Covert IRS-UAV Networks Empowered by Deep Reinforcement Learning

Esraa M. Ghourab, *Senior Member, IEEE*, Omar Alhussein, *Senior Member, IEEE*, De Mi, *Senior Member, IEEE*, Qiang Ye, *Senior Member, IEEE*, and Sami Muhaidat, *Senior Member, IEEE*

***Abstract*—Covert wireless communication ensures both information confidentiality and transmission untraceability, which is increasingly vital for mission-critical extended reality (XR) services. While uncrewed aerial vehicles (UAVs) provide mobility and flexible coverage, and intelligent reflecting surfaces (IRSs) enable energy-efficient signal manipulation, their joint use for covert communications has not yet been sufficiently explored. This paper proposes a novel UAV-mounted IRS system for covert communications that passively reflects source signals toward a legitimate receiver while minimizing detection by an adversary warden. In contrast to previous work that treats trajectory design, beamforming, and power control in isolation, the proposed work develops a unified framework based on double deep Q-networks (DDQN) to jointly optimize the UAV trajectory, power allocation, and IRS phase shifts under covert constraints. We analytically derive the optimal detection threshold and the minimum detection error probability, which are dynamically integrated into the learning framework. The optimization problem is formulated as a constrained Markov decision process, which allows the agent to adaptively learn optimal policies in dynamic environments without relying on perfect channel knowledge. Simulation results demonstrate that the proposed framework significantly improves covert rate and energy efficiency compared with the iterative and random benchmark schemes, while also providing insights into the impact of system parameters on performance.**

***Index Terms*—Covert communication, double deep Q-network, reinforcement learning, physical layer security.**

I. Introduction

WIRELESS communication links are inherently susceptible to information leakage due to their broadcast nature [1]. While traditional cryptographic techniques rely on secret key exchanges to ensure confidentiality, physical layer security (PLS) offers a lower-complexity alternative by exploiting the intrinsic randomness of wireless channels [2]. Specifically, PLS enhances security by ensuring that the channel quality between legitimate users exceeds that of the eavesdropper, thereby minimizing the potential for information leakage [3]. However, PLS mainly protects the content of transmissions but does not conceal its existence. As a result, adversaries may still detect or localize transmissions, which is highly problematic for mission-critical scenarios such as military operations and secure extended reality (XR) services [4].

To address this vulnerability, covert communication techniques have been proposed to conceal the existence of transmission while ensuring reliable reception at legitimate receivers [5], [6]. By making the received signal at the warden statistically indistinguishable from noise or interference, covert communication significantly reduces the risk of detection. Foundational work such as Bash et al. [7] introduced the "square root law," establishing fundamental covert capacity limits. This seminal result triggered extensive research into covert systems. For instance, the authors in [8] and [9] considered more practical scenarios involving uncertain warden locations and cooperative strategies using friendly jamming to confuse the warden. However, most existing solutions assume terrestrial infrastructures, often subject to channel blockages, mobility constraints, and limited coverage, thereby restricting the robustness of covert communications in dynamic environments.

Uncrewed aerial vehicles (UAVs) provide new opportunities to overcome these issues. With their high mobility, flexible deployment, and strong line-of-sight (LoS) links [10], [11], UAVs enhance reliability and covert performance compared to ground-based systems [12]. In wireless networks, UAVs can serve as relays [1] or operate as aerial base stations (BS) to extend the coverage of ground infrastructure [13]. Prior research has used UAVs as transmitters [14] or relays [11] to confuse wardens by exploiting spatial and temporal mobility [15]. Yet, UAV transmissions remain energy-constrained and detectable due to their strong LoS channels [16].

Recently, intelligent reflecting surfaces (IRSs) have emerged as a promising energy-efficient solution for controlling the radio propagation environment [17], [18]. An IRS is a software-controlled surface composed of a large number of low-cost passive reflective elements, each capable of dynamically adjusting the phase, amplitude, or polarization of incident electromagnetic waves. By intelligently configuring phase-shift matrices, IRSs can perform passive beamforming to

Received 30 April 2025; revised 13 September 2025; accepted 23 November 2025. Date of publication 10 December 2025; date of current version 18 February 2026. This work was supported in part by the P.R.C. Shenzhen Science and Technology Program under Grant GJHZ20220913143013024, in part by the P.R.C. Ministry of Education under Grant HZKY20220106202202120, in part by the ORAN-TWIN and ORAN-TWIN-X Projects of Communications Hub for Empowering Distributed Cloud Computing Applications and Research (CHEDDAR), funded by the UK EPSRC under Grant EP/Y037421/1 and Grant EP/X040518/1, and in part by the SIGHT-AI project funded by the Eureka Globalstars Innovate UK under Grant 10150077. *(Corresponding author: De Mi.)*

Esraa M. Ghourab and Omar Alhussein are with the KU 6G Research Centre, Department of Computer Science, Khalifa University, Calgary, United Arab Emirates (e-mail: essra.ghourab@ku.ac.ae; omar.alhussein@ku.ac.ae).

De Mi is with the College of Computing, Birmingham City University, B4 7XG Birmingham, U.K. (e-mail: de.mi@bcu.ac.uk).

Qiang Ye is with the Department of Electrical and Software Engineering, University of Calgary, Calgary, AB T2N 1N4, Canada (e-mail: qiang.ye@ucalgary.ca).

Sami Muhaidat is with the KU 6G Research Centre, Department of Computer and Information Engineering, Khalifa University, Calgary, United Arab Emirates (e-mail: sami.muhaidat@ku.ac.ae).

Digital Object Identifier 10.1109/JSAC.2025.3642218

enhance the quality of desired links and suppress interference without requiring active transmission power [19], [20]. These features make IRSs highly attractive for improving covert communication, where both power efficiency and signal concealment are critical. Most existing studies, however, focus on static ground-mounted IRSs that reflect UAV-transmitted signals, which do not alleviate the transmit power burden of UAVs. In contrast, integrating IRSs with UAV platforms opens new design possibilities for covert systems. Particularly, mounting an IRS on a UAV enables adaptive aerial repositioning and dynamic reconfiguration of the wireless environment, thereby overcoming LoS blockages [10], improving received signal power at legitimate users [21], as well as enhancing covertness by introducing spatial and phase-induced uncertainty at the warden [22].

Several recent studies have investigated IRS-UAV integration for secure and efficient communication. In [23], a fixed-position IRS-UAV was adopted to mitigate jamming and enhance legitimate transmission. In [24], the IRS-UAV altitude was optimized for cooperative relaying under fixed horizontal positioning. The work in [12] formulated a throughput maximization problem for a mobile IRS-UAV, jointly optimizing UAV trajectory, IRS beamforming, and power allocation via alternating optimization. In [16], the authors exploited phase-shift uncertainty in UAV-mounted IRS to confuse the warden under a binary hypothesis testing framework and optimized the transmit-to-jamming power ratio. A stochastic channel model with adaptive phase-shift control was proposed in [25] to enhance communication reliability, while a learning-based secure transmission scheme using twin-deep deterministic policy gradient (twin-DDPG) was presented in [26] to jointly optimize the UAV trajectory, beamforming, and IRS phase configuration under imperfect CSI conditions. In [1], the minimization of age of information (AoI) under covert constraints was studied using DQN- and DDPG-based approaches.

Despite promising results in UAV-IRS integration, e.g., enhancing throughput [12], mitigating jamming [23], or introducing uncertainty for covert transmission [16], three critical limitations remain in the literature. First, many works assume static or restricted UAV mobility and optimize only one dimension (e.g., UAV trajectory, IRS phase configuration, or transmit power) in isolation [1], [23], [24]. Second, most optimization frameworks rely on predefined models or idealized channel assumptions, limiting adaptability in dynamic wireless environments. Third, while some learning-based methods (e.g., DQN, DDPG) have been applied to secure transmissions [26], they do not explicitly incorporate covert communication constraints or enable joint optimization across UAV trajectory, power allocation, and IRS configuration. To the best of our knowledge, a unified framework that simultaneously addresses these aspects under covert constraints remains unexplored.

To fill this gap, we propose a double deep Q-network (DDQN)-based covert communication framework for UAV-mounted IRS systems. Our approach jointly optimizes UAV trajectory, transmit/jamming power, and IRS passive beamforming under stringent covert constraints. A key novelty lies in the analytical derivation and integration of the optimal detection threshold and minimum detection error probability at the warden into the reinforcement learning process. By formulating the problem as a constrained Markov decision process (CMDP), the proposed method dynamically adapts to time-varying environments and learns optimal covert transmission policies online without relying on idealized assumptions.

Extensive simulation results validate the effectiveness of the proposed method and provide new insights into covert system design, including: the impact of the covert constraint parameter ϵ on average covert rate, sensitivity analysis of reward-shaping penalties, trade-offs among UAV mobility, power budget constraints, and detectability, and performance evaluation across different IRS sizes. These findings highlight the robustness and practical advantages of the proposed framework for next-generation secure UAV-assisted wireless communication systems.

The main contributions of this work are summarized as follows:

- We propose a novel IRS-UAV-enabled cooperative covert communication system, in which the IRS is mounted on a UAV to passively reflect the source signal toward the destination. This design enables energy-efficient relaying without active forwarding. We analyze detection performance at the warden via binary hypothesis testing and derive the optimal detection threshold with the corresponding minimum detection error probability.
- We formulate a covert rate maximization problem that jointly optimizes transmit power, UAV trajectory, and IRS passive beamforming. To address the high-dimensional, dynamic, and non-convex nature of the problem, we develop a DDQN-based reinforcement learning framework, enabling real-time decision-making and intelligent spatial diversification.
- We conduct extensive simulations to validate the proposed approach, demonstrating significant improvements in covert communication performance. Our results quantify the impact of system parameters, such as IRS size, covert constraint level, and power budgets, on the average covert rate, revealing important trade-offs and guiding system design choices.

The remainder of this article is organized as follows: The system model and performance metrics are presented in Section II. Section III analyzes the performance metric to formulate the covert rate maximization problem. Section IV introduces the proposed DDQN-based approach. The numerical results are provided in Section V, and the conclusion is drawn in Section VI.

II. System Model

A. Network Model

As illustrated in Fig. 1, we consider a typical cooperative covert communication scenario. The system comprises a source node (Alice), a UAV-mounted intelligent reflecting surface (IRS) acting as a mobile relay, a legitimate XR user (Bob), and a friendly jammer assisting in disrupting the detection capabilities of an adversary (Willie). The UAV-mounted IRS passively reflects Alice's signal toward Bob, enabling energy-efficient relaying. All ground nodes (Alice, Bob, Willie, and the jammer) are equipped with a single

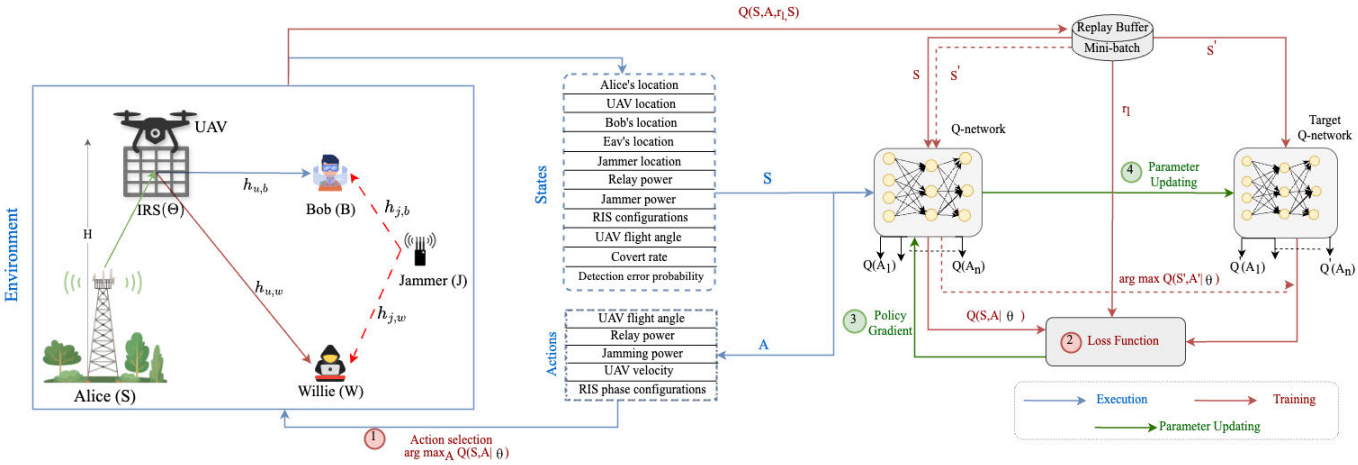

Fig. 1. UAV-IRS assisted covert communication system model.

transmit/receive antenna and operate in half-duplex mode, while the UAV carries a passive IRS to support cooperative transmission. Due to obstacles such as buildings and trees, Alice cannot transmit directly to Bob. Therefore, a UAV is deployed to assist in delivering the requested data to the legitimate XR receiver. To ensure safety, the UAV-IRS hovers at a sufficiently high altitude H to avoid collisions.

