Quantum-Enhanced DRL Optimization for DoA Estimation and Task Offloading in ISAC Systems

Anal Paul, *Member, IEEE*, Keshav Singh, *Member, IEEE*, Aryan Kaushik, *Member, IEEE*, Chih-Peng Li, *Fellow, IEEE*, Octavia A. Dobre, *Fellow, IEEE*, Marco Di Renzo, *Fellow, IEEE*, and Trung Q. Duong, *Fellow, IEEE*

***Abstract*—This work proposes a quantum-aided deep reinforcement learning (DRL) framework designed to enhance the accuracy of direction-of-arrival (DoA) estimation and the efficiency of computational task offloading in integrated sensing and communication systems. Traditional DRL approaches face challenges in handling high-dimensional state spaces and ensuring convergence to optimal policies within complex operational environments. The proposed quantum-aided DRL framework that operates in a military surveillance system exploits quantum computing's parallel processing capabilities to encode operational states and actions into quantum states, significantly reducing the dimensionality of the decision space. For the very first time in literature, we propose a quantum-enhanced actor-critic method, utilizing quantum circuits for policy representation and optimization. Through comprehensive simulations, we demonstrate that our framework improves DoA estimation accuracy by 91.66% and 82.61% over existing DRL algorithms with faster convergence rate, and effectively manages the trade-off between sensing and communication and optimizing task offloading decisions under stringent ultra-reliable low-latency communication requirements. Comparative analysis also reveals that our approach reduces the overall task offloading latency by 43.09% and 32.35% compared to the DRL-based deep deterministic policy gradient and proximal policy optimization algorithms, respectively.**

A. Paul, K. Singh, and C-P. Li are with the Institute of Communications Engineering, National Sun Yat-sen University, Kaohsiung 80424, Taiwan (Email: apaul@ieee.org, keshav.singh@mail.nsysu.edu.tw, cpli@faculty.nsysu.edu.tw).

The work of K. Singh and C.-P. Li was supported in part by the National Science and Technology Council of Taiwan under Grants NSTC 112-2221-E-110-038-MY3 and NSTC 112-2221-E-110-029-MY3 and also supported in part by the Sixth Generation Communication and Sensing Research Center funded by the Higher Education SPROUT Project, the Ministry of Education of Taiwan. The work of O. A. Dobre was supported by Canada Research Chairs Program CRC-2022-00187. The work of A. Kaushik was supported by the Higher Education Innovation Fund supported project "Green and Intelligent 6G Connectivity: AI and Holographic Surfaces-assisted Integrated Sensing and Communications". The work of M. Di Renzo was supported in part by the European Commission through the Horizon Europe project titled COVER under grant agreement number 101086228, the Horizon Europe project titled UNITE under grant agreement number 101129618, and the Horizon Europe project titled INSTINCT under grant agreement number 101139161, as well as by the Agence Nationale de la Recherche (ANR) through the France 2030 project titled ANR-PEPR Networks of the Future under grant agreement NF-YACARI 22-PEFT-0005, and by the CHIST-ERA project titled PASSIONATE under grant agreements CHIST-ERA-22-WAI-04 and ANR-23-CHR4-0003-01. The work of T. Q. Duong was supported in part by the Canada Excellence Research Chair (CERC) Program CERC-2022-00109. (*Corresponding author: Keshav Singh.*)

A. Kaushik is with the School of Engineering and Informatics, University of Sussex, Brighton BN1 9RH, UK (E-mail: aryan.kaushik@sussex.ac.uk).

O. A. Dobre is with the Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's, NL A1C 5S7, Canada (E-mail: odobre@mun.ca).

M. Di Renzo is with Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des Signaux et Systèmes, 3 Rue Joliot-Curie, 91192 Gif-sur-Yvette, France. (marco.di-renzo@universite-paris-saclay.fr)

T. Q. Duong is with the Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's, NL A1C 5S7, Canada, and also with the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, BT7 1NN Belfast, U.K. (e-mail: tduong@mun.ca).

***Index Terms*—Quantum computing, deep reinforcement learning, direction-of-arrival estimation, vehicular task offloading, surveillance systems, ultra-reliable low-latency communication.**

I. INTRODUCTION

THE advent of integrated sensing and communication (ISAC) systems marks a transformative era in military surveillance, essential for modern warfare [1]. The integration of ground, aerial, and space networks in the sixth-generation (6G) communications is a game-changer that delivers unmatched levels of global connectivity, low-latency communication, accurate sensing capabilities, and distributed task offloading [2]–[4]. These capabilities are instrumental in time-sensitive military surveillance systems, making them an essential component for the next level of military operations. ISAC systems are essential in achieving a dual-purpose goal: real-time environmental sensing for threat detection and dynamic communication for command and control [5]. The authors in [5] proposed reconfigurable intelligent surfaces (RISs) RIS-aided ISAC system to maximize the weighted performance metrics while maintaining robust communication links. Pan *et al.* investigated the use of unmanned aerial vehicles (UAVs) to provide ISAC services [6]. They take advantage of UAV mobility to improve accuracy in target location estimation and ensure quality-of-service (QoS) in communication [6], [7]. Further study revealed that the UAV-mounted RIS is capable of providing reliable coverage for ISAC systems [8], [9].

The direction-of-arrival (DoA) estimation is crucial in the sensing component of ISAC systems, particularly if we consider military surveillance applications [6], [10]. It involves determining the angle at which a received signal arrives, which is crucial for accurately localizing and tracking unauthorized flying objects (UFOs) or potential threats. In their research, Chen et al. [11] investigated the use of passive beamforming with RIS for estimating the DoA of ground vehicles. Multiple measurements were analyzed to determine the ISAC system's theoretical Cramer-Rao lower bound (CRLB) in estimating the DoA [6], [11], [12].

The authors in [13], [14] found that the ISAC-aided military surveillance systems require precise DoA estimation as well as effective task offloading to mobile edge computing (MEC) due to the high computational intensity of processing tasks. Task offloading to the MEC node for the armored vehicles is essential as those vehicles are not equipped with specialized hardware units to process some computationally heavy tasks [14]. This scenario is further complicated in operations where ultra-reliable low-latency communications (URLLC) are essential, necessitating swift processing and dissemination of critical information with minimal delay [15]. Therefore, effective task offloading becomes not just a strategic choice but a necessity to meet the stringent URLLC service constraints. Herein, MEC-enabled non-terrestrial networks (i.e., satellite communication) emerge as a pivotal element, offering seamless integration with terrestrial networks and facilitating the effective task [16]. The satellite-terrestrial-integrated ISAC framework differs from the MEC framework in key ways. ISAC integrates sensing and communication for environmental monitoring, while MEC focuses on reducing latency by offloading tasks to edge servers. ISAC uses a unified infrastructure for efficient resource use, whereas MEC employs distributed edge nodes. ISAC is suited for real-time monitoring and autonomous systems like self-driving cars and drones, while MEC is ideal for low-latency applications such as AR/VR and IoT services. ISAC optimizes spectrum efficiency and ensures accurate object detection, while MEC focuses on reducing latency and improving offloading efficiency.

To optimize military surveillance systems with limited resources for sensing, communication, and computation, implementing a robust resource allocation strategy within the ISAC framework is crucial. The evolution of artificial intelligence and machine learning, followed by advancements in deep learning (DL) strategy, significantly overcomes the hurdles of traditional optimization techniques [17], [18]. AI and ML algorithms introduce unprecedented adaptability, learning capability, and predictive analysis, enabling sophisticated and efficient resource allocation strategies that were previously unattainable [19], [20]. Wang *et al.* demonstrated that the DL technique for ISAC-enabled predictive beamforming outperforms traditional methods by bypassing intermediate state parameter estimation [17]. The authors in [18] used advanced deep reinforcement learning (DRL) algorithms in securing ISAC systems, balancing communication efficacy with security against eavesdropping threats. DRL algorithms excel in sequential decision-making environments, making them ideal for dynamic, uncertain systems through continuous interaction with the environment [18]–[20]. Using a DRL-based deep deterministic policy gradient (DDPG) algorithm, Gong *et al.* solved the joint optimization problem of vehicular task scheduling and resource allocation in an ISAC framework [19]. In another work, Liu *et al.* explored a DRL-based proximal policy optimization (PPO) algorithm in a multi-user multiple-input single-output (MISO) scenario to maximize system capacity in an RIS-aided ISAC system [20].

A. Motivations and Contributions

The proliferation of non-terrestrial networks using 6G technology, encompassing both satellite and aerial platforms, offers a new frontier for enhancing military surveillance capabilities [2]. The potential absence of direct line-of-sight (LoS) from the ground radar system to UFOs due to several obstructions or geographical constraints necessitates innovative solutions [5]. Installing RIS in a high-rising building or deploying UAVs could be helpful in the DoA estimation process [6]. The application of RIS in DoA estimation is well investigated, but their fixed point installation limits full exploitation in such scenarios [5]. Conversely, some existing works employ UAVs for sensing in DoA estimation, yet the continuous hovering and sensing operations quickly deplete the UAVs' battery power. Therefore, UAV-mounted passive RIS emerges as a promising approach for enhancing the DoA estimation process by combining the UAV's mobility and RIS's passive beamforming capabilities while reporting to the ground radar system [21]. However, integrating these technologies into a single framework that supports robust multiple target detection (i.e., enemy UFOs) in its airspace and efficient task offloading under URLLC stringent QoS requirements raises a significant challenge for the optimization algorithms. While DRL offers a promising solution for the optimization of this joint framework, it faces training time limitation and scalability for high-dimensional computational space environments [22].

Quantum computing emerges as a potential game-changer for DRL, offering a new prospect to address these limitations [23]. Its unparalleled processing power and ability to handle complex computations simultaneously (parallelism) hold immense promise to accelerate DRL training and improve its efficiency [24]. In this study, we propose a quantum-assisted DRL framework, carefully developed to enhance the accuracy of DoA estimation of UFOs and to optimize non-terrestrial task offloading for URLLC demands within military surveillance systems using a comprehensive ISAC framework. Our solution relies on the exceptional computational power of quantum computing. By translating operational states and actions into quantum states, it substantially compresses the decision space, facilitating a more efficient learning trajectory. Furthermore, the integration of a quantum-enhanced actor-critic algorithm, employing quantum circuits for policy representation and optimization, exhibits the cutting-edge application of quantum computing to address the multifaceted optimization challenges encountered in military surveillance operations. The main contributions of this proposed work are summarized as follows:

- **Optimized ISAC System Balance:** The framework achieves an optimal balance between sensing for DoA estimation and communication for task offloading. Employing UAV-mounted passive RIS enables simultaneous DoA estimation and sensing of UFOs. The framework's novel integration of root mean squared error (RMSE), CRLB, and semidefinite programming (SDP) establishes a statistically robust foundation for DoA estimation, enhancing accuracy and reducing potential errors within a quantum-DRL environment.
- **Sophisticated Task Offloading Scheme:** By utilizing non-terrestrial networks, including satellites and aerial platforms, the proposed scheme addresses terrestrial network limitations and facilitates the execution of computationally demanding tasks within tolerable latency. The framework uniquely minimizes task offloading latency and DoA estimation errors by incorporating a novel Nash equilibrium-based reward mechanism.
- **Quantum-Enhanced Actor-Critic Method:** The novel methodology of quantum-enhanced DRL framework, along with actor-critic networks, employs quantum circuits for policy representation and optimization within a multi-agent setting. Demonstrated through extensive simulations, this approach significantly enhances DoA estimation accuracy and task offloading efficiency, meeting URLLC's rigorous demands. The proposed framework fastens the convergence speed during the training process, surpassing the perfor-

mance of existing conventional DRL-based DDPG [19] and PPO [20] algorithms. The further comparative analysis highlights substantial task offloading latency reductions over traditional DRL algorithms, emphasizing quantum computing's transformative impact on military surveillance operations.

The rest of the paper is organized as follows: Section II outlines the proposed system model and its application to ISAC within military surveillance. Section III provides insight into problem formulation for DoA estimation and task offloading. Section IV unveils our novel quantum-enhanced DRL algorithm and its implementation. Section V assesses our framework through simulations, metrics, and comparisons. Section VI concludes our contributions.

II. System Model

Fig. 1 depicts a complex military surveillance system that operates in urban, geographically challenging environments. The system comprises a set of ground vehicles, denoted as $\mathcal{V} = \{1, \ldots, V\}$, strategically deployed in a spatially distributed area. Each vehicle $v \in \mathcal{V}$ is equipped with a ground radar detection system, which is enhanced through the deployment of $\mathcal{U} = \{1, \ldots, U\}$ UAVs. The UAVs are tasked with scanning the airspace to detect $\mathcal{W} = \{1, 2, \ldots, W\}$ UFOs in its airspace. Inspired by the energy efficiency achieved by RIS for the rate demands of the end-users [25], we consider that each UAV $u \in \mathcal{U}$ is equipped with a passive 2-bit RIS consisting of $\mathcal{N} = \{1, \ldots, N\}$ elements. For the reader's convenience, Table I provides a comprehensive list of symbols. Please note that only a few symbols are redefined with proper mentioning for multiple purposes within this manuscript, but their scopes are limited to the specific contexts discussed.

The total operational timeframe, T, is dynamically partitioned into two segments: τT and $(1-\tau)T$, as shown in Fig. 2. During the τT phase, each vehicle-based ground radar system concludes the presence or absence of UFOs and estimates their DoA in the airspace. Simultaneously, the ground vehicles collect data from $\mathcal{G} = \{1, \ldots, G\}$ neighboring sensors for its next military operations. The accumulated data from all the sensors requires heavy processing, which demands dedicated computational resources. However, not all military vehicles have high-end processing units due to cost, resource, and risk feasibility constraints. To address this, we use a task-offloading technology [16] wherein ground vehicles offload tasks to the UAVs and a network of $\mathcal{L} = \{1, \ldots, L\}$ satellites during the sub-slot of $(1-\tau)T$, as depicted in Fig. 2. The UAVs assist in quick task caching, reducing the need for hardware-based task processing, while the satellites perform heavy computation in their onboard edge processing units. It is worth mentioning here that if the radar system detects the presence of a UFO in the airspace, the ground vehicle halts the task offloading process due to the potential threat of intercepting the sensitive data.

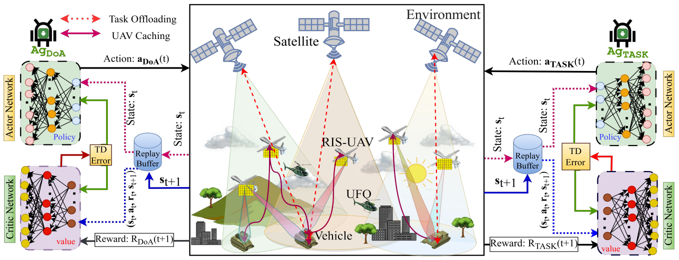


Fig. 1: DoA estimation and task offloading using quantum-DRL.

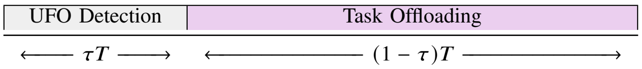


Fig. 2: Time-frame model of surveillance system framework.

A. UFO Sensing and 3D DoA Estimation

The vehicular ground-based radar system initiates the DoA estimation by transmitting a scanning signal $s(t)$ towards the UAV-mounted RIS in the present surveillance framework. This signal, represented by $s(t) = A(t)e^{j(2\pi f_c t + \psi(t))}$, where $A(t)$ is the amplitude, f_c the carrier frequency, and $\psi(t)$ the phase, propagates through the airspace and is received by UAVs equipped with 2-bit passive RIS. Each UAV's RIS, consisting of $\mathcal{N} = \{1, 2, \ldots, N\}$ elements, manipulates this signal via discrete phase adjustments corresponding to the 2-bit control. The phase-shifted signal from the n-th element of the RIS is expressed as $s_n(t) = s(t)e^{j\phi_n}$, where ϕ_n is the phase shift induced by the RIS.

