enhanced actor-critic structure, it outperformed conventional DRL in reducing task offloading latency. A quantum multi-agent RL scheme for cooperative mobile access in STIN was also proposed in [18], where the QNN architecture provided robust reward convergence in large action spaces and improved overall system performance compared to DQN agents. It is important to note that all existing QRL frameworks are deployed on ground stations. Running QRL onboard LEO satellites remains infeasible, as most satellites lack quantum hardware. Even pioneering missions such as the first quantum satellite computer, ROQuET [19], cannot yet support QRL workloads in orbit. While ROQuET represents a milestone in space-based quantum technologies, its utility for practical QRL deployment has not been realized. Motivated by the limitations of classical DRL in satellite contexts and the potential of QRL for expedited learning, this article proposes a hybrid solution: training the QRL agent offline by leveraging quantum acceleration, and then exporting the learned policy into a classical representational format suitable for deployment on LEO satellites with only classical inference capabilities. This approach enhances training efficiency, reduces retraining latency, and improves adaptability in dynamic orbital environments, while remaining full compatible with satellites' classical hardware constraints. The main contributions can be summarized as follows:

- **Fully quantum actor-critic design**: We propose a quantum deep deterministic policy gradient (Q-DDPG) framework in which both the actor and critic are realized via variational quantum circuits (VQCs). In contrast to existing quantum DDPG approaches, which combine a quantum actor with a classical critic [20], [21], or use hybrid classical-quantum layer structures [22], our design is fully quantum on the training side, thereby exploiting the enhanced representational capacity and learning efficiency of quantum models.
- **Policy translator for classical deployment**: We introduce a classical neural network-based policy translator that maps the trained quantum policy into a purely classical format. This enables inference on LEO satellites that lack quantum hardware, bridging the gap between quantum-enhanced training and classical-only deployment.
- **First deployment-ready quantum DRL framework**: To the best of our knowledge, this is the first work in the literature that achieves training a QRL agent and subsequently deploying it on a classical DNN device. This unique capability ensures that quantum acceleration can be leveraged during training, while maintaining lightweight and practical deployment on existing satellite systems.

Table I highlights the distinctions and contributions of our proposed QRL framework relative to existing literature. These contributions establish a new paradigm for QRL in satellite networks: quantum-enhanced training coupled with classical-compatible deployment. The remainder of this article is organized as follows: Section II describes the system model and outlines the analytical framework. Section III details the resource allocation for LEO satellite with the problem formulation and the proposed RL approach. Then, Section IV presents the proposed quantum RL solution with classical deployment on LEO satellite. Section V has the numerical results. Section VI discusses the practical considerations and scope of quantum reinforcement learning. Lastly, Section VII concludes the article.

II. System Model

Fig. 1 depicts a coexisting GEO-LEO multi-beam satellite network consisting of a GEO satellite and a LEO satellite. The GEO satellite S_G operates in multi-beam mode with an ι_G-color frequency reuse scheme ($\iota_G > 1$), where its total bandwidth B_G is divided into ι_G sub-bands. Each GEO spot beam covers a circular ground cell of radius R_G and serves a single GEO user U_{G_j}, $j = 1, \cdots, N_G$. In the LEO segment, a LEO satellite S_L serves its users U_{L_i} $i = 1, \cdots, N_L$, using the GEO spectrum. The LEO satellite employs phased-array antennas to form steerable narrow beams while maintaining a fixed orientation. The LEO satellite is assumed to have a circular footprint of radius F_L ($F_L > R_G$), within which non-overlapping hexagonal spot beam ground coverage of radius R_L ($R_L < R_G$) are deployed. The coverage area of interest is defined as one LEO footprint. The number of GEO and

TABLE I: Contribution comparison to relevant quantum reinforcement learning literature

	Network type		Quantum RL framework	Inference deployment	
	Satellite	Terrestrial		Ground	LEO satellite
[16]	✓		Hybrid quantum double deep Q-learning (DQN and quantum NN)	Quantum device	
[17]	✓		Quantum-enhanced actor-critic (using variational quantum circuit)	Quantum device	
[18]	✓		Multi-agent quantum neural network	Quantum device	
[20], [21]		✓	Quantum-enhanced DDPG (quantum actor with parameterized quantum circuit and classical DNN critic)	Quantum device	
[22]		✓	Quantum-enhanced DDPG (both actor and critic networks have hybrid classical-quantum layers)	Quantum device	
This work	✓		Fully quantum-DDPG (quantum actor and quantum critic) with a policy translator for classical deployment		Classical device

LEO spot beams within this area are $N_G = 2\pi F_L^2/2\pi R_G^2$ and $N_L = 2\pi F_L^2/3\sqrt{3}R_L^2$, respectively. Each beam serves a single user, and all GEO and LEO users are equipped with single-antenna terminals.

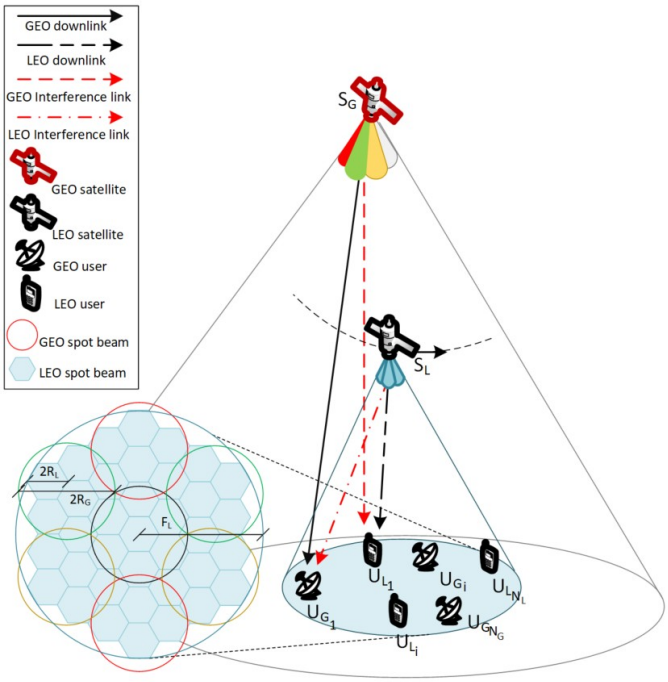

Fig. 1: An illustration of a multi-beam GEO-LEO satellite network, where the GEO satellite employs a 4-color frequency reuse scheme ($\iota_G = 4$).

A. Signal Model

Let s_{G_j} and s_{L_i} denote the normalized data symbols requested by the GEO and LEO users, respectively, such that $\mathbb{E}\left[|s_{G_j}|^2\right] = \mathbb{E}\left[|s_{L_i}|^2\right] = 1$. Let t represent the transmission time. Accordingly, the received signal at a GEO user is given by:

$$y_{U_{G_j}}(t) = \sqrt{\frac{P_{G_j}G_{G_j}}{L_{GU_{G_j}}}}h_{GU_{G_j}}(t)s_{G_j} + n_{U_G}(t) + \underbrace{\sum_{i=1}^{N_{GL}}\sqrt{\frac{P_{L_i}G_{L_i}}{L_{LU_{G_j}}}}h_{LU_{G_j}}(t)s_{L_i}e^{j2\pi\Delta f_i^D(t)t}}_{\text{inter-system interference}}, \quad (1)$$

and the received signal at a LEO user is given by:

$$y_{U_{L_i}}(t) = \sqrt{\frac{P_{L_i}G_{L_i}}{L_{LU_{L_i}}}}h_{LU_{L_i}}(t)s_{L_i}e^{j2\pi\Delta f_i^D(t)t} + n_{U_L}(t) + \underbrace{\sqrt{\frac{P_{G_j}G_{G_j}}{L_{GU_{L_i}}}}h_{GU_{L_i}}(t)s_{G_j}}_{\text{inter-system interference}} + \underbrace{\alpha_f\sum_{l,l\neq i}^{N_{GL}}\sqrt{\frac{P_{L_l}G_{L_l}}{L_{LU_{L_l}}}}h_{LU_{L_l}}(t)s_{L_l}e^{j2\pi\Delta f_l^D(t)t}}_{\text{LEO intra-system interference}}. \quad (2)$$

Here, P_{G_j} and P_{L_i} are the j-th GEO and the i-th LEO spot beam transmit power, respectively.

G_{G_j} and G_{L_i} are the corresponding spot beam gain and are modeled as

$$G_{G_j/L_i}(\phi) = G_{G_j/L_i}^{max}\left(\frac{J_1(\mu)}{2\mu} + \frac{36J_3(\mu)}{\mu^3}\right)^2, \quad (3)$$

where ϕ is the off-boresight angle from satellite to its user; ϕ_{3dB} is the off-boresight angle corresponding to the 3 dB beamwidth; J_1 and J_3 respectively denote the first order and third order Bessel functions; $\mu = 2.07123\sin(\phi)/\sin(\phi_{3dB})$; $G_{G_j/L_i}^{max} = 4\pi A\eta/(c/f_c)^2$ is the maximum antenna gain at zero off-boresight angle with A and η being the antenna area and efficiency, respectively, c being the speed of light, and f_c being the carrier frequency.

$h_{GU_{G_j}}$, $h_{GU_{L_i}}$, $h_{LU_{G_j}}$, $h_{LU_{L_l}}$, and $h_{LU_{L_i}}$ denote the channel between GEO and LEO satellites to the corresponding users, respectively. The channel coefficients are modeled using Shadowed-Rician distribution. Note that perfect channel state information is assumed.

