entropy

MDPI

*Article*

# Secure Cooperative Communications in 6G Networks: A Constrained Hierarchical Reinforcement Learning Framework with Hybrid Action Space

**Xiaosi Tian, Zulin Wang and Yuanhan Ni ***

School of Electronic and Information Engineering, Beihang University, Beijing 100191, China; xiaosi_tian@buaa.edu.cn (X.T.); wzulin@buaa.edu.cn (Z.W.)
* Correspondence: yuanhanni@buaa.edu.cn

**Abstract**

With the rapid evolution toward 6G networks, ensuring robust physical layer security (PLS) in highly dynamic and heterogeneous wireless environments has become a key challenge. Traditional security methods often struggle to adapt to time-varying channels, especially in the absence of perfect channel state information. Furthermore, the dynamic nature of node selection and power allocation in heterogeneous networks creates a complex hybrid action space operating across multiple timescales, significantly complicating the design of efficient and adaptive security strategies. To address this, this paper proposes a novel constrained hierarchical reinforcement learning (CHRL) framework for secure cooperative communications in next-generation wireless systems. The framework is designed to optimize secrecy performance within a hybrid action space comprising both discrete node selection and continuous power allocation, operating at different timescales. By hierarchically decoupling the joint optimization problem, the upper layer performs risk-aware node selection to maximize long-term secrecy capacity (SC) while guaranteeing a stable and secure link. At the lower layer, we develop a constrained MiniMax Multi-objective Deep Deterministic Policy Gradient (M3DDPG) algorithm that optimizes power allocation considering worst-case conditions. Lagrange multipliers are integrated to enforce a strictly positive SC constraint throughout transmission, effectively preventing security outages. Simulation results under time-varying Rayleigh fading channels demonstrate that the proposed CHRL framework outperforms existing HRL methods, achieving up to 17% improvement in SC while strictly maintaining security constraints. These results validate the effectiveness of the proposed approach for enhancing PLS in next-generation cooperative wireless networks.

**Keywords:** hierarchical reinforcement learning; cooperative communication; secrecy capacity; relay selection; power allocation



Academic Editor: Giorgio Taricco

Received: 14 February 2026
Revised: 31 March 2026
Accepted: 3 April 2026
Published: 4 April 2026



## 1. Introduction

Next-generation wireless systems envisioned for beyond-5G/6G networks are evolving toward highly dynamic and heterogeneous architectures characterized by ultra-dense device connectivity, diverse service requirements, and rapidly time-varying network topologies [1,2]. The coexistence of massive Internet-of-Things devices and high-mobility terminals significantly increases environment complexity. Consequently, this openness introduces severe challenges for secure confidential transmission [3]. Since conventional cryptographic mechanisms often struggle with real-time threats and channel uncertainty,

physical layer security (PLS) has emerged as a vital solution. By exploiting the inherent randomness of wireless channels, PLS effectively enhances communication confidentiality in these complex scenarios [4,5].

Extensive studies have investigated PLS in communication systems [6–8]. One effective approach is to introduce relays and jammers as intermediate nodes in wireless communication networks [9–11]. Jammers transmit artificial interference to degrade the eavesdropping channel, while relays enhance the legitimate channel by forwarding source signals to the intended destination. In this way, the main channel can be strengthened relative to the eavesdropping channel, thereby improving the secrecy capacity (SC). To maximize secrecy performance, it is essential to select appropriate relays and jammers from multiple candidates and dynamically adjust their transmit power strategies. To achieve this, various optimization-based methods have been employed in the literature. For example, authors in [6] studied cooperative networks with selfish but friendly intermediate nodes and proposed a pricing-based relay and jammer selection algorithm to maximize SC. In [12], an optimal power allocation strategy was derived for simple multi-hop networks under system constraints. The work in [13] introduced an antenna selection scheme to leverage spatial diversity for enhanced security. Despite their contributions, these approaches critically depend on the availability of perfect channel state information (CSI).

However, the dynamic nature of wireless channels, often modeled as time-varying fading processes, introduces considerable uncertainty into the system [14–16]. Under such conditions, traditional optimization-based methods for node selection and power allocation become unreliable, as they generally need prior knowledge of the environment. Moreover, the joint optimization of node selection and power allocation creates a hybrid action space comprising both discrete and continuous variables that operates on different timescales. This hybrid action space at different timescales further complicates the design of efficient and adaptive security strategies [17].