The echo signal $\mathbf{s}_{\text{echo}}(t) \in \mathbb{C}^{M\times 1}$ received by the ground radar is an aggregate of signals that are reflected from multiple UFOs $w \in \mathcal{W}$ towards the UAVs. The phase-shifted signal after interacting with the UFO is given by $\mathbf{x}_w(t) = \mathbf{s}(t)e^{j\Phi_u(t)}$, where $\Phi_u(t)$ encompasses the collective phase shifts introduced by the RIS on UAV u. The incident echo signal's azimuth angle θ_w and elevation angle ϕ_w are critical for three-dimensional (3D) DoA estimation. Incorporating the mobility of the vehicle-based ground radar, UAVs, and UFOs, the Doppler effect is integrated into the signal model. The received echo signal is written by:

$$\mathbf{s}_{\text{echo}}(t) = \sum_{u=1}^{U}\sum_{w=1}^{W} \mathbf{G}_{vu}(t)\mathbf{h}_{uw}(t)\mathbf{a}_u(\theta_w, \phi_w)(t)\mathbf{\Phi}^u_{\text{RIS}}(t) \\ \times \mathbf{x}_w(t - \boldsymbol{\tau}_{uw}(t))e^{j2\pi f_{d_{uw}} t} + \boldsymbol{\eta}_v(t), \tag{1}$$

where $\mathbf{G}_{vu}(t) \in \mathbb{C}^{M\times N}$ is the channel gain matrix from the u-th UAV to the ground radar v with M antennas, $\mathbf{h}_{uw}(t) \in \mathbb{C}^{N\times 1}$ represents the channel gain vector from the w-th UFO to the RIS elements on the u-th UAV, $\mathbf{a}_u(\theta_w, \phi_w)(t) \in \mathbb{C}^{N\times 1}$ is the steering vector of the RIS incorporating the azimuth and elevation angles, and $\mathbf{\Phi}^u_{\text{RIS}}(t) \in \mathbb{C}^{N\times N}$ is the RIS phase shift matrix on UAV u. For a given measurement instance t during the operational phase τT, the phase shift matrix is represented as:

$$\mathbf{\Phi}^u_{\text{RIS}}(t) = \text{diag}\left(e^{j\varphi_1}, e^{j\varphi_2}, \ldots, e^{j\varphi_N}\right), \tag{2}$$

where diag(.) denotes a diagonal matrix with N elements, $e^{j\varphi_n}$ corresponds to the phase shift induced by the n-th element of the RIS on UAV u, and φ_n is the phase shift value for the n-th RIS element at time t. The term $\mathbf{x}_w(t)$ signifies the phase-shifted radar signal incident on UFO w, $\boldsymbol{\tau}_{uw}(t)$ is the vector of propagation delays from the RIS elements on UAV u to UFO w, $f_{d_{uw}}$ denotes the Doppler frequency shift for the link between UAV u and UFO w, and $\boldsymbol{\eta}_v(t) \in \mathbb{C}^{M\times 1}$ is the noise vector at the radar.

TABLE I: A comprehensive symbol table for the proposed work.

Symbol	Description	Symbol	Description	Symbol	Description
c	Speed of light	Z	Number of measurements	d	Inter-element spacing
f_c	Carrier frequency	A	Amplitude of the scanning signal	ψ	Phase of the scanning signal
M	Number of antennas in vehicle	κ	Rician factor	λ	Wavelength
$\mathbf{x}_v$	Position of vehicle	$\mathbf{x}_u$	Position of UAV	$\mathbf{X}_l$	Position of satellite
$\mathbf{v}_v$	Velocity vector of vehicle	$\mathbf{v}_u$	Velocity vector of UAV	$\mathbf{V}_l$	Velocity vector of satellite
$\mathbf{w}_v$	Waypoint vector of vehicle	$\mathbf{w}_u$	Waypoint vector of UAV	θ_v	Heading direction of vehicle
θ_u	Horizontal direction of UAV	θ_l	Horizontal direction of satellite	ϕ_u	Vertical inclination of UAV
θ_w	Azimuth angle from UFO to UAV	ϕ_w	Elevation angle from UFO to UAV	$\phi_{i,uw}$	Phase shift for NLoS
Δt	Time interval	β_1	Proportion for local processing	β_2	Proportion for UAV caching
β_3	Proportion for satellite offloading	T	Total time duration	f_l	CPU capability of satellite
f_v	CPU capability of vehicle	T_{Lat}	Total processing latency	$\alpha_{i,uw}$	Amplitude gain for NLoS
s	Scanning signal	$\epsilon_{\max}$	Max error norm	f_D^{vum}	Doppler shift from vehicle to UAV
PL	Path loss	PL_0	Reference path loss	X_σ	Shadow fading
$\mathbf{R}_n$	Noise covariance matrix	T_{com}^{vj}	Computation latency	T_{tra}^{vj}	Transmission latency
Γ_v^u	SINR vehicle to UAV	Γ_v^l	SINR vehicle to satellite	υ_{vl}	Channel dispersion
υ_{vu}	Channel dispersion vehicle to UAV	ς_{vu}	Codeword length vehicle to UAV	ς_{vl}	Codeword length vehicle to satellite
ξ_{vu}	QoS for URLLC vehicle to UAV	ξ_{vl}	QoS for URLLC vehicle to satellite	$\mathbf{h}_{vu}$	Channel vector vehicle to UAV
$\mathbf{h}_{vl}$	Channel vector vehicle to satellite	$\mathbf{h}_{vu}^{\text{LoS}}$	LoS channel vector vehicle to UAV	$\mathbf{h}_{vu}^{\text{NLoS}}$	NLoS channel from vehicle to UAV
$\mathbf{h}_{vl}^{\text{LoS}}$	LoS channel from vehicle to satellite	$\mathbf{h}_{vl}^{\text{NLoS}}$	NLoS channel	$\mathbf{e}_{vu}$	Channel estimation error
$\mathbf{e}_{vl}$	Channel estimation error in satellite	D_v	Task of vehicle	Ω_{vj}	Sub-task data size
$\mathcal{V}$	Set of vehicles	$\mathcal{U}$	Set of UAVs	$\mathcal{W}$	Set of UFOs
$\mathcal{N}$	Set of RIS elements	$\mathcal{B}$	Set of base stations	$\mathcal{J}$	Set of sub-tasks
$I(\theta, \phi)$	Fisher information matrix	$\mathbf{G}_{vu}$	Channel gain UAV to radar	$\mathbf{R}_{\mathbf{s}_{\text{echo}}}$	Covariance matrix of the echo signal
$\mathbf{\Phi}_{\text{RIS}}^u$	RIS phase shift matrix	$\mathbf{x}_w$	Phase-shifted radar signal	$\mathbf{a}_u$	Steering vector of RIS
φ_n	Phase shift induced by RIS element	$\boldsymbol{\tau}_{uw}$	Propagation delays	$f_{d_{uw}}$	Doppler shift
δ_θ	Adjustment factor for orbital	θ_{vum}	AoD from vehicle to UAV	$\mathbf{W}$	Beamforming matrix
P_v	Transmission power of vehicle	$\mathscr{B}$	Communication bandwidth	c_{vj}	Computational complexity
Θ_v	Heading update for vehicle	Θ_u	Desired horizontal angle for UAV	Φ_u	Desired vertical angle for UAV
$\mathcal{S}_{\mathtt{oa}}$	Whole operational state space	$\mathbf{s}_{\mathtt{oa}}$	Temporal operational state	$\mathcal{A}_{\mathtt{oa}}$	Operational action space
$\mathbf{a}_{\mathtt{oa}}$	Operational action	$\mathbf{s}_{\mathtt{DoA}}$	DoA estimation state	$\mathbf{s}_{\mathtt{TASK}}$	Task offloading state
$\mathbf{a}_{\mathtt{DoA}}$	DoA estimation action	$\mathbf{a}_{\mathtt{TASK}}$	Task offloading action	$r_{\mathtt{oa}}$	Reward
τ	Time step	$\mathcal{E}$	Quantum environment	$\mathbf{s}_{\mathtt{oa}}^{Q}$	Quantum-encoded operational state
$\mathbf{a}_{\text{opt}}$	Optimal action	$\mathbf{R}_{(\theta_w, \phi_w)}$	DoA estimation decision	$\mathbf{\Phi}_{\text{RIS}}$	RIS configuration
$\widetilde{\mathbb{R}}_{vu}$	Communication rate from v to u	$\widetilde{\mathbb{R}}_{vl}$	Communication rate from v to l	λ_{D}	Penalty coefficient for DoA
C_{D}	Cost function for DoA estimation	λ_{T}	Penalty term for task offloading	C_{T}	Cost function for task offloading
γ	Learning rate	$\mathbf{a}$	Action vector	$\mathbb{E}$	Expectation operator
$\lvert\psi\rangle$	Quantum state	$\mathcal{U}$	Unitary operation	$\mathbf{M}$	Measurement operator
Q	Quantum circuit	ϱ_{oa}	Quantum parameter	α	Learning rate
$\theta_{\mathtt{oa}}^{\pi}$	Actor network parameter	$\theta_{\mathtt{oa}}^{Q}$	Critic network parameter	$\mathcal{D}$	Replay buffer
$\delta_{\mathcal{D}}$	Temporal difference error	$V_{\theta_{\mathtt{oa}}^{V}}$	Value function	D_{KL}	Kullback-Leibler divergence
$\mathcal{L}$	Loss function	β	Balancing coefficient	π	Policy

1) Steering Vector Formulation: Considering a uniform linear array of RIS elements with inter-element spacing d. For an incident signal with wavelength λ, the steering vector $\mathbf{a}_u(\theta_w, \phi_w)(t)(\theta_w, \phi_w)$ for the u-th UAV is expressed as a function of the azimuth angle θ_w and elevation angle ϕ_w. The phase shift at the n-th RIS element, due to a signal arriving from direction (θ_w, ϕ_w), is given by $\frac{2\pi}{\lambda} nd \sin(\theta_w) \sin(\phi_w)$. Therefore, the steering vector is formulated as:

$$\mathbf{a}_u(\theta_w, \phi_w)(t) = \Big[1, e^{j\frac{2\pi}{\lambda} d \sin(\theta_w(t)) \sin(\phi_w(t))}, \ldots, \\ e^{j(N-1)\frac{2\pi}{\lambda} d \sin(\theta_w(t)) \sin(\phi_w(t))}\Big]^{\text{T}}, \tag{3}$$

where $[\cdot]^{\text{T}}$ indicates the transpose operation.

2) Imperfect Channel and Pathloss Model: Channel imperfections, such as multipath fading, shadowing, and environmental obstructions, along with imperfect channel state information (CSI), can significantly distort the transmitted signal. The channels from UAVs to the ground radar ($\mathbf{G}_{vu}(t)$) and from UFOs to UAVs ($\mathbf{h}_{uw}(t)$) are modelled using a Rician fading channel. This model includes LoS and non-LoS (NLoS) components and is expressed as:

$$\mathbf{h}_{uw}(t) = \sqrt{\frac{K}{K+1}}\mathbf{h}_{uw}^{\text{LoS}}(t) + \sqrt{\frac{1}{K+1}}\mathbf{h}_{uw}^{\text{NLoS}}(t), \tag{4}$$

where K is the Rician factor. The LoS component is modelled as:

$$\mathbf{h}_{uw}^{\text{LoS}}(t) = \frac{\lambda}{4\pi d_{uw}(t)} e^{-j\frac{2\pi}{\lambda} d_{uw}(t)}, \tag{5}$$

and the NLoS component is modelled as [26]:

$$\mathbf{h}_{uw}^{\text{NLoS}}(t) = \sum\nolimits_{i=1}^{N} \alpha_{i,uw}(t) e^{-j\phi_{i,uw}(t)}, \tag{6}$$

where $\alpha_{i,uw}$ and $\phi_{i,uw}$ are the amplitude gains and the phase shifts for each NLoS multipath component, respectively. The channel between UAVs and the ground radar, $\mathbf{G}_{vu}(t)$, follows a similar Rician model to $\mathbf{h}_{uw}(t)$ as given in (5). The path loss for both UAV-ground radar and UAV-UFO links is expressed as:

$$\text{PL}(d) = \text{PL}_0 + 10\gamma \log_{10}\left(\frac{d}{d_0}\right) + X_\sigma, \tag{7}$$

where X_σ is the shadowing component, which is a Gaussian random variable with zero mean and standard deviation σ.

For imperfect CSI, we model the actual channel $\tilde{\mathbf{h}}_{uw}(t)$ as the sum of the estimated channel and an error term ($\mathbf{e}_{uw}$):

$$\tilde{\mathbf{h}}_{uw}(t) = \mathbf{h}_{uw}(t) + \mathbf{e}_{uw}(t), \tag{8}$$

with the error norm bounded by $\epsilon_{\max}$:

$$0 < \|\mathbf{e}_{uw}(t)\| \leq \epsilon_{\max}. \tag{9}$$

3) Estimation with Temporally Distributed Measurements: In our DoA estimation process, the echo signal $\mathbf{s}_{\text{echo}}(t)$ received during the τT phase is systematically measured Z times. These measurements, spaced at regular intervals over the entire duration of the τT phase, enable a time-distributed capture of the signal dynamics. Thus the sample duration can be written as $t_z = \frac{\tau T}{Z}$ and $t_z < \tau T < T$. This measurement strategy is crucial to accurately characterizing the signal reflections from the UFOs, considering their potential movement and environmental variations. The aggregated echo signal (using Z measurements), during τT phase, is represented as follows:

$$\mathbf{s}_{\text{echo}}(t) = \sum_{z=1}^{Z} \Bigg(\sum_{u=1}^{U} \sum_{w=1}^{W} \mathbf{G}_{vu}(t_z) \mathbf{h}_{uw}(t_z) \mathbf{a}_u(\theta_w, \phi_w)(t_z) \\ \mathbf{\Phi}_{\text{RIS}}^{u}(t_z) \mathbf{x}_w(t_z - \tau_{uw}(t_z)) e^{j2\pi f_{d_{uw}} t_z} + \boldsymbol{\eta}_v(t_z) \Bigg). \quad (10)$$

B. Vehicular Communication Model with UAVs and Satellites

In our urban-focused vehicular network system, an MISO uplink communication model is implemented. Each vehicle $v \in \mathcal{V}$ is equipped with $\mathcal{M} = \{1, 2, \ldots, M\}$ transmitting antennas and is complemented by a fleet of UAVs and a satellite, each having a single receiving antenna.

1) Mobility Model: Our proposed model extends the classical random waypoint model [27] incorporating advanced dynamics to simulate the movement of three types of entities: ground vehicles, UAVs, and satellites. The entities are initially distributed randomly within a 3D space bounded by $\mathcal{X} \in (-X_{\min}, +X_{\max})$, $\mathcal{Y} \in (-Y_{\min}, +Y_{\max})$, and $\mathcal{Z} \in (-Z_{\min}, +Z_{\max})$. Their positions are considered static within the interval $[\Delta_{t,t-1}]$.

The mobility model for ground vehicles integrates urban mobility factors, considering the complex movement patterns in city environments. The position of vehicle v at time slot t, denoted as $\mathbf{x}_v(t) = [x_v(t), y_v(t)]^{\mathrm{T}}$, is determined by its velocity vector $\mathbf{v}_v(t) = [v_x, v_y]$ and heading direction $\theta_v(t)$. The velocity and directional updates for ground vehicles at each time slot are given by:

$$\mathbf{v}_v(t) = \lambda \mathbf{v}_v(t-1) + (1-\lambda)\mathbf{w}_v, \quad (11)$$

$$\theta_v(t) = \mu \theta_v(t-1) + (1-\mu)\Theta_v, \quad (12)$$

where λ and μ are momentum and adaptation factors in range of $(0.1, 0.2)$. $\mathbf{w}_v$ is the waypoint vector, and Θ_v denotes the heading direction. To translate the velocity and direction into positional changes, we incorporate trigonometric functions as:

$$x_v(t) = x_v(t-1) + \Delta t \mathbf{v}_v(t) \cos(\theta_v(t)), \quad (13)$$

$$y_v(t) = y_v(t-1) + \Delta t \mathbf{v}_v(t) \sin(\theta_v(t)), \quad (14)$$

where Δt is the time interval between updates.

The mobility model for UAVs accounts for three-dimensional space, reflecting their operational dynamics comprehensively. The model includes horizontal movements, altitude changes, and responses to atmospheric fluctuations. The UAV's position at time slot t, denoted as $\mathbf{x}_u(t) = [x_u(t), y_u(t), z_u(t)]^{\mathrm{T}}$, is influenced by its velocity vector $\mathbf{v}_u(t)$ and heading angles $(\theta_u(t), \phi_u(t))$, where $\theta_u(t)$ represents the horizontal direction and $\phi_u(t)$ indicates the vertical inclination. The velocity and direction of the UAV at each time slot are updated as follows:

$$\mathbf{v}_u(t) = \alpha_u \mathbf{v}_u(t-1) + (1-\alpha_u)\mathbf{w}_u + \boldsymbol{\vartheta}_v, \quad (15)$$

$$\theta_u(t) = \beta_u \theta_u(t-1) + (1-\beta_u)\Theta_u + \vartheta_\theta, \quad (16)$$

$$\phi_u(t) = \gamma_u \phi_u(t-1) + (1-\gamma_u)\Phi_u + \vartheta_\phi, \quad (17)$$

where $\alpha_u, \beta_u, \gamma_u$ are persistence factors, $\mathbf{w}_u$ is the waypoint vector for velocity, Θ_u and Φ_u are the desired horizontal and vertical angles, and $\boldsymbol{\vartheta}_v, \vartheta_\theta, \vartheta_\phi$ are random perturbation terms. To translate these velocities and angles into the UAV's positional changes, we use trigonometric functions:

$$x_u(t) = x_u(t-1) + \Delta t \mathbf{v}_u(t) \cos(\theta_u(t)) \cos(\phi_u(t)), \quad (18)$$

$$y_u(t) = y_u(t-1) + \Delta t \mathbf{v}_u(t) \sin(\theta_u(t)) \cos(\phi_u(t)), \quad (19)$$

$$z_u(t) = z_u(t-1) + \Delta t \mathbf{v}_u(t) \sin(\phi_u(t)), \quad (20)$$

where Δt is the time interval between updates. These equations allow for a precise and realistic representation of the UAV's trajectory in 3D space, accommodating complex flight patterns and environmental influences.