$L_{GU_{G_j}}$, $L_{LU_{G_j}}$, $L_{GU_{L_i}}$, $L_{LU_{L_i}}$, and $L_{LU_{L_l}}$ are the path loss between satellites and ground users, and are modeled as free space propagation loss as follows:

$$L = \left(\frac{4\pi f_c}{c}\cdot 2R_E\cdot \arcsin\left(\sqrt{\sin^2\left(\frac{l_U^a - l_S^a}{2}\right) + \cos(l_S^a)\cos(l_U^a)\sin^2\left(\frac{l_U^o - l_S^o}{2}\right)}\right)\right)^2, \quad (4)$$

where R_E is the Earth's radius; $\{l_S^a, l_S^o\}$ and $\{l_U^a, l_U^o\}$ are the latitude and longitude of satellite and user, respectively.

Δf_i^D and Δf_l^D represent the residual Doppler frequency offset due to LEO satellite movement after perfect Doppler

precompensation. Perfect Doppler precompensation technique can only compensate Doppler at a beam center. For any user locating off beam center, there is a residual Doppler frequency offset as below:

$$\Delta f_i^D(t) = \frac{V_L(t)}{c} f_c \left(\cos\varphi_{oc}(t) - \cos\varphi_{ctr}(t)\right), \quad (5)$$

where V_L is the LEO velocity, φ_{oc} and φ_{ctr} are the angle between the LEO velocity vector and the line of sight to user and beam center, respectively. When a user is at the beam edge, the upper bound residual Doppler frequency offset is

$$\sup\left\{\left|\Delta f_i^D\right|\right\} = \frac{V_{L_i}^*}{c} f_c \left(\frac{R_L}{\sqrt{R_L^2 + A_L^2}}\right), \quad (6)$$

where $V_{L_i}^*$ is the LEO velocity when it is directly overhead the beam center and A_L is the LEO altitude.
$N_{GL} = 2\pi R_G^2 / 3\sqrt{3} R_L^2$ denotes the number of LEO beams inside a GEO beam. $\alpha_f \in \{0, 1\}$ denotes the frequency reuse scheme adopted by LEO satellite, where $\alpha_f = 0$ indicates no frequency reuse (no LEO intra-system interference), and $\alpha_f = 1$ corresponds to a full frequency reuse scheme (N_{GL} LEO beams use the same frequency sub-band as their masked GEO beam).
Lastly, n_{U_G} and n_{U_L} are the additive white Gaussian noise (AWGN) at GEO and LEO users with zero mean and variance $\sigma_{n_G}^2$, $\sigma_{n_L}^2$, respectively.

B. Performance Metric

1) Effective Bandwidth: The bandwidth per beam available for LEO transmission is given as

$$B_{pb} = \frac{B_G}{\iota_G \iota_L}, \quad (7a)$$

$$\iota_L = \begin{cases} 1, \text{ full frequency reuse} \\ N_{GL}, \text{ no frequency reuse} \end{cases} \quad (7b)$$

where ι_L is the LEO frequency reuse scheme factor. Under full frequency reuse, all LEO spot beams located within a GEO beam share the same frequency band as that GEO beam. In contrast, with no frequency reuse, the GEO beam's bandwidth is evenly partitioned among the LEO spot beams it contains.

Let B_{pb}^e represent the effective bandwidth available per LEO beam, considering the impact of residual Doppler frequency offset. It can be written as

$$B_{pb} - 2\sup\left\{\left|\Delta f_i^D\right|\right\} \le B_{pb}^e \le B_{pb}. \quad (8)$$

2) Signal-to-Interference-plus-Noise Ratio (SINR): The SINR per bandwidth at a GEO user can be expressed as

$$\Gamma_{U_{G_j}} = \frac{P_{G_j} G_{G_j} L_{GU_{G_j}}^{-1} \left|h_{GU_{G_j}}\right|^2 \iota_G / B_G}{\sum_{i=1}^{N_{GL}} P_{L_i} G_{L_i} L_{LU_{G_j}}^{-1} \left|h_{LU_{G_j}}\right|^2 / B_{pb}^e + \sigma_{n_G}^2}. \quad (9)$$

The SINR per bandwidth at a LEO user can be expressed as

$$\Gamma_{U_{L_i}} = \frac{P_{L_i} G_{L_i} L_{LU_{L_i}}^{-1} \left|h_{LU_{L_i}}\right|^2 / B_{pb}^e}{P_{G_j} G_{G_j} L_{GU_{L_i}}^{-1} \left|h_{GU_{L_i}}\right|^2 \iota_G / B_G + \sigma_{n_{U_L}}^2 + \alpha_f \sum_{l, l\neq i}^{N_{GL}} P_{L_l} G_{L_l} L_{LU_{L_l}}^{-1} \left|h_{LU_{L_l}}\right|^2 / B_{pb}^e}. \quad (10)$$

3) LEO Spectral Efficiency: Let η_L denote the LEO spectral efficiency. η_L can be defined as the ergodic channel capacity in nat/sec/Hz,

$$\eta_L = \sum_{i=1}^{N} \ln\left(1 + \Gamma_{U_{L_i}}\right). \quad (11)$$

C. Extension to Multi-LEO scenario

Although the system model considers a single LEO satellite footprint within a GEO coverage area for clarity of exposition, the formulation readily extends to the case of multiple LEO satellites simultaneously illuminating the same GEO footprint. In such a case, each LEO satellite forms an independent multi-beam footprint, and additional inter-satellite interference arises only in the geographical regions where neighboring LEO footprints overlap, i.e., at the edge beams. This inter-LEO interference can be explicitly included in the SINR expressions by augmenting the aggregate interference term, without changing the structure of the proposed beam-size and power optimization problem.

III. Problem Formulation

This work focuses on characterizing the physical-layer performance limits of GEO-LEO coexistence under residual Doppler and interference constraints. Therefore, the objective is formulated as the maximization of the LEO spectral efficiency, which captures the efficiency of spectrum utilization in an interference-limited, frequency-reuse multi-beam system. Traffic demand is assumed to be sufficiently backlogged in each active beam so that the achievable rate is primarily constrained by Doppler-induced spectral leakage and GEO protection requirements, rather than by instantaneous queue states. While practical satellite systems exhibit non-uniform and time-varying traffic, incorporating traffic awareness (e.g., beam-dependent weights, minimum rate constraints, or queue-aware scheduling) would mainly affect the higher-layer resource scheduling and can be naturally integrated into the proposed framework by replacing the spectral efficiency with a weighted sum-rate or demand-constrained utility. Such an extension is left for future work, as the present work aims to isolate and quantify the impact of residual Doppler and interference coupling on the fundamental resource allocation design. The optimization problem can thus be expressed as:

$$\max_{R_L, P_{L_i}} \sum_{i=1}^{N} \ln\left(1 + \Gamma_{U_{L_i}}\right) \quad (12a)$$

$$\text{s.t.} \sum_{i=1}^{N} P_{L_i} \le P_{L,\max}, \quad (12b)$$

$$\Gamma_{U_{G_j}} \ge \bar{\Gamma}_{U_{G_j}},\ j = \{1, \cdots, N_G\}, \quad (12c)$$

$$B_{pb} - 2\sup\left\{\left|\Delta f_i^D\right|\right\} \le B_{pb}^e \le B_{pb}, \quad i = \{1, \cdots, N_L\}, \quad (12d)$$

$$\inf\{R_L\} \le R_L \le R_G, \quad (12e)$$

where $P_{L,\max}$ is the LEO total transmit power budget, $\bar{\Gamma}_{U_{G_j}}$ denotes the GEO user SINR threshold, and $\inf\{R_L\} =$

$\sqrt{B_{pb}^2 c^2 A_L^2/(4V_{L_i}^{*2} f_c^2 - B_{pb}^2 c^2)}$ denotes the minimum feasible LEO beam size, determined by accounting for the residual Doppler frequency offset to ensure reliable transmission.

The constraint (12b) enforces the maximum power budget of the LEO satellite. The constraint (12c) guarantees adequate protection for GEO users, which is a strict requirement in GEO–LEO satellite networks. Finally, the constraints (12d) and (12e) ensure that the residual Doppler frequency offset remains within acceptable limits for feasible LEO transmission.

The optimization problem (12) is non-convex due to the logarithmic dependence on SINR and nonlinear SINR-Doppler constraints. Conventional convex optimization methods, which rely on relaxations or approximations, fail to generalize in dynamic satellite environments, while heuristic approaches lack the adaptability and scalability needed for real-time operation. To overcome these limitations, we adopt an RL framework with a continuous action space. This design offers three key benefits: (i) RL adaptively learns from CSI and Doppler variations, enabling robust performance in dynamic GEO-LEO environments; (ii) continuous actions provide fine-grained control over beam size and power allocation, which is difficult to achieve with discrete RL or conventional optimization; and (iii) quantum RL techniques are incorporated during training to improve exploration, scalability, and convergence efficiency, while the final trained policy is expressed in classical form for deployment on existing satellites.

A. Reinforcement Learning-Based Problem Reformulation

Reinforcement learning is a machine learning framework in which an agent acquires decision-making strategies through iterative interactions with its environment, with the objective of maximizing cumulative long-term rewards. At each time step, the agent observes the current environmental state, selects an action according to a policy, and receives evaluative feedback in the form of a reward. The environment subsequently transitions to a new state as a consequence of the chosen action. Through repeated interactions, the agent seeks to optimize its policy to achieve maximal long-term reward. By applying suitable algorithms, the agent continually updates and improves its policy, thereby enhancing its decision-making ability over time.

Here, problem (12) is reformulated as an RL problem modeled as a Markov decision process (MDP) within the context of a multi-beam GEO-LEO satellite system. The key components of the MDP are outlined below.