The IRS is modeled as a uniform planar array (UPA) consisting of M passive reflecting elements, with the first element as the reference point. The distances from Alice, Bob, and Willie to the reference element of the UAV-IRS are denoted as d_{su}, d_{ub}, and d_{uw}, respectively. The IRS adjusts only the phase shifts of its elements at each time slot, to satisfy system constraints, represented by a diagonal matrix

$$\mathbf{\Theta} = \text{diag}\{e^{j\theta_1}, e^{j\theta_2}, \cdots, e^{j\theta_M}\},$$

where $\theta_m \in [0, 2\pi)$ denotes the phase shift applied by the m^{th} reflector element, $m = 1, 2, \ldots, M$.

We assume that the UAV follows a predefined mission trajectory, traveling from an initial to a final location while assisting the source-destination (Alice-Bob) communication. The total flight duration T is discretized into N time slots with slot length $\Delta t = T/N$. Since Δt is sufficiently small, the UAV's position is assumed constant within each slot. A three-dimensional Cartesian coordinate system is adopted, where the UAV-IRS flies at a fixed altitude H, and all ground-based nodes (Alice, Bob, Willie, and jammer) are located on the ground with a Z-axis coordinate of zero (i.e., $z = 0$).

The horizontal 2D position of the UAV at the slot n is denoted by $\mathbf{q}_U[n] = (q_{x_u}[n], q_{y_u}[n]), \forall n \in \mathcal{N}, \mathcal{N} = \{1, 2, \ldots, N\}$. The ground nodes move much slower than the UAV; their locations are considered static with coordinates $\mathbf{q}_B = (q_{x_b}, q_{y_b})$, $\mathbf{q}_W = (q_{x_w}, q_{y_w})$, $\mathbf{q}_J = (q_{x_j}, q_{y_j})$, and $\mathbf{q}_S = (q_{x_s}, q_{y_s})$. Since the UAV's mission is executed between dock stations, the UAV's initial and final locations are predefined as $\mathbf{q}_U[0] = \mathbf{q}_I$ and $\mathbf{q}_U[N] = \mathbf{q}_F$, respectively. The UAV velocity is limited by a maximum speed $V_{\max}$, so the displacement per slot is constrained as

$$\|\mathbf{q}_U[n+1] - \mathbf{q}_U[n]\| \le V_{\max}\Delta t, \quad \forall n \in \mathcal{N} \setminus \{N\}, \tag{1}$$

where $\|\cdot\|$ denotes the Euclidean norm.

The source transmit power allocation $\{P_s[n]\}$ strongly affects covert communication performance and must satisfy both average and peak power constraints. Let P_{avg} and $P_{\max}$ denote the average and maximum transmit powers, respectively. The power constraints are

$$\frac{1}{N}\sum_{n=1}^{N} P_s[n] \le P_{\text{avg}}, \tag{2}$$

$$0 \le P_s[n] \le P_{\max}, \quad \forall n \in \mathcal{N}. \tag{3}$$

For non-triviality, we assume $P_{\text{avg}} < P_{\max}$. Equivalently, (2) can be expressed as follows.

$$\sum_{n=1}^{N} P_s[n] \le \bar{P}_s, \tag{4}$$

where $\bar{P}_s \triangleq NP_{\text{avg}}$ is the total transmit power budget over the whole UAV mission.

B. Channel Model

Due to obstacles such as buildings and trees, the direct communication link between Alice and Bob is obstructed in the covert communication network illustrated in Fig. 1. To enable covert transmission against Willie, a UAV-mounted IRS is deployed as a mobile relay between Alice and Bob. Owing to the UAV's elevated position, the Alice-UAV link is predominantly line-of-sight (LoS) with high probability [1]. The corresponding channel gain is given by:

$$h_{su}[n] = \sqrt{\beta_0 d_{su}^{-\rho}[n]} \left(1, e^{-j\frac{2\pi}{\lambda}d\phi_{su}[n]}, \cdots, e^{-j\frac{2\pi}{\lambda}(M-1)d\phi_{su}[n]}\right)^T, \tag{5}$$

where β_0 denotes the average path loss at a reference distance of 1 m, $\rho = 2$ is the path loss exponent, λ is the carrier wavelength, and d is the IRS element spacing (with $d \le \lambda/2$ to avoid spatial aliasing). The cosine of the angle of departure (AoD) from Alice to the UAV-IRS is

$$\phi_{su}[n] = \frac{q_{x_u}[n] - q_{x_s}}{d_{su}[n]},$$

and the distance between Alice and the UAV-IRS at slot n is

$$d_{su}[n] = \sqrt{\|\mathbf{q}_U[n] - \mathbf{q}_S\|^2 + H^2}.$$

Similarly, the UAV-IRS-Bob link is LoS-dominated, with a channel gain:

$$h_{ub}[n] = \sqrt{\beta_0 d_{ub}^{-2}[n]} \left(1, e^{-j\frac{2\pi}{\lambda}d\phi_{ub}[n]}, \cdots, e^{-j\frac{2\pi}{\lambda}(M-1)d\phi_{ub}[n]}\right)^T, \quad (6)$$

where

$$d_{ub}[n] = \sqrt{\|\mathbf{q}_U[n] - \mathbf{q}_B\|^2 + H^2}, \quad \phi_{ub}[n] = \frac{q_{x_u}[n] - q_{x_b}}{d_{ub}[n]}.$$

Willie is considered an adversary of both Alice and the UAV-IRS, and due to obstacles, no direct link exists between the ground BS and Willie. For the UAV-IRS-Willie channel, coherent phase alignment cannot be guaranteed. Therefore, UAV-IRS-Willie channel is modeled as Rayleigh fading with large-scale distance-based attenuation:

$$h_{uw}[n] = \sqrt{\beta_0 d_{uw}^{-2}[n]} \left(1, e^{-j\frac{2\pi}{\lambda}d\phi_{uw}[n]}, \cdots, e^{-j\frac{2\pi}{\lambda}(M-1)d\phi_{uw}[n]}\right)^T, \quad (7)$$

where

$$d_{uw}[n] = \sqrt{\|\mathbf{q}_U[n] - \mathbf{q}_W\|^2 + H^2}, \quad \phi_{uw}[n] = \frac{q_{x_u}[n] - q_{x_w}}{d_{uw}[n]}.$$

The channels from the friendly jammer to Bob and Willie are modeled as large-scale path loss with block Rayleigh fading as follows.

$$h_{jb}[n] = \sqrt{\beta_0 d_{jb}^{-\rho_j}[n]}\, g_{jb}, \quad (8)$$

$$h_{jw}[n] = \sqrt{\beta_0 d_{jw}^{-\rho_j}[n]}\, g_{jw}, \quad (9)$$

where $g_{jb}, g_{jw} \sim \mathcal{CN}(0,1)$ are independent fading coefficients, and ρ_j is the path loss exponent for jamming links.

The received signal at the UAV-mounted IRS at time slot n is given as follows.

$$y_u[n] = \sqrt{P_s[n]}\, h_{su}[n]\, x_s + n_u[n], \quad (10)$$

where x_s is Alice's information-bearing symbol with $\mathbb{E}[|x_s|^2] = 1$, where $\mathbb{E}\{.\}$ denotes the expectation operator. The term $h_{su}[n]$ represents the channel gain from Alice to the UAV-mounted IRS as defined in (5). Moreover, $P_s[n]$ is Alice's transmit power, and $n_u[n] \sim \mathcal{CN}(0, \sigma^2)$ is complex additive white Gaussian noise (AWGN) at the UAV.

After reflection by the UAV-mounted IRS, the received signal at Bob is expressed as follows.

$$y_b[n] = \sqrt{P_s[n]}\, h_{\text{eff},b}[n]\, x_s + \sqrt{\kappa P_j[n]}\, h_{jb}[n]\, x_j + n_b[n], \quad (11)$$

where $h_{\text{eff},b}[n] \triangleq h_{ub}^H[n]\mathbf{\Theta}[n]h_{su}[n]$ is the effective cascaded channel, with $h_{ub}[n]$ and $h_{jb}[n]$ are the channel gains from the UAV-mounted IRS to Bob and from the jammer to Bob, respectively, while $\mathbf{\Theta}[n]$ represents the passive beamforming matrix of the IRS at time slot n. x_j is the jammer's artificial noise (AN) symbol with $\mathbb{E}[|x_j|^2] = 1$, $n_b[n] \sim \mathcal{CN}(0, \sigma^2)$ is AWGN at Bob, $P_j[n]$ is jammer power, and $\kappa \in [0,1]$ represents residual AN leakage at Bob due to imperfect cancellation (with $\kappa = 0$ for perfect cancellation [18], [27], $\kappa = 1$ for no cancellation [16]).

Accordingly, the signal-to-interference-plus-noise ratio (SINR) for the legitimate link is given as follows.

$$\gamma_b[n] = \frac{P_s[n]|h_{\text{eff},b}[n]|^2}{\kappa P_j[n]|h_{jb}[n]|^2 + \sigma^2}. \quad (12)$$

The instantaneous achievable transmission rate R_{sb} in (bits/s/Hz) is written as follows [12].

$$R_{sb}[n] = \log_2(1 + \gamma_b[n]). \quad (13)$$

To ensure reliable communication, a minimum rate constraint R_{Th} is imposed. Therefore, Bob successfully decodes data only if

$$R_{sb}[n] \geq R_{\text{Th}}, \ \forall n. \quad (14)$$

C. System Hypothesis Test

Willie attempts to detect covert transmissions by performing a binary hypothesis test on his received signals. Specifically, the null hypothesis (H_0) happens when Alice is silent and the UAV-IRS is inactive; only jamming and noise are present, while the alternative hypothesis (H_1) indicates that Alice transmits and the UAV-IRS forwards the covert signal to Bob, with the presence of a jamming signal. The presence of the friendly jammer makes detection at Willie more difficult. We adopt the infinite blocklength assumption, i.e., Willie can average over infinitely many samples per time slot [16], which yields a worst-case detection capability. This worst-case scenario strengthens the rigor of the covert communication analysis, as it enables Willie to perform highly accurate detection by averaging over unlimited received samples.

The received signal at Willie for the i-th observation in slot n is written as follows.

$$y_w^{(i)}[n] = \begin{cases} \sqrt{P_s[n]}h_{\text{eff,w}}[n]x_s + \sqrt{P_j[n]}h_{jw}[n]x_j + n_w^{(i)}[n], & H_1 \\ \sqrt{P_j[n]}h_{jw}[n]x_j + n_w^{(i)}[n], & H_0 \end{cases}$$

where x_s and x_j are unit-power signals from Alice and the jammer, respectively; $h_{\text{eff},w}[n] \triangleq h_{uw}^H[n]\mathbf{\Theta}[n]h_{su}[n]$ is the effective cascaded Alice-UAV-Willie channel; $h_{jw}[n]$ is the jammer-Willie channel; and $n_w^{(i)}[n] \sim \mathcal{CN}(0, \sigma_w^2)$ is AWGN.