The satellite's position at time slot t, represented as $\mathbf{X}_l(t) = [x_l(t), y_l(t), z_l(t)]$, is characterized by a constant altitude (z-coordinate) due to its fixed lower earth orbit. The updates for the horizontal movement of the satellite can be represented as:

$$\mathbf{V}_l(t) = \mathbf{V}_l(t-1), \quad (21)$$

$$\theta_l(t) = \theta_l(t-1) + \delta_\theta \Delta t \theta_l, \quad (22)$$

where δ_θ is an adjustment factor for orbital movements. Similar to the ground vehicle positioning model, the new satellite coordinates $\mathbf{x}_l(t) = [x_l(t), y_l(t), z_l(t)]^{\mathrm{T}}$ at each time slot are calculated while keeping the z-coordinate constant.

2) Channel Model: The Rician channel model is employed to describe the LoS and NLoS propagation components, taking into account the angle-of-departure (AoD) and Doppler shift phenomena. The channel vector for the vehicle-to-UAV link from vehicle $v \in \mathcal{V}$ to UAV $u \in \mathcal{U}$, denoted as $\mathbf{h}_{vu}^{LoS}(t)$, is modeled for each transmitting antenna as:

$$\mathbf{h}_{vu}^{LoS}(t) = \Bigg[\frac{\lambda}{4\pi d_{vu}} e^{-j2\pi \frac{d_{vu}}{\lambda}} e^{-j2\pi f_c \frac{v_{vu}}{c} \cos(\theta_{vu1}) t}, \\ \ldots, \frac{\lambda}{4\pi d_{vu}} e^{-j2\pi \frac{d_{vu}}{\lambda}} e^{-j2\pi f_c \frac{v_{vu}}{c} \cos(\theta_{vuM}) t} \Bigg]^{\mathrm{T}}, \quad (23)$$

where λ is the wavelength, d_{vu} the distance, v_{vu} the relative velocity, c the speed of light, and θ_{vum} the AoD for the m-th antenna. The Doppler shift for each antenna is expressed as:

$$f_D^{vum} = \frac{v_{vu} f_c \cos(\theta_{vum})}{c}. \quad (24)$$

An analogous vector formulation applies from v-th vehicle to l-th satellite link ($v \rightarrow l$), $\mathbf{h}_{vl}^{LoS}(t)$, incorporating the parameters d_{vl}, v_{vl}, and θ_{vlm} for each transmitting antenna.

In our MISO communication framework, the aggregate channel gain vector from the v-th vehicle to u-th UAV link is derived using the Rician model and expressed as:

$$\mathbf{h}_{vu}(t) = \sqrt{PL_{vu}} \left(\sqrt{\frac{\kappa}{\kappa+1}} \mathbf{h}_{vu}^{LoS}(t) + \sqrt{\frac{1}{\kappa+1}} \mathbf{h}_{vu}^{NLoS}(t) \right), \quad (25)$$

where $\mathbf{h}_{vu}^{LoS}(t)$ and $\mathbf{h}_{vu}^{NLoS}(t)$ represent the LoS and NLoS components, respectively, for the vehicle-to-UAV link. The symbol PL_{vu} represents the path loss model defined in (27). The channel vector $\mathbf{h}_{vu}^{NLoS}(t)$ for the vehicle-to-UAV link is modeled as a complex Gaussian random vector, being mathematically represented as:

$$\mathbf{h}_{vu}^{NLoS}(t) = \left[h_{vu1}^{nlos}(t), h_{vu2}^{nlos}(t), \ldots, h_{vuM}^{nlos}(t)\right]^{\mathrm{T}}, \tag{26}$$

where each component $h_{vum}^{nlos}(t)$, $m = 1, 2, \ldots, M$, is modeled as a complex Gaussian random variable, $h_{vum}^{nlos}(t) \sim \mathcal{CN}(0, \sigma_{nlos}^2)$, where $\mathcal{CN}$ represents the complex normal distribution with zero mean and variance σ_{nlos}^2, encapsulating the NLoS propagation characteristics. The expression for the vehicle-to-satellite link $\mathbf{h}_{vl}(t)$ follows a similar structure, accounting for the respective parameters of the satellite link. The path loss for the vehicle-to-UAV link is modeled as follows:

$$PL_{vu} = \left(\frac{4\pi d_{vu}}{\lambda}\right)^2. \tag{27}$$

The path loss model for the vehicle-to-satellite link PL_{vl} adheres to a similar formulation, with d_{vl} indicating the distance to the satellite.

3) Imperfect CSI and SINR Calculation: Considering imperfect CSI, the estimated channel vectors $\hat{\mathbf{h}}_{vu}$ and $\hat{\mathbf{h}}_{vl}$ are:

$$\hat{\mathbf{h}}_{vu} = \mathbf{h}_{vu}(t) + \mathbf{e}_{vu}, \quad \hat{\mathbf{h}}_{vl} = \mathbf{h}_{vl}(t) + \mathbf{e}_{vl}, \tag{28}$$

with $\mathbf{e}_{vu}$ and $\mathbf{e}_{vl}$ as the estimation error vectors, each element of which is a complex Gaussian random variable. The signal-to-interference plus noise ratio (SINR) for vehicle v communicating with UAV u and satellite l under non-orthogonal multiple access (NOMA) is derived as:

$$\Gamma_v^u(t) = \frac{P_v \|\hat{\mathbf{h}}_{vu}(t)\|^2}{\sum_{k=v+1}^{V} P_k \|\hat{\mathbf{h}}_{ku}(t)\|^2 + \sigma_{n_u}^2}, \tag{29}$$

$$\Gamma_v^l(t) = \frac{P_v \|\hat{\mathbf{h}}_{vl}(t)\|^2}{\sum_{k=v+1}^{V} P_k \|\hat{\mathbf{h}}_{kl}(t)\|^2 + \sigma_{n_l}^2}, \tag{30}$$

where $\|\hat{\mathbf{h}}_{vu}\|$ and $\|\hat{\mathbf{h}}_{vl}\|$ denote the norms of the estimated channel vectors, capturing the combined effect of all transmitting antennas of the vehicle. In this NOMA setup, successive interference cancellation is employed for the decoding process.

4) Data Rate Calculation under URLLC Requirements: The data rate for a vehicle v communicating with a UAV u or satellite s at time t, considering URLLC requirements and imperfect CSI, is expressed as follows:
For the vehicle-to-UAV link:

$$\mathrm{R}_{vu}(t) = \mathscr{B}\left(\log_2\left(1 + \Gamma_v^u(t)\right) - \sqrt{\frac{\upsilon_{vu}(\Gamma_v^u)(t)}{\varsigma_{vu}(t)}}\,\xi_{vu}(t)\right), \tag{31}$$

where $\mathscr{B}$ represents the bandwidth of the communication channel. The term $\varsigma_{vu}(t)$ signifies the codeword/block length, and $\xi_{vu}(t)$ is a QoS parameter adjusting the data rate to meet the URLLC reliability requirement, defined as:

$$\xi_{vu}(t) = \frac{\mathrm{Q}^{-1}(\nu_{vu})}{\log_e 2}, \tag{32}$$

where $\mathrm{Q}^{-1}(\nu_{vu})$ is the inverse of the Q-function with parameter ν_{vu} (i.e., represents a packet error rate), used for calculating the decoding error probability. The channel dispersion $\upsilon_{vu}(\Gamma_v^u)(t)$ is given by:

$$\upsilon_{vu}(\Gamma_v^u)(t) = 1 - \left(1 + \Gamma_v^u(t)\right)^{-2}. \tag{33}$$

Similarly, for the vehicle-to-satellite link:

$$\mathrm{R}_{vl}(t) = \mathscr{B}\left(\log_2\left(1 + \Gamma_v^l(t)\right) - \sqrt{\frac{\upsilon_{vl}(\Gamma_v^l)(t)}{\varsigma_{vl}(t)}}\,\xi_{vl}(t)\right). \tag{34}$$

C. Partial Task Offloading with UAV Caching

In parallel, the ground vehicles are tasked with receiving critical information from a network of military sensors deployed in their vicinity. Processing this cumulative sensory data, coupled with the radar information, requires substantial computational resources. Given practical constraints such as cost, maintenance, and sensitivity, installing high-end processing units in each ground military vehicle is unfeasible. We propose a hybrid task offloading mechanism involving UAVs and satellites to tackle this. Specifically, ground vehicles offload their latency-sensitive URLLC-enabled computational tasks to the UAVs for task caching and a satellite for processing. The task of the i-th vehicle, denoted as D_v, is decomposed into a series of smaller sub-tasks as follows:

$$D_v = \bigcup_{j=1}^{J} \Omega_{vj}, \tag{35}$$

where Ω_{vj} signifies the data size of j-th sub-task of the v-th vehicle, and $\mathcal{J} = \{1, 2, \ldots, J\}$ is the count of sub-tasks partitioning the original task D_v.

Each sub-task Ω_{vj} can either be processed locally within the vehicle, offloaded for computation to a proximate node, or cached in the UAVs. The offloading decision is modeled as a binary variable x_{vj}^l for sub-task Ω_{vj}, where $x_{vj}^l = 1$ indicates offloading to the satellite, and $x_{vj}^l = 0$ denotes local processing. The caching decision is modeled as a binary variable x_{vj}^u for sub-task Ω_{vj}, where $x_{vj}^u = 1$ indicates caching to the u-th UAV and $x_{vj}^u = 0$ signifies no caching service. Here to note that $(x_{vj}^l + x_{vj}^u) \leq 1$, where $l \in \mathcal{L}$ and $u \in \mathcal{U}$.

In our task offloading strategy, we ensure an equitable distribution of computational tasks across local processing units, UAVs, and satellites by imposing constraints on the offloading proportions. Specifically, we have $\sum_{j=1}^{J} x_{vj}^l = \beta_1 J$, $\sum_{j=1}^{J} x_{vj}^u = \beta_2 J$, and $\sum_{j=1}^{J}(1 - x_{vj}^l - x_{vj}^u) = \beta_3 J$, where β_1, β_2, and β_3 denote the proportions of tasks allocated for local processing, UAV caching, and satellite offloading, respectively. To ensure these proportions are both feasible and optimized, we establish the following boundary conditions $0 < \beta_{\min} \leq \beta_i < \beta_{\max} \leq 1$, $i = 1, 2, 3$, under the constraint $\beta_1 + \beta_2 + \beta_3 = 1$. This approach guarantees a balanced task allocation, optimizing resource utilization within our system's operational parameters.

The computation latency T_{com}^{vj} for a sub-task $\forall \Omega_{vj} \in D_v$ is conditionally determined based on the caching decision:

$$T_{com}^{vj}(t) = \begin{cases} 0 & \text{if } x_{vj}^u = 1, \\ x_{vj}^l \frac{c_{vj}(t)}{f_l(t)} + (1 - x_{vj}^l) \frac{c_{vj}(t)}{f_v(t)} & \text{otherwise.} \end{cases} \tag{36}$$

Here, c_{vj} denotes the computational complexity of sub-task Ω_{vj}. The symbols f_l and f_v represent the available CPU processing capability (i.e., $f_{\min}(t) < f_i(t) < f_{\max}(t)$) of the satellite and vehicle, respectively. The communication transmission latency T_{tra}^{vj} for offloading Ω_{vj} is dependent on the data rate $\mathrm{R}_{vu}(t)$ or

$\mathrm{R}_{vl}(t)$, and is expressed as:

$$T^{vj}_{tra}(t) = \left(x^{l}_{vj}\frac{\Omega_{vj}}{\mathrm{R}_{vl}(t)} + x^{u}_{vj}\frac{\Omega_{vj}}{\mathrm{R}_{vu}(t)}\right), \qquad \exists\Omega_{vj} \in D_v. \tag{37}$$

The total latency T_{Lat} for processing D_v is a function of the computation and communication latencies, expressed as:

$$T_{\mathrm{Lat}}(t) = \sum\nolimits_{j=1}^{J} T^{vj}_{\mathrm{Lat}}(t) = \left(T^{vj}_{tra}(t) + T^{vj}_{com}(t)\right), \forall j \in D_v. \tag{38}$$

We focus on optimizing the task offloading and computational resource distribution, excluding detailed consideration of response times due to their minimal impact on the system's overall performance, given the disparity in data size between response payloads and processing tasks.

III. Problem Formulation

In this integrated sensing and communication system, our primary objective is to optimize the operational efficiency of the military surveillance network while ensuring robust and reliable communication between ground vehicles, UAVs, and satellites. The problem encompasses several key components: accurate 3D DoA estimation for detecting UFOs, efficient management of communication resources, and effective URLLC-enabled task-offloading strategies to balance the computational load.

A. DoA Estimation and Obtaining Sensing Decision

The primary sensing objective is to enhance the 3D DoA estimation for UFO detection using the UAVs' RIS-enhanced radar systems. Given the received echo signal $\mathbf{s}_{\mathrm{echo}}(t_z)$ for each measurement instance t_z within the τT phase, we can define Z measurements using (10) as $\mathbf{s}_{\mathrm{echo}}(t) = \mathbf{a}_u(\theta_w, \phi_w)(t)\mathbf{s}(t) + \boldsymbol{\eta}_v(t)$, where

$$\mathbf{s}(t) = \sum\nolimits_{z=1}^{Z}\left(\sum\nolimits_{u=1}^{U}\sum\nolimits_{w=1}^{W}\mathbf{G}_{vu}(t_z)\mathbf{h}_{uw}(t_z)\mathbf{\Phi}^{u}_{\mathrm{RIS}}(t_z)\right.$$
$$\left.\mathbf{x}_w(t_z - \boldsymbol{\tau}_{uw}(t_z))e^{j2\pi f_{d_{uw}} t_z}\right). \tag{39}$$

To facilitate DoA estimation, the covariance matrix of the echo signal, $\mathbf{R}_{\mathbf{s}_{\mathrm{echo}}}$, is calculated as:

$$\begin{aligned}\mathbf{R}_{\mathbf{s}_{\mathrm{echo}}}(t) &= \mathbb{E}[\mathbf{s}_{\mathrm{echo}}(t)\mathbf{s}^{\mathrm{H}}_{\mathrm{echo}}(t)]\\ &= \mathbf{a}_u(\theta_w, \phi_w)(t)\mathbf{R}_s(t)\mathbf{a}^{\mathrm{H}}_u(\theta_w, \phi_w)(t) + \mathbf{R}_n(t),\end{aligned} \tag{40}$$

where $\mathbf{R}_s$ is the covariance matrix of the signal and $\mathbf{R}_n$ is the noise covariance matrix. To refine the accuracy of DoA estimation, we deploy three key methodologies: RMSE, CRLB [6], and SDP. Each method offers unique insights and optimization capabilities for our DoA estimation scenario.

1) Formulation of DoA Estimation: The RMSE is our initial step in quantifying the accuracy of DoA estimation. It directly measures the average deviation between estimated and true DoA values. Mathematically, the RMSE is defined as:

$$\mathrm{RMSE} = \sqrt{\mathbb{E}\left[(\theta_{\mathrm{true}} - \hat{\theta})^2 + (\phi_{\mathrm{true}} - \hat{\phi})^2\right]}, \tag{41}$$

where θ_{true} and ϕ_{true} are the true azimuth and elevation angles of the UFOs, while $\hat{\theta}$ and $\hat{\phi}$ are their estimated counterparts. This RMSE metric serves as a practical gauge for the performance of our estimation process.