State space $\mathcal{S}$: Let $s_t \in \mathcal{S}$ represent the state at a time step t. $s(t)$ composes of (i) the CSIs of GEO and LEO users, denoted by $\mathbf{h}_t = \left[h_{G_1,t}, \ldots, , h_{G_{N_G},t}, h_{L_1,t}, \ldots, h_{L_{N_L},t}\right]$, in which $h_{G_1,t} = G_{G_1}^{1/2} L_{GU_{G_1},t}^{-1/2} h_{GU_{G_1},t}$, and (ii) the upper bound residual Doppler frequency offset, denoted by $\mathbf{\Delta f}_{D,t} = [\sup\{|\Delta f_1^D|\}, \ldots, \sup\{|\Delta f_{N_L}^D|\}]$. The state s_t is defined as

$$s_t \triangleq \left\{\mathbf{h}_t, \mathbf{\Delta f}_{D,t}\right\}. \tag{13}$$

Action space $\mathcal{A}$: Let $a_t \in \mathcal{A}$ be the action at a time step t. As problem (12) aims to design the beam size and transmit power distribution for LEO satellite, the action a_t includes the LEO beam radius $R_{L,t}$ and transmit power $\mathbf{P}_{L,t} = \left[P_{L_1,t}, \ldots, P_{L_{N_L},t}\right]$, defined as

$$a_t \triangleq \left\{R_{L,t}, \mathbf{P}_{L,t}\right\}. \tag{14}$$

Reward function: The reward function measures the effectiveness of the LEO beam design and power allocation following the agent's action. Since the objective of problem (12) is to maximize LEO spectral efficiency, the reward at a time step t is defined as a function of spectral efficiency:

$$r_t = \begin{cases} \eta_{L,t}, & \text{if no constraint is violated,} \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

Let $\kappa \in [0, 1)$ denote the reward discount factor, the cumulative reward can be expressed as

$$\mathfrak{R} = \sum_{i=t}^{\infty} \kappa^{i-t} r_i. \tag{16}$$

The reward in (15) follows a feasibility-first design. Constraints (12b)-(12e) define the admissible action set, including the satellite power budget, GEO user protection, and Doppler compensation limits. Actions violating any of these constraints correspond to physically or regulatorily infeasible transmission states. Therefore, the reward is defined as the LEO spectral efficiency only when all constraints are satisfied, and zero otherwise. This hard-constraint formulation avoids allowing the learning agent to trade off spectral efficiency against constraint violations, which would be inconsistent with the underlying system requirements. While soft-penalty formulations could be used to smooth the learning signal, they would relax strict feasibility and are thus not adopted in this work.

B. Quantum Representation of Environment State

To leverage quantum learning, the classical state vector s_t in (13) of dimension N_s is embedded into a quantum state of n_s qubits using amplitude encoding [23]. Each qubit is a two-level quantum system, and together n_s qubits span a Hilbert space of dimension $N_{\mathcal{H}} = 2^{n_s}$. First, s_t is normalized as $\hat{s}_t = s_t/\|s_t\|_2$, where $\|s_t\|_2 = \sqrt{\sum_{i=0}^{N_s-1} |s_{i,t}|^2}$. The number of qubits required is $n_s = \lceil \log_2 N_s \rceil$. If N_s is not a power of two, the vector s_t is extended to length $N_{\mathcal{H}}$ by padding with zeros. The normalized entries are then mapped to the amplitudes of the n_s-qubit state. Hence, the quantum-encoded state $|s_t^Q\rangle$[1] at a time step t can be represented as

$$|s_t^Q\rangle = \sum_{i=0}^{N_{\mathcal{H}}-1} \hat{s}_{i,t} |i\rangle, \quad |s_t^Q\rangle \in \mathbb{C}^{2^{n_s}}, \tag{17}$$

where $|i\rangle$ denotes the computational basis states of the n_s-qubit Hilbert space, and $\mathbb{C}$ denotes the complex value space. This encoding ensures that both the channel state information $\mathbf{h}_t$ and the Doppler residual offsets $\mathbf{\Delta f}_{D,t}$ are compactly represented in a quantum state, enabling efficient manipulation and learning within the QRL framework.

[1]In this work, we use Dirac notation where $|\psi\rangle$ denotes a quantum state vector (a normalized element of a Hilbert space $\mathcal{H}$), and $\langle\psi|$ is its conjugate transpose.

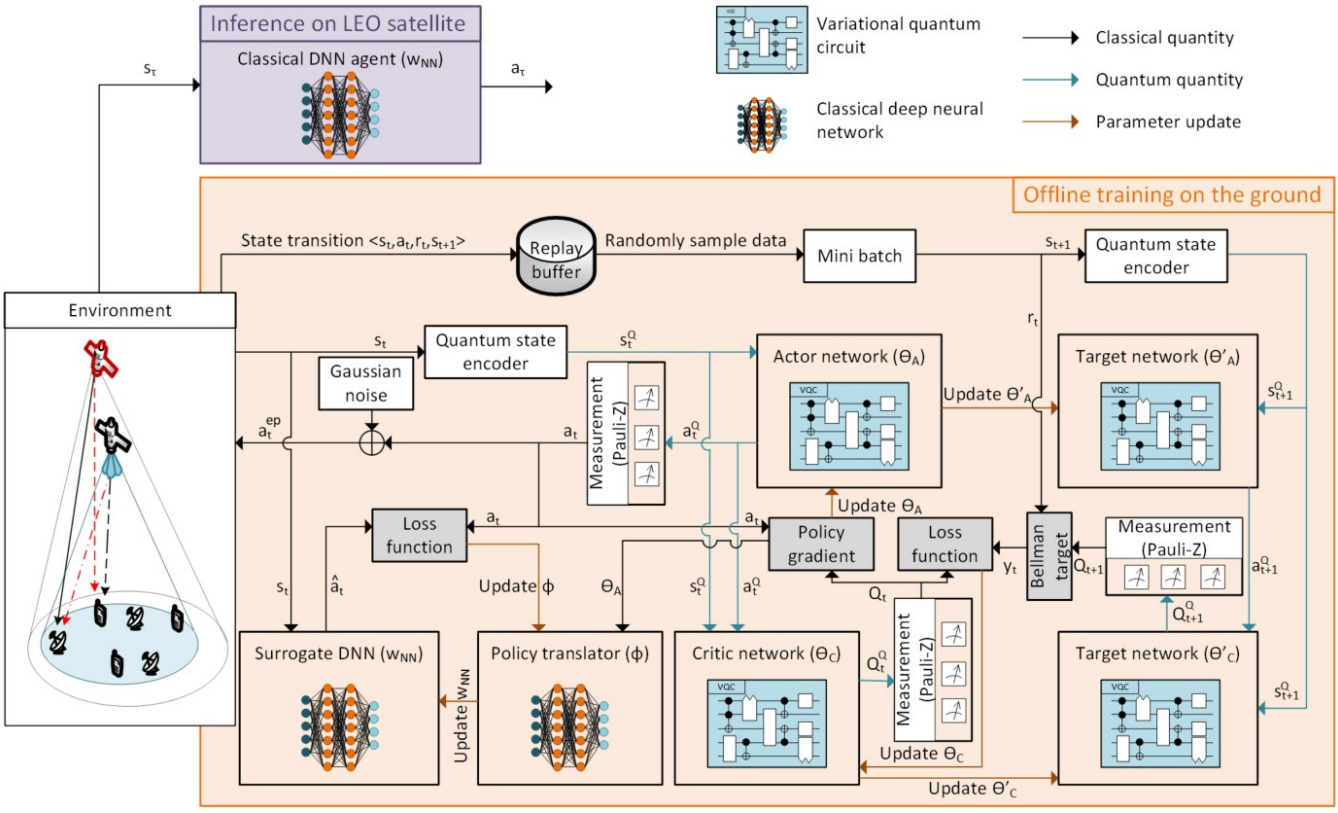

Fig. 2: Proposed quantum deep deterministic policy gradient with classical policy deployment (Q-DDPG-CPD) framework architecture.

IV. Quantum Deep Deterministic Policy Gradient with Classical Policy Deployment for LEO Satellite Resource Allocation

This section describes the proposed quantum deep deterministic policy gradient with classical policy deployment (Q-DDPG-CPD) framework for solving problem (12). In Q-DDPG-CPD, both the actor and critic are implemented as VQCs. Each VQC takes a quantum-encoded input state and outputs another quantum state, with the trainable gate rotation angles serving as network parameters. Parameter updates are performed using classical optimization methods. To interface the quantum outputs with the classical RL loop, a Pauli-Z measurement is applied to obtain expectation values, which are then mapped into classical format, producing continuous actions from the actor and scalar Q-value estimates from the critic. For inference deployment on LEO satellites, where the agent must be implemented as a classical DNN, the actor of Q-DDPG-CPD incorporates a policy translator. This translator, realized as a DNN, is trained jointly with the VQC actor to map the quantum actor's parameters into the corresponding parameters of the classical DNN agent executed on the satellite. The overall architecture of the proposed framework is illustrated in Fig. 2 with key notions summarized in Table II.