Reinforcement learning (RL) offers a powerful model-free framework for addressing adaptive decision-making problems under uncertainty [18–20]. By continuously interacting with the environment, RL agents are able to learn transmission strategies without requiring explicit system models, making RL particularly attractive for dynamic and complex wireless networks [21]. In the application of RL in complex 6G heterogeneous scenarios [22], hierarchical RL (HRL) has been introduced to address the challenges of hybrid action space optimization, specifically involving joint node selection and power allocation for PLS [23,24]. By decomposing intricate resource allocation tasks into distinct sub-levels, this framework achieves superior convergence performance compared to conventional flat RL architectures [25–27]. In [25], a HRL approach is proposed, performing relay selection and power allocation at separate levels. However, a key limitation of these methods lies in the discretization of power allocation actions, which prevents the model from attaining theoretically optimal power values. Additionally, existing studies mainly focus on SC maximization [28] or secrecy outage probability minimization [25,28]. Most of these studies did not impose strict positive SC constraints, making it difficult to guarantee the requirement of stable secure communication.

Furthermore, conducting this research highlights several key challenges in designing adaptive security strategies to address these limitations. First, the joint optimization of topology management and resource allocation creates a complex hybrid action space. Directly applying conventional RL algorithms often leads to a combinatorial explosion, rendering the training process unstable and computation strictly infeasible for real-time deployment. Second, maintaining a strictly positive SC at all times is exceedingly difficult in highly dynamic environments where perfect instantaneous CSI is unavailable. Finally, achieving robust continuous power control requires striking a delicate balance in a competitive scenario.

The cooperative jammers must sufficiently degrade the eavesdropping channels without inadvertently collapsing the legitimate transmission links, a multi-objective trade-off that is highly sensitive to parameter oscillations during model training.

To address the above challenges, this paper proposes a constrained HRL (CHRL) framework for hybrid action space optimization in secure cooperative communication. The main contributions of this paper are summarized as follows.

- Hybrid action hierarchical learning framework
  Different from conventional HRL approaches that discretize continuous variables or optimize a single decision layer, we develop a hybrid action hierarchical framework that jointly models discrete node selection and continuous power allocation across multiple timescales. The proposed framework not only resolves the combinatorial explosion problem inherent in hybrid discrete-continuous action spaces but also overcomes the performance degradation caused by power allocation discretization seen in conventional methods.
- Strict secrecy constraint enforcement
  Existing studies primarily focus on maximizing SC or minimizing outage probability, often failing to impose strict positive SC constraints. To address this, our framework enforces security guarantees across both decision levels. Our upper-level policy explicitly models secrecy violations as long-term risk signals, enabling a reward–risk coupled policy for discrete node selection that guarantees a strictly positive SC under dynamic channel variations. Complementing this, the lower-level policy employs a Lagrange multiplier-based constraint mechanism, ensuring that continuous power control is optimized while maintaining secure communication at all times.
- Constrained MiniMax continuous power allocation under worst-case channel uncertainty
  In jammer-assisted cooperative communication, the lower-level power allocation is inherently a competitive game: friendly jammers must degrade eavesdropping channels while minimizing their interference to legitimate receivers. Existing methods often fail to fully account for this delicate balance, leading to vulnerable strategies that severely compromise system robustness. To address this, we introduce a virtual adversary to represent worst-case channel conditions and the potential degradation of the legitimate link caused by jammers. By modeling this competitive scenario as a constrained zero-sum game, we deploy a constrained MiniMax Multi-objective Deep Deterministic Policy Gradient (M3DDPG) algorithm to achieve robust and strictly secure continuous power allocation at the physical layer.

In summary, the proposed framework jointly integrates hierarchical hybrid-action optimization, risk-aware node selection, and constrained minimax power control into a unified security-aware learning architecture. Numerical results show our framework outperforms existing methods, improving SC by 6%, 17%, and 49% over H-RL [25], hierarchical Q/E networks (H-Q/E) [23], and M3DDPG [29], respectively, while guaranteeing strict adherence to security constraints.