Building upon the insights from RMSE, the CRLB provides the lower bound on the variance of any unbiased estimator, which, in our case, relates to DoA parameters θ and ϕ. The probability density function of $\mathbf{R}_{\mathbf{s}_{\mathrm{echo}}}$ is expressed as

$$p(\mathbf{s}_{\mathrm{echo}}(t)) = \frac{1}{(2\pi)^{\frac{Z}{2}}\sqrt{\det(\mathbf{R}_{\mathbf{s}_{\mathrm{echo}}})}}$$
$$\exp\left(-\frac{1}{2}(\mathbf{s}_{\mathrm{echo}}(t) - \mu)^{\mathrm{H}}\mathbf{R}^{-1}_{\mathbf{s}_{\mathrm{echo}}}(\mathbf{s}_{\mathrm{echo}}(t) - \mu)\right), \tag{42}$$

where $\mu = \mathbb{E}[\mathbf{s}_{\mathrm{echo}}(t)]$. The Fisher information matrix is derived as follows [28]:

$$I(\theta, \phi) = -\mathbb{E}\begin{bmatrix}\frac{\partial^2 \ln p}{\partial\theta^2} & \frac{\partial^2 \ln p}{\partial\theta\partial\phi}\\ \frac{\partial^2 \ln p}{\partial\phi\partial\theta} & \frac{\partial^2 \ln p}{\partial\phi^2}\end{bmatrix}. \tag{43}$$

The CRLB is then given by:

$$\mathrm{CRLB}(\theta, \phi) = \mathrm{diag}(I^{-1}(\theta, \phi)). \tag{44}$$

The CRLB acts as a benchmark to assess the effectiveness of our estimation methods against the theoretical best possible performance.

Finally, we employ SDP to optimize our DoA estimation process, striving to achieve performance as close to the CRLB as possible. SDP excels in handling complex-valued matrices, making it highly suitable for DoA estimation. We formulate the optimization problem as follows:

$$\begin{aligned}&\min_{\mathbf{R}_{(\theta,\phi)}}\ \mathrm{trace}(\mathbf{R}_{(\theta,\phi)}\mathbf{s}_{\mathrm{echo}}(t)\mathbf{s}^{\mathrm{H}}_{\mathrm{echo}}(t))\\ &\text{s.t: } \mathbf{R}_{(\theta,\phi)} \succeq 0,\\ &\quad\ \mathrm{rank}(\mathbf{R}_{(\theta,\phi)}) = 1, \qquad \mathrm{trace}(\mathbf{R}_{(\theta,\phi)}) = 1,\end{aligned} \tag{45}$$

where $\mathbf{R}_{(\theta,\phi)}$ is the covariance matrix associated with the estimated DoA parameters. This matrix is optimized to minimize the difference between the projected and actual covariance matrices of the echo signal.

B. Detailed derivation process for $\mathbf{R}_{(\theta,\phi)}$

The covariance matrix of the echo signal, $\mathbf{R}_{\mathbf{s}_{\mathrm{echo}}}$, is defined as:

$$\mathbf{R}_{\mathbf{s}_{\mathrm{echo}}}(t) = \mathbb{E}[\mathbf{s}_{\mathrm{echo}}(t)\mathbf{s}^{\mathrm{H}}_{\mathrm{echo}}(t)]. \tag{46}$$

Substituting $\mathbf{s}_{\mathrm{echo}}(t)$ into the above equation, we get:

$$\begin{aligned}\mathbf{R}_{\mathbf{s}_{\mathrm{echo}}}(t) &= \mathbb{E}\left[(\mathbf{a}_u(\theta_w, \phi_w)(t)\mathbf{s}(t) + \boldsymbol{\eta}_v(t))\right.\\ &\qquad\left.(\mathbf{a}_u(\theta_w, \phi_w)(t)\mathbf{s}(t) + \boldsymbol{\eta}_v(t))^{\mathrm{H}}\right]\\ &= \mathbb{E}\left[\mathbf{a}_u(\theta_w, \phi_w)(t)\mathbf{s}(t)\mathbf{s}^{\mathrm{H}}(t)\mathbf{a}^{\mathrm{H}}_u(\theta_w, \phi_w)(t)\right] + \mathbb{E}\left[\boldsymbol{\eta}_v(t)\boldsymbol{\eta}^{\mathrm{H}}_v(t)\right]\\ &= \mathbf{a}_u(\theta_w, \phi_w)(t)\mathbb{E}\left[\mathbf{s}(t)\mathbf{s}^{\mathrm{H}}(t)\right]\mathbf{a}^{\mathrm{H}}_u(\theta_w, \phi_w)(t) + \mathbf{R}_n(t),\end{aligned} \tag{47}$$

where $\mathbf{R}_n(t) = \mathbb{E}\left[\boldsymbol{\eta}_v(t)\boldsymbol{\eta}^{\mathrm{H}}_v(t)\right]$ is the noise covariance matrix. Let $\mathbf{R}_s(t) = \mathbb{E}\left[\mathbf{s}(t)\mathbf{s}^{\mathrm{H}}(t)\right]$ be the covariance matrix of the signal $\mathbf{s}(t)$. Thus, the covariance matrix of the echo signal is expressed as (40). To get further insights into the derivation, let's consider the structure of the covariance matrix $\mathbf{R}_s(t)$. The signal $\mathbf{s}(t)$ is a superposition of multiple components, each subject to different time delays $\boldsymbol{\tau}_{uw}(t_z)$ and Doppler shifts $f_{d_{uw}}$. The covariance matrix $\mathbf{R}_s(t)$ captures these effects:

$$\mathbf{R}_s(t) = \mathbb{E}\left[\left(\sum_{z=1}^{Z}\sum_{u=1}^{U}\sum_{w=1}^{W}\mathbf{G}_{vu}(t_z)\mathbf{h}_{uw}(t_z)\mathbf{\Phi}^{u}_{\mathrm{RIS}}(t_z)\mathbf{x}_w(t_z - \boldsymbol{\tau}_{uw}(t_z))\right.\right.$$
$$\left.e^{j2\pi f_{d_{uw}} t_z}\right)\left(\sum_{z'=1}^{Z}\sum_{u'=1}^{U}\sum_{w'=1}^{W}\mathbf{G}_{vu}(t_{z'})\mathbf{h}_{uw}(t_{z'})\mathbf{\Phi}^{u}_{\mathrm{RIS}}(t_{z'})\right.$$

$$\mathbf{x}_w(t_{z'} - \boldsymbol{\tau}_{uw}(t_{z'}))e^{j2\pi f_{d_{uw}} t_{z'}}\Bigg)^{\mathrm{H}}\Bigg], \tag{48}$$

$$= \sum_{z=1}^{Z}\sum_{z'=1}^{Z}\sum_{u=1}^{U}\sum_{u'=1}^{U}\sum_{w=1}^{W}\sum_{w'=1}^{W} \mathbf{G}_{vu}(t_z)\mathbf{h}_{uw}(t_z)\boldsymbol{\Phi}^{u}_{\mathrm{RIS}}(t_z)$$
$$\mathbf{R}_{x_w}(t_z - \boldsymbol{\tau}_{uw}(t_z), t_{z'} - \boldsymbol{\tau}_{uw}(t_{z'}))\boldsymbol{\Phi}^{u'}_{\mathrm{RIS}}(t_{z'})^{\mathrm{H}}\mathbf{h}_{u'w'}(t_{z'})^{\mathrm{H}}$$
$$\mathbf{G}_{v'u'}(t_{z'})^{\mathrm{H}}e^{j2\pi f_{d_{uw}} t_z}e^{-j2\pi f_{d_{u'w'}} t_{z'}}. \tag{49}$$

The matrix $\mathbf{R}_{x_w}(t_z - \boldsymbol{\tau}_{uw}(t_z), t_{z'} - \boldsymbol{\tau}_{uw}(t_{z'}))$ represents the cross-correlation of the transmitted signals at different delays and Doppler shifts. This complex structure accounts for the varying propagation paths and their impacts on the received signal. Finally, incorporating these details, the covariance matrix of the echo signal becomes:

$$\mathbf{R}_{\mathbf{s}_{\mathrm{echo}}}(t) = \mathbf{a}_u(\theta_w, \phi_w)(t)\Bigg(\sum_{z=1}^{Z}\sum_{z'=1}^{Z}\sum_{u=1}^{U}\sum_{u'=1}^{U}\sum_{w=1}^{W}\sum_{w'=1}^{W} \mathbf{G}_{vu}(t_z)\mathbf{h}_{uw}(t_z)$$
$$\boldsymbol{\Phi}^{u}_{\mathrm{RIS}}(t_z)\mathbf{R}_{x_w}(t_z - \boldsymbol{\tau}_{uw}(t_z), t_{z'} - \boldsymbol{\tau}_{uw}(t_{z'}))\boldsymbol{\Phi}^{u'}_{\mathrm{RIS}}(t_{z'})^{\mathrm{H}}$$
$$\mathbf{h}_{u'w'}(t_{z'})^{\mathrm{H}}\mathbf{G}_{v'u'}(t_{z'})^{\mathrm{H}}e^{j2\pi f_{d_{uw}} t_z}e^{-j2\pi f_{d_{u'w'}} t_{z'}}\Bigg)$$
$$\mathbf{a}^{\mathrm{H}}_u(\theta_w, \phi_w)(t) + \mathbf{R}_n(t). \tag{50}$$

This derivation shows how the covariance matrix $\mathbf{R}_{(\theta,\phi)}(t) = \mathbf{R}_{\mathbf{s}_{\mathrm{echo}}}(t)$ is obtained from the received echo signal model. The primary goal in our DoA estimation framework is to minimize the difference between the projected covariance matrix, $\mathbf{R}_{(\theta_w,\phi_w)}$, and the actual covariance matrix, $\mathbf{R}_{\mathbf{s}_{\mathrm{echo}}}$. This approach aims to reduce the RMSE in our estimations. The optimization problem is formulated as follows:

$$\mathrm{RMSE} \propto \min_{\mathbf{R}_{(\theta_w,\phi_w)}} \mathrm{trace}(\mathbf{R}_{\mathbf{s}_{\mathrm{echo}}} - \mathbf{R}_{(\theta_w,\phi_w)}\mathbf{s}_{\mathrm{echo}}(t)\mathbf{s}^{\mathrm{H}}_{\mathrm{echo}}(t)). \tag{51}$$

1) Obtaining UFO Sensing Decision: UFO detection is approached as a binary hypothesis testing problem in our radar-based surveillance system. We assess the presence (H_1) or absence (H_0) of UFOs based on the received echo signal $\mathbf{s}_{\mathrm{echo}}(t)$. The hypotheses are defined using the indicator function Ψ, where $\Psi = 1$ indicates the UFO presence and $\Psi = 0$ indicates absence:

$$H_0(\Psi = 0): \quad \mathbf{s}_{\mathrm{echo}}(t) = \boldsymbol{\eta}_v(t), \tag{52}$$

$$H_1(\Psi = 1): \quad \mathbf{s}_{\mathrm{echo}}(t) = \mathbf{a}_u(\theta_w, \phi_w)(t)s(t) + \boldsymbol{\eta}_v(t). \tag{53}$$

To facilitate the detection decision, a threshold λ is established based on the desired probability of false alarm (P_{fa}). The test statistic for energy detection, denoted by Y, is given by:

$$Y = \frac{1}{Z}\sum\nolimits_{z=1}^{Z} |\mathbf{s}_{\mathrm{echo}}(z)|^2, \tag{54}$$

where Z represents the number of temporally distributed measurements. The probability of detection (P_d) is the likelihood of correctly detecting a UFO when it is present ($\Psi = 1$). It is calculated as [29]:

$$P_d = Q\left(\left(\frac{\lambda}{P_n} - \gamma - 1\right)\frac{\sqrt{Z}}{\gamma + 1}\right), \tag{55}$$

where γ denotes the signal-to-noise ratio (SNR), P_n is the noise power which is calculated as $P_n = \mathbb{E}[\nu(t)\nu^{\mathrm{H}}(t)]$, and $Q(.)$ is the Q-function. Similarly, the probability of a false alarm (P_{fa}) is the likelihood of incorrectly detecting a UFO when it is absent ($\Psi = 0$). This probability is determined as [29]:

$$P_{fa} = Q\left(\left(\frac{\lambda}{P_n} - 1\right)\sqrt{Z}\right). \tag{56}$$

C. Communication for Task Offloading

In our military surveillance system, the communication strategy is optimized for data throughput and latency minimization across vehicle-to-UAV and vehicle-to-satellite links within the remaining sub-slots $(T - \tau T)$, emphasizing the URLLC requirements. Task offloading decisions are dynamically adjusted based on the UFO detection, with operational logic that suspends offloading during UFO presence ($\mathcal{P}(H_1)$) to prioritize surveillance, except under the false detection scenarios characterized by $(1 - P_d)$. Conversely, normal offloading resumes in the absence of UFO detection ($\mathcal{P}(H_0)$), provided no false alarm occurs, as indicated by $(1 - P_{fa})$. Given that each vehicle's task is divided into J sub-tasks, the effective throughput for the vehicle-to-UAV link, denoted as $\mathbb{R}_{vu}$ in (31), and the vehicle-to-satellite link, denoted as $\mathbb{R}_{vl}$ in (34), are recalculated to accommodate this division. The adjusted throughputs are expressed as:

$$\widetilde{\mathbb{R}}_{vu}(t) = \left(\frac{T - \tau T}{JT}\right)\mathcal{P}(H_0)(1 - P_{fa})\mathbb{R}_{vu}(t), \tag{57}$$

$$\widetilde{\mathbb{R}}_{vl}(t) = \left(\frac{T - \tau T}{JT}\right)\mathcal{P}(H_0)(1 - P_{fa})\mathbb{R}_{vl}(t). \tag{58}$$

D. Integrated Optimization Problem

The underlying problem is integrating the sensing, communication, and computational aspects into a unified optimization framework. This involves jointly optimizing the DoA estimation process, UFO sensing performance, communication link parameters, and task offloading decisions to achieve the best overall system performance.

In the context of DoA estimation, the SNR at the receiver significantly influences the accuracy of DoA estimation, directly affecting sensing performance. The power of the received signal from (10), P_s, is defined as $P_s = \mathbb{E}[\mathbf{s}_{\mathrm{echo}}(t)\mathbf{s}^{\mathrm{H}}_{\mathrm{echo}}(t)] - \mathbb{E}[\boldsymbol{\eta}_v(t)\boldsymbol{\eta}^{\mathrm{H}}_v(t)]$. The noise power, P_n, is given by $P_n = \mathbb{E}[\boldsymbol{\eta}_v(t)\boldsymbol{\eta}^{\mathrm{H}}_v(t)]$, with $\boldsymbol{\eta}_v(t)$ denoting the noise vector at the radar receiver. Finally, the SNR, a critical metric for assessing the received signal quality relative to background noise, is defined as $\mathrm{SNR} = \frac{P_s}{P_n}$. The τT sub-slot in Fig. 2 aims to effectively ensure DoA optimization while adhering to stringent detection performance standards. The optimization task, focused on refining the accuracy of DoA estimation, is formulated as:

$$f_{\mathrm{DoA}} = \min_{\mathbf{R}_{(\theta_w,\phi_w)}, Z, \boldsymbol{\Phi}_{\mathrm{RIS}}} \left\|\mathbf{R}_{\mathbf{s}_{\mathrm{echo}}} - \mathbf{R}_{(\theta_w,\phi_w)}\mathbf{s}_{\mathrm{echo}}(t)\mathbf{s}^{\mathrm{H}}_{\mathrm{echo}}(t)\right\|^2_{\mathrm{F}},$$

$$\text{subject to} \tag{59}$$

$$\text{C1: } \mathrm{SNR} \geq \mathrm{SNR}_{\min},$$
$$\text{C2: } \varphi_n \in \{0, \pi/2, \pi, 3\pi/2\}, \quad \forall n \in \mathcal{N},$$
$$\text{C3: } \theta \in [0, 2\pi], \quad \phi \in [0, \pi],$$
$$\text{C4: } P_d \geq P^{th}_d,$$
$$\text{C5: } P_{fa} \leq P^{th}_{fa},$$
$$\text{C6: } 0 < Z \leq Z_{\max}.$$

where the constraint C1 guarantees a minimum $\mathrm{SNR}_{\min}$ at the receiver, ensuring the received signal's integrity for dependable DoA estimation. Constraint C2 precisely controls the phase shifts introduced by the RIS. Constraint C3 confines the optimization search within realistic azimuth (θ) and elevation (ϕ) angle bounds, ensuring that the DoA estimation adheres to feasible operational domains. Furthermore, constraint C4 ensures that the probability of detecting a UFO (P_d) surpasses a predetermined threshold P_d^{th}. Conversely, C5 ensures that the system's P_{fa} does not exceed an upper limit P_{fa}^{th}. Constraint $C6$ bounds the maximum number of measurements (i.e., $\tau T = Z \times t_z$) in the DoA estimation process.