1) Energy Detector (Infinite Averaging): Based on probability theory and the statistical properties of the received signals, Willie can determine an optimal detection threshold that minimizes the detection error probability (DEP) [28]. With $K \to \infty$ observations, Willie's test statistic is the time-averaged received power.

$$T(y_w)[n] = \lim_{K\to\infty} \frac{1}{K} \sum_{i=1}^{K} \left|y_w^{(i)}[n]\right|^2 \underset{D_0}{\overset{D_1}{\gtrless}} P_{\text{th}}[n], \quad (15)$$

where $T(y_w)[n]$ denotes the average received signal power and $P_{\text{th}}[n]$ is Willie's decision threshold. If $T[y_w][n] > P_{\text{th}}[n]$, Willie confirms that Alice is transmitting D_1. Otherwise, Willie considers that Alice is being silent D_0.

The probabilities of a false alarm (P_{FA}) and missed detection (P_{MD}) are defined as follows.

$$P_{\text{FA}} = \Pr(D_1|H_0), \quad P_{\text{MD}} = \Pr(D_0|H_1).$$

The total detection error probability is $\zeta = P_{\text{FA}} + P_{\text{MD}}$. Willie seeks to minimize ζ by choosing P_{th}, while Alice, the UAV-IRS, and Bob aim to ensure $\zeta \geq 1 - \epsilon$ for some small $\epsilon > 0$, which guarantees covertness [29].

Therefore, the received power, $T(y_w)[n]$, can be written using independence and the law of large numbers,

$$T(y_w)[n] = \begin{cases} P_s[n]|h_{\text{eff,w}}[n]|^2 + P_j[n]|h_{jw}[n]|^2 + \sigma_w^2, & H_1 \\ P_j[n]|h_{jw}[n]|^2 + \sigma_w^2, & H_0 \end{cases} \tag{16}$$

In addition to channel noise, uncertainties arise from the random phase shifting of the UAV-mounted IRS and the jamming signal generated by the friendly jammer. Then, the probability distribution of $T(y_w)$ is analyzed in *Lemma* 1 under the two cases of H_0 and H_1.

Lemma 1: The CDF and PDF of the centered statistic can be derived, respectively, as

$$F_Z(z) = \begin{cases} 1 + \dfrac{\lambda_j}{\lambda_u - \lambda_j} e^{-\lambda_u z} - \dfrac{\lambda_u}{\lambda_u - \lambda_j} e^{-\lambda_j z}, & H_1, \\ 1 - e^{-\lambda_j z}, & H_0, \end{cases} \tag{17}$$

and

$$f_Z(z) = \begin{cases} \dfrac{\lambda_u \lambda_j}{\lambda_u - \lambda_j} \left(e^{-\lambda_j z} - e^{-\lambda_u z}\right), & H_1, \\ \lambda_j e^{-\lambda_j z}, & H_0, \end{cases} \quad z \geq 0, \tag{18}$$

where $Z \triangleq T(y_w)[n] - \sigma_w^2$

Proof: from (16), we define $A \triangleq P_s[n]|h_{\text{eff},w}[n]|^2$ and $B \triangleq P_j[n]|h_{jw}[n]|^2$. It can be seen that under the IRS phase-randomization assumption and Rayleigh fading, the random variables A, and B are independently exponential random variables with means

$$\mathbb{E}[A] = P_s[n]\, G_u[n], \qquad \mathbb{E}[B] = P_j[n]\, G_j[n], \tag{19}$$

with

$$G_u[n] = \eta_u[n]\, \beta_0^2\, d_{uw}^{-2}[n]\, d_{su}^{-2}[n], \quad G_j[n] = \beta_0\, d_{jw}^{-\rho_j}[n]. \tag{20}$$

where $\eta_u[n]$, satisfying $0 \leq \eta_u[n] \leq M^2$, denotes the IRS combining gain toward Willie, with the upper bound M^2 achieved under perfectly coherent phasing. Accordingly, the pdf of A and B are respectively given as

$$f_A(a) = \lambda_u e^{-\lambda_u a}, \quad a \geq 0, \tag{21}$$

$$f_B(b) = \lambda_j e^{-\lambda_j b}, \quad b \geq 0. \tag{22}$$

where, $\lambda_u = 1/P_s[n]G_u[n]$ and $\lambda_j = 1/P_j[n]G_j[n]$. Given $Z = A + B$ and $z \geq 0$ with $\lambda_u \neq \lambda_j$, the centered statistic Z is written as follows:

$$Z = \begin{cases} A + B, & H_1, \\ B, & H_0, \end{cases} \tag{23}$$

With $Z \triangleq T(y_w)[n] - \sigma_w^2$, the CDF of Z under H_1 can be derived as

$$\begin{aligned} F_{Z|H_1}(z) &= \Pr(Z \leq z) = \int_0^z \int_0^{z-a} \lambda_u e^{-\lambda_u a}\, \lambda_j e^{-\lambda_j b}\, db\, da \\ &= \left(1 - e^{-\lambda_u z}\right) - \lambda_u e^{-\lambda_j z} \int_0^z e^{-(\lambda_u - \lambda_j)a} da \\ &= 1 + \frac{\lambda_j}{\lambda_u - \lambda_j} e^{-\lambda_u z} - \frac{\lambda_u}{\lambda_u - \lambda_j} e^{-\lambda_j z}. \end{aligned} \tag{24}$$

Similarly, the CDF of Z under H_0 is given as

$$F_{Z|H_0}(z) = 1 - e^{-\lambda_j z}. \tag{25}$$

Therefore, the CDF of Z under the both hypotheses is given as (17). By taking the first-order derivative of (24) and (25), the PDFs of the random variable Z under H_0 and H_1 can be written as in (18).

Lemma 1 is proved.

These distributions are used to derive P_{FA}, P_{MD}, and the minimum detection error probability $\zeta^\star[n]$ in the following subsection.

III. Performance Constraints

A. Covert Constraint

Willie may mis-detect covert transmissions due to channel randomness, IRS phase uncertainty, and jamming interference. Two detection errors are defined as, P_{FA} is the probability that Alice is silent, but Willie decides that Alice is transmitting (D_1) according to the received power. While P_{MD} is the probability that Alice is transmitting, Willie decides that Alice is silent (D_0). Therefore, the total detection error probability is concluded as the probability summation of both probabilities, i.e., $\zeta[n] \triangleq P_{FA}[n] + P_{MD}[n]$.

Let $x \triangleq P_{\text{th}}[n] - \sigma_w^2$. Willie's decision rule becomes

$$\delta(Z) = \begin{cases} D_1, & Z \geq x, \\ D_0, & Z < x. \end{cases} \tag{26}$$

Therefore, the error probabilities of FA and MD are formulated, respectively, as follows

$$P_{FA}[n] = \Pr(D_1|H_0) = \Pr(Z \geq x|H_0), \tag{27}$$

$$P_{MD}[n] = \Pr(D_0|H_1) = \Pr(Z < x|H_1). \tag{28}$$

Because $Z \geq 0$, if $x \leq 0$, then $P_{FA}[n] = 1$ and $P_{MD}[n] = 0$, giving $\zeta[n] = 1$. The nontrivial regime is $x > 0$.

Using Lemma 1, the total detection error probability, $\zeta[n]$, can be written as

$$\zeta[n] = 1 + \frac{\lambda_j}{\lambda_u - \lambda_j} e^{-\lambda_u x} - \frac{\lambda_u}{\lambda_u - \lambda_j} e^{-\lambda_j x} + e^{-\lambda_j x}, \tag{29}$$

where $x = P_{\text{th}}[n] - \sigma_w^2 > 0$.

Proposition 1: The optimal detection threshold that minimizes ζ at Willie can be given as

$$P_{\text{th}}^*[n] = \frac{\ln \lambda_u - \ln \lambda_j}{\lambda_u - \lambda_j} + \sigma_w^2, \tag{30}$$

and the minimum detection error probability is

$$\zeta^\star = 1 - \left(\frac{\lambda_j}{\lambda_u}\right)^{\lambda_u/(\lambda_u - \lambda_j)}. \tag{31}$$

Proof: In order to minimize the error probability ζ, Willie needs to set the detection threshold P_{th} carefully. Therefore,

to derive the optimal P_{th}, we analyze the monotonicity of ζ. By taking the first-order derivative of (29), we have

$$\zeta'[n] = \frac{\lambda_u \lambda_j}{\lambda_u - \lambda_j}\left(e^{-\lambda_j x} - e^{-\lambda_u x}\right) - \lambda_j e^{-\lambda_j x} \tag{32}$$

After some mathematical simplification, (32) can be rewritten as

$$\zeta'[n] = \frac{\lambda_j}{\lambda_u - \lambda_j}\left(\lambda_j e^{-\lambda_j x} - \lambda_u e^{-\lambda_u x}\right). \tag{33}$$

In order to find the optimal detection threshold, we solve $\zeta'[n] = 0$, resulting in $\lambda_j e^{-\lambda_j x} = \lambda_u e^{-\lambda_u x}$. Thus, the optimum detection threshold can be given as

$$x^* = \frac{\ln \lambda_u - \ln \lambda_j}{\lambda_u - \lambda_j}. \tag{34}$$

Noting that $P^*_{\text{th}}[n] = x^* + \sigma^2$, then $P^*_{\text{th}}[n]$ can be expressed as in (30).

By substituting $x = x^*$ into $\zeta[n]$, and after some mathematical manipulation, the optimal error detection probability ζ^* can be described as in (31), which completes the proof.

Remark 1: It is important to highlight that we seek to find the optimal judgement threshold that minimizes $\zeta[n]$.

Case 1: When $P_{\text{th}}[n] \le \sigma_w^2$, then $x \le 0$ and $\zeta[n] = 1$, which means that Willie always decides D_1, no optimal $\zeta[n]$.

Case 2: The worst case for the legitimate system is when Willie chooses $P^*_{\text{th}}[n]$, attaining $\zeta^\star$. To guarantee covertness, the system enforces

$$\zeta[n] \ \ge \ 1 - \epsilon, \ \forall n, \tag{35}$$

for some small $\epsilon \in (0, 1)$ [29]. A conservative formulation requires $\zeta^\star \ge 1 - \epsilon$, which holds even if Willie optimally adjusts his threshold.

Remark 2: By examining λ_u and λ_j, all results above are explicit functions of powers, distances, IRS size M, and path-loss exponents. Increasing M or moving the UAV to improve $G_u[n]$ raises the ratio λ_j/λ_u, which *increases* $\zeta^\star$ (improves covertness).

B. Covert Capacity

To evaluate the average covert capacity (C_c), we first assess the probability that Alice's transmission via the UAV-mounted IRS avoids outage. An outage occurs when the achievable rate at Bob falls below the target threshold R_{Th}, i.e., $R_{sb}[n] < R_{\text{Th}}$, in which case Bob cannot decode reliably. Therefore, reliable communication is ensured if $R_{sb}[n] \ge R_{\text{Th}}$. The non-outage probability is written as follows [29].

$$\delta_{srd}[n] = \Pr\big(\log_2(1 + \gamma_b[n]) \ge R_{\text{Th}}\big) = \Pr(\gamma_b[n] \ge \Gamma_{\text{Th}}), \tag{36}$$

where $\Gamma_{\text{Th}} = 2^{R_{\text{Th}}} - 1$ is the reliability SINR threshold. The *average covert capacity* over the UAV mission of N slots is then defined as [29].

$$C_c = \frac{1}{N}\sum_{n\in\mathcal{N}} R_{sb}[n]\, \delta_{srd}[n], \tag{37}$$

which quantifies the average rate at Bob while satisfying covertness against Willie.

Worth mentioning that a *positive average covert capacity* $C_c > 0$ requires (1) a non-zero probability of non-outage at Bob and (2) feasible covertness against Willie. We derive simple, checkable per-slot conditions that guarantee both simultaneously.