The rest of this paper is organized as follows. Section 2 presents the system model, including the Rayleigh time-varying channel and the cooperative communication setup with relays and jammers. Section 3 formulates the aforementioned secure cooperative communication scenario as a constrained Markov decision process (CMDP). Section 4 introduces the proposed CHRL framework, detailing the two-level structure for relay/jammer selection and power allocation. The experimental results are presented in Section 5 and the conclusion is drawn in Section 6.

## 2. System Model

In this section, we first establish a Rayleigh time-varying channel using stochastic differential equations (SDEs). Under this channel, we propose a secure cooperative wireless communication system with relays and jammers.

### *2.1. Rayleigh Time-Varying Channel Using SDEs*

To better capture the characteristics of real-world wireless channels, paper [16] employs SDEs to model the random behavior of the signal envelope variations in the time domain. The in-phase and quadrature components, denoted as $I(t), Q(t)_{t\geq 0}$, are considered as Markov uncorrelated Gaussian random processes with time-varying mean and variance. Therefore, $I(t)$ and $Q(t)$ are solutions to the following identically distributed Ornstein–Uhlenbeck SDEs for $t > 0$:

$$\begin{cases} dI(t) = -\frac{1}{2}BI(t)dt + \frac{\sqrt{2}}{2}B^{\frac{1}{2}}\sigma dW^{(I)}(t), & t > 0, \\ dQ(t) = -\frac{1}{2}BQ(t)dt + \frac{\sqrt{2}}{2}B^{\frac{1}{2}}\sigma dW^{(Q)}(t), & t > 0, \\ I(0) = I_0, \\ Q(0) = Q_0, \end{cases} \tag{1}$$

where $B = 2k$ and $\sigma = \frac{\beta}{\sqrt{k}}$. $W^{(I)}(t)$ and $W^{(Q)}(t)$ are independent Wiener processes. In this context, let $g(t)$ be the Markovian projection of $g(t) = I^2(t) + Q^2(t)$, which is obtained by solving the following SDE [16]:

$$\begin{cases} dg(t) & = B\big(\sigma^2 - g(t)\big)dt + \sigma\sqrt{2Bg(t)}dW(t), \\ g(0) & = g(0). \end{cases} \tag{2}$$

Using the Fokker–Planck equation, the limit distribution $p(.)$ of the projected process $g(t)$ satisfies:

$$\frac{\partial}{\partial y}(yp(y)) = \Big(1 - \frac{y}{\sigma^2}\Big)p(y), \quad y > 0. \tag{3}$$

The solution of Equation (3) is:

$$p(y) = \frac{1}{\sigma^2}e^{\frac{-y}{\sigma^2}}, \quad y > 0. \tag{4}$$

Therefore, $g(t)$ follows an asymptotic exponential distribution with parameter $\frac{1}{\sigma^2}$, which is the exact asymptotic distribution of $g(t)$. We solved the original SDE Equation (1) and the projected SDE Equation (2), where $T$ is the final time. The histograms of $g(T)$ and $I^2(T) + Q^2(T)$ are compared in Figure 1. As shown in the figure, the samples of $g(T)$ produce a histogram that closely resembles $I^2(T) + Q^2(T)$.

Additionally, Figure 2 shows the variation in time-varying Rayleigh channel coefficients modeled by the SDE over time. We note that the waveform manifests as randomly occurring pulses or spikes, with the amplitudes of these pulses adhering to an exponential distribution.

This channel modeling approach implies that the channel coefficients cannot be accurately predicted in advance during node selection and power allocation. Consequently, traditional optimization-based methods that rely on perfect CSI become inapplicable under this setting.

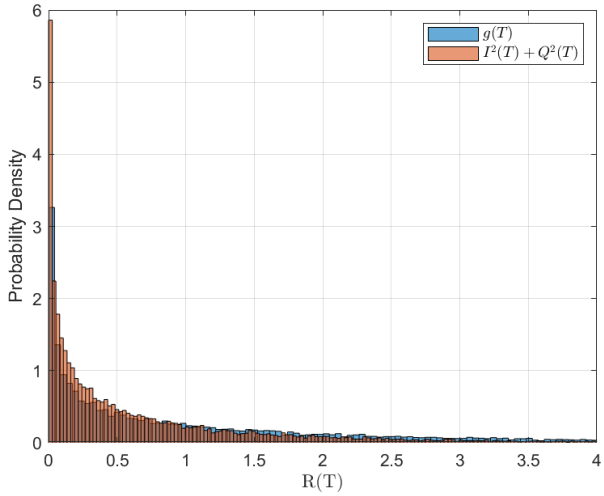


**Figure 1.** Rayleigh Fading: Histogram of $g(T)$ and $I^2(T) + Q^2(T)$ using $10^6$ samples with the following parameters: $T = 4$, $N = 100$, $I_0 = 0$, $Q_0 = 0$, $B = 1$, and $\sigma = 1$.