On the other hand, given the stringent requirements of URLLC services, our optimization problem focuses on minimizing the total latency involved in task offloading, comprising both communication and computation latencies, while ensuring that the reliability requirements are met. The objective function, alongside the constraints for URLLC services, is formulated using (38) as follows:

$$f_{\text{TASK}} = \min_{P_v(t),\widetilde{\mathrm{R}}_{vu}(t),\widetilde{\mathrm{R}}_{vl},f_v,f_l} \sum\nolimits_{j=1}^{J} \left(T_{tra}^{vj}(t) + T_{com}^{vj} \right),$$

$$\begin{aligned} &\text{subject to} \\ &\quad \text{C7: } T_{com}^{vj} \leq T_{com}^{th}, \forall j, \\ &\quad \text{C8: } T_{tra}^{vj} \leq T_{tra}^{th}, \forall j, \\ &\quad \text{C9: } \widetilde{\mathrm{R}}_{vu}(t) \geq \mathrm{R}_{vu}^{min}, \forall t, \\ &\quad \text{C10: } \widetilde{\mathrm{R}}_{vl}(t) \geq \mathrm{R}_{vl}^{min}, \forall t, \\ &\quad \text{C11: } P_v(t) \leq P_{\max}, \forall t, \end{aligned} \tag{60}$$

where constraints C7 and C8 ensure that the computation and communication latencies for each sub-task j do not exceed their respective thresholds, T_{com}^{th} and T_{tra}^{th}, crucial for maintaining the URLLC's low-latency requirements. Constraint C9 mandates the adjusted data rate for the vehicle-to-UAV link $\widetilde{\mathrm{R}}_{vu}(t)$ to meet or surpass a minimum R_{vu}^{min}, vital for link reliability and meeting the URLLC latency criteria. Similarly, C10 requires the vehicle-to-satellite link's adjusted data rate $\widetilde{\mathrm{R}}_{vl}(t)$ to exceed R_{vl}^{min}, ensuring efficient data transmission aligning with the URLLC standards. Constraint C11 ensures that the power used for transmission $P_v(t)$ does not exceed a maximum power budget $P_{\max}$ at any given time t.

To achieve an optimal balance between sensing accuracy and communication efficiency, we propose a joint optimization framework formulated as a multi-objective optimization problem:

$$\begin{aligned} &\underset{\substack{\mathbf{R}_{(\theta_w,\phi_w)},Z,\mathbf{\Phi}_{\text{RIS}},\\ P_v(t),\widetilde{\mathrm{R}}_{vu}(t),\widetilde{\mathrm{R}}_{vl}(t),f_v,f_l}}{\text{minimize}} \quad f = \omega_1 f_{\text{DoA}} + (1-\omega_1) f_{\text{TASK}}, \\ &\text{s.t.:} \quad \text{C1 - C11,} \end{aligned} \tag{61}$$

where the weight $0 < \omega_1 < 1$ adjusts the importance of each function to the overall objective, allowing for flexibility in prioritizing between sensing accuracy and communication efficiency based on the application's needs.

IV. Quantum-Aided Multi-Agent DRL Solution

In addressing the intricate and high-dimensional state space challenges of military surveillance systems, we introduce a quantum-aided multi-agent DRL solution. This approach utilizes the parallel processing power of quantum computing, utilizing phenomena such as superposition and entanglement to transcend the limitations of traditional optimization methods [23]. The fusion of quantum computing with multi-agent DRL facilitates enhanced distributed decision-making and learning, strengthening the system's adaptability, resilience, and performance in dynamic operational scenarios [24], [30].

Our proposed architecture introduces a dual-tiered agent framework, as depicted in Fig. 1. Within this framework, we define two distinct operational agents, denoted by $\mathtt{Ag}_{\mathtt{DoA}}$ and $\mathtt{Ag}_{\mathtt{TASK}}$, which are responsible for the immediate execution of DoA estimation and task offloading decisions, respectively. These agents interact directly with the operational environment to fulfill their designated tasks.

A. MDP Model for Operational Agents

The decision-making processes of the operational agents are modeled using an MDP framework to apply the DRL model [20]. Each operational agent has a distinct role and operates within its unique state and action spaces, which are outlined as follows:

1) ***Operational State Space*** *($\mathcal{S}_{\mathtt{oa}}$):* At any discrete time instance t, the operational state space for an agent is represented by $\mathbf{s}_{\mathtt{oa}}(t) \in \mathcal{S}_{\mathtt{oa}}$, which is composed of the state vectors for DoA estimation and task offloading decision processes. The state space is formally given by

$$\mathbf{s}_{\mathtt{oa}}(t) = \{\mathbf{s}_{\mathtt{DoA}}(t), \mathbf{s}_{\mathtt{TASK}}(t) \mid \mathtt{DoA}, \mathtt{TASK} \in \mathtt{oa}\}, \tag{62}$$

where $\mathbf{s}_{\mathtt{DoA}}(t)$ is defined for DoA estimation as $\mathbf{s}_{\mathtt{DoA}}(t) = \{s(t), \mathbf{G}_{vu}, \mathbf{h}_{uw}, \boldsymbol{\eta}_v, \mathbf{x}_v, \mathbf{v}_v, \mathbf{w}_v, \Theta_v, \mathbf{x}_u, \mathbf{v}_u, \mathbf{w}_u, \Theta_u\}$. The state vector for task offloading, $\mathbf{s}_{\mathtt{TASK}}(t)$, comprises $\mathbf{s}_{\mathtt{TASK}}(t) = \{\mathbf{h}_{vu}, \mathbf{h}_{vl}, D_v, \xi_{vu}, \xi_{vl}, c_{vj}, f_v, \sigma_{n_u}^2, x_{vj}^u, \sigma_{n_l}^2, x_{vj}^l, f_l\}$.

2) ***Operational Action Space*** *($\mathcal{A}_{\mathtt{oa}}$):* The action space for each operational agent is defined to align with its specific operational role within the surveillance system. The collective action space, denoted by $\mathbf{a}_{\mathtt{oa}}(t) \in \mathcal{A}_{\mathtt{oa}}$, is composed of the individual action sets for agents involved in DoA estimation and task offloading, respectively:

$$\mathbf{a}_{\mathtt{oa}}(t) = \{\mathbf{a}_{\mathtt{DoA}}(t), \mathbf{a}_{\mathtt{TASK}}(t) \mid \mathtt{DoA}, \mathtt{TASK} \in \mathtt{oa}\}, \tag{63}$$

where the action vector $\mathbf{a}_{\mathtt{DoA}}(t)$ relevant to agent $\mathtt{Ag}_{\mathtt{DoA}}$ at time t includes decisions related to DoA estimation such as $\mathbf{a}_{\mathtt{DoA}}(t) = \{\mathbf{R}_{(\theta_w,\phi_w)}, \mathbf{\Phi}_{\text{RIS}}, Z\}$. Concurrently, the action set for task offloading decisions, $\mathbf{a}_{\mathtt{TASK}}$, for agent $\mathtt{Ag}_{\mathtt{TASK}}$ is constituted by $\mathbf{a}_{\mathtt{TASK}}(t) = \{P_v(t), \widetilde{\mathrm{R}}_{vu}(t), \widetilde{\mathrm{R}}_{vl}(t), f_v(t), f_l(t)\}$, encompassing power allocation, communication rate adjustments, and computational resource management.

3) ***Nash Equilibrium-based Rewards Calculation****:* The multi-agent framework employs a reward structure grounded in a joint objective function, aiming to balance DoA estimation accuracy and task offloading efficiency. This balance is regulated by dynamically allocating time between DoA estimation (τ) and task offloading, directly influencing the system performance as described in the joint minimization problem in (61). To achieve Nash equilibrium [31], where no agent benefits from unilaterally changing its strategy, we define dynamic penalty coefficients and cost functions. These components are designed to penalize deviations from desired performance thresholds, thus incentivizing agents toward optimal behaviour:

$$r_{\mathtt{oa}}(t) = \begin{cases} \frac{\omega_1}{f_{\text{DoA}}(t)} - \lambda_{\text{D}}(t) C_{\text{D}}(t), \text{ for } \mathtt{Ag}_{\mathtt{DoA}} \\ \frac{(1-\omega_1)}{f_{\text{TASK}}(t)} - \lambda_{\text{T}}(t) C_{\text{T}}(t), \text{ for } \mathtt{Ag}_{\mathtt{TASK}} \end{cases}, \tag{64}$$

where $\lambda_{\mathrm{D}}(t) = \max(0, \sum_{c=1}^{6} \kappa_c \mathbb{I}(\mathrm{C}_c \text{ violation}))$ and $\lambda_{\mathrm{T}}(t) = \max(0, \sum_{c=7}^{12} \kappa_c \mathbb{I}(\mathrm{C}_c \text{ violation}))$ denote the dynamic penalty coefficients for DoA estimation and task offloading, respectively. Here, κ_c and κ_c are the penalty weights for constraint violations, and $\mathbb{I}(.)$ indicates a constraint violation. The cost functions are defined as $C_{\mathrm{D}}(t) = |\text{estimated DoA} - \text{actual DoA}|^2$ for DoA estimation error, and $C_{\mathrm{T}}(t) = \max(0, T_{\mathrm{lat}^{\max}}(t) - T_{\mathrm{lat}}(t))$ for task offloading latency, ensuring penalties are directly tied to the magnitude of performance deviation. This structured approach to reward calculation drives the system toward a Nash equilibrium, optimizing the overall surveillance operation. Therefore, the cumulative reward can be expressed as:

$$r_{\mathtt{oa}}^{\mathrm{sum}}(t) = \frac{\omega_1}{f_{\mathrm{DoA}}(t)} - \lambda_{\mathrm{D}}(t) C_{\mathrm{D}}(t) + \frac{(1-\omega_1)}{f_{\mathrm{TASK}}(t)} - \lambda_{\mathrm{T}}(t) C_{\mathrm{T}}(t). \quad (65)$$

Algorithm 1 Quantum State Encoding and Initialization

```
Require: Classical state vectors $\mathbf{s}_{\mathtt{oa}}(t)$
Ensure: Quantum-encoded operational state $\mathbf{s}_{\mathtt{oa}}^{Q}(t)$
1: Normalize $\mathbf{s}_{\mathtt{DoA}}(t)$ and $\mathbf{s}_{\mathtt{TASK}}(t)$ to $\|\mathbf{s}(t)\|_2 = 1$
2: for $\mathbf{s} \in \{\mathbf{s}_{\mathtt{DoA}}(t), \mathbf{s}_{\mathtt{TASK}}(t)\}$ do
3:     Encode $\mathbf{s}$ into $|\psi_{\mathtt{oa}}(t)\rangle = \sum_{i=0}^{N-1} \frac{s_i(t)}{\|\mathbf{s}_{\mathtt{oa}}(t)\|_2} |i\rangle$
4: end for
5: Set $\mathbf{s}_{\mathtt{oa}}^{Q}(t) = \{|\psi_{\mathtt{DoA}}(t)\rangle, |\psi_{\mathtt{TASK}}(t)\rangle\}$
6: Initialize quantum system with $\mathbf{s}_{\mathtt{oa}}^{Q}(t)$
7: return $\mathbf{s}_{\mathtt{oa}}^{Q}(t)$
```

B. Quantum-Aided Multi-agent DRL Framework

By leveraging the power of quantum computation, our approach aims to address the high-dimensional challenges prevalent in military surveillance systems, improving both the DoA estimation accuracy and computational task offloading efficiency. The proposed framework's procedural details and operational insights are thoroughly discussed, with **Algorithm 3** serving as the core for our quantum-aided DRL optimization process.

1) ***Quantum-Encoded Operational State Space*** *($\mathcal{S}_{\mathtt{oa}}^{Q}$):* For a given discrete time instant t, the quantum-encoded operational state space [30], [32], denoted by $\mathbf{s}_{\mathtt{oa}}^{Q}(t) \in \mathcal{S}_{\mathtt{oa}}^{Q}$, describes the quantum states relevant to DoA estimation and task offloading decisions. Specifically, $\mathbf{s}_{\mathtt{oa}}^{Q}(t)$ comprises:

$$\mathbf{s}_{\mathtt{oa}}^{Q}(t) = \{|\psi_{\mathtt{DoA}}(t)\rangle, |\psi_{\mathtt{TASK}}(t)\rangle\}, \quad (66)$$

where $|\psi_{\mathtt{DoA}}(t)\rangle$ and $|\psi_{\mathtt{TASK}}(t)\rangle$ represent the quantum-encoded states (e.g., here N defines the size of the quantum state space) derived from their classical counterparts, $\mathbf{s}_{\mathtt{DoA}}(t)$ and $\mathbf{s}_{\mathtt{TASK}}(t)$, through amplitude encoding [33]. This encoding process initiates with normalizing the classical vectors to the unit norm, followed by the amplitude encoding [33], which maps each vector $\mathbf{s}$ into a quantum state $|\psi\rangle$, as detailed in **Algorithm 1**. This quantum-encoded state space, exploiting quantum superposition, affords a quantum computational advantage by enabling parallel processing of multiple states.

Lemma 1. *Quantum encoding of operational states and actions into quantum states significantly reduces the dimensionality of the decision space, thereby enhancing the efficiency of the learning process in the quantum-aided DRL framework.*

Proof. For an n-qubit quantum system, operational agent states and actions are encoded into a quantum state $|\psi(t)\rangle$ within a Hilbert space $\mathcal{H}$ of dimension 2^n [34]. Utilizing the principle of superposition, this encoding is represented mathematically as:

$$|\psi(t)\rangle = \sum_{i=0}^{2^n-1} \alpha_i |i\rangle, \quad \text{with} \quad \sum_{i=0}^{2^n-1} |\alpha_i|^2 = 1, \quad (67)$$

where $\alpha_i \in \mathbb{C}$ are probability amplitudes, indicating the complex likelihood of the system being found in each basis state upon measurement. The set $\{|i\rangle\}_{i=0}^{2^n-1}$ denotes the computational basis, where each basis state $|i\rangle$ is a direct representation of the binary equivalent of the integer i. The computational basis can be formally defined as: $\mathbf{b} = \{|i\rangle : i \in \{0, 1, \ldots, 2^n - 1\}\}$.

Unitary transformations $\mathcal{U}(t)$ evolve $|\psi(t)\rangle$ into:

$$|\psi'(t)\rangle = \mathcal{U}(t)|\psi(t)\rangle = \sum_{i=0}^{2^n-1} \alpha_i' |i\rangle, \quad (68)$$

with $\alpha_i' = \mathcal{U}(t)\alpha_i$, indicating that the complexity of operations scales as $O(poly(n))$. The Grover search algorithm [35] highlights quantum computational advantages by requiring $O(\sqrt{2^n})$ queries to identify a marked item in a search space S, contrasting with the classical search complexity of $O(2^N)$, where $n \ll N$. □

2) ***Quantum-Enhanced Actor-Critic Framework:*** Building upon the quantum-encoded operational state spaces, our approach employs a quantum-enhanced actor-critic method for each operational agent. This method employs separate networks for the policy (actor) and value function (critic), optimized to work within the quantum computing paradigm.

a) Quantum Circuit Initialization and Actor Network: Quantum circuits, parameterized by action vectors $\mathbf{a}_{\mathtt{oa}}(t)$, process quantum-encoded operational state spaces $\mathbf{s}_{\mathtt{oa}}^{Q}(t)$ through unitary transformations $\mathtt{U}(\varrho_{\mathtt{oa}}(t))$, reflecting the decision-making policies. The initialization of these quantum circuits ($Q_{\mathtt{oa}}$) is formalized as follows [36]:

$$Q_{\mathtt{oa}}(t, \mathbf{a}_{\mathtt{oa}}) = \mathtt{U}(\varrho_{\mathtt{oa}}(t))|\psi_{\mathtt{oa}}(t)\rangle, \quad (69)$$

where $\mathtt{U}(\varrho_{\mathtt{oa}}(t))$ represents the quantum equivalent of actions, optimized to maximize the expected reward. This unitary operation transforms the quantum-encoded states according to agent-specific actions $\mathbf{a}_{\mathtt{oa}}(t)$, mapping the initial state $|\psi_{\mathtt{oa}}(t)\rangle$ to a new state $|\psi_{\mathtt{oa}}'(t)\rangle$ as:

$$\mathtt{U}(\varrho_{\mathtt{oa}}(t)) : |\psi_{\mathtt{oa}}(t)\rangle \mapsto |\psi_{\mathtt{oa}}'(t)\rangle. \quad (70)$$

The evolution of these states under the influence of actions is governed by the Hamiltonian $\mathrm{H}_{\mathtt{oa}}(\varrho_{\mathtt{oa}})$, with the unitary operation expressed as $\mathtt{U}(\varrho_{\mathtt{oa}}(t)) = e^{-i\mathrm{H}_{\mathtt{oa}}(\varrho_{\mathtt{oa}}(t))}$, which encodes the total energy of the system. This is not about the physical energy in the conventional sense but rather a mathematical representation of the system's energy states and transitions within the quantum computational model. Mathematically, $\mathrm{H}_{\mathtt{oa}}(\varrho_{\mathtt{oa}})$ is defined as:

$$\mathrm{H}_{\mathtt{oa}} = \sum_i \epsilon_i |i\rangle\langle i| + \sum_{i \neq j} \tau_{ij} (|i\rangle\langle j| + |j\rangle\langle i|), \quad (71)$$

where ϵ_i represents the energy associated with the system being in a particular state $|i\rangle$, and τ_{ij} represents the transition energy between states $|i\rangle$ and $|j\rangle$.