A. Quantum Actor

Let θ_A denote the gate rotation angles of the actor VQC. The actor VQC transforms a quantum-encoded state $|s_t^Q\rangle$ into a quantum action $|a_t^Q\rangle$, using the same number of qubits as its quantum state input. The quantum action can be expressed as

$$|a_t^Q\rangle = \mathcal{U}(\theta_A)|s_t^Q\rangle, \quad |a_t^Q\rangle \in \mathbb{C}^{2^{n_s}}, \tag{18}$$

where $\mathcal{U}(\theta_A) = e^{-i\hat{\mathcal{H}}(\theta_A)}$ is a unitary operator representing the system's energy dynamics in the quantum computational model. The Hamiltonian operator $\hat{\mathcal{H}}(\theta_A)$ is defined as

$$\hat{\mathcal{H}}(\theta_A) = \sum_i \epsilon_i |i\rangle\langle i| + \sum_{i\neq j} \mathcal{T}_{ij}(|i\rangle\langle j| + |j\rangle\langle i|), \tag{19}$$

where ϵ_i is the energy associated with the system being in state $|i\rangle$, and $\mathcal{T}_{ij}$ is the transition energy between state $|i\rangle$ and state $|j\rangle$. This formulation explicitly connects the trainable parameters of the actor VQC, θ_A, to the evolution of the quantum action.

To interface with the classical environment, the quantum action $|a_t^Q\rangle$ is measured using the Pauli-Z operator [24]. The expectation value of the measurement produces real numbers, which are then mapped to a continuous classical action a_t of dimension N_a. The extraction of N_a continuous action components via Pauli-Z expectation values can be expressed as

$$a_t[k] = f_k\left(\sum_{j=1}^{n_s} \omega_{k,j}\langle a_t^Q|\mathbf{Z}_j|a_t^Q\rangle\right), \quad k = 1, \cdots, N_a, \tag{20}$$

where $f_k(\cdot)$ is a scaling function mapping the measured expectation to the environment's action range, $\langle a_t^Q|\mathbf{Z}_j|a_t^Q\rangle$ is

TABLE II: Summary of key notations

Notation	Description
$\mathcal{S}$	Classical state space
$\mathcal{A}$	Classical action space
s_t	Classical state
a_t	Classical action
Q_t	Classical Q-value
r_t	Immediate reward
κ	Reward discount factor
$\mathfrak{R}$	Cumulative reward
N_s	Classical state vector dimension
N_a	Classical action vector dimension
$N_{\mathcal{H}}$	Hilbert space dimension
n_s	Number of qubits used in actor VQC
s_t^Q	Quantum-encoded state
a_t^Q	Quantum action
Q_t^Q	Quantum Q-Value
$\mathcal{U}(\cdot)$	Unitary operator
$\hat{\mathcal{H}}(\cdot)$	Hamiltonian operator
$\mathbf{Z}$	Pauli-Z operator
$\omega_{k,j}$	Pauli-Z linear weights
a_t^{ep}	Explored action
$\hat{a}_t$	Surrogate action
θ_A	Actor VQC parameters
θ_C	Critic VQC parameters
θ'_A	Target actor VQC parameters
θ'_C	Target critic VQC parameters
ϕ	Policy translator DNN parameters
w_{NN}	Surrogate DNN parameters
ξ	Optimizer learning rate
F	Soft update coefficient
L_A	Number of layers in actor VQC
G_A	Number of parameterized gates per layer in actor VQC
L_C	Number of layers in critic VQC
G_C	Number of parameterized gates per layer in critic VQC
L_P	Number of hidden layer of policy translator DNN
L_S	Number of hidden layer of surrogate DNN

the Pauli-Z expectation on qubit j, and $\omega_{k,j}$ are linear weights that allow constructing an N_a-dimensional action vector from the n_s-qubit output of the actor VQC. If $\log_2 N_a \leq n_s$, each $\mathbf{Z}_j$ acts on a single qubit: $\mathbf{Z}_j = [[1,0],[0,-1]]$. If $\log_2 N_a > n_s$ or higher expressivity is desired, $\mathbf{Z}_j$ can be generalized to act on multi-qubit operators, called Pauli string, allowing linear combinations of qubit expectation values to produce the required N_a outputs.

B. Quantum Critic

Let θ_C denote the gate rotation angles of the critic VQC. The critic receives as input the quantum-encoded state together with the quantum action, and transforms them into a quantum Q-value $|Q_t^Q\rangle$:

$$|Q_t^Q\rangle = \mathcal{U}(\theta_C)\left(|s_t^Q\rangle \otimes |a_t^Q\rangle\right), \quad |Q_t^Q\rangle \in \mathbb{C}^{2^{2n_s}}, \tag{21}$$

where the joint quantum state fed into the critic VQC is written as a tensor product: $|s_t^Q\rangle \otimes |a_t^Q\rangle = \sum_{i,j} s_i a_j |ij\rangle$.

To obtain a scalar value suitable for RL updates, the quantum Q-value $|Q_t^Q\rangle$ is measured with the Pauli-Z operator, yielding the classical Q-value estimate:

$$Q_t = \langle Q_t^Q | \mathbf{Z} | Q_t^Q \rangle. \tag{22}$$

This classical Q-value is used in the Bellman update and policy gradient computation, which in turn drive the classical optimizers to update both the actor parameters θ_A and critic parameters θ_C.

C. Replay, Exploration, and Parameter Updates

1) Replay Buffer and Mini-Batch Sampling: In Q-DDPG-CPD, experience replay and action exploration are implemented in the classical domain. The agent's interactions with the environment, called transitions (s_t, a_t, r_t, s_{t+1}), are stored in a replay buffer. During training, a mini-batch of experiences is sampled randomly from this buffer to break temporal correlations and stabilize learning. Continuous action exploration is achieved by adding Gaussian noise to the actor's deterministic output a_t:

$$a_t^{ep} = a_t + \mathcal{N}(0, \sigma_a^2), \tag{23}$$

where $\mathcal{N}(0, \sigma_a^2)$ is a zero-mean Gaussian with variance σ_a^2.

2) Quantum Actor-Critic Forward Pass: For each transition in the mini-batch, the next state s_{t+1} is encoded into a quantum state $|s_{t+1}^Q\rangle$ using a quantum amplitude encoder, similarly to (17). This quantum state is fed into the target actor VQC, parameterized by θ'_A, to generate the quantum next action $|a_{t+1}^Q\rangle$. The quantum next action is then passed to the target critic VQC, parameterized by θ'_C, along with $|s_{t+1}^Q\rangle$ to produce the quantum target Q-value $|Q_{t+1}^Q\rangle$:

$$|Q_{t+1}^Q\rangle = \mathcal{U}(\theta'_C)\left(|s_{t+1}^Q\rangle \otimes |a_{t+1}^Q\rangle\right). \tag{24}$$

3) Training Objective and Parameter Updates: The quantum target Q-value $|Q_{t+1}^Q\rangle$ is measured using a Pauli-Z expectation to obtain the classical scalar, which is used in the Bellman target computation:

$$y_t = r_t + \kappa \langle Q_{t+1}^Q | \mathbf{Z} | Q_{t+1}^Q \rangle. \tag{25}$$

The critic VQC parameters θ_C are updated by minimizing the mean squared error (MSE) between the measured Q-value and the Bellman target:

$$\mathcal{L}(\theta_C) = (Q_t - y_t)^2. \tag{26}$$

Let $\nabla_{\theta_C}\mathcal{L}$ denote the gradient vector of the loss with respect to all parameters θ_C. The gradient of the loss with respect to a parameter $\theta_{C,j}$ of θ_C is computed using the parameter-shift rule [25] as follows:

$$\frac{\partial \mathcal{L}}{\partial \theta_{C,j}} = \frac{\partial \mathcal{L}}{\partial Q_t} \cdot \frac{\partial Q_t}{\partial \theta_{C,j}}, \tag{27a}$$

$$\frac{\partial Q_t}{\partial \theta_{C,j}} = \frac{1}{2}\left(Q_t\left(\theta_{C,j} + \frac{\pi}{2}\right) - Q_t\left(\theta_{C,j} - \frac{\pi}{2}\right)\right). \tag{27b}$$

The critic parameters are then updated via classical gradient descent:

$$\theta_C \leftarrow \theta_C - \xi \nabla_{\theta_C}\mathcal{L}, \tag{28}$$

where ξ is the learning rate in the optimizer.

The actor VQC parameters θ_A are updated using the deterministic policy gradient, aiming to maximize the critic's Q-value estimate for the actor's proposed action:

$$\nabla_{\theta_A}\mathcal{J} \approx \mathbb{E}_{s_t \sim \mathcal{D}}\left[\nabla_{a_t} Q(s_t, a_t; \theta_C) \nabla_{\theta_A} a_t\right], \tag{29}$$

where $\mathcal{J}$ represents the expected return under the current policy, $\mathcal{D}$ is the replay buffer distribution, and $\nabla_{\theta_A} a_t$ is computed via the parameter-shift rule similar to (27).

Algorithm 1: Proposed Q-DDPG-CPD algorithm for LEO satellite resource allocation.