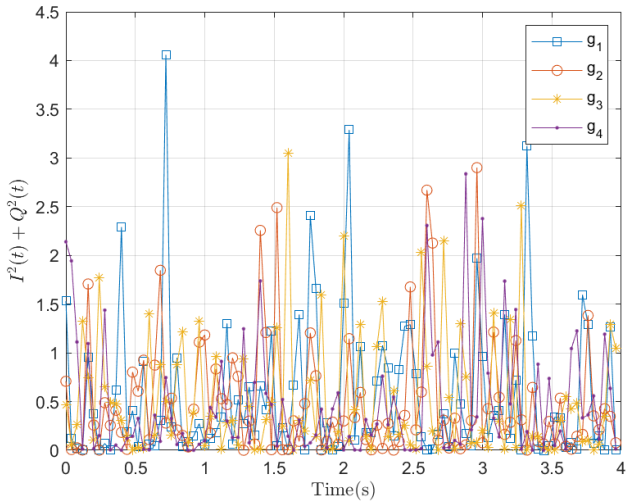


**Figure 2.** Samples of $I^2(t) + Q^2(t)$ with the following parameters: $T = 4$, $N = 100$, $I_0 = 0$, $Q_0 = 0$, $B = 1$, and $\sigma = 1$.

### *2.2. Multi-Hop Secure Cooperative Communication Framework*

We delineate a sophisticated multi-hop cooperative communication network architecture designed for secure information transmission in complex and dynamic environments, shown in Figure 3. The network resides in a wireless setting comprising a source node $S$, a set of $N$ legitimate destination (Bobs) denoted by $\mathcal{D} = \{D_1, D_2, \ldots, D_N\}$, and a group of $Q$ external eavesdroppers (Eves) represented by $\mathcal{E} = \{E_1, \ldots, E_Q\}$. To facilitate reliable communication over extended distances and overcome channel fading, we consider a pool of $K$ candidate nodes $\mathcal{R} = \{R_1, R_2, \ldots, R_K\}$, which serve as the relaying infrastructure. For each destination $D_n$, a specific multi-hop transmission path $\mathcal{P}_n$ is established, represented by an ordered sequence of nodes $\mathcal{P}_n = \{n_{0,n}, n_{1,n}, \ldots, n_{M_n,n}\}$, where $n_{0,n} = S$ is the source and $n_{M_n,n} = D_n$ is the $n$-th destination. A defining characteristic of this model is the presence of simultaneous internal and external security threats, where the intermediate nodes $\{n_{i,n}\}_{i=1}^{M_n-1} \subset \mathcal{R}$ are categorized as untrusted relays. These nodes are essential for forwarding confidential messages using a decode-and-forward (DF) protocol but also act as potential internal eavesdroppers. To counteract these threats, a subset of idle nodes $\mathcal{J}(t) \subset \mathcal{R}$ is dynamically selected as cooperative jammers to transmit jamming signals to degrade the channels of both external and internal eavesdroppers. Regarding hardware and channel properties, all nodes are equipped with a single antenna and operate in half-duplex

mode, while the wireless channels are modeled as independent Rayleigh time-varying fading processes.