The actor-network employs variational quantum circuits, parameterized by $\theta_{\mathtt{oa}}^{\pi}$, to efficiently explore action probabilities through quantum superposition and entanglement [37], written as:

$$\pi_{\theta^{\pi}_{\mathtt{oa}}}(\mathbf{a}_{\mathtt{oa}}(t)|\mathbf{s}^{Q}_{\mathtt{oa}}(t)) = \langle\psi_{\mathtt{oa}}(t)|\mathtt{U}^{\dagger}(\varrho_{\mathtt{oa}}(t)) \\ \mathtt{U}(\varrho_{\mathtt{oa}}(t))|\psi_{\mathtt{oa}}(t)\rangle, \tag{72}$$

where $\mathtt{U}^{\dagger}(\varrho_{\mathtt{oa}}(t))$ is its Hermitian adjoint, ensuring reversibility and the preservation of quantum state properties during policy application. To optimize $\theta^{\pi}_{\mathtt{oa}}$, we use the parameter shift rule to estimate the gradient as follows:

$$\nabla_{\theta^{\pi}_{\mathtt{oa}}} J(\theta^{\pi}_{\mathtt{oa}}) = \frac{1}{2}\Big(\langle\partial_{\theta^{\pi}_{\mathtt{oa}}}\psi_{\mathtt{oa}}(t)|\mathtt{U}^{\dagger}(\varrho_{\mathtt{oa}}(t))\mathtt{U}(\varrho_{\mathtt{oa}}(t))|\psi_{\mathtt{oa}}(t)\rangle \\ + \langle\psi_{\mathtt{oa}}(t)|\mathtt{U}^{\dagger}(\varrho_{\mathtt{oa}}(t))\mathtt{U}(\varrho_{\mathtt{oa}}(t))|\partial_{\theta^{\pi}_{\mathtt{oa}}}\psi_{\mathtt{oa}}(t)\rangle\Big). \tag{73}$$

b) Critic Network Evaluation: According to the current policy, the critic network, parameterized by $\theta^{Q}_{\mathtt{oa}}$, evaluates the expected return of taking an action **a** in-state **s**. This evaluation guides the policy improvement by providing feedback on the action value. This evaluation is quantitatively expressed as:

$$Q_{\theta^{Q}_{\mathtt{oa}}}(\mathbf{s}^{Q}_{\mathtt{oa}}(t), \mathbf{a}_{\mathtt{oa}}(t)) = \langle\psi'_{\mathtt{oa}}(t)|M_{\theta^{Q}_{\mathtt{oa}}}|\psi'_{\mathtt{oa}}(t)\rangle, \tag{74}$$

where $M_{\theta^{Q}_{\mathtt{oa}}}$ represents the measurement operator parameterized by the critic network.

c) Replay Buffer: A quantum-enhanced replay buffer is utilized to store experience tuples $(\mathbf{s}, \mathbf{a}, r, \mathbf{s}')$, collected from interactions with the environment. This buffer serves as a database for sampling mini-batches of experiences [19], reducing the correlation in the observation sequence and improving the stability and efficiency of learning:

$$\mathcal{D} = \{(\mathbf{s}^{Q}_{\mathtt{oa}}(t_i), \mathbf{a}(t_i), r(t_i), \mathbf{s}^{Q}_{\mathtt{oa}}(t_{i+1}))\}_{i=1}^{N}, \tag{75}$$

where $\mathcal{D}$ denotes the replay buffer containing N experiences, facilitating the training of both the actor and critic networks within the quantum-augmented DRL framework.

d) Learning with Quantum-Enhanced TD Error: The optimization of actor and critic networks within the quantum-augmented DRL framework utilizes a quantum-enhanced temporal difference (TD) learning approach [38]. This involves computing the TD error in a manner that accounts for the quantum-encoded states and the probabilistic nature of quantum measurements. Given a quantum-encoded state $\mathbf{s}^{Q}_{\mathtt{oa}}(t)$ and its successor $\mathbf{s}^{Q}_{\mathtt{oa}}(t+1)$, along with the reward $r(t)$. The quantum-enhanced TD error, accounting for experiences sampled from the replay buffer, is defined as [38]:

$$\delta_{\mathcal{D}}(t) = r_{\mathtt{oa}}(t) + \gamma\langle\psi_{\mathbf{s}'^{Q}_{\mathtt{oa}}}|Q_{\theta^{Q}_{\mathtt{oa}}}|\psi_{\mathbf{s}'^{Q}_{\mathtt{oa}}}\rangle - \langle\psi_{\mathbf{s}^{Q}_{\mathtt{oa}}}|Q_{\theta^{Q}_{\mathtt{oa}}}|\psi_{\mathbf{s}^{Q}_{\mathtt{oa}}}\rangle, \tag{76}$$

where $\psi_{\mathbf{s}^{Q}_{\mathtt{oa}}}$ and $\psi_{\mathbf{s}'^{Q}_{\mathtt{oa}}}$ denote the quantum-encoded states of the current and next states sampled from the replay buffer $\mathcal{D}$, enhancing the learning stability and efficiency.

To incorporate experiences from the replay buffer in the optimization of critic network parameters $\theta^{Q}_{\mathtt{oa}}$, the loss function is defined as:

$$\mathcal{L}_{\mathcal{D}}(\theta^{Q}_{\mathtt{oa}}) = \mathbb{E}_{(\mathbf{s},\mathbf{a},r,\mathbf{s}')\sim\mathcal{D}}\left[\delta_{\mathcal{D}}(t)^2\right] + \lambda_1\|\theta^{Q}_{\mathtt{oa}}\|_2^2 \\ - \lambda_2\mathbb{E}_{(\mathbf{s},\mathbf{a},r,\mathbf{s}')\sim\mathcal{D}}\left[\mathcal{F}(\rho_{\psi_{\mathbf{s}'^{Q}_{\mathtt{oa}}}}, \sigma_{\psi_{\mathbf{s}'^{Q}_{\mathtt{oa}}}}(\theta^{Q}_{\mathtt{oa}}))\right], \tag{77}$$

where the expectations are over the distribution of experiences $(\mathbf{s}, \mathbf{a}, r, \mathbf{s}')$ sampled from the replay buffer $\mathcal{D}$, facilitating the training of actor and critic networks within the quantum-augmented DRL framework, and $\mathbb{E}[\delta(t)^2]$ denotes the expected squared TD error to minimize the discrepancy in predicted versus actual rewards. The L2 regularization term, $\lambda_1\|\theta^{Q}_{\mathtt{oa}}\|_2^2$ is used to prevent overfitting by penalizing large weights. We use a quantum fidelity term, $\mathcal{F}$, to encourage the critic network to accurately reflect the underlying quantum state dynamics by maximizing the fidelity between the target and predicted quantum states.

Algorithm 2 Quantum State Optimization with Feedback

```
Require: ${\mathbf{s}^{Q}_{\mathtt{oa}}(t)}_{\mathtt{oa}=1}^{N}$, $\mathcal{A}_{\mathtt{oa}}$.
Ensure: $\mathbf{a}_{\text{opt}}(t)$.
$\Psi_{\text{init}} \leftarrow \bigotimes_{\mathtt{oa}=1}^{N} |\psi_{\mathtt{oa}}(t)\rangle$
for $i \leftarrow 1$ to $N$ do
    $Q_{\mathtt{Ag}_i} \leftarrow U_{\text{encode}}(\mathbf{s}_{\mathtt{Ag}_i})|\mathbf{0}\rangle^{\otimes n}$
end for
$\mathcal{H}_{\text{global}} \leftarrow \sum_{\mathtt{oa},\mathtt{oa}'} H_{\mathtt{oa},\mathtt{oa}'}$
Prepare an initial quantum state $|\Psi_{\text{init}}\rangle$
Define $Q(\varrho_{\text{oa}}) = \mathrm{U}_n(\varrho_{\text{oa}_n})\mathrm{U}_{n-1}(\varrho_{\text{oa}_{n-1}})\ldots\mathrm{U}_1(\varrho_{\text{oa}_1})$
while not converged do
    Apply $Q(\varrho_{\text{oa}})$ to $|\Psi_{\text{init}}\rangle$ to get $|\Psi(\varrho_{\text{oa}})\rangle$
    Measure $\mathbb{E}(\varrho_{\text{oa}}) = \langle\Psi(\varrho_{\text{oa}})|\mathcal{H}_{\text{global}}|\Psi(\varrho_{\text{oa}})\rangle$
    $\varrho_{\min} = \operatorname{argmin}_{\varrho_{\text{oa}}} \mathbb{E}(\varrho_{\text{oa}})$
    Update $\varrho_{\text{oa}} \leftarrow \varrho_{\min}$
end while
$\Psi_{\text{ground}} \leftarrow |\Psi(\varrho_{\min})\rangle$
$\mathbf{a}_{\text{opt}}(t) \leftarrow \mathbf{a}_{\text{op}}(t) \leftarrow$ Measure$(\Psi_{\text{ground}})$
return $\mathbf{a}_{\text{opt}}(t)$
```

e) Quantum State Optimization with Feedback Loop: In **Algorithm 2**, each $|\psi_{\mathtt{oa}}(t)\rangle$ represents the quantum-encoded state of an individual agent at time t. By taking the tensor product $\bigotimes_{\mathtt{oa}=1}^{N}|\psi_{\mathtt{oa}}(t)\rangle$, we construct a multi-agent quantum state that encompasses the entire system's state information. The optimization of quantum states, coupled with a feedback mechanism, plays a pivotal role in enhancing the performance of DRL agents in dynamic environments [38], [39]. The algorithmic framework outlined in **Algorithm 2** directs the optimization process based on quantum principles. The core of the optimization lies in the application of unitary transformations $Q(\varrho_{\text{oa}})$, which evolve the quantum state to explore the decision space. These transformations are defined as:

$$Q(\varrho_{\text{oa}}) = \mathrm{U}_n(\varrho_{\text{oa}_n})\mathrm{U}_{n-1}(\varrho_{\text{oa}_{n-1}})\ldots\mathrm{U}_1(\varrho_{\text{oa}_1}), \tag{78}$$

where each $\mathrm{U}_i(\varrho_{\text{oa}_i})$ represents a parameterized unitary operation, reflecting the decision-making policy. The objective is to find the optimal parameters ϱ_{oa} that maximize the reward, as quantified by the measurement

$$\mathbb{E}(\varrho_{\text{oa}}) = \langle\Psi(\varrho_{\text{oa}})|\mathcal{H}_{\text{global}}|\Psi(\varrho_{\text{oa}})\rangle, \tag{79}$$

where $\mathcal{H}_{\text{global}}$ represents the global Hamiltonian of the system, encapsulating the interaction between agents and the environment. The optimization process iteratively adjusts ϱ_{oa} to find the minimum of $\mathbb{E}(\varrho_{\text{oa}})$, indicative of the optimal decision-making strategy.

The efficacy of actions is evaluated through projective measurement operators $\{M_m\}$, which, when applied to the post-action quantum states [38]. The feedback for actions taken by any agent $\mathtt{Ag}_{\mathtt{oa}}$ is computed as:

$$b_{\mathtt{oa}}(t) = \sum\nolimits_m m\langle\psi'_{\mathtt{oa}}(t)|M_m^{\dagger}M_m \otimes |\mathbf{a}_{\text{opt}}(t)\rangle\langle\mathbf{a}_{\text{opt}}|\psi'_{\mathtt{oa}}\rangle, \tag{80}$$

where $\mathbf{a}_{\text{opt}}(t)$ in the measurement process, allowing for the evaluation of action efficacy $b_{\mathtt{oa}}(t)$ for the post-action quantum states and the optimized actions taken by the agents. $b_{\mathtt{oa}}(t)$ denotes the weighted sum of all possible measurement outcomes for actions undertaken by the agent $\mathtt{Ag}_{\mathtt{oa}}$. Higher values of $b_{\mathtt{oa}}(t)$ indicate favourable actions, while lower values suggest

Algorithm 3 Quantum-aided DRL with Actor-Critic Networks

```
Initialize actor network $\theta^{\pi}_{\mathtt{oa}}$, critic network $\theta^{Q}_{\mathtt{oa}}$, and replay buffer $\mathcal{D}$
Prepare initial quantum-encoded state $\mathbf{s}^{Q}_{\mathtt{oa}}(t)$ using **Algorithm 1**
**while** not converged **do**
    **for** each timestep $t$ **do**
        Sample a mini-batch $\mathcal{B}_t = \{(\mathbf{s}_i, \mathbf{a}_i, r_i, \mathbf{s}'_i)\}_{i=1}^{N_b} \in \mathcal{D}$
**Actor Network Optimization:**
        Compute policy $\pi_{\theta^{\pi}_{\mathtt{oa}}}(\mathbf{a}_{\mathtt{oa}}(t)|\mathbf{s}^{Q}_{\mathtt{oa}}(t))$ using (72)
        Estimate gradient using the rule in (73)
        Update: $\theta^{\pi}_{\mathtt{oa}} \leftarrow \theta^{\pi}_{\mathtt{oa}} + \alpha \nabla_{\theta^{\pi}_{\mathtt{oa}}} J(\theta^{\pi}_{\mathtt{oa}})$
**Critic Network Evaluation:**
        Compute action value $Q_{\theta^{Q}_{\mathtt{oa}}}(\mathbf{s}^{Q}_{\mathtt{oa}}(t), \mathbf{a}_{\mathtt{oa}}(t))$ using (74)
        Calculate TD error for critic network using (76)
        Update $\theta^{Q}_{\mathtt{oa}}$ by minimizing $\mathcal{L}(\theta^{Q}_{\mathtt{oa}})$ in (77)
**Feedback Loop via Quantum Measurement:**
        Call **Algorithm 2**
        Evaluate action efficacy $b_{\mathtt{oa}}(t)$ via (80)
**Optimization of Action Parameters:**
        Update action parameters $\varrho_{\mathtt{oa}}$ using (81)
    **end for**
**end while**
```

less desirable actions.

f) Optimization of Action Parameters: The iterative refinement of the action parameters $\varrho_{\mathtt{oa}}$ for each agent is instrumental. The following expression guides this refinement process:

$$\varrho_{\mathtt{oa}}(t+1) = \varrho_{\mathtt{oa}}(t) - \alpha \nabla_{\varrho_{\mathtt{oa}}} \mathcal{L}_{\text{act}}(\mathbf{s}_{\mathtt{oa}}(t), \mathbf{a}_{\mathtt{oa}}(t), \varrho_{\mathtt{oa}}(t)), \quad (81)$$

where α represents the learning rate, and $\nabla_{\varrho_{\mathtt{oa}}}$ denotes the gradient of the loss function, $\mathcal{L}_{\text{act}}$, which is defined as:

$$\begin{aligned}\mathcal{L}_{\text{act}}(\mathbf{s}_{\mathtt{oa}}(t), \mathbf{a}_{\mathtt{oa}}(t), \varrho_{\mathtt{oa}}(t)) = \beta(Q_{\theta^{Q}_{\mathtt{oa}}}(\mathbf{s}_{\mathtt{oa}}(t), \mathbf{a}_{\mathtt{oa}}(t)) \quad (82)\\ - V_{\theta^{V}_{\mathtt{oa}}}(\mathbf{s}_{\mathtt{oa}}(t)))^2 + (1-\beta)\mathrm{D}_{\mathrm{KL}}(\pi_{\theta^{\pi}_{\mathtt{oa}}} \| \pi_{\theta^{\pi}_{\mathtt{oa}}{}'}),\end{aligned}$$

where β is a balancing coefficient. $V_{\theta^{V}_{\mathtt{oa}}}$ denotes the critic's estimate of the expected return from state $\mathbf{s}_{\mathtt{oa}}(t)$, parameterized by the weights $\theta^{V}_{\mathtt{oa}}$. D_{KL} is the Kullback-Leibler divergence measuring the difference between the current policy $\pi_{\theta^{\pi}_{\mathtt{oa}}}$ and a target policy $\pi_{\theta^{\pi}_{\mathtt{oa}}{}'}$.