Input : $\mathbf{h}_t, \mathbf{\Delta f}_D$
Output: R_L^*, $\mathbf{P}_L^*$, w_{NN}
Initialize actor and critic VQC parameters, target actor and critic VQC parameters, policy translator and surrogate DNN parameters, replay buffer, soft update coefficient, learning rate.
for *episode* $ep = 1 : N_{ep}$ **do**
  Receive initial state s_1;
  for *time-step* $t = 1 : T$ **do**
    Encode state s_t into quantum state s_t^Q using (17);
    Select a quantum action a_t^Q using the actor VQC as in (18), measure the quantum action to get classical action a_t using (20), and do action exploration using (23);
    Perform a_t and get reward r_t;
    Update quantum state-action Q-value function Q_t^Q using (21), and measure the quantum Q-value using (22);
    Observe the transition from s_t to s_{t+1};
    Store experiences $< s_t, a_t, r_t, s_{t+1} > \rightarrow$ replay buffer;
    if *replay buffer is full* **then**
      Delete the oldest experiences;
    end if
    Sample a mini batch of experiences;
    Encode state s_{t+1} into quantum state s_{t+1}^Q using (17);
    Compute Bellman target values as in (25);
    Update critic VQC network using (28);
    Update actor VQC network using (30);
    Update target actor and critic VQC networks using (31);
    Obtain the surrogate DNN parameters w_{NN} using the policy translator as in (32);
    Update the surrogate DNN parameters;
    Obtain the surrogate action $\hat{a}_t$ using the surrogate DNN;
    Update the policy translator DNN using (34);
  end for
end for

The actor parameters are updated as follows:

$$\theta_A \leftarrow \theta_A - \xi \nabla_{\theta_A} \mathcal{J}. \tag{30}$$

The target actor and critic VQCs are updated using a soft update to stabilize training:

$$\theta'_A \leftarrow \digamma \theta_A + (1 - \digamma)\theta'_A, \tag{31a}$$
$$\theta'_C \leftarrow \digamma \theta_C + (1 - \digamma)\theta'_C, \tag{31b}$$

where $0 < \digamma \ll 1$ is the soft update coefficient.

D. Policy Translator and Surrogate DNN

The policy translator is a classical DNN $g(\cdot;\phi)$ parameterized by ϕ, whose role is to map the actor VQC parameters θ_A into the weights of a surrogate classical actor network. Formally, the translator outputs

$$w_{NN} = g\left(\theta_A; \phi\right), \tag{32}$$

where w_{NN} represents the parameter vector of the surrogate DNN. The surrogate DNN has the same structure as the deployable inference agent onboard the LEO satellite, and it generates a surrogate action $\hat{a}_t$ given the environment state s_t.

The policy translator is trained by minimizing the MSE between the quantum actor measured action a_t and the surrogate action $\hat{a}_t$:

$$\mathcal{L}(\phi) = \mathbb{E}\left[\|a_t - \hat{a}_t\|^2\right]. \tag{33}$$

The policy translator parameters are updated as

$$\phi \leftarrow \phi - \xi \nabla_\phi \mathcal{L}(\phi). \tag{34}$$

This supervised alignment ensures that the surrogate DNN learns to reproduce the quantum actor's actions, and the policy translator provides a direct mapping $\theta_A \rightarrow w_{NN}$, enabling seamless deployment of the trained policy for LEO onboard inference without requiring quantum computation. The operation of Q-DDPG-CPD is described in Algorithm 1.

Remark 1. *For multi-LEO deployments, the proposed quantum-DRL-based beam size and power control framework can be generalized to a multi-agent setting, where each LEO satellite acts as an agent optimizing its local beams while treating the interference from neighboring satellites as part of the environment. Coordination can be realized via centralized training with distributed execution or limited inter-satellite information exchange, which preserves scalability while accounting for inter-LEO coupling.*

E. Convergence Analysis

The convergence of the proposed Q-DDPG-CPD is discussed in the following theoretical analysis.

With $Q_t = \langle Q_t^Q | \mathbf{Z} | Q_t^Q \rangle$ is the measured classical Q-value of the critic VQC at time t and y_t as in (25) is the classical Bellman target obtained from the next quantum state, for each state s_t, let's define a mapping $\mathcal{F}_{s_t}(Q_t)$ as

$$\mathcal{F}_{s_t}(Q_t) = \mathbb{E}\left[r_t + \kappa Q_{t+1} | s_t, a_{t+1} = \underset{\langle a^Q | \mathbf{Z} | a^Q \rangle}{\arg\max}\, Q_{t+1}(s_{t+1}, a^Q)\right]. \tag{35}$$

Here, a_{t+1} is the classical action obtained from measuring the optimal quantum action a^Q from the actor VQC.

The critic update in (28) can be viewed as a discrete-time iteration approximating the fixed point of $\mathcal{F}$. Convergence of the critic is guaranteed under the following standard conditions: (i) smoothness of measured Q-values with respect to θ_C:

$$\exists \mathcal{L}(\theta_C) > 0 \quad \text{such that}$$
$$\|\nabla_{\theta_C} Q_t(\theta_C) - \nabla_{\theta_C} Q_t(\theta'_C)\| \leq \mathcal{L}(\theta_C)\|\theta_C - \theta'_C\|,\ \forall \theta_C, \theta'_C; \tag{36}$$

(ii) bounded rewards; and (iii) learning rate conditions:

$$\xi_t > 0, \quad \sum_{t=1}^{\infty} \xi_t = \infty, \quad \sum_{t=1}^{\infty} \xi_t^2 < \infty. \tag{37}$$

Under these conditions, the measure Q-value sequence $\{Q_t\}$ converges to a stationary point of the Bellman operator with probability 1, which implies that the critic VQC parameters θ_C converge in the sense that they produce Q-values consistent with the Bellman target.

The actor VQC parameters θ_A are updated using the deterministic policy gradient in (30). Since the critic has approximately converged, the actor updates follow the gradient of the expected return with respect to the measured actions a_t. Under smoothness and boundedness of the quantum unitaries, the sequence of actor parameters $\{\theta_A\}$ converges to a stationary point of the expected return function, in the sense that the measured actions a_t stabilize.

Because the actor evolves on a slower timescale than the critic, and the policy translator DNN tracks the actor outputs, the translator parameters ϕ converge to a stationary mapping reproducing the measured quantum actor outputs under the Lipschitz continuity:

$$\|g(\theta_A;\phi) - g(\theta'_A;\phi)\| \leq \mathcal{L}(\phi)\|\theta_A - \theta'_A\|, \; \forall \theta_A, \theta'_A. \tag{38}$$

F. Computational Complexity Analysis

The computational complexity analysis in this section assumes that any operation whose input or output is a quantum state must be executed on a quantum computer, while classical operations are performed on a classical computer.

The quantum amplitude encoding requires N_s elementary gates, which act in parallel on superpositions, resulting in a time complexity of $\mathcal{O}(N_s)$. Let L_A and L_C be the number of layers in the actor and critic VQCs, respectively, and let G_A and G_B denote the number of parameterized gates per layer in the actor and critic VQCs, respectively. The actor VQC forward pass results in a complexity of $\mathcal{O}(L_A G_A)$ assuming each gate application is counted as one basic operation. The Pauli-Z expectation measurements are performed on each of the n_s qubits that produces N_a action components in parallel from the same actor VQC execution. Since the measurement itself is a projective readout, its asymptotic cost is treated as $\mathcal{O}(1)$ relative to the circuit preparation. Similarly, the critic VQC forward yields a complexity of $\mathcal{O}(L_C G_C)$.

For VQC parameter update, each parameterized gate in the VQC requires two forward evaluations per gradient computation when using the parameter-shift rule. Thus, the parameter update for both actor and critic VQCs results in a complexity of $\mathcal{O}(2L_A G_A + 2L_C G_C)$.

Let L_P and L_S denote the number of hidden layers in the policy translator and the surrogate DNN, respectively, and let Z_l denote the number of neurons of hidden layer l. The policy translator DNN forward pass costs $\mathcal{O}(|\theta_A|Z_1 + \sum_{l=1}^{L_P} Z_l Z_{l+1})$, and the policy translator parameter update has a complexity of $\mathcal{O}(|\phi|)$. Lastly, the complexity of the surrogate DNN forward pass is $\mathcal{O}(|\mathcal{S}|Z_1 + \sum_{l=1}^{L_S} Z_l Z_{l+1})$, and the complexity for parameter update is $\mathcal{O}(|w_{NN}|)$.

Total complexity of the proposed Q-DDPG-CPD can be expressed as

$$\begin{aligned} \mathcal{O}\Bigg(N_{ep} T \mathfrak{B} \Big(& 2N_s + 4\left(L_A G_A + L_C G_C\right) \\ & + |\theta_A| Z_1 + \sum_{l=1}^{L_P} Z_l Z_{l+1} + |\phi| \\ & + |\mathcal{S}| Z_1 + \sum_{l=1}^{L_S} Z_l Z_{l+1} + |w_{NN}| \Big) \Bigg), \end{aligned} \tag{39}$$

where N_{ep} is the number of training episodes, T is the number of time steps per episode, and $\mathfrak{B}$ is the mini batch size.

V. Numerical Results

This section presents the performance evaluation of the proposed Q-DDPG-CPD and compares it with several classical DRL benchmarks designed to accelerate the learning process of DDPG:

- **Benchmark 1: Conventional DDPG** [26] - the standard DDPG algorithm without any modification;
- **Benchmark 2: DPDS-DPG** [27] - incorporates the post-decision state into DDPG to enhance learning speed;
- **Benchmark 3: FPDS-DDPG** [11] - builds on DPDS-DPG by using fuzzy logic instead of random action exploration to further accelerate the agent's learning;

and some non-DRL benchmarks as follows:

- **Benchmark 4: Beam size design without Doppler** [28] - LEO beam radius optimization based on elevation angle without considering the effect of residual Doppler frequency offset;
- **Benchmark 5: LEO beam transmit power control** [29] - continuous LEO transmit power allocation optimization.

A. Environment Setup

The simulation environment is constructed as a multi-beam GEO-LEO satellite network, configured according to the parameters summarized in Table III. The coverage area corresponds to a single LEO footprint, modeled as a circular region with radius of $2,350 km$ and centered at the simulator's origin. The GEO satellite is configured with seven beams, each having a radius of $555\ km$, and employs a four-color frequency reuse pattern. One GEO user is assigned to each GEO beam, while one LEO user is associated with each LEO beam. The satellite-to-ground communication channels are characterized by Shadowed-Rician fading, parameterized as $\{m, b, w\} = \{4, 0.126, 0.835\}$. The Doppler frequency offset is computed based on real-time orbital information from the LEO satellite Iridium-NEXT 914 [30].