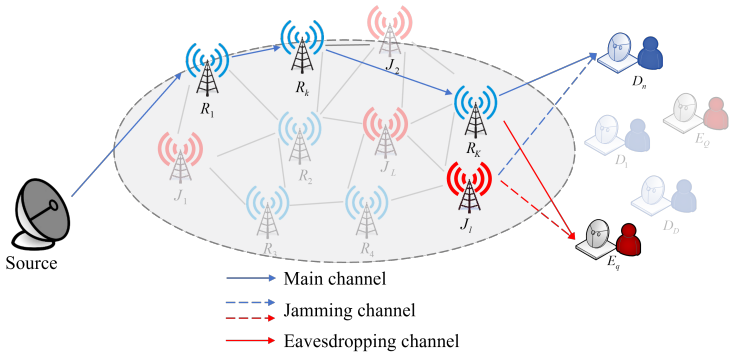


**Figure 3.** System model of cooperative communication.

To focus our analysis on the relaying links, we assume that there is no direct communication path from the source to Bobs and Eves. This means that the destination can only receive information from the source through the relay nodes, and the Eve cannot intercept any information from the broadcast channel between the source and the intermediate nodes [6]. Consequently, the SC is solely determined by the cooperative channel.

### *2.3. Multi-Hop Communication Process*

The end-to-end information transmission from the source $S$ to each designated destination $D_n$ is characterized by a sequential multi-hop transmission procedure structured into $M_n$ consecutive time slots. This process initiates with a broadcast phase, where the source node $S$, designated as $n_{0,n}$, transmits the confidential signal $x_1$ to the primary relay $n_{1,n}$ within the predefined path $\mathcal{P}_n$. To mitigate the eavesdropping risk from both the $Q$ Eves $\mathcal{E}$ and the idle untrusted relays within the pool $\mathcal{R}$, a subset of nodes $\mathcal{J}$ is dynamically activated as cooperative jammers. These jammers emit jamming signal $s_j$ concurrently with the source's broadcast, effectively degrading the signal-to-interference-plus-noise ratio (SINR) at any unauthorized tapping point while the first relay attempts to reliably decode the transmission.

Following the initial broadcast, the communication enters the intermediate relaying phases, which span from hop $i = 2$ to $M_n - 1$. Each time slot $i$ within this sequence utilizes a DF protocol, where the $i$-th node in the path which acted as the receiver in the $(i-1)$-th slot first decodes the confidential message received previously. Upon successful decoding, this relay forwards the re-encoded signal to the subsequent node $n_{i,n}$ in the path sequence.

The final stage of the communication procedure is the delivery phase, occurring in the $M_n$-th time slot. In this phase, the final intermediate relay in the path $\mathcal{P}_n$ transmits the processed signal to the ultimate destination $D_n$, designated as $n_{M_n,n}$. Similar to the preceding hops, this concluding transmission is conducted under the protection of cooperative jamming to ensure that the cumulative SC is not compromised at the terminal link.

### *2.4. Signal Model at the i-th Hop*

Within the established multi-hop network, the transmission from the source to each destination $D_n$ is decomposed into a sequence of discrete relaying intervals. We consider the $i$-th hop within the specific path $\mathcal{P}_n$, where node $u = n_{i-1,n}$ serves as the active transmitter (either the source or a preceding relay) and $v = n_{i,n}$ serves as the intended legitimate receiver. In this complex wireless environment, the transmission is characterized by the dynamic interplay between the confidential signal, the artificial interference generated by cooperative jammers $\mathcal{J}$, and ubiquitous environmental noise.