1) Bob-Side Reliability Bound: With $\gamma_b[n] = P_s[n]|h_{\text{eff,b}}[n]|^2/\kappa P_j[n]|h_{jb}[n]|^2 + \sigma^2$, the non-outage event $\gamma_b[n] \ge \Gamma_{\text{Th}}$ is equivalent to $P_s[n]|h_{\text{eff,b}}[n]|^2 \ge \Gamma_{\text{Th}}\big(\kappa P_j[n]|h_{jb}[n]|^2 + \sigma^2\big)$. Solving for $P_j[n]$ gives the *Bob-side admissible upper bound* as follows.

$$P_j[n] \ \le \ U[n] \ \triangleq \ \frac{P_s[n]|h_{\text{eff,b}}[n]|^2 - \Gamma_{\text{Th}}\sigma^2}{\Gamma_{\text{Th}}\kappa|h_{jb}[n]|^2}. \tag{38}$$

This requires $P_s[n]|h_{\text{eff,b}}[n]|^2 > \Gamma_{\text{Th}}\sigma^2$; otherwise $U[n] \le 0$ and Bob cannot avoid outage.

2) Feasible Covertness Against Willie: Willie-side covertness lower bound on P_j is proposed in *Lemma* 2.

Lemma 2: Considering the lower bound ε on Willie's detection error probability, the covertness lower bound on P_j at Willie's side is given by

$$P_j[n] \ \ge \ c(\varepsilon)\,\frac{G_u[n]}{G_j[n]}\,P_s[n] \triangleq \ L[n], \tag{39}$$

where $c(\varepsilon) = 1 + \frac{1}{\varepsilon}$.

Proof: From the minimum-error energy detector $\zeta^*[n]$, we define the power ratio as

$$r[n] \triangleq \frac{\lambda_u}{\lambda_j} = \frac{\mathbb{E}[B]}{\mathbb{E}[A]} = \frac{P_j[n]\,G_j[n]}{P_s[n]\,G_u[n]} > 0. \tag{40}$$

By substituting (40) in (31), Willie's minimum detection error probability under the optimal threshold can be rewritten as follows:

$$\zeta^\star[n] \ = \ 1 - \left(\frac{\lambda_j}{\lambda_u}\right)^{\lambda_u/(\lambda_u-\lambda_j)} = 1 - \left(\frac{1}{r[n]}\right)^{r[n]/(r[n]-1)} \tag{41}$$

Imposing the covert requirement $\zeta^\star[n] \ge 1 - \varepsilon$ yields the inequality,

$$r[n]^{-\,r[n]/(r[n]-1)} \ \le \ \varepsilon, \ \varepsilon \in (0, 1). \tag{42}$$

which can be altered to

$$\frac{r[n]}{r[n]-1}\ln\frac{1}{r[n]} \ \le \ \ln\varepsilon, \ \varepsilon \in (0, 1) \tag{43}$$

Owing to $\varepsilon \in (0, 1)$, we have $\ln\varepsilon < 0$. Accordingly, for $r[n] > 1$ (i.e., the jammer-dominant regime), (43) can be solved to yield

$$r[n] \ \ge \ \frac{\ln\varepsilon}{W_0\big((\ln\varepsilon)\,\varepsilon\big)} \ \triangleq \ c_\star(\varepsilon), \quad r[n] > 1, \tag{44}$$

where $W_0(\cdot)$ denotes the principal branch of the Lambert-W function. Therefore, if the system operates with $r[n] > 1$ (jammer-dominant regime) or if $\varepsilon \in (0, e^{-1}]$ [30], [31], then the inequality $r[n]^{-\,r[n]/(r[n]-1)} \le \varepsilon$ holds, implying $r[n] \ge c_\star(\varepsilon)$. Equivalently, the covertness *exact* lower bound on $P_j[n]$ is

$$P_j[n] \ \ge \ L_\star[n] \ \triangleq \ c_\star(\varepsilon)\,\frac{G_u[n]}{G_j[n]}\,P_s[n]. \tag{45}$$

To obtain a closed-form, easy-to-check sufficient condition valid for all $\varepsilon \in (0,1)$, we upper-bound the exact requirement by enforcing

$$r[n] \;\geq\; c(\varepsilon) \;\triangleq\; 1+\frac{1}{\varepsilon} \tag{46}$$

holds for $r[n]^{-\frac{r[n]}{r[n]-1}} \leq \varepsilon$. Consequently,

$$P_j[n] \;\geq\; L[n] \;\triangleq\; \left(1+\tfrac{1}{\varepsilon}\right)\frac{G_u[n]}{G_j[n]}\,P_s[n] \tag{47}$$

is an easy-to-check sufficient condition. Moreover, as ε tends to 0, $c(\varepsilon)$ is asymptotically tight with respect to $c_\star(\varepsilon)$.

Solving $\zeta^\star[n] \geq 1-\varepsilon$ for $r[n]$ yields

$$r[n] \;\geq\; \left(1+\frac{1}{\varepsilon}\right) \triangleq\; c(\varepsilon), \tag{48}$$

Therefore, enforcing $P_j[n] \geq L[n]$ in (39) guarantees $\zeta^\star[n] \geq 1-\varepsilon$ for all $\varepsilon \in (0,1)$, while (45) provides the exact threshold in the jammer-dominant regime.

Lemma 2 is proved.

3) Feasible Jammer Band: By combining the reliability upper bound in (38) with the covertness lower bound in (39), we obtain a feasible jammer power interval

$$P_j[n] \;\in\; [\,L[n],\, U[n]\,],$$

which exists if and only if the following conditions are satisfied:

$$P_s[n]|h_{\text{eff,b}}[n]|^2 \;>\; \Gamma_{\text{Th}}\sigma^2, \qquad L[n] \;\leq\; U[n]. \tag{49}$$

Remark 3: The first condition in (49) ensures that Bob's received SNR exceeds the threshold Γ_{Th}, thereby guaranteeing $U[n] > 0$. The second condition guarantees that the feasible set for $P_j[n]$ is nonempty, i.e., there exists at least one jammer power level that is simultaneously large enough to ensure covertness ($P_j[n] \geq L[n]$) and small enough to preserve Bob's reliability ($P_j[n] \leq U[n]$). If condition (49) holds with nonzero probability across slots, then (i) Bob achieves non-outage with positive probability, and (ii) the covertness constraint is simultaneously satisfied. In this case, the covert capacity defined in (37) satisfies $C_c > 0$. Conversely, if condition (49) fails for all slots, then no feasible jammer power exists, and hence $C_c = 0$.

C. Covert Rate Maximization

The objective is to maximize the average covert communication rate while minimizing Willie's detection capability. This requires joint optimization of the source power allocation $\boldsymbol{P_s} = \{P_s[n], n \in \mathcal{N}\}$, the jammer power allocation $\boldsymbol{P_j} = \{P_j[n], n \in \mathcal{N}\}$, the IRS passive beamforming phases $\boldsymbol{\Phi} = \{\boldsymbol{\theta}[n], n \in \mathcal{N}\}$, and the UAV trajectory $\boldsymbol{q} = \{\mathbf{q}_U[n], n \in \mathcal{N}\}$. The optimization problem is formulated as follows.

$$\underset{\boldsymbol{q},\boldsymbol{\Theta},\boldsymbol{P_s},\boldsymbol{P_j}}{\text{maximize}} \quad C_c = \frac{1}{N}\sum_{n\in\mathcal{N}} R_{sb}[n]\delta_{srd}[n] \tag{P1}$$

$$\text{subject to} \quad \mathbf{q}_U[1] = \mathbf{q}_I, \mathbf{q}_U[N] = \mathbf{q}_F, \tag{P1.a}$$

$$\|\mathbf{q}_U[n+1] - \mathbf{q}_U[n]\| \leq V_{\max}\Delta t, \quad \forall n \tag{P1.b}$$

$$\mathbf{q}_U[n] \in \mathcal{L}, \quad \forall n \tag{P1.c}$$

$$\tfrac{1}{N}\sum_{n=1}^{N} P_s[n] \leq P_{\text{avg}} \tag{P1.d}$$

$$P_s[n]|h_{\text{eff,b}}[n]|^2 > \Gamma_{\text{Th}}\sigma^2, \quad \forall n \tag{P1.e}$$

$$0 \leq P_s[n] \leq P_{\max}, \quad \forall n \tag{P1.f}$$

$$\max\{0, L[n]\} \leq P_j[n] \leq \min\{U[n], P_{\max}\}, \quad \forall n \tag{P1.g}$$

$$\zeta^\star[n] \geq 1-\epsilon,\; 0<\epsilon<1 \tag{P1.h}$$

$$R_{sb}[n] \geq R_{\text{Th}}, \quad \forall n \tag{P1.i}$$

$$0 < \theta_i[n] < 2\pi, \quad \forall i,n \tag{P1.j}$$

Constraints (P1.a)-(P1.b) enforce UAV mobility feasibility: the UAV must begin and end at designated dock stations and respect its maximum speed $V_{\max}$. Constraint (P1.c) restricts the UAV trajectory to the mission region $\mathcal{L}$. Constraints (P1.d)-(P1.g) regulate power allocations. Specifically, (P1.d) enforces the average source power budget, (P1.e) ensures Bob can avoid outage, (P1.f) sets per-slot peak source power, and (P1.g) enforces the feasible jammer power band derived from the reliability and covertness conditions. Constraint (P1.h) imposes the covert requirement that Willie's minimum detection error probability $\zeta^\star$ remains above $1-\epsilon$. Constraint (P1.i) guarantees Bob's rate meets the target threshold R_{Th}, while (P1.j) ensures feasible IRS phase shifts.

Problem (P1) is highly non-convex due to: (i) multiplicative coupling between IRS phases $\boldsymbol{\Theta}$ and cascaded channels in $R_{sb}[n]$; (ii) non-convex dependence of channel gains $G_u[n], G_j[n]$ on UAV trajectory $\boldsymbol{q}$; (iii) the nonlinear covert constraint $\zeta^\star(\lambda_u,\lambda_j) \geq 1-\epsilon$. To address this, we adopt a constrained Markov decision process (CMDP) framework [32], solved via deep reinforcement learning (DRL). DRL is well-suited because it (i) scales to high-dimensional decision spaces (trajectory, IRS phases, power allocations), (ii) adapts to time-varying wireless environments, and (iii) can incorporate covert constraints through either hard penalties or primal-dual multipliers [33]. In particular, double deep Q-networks (DDQN) with Lagrangian penalty updates provide efficient approximations to near-optimal solutions in this covert communication setting [32], [33].

IV. DRL-Based Optimization Algorithm

A. Constrained Markov Decision Process Modeling

A Markov decision process (MDP) provides a mathematical framework for modeling discrete-time stochastic control problems, where system evolution is determined by both random factors and an agent's actions. It assumes a fully observable environment and forms the foundation for reinforcement learning (RL). To enable the UAV to jointly optimize its trajectory, IRS configuration, and power allocation under the complex environment described in (P1), the problem is reformulated as a finite discrete-time constrained MDP (CMDP).

Formally, the CMDP is represented as the tuple

$$\mathcal{M} := (\boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{S}}, \pi),$$

where: $\boldsymbol{\mathcal{S}}$ is the state space, describing the environment at each decision step; $\boldsymbol{\mathcal{S}}$ is the action space, containing feasible agent actions; $\boldsymbol{\mathcal{S}} : \boldsymbol{\mathcal{S}} \times \boldsymbol{\mathcal{S}} \to \mathbb{R}$ is the reward function, quantifying

the immediate benefit of an action in a given state; and π is the policy mapping states to actions. At each iteration k and time slot n, the state, action, and reward are defined as follows.

1) State Space $s^{(k)}[n]$: The state space includes essential information required for decision-making. At time slot n, the state is defined as follows.

$$s[n]=\Big[\boldsymbol{q}_U[n],\, P_s[n],\, P_j[n],\, \boldsymbol{\Theta}[n-1],\, \zeta^\star[n-1],\, C_c[n-1]\Big]^\top, \tag{50}$$

where $\boldsymbol{q}_U[n]$ is the UAV position at slot n within a 200×200 m horizontal plane at a fixed altitude of 50 m; $P_s[n]$ and $P_j[n]$ denote the BS and jammer transmit powers; $\boldsymbol{\Theta}[n-1]$ denote the IRS phase *codeword* selected in the previous slot (i.e., as formulated in (53)); $\zeta^\star[n-1]$ is the previous minimum detection error probability at Willie; and $C_c[n-1]$ is the previous average covert capacity.