C. Computational Complexity of Quantum-aided DRL

Given the quantum-encoded state space $\mathcal{S}^{Q}_{\mathtt{oa}}$ for each agent $\mathtt{Ag}_{\mathtt{DoA}}$ and $\mathtt{Ag}_{\mathtt{TASK}}$, the quantum state encoding exhibits a complexity of $\mathcal{O}(\log \mathrm{D})$, utilizing amplitude encoding within an n-qubit system, where $\mathrm{D} = 2^n$. The quantum-encoded operational state $|\psi^{Q}_{\mathtt{oa}}(t)\rangle$ is defined as $|\psi^{Q}_{\mathtt{oa}}(t)\rangle = \sum_{i=0}^{2^n-1} \alpha_i |i\rangle$, with the normalization condition $\sum_{i=0}^{2^n-1} |\alpha_i|^2 = 1$. The computational complexity associated with the quantum decision-making, facilitated by the unitary transformations $\mathtt{U}(\varrho_{\mathtt{oa}}(t))$, is $\mathcal{O}(\mathrm{G}_{\mathrm{U}} n_{\mathrm{U}})$, where G_{U} represents the gate count and n_{U} denotes the qubit count involved in $\mathtt{U}$. The optimization process, involving iterative adjustments over I number of iterations of parameters $\varrho_{\mathtt{oa}}$, refers to (81). Therefore, the total computational complexity is expressed as

$$C_{\text{total}} = \mathcal{O}(\log \mathrm{D} + \mathrm{G}_{\mathrm{U}} n_{\mathrm{U}} + (\mathrm{I} \times \mathtt{U} \times n)). \quad (83)$$

D. Convergence Analysis

To establish the reliability and effectiveness of the proposed quantum-aided multi-agent DRL framework, we present a theoretical analysis demonstrating the convergence of our solution. The proof is predicated on the principles of quantum computation and RL theory, ensuring a systematic approach towards achieving an optimal policy.

Theory 1. *Given a quantum-aided multi-agent DRL framework with the Hilbert space $\mathcal{H}$ for quantum state encodings $|\psi^{Q}_{\mathtt{oa}}(t)\rangle$, and unitary operations $\mathtt{U}(\varrho_{\mathtt{oa}}(t))$ for policy representation, the framework converges to an optimal policy π^*.*

Proof. Consider a quantum-aided DRL framework wherein the state of each agent at time t is quantum-encoded as $|\psi^{Q}_{\mathtt{oa}}(t)\rangle \in \mathcal{H}$, with actions executed through parameterized unitary operations $\mathtt{U}(\varrho_{\mathtt{oa}}(t))$. The evolution under action a is described by:

$$|\psi'^{Q}_{\mathtt{oa}}(t)\rangle = \mathtt{U}(\varrho_{\mathtt{oa}}(t))|\psi^{Q}_{\mathtt{oa}}(t)\rangle. \quad (84)$$

The policy $\pi_{\theta^{\pi}_{\mathtt{oa}}}$, parameterized by $\theta^{\pi}_{\mathtt{oa}}$, is optimized by updating $\theta^{\pi}_{\mathtt{oa}}$ to maximize the expected cumulative reward. The policy gradient, derived using the parameter shift rule, is:

$$\nabla_{\theta^{\pi}_{\mathtt{oa}}} J = \frac{1}{2}\left(\langle \partial_{\theta^{\pi}_{\mathtt{oa}}} \psi_{\mathtt{oa}} | \mathtt{U}^{\dagger}\mathtt{U} | \psi_{\mathtt{oa}} \rangle + c^*\right), \quad (85)$$

where c^* denotes the complex conjugate, the Born rule provides feedback for policy updates after quantum measurement collapses $|\psi^{Q}_{\mathtt{oa}}(t)\rangle$ to classical outcomes [36].

Define $V^{\pi}(s)$ as the expected return from state s under policy π, and $Q^{\pi}(s,a)$ as the expected return from taking action a in state s and following π. The Bellman optimality equations are given by

$$V^*(|\psi^{Q}_{s}\rangle) = \max_{a \in \mathcal{A}} Q^*(|\psi^{Q}_{s}\rangle, a), \quad (86)$$

$$Q^*(|\psi^{Q}_{s}\rangle, a) = \mathbb{E}\left[\mathcal{O}_{R(s,a)} + \gamma V^*(|\psi^{Q}_{s'}\rangle) | |\psi^{Q}_{s}\rangle, a\right], \quad (87)$$

where $\mathcal{O}_{R(s,a)}$ represents the quantum observable corresponding to the reward for taking action a in state $|\psi^{Q}_{s}\rangle$ is defined by a Hermitian operator that acts on the Hilbert space $\mathcal{H}$ which can be expressed as

$$\mathbb{E}_{\mathcal{O}_{R(s,a)}} = \langle \psi^{Q}_{s} | \mathcal{O}_{R(s,a)} | \psi^{Q}_{s} \rangle, \quad (88)$$

The compactness of $\mathcal{H}$ and continuity of $\mathtt{U}(\varrho_{\mathtt{oa}}(t))$ imply that for any $\epsilon > 0$, there exists a $\delta > 0$ such that $\|\mathtt{U}(\varrho_{\mathtt{oa}}(t)) - \mathtt{U}(\varrho_{\mathtt{oa}}(t+\delta))\| < \epsilon$ for all t. Thus, as $t \to \infty$, we have:

$$\lim_{t \to \infty} \|\nabla_{\theta^{\pi}_{\mathtt{oa}}} J(\theta^{\pi}_{\mathtt{oa}}(t))\| = 0, \quad (89)$$

ensuring convergence of $V^{\pi}(|\psi^{Q}_{s}\rangle)$ to $V^*(|\psi^{Q}_{s}\rangle)$ and $Q^{\pi}(|\psi^{Q}_{s}\rangle, a)$ to $Q^*(|\psi^{Q}_{s}\rangle, a)$ for all $|\psi^{Q}_{s}\rangle \in \mathcal{H}$ and $a \in \mathcal{A}$, thereby establishing convergence to the optimal policy π^*. □

From (65), we consider the reward sequence $\{r_i = r^{\text{sum}}_{\mathtt{oa}}(t)\}_{t=1}^{T}$ where T is the total number of episodes. The moving average μ_t and variance σ_t^2 over a window of size W are defined as:

$$\mu_t = \frac{1}{W} \sum_{i=t-W+1}^{t} r_i, \quad \sigma_t^2 = \frac{1}{W} \sum_{i=t-W+1}^{t} (r_i - \mu_t)^2. \quad (90)$$

Convergence is determined when σ_t^2 remains below the threshold for the last W episodes. In our analysis, we set $W = 100$. The reward variance σ_t^2 over the final window can be expressed as:

$$\sigma^2_{T-W+1} = \frac{1}{100} \sum_{i=T-99}^{T} (r_i - \mu_{T-W+1})^2. \quad (91)$$

If $\sigma^2_{T-W+1} < 0.05$, we conclude the algorithm has converged.

V. Numerical Results and Analysis

The present military surveillance system operates within a simulated 4 km^2 urban environment [4], with UAVs, satellites,

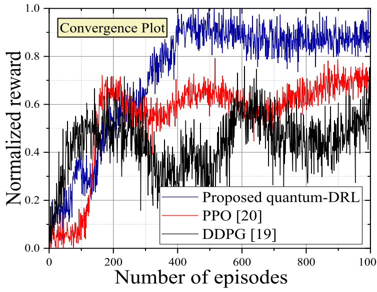
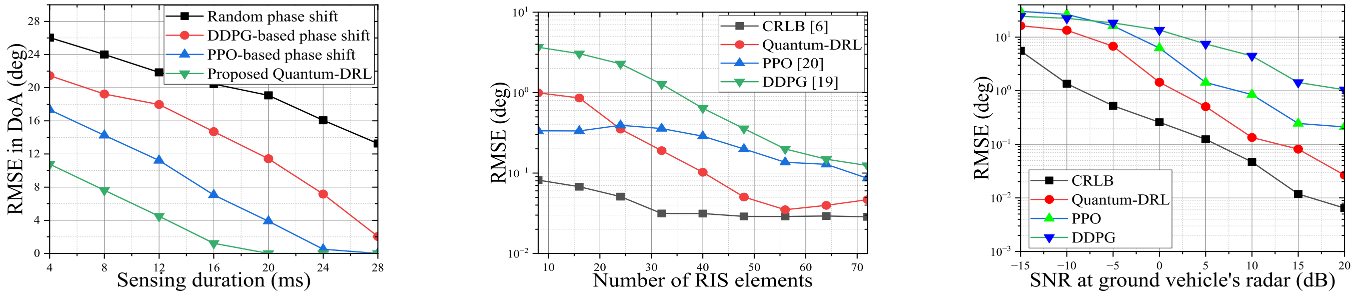


Fig. 3: Convergence plot: reward vs. episode.

and ground vehicles working through LoS and NLoS conditions under consideration of Rician factor $\kappa = 6.46$ [16] and $\epsilon_{\max} = 0.05$. We further consider a 100 ms time frame duration, and dynamic altitudes for UAVs are up to 100 meters [2]. UAV and ground vehicle velocities are limited to 15 m/s and 10 m/s, respectively, while satellites maintain a fixed orbit at 550 km [4] with a velocity of 7.8 km/s. The number of UAVs, UFOs, satellites, and ground vehicles is set to 16, 4, 8, and 2, respectively, with each UAV equipped with a 64-element-aided RIS [11]. The number of maximum measurements ($Z_{\max}$) is fixed to 32 [11]. The carrier frequency for the radar signals is 2.4 GHz, and the allocated bandwidth to each military vehicle is 20 MHz [4], [11]. The URLLC packet error rate is fixed to 10^{-5}, and the block length is set to 256 bit. The computational task size is 1 Mbits [2], with computational complexities $(100, 300)$ cycles/bit [2]. The satellites and vehicles have computational capacities of 12 GHz [2] and 4.8 GHz, respectively. We set the threshold for $P_d^{th} = 0.9$, and $P_{fa}^{th} = 0.1$ [29]. The maximum transmission power of vehicles is fixed at 2 W [2]. We also set the minimum required data rate for communications from vehicles to UAVs and satellites at 20 Mbps [4]. The framework imposes a maximum tolerable latency of 15 ms and 20 ms for task transmission and computing processes, respectively. Noise levels at the receivers are consistently maintained at −130 dBm [2]. For all simulation runs except the plot in Fig. 3, the results are averaged over 100 simulation runs to ensure statistical reliability and smooth out any fluctuations and irregularities.

We employ an 8-qubit quantum system on Google Cirq and QSimhSimulator for constructing variational quantum circuits that construct the actor-critic networks essential to our DRL framework. The variational circuits are composed of four layers, each layer hosting 128 neurons and employing ReLU activation functions aimed at simulating the policy and value function estimations. These layers utilize parameterized quantum gates, including R_x, R_y, R_z, and CNOT gates, to facilitate the encoding of the system's state and the execution of complex network interactions through quantum gradient descent. For the quantitative analysis, the framework utilizes a discount factor (γ) of 0.99 [2] and a learning rate (α) of 0.0001 [4]. The replay buffer accommodates $10,000$ experiences; batch size is 32, and the maximum number of training episodes is 100 [2], ensuring a robust dataset for network training and updates. Comparative benchmarks include DDPG [19] and PPO [20] algorithms executed on a computing setup featuring an Intel i9 processor, 64GB RAM, and an NVIDIA RTX GPU.

Fig. 3 illustrates the superior performance enhancements achieved by our proposed quantum-DRL framework when applied to present military surveillance systems, in comparison to conventional DRL techniques such as DDPG [19] and PPO [20]. For better visualization, we normalize the actual reward of all outcomes in the range of $(0, 1)$, using $r_{norm_i} = (r_{\text{i}} - r_{min}) / (r_{max} - r_{min})$. Quantum-DRL outperforms the DDPG and PPO by approximately 76.32% and 48.73% in normalized reward at episode 1000, respectively. This noteworthy enhancement highlights the quantum-DRL's superior capability in navigating complex state-action spaces efficiently. Furthermore, a critical observation from our experiments is the convergence rate of quantum-DRL. For our proposed quantum-DRL algorithm, we observe that the reward reaches a relatively stable state after approximately 650 episodes, although minor oscillations are present. These oscillations are attributed to the inherent exploration-exploitation trade-off in reinforcement learning, particularly in complex environments. For a more formal and quantitative criterion, we consider an algorithm to have converged if the variance of the normalized reward over the last 100 episodes falls below a predefined threshold. Specifically, we measure the moving average and standard deviation of the reward over a sliding window of 100 episodes. Convergence is achieved when the standard deviation remains consistently low (we consider below 0.05) over the window. In the case of the proposed quantum-DRL, the normalized reward stabilizes with minor fluctuations, suggesting effective learning and adaptation. Comparatively, the benchmarks PPO [20] and DDPG [19], also exhibit a slower convergence, but with differing stability levels. PPO stabilizes earlier, while DDPG shows more variability. This faster convergence rate of quantum-DRL is attributed to its quantum-enhanced decision-making process, which utilizes quantum parallelism and entanglement to explore and exploit the decision space more comprehensively than classical approaches.

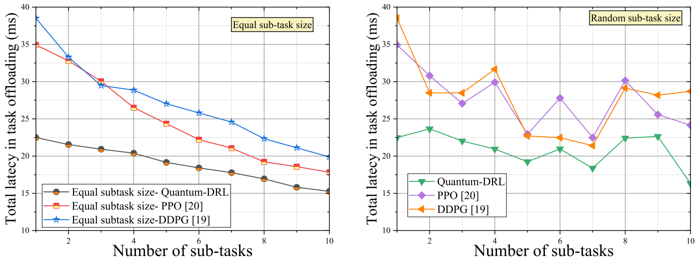


(a) RMSE in DoA vs. sensing duration.

(b) RMSE in DoA vs. RIS elements.

(c) RMSE in DoA vs. SNR level.

Fig. 4: Comparison of RMSE in DoA under various conditions.



(a) Equal sub-task size portioning. (b) Random sub-task size portioning.

Fig. 5: Offloading latency vs. no. of sub-tasks.

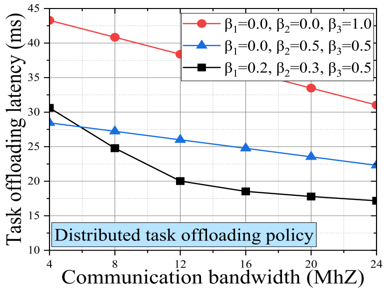


Fig. 6: Offloading latency vs. wireless bandwidth.

The unique ability of the quantum framework to process and encode high-dimensional data allows for a deeper understanding of the operational environment, thereby significantly contributing to the overall system performance.

In Fig. 4a, the efficacy of the quantum-DRL-based RIS phase shift design over its counterparts is distinctly evident through the substantial reduction in RMSE for DoA estimation across sensing durations. The quantum-DRL-based approach yields a reduction in RMSE compared to random-phase shift-based RIS design, DDPG, and PPO-based phase shift designs by 94.10%, 91.66%, and 82.61% respectively at a sensing duration of 16 ms. Such efficiency highlights the quantum DRL framework's superior capability to efficiently explore and optimize the complex, high-dimensional solution space.

Fig. 4b evaluates the performance of the quantum-DRL method against the theoretical lower bound for RMSE in DoA estimation, represented by the CRLB [6]. The reduction in RMSE with the quantum-DRL approach compared to PPO [20] and DDPG [19] methods demonstrates its superior efficiency and closer adherence to the CRLB. The performance gain in RMSE reduction for the quantum-DRL approach compared to the PPO method is approximately 69.16%, and compared to the DDPG method, it is approximately 73.37%, when number of elements in RIS is 64. These findings underscore the quantum-DRL framework's enhanced efficacy in approaching the theoretical accuracy limits set by the CRLB.

In Fig. 4c, the quantum-DRL approach significantly outperforms its counterparts in reducing the RMSE for DoA estimation. At 5 dB SNR, quantum-DRL reduces the RMSE by approximately 64.59% and 93.23% compared to PPO and DDPG, respectively, showcasing its significant advantage in minimizing estimation errors. This performance highlights quantum-DRL's capabilities in optimizing RIS phase shift design even in low SNR scenarios, demonstrating its capacity for near-theoretical accuracy and robustness to noise.

Fig. 5a presents a comparative analysis of task latency reductions across different partitioning schemes. Our quantum-DRL approach demonstrates superior performance at the partitioning level of actual main task to 10 subtasks, showcasing a notable decrease in task latency over DDPG [19] and PPO [20]. In equal subtask sizing as shown in Fig. 5b, quantum-DRL achieves a performance gain of 23.18% and 14.36% compared to DDPG and PPO, respectively. When employing a random subtask sizing strategy, the efficiency of quantum-DRL is further accentuated, yielding performance gains of 43.09% over DDPG and 32.35% over PPO.