- Legitimate Signal Model
  During the $i$-th time slot, the signal received at the legitimate node $v$ is modeled as a superposition of the desired information-bearing signal and the aggregated jamming signals designed to confuse potential eavesdroppers. The received signal $y_{v,i}^{(n)}$ is mathematically expressed as:
  $$y_{v,i}^{(n)} = \sqrt{P_u} h_{u,v} x_i + \sum_{j \in \mathcal{J}} \sqrt{P_j} h_{j,v} s_j + v_v, \tag{5}$$
  where $P_u$ and $P_j$ denote the instantaneous transmit powers of the transmitting node and the $j$-th jammer, respectively. The term $h_{u,v}$ represents the complex channel fading coefficient between the transmitter and the receiver, while $h_{j,v}$ represents the channel coefficient between the $j$-th jammer and the receiver. $x_i$ is the normalized confidential message symbol. The summation term accounts for the deliberate interference $s_j$ emitted by the cooperative jammers, which serves to degrade the eavesdropping channels. Finally, $v_v \sim \mathcal{CN}(0, \sigma^2)$ represents the additive white Gaussian noise at the receiver.
- External Eavesdropping Model
  To account for external security threats, we assume the presence of $Q$ non-colluding Eves that monitor the transmission medium from various geographical locations. For any specific Eve $E_q$, the intercepted signal is subject to the same broadcast environment but experiences different channel fading:
  $$y_{E_q,i}^{(n)} = \sqrt{P_u} h_{u,E_q} x_i + \sum_{j \in \mathcal{J}} \sqrt{P_j} h_{j,E_q} s_j + v_{E_q}. \tag{6}$$
  The effectiveness of the secrecy strategy depends on the ability of the jammers to ensure that the interference power at $E_q$ significantly outweighs the signal power, thereby minimizing the information leakage.
- Internal Untrusted Relay Eavesdropping
  A unique challenge in this network is the untrusted nature of the relay pool $\mathcal{R}$. Any relay node $R_k \in \mathcal{R}$ that is not explicitly assigned as the transmitter or receiver for the current hop $i$ is regarded as a potential internal eavesdropper. These nodes may attempt to decode the message they are tasked to forward in other slots. The signal intercepted by an untrusted relay $R_k$ is modeled as:
  $$y_{R_k,i}^{(n)} = \sqrt{P_u} h_{u,R_k} x_i + \sum_{j \in \mathcal{J}} \sqrt{P_j} h_{j,R_k} s_j + v_{R_k}. \tag{7}$$
  This modeling approach necessitates a sophisticated power allocation strategy. While increasing $P_j$ can effectively jam Eves, it may inadvertently interfere with the legitimate receiver $v$ or provide varying levels of protection against different untrusted relays depending on their spatial distribution.

### *2.5. Secrecy Capacity Formulation*

Based on the multi-destination and multi-eavesdropper signal models established in the previous section, the secrecy performance of each hop must be rigorously quantified to account for the heterogeneous risks posed by internal and external adversaries.

- Instantaneous Hop Channel Capacity
  For the $i$-th hop of the $n$-th path, where node $u$ transmits to node $v$, the legitimate channel capacity $C_{v,i}^{(n)}$ is derived from the Shannon capacity formula, considering the interference from the cooperative jammers $\mathcal{J}$:

$$C_{v,i}^{(n)} = \log_2\left(1 + \frac{P_u g_{u,v}}{\sigma^2 + \sum_{j\in\mathcal{J}} P_j g_{j,v}}\right), \tag{8}$$

where $g_{a,b} = |h_{a,b}|^2$ represents the instantaneous channel power gain between any two nodes $a$ and $b$ at time $t$ modeled in Section 2.1.

Simultaneously, the information leakage risk is characterized by the maximum achievable rate among all potential Eves. In this scenario, the total information leakage rate $C_{leak,i}^{(n)}$ is determined by the most favorable channel conditions available to either Eves or idle untrusted relays:

$$C_{leak,i}^{(n)} = \max\left\{\max_{q\in E} \log_2\left(1+\Gamma_{E_q,i}^{(n)}\right), \max_{R_k\in R_{\{u,v\}}} \log_2\left(1+\Gamma_{R_k,i}^{(n)}\right)\right\}, \tag{9}$$

where $\Gamma_{E,i}^{(n)}$ and $\Gamma_{int,i}^{(n)}$ denote the corresponding SINR at the Eve $E_q$ and the idle untrusted relay $R_k$, respectively.

- End-to-End Secrecy Capacity

  Due to the utilization of the DF protocol across the multi-hop architecture, the end-to-end performance is subject to the inter-hop performance interdependencies. Specifically, the end-to-end transmission rate is restricted by the link with the minimum instantaneous capacity, often referred to as the restrictive throughput constraint of the multi-hop chain. Consequently, the end-to-end SC for destination $D_n$, denoted as $C_{S,n}$, is defined as the difference between the bottleneck legitimate rate and the peak leakage rate encountered along the entire path:

$$C_{S,n} = \min_{i\in\{1,\ldots,M_n\}} C_{v,i}^{(n)} - \max_{i\in\{1,\ldots,M_n\}} C_{leak,i}^{(n)}, \tag{10}$$