B. Quantum RL Agent Setup

The quantum RL agent is implemented using VQCs constructed on a 16-qubit system simulated via the Qiskit backend within the PennyLane hybrid quantum-classical framework. The agent employs PyTorch double-precision tensors to ensure numerical stability and facilitate gradient-based optimization.

TABLE III: Environment simulator's parameters

Parameter	Value
Carrier frequency	1.53 GHz
Speed of light	3×10^5 km/s
GEO/LEO altitude	35,786/785 km
LEO footprint radius	$2,350$ km
GEO spot beam radius	555 km
GEO frequency reuse factor	4
GEO bandwidth	800 kHz
GEO transmit power per beam	100 W
GEO/LEO antenna diameter	2.4/0.2 m
Antenna efficiency	0.55
Noise power	-120 dBm/Hz

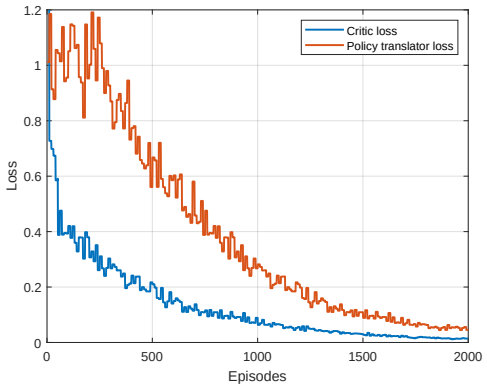

Fig. 3: Critic loss and policy translator loss of the proposed Q-DDPG-CPD agent.

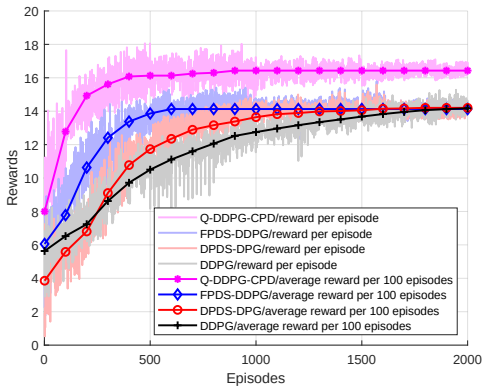

Fig. 4: Reward convergence comparison between the proposed quantum DRL and classical DRL benchmarks.

The circuit architecture begins with an amplitude encoder, which embeds classical state to the quantum state space by mapping normalized feature amplitudes through a series of parameterized R_X and R_Z rotations applied to each qubit. The variational core of the model comprises four layers, each consisting of a chain of controlled-NOT gates that entangle adjacent qubits, followed by 128 trainable rotational gates distributed across all qubits. These parameterized single-qubit rotations introduce nonlinear transformations and enhance the expressive capacity of the circuit. The policy translator and surrogate DNNs both consist of two fully connected hidden layers, in which each layer of the policy translator DNN contains 256 neurons and the surrogate DNN has 256 and 128 neurons, respectively. The agent is trained using a learning rate of 0.001, a reward discount factor of 0.99, a mini batch of size 64, a time step duration of 5 ms, and the maximum time step per training episode of 10000.

C. Simulation Results

Note that in the simulation setup, the surrogate DNN of the proposed Q-DDPG-CPD and the actor networks of all classical DRL benchmarks are configured with identical architectures (same number of layers, neurons, and activation functions), ensuring strictly capacity-matched models. Consequently, all performance comparisons at the inference stage are conducted under the same network sizes and configurations, so that any observed gains cannot be attributed to increased classical model capacity.

1) Proposed Quantum Learning Performance: Fig. 3 illustrates the convergence behavior of the critic loss and the policy translator loss of the proposed Q-DDPG-CPD agent over 2000 training episodes. Both losses exhibit a consistent downward trend, indicating stable learning and effective policy optimization and translation. The critic loss decreases rapidly during the early training phase within the first 300 episodes, reflecting efficient value-function estimation and accurate Q-value prediction. In contrast, the policy translator loss starts at a higher magnitude and shows a more gradual decline, signifying that the mapping from the actor VQC parameters into the weights of the surrogate DNN requires a longer adaptation period. After approximately 1500 episodes, both losses reach convergence, maintaining low and stable values.

Fig. 4 presents the reward convergence profiles of the proposed quantum RL agent and the DRL benchmark agents. The proposed quantum agent achieves the highest cumulative and average rewards throughout training, demonstrating superior policy learning efficiency and stability. The reward curve of Q-DDPG-CPD rises sharply during the initial 500 episodes, indicating rapid adaptation and effective exploitation of the quantum state-action representation. After approximately 1000 episodes, the reward stabilizes at a higher level compared to the other baselines, confirming improved convergence performance. In contrast, the classical DDPG and DPDS-DDPG agents exhibit slower reward growth and lower asymptotic performance, implying less efficient exploration and suboptimal policy updates. The FPDS-DDPG variant shows moderate improvement over DDPG but still lags behind Q-DDPG-CPD.

2) Effect of LEO Beam Radius and Transmit Power on SINR at GEO User: Figs. 5 and 6 show the effect of LEO beam radius and transmit power on the SINR at GEO user when the LEO system employs full frequency reuse and no frequency reuse schemes, respectively. Under full frequency reuse scheme, the SINR at GEO user varies from 5.97 dB to 6.73 dB, while under no frequency reuse scheme, it varies from -4.45 dB to 8.48 dB.

3) Spectral Efficiency of LEO System: In this section, quantum-DDPG without policy translator agent refers to the proposed quantum-DDPG agent that requires quantum device during inference deployment stage.

Figs. 7 to 10 present the spectral efficiency of LEO system under the optimal beam radius and transmit power at an

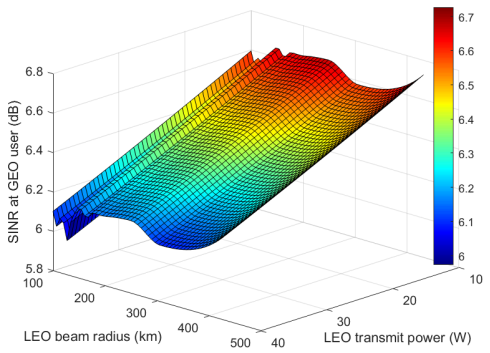

Fig. 5: SINR at GEO user when LEO system employs full frequency reuse scheme.

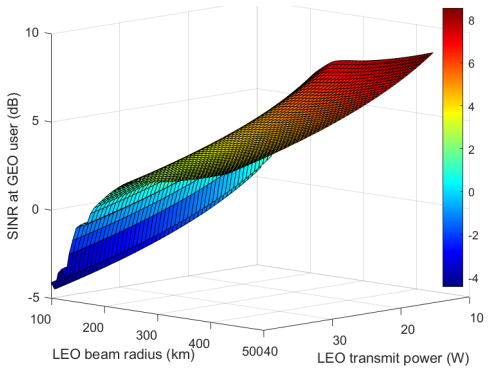

Fig. 6: SINR at GEO user when LEO system employs no frequency reuse scheme.

average SINR at LEO users of $11.72\,dB$. The optimal beam radius shows slight variation between the full frequency reuse and no frequency reuse schemes. Nonetheless, both schemes yield the same optimal number of LEO beams per GEO beam, represented as N^*_{GL} in Figs. 7 and 8. In Figs. 9 and 10, the optimal transmit power for both frequency reuse schemes is similar when using DRL and non-DRL benchmarks, at $18.32\,W$ and $13.74\,W$, respectively. In contrast, employing the quantum-DDPG and Q-DDPG-CPD agents under the full frequency reuse scheme yields $20.57\,W$ and $19.90\,W$, respectively, while under the no frequency reuse scheme, the results are $20.18\,W$ and $19.76\,W$, respectively.

Figs. 11 and 12 provide an overview of the performance of the proposed quantum RL compared to DRL and non-DRL benchmarks. Under the full frequency reuse scheme, the quantum-DDPG agent achieves the highest spectral efficiency, followed by the proposed Q-DDPG-CPD agent with 16.05% lower efficiency, the DRL agent with 38.77% lower, and the non-DRL benchmark with 49.38% lower. Under no frequency reuse scheme, the quantum-DDPG agent again performs best, with the Q-DDPG-CPD agent at 15.21% lower, the DRL agent at 45.30% lower, and non-DRL benchmark at 57.49% lower. This performance trend is expected because the Q-DDPG-CPD agent employs a policy translator that maps the quantum-DDPG agent's performance from quantum devices to classical devices. The effectiveness of this policy translator is limited, achieving approximately 85% of the original performance.

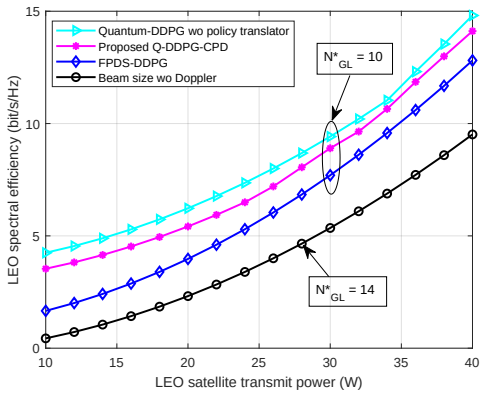

Fig. 7: Performance comparison under different LEO power budget when employing full frequency reuse scheme.