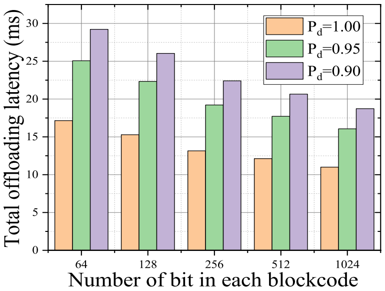


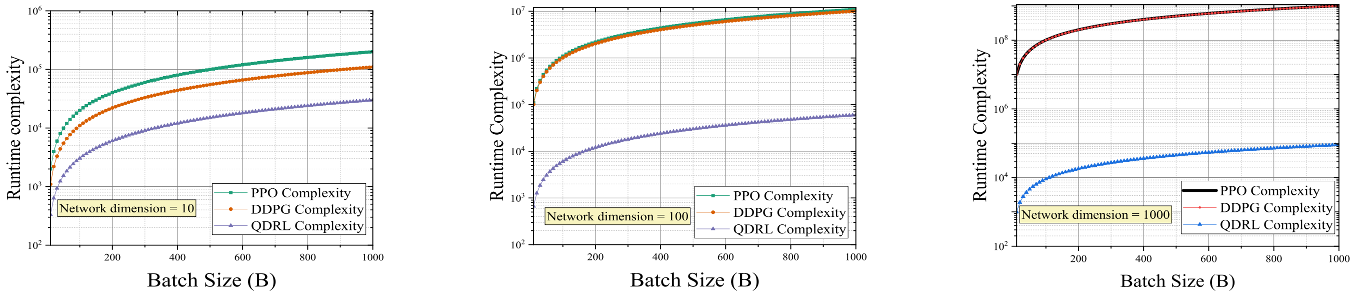
Fig. 7: Offloading latency vs. blockcode length.

Fig. 6, evaluates the task offloading latency versus allocated system bandwidth by various task distribution strategies, as detailed in Section II-C. The setting employing a balanced distribution with $\beta_1 = 0.2$, $\beta_2 = 0.3$, and $\beta_3 = 0.5$ demonstrates a compelling performance gain by reducing the task offloading latency by 47.83% while bandwidth is 20 mHz compared to a satellite MEC-only scenario ($\beta_3 = 1$), which does not utilize local processing or UAV caching. Moreover, an arrangement with $\beta_1 = 0$, $\beta_2 = 0.5$, and $\beta_3 = 0.5$ enhances this efficiency, showcasing a latency reduction of 29.82% relative to the satellite MEC-only configuration. This configuration demonstrates the critical interaction between UAV caching and satellite MEC, emphasizing the significance of strategic task distribution.

Fig. 7 reveals how the P_d significantly influences vehicular task offloading latency with the changes in block length. With a block length of 256 bits, setting $P_d = 1$ showcases substantial performance improvements, yielding a latency reduction of approximately 31.56% compared to a $P_d = 0.95$, and an even more pronounced reduction of about 41.30% when set against a P_d threshold of 0.9. This trend is consistent with larger block lengths, where increased P_d consistently correlates with lower task offloading latency. The reduction in latency becomes more substantial as the block length increases. This behavior highlights the critical importance of accurate detection of UFOs in enhancing system performance, as higher P_d values directly contribute to more efficient task processing and reduced task offloading latency in vehicular networks, emphasizing the need for optimized detection.

A. Runtime Complexity Analysis

An analysis of the runtime for each algorithm can offer valuable insights into their efficiency and practicality. However, directly presenting the programming running time can be influenced by various external factors such as programming language, hardware architecture, and coding styles. To address these variabilities, we propose presenting a runtime complexity analysis of the key algorithms depicted in Fig. 4, namely PPO, DDPG, and QDRL. For PPO, during the policy update, it requires gradient



(a) Runtime complexity at $d = 10$.

(b) Runtime complexity at $d = 100$.

(c) Runtime complexity at $d = 1000$.

Fig. 8: Comparative analysis of runtime complexity across varying network dimensions for PPO, DDPG, and QDRL algorithms.

computation with a complexity of $O(TBd)$, where T is the time steps per update, B is the batch size, and d is the policy network dimension. The clipped objective function adds a complexity of $O(Bd^2)$. The value function update, which has a similar complexity to the policy update, is $O(TBd)$. Therefore, the minimum runtime complexity for PPO becomes $O(TBd + Bd^2)$. On the other hand, for DDPG, the actor update complexity through gradient computation is $O(Bd)$. For the critic network update, the gradient computation complexity is $O(Bd^2)$, along with the value network update of $O(Bd^2)$. Considering the replay buffer sampling with a complexity of $O(B)$, the overall runtime complexity of DDPG becomes $O(B + Bd + Bd^2 + Bd^2)$, based on neural network dimension which simplifies to $O(Bd + Bd^2)$. For QDRL, the runtime complexity is given in (83). Here, the state space dimension D is similar to the network dimension d in PPO and DDPG. Since D = 2^n, we can set n such that D is comparable to d. The number of iterations (I) and unitary transformations ($\mathbb{U}$) in QDRL should correspond to the number of updates in PPO and DDPG. Additionally, the gate count ($G_{\mathbb{U}}$) and qubit count ($n_{\mathbb{U}}$) can be aligned with the network dimensions and update mechanisms of PPO and DDPG. Given these considerations, we can adjust the QDRL parameters for better alignment: let $n = \log_2(d)$, making D equivalent to the network dimension d, and let I and $\mathbb{U}$ correspond to the time steps T and batch size B for PPO and DDPG. With these adjustments, we redefine the complexity equations and generate a more accurate runtime complexity plot as shown in Fig. 8 for a comparative analysis.

Fig. 8 demonstrates the comparative analysis of runtime complexity across varying network dimensions (i.e., $d = 10$ in Fig. 8a, $d = 100$ in Fig. 8b, and $d = 1000$ in Fig. 8c) for PPO, DDPG, and QDRL algorithms. PPO and DDPG have higher complexities compared to QDRL for smaller batch sizes due to the quadratic term Bd^2. As batch size increases, QDRL's complexity rises more gradually compared to PPO and DDPG, showing potential efficiency for larger batch sizes. For larger network dimensions ($d = 1000$), the relative performance difference becomes more pronounced, with QDRL maintaining lower complexity increases compared to PPO and DDPG.

VI. Conclusions

This work presented a quantum-aided DRL framework to enhance DoA estimation accuracy and computational task offloading latency in ISAC systems for military surveillance. By utilizing quantum computing's parallelism, it reduces the decision space dimensionality by encoding operational states and actions into quantum states, introducing a quantum-enhanced actor-critic method for policy optimization. Comparative analyses demonstrated significant outperformance, with faster convergence and a 76.32% and 48.73% improvement in normalized reward over DDPG and PPO, respectively. The quantum-DRL approach notably reduced RMSE in DoA estimation by over 94.10% compared to the random phase shift method, and by 91.66% and 82.61% against DDPG and PPO, respectively. Additionally, it minimized task offloading latency under URLLC requirements, achieving up to 43.09% latency reduction compared to DDPG and 32.35% against PPO, evidencing its efficacy.

References

[1] A. Aubry, A. D. Maio, and L. Pallotta, "Power-aperture resource allocation for a MPAR with communications capabilities," *IEEE Trans. Veh. Technol.*, pp. 1–14, 2024.

[2] D. S. Lakew, A.-T. Tran, N.-N. Dao, and S. Cho, "Intelligent self-optimization for task offloading in LEO-MEC-assisted energy-harvesting-UAV systems," *IEEE Trans. Netw. Sci. Eng.*, pp. 1–14, 2024.

[3] G. Geraci *et al.*, "What will the future of UAV cellular communications be? a flight from 5G to 6G," *IEEE Commun. Surv. Tutor.*, vol. 24, no. 3, pp. 1304–1335, 3rd Quart., 2022.

[4] D. Han *et al.*, "Two-timescale learning-based task offloading for remote IoT in integrated satellite–terrestrial networks," *IEEE Internet Things J.*, vol. 10, no. 12, pp. 10 131–10 145, Jun. 2023.

[5] Y. Xu, Y. Li, J. A. Zhang, M. Di Renzo, and T. Q. S. Quek, "Joint beamforming for RIS-assisted integrated sensing and communication systems," *IEEE Trans. Commun.*, pp. 1–1, 2023.

[6] Y. Pan, R. Li, X. Da, H. Hu, M. Zhang, D. Zhai, K. Cumanan, and O. A. Dobre, "Cooperative trajectory planning and resource allocation for UAV-enabled integrated sensing and communication systems," *IEEE Trans. Veh. Technol.*, pp. 1–16, 2023.

[7] S. Li, B. Duo, X. Yuan, Y.-C. Liang, and M. Di Renzo, "Reconfigurable intelligent surface assisted UAV communication: Joint trajectory design and passive beamforming," *IEEE Wirel. Commun. Lett.*, vol. 9, no. 5, pp. 716–720, May 2020.

[8] R. Liu, M. Li, H. Luo, Q. Liu, and A. L. Swindlehurst, "Integrated sensing and communication with reconfigurable intelligent surfaces: Opportunities, applications, and future directions," *IEEE Wirel. Commun.*, vol. 30, no. 1, pp. 50–57, Feb. 2023.

[9] A. Magbool *et al.*, "A survey on integrated sensing and communication with intelligent metasurfaces: Trends, challenges, and opportunities," Jan. 2024.

[10] A. M. Elbir, K. V. Mishra, M. R. B. Shankar, and S. Chatzinotas, "The rise of intelligent reflecting surfaces in integrated sensing and communications paradigms," *IEEE Netw.*, pp. 1–8, 2022.

[11] Z. Chen, P. Chen, Z. Guo, Y. Zhang, and X. Wang, "A RIS-based vehicle DOA estimation method with integrated sensing and communication system," *IEEE Trans. Intell. Transp. Syst.*, pp. 1–13, 2023.

[12] X. Wang, Z. Fei, J. Huang, and H. Yu, "Joint waveform and discrete phase shift design for RIS-assisted integrated sensing and communication system under cramer-rao bound constraint," *IEEE Trans. Veh. Technol.*, vol. 71, no. 1, pp. 1004–1009, Jan. 2022.

[13] Z. Fei, X. Wang, N. Wu, J. Huang, and J. A. Zhang, "Air-ground integrated sensing and communications: Opportunities and challenges," *IEEE Commun. Mag.*, vol. 61, no. 5, pp. 55–61, May 2023.

[14] Q. Liu, R. Luo, H. Liang, and Q. Liu, “Energy-efficient joint computation offloading and resource allocation strategy for ISAC-aided 6G V2X networks,” *IEEE Trans. Green Commun. Netw.*, vol. 7, no. 1, pp. 413–423, Mar. 2023.

[15] Q. Wu, J. Xu, Y. Zeng, D. W. K. Ng, N. Al-Dhahir, R. Schober, and A. L. Swindlehurst, “A comprehensive overview on 5G-and-beyond networks with UAVs: From communications to sensing and intelligence,” *IEEE J. Sel. Areas Commun.*, vol. 39, no. 10, pp. 2912–2945, Oct. 2021.

[16] A. Paul, K. Singh, M.-H. T. Nguyen, C. Pan, and C.-P. Li, “Digital twin-assisted space-air-ground integrated networks for vehicular edge computing,” *IEEE J. Sel. Top. Signal Process.*, pp. 1–16, 2023.

[17] Z. Wang and V. W. Wong, “Deep learning for isac-enabled end-to-end predictive beamforming in vehicular networks,” in *Proc. IEEE International Conference on Communications*, Oct. 2023, pp. 5713–5718.

[18] Q. Liu, Y. Zhu, M. Li, R. Liu, Y. Liu, and Z. Lu, “DRL-based secrecy rate optimization for RIS-assisted secure ISAC systems,” *IEEE Trans. Veh. Technol.*, vol. 72, no. 12, pp. 16 871–16 875, Dec. 2023.

[19] Y. Gong, Y. Wei, Z. Feng, F. R. Yu, and Y. Zhang, “Resource allocation for integrated sensing and communication in digital twin enabled internet of vehicles,” *IEEE Trans. Veh. Technol.*, vol. 72, no. 4, pp. 4510–4524, 2023.

[20] X. Liu, H. Zhang, K. Long, M. Zhou, Y. Li, and H. V. Poor, “Proximal policy optimization-based transmit beamforming and phase-shift design in an IRS-aided ISAC system for the THz band,” *IEEE J. Sel. Areas Commun.*, vol. 40, no. 7, pp. 2056–2069, Jul. 2022.

[21] M. Di Renzo, A. Zappone, M. Debbah, M.-S. Alouini, C. Yuen, J. de Rosny, and S. Tretyakov, “Smart radio environments empowered by reconfigurable intelligent surfaces: How it works, state of research, and the road ahead,” *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2450–2525, Nov. 2020.

[22] W. Chen, X. Qiu, T. Cai, H.-N. Dai, Z. Zheng, and Y. Zhang, “Deep reinforcement learning for internet of things: A comprehensive survey,” *IEEE Commun. Surv. Tutor.*, vol. 23, no. 3, pp. 1659–1692, 3rd Quart., 2021.

[23] R. Yan, Y. Wang, Y. Xu, and J. Dai, “A multiagent quantum deep reinforcement learning method for distributed frequency control of islanded microgrids,” *IEEE Trans. Control Netw. Syst.*, vol. 9, no. 4, pp. 1622–1632, Dec. 2022.

[24] Silvirianti, B. Narottama, and S. Y. Shin, “Layerwise quantum deep reinforcement learning for joint optimization of UAV trajectory and resource allocation,” *IEEE Internet Things J.*, vol. 11, no. 1, pp. 430–443, Jan. 2024.

[25] K. Wang, N. Qi, H. Liu, A.-A. A. Boulogeorgos, T. A. Tsiftsis, M. Xiao, and K.-K. Wong, “Reconfigurable intelligent surfaces aided energy efficiency maximization in cell-free networks,” *IEEE Wireless Commun. Lett.*, vol. 13, no. 6, pp. 1596–1600, 2024.

[26] J. Xie, W. Wang, X. Liu, I. Rashdan, C. Di, and J. Qin, “Identification of NLOS based on soft decision method,” *IEEE Wireless Commun. Lett.*, vol. 12, no. 4, pp. 703–707, Apr. 2023.

[27] H. A. Ammar, R. Adve, S. Shahbazpanahi, G. Boudreau, and K. V. Srinivas, “RWP+: A new random waypoint model for high-speed mobility,” *IEEE Commun. Lett.*, vol. 25, no. 11, pp. 3748–3752, Nov. 2021.

[28] M. Chen, Q. Li, L. Huang, L. Feng, and M. Rihan, “One-bit cramér–rao bound of direction of arrival estimation for deterministic signals,” *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 71, no. 2, pp. 957–961, Feb. 2024.

[29] A. Paul and S. P. Maity, “Outage analysis in cognitive radio networks with energy harvesting and Q-routing,” *IEEE Trans. Veh. Technol.*, vol. 69, no. 6, pp. 6755–6765, Jun. 2020.

[30] A. Paul, K. Singh, C.-P. Li, O. A. Dobre, and T. Q. Duong, “Digital twin-aided vehicular edge network: A large-scale model optimization by quantum-DRL,” *IEEE Trans. Veh. Technol.*, pp. 1–17, 2024.

[31] N. Yang, L. Han, R. Liu, Z. Wei, H. Liu, and C. Xiang, “Multiobjective intelligent energy management for hybrid electric vehicles based on multi-agent reinforcement learning,” *IEEE Trans. Transp. Electrification*, vol. 9, no. 3, pp. 4294–4305, Sept. 2023.

[32] F. Metz and M. Bukov, “Self-correcting quantum many-body control using reinforcement learning with tensor networks,” *Nat. Mach. Intell.*, vol. 5, no. 7, pp. 780–791, Jul. 2023.

[33] K. Miyamoto and H. Ueda, “Extracting a function encoded in amplitudes of a quantum state by tensor network and orthogonal function expansion,” *Quantum Information Processing*, vol. 22, no. 6, p. 239, Jun. 2023.

[34] M. Schuld and N. Killoran, “Quantum machine learning in feature hilbert spaces,” *Physical Review Letters*, vol. 122, no. 4, p. 040504, Feb. 2019.

[35] Z. Qu and H. Sun, “A secure information transmission protocol for healthcare cyber based on quantum image expansion and grover search algorithm,” *IEEE Trans. Netw. Sci. Eng.*, vol. 10, no. 5, pp. 2551–2563, Sept.-Oct. 2023.

[36] M. S. Rudolph *et al.*, “Synergistic pretraining of parametrized quantum circuits via tensor networks,” *Nat. Commun.*, vol. 14, no. 1, p. 8367, Dec. 2023.

[37] Z. Li, K. Xue, J. Li, L. Chen, R. Li, Z. Wang, N. Yu, D. S. L. Wei, Q. Sun, and J. Lu, “Entanglement-assisted quantum networks: Mechanics, enabling technologies, challenges, and research directions,” *IEEE Commun. Surv. Tutor.*, vol. 25, no. 4, pp. 2133–2189, 4th Quart., 2023.

[38] J. A. Ansere, E. Gyamfi, V. Sharma, H. Shin, O. A. Dobre, and T. Q. Duong, “Quantum deep reinforcement learning for dynamic resource allocation in mobile edge computing-based IoT systems,” *IEEE Trans. Wireless Commun.*, pp. 1–1, 2023.

[39] J. A. Ansere, D. T. Tran, O. A. Dobre, H. Shin, G. K. Karagiannidis, and T. Q. Duong, “Energy-efficient optimization for mobile edge computing with quantum machine learning,” *IEEE Wireless Commun. Lett.*, pp. 1–1, 2023.