- SC Maximization Problem

  The ultimate goal of the system is to determine an optimal policy $\pi$ that maximizes the expected discounted sum of SC over the entire communication duration $T$ for all $N$ Bobs while strictly adhering to security and power constraints:

$$\max_{\pi} \quad \mathbb{E}_\pi\left[\sum_{t=0}^{T}\gamma^t \sum_{n=1}^{N} C_{S,n}(t)\right] \tag{11a}$$

$$\text{s.t.} \quad C_{S,n}(t) > 0, \forall n, \forall t, \tag{11b}$$

$$0 \le P_u(t) \le P_{max,u},\ 0 \le P_j(t) \le P_{max,j}, \forall t, \tag{11c}$$

  where $\gamma \in [0,1]$ is the discount factor representing the importance of future rewards. This formulation ensures that the proposed framework not only maximizes throughput but also maintains a persistent security guarantee across the entire multi-hop network.

## 3. Constrained Markov Decision Process Formulation

In this section, the secure resource allocation problem in multi-hop, multi-destination networks is modeled as a CMDP. This framework is essential because traditional MDPs cannot inherently guarantee the strict requirements $C_{S,n}(t) > 0$ necessary for PLS. The CMDP is defined by the quintuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \pi, r, c)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is hybrid action space, $\mathcal{P}$ is the state transition probability, $\pi$ is the policy, $r$ is the reward function, and $c$ is the cost function.

### *3.1. Comprehensive State Space ($\mathcal{S}$)*

The state space $\boldsymbol{s}_t \in \mathcal{S}$ is designed to provide a Markovian representation of the environment, encapsulating the high-dimensional CSI and the hardware-specific constraints of the pool. The state vector at time $t$ is defined as:

$$\boldsymbol{s}_t = \{\mathbf{G}_{leg}(t), \mathbf{G}_{int}(t), \mathbf{G}_{ext}(t), \mathbf{P}_{max}, \boldsymbol{\Phi}(t)\}. \tag{12}$$

The specific expressions for each component of Equation (12) are detailed as follows:

- Legitimate Channel Gain Set $\mathbf{G}_{leg}(t)$
  This set contains the instantaneous channel power gains for every hop in the predefined paths $\mathcal{P}_n$ for all $N$ Bobs. In the RL setting, the agent cannot access instantaneous CSI and must rely on delayed CSI from the previous time slot. Thus, the state space is defined as the historical channel gains between all node pairs, expressed as:

$$\mathbf{G}_{leg}(t) = \left\{ g_{n_{i-1,n}, n_{i,n}}(t-1) \mid n = 1, \ldots, N; i = 1, \ldots, M_n \right\}, \tag{13}$$

  where $g_{u,v}(t) = |h_{u,v}(t)|^2$ is the gain between node $u$ and $v$ modeled in Section 2.1. The total dimension of this set is $\sum_{n=1}^{N} M_n$, representing the full connectivity status of the legitimate network.
- Internal Leakage Channel Matrix $\mathbf{G}_{int}(t)$
  This component characterizes the channel gains between the current transmitters and all potential internal eavesdroppers. For a transmitter $u$ in the $i$-th hop, it is defined as:

$$\mathbf{G}_{int}(t) = \left[ g_{u,R_k}(t-1) \right] \in \mathbb{R}^{1 \times |\mathcal{R}_{\{u,v\}}|}, \tag{14}$$

  where $R_k \in \mathcal{R}$ and $R_k \neq u, v$. This allows for the agent to evaluate the risk of information interception by the very nodes meant to assist the network.
- External Leakage Channel Matrix $\mathbf{G}_{ext}(t)$
  This matrix captures the channel gains from both the active transmitters and the selected jammers to the $Q$ Eves:

$$\mathbf{G}_{ext}(t) = \begin{bmatrix} g_{u,E_1}(t-1) & g_{u,E_2}(t-1) & \cdots & g_{u,E_Q}(t-1) \\ g_{j_1,E_1}(t-1) & g_{j_1,E_2}(t-1) & \cdots & g_{j_1,E_Q}(t-1) \\ \vdots & \vdots & \ddots & \vdots \\ g_{j_{|\mathcal{J}|},E_1}(t-1) & g_{j_{|\mathcal{J}|},E_2}(t-1) & \cdots & g_{j_{|\mathcal{J}|},E_Q}(t-1). \end{bmatrix} \tag{15}$$