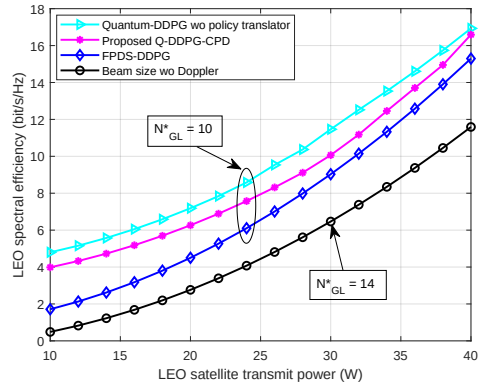

Fig. 8: Performance comparison under different LEO power budget when employing no frequency reuse scheme.

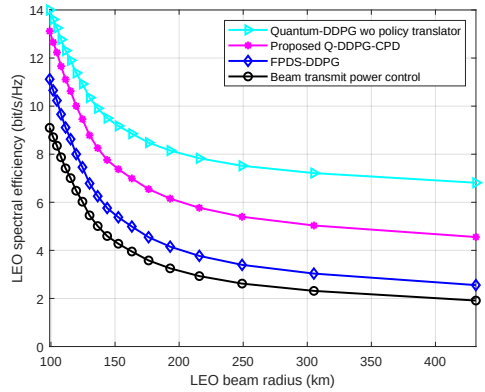

Fig. 9: Performance comparison under different LEO beam radius when employing full frequency reuse scheme.

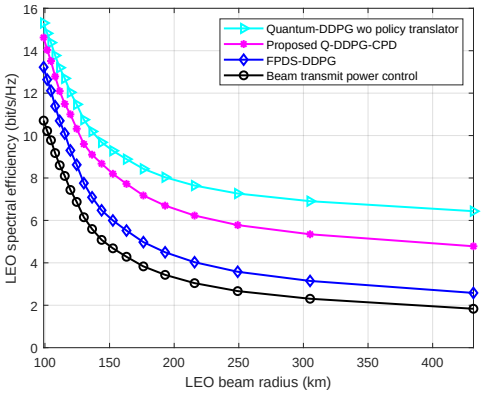

Fig. 10: Performance comparison under different LEO beam radius when employing no frequency reuse scheme.

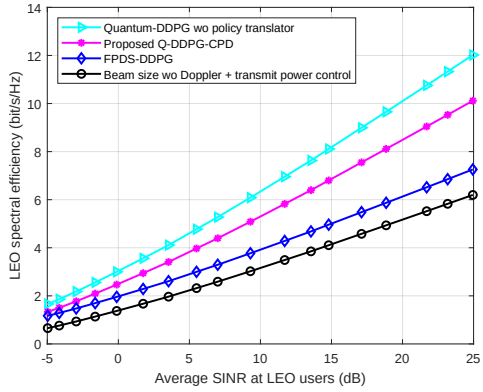

Fig. 11: Performance comparison under different LEO SINR when employing full frequency reuse scheme.

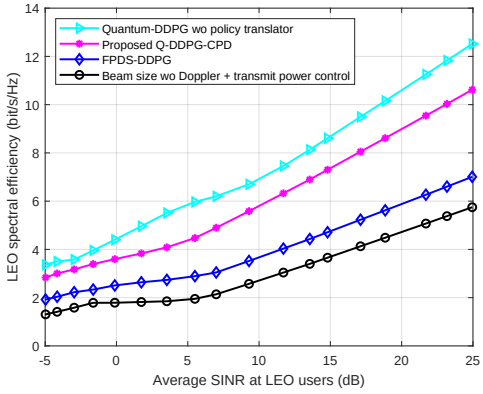

Fig. 12: Performance comparison under different LEO SINR when employing no frequency reuse scheme.

4) Deployment Feasibility on LEO Satellite Hardware: To access the deployment readiness of the proposed Q-DDPG-CPD framework on LEO satellites, the onboard inference latency of the translated surrogate DNN is analyzed for different network sizes and representative space-qualified processing platforms. Note that only the lightweight classical surrogate policy network, rather than the quantum training module, is required to be executed on board. Table IV reports the estimated forward-pass latency for four surrogate DNN configurations on typical LEO satellite onboard computers and data processing units [31].

The inference latency values are obtained from analytical floating-point operation (FLOP) counting of fully connected layers, assuming single-precision arithmetic and conservative peak throughput of each processor. This provides an upper-bound estimate under worst-case execution without aggressive parallelization or hardware acceleration. Even for the largest considered network, the resulting inference delay remains in the order of microseconds on radiation-tolerant CPUs and sub-microseconds on modern space-grade SoCs and AI accelerators, which is several orders of magnitude smaller than the adopted DRL decision time step of 5 ms. For L-band LEO downlinks, after Doppler precompensation, the residual Doppler results in a channel coherence time on the order of 5–30 ms, indicating that the instantaneous CSI can be regarded as quasi-static within one decision interval and does not become significantly outdated during policy inference. Therefore, the end-to-end processing time of the proposed algorithm, including state acquisition and surrogate DNN inference, is well within both the channel coherence time and the MAC-layer resource reconfiguration timescale. These results confirm that the proposed surrogate DNN satisfies the real-time and memory constraints of current LEO satellite hardware, thereby supporting the practical deployability of the Q-DDPG-CPD framework.

VI. Practical Considerations and Scope of Quantum Reinforcement Learning

While most studies on quantum-enhanced DRL employ amplitude encoding and variational quantum circuits (VQCs) [16]–[18], [20]–[22], these approaches currently face significant practical constraints on near-term quantum hardware. In particular, state preparation for amplitude encoding requires large circuit depth, and existing noisy intermediate-scale quantum (NISQ) devices are limited by a small number of qubits, short coherence times, and imperfect gate fidelities [32]. Consequently, full-scale experimental implementation remains largely unfeasible, and most existing results are obtained using simulated quantum backends.

To bridge the gap between algorithmic potential and hardware feasibility, several hardware-aware alternatives have been proposed. Angle encoding or basis encoding can reduce state preparation overhead by directly mapping classical data to qubit rotation angles or computational basis states, thereby enabling shallower circuits [33]. Similarly, hardware-efficient ansatz designs allow VQCs to be implemented with reduced depth and fewer gates, trading off some expressivity for

TABLE IV: Inference latency estimation on representative LEO satellite hardware

LEO hardware [31]	**Assumed FP32 throughput**	**Surrogate DNN structure / inference latency**			
		128-64	**256-128**	**512-256**	**1024-512**
		70.27 kFLOPs	173.31 kFLOPs	0.48 MFLOPs	1.48 MFLOPs
Sirius OBC – LEON3FT at 100 MHz	100 MFLOPS	0.00070 ms	0.00173 ms	0.00478 ms	0.0148 ms
Gen-2 OBC – Cortex-A72 (2.0 GHz)	10 GFLOPS	0.000007 ms	0.000017 ms	0.000048 ms	0.000148 ms
Antelope DPU – Zynq Ultrascale+ A53	6 GFLOPS	0.000012 ms	0.000029 ms	0.000080 ms	0.000246 ms
RDP-23FV – Versal AI Engine	1 TFLOPS	$7.0e^{-8}$ ms	$1.7e^{-7}$ ms	$4.8e^{-7}$ ms	$1.5e^{-6}$ ms

practical implementability [34]. These strategies demonstrate potential pathways toward realizing quantum DRL on near-term devices, albeit with limitations in scalability and representational power.

This work focuses on algorithmic feasibility and performance evaluation under idealized simulation conditions. We aim to explore the potential benefits of quantum RL in terms of convergence speed and policy expressivity, rather than immediate real-world deployment. Experimental realization on physical quantum processors is therefore identified as future work, contingent on advances in quantum hardware and error mitigation techniques.

VII. Conclusions

This article addressed the resource allocation problem for LEO satellites by jointly optimizing beam size and transmit power, while considering GEO interference constraints, residual Doppler frequency offsets, and frequency reuse strategies. The objective was to maximize the spectral efficiency of LEO systems operating within multi-beam GEO-LEO networks. We proposed a quantum RL framework that is deployable on LEO satellites with classical hardware. The proposed quantum agent, equipped with a policy translator for classical deployment, achieved approximately 85% of the quantum-DDPG performance, enabling practical onboard LEO implementation without significant loss. These results confirm that leveraging quantum acceleration during offline training can significantly enhance spectral efficiency while providing a deployment-ready classical representation compatible with existing LEO satellite hardware. The proposed framework not only reduces retraining latency but also offers a scalable solution for future multi-beam GEO-LEO networks, demonstrating the practical potential of quantum RL in satellite communications.

References

[1] Q. T. Ngo, Z. Tang, B. Jayawickrama, and et. al., "Timeliness of information in 5G non-terrestrial networks: A survey," *IEEE Internet Things J.*, vol. 11, no. 21, pp. 34 652–34 675, 2024.

[2] International Telecommunication Union, "Operation of earth stations in motion communicating with geostationary space stations in the fixed-satellite service allocations," 2019, ITU-R.

[3] International Telecommunications Union, "Radio regulations. Chapter VI, Provision of servies and stations. Article 22, Space Services," (2016). ITU-R.

[4] N. Heydarishahreza, T. Han, and N. Ansari, "Spectrum sharing and interference management for 6G LEO satellite-terrestrial network integration," *IEEE Commun. Surveys Tuts.*, pp. 1–1, 2024.

[5] A. Guidotti, A. Vanelli-Coralli, and et. al., "Role and evolution of non-terrestrial networks toward 6G systems," *IEEE Access*, vol. 12, pp. 55 945–55 963, 2024.