2) Action Space $a^{(k)}[n]$: At each slot n, the UAV agent selects:

$$a[n]=\Big[\beta[n],\, v[n],\, \Delta P_s[n],\, \Delta P_j[n],\, \boldsymbol{\Theta}[n]\Big]^\top, \tag{51}$$

where $\beta[n]\in[0,2\pi]$ is the UAV's flight angle (discretized into 10 intervals); $v[n]\in[0,V_{\max}]$ is the UAV speed, discretized with step size 1 m/s; and $\Delta P_s[n]$ and $\Delta P_j[n]$ are incremental adjustments to the BS transmission power and jammer power, respectively, chosen from a small discrete set (e.g., $\{\pm1,0\}$). The absolute powers are then updated with feasibility clipping:

$$\begin{aligned} P_s[n] &\leftarrow \min\{\max\{P_s[n-1]+\Delta P_s[n],\,0\},\,P_{\max}\},\\ P_j[n] &\leftarrow \min\{\max\{P_j[n-1]+\Delta P_j[n],\,L[n]\},\,U[n]\}. \end{aligned} \tag{52}$$

This ensures that the CMDP respects the per-slot feasibility band of $(P_s[n],P_j[n])$ defined in problem (P1). where $L[n]$ and $U[n]$ are the covertness lower bound and reliability upper bound derived in Sec. III-B. Finally, the IRS phase configuration at time slot n is selected from a precomputed 2^{bit}-quantized *codebook* $\mathcal{C}=\{\boldsymbol{\theta}^{(1)},\ldots,\boldsymbol{\theta}^{(K)}\}$. Accordingly,

$$\boldsymbol{\Theta}[n]=\boldsymbol{\theta}^{(c[n])},\;\; c[n]\in\{1,\ldots,K\}, \tag{53}$$

where each codeword $\boldsymbol{\theta}^{(k)}\in\left\{0,\frac{2\pi}{2^{\text{bit}}},\ldots,\frac{2\pi(2^{\text{bit}}-1)}{2^{\text{bit}}}\right\}^M$ respects the hardware quantization limitation into 2^{bit} discrete phase levels within $[0,2\pi)$ [34].

3) Reward Function $R^{(k)}[n]$: The reward is designed to maximize the average covert communication rate while penalizing constraint violations. In the proposed model, the reward is formulated as follows.

$$R(s[n],a[n])=\tfrac{1}{N}\sum_{n\in\mathcal{N}}R_{sb}[n]\delta_{srd}[n]+h_{\text{penalty}}, \tag{54}$$

where $R_{sb}[n]\delta_{srd}[n]$ is the covert throughput at slot n, and h_{penalty} comprises penalty terms addressing system constraint violations, including

- **Trajectory violation:** $-\Lambda_p\|\boldsymbol{q}_U[n]-\mathcal{L}\|^2$ penalizes UAV deviation from the mission area $\mathcal{L}$.
- **Concealment violation:** $-\Lambda_C|\zeta^\star[n]-(1-\epsilon)|$ penalizes deviation from the covert constraint.
- **QoS violation:** $-\Lambda_b|R_{sb}[n]-R_{\text{Th}}|$ penalizes insufficient rate for Bob.

Therefore, the detailed reward at time slot n becomes:

$$\begin{aligned} R(s[n],a[n]) = \tfrac{1}{N}\sum_{n\in\mathcal{N}}R_{sb}[n]\delta_{srd}[n] &- \Lambda_p\|\boldsymbol{q}_U[n]-\mathcal{L}\|^2\\ &- \Lambda_C\big|\zeta^\star[n]-(1-\epsilon)\big| - \Lambda_b\big|R_{sb}[n]-R_{\text{Th}}\big|. \end{aligned} \tag{55}$$

The key challenge in modeling covert communication as a CMDP lies in ensuring that all constraints (trajectory feasibility, covertness, and QoS) are respected *throughout training*, rather than only in expectation. To address this, the proposed framework incorporates an adaptive penalty mechanism inspired by primal-dual optimization [35]. After each episode:

$$\Lambda_C \leftarrow \big[\Lambda_C+\eta_C\big((1-\epsilon)-\zeta^\star[n]\big)\big]^+, \tag{56}$$

$$\Lambda_b \leftarrow \big[\Lambda_b+\eta_b\big(R_{\text{Th}}-R_{sb}[n]\big)\big]^+, \tag{57}$$

$$\Lambda_p \leftarrow \big[\Lambda_p+\eta_p\big(\|\boldsymbol{q}_U[n]-\mathcal{L}\|-d_{\max}\big)\big]^+, \tag{58}$$

where $\eta_C,\eta_b,\eta_p>0$ are step sizes and $[x]^+=\max(0,x)$.

Algorithm 1 CMDP Training With Adaptive Penalties for Covert Communication

Input: Learning rates η_C,η_b,η_p, tolerance ϵ, max episodes K, horizon N
Output: Optimized policy $\pi^\star$

Initialize policy π, penalty weights $(\Lambda_C,\Lambda_b,\Lambda_p)\leftarrow(0,0,0)$
for *episode* $k=1,\ldots,K$ **do**
  Initialize environment (UAV position, IRS phases, powers)
  for *time slot* $n=1,\ldots,N$ **do**
    Observe state $s[n]$
    Select action $a[n]\sim\pi(s[n])$
    Execute $a[n]$, update powers $P_s[n]$, $P_j[n]$ (with clipping to feasible ranges)
    Compute covert throughput $R_{sb}[n]\delta_{srd}[n]$
    Evaluate constraints:
    (i) covertness: $\zeta^\star[n]\geq1-\epsilon$
    (ii) reliability: $R_{sb}[n]\geq R_{\text{Th}}$
    (iii) trajectory feasibility: $\boldsymbol{q}_U[n]\in\mathcal{L}$
    Update penalties (primal-dual step):
      $\Lambda_C\leftarrow[\Lambda_C+\eta_C((1-\epsilon)-\zeta^\star[n])]^+$
    $\Lambda_b\leftarrow[\Lambda_b+\eta_b(R_{\text{Th}}-R_{sb}[n])]^+$
    $\Lambda_p\leftarrow[\Lambda_p+\eta_p(\|\boldsymbol{q}_U[n]-\mathcal{L}\|-d_{\max})]^+$
    Compute reward according to (56)
    Store $(s[n],a[n],R,s[n+1])$ in replay buffer
  end
  Update policy π via DDQN (with stored transitions)
end
Return optimized policy $\pi^\star$

Specifically, each system constraint is associated with a *penalty weight* $(\Lambda_C,\Lambda_b,\Lambda_p)$ that is dynamically updated during learning. At each training step, if a constraint is *violated*, the corresponding penalty weight is *increased*, amplifying its contribution in the reward function. Conversely, when the constraint is satisfied, the weight is gradually reduced or stabilized as described in **Algorithm** 1. This dynamic adjustment ensures that penalties increase when constraints are violated and decrease when satisfied, enabling the agent to balance covert rate maximization with covertness, trajectory feasibility, and QoS guarantees.

B. Proposed DDQN-Based Method

Deep Q-network (DQN) is a reinforcement learning algorithm designed to extend classical Q-learning to high-dimensional and continuous state spaces. While Q-learning is effective in small, discrete settings, it becomes unstable when combined with nonlinear function approximators [1]. DQN

mitigates this issue by integrating a deep neural network to approximate the Q-function, where each input corresponds to a state and each output corresponds to the Q-value of a feasible action [36]. During training, the weights of this neural network are updated to minimize the discrepancy between predicted and target Q-values, thus improving decision-making in large-scale complex environments [33].

However, standard DQN tends to *overestimate* action values in complex problems. To overcome this limitation, the Double DQN (DDQN) algorithm was introduced [33]. DDQN improves stability by decoupling action selection and action evaluation through two separate networks: (i) the *online* network $Q(\mathcal{S}, \mathcal{A}|\theta)$, and (ii) the *target* network $\hat{Q}(\mathcal{S}, \mathcal{A}|\hat{\theta})$, where θ and $\hat{\theta}$ denote their parameters. The online network selects actions, while the target network evaluates them, thereby reducing over-optimism.

Moreover, rather than updating the network parameters using all available data at each iteration, DDQN employs a more efficient training method by utilizing stochastic gradient descent (SGD) on a randomly sampled mini-batch from a replay buffer (RB). Particularly, to further stabilize learning, DDQN employs a *replay buffer* $\mathcal{D}$ of finite size D_{rm}, which stores transition tuples $e = (\mathcal{S}, \mathcal{A}, r, \hat{\mathcal{S}})$. By sampling mini-batches D'_{rm} uniformly from $\mathcal{D} = (e^{(1)}, e^{(2)}, \ldots, e^{(D_{\text{rm}})})$, correlation between consecutive updates is reduced, leading to more robust convergence.

1) Loss Function: The online network selects the action $\arg\max_{\mathcal{A}} Q(\hat{\mathcal{S}}, \mathcal{A}|\theta)$, while the target network evaluates it, decoupling action selection and value estimation. The DDQN temporal-difference (TD) loss is formulated as the mean squared error (MSE) over a mini-batch and is defined as follows [37].

$$\mathcal{L} = \mathbb{E}_{D'_{\text{rm}}} \left(r + \underbrace{\Gamma_{rl} \hat{Q}\left(\hat{\mathcal{S}}, \arg\max_{\mathcal{A}} Q(\hat{\mathcal{S}}, \mathcal{A}|\theta) \mid \hat{\theta}\right)}_{A_1} - \underbrace{Q(\mathcal{S}, \mathcal{A}|\theta)}_{A_2} \right)^2, \tag{59}$$

where Γ_{rl} is the discount factor. The expectation $\mathbb{E}_{D'_{\text{rm}}}$ is taken over the mini-batch samples. where A_1 represents the estimated Q-value of the next state-action pair, discounted by the factor Γ_{rl} and evaluated by the target network $\hat{Q}(\cdot|\hat{\theta})$. The next action is selected by the online network, denoted as $\arg\max_{\boldsymbol{\mathcal{S}}} Q$. A_2 corresponds to the predicted Q-value of the current state-action pair by the online network.

2) Parameter Updates: The online network parameters θ are updated by stochastic gradient descent:

$$\theta^{k+1} = \theta^k - \epsilon_\theta \nabla_\theta \mathcal{L},$$

and the target network is updated periodically every τ iterations via hard updates $\hat{\theta} \leftarrow \theta$.

3) Integration With Covert Communication: In the proposed framework, DDQN is responsible for jointly optimizing BS transmission power $P_s[n]$, jammer power $P_j[n]$, and IRS phase configuration $\boldsymbol{\Theta}[n]$. The reward function used in training incorporates covert throughput, reliability, and covertness constraints with adaptive penalties as described in Sec. IV-A, ensuring that the learned policy balances achievable covert capacity with constraint satisfaction. Therefore, once trained, the DDQN policy yields the optimal joint control strategy for each time slot.

Algorithm 2 Proposed DDQN-Based Approach

Input: CMDP environment model (Algorithm 1), Replay buffer $\mathcal{D} = (e^{(1)}, e^{(2)}, \ldots, e^{(D_{\text{rm}})})$, UAV transmission powers P_s, jammer power P_J, IRS phase configuration $\boldsymbol{\Theta}$, flight angle β, UAV velocity v, UAV trajectory $\boldsymbol{q}$, detection error probability ζ^*
Output: Best solution $(\mathcal{S}^{(k)}, \mathcal{A}^{(k)}, r^{(k)})$

```
/* Initialization */
Initialize replay buffer D
Initialize online Q-network parameters θ
Initialize target Q-network θ̂ ← θ
Initialize state S_0
Initialize noise N
Set exploration parameter ϵ
/* Main Training Loop */
for episode k = 1, ..., K do
    Initialize environment (UAV position, IRS config, power levels)
    for slot n = 1, ..., N do
        Select action a[n] using ϵ-greedy policy
        Execute a[n], update (P_s[n], P_j[n], Θ[n], q_U[n])
        Compute covert throughput C_c[n] and penalties
        Interact with CMDP environment (Algorithm 1) to obtain next state
          Ŝ and reward r[n]^(k) (56)
        Store transition (S_n, a_n, R, S_{n+1}) in buffer D
        Sample mini-batch from D and compute TD loss (60) using online
          and target Q-networks
        Update online network parameters θ using SGD
        Every τ steps: update target network parameters θ̂ ← θ
    end
end
return optimized policy π*
```

The overall DDQN-based algorithm is summarized in **Algorithm 2**.

4) Complexity Analysis: The computational complexity of the proposed DDQN-based method arises from three main components: (i) neural network forward/backward propagation, (ii) replay buffer, and (iii) action cardinality (discretization and IRS codebook).

5) Neural Network Complexity: Let L denote the number of layers, n_ℓ the number of neurons in layer ℓ, B the mini-batch size, and $|\mathcal{A}|$ the (discrete) action *cardinality*. For a standard DQN/DDQN head that outputs $Q(s, \cdot) \in \mathbb{R}^{|\mathcal{A}|}$, the per-update (online network) cost scales as

$$\mathcal{O}\left(B\left[\sum_{\ell=1}^{L-2} n_\ell n_{\ell+1} + n_{L-1}|\mathcal{A}|\right]\right),$$

where the last term reflects the output layer sized by $|\mathcal{A}|$ [38]. DDQN adds one *target-network* forward pass of the same order. Since B and L are moderate (tens to hundreds), this component is tractable provided $|\mathcal{A}|$ is kept reasonable.

6) Replay Buffer Operations: Uniform sampling from a buffer of size D_{rm} costs $\mathcal{O}(B)$ per update (negligible vs. the network). Storing transitions $(s, a, r, s', \text{done})$ requires

$$\mathcal{O}\big(D_{\text{rm}}(2d_s + d_a)\big),$$

where d_s and d_a are the tensor sizes used in code.

7) Action Cardinality and IRS Quantization: With discretized heading/speed/power increments and an IRS codebook of size K,

$$|\mathcal{A}| = |\mathcal{B}_\beta| \cdot |\mathcal{B}_v| \cdot |\mathcal{B}_{\Delta P_s}| \cdot |\mathcal{B}_{\Delta P_j}| \cdot K.$$

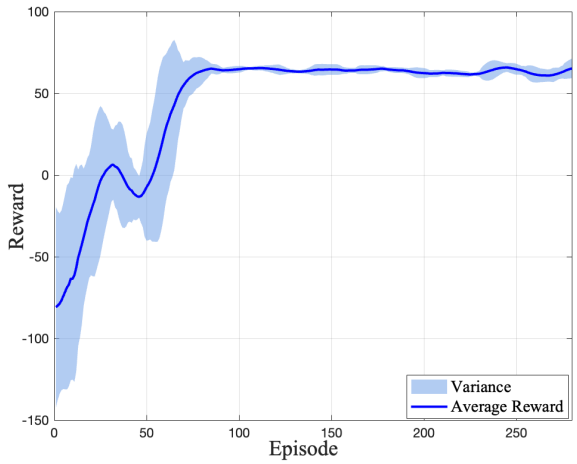

Fig. 2. Training convergence of the DDQN agent: per-episode average reward over training.

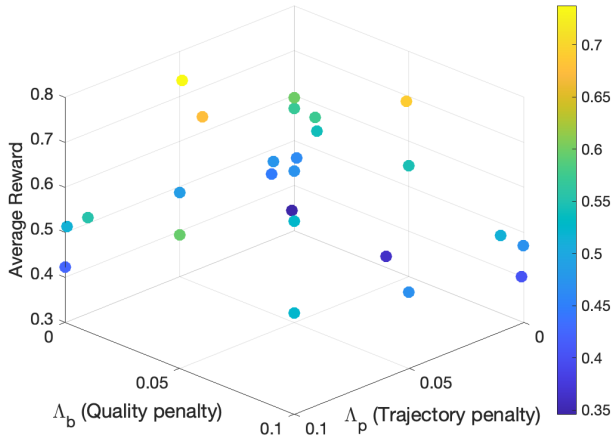

Fig. 3. Average reward versus penalty weights Λ_p (trajectory) and Λ_b (QoS), with Λ_C (covertness) held at 0.01 for initialization.

Without a codebook, b-bit per-element phases yield an IRS factor L^M with $L = 2^b$ (exponential in M). The codebook is replaced L^M by K, keeping $|\mathcal{A}|$ tractable.

8) Environment/Channel Update: Each step computes the cascaded IRS gain and SNR/covert metrics. Applying a codeword $\boldsymbol{\Theta}[n] = \boldsymbol{\theta}^{(c[n])}$ and forming the cascaded channel is $\mathcal{O}(M)$, while computing the feasible band $[L[n], U[n]]$ and $\zeta^\star[n]$ is $\mathcal{O}(1)$ once channel summaries are available. Therefore, this ensures scalability of the algorithm for real-time UAV-IRS control on modern GPUs or edge servers.

V. Numerical Results

In this section, we evaluate the effectiveness of the proposed UAV-IRS-assisted covert communication scheme. Unless otherwise stated, the simulation parameters are as follows. The transmitter (Alice), receiver (Bob), warden (Willie), and friendly jammer are placed at $\boldsymbol{q}_S = (0,0,0)\,\mathrm{m}$, $\boldsymbol{q}_B = (200,0,0)\,\mathrm{m}$, $\boldsymbol{q}_W = (150,0,0)\,\mathrm{m}$, and $\boldsymbol{q}_J = (170,0,0)\,\mathrm{m}$, respectively. The UAV-IRS flies at a fixed altitude of 70 m with initial and final waypoints $\boldsymbol{q}_U^{\text{init}} = (20,50,70)\,\mathrm{m}$ and $\boldsymbol{q}_U^{\text{final}} = (200,50,70)\,\mathrm{m}$.

The transmit-power budgets are $P_{s,\max} = P_{j,\max} = 20$ W. The receiver noise powers are identical at Bob and Willie: $\sigma^2 = \sigma_w^2 = -110dBm$. The reference channel power gain at 1 m is $\beta_0 = -30$ dB. The large-scale path-loss exponents are $\rho = 3$ (air-to-ground) and $\rho_j = 4$ (ground-to-ground). The IRS size is $M \in \{32, 64, 128\}$, and the covertness budget is $\varepsilon = 0.005$ (i.e., we require $\zeta^\star \geq 1 - \varepsilon$). Unless noted, IRS phases are selected from a 2^{bit}-level quantized *codebook* as described earlier. The policy/value network is a feed-forward neural network with three hidden layers of 128 neurons each. All the system parameters, unless otherwise noted, are summarized in Tbale I.

A. Training Performance and Constraint-Violation Analysis

Fig. 2 shows the evolution of the reward during training. Early episodes yield lower (sometimes negative) rewards due to exploration. After roughly 80 episodes, the learning stabilizes and the average reward settles around ≈ 70, indicating convergence of the policy.

To assess the impact of the reward penalties, Fig. 3 plots the average reward under different trajectory weights Λ_p and QoS weights Λ_b while holding the covertness weight at $\Lambda_C = 0.01$ for initialization. Increasing either penalty decreases the achieved reward, as the agent prioritizes constraint satisfaction more strictly. The effect of Λ_p is more pronounced than that of Λ_b, indicating that trajectory feasibility (area/speed constraints) exerts a stronger influence on achievable covert rate than modest variations around the QoS threshold. These observations provide practical guidance for setting penalty initializations before adaptive updates.[1] The recommended

TABLE I

Simulation Parameters (Default Unless Stated)

Symbol	Description	Value / Units
h_U	UAV altitude	70 m
$P_{s,\max},\ P_{j,\max}$	Max source / jammer power	20 W, 20 W
$\sigma^2,\ \sigma_w^2$	Noise power at Bob / Willie	-110 dBm
β_0	Ref. channel power gain @ 1 m	-30 dB
$\rho,\ \rho_j$	Path-loss exponents (A2G / G2G)	3 / 4
M	IRS elements	$\{32, 64, 128\}$
b	IRS phase quantization	2 bits (unless noted)
K	IRS codebook size	32 (unless noted)
ε	Covertness budget (DEP gate)	0.005 $(\zeta^\star \geq 1-\varepsilon)$
κ	Residual jammer leakage at Bob	10^{-4}
$V_{\max},\ \Delta t$	UAV speed limit, slot length	15 m/s, 1 s
$R_{\text{Th}},\ \Gamma_{\text{Th}}$	QoS rate & SINR threshold	1 bps/Hz, $2^{R_{\text{Th}}}-1$

TABLE II

Penalty Weight *initializations* for Reward Tuning (Used With Adaptive Updates)

Penalty	Init. Range	Remarks
Λ_p (Trajectory)	$[0.01,\ 0.05]$	Strong influence on reward; keep moderate to avoid over-penalizing path flexibility.
Λ_b (QoS)	$[0.005,\ 0.02]$	Less sensitive; balances rate target vs. exploration near the threshold.
Λ_C (Covertness)	0.01 (init.)	Initialized small; then adapted to enforce $\zeta^\star \geq 1-\varepsilon$ via primal-dual updates.

[1]The values in Table II are *initializations*; during training we update $(\Lambda_p, \Lambda_b, \Lambda_C)$ using the adaptive/primal-dual rules in Sec. IV-A.

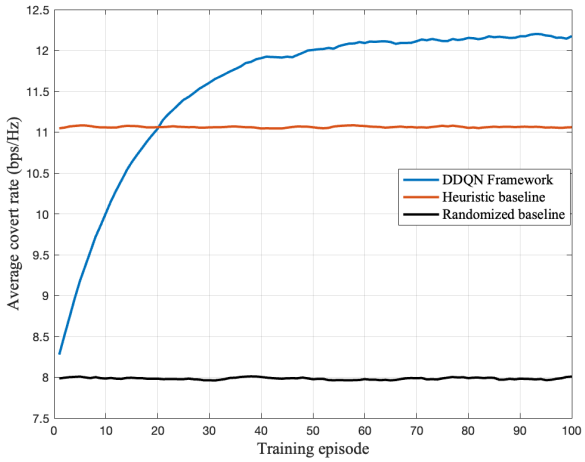

Fig. 4. Convergence of the proposed DDQN policy under covertness constraints. Average covert rate (bps/Hz) versus training episode for the DDQN agent compared with two system-level baselines: (i) *Heuristic* straight-line trajectory at $V_{\max}$ with fixed feasible (P_s, P_j) and a Bob-steering IRS codeword, and (ii) *Randomized* joint policy (trajectory, powers, and IRS codeword sampled within constraints).

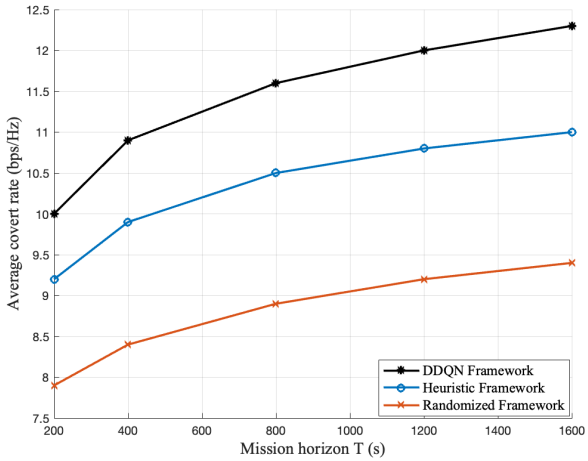

Fig. 5. Average covert rate versus mission horizon T for the proposed DDQN policy and two non-learning baselines, (i) *Heuristic*; (ii)and *Randomized* under identical channels and constraints.

settings for the penalty parameters are summarized in Table II, based on the sensitivity analysis findings.