[6] Z. Zheng and et. al., "Cooperative multi-satellite and multi-RIS beamforming: Enhancing LEO satcom and mitigating LEO-GEO intersystem interference," *IEEE J. Select. Areas Commun.*, vol. 43, no. 1, pp. 279–296, 2025.

[7] M. He, G. Cui, M. Wu, and W. Wang, "Collaborative interference avoidance technology in GEO-LEO co-existing satellite system," *Int. J. Satell. Commun. Network.*, vol. 42, no. 4, pp. 257–272, 2022.

[8] D. Yan, Y. He, and H. Fu, "Interference analysis of NGSO constellation to GEO satellite communication system based on spatio–temporal slices," *IEEE Internet Things J.*, vol. 10, no. 18, pp. 16 605–16 616, 2023.

[9] B. Li, J. Park, A. Al-Hourani, S. R. Pokhrel, and J. Choi, "A novel frequency reuse model for co-existing LEO and GEO satellites," *IEEE Wireless Commun. Lett.*, vol. 13, no. 4, pp. 1024–1028, 2024.

[10] K. Weththasinghe, Q. T. Ngo, Y. He, and B. Jayawickrama, "Optimising beam size in multibeam LEO satellite networks: Addressing interbeam interference, doppler shift, and frequency reuse," *IEEE Trans. Aerospace and Electronics Syst.*, vol. 61, no. 3, pp. 5871–5884, 2025.

[11] Q. T. Ngo and et. al., "A fast fuzzy DRL-based joint beam design and power allocation for multi-beam GEO-LEO coexisting satellite networks," *IEEE Trans. Wireless Commun.*, vol. 24, no. 10, pp. 1558–2248, 2025.

[12] Q. T. Ngo, B. Jayawickrama, Y. He, and E. Dutkiewicz, "A novel satellite-based REM construction in cognitive GEO-LEO satellite IoT networks," *IEEE Internet Things J.*, vol. 12, no. 6, pp. 7532–7548, 2025.

[13] W. Fan and et. al., "Satellite edge intelligence: DRL-based resource management for task inference in LEO-based satellite-ground collaborative networks," *IEEE Trans. Mobile Comput.*, pp. 1–18, 2025.

[14] S. C. Prabhashana and et. al., "Machine learning-based resource allocation in 6G integrated space and terrestrial networks-aided intelligent autonomous transportation," *IEEE Trans. Intell. Transportation Syst.*, pp. 1–13, 2025.

[15] Q. T. Ngo, Y. He, B. Jayawickrama, and E. Dutkiewicz, "Multi-agent DDPG-based joint beam hopping and resource allocation for cognitive GEO-LEO satellite networks," *IEEE Netw. Lett.*, pp. 1–1, 2025.

[16] S. Huang and et. al., "Edge intelligence in satellite-terrestrial networks with hybrid quantum computing," *IEEE Wireless Commun. Lett.*, vol. 14, no. 5, pp. 1341–1345, 2025.

[17] A. Paul and et. al., "Quantum-enhanced DRL optimization for DoA estimation and task offloading in ISAC systems," *IEEE J. Select. Areas Commun.*, vol. 43, no. 1, pp. 364–381, 2025.

[18] G. S. Kim and et. al., "Quantum multi-agent reinforcement learning for cooperative mobile access in space-air-ground integrated networks," *IEEE Trans. Mobile Comput.*, pp. 1–18, 2025.

[19] M. Frackiewicz, "Orbital quantum leap: First photonic edge-computing satellite set to transform space data processing," TS2 Space, 2025. [Online]. Available: https://ts2.tech/en/orbital-quantum-leap-first-photonic-edge-computing-satellite/

[20] W. Shi and et. al., "Optimal energy management for multistack fuel cell vehicles based on hybrid quantum reinforcement learning," *IEEE Trans. Transportation Electrification*, vol. 11, no. 3, pp. 8500–8511, 2025.

[21] J. A. Ansere, S. C. Prabhashana, O. A. Dobre, and T. Q. Duong, "Quantum machine learning DDPG for digital twin semantic vehicular networks," in *Proc. IEEE Int. Conf. Machine Learning Commun. Netw.*, 2025, pp. 1–6.

[22] X. Wei and et. al., "A quantum reinforcement learning approach for joint resource allocation and task offloading in mobile edge computing," *IEEE Trans. Mobile Comput.*, vol. 24, no. 4, pp. 2580–2593, 2025.

[23] J. Gonzalez-Conde, T. W. Watts, P. Rodriguez-Grasa, and M. Sanz, "Efficient quantum amplitude encoding of polynomial functions," *Quantum*, vol. 8, p. 1297, 2024.

[24] O. Crawford, B. v. Straaten, D. Wang, T. Parks, E. Campbell, and S. Brierley, "Efficient quantum measurement of Pauli operators in the presence of finite sampling error," *Quantum*, vol. 5, p. 385, 2021.

[25] G. E. Crooks, "Gradients of parameterized quantum gates using the parameter-shift rule and gate decomposition," *arXiv e-prints*, pp. arXiv–1905, 2019.
[26] T. Lillicrap, J. Hunt, A. Pritzel, and N. Heess, "Continuous control with deep reinforcement learning," in *Proc. Int. Conf. Learning Representations*, 2016, pp. 1–6.
[27] Q. T. Ngo, K. Phan, A. Mahmood, and W. Xiang, "Hybrid IRS-assisted secure satellite downlink communications: A fast deep reinforcement learning approach," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 8, no. 4, pp. 2858–2869, 2024.
[28] S. Han, W. Shin, and J.-H. Kim, "Evaluation for elevation angle-dependent beam size in NR NTN systems," in *Proc. Int. Conf. Information Commun. Technol. Convergence (ICTC)*, 2023, pp. 321–323.
[29] P. Gu, R. Li, C. Hua, and R. Tafazolli, "Dynamic cooperative spectrum sharing in a multi-beam LEO-GEO co-existing satellite system," *IEEE Trans. Wireless Commun.*, vol. 21, no. 2, pp. 1170–1182, 2022.
[30] Space-Track, "Satellite catalog," Sep 2025. [Online]. Available: https://space-track.org/catalog.
[31] National Aeronautics and Space Administration (NASA), "State-of-the-art of small spacecraft technology," March 2025. [Online]. Available: https://www.nasa.gov/smallsat-institute/sst-soa/small-spacecraft-avionics/.
[32] D. D. Awschalom, H. Bernien, R. Hanson, W. D. Oliver, and J. Vučković, "Challenges and opportunities for quantum information hardware," *Science*, vol. 390, no. 6777, pp. 1004–1010, 2025.
[33] M. A. Khan, M. N. Aman, and B. Sikdar, "Beyond bits: A review of quantum embedding techniques for efficient information processing," *IEEE Access*, vol. 12, pp. 46 118–46 137, 2024.
[34] D. A. Fedorov, B. Peng, N. Govind, and Y. Alexeev, "VQE method: a short survey and recent developments," *Materials Theory*, vol. 6, no. 1, p. 2, 2022.

QUYNH TU NGO (Senior Member, IEEE) received a B.Sc. in Electrical Engineering (Magna Cum Laude) from California State University Los Angeles, USA, in 2013; an M.Sc. in Telecommunications from University of Sciences, Vietnam, in 2016; and a Ph.D. in Telecommunications and Networks from La Trobe University, Australia, in 2023. She is currently a Postdoctoral Research Fellow at the School of Electrical and Data Engineering, University of Technology Sydney, Australia. Her research interests include satellite communications, IoT networks, intelligent non-terrestrial networks, and AI/ML in 5G/6G.

YING HE (Senior Member, IEEE) received the B.Eng. degree in Telecommunications Engineering from Beijing University of Posts and Telecommunications, China, in 2009, and the Ph.D. degree in Telecommunications Engineering from the University of Technology Sydney, Australia, in 2017. She is currently a Senior Lecturer with the School of Electrical and Data Engineering, University of Technology Sydney. Her research interests are physical layer algorithms in wireless communication with machine learning, vehicular communication, spectrum sharing and satellite communication.

BEESHANGA JAYAWICKRAMA (Senior Member, IEEE) received the B.E. degree (Hons. I) in Telecommunications Engineering and the Ph.D. degree in Electronic Engineering from Macquarie University, Australia, in 2011 and 2015, respectively. He is currently affiliated as a Visiting Fellow at University of Technology Sydney, Australia. He was extensively involved in spectrum sensing and interference mitigation research for spectrum access systems. His research interests include non-terrestrial networks, 5G/6G, cognitive radio, and signal processing.

ERYK DUTKIEWICZ (Senior Member, IEEE) received his B.E. degree in Electrical and Electronic Engineering in 1988 and his M.Sc. degree in Applied Mathematics in 1992 from the University of Adelaide, Australia, and his Ph.D. in Telecommunications from the University of Wollongong, Australia, in 1996. His industry experience includes management of the Wireless Research Laboratory at Motorola in early 2000's. Prof. Dutkiewicz is currently Associate Dean International in the Faculty of Engineering and IT at University of Technology Sydney, Australia. He also holds a professorial appointment at Hokkaido University in Japan. His current research interests cover 5G/6G and IoT networks.

SHIVA RAJ POKHREL (Senior Member, IEEE) received the Ph.D. degree in ICT engineering from Swinburne University of Technology, Melbourne, VIC, Australia, in 2017. He is a Distinguished Marie Curie Fellow and a Senior Lecturer with Deakin University, Melbourne. He brings expertise in experimental distributed quantum computing, communication, and large-scale quantum experiments. He specializes in mobile and quantum computing, focusing on machine learning/deep learning architectures, distributed training, and hyperparameter optimization across diverse environments, including public clouds and HPC centers.