Fig. 4 reports the average covert rate (bps/Hz) versus training episode for the proposed DDQN agent and two system-level baselines under identical channels and constraints. The DDQN curve increases rapidly during the first 30-40 episodes as the agent learns to: (i) select Bob-steering IRS codewords from the K-size codebook; (ii) exploit geometry to reduce the lower bound of the feasibility band $[L[n], U[n]]$ on $P_j[n]$; and (iii) allocate power to raise while respecting the covertness constraint. The policy surpasses the heuristic straight-line baseline around the episode ∼18 and stabilizes near 12.2 bps/Hz by ∼50 episodes, yielding gains of ≈ 1.1 bps/Hz over the heuristic and ≈ 4.2 bps/Hz over the randomized joint baseline. The flat baseline lines reflect episode-invariant reference policies; the low variance of the DDQN trace after convergence indicates stable training and consistent constraint satisfaction.

Fig. 5 provides a system-level comparison for average covert rate versus mission horizon T for the proposed DDQN policy and two non-learning baselines under identical channels and constraints. As the mission horizon increases from $T = 200$ to 1600 s, all methods exhibit a monotone rise in average covert rate because the UAV can reach longer in favorable geometries near Bob while respecting the per-slot feasibility band $P_j[n] \in [L[n], U[n]]$. The DDQN policy consistently dominates, improving from ≈ 10.0 to ≈ 12.3 bps/Hz; the heuristic increases from ≈ 9.2 to ≈ 11.0 bps/Hz; and the randomized baseline from ≈7.9 to ≈9.4 bps/Hz. The persistent DDQN advantage (about 1.3-2.9 bps/Hz over randomized and 1.0-1.3 bps/Hz over the heuristic across T) stems from joint, state-aware coordination of trajectory, power allocation, and IRS codebook selection under the CMDP constraints.

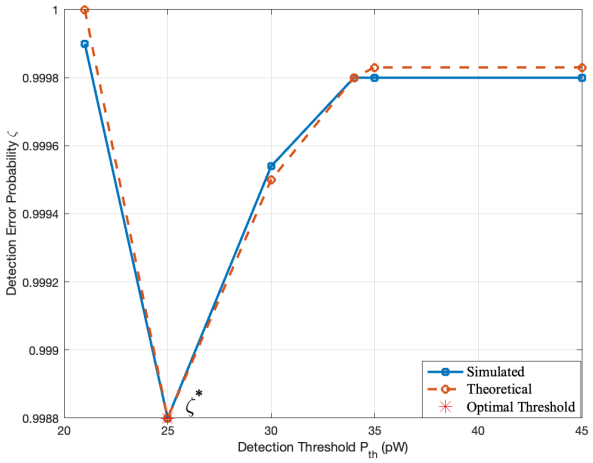

Fig. 6. Detection error probability ζ versus detection threshold P_{th} for the UAV-IRS-assisted covert system when $M = 32$.

B. Performance Analysis

Maintaining a high detection error probability is essential for covert communication in XR scenarios, as it renders transmissions statistically indistinguishable from noise and prevents an adversary from inferring the presence of immersive traffic. In Fig. 6, we analyze the total DEP at Willie, $\zeta = P_{\text{FA}} + P_{\text{MD}}$, as a function of the detector threshold P_{th} for $(P_j, P_s) = (10\text{ W}, 20\text{ W})$. As predicted by the closed-form analysis (Sec. III-A), ζ decreases with P_{th}, attains a unique minimum at the theoretical optimum $P_{\text{th}}^{\star}$, and then increases as P_{th} grows further. In our setup, the minimum occurs near $P_{\text{th}} \approx 25$ pW, matching the theory and validating the DEP model.

Fig. 7 compares covert rate performance for three IRS strategies with $M = 32$: (i) random phase selection, (ii) greedy iterative refinement, and (iii) the proposed DDQN-optimized policy. The DDQN agent consistently dominates across runs, reaching ≈ 10 bps/Hz after 30 runs, while greedy converges

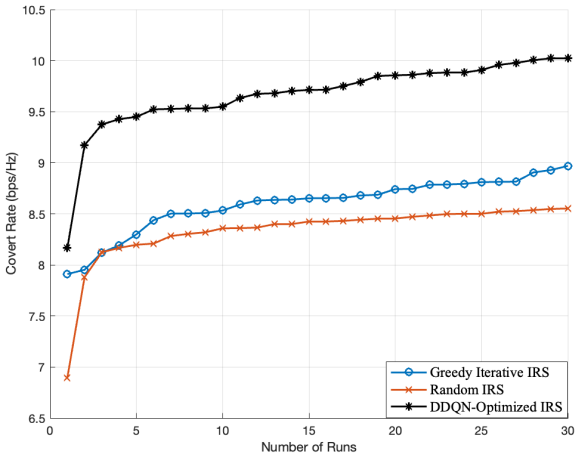

Fig. 7. Covert rate versus number of runs for IRS configuration strategies: random, greedy iterative, and DDQN-optimized (UAV-IRS system, $M = 32$).

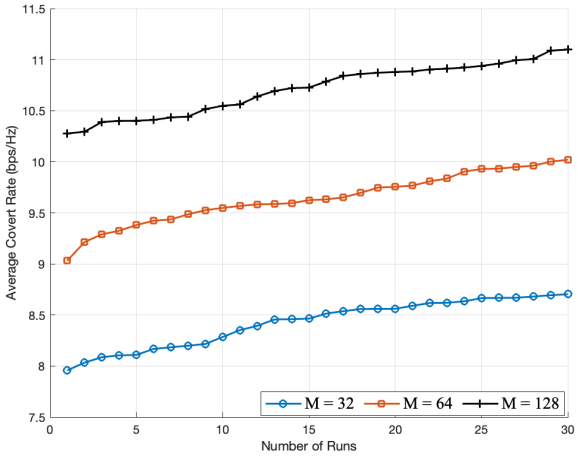

Fig. 8. Average covert rate versus number of runs for different IRS sizes $M \in \{32, 64, 128\}$ under the proposed DDQN-based optimization.

near 9 bps/Hz and random saturates around 8.5 bps/Hz. Random selection improves marginally and then stalls (no state-aware adaptation), and greedy often gets trapped in local optima. In contrast, DDQN leverages the learning capability to continuously exploit trajectory, power, and codebook actions jointly, yielding an additional $\approx$ 1.5-2 bps/Hz covert rate gain. This improvement is particularly critical for maintaining secure XR communications in dynamic environments.

Fig. 8 shows how the IRS size M impacts the average covert rate under the proposed DDQN framework. Three different configurations are considered: $M = 32$, $M = 64$, and, $M = 128$ reflecting small, medium, and large IRS sizes, respectively. As runs progress, all configurations improve, confirming sample-efficient learning of the IRS codebook policy. Larger IRS arrays provide higher rates due to stronger coherent combining, with $M = 128$ the average covert rate exceeds 11 bps/Hz, compared with $\approx$ 10 bps/Hz for $M = 64$ and $\approx$ 8.7 bps/Hz for $M = 32$. Particularly, the relative covert rate gain confirms the theoretical expectation that the beamforming capability

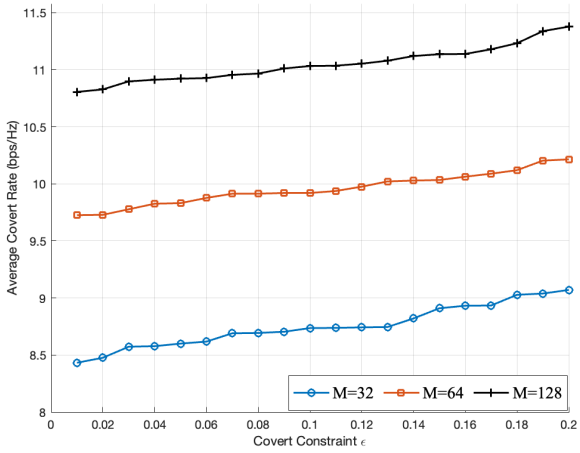

Fig. 9. Average covert rate versus covertness budget ε for IRS sizes $M \in \{32, 64, 128\}$ under the proposed DDQN policy. Larger ε relaxes the constraint $\zeta^\star \geq 1 - \varepsilon$.

of the IRS improves with the number of elements, thereby enhancing the effective signal power at the legitimate receiver while maintaining covertness against the adversary.

Note that, while increasing M boosts covert throughput, it also raises optimization complexity, hardware cost, and UAV payload/energy demands. In many UAV-assisted covert deployments, $M = 64$ strikes a favorable balance between performance and complexity; $M = 128$ is attractive when maximizing covert throughput is critical and the platform can accommodate the added burden.

Fig. 9 shows the achievable average covert rate as the covertness budget ε varies for $M = \{32, 64, 128\}$. By definition, ε sets the tolerated detection-error shortfall through $\zeta^\star \geq 1 - \varepsilon$, where increasing ε *relaxes* the covertness requirement, allowing the policy to either (i) reduce the friendly-jammer power toward its lower bound $L(\varepsilon)$ or (ii) select more Bob-centric IRS codewords, both of which raise Bob's SINR and thus the covert rate. Consistent with this, the average covert rate increases monotonically with ε. For example, with $M = 128$ the average rate rises from $\approx$ 10.8 bps/Hz at $\varepsilon \approx 0$ to $\approx$ 11.4 bps/Hz at $\varepsilon = 0.2$; the $M = 64$ and $M = 32$ cases follow the same trend, reaching $\approx$ 10.2 bps/Hz and $\approx$ 9.1 bps/Hz at $\varepsilon = 0.2$, respectively. The insights confirm the effectiveness of the DDQN-based policy in adapting to varying covert constraints, and the careful design of IRS size and covert constraint thresholds is crucial for optimizing the covert capacity of UAV-assisted wireless systems.

Fig. 10 shows the impact of the available travel time $T = N\Delta t$ on the *average covert rate* under three trajectory policies: random, greedy iterative, and the proposed DDQN-optimized policy. As T increases, the average covert rate improves because the UAV can reposition more effectively (more slots N) to strengthen the cascaded Alice→UAV→Bob link while keeping the covertness constraints satisfied. The DDQN-optimized trajectory consistently dominates: e.g., at $T = 1600$ s it achieves $\approx$ 12.3 bps/Hz versus $\approx$ 10.2 bps/Hz (greedy) and $\approx$ 9.7 bps/Hz (random). The gap widens with

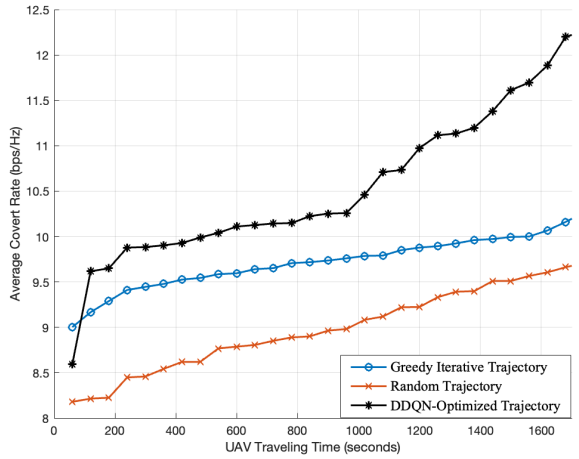

Fig. 10. Average covert rate (bps/Hz) versus UAV travel time for random, greedy, and DDQN-optimized trajectory strategies (IRS size $M = 32$).

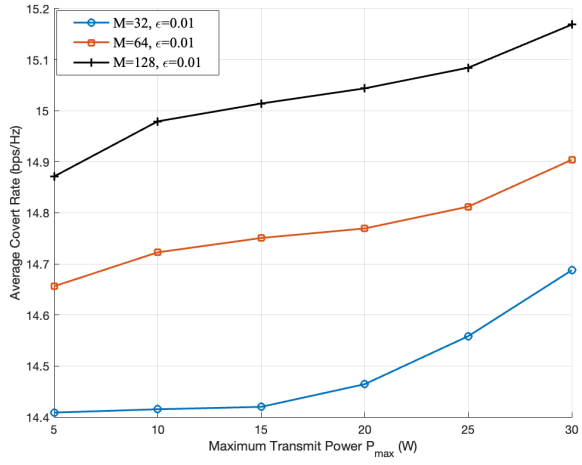

Fig. 11. Average covert rate (bps/Hz) versus maximum transmit power $P_{\max}$ for IRS sizes $M \in \{32, 64, 128\}$, with covertness budget $\varepsilon = 0.01$.

T because DDQN jointly optimizes heading/speed, power steps, and IRS codebook index across time, whereas greedy is locally greedy, and random lacks state awareness. These results validate the effectiveness of the DDQN-based optimization framework in adaptively adjusting UAV trajectories, highlighting the crucial role of intelligent mobility control in enhancing covert communication performance in UAV-IRS-assisted systems. Allowing longer UAV traveling times significantly improves covert communication performance, but mission-specific constraints (such as energy limits or real-time latency requirements) must be considered when planning trajectories.

Fig. 11 illustrates the effect of the power budget on performance for different IRS sizes. Increasing $P_{\max}$ raises the achievable average covert rate across all M because the DDQN agent can exploit stronger coherent combining at Bob while respecting the covertness constraint. Larger M systematically yields higher covert rates, confirming the effectiveness of IRS-assisted beamforming in enhancing covert transmission capabilities. At high levels of $P_{\max}$, the curves saturate, especially for $M = 128$, because the covertness requirement effectively limits the usable signal power and may force higher jamming (or conservative codebook choices), limiting further growth. These results validate the significant role of IRS and power control in optimizing covert communications under security constraints.

In summary, the simulation results validate the effectiveness of the proposed DDQN-based framework in significantly enhancing covert communication performance. Intelligent UAV mobility control, IRS phase adjustment, and adaptive power allocation collectively contribute to achieving high covert capacity while maintaining stringent covertness constraints. *Takeaways.* (i) A larger IRS (M) consistently improves the achievable covert rate because of the stronger coherent combining toward Bob; (ii) increasing ε trades covertness for throughput in a predictable, monotone way; and (iii) gains saturate once the jammer power is no longer binding and performance becomes limited by geometry, power budgets, and codebook quantization.

VI. Conclusion

This paper presented a DDQN-based framework for covert communication in UAV-based IRS systems that jointly optimizes source and jamming power control, UAV trajectory, and IRS phase configuration to improve covert throughput under stringent detection constraints. By incorporating a derived analytical expression for the optimal detection threshold and the minimum detection error probability into the system model, the proposed method bridges learning-based decision-making with physical layer guarantees for covert performance. In contrast to conventional approaches where the system parameters are optimized separately, our CMDP-based formulation enables an adaptive strategy that considers time-varying environments and learning conflicts. Simulation results verify that the proposed method significantly outperforms baseline methods in terms of average covert rate with controlled detection error and energy-efficient UAV behavior. Moreover, sensitivity analyses highlight the importance of balancing trajectory and quality penalties in the reward function. This work offers a promising direction for next-generation secure UAV communication systems, particularly in environments where covertness and mobility are critical.

References

[1] D. Wang, L. Cao, W. Shen, Z. Li, and Q. Li, "Age-of-information minimization in aerial-IRS-assisted covert communication for Internet of Things networks," *IEEE Internet Things J.*, vol. 12, no. 8, pp. 9583–9595, Apr. 2025.

[2] M. H. Khoshafa et al., "RIS-assisted physical layer security in emerging RF and optical wireless communication systems: A comprehensive survey," *IEEE Commun. Surv. Tutor.*, vol. 27, no. 4, pp. 2156–2203, Apr. 2025.

[3] J. M. Hamamreh, H. M. Furqan, and H. Arslan, "Classifications and applications of physical layer security techniques for confidentiality: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 2, pp. 1773–1828, 2nd Quart., 2019.

[4] Y. Jiang, L. Wang, H.-H. Chen, and X. Shen, "Physical layer covert communication in B5G wireless networks—Its research, applications, and challenges," *Proc. IEEE*, vol. 112, no. 1, pp. 47–82, Jan. 2024.

[5] X. Chen et al., "Covert communications: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 2, pp. 1173–1198, 2nd Quart., 2023.

[6] D. Wang, P. Qi, N. Zhang, J. Si, Z. Li, and N. Al-Dhahir, "Covert wireless communication with spectrum mask in Internet of Things networks," *IEEE Trans. Commun.*, vol. 69, no. 12, pp. 8402–8415, Dec. 2021.

[7] B. A. Bash, D. Goeckel, D. Towsley, and S. Guha, "Hiding information in noise: Fundamental limits of covert wireless communication," *IEEE Commun. Mag.*, vol. 53, no. 12, pp. 26–31, Dec. 2015.

[8] X. Chen et al., "Multi-antenna covert communication via full-duplex jamming against a warden with uncertain locations," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 5467–5480, Aug. 2021.

[9] T.-X. Zheng, Z. Yang, C. Wang, Z. Li, J. Yuan, and X. Guan, "Wireless covert communications aided by distributed cooperative jamming over slow fading channels," *IEEE Trans. Wireless Commun.*, vol. 20, no. 11, pp. 7026–7039, Nov. 2021.

[10] X. Lu, E. Hossain, T. Shafique, S. Feng, H. Jiang, and D. Niyato, "Intelligent reflecting surface enabled covert communications in wireless networks," *IEEE Netw.*, vol. 34, no. 5, pp. 148–155, Sep. 2020.

[11] X. Chen et al., "UAV relayed covert wireless networks: Expand hiding range via drones," *IEEE Netw.*, vol. 36, no. 4, pp. 226–232, Jul. 2022.

[12] X. Liu, Y. Yu, F. Li, and T. S. Durrani, "Throughput maximization for RIS-UAV relaying communications," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 19569–19574, Oct. 2022.

[13] H. Shen, Q. Ye, W. Zhuang, W. Shi, G. Bai, and G. Yang, "Drone-small-cell-assisted resource slicing for 5G uplink radio access networks," *IEEE Trans. Veh. Technol.*, vol. 70, no. 7, pp. 7071–7086, Jul. 2021.

[14] S. Yan, S. V. Hanly, and I. B. Collings, "Optimal transmit power and flying location for UAV covert wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 11, pp. 3321–3333, Nov. 2021.

[15] X. Jiang, Z. Yang, N. Zhao, Y. Chen, Z. Ding, and X. Wang, "Resource allocation and trajectory optimization for UAV-enabled multi-user covert communications," *IEEE Trans. Veh. Technol.*, vol. 70, no. 2, pp. 1989–1994, Feb. 2021.

[16] X. Chen, Z. Chang, M. Liu, N. Zhao, T. Hämäläinen, and D. Niyato, "UAV-IRS assisted covert communication: Introducing uncertainty via phase shifting," *IEEE Wireless Commun. Lett.*, vol. 13, no. 1, pp. 103–107, Jan. 2024.

[17] Q. Wu and R. Zhang, "Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 106–112, Jan. 2020.

[18] Q. Wang et al., "STAR-RIS aided covert communication in UAV air-ground networks," *IEEE J. Sel. Areas Commun.*, vol. 43, no. 1, pp. 245–259, Jan. 2025.

[19] Y. Wu, S. Wang, J. Luo, and W. Chen, "Passive covert communications based on reconfigurable intelligent surface," *IEEE Wireless Commun. Lett.*, vol. 11, no. 11, pp. 2445–2449, Nov. 2022.

[20] M. Cui, G. Zhang, and R. Zhang, "Secure wireless communication via intelligent reflecting surface," *IEEE Wireless Commun. Lett.*, vol. 8, no. 5, pp. 1410–1414, Oct. 2019.

[21] C. Wang et al., "Covert communication assisted by UAV-IRS," *IEEE Trans. Commun.*, vol. 71, no. 1, pp. 357–369, Jan. 2023.

[22] M. Tatar Mamaghani and Y. Hong, "Aerial intelligent reflecting surface-enabled terahertz covert communications in beyond-5G Internet of Things," *IEEE Internet Things J.*, vol. 9, no. 19, pp. 19012–19033, Oct. 2022.

[23] X. Tang, D. Wang, R. Zhang, Z. Chu, and Z. Han, "Jamming mitigation via aerial reconfigurable intelligent surface: Passive beamforming and deployment optimization," *IEEE Trans. Veh. Technol.*, vol. 70, no. 6, pp. 6232–6237, Jun. 2021.

[24] T. Shafique, H. Tabassum, and E. Hossain, "Optimization of wireless relaying with flexible UAV-borne reflecting surfaces," *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 309–325, Jan. 2021.

[25] T. L. Nguyen, G. Kaddoum, T. N. Do, and Z. J. Haas, "Channel characterization of UAV-RIS-aided systems with adaptive phase-shift configuration," *IEEE Wireless Commun. Lett.*, vol. 12, no. 12, pp. 2058–2062, Dec. 2023.

[26] X. Guo, Y. Chen, and Y. Wang, "Learning-based robust and secure transmission for reconfigurable intelligent surface aided millimeter wave UAV communications," *IEEE Wireless Commun. Lett.*, vol. 10, no. 8, pp. 1795–1799, Aug. 2021.

[27] B. He, Y. She, and V. K. N. Lau, "Artificial noise injection for securing single-antenna systems," *IEEE Trans. Veh. Technol.*, vol. 66, no. 10, pp. 9577–9581, Oct. 2017.

[28] Q. Zhao, C. Gao, D. Zheng, Y. Li, and X. Zheng, "Covert communication in a multirelay-assisted wireless network with an active warden," *IEEE Internet Things J.*, vol. 11, no. 9, pp. 16450–16460, May 2024.

[29] C. Gao, B. Yang, X. Jiang, H. Inamura, and M. Fukushi, "Covert communication in relay-assisted IoT systems," *IEEE Internet Things J.*, vol. 8, no. 8, pp. 6313–6323, Apr. 2021.

[30] X. Chen, Z. Chang, and T. Hämäläinen, "Enhancing covert secrecy rate in a zero-forcing UAV jammer-assisted covert communication," *IEEE Wireless Commun. Lett.*, vol. 13, no. 12, pp. 3375–3379, Dec. 2024.

[31] D. A. Barry, J.-Y. Parlange, L. Li, H. Prommer, C. J. Cunningham, and F. Stagnitti, "Analytical approximations for real values of the Lambert W-function," *Math. Comput. Simul.*, vol. 53, nos. 1–2, pp. 95–103, Aug. 2000.

[32] T. K. Rodrigues, K. Suto, H. Nishiyama, J. Liu, and N. Kato, "Machine learning meets computation and communication control in evolving edge and cloud: Challenges and future perspective," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 1, pp. 38–67, 1st Quart., 2020.

[33] H. V. Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2016, vol. 30, no. 1, pp. 2094–2100.

[34] Q. Wu, S. Zhang, B. Zheng, C. You, and R. Zhang, "Intelligent reflecting surface-aided wireless communications: A tutorial," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 3313–3351, May 2021.

[35] D. Jakovetic, D. Bajovic, J. Xavier, and J. M. F. Moura, "Primal-dual methods for large-scale and distributed convex optimization and data analytics," *Proc. IEEE*, vol. 108, no. 11, pp. 1923–1938, Nov. 2020.

[36] Y. Zhang, Z. Mou, F. Gao, J. Jiang, R. Ding, and Z. Han, "UAV-enabled secure communications by multi-agent deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 11599–11611, Oct. 2020.

[37] A. Gao, Q. Wang, Y. Hu, W. Liang, and J. Zhang, "Dynamic role switching scheme with joint trajectory and power control for multi-UAV cooperative secure communication," *IEEE Trans. Wireless Commun.*, vol. 23, no. 2, pp. 1260–1275, Feb. 2025.

[38] C. Qiu, Y. Hu, Y. Chen, and B. Zeng, "Deep deterministic policy gradient (DDPG)-based energy harvesting wireless communications," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 8577–8588, Oct. 2019.