  This $(|\mathcal{J}|+1) \times Q$ matrix is vital for the lower-level controller to balance jamming power against signal leakage.
  It should be noted that neither Eves nor Bobs can access exact instantaneous CSI during decision-making due to the highly dynamic time-varying Rayleigh fading. Instead, the agent can only observe the delayed CSI from the previous time slot [30,31], which is treated as part of the environment state. This modeling choice maintains the Markov property while avoiding the unrealistic assumption of obtaining instantaneous CSI from Eves.
- Heterogeneous Power Constraint Vector $\mathbf{P}_{max}$
  This is a static hardware-profile vector that informs the agent of the physical limitations of each node in the pool:

$$\mathbf{P}_{max} = [P_{max,R_1}, P_{max,R_2}, \ldots, P_{max,R_K}]^T. \tag{16}$$

Including this in the state space allows for the RL model to generalize its power allocation policy across nodes with different battery levels or hardware capabilities.

- Topology and Progress Indicator $\mathbf{\Phi}(t)$
  This vector tracks the real-time status of the multi-hop process and node availability:

$$\mathbf{\Phi}(t) = [i_1(t), i_2(t), \ldots, i_N(t), \mathbf{I}_{occ}(t)], \tag{17}$$

  where $i_n(t) \in \{1, \ldots, M_n\}$ is the current hop index for destination $D_n$. $\mathbf{I}_{occ}(t) \in \{0,1\}^K$ is the node occupancy vector, where $I_k = 1$ indicates that relay $R_k$ is currently busy forwarding data and cannot be selected as a jammer.

### *3.2. Hierarchical Action Space $\mathcal{A}$*

Due to the combinatorial complexity of selecting optimal paths and jammers alongside the continuous nature of power control, the action space is partitioned into two distinct sub-spaces.

- Discrete Joint Relay and Jammer Selection ($a_t^D$)
  The upper-level policy, acting as a meta-controller, manages the network topology by making discrete selections from the candidate pool $\mathcal{R}$. The action is a composite decision process:
  Relay Selection: Based on the delayed legitimate channel gain set $\mathbf{G}_{leg}(t)$, the agent identifies the optimal relay nodes $\{n_{i,n}\}$ to form the multi-hop transmission path $\mathcal{P}_n$ for each Bob.
  Jammer Selection: From the subset of idle nodes $\mathcal{F}(t) \subset \mathcal{R}$, where $I_k = 0$ in the node occupancy vector, the policy selects a binary vector $z_t \in \{0,1\}^K$ to designate cooperative jammers $\mathcal{J}(t)$. These nodes are strategically chosen to degrade the SINR at both external and internal eavesdropping points.
- Continuous Power Allocation ($a_t^C$)
  The lower-level policy determines the precise power levels for the selected relays and jammers. The action is a continuous vector $\mathbf{p}_t = [P_u(t), P_{j,1}(t), \ldots, P_{j,|\mathcal{J}|}(t)]$, where $P_u$ is the power of the active transmitter in path $\mathcal{P}_n$ and $P_j$ represents the power of the $j$-th jammer.

### *3.3. Transition Probability ($\mathcal{P}$)*

The transition probability $\mathcal{P}(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, a_t)$ defines the evolution of the environment. In the context of a multi-hop secure network, this transition is governed by two distinct processes.

- Wireless Channel Dynamics
  The channel gains $\mathbf{G}_{leg}(t)$, $\mathbf{G}_{int}(t)$, and $\mathbf{G}_{ext}(t)$ are governed by independent stochastic processes. While these transitions follow the temporal correlation defined by SDEs, they remain exogenous and highly dynamic, rendering the instantaneous channel power gains unpredictable for the agent at the time of decision-making. Consequently, the agent must rely on delayed CSI from the previous time slot.
- Protocol-Driven State Evolution
  The topology and progress indicator $\mathbf{\Phi}(t)$ evolves deterministically based on the agent's actions. When a transmission for destination $D_n$ successfully completes a hop, meaning that the SC constraint is satisfied and the message is correctly decoded, the hop index $i_n$ increments from $i$ to $i+1$. Once $i_n = M_n$, the node occupancy vector $\mathbf{I}_{occ}$ is updated to release the nodes associated with that path back into the pool $\mathcal{R}$.

### *3.4. Policy Mapping ($\pi$)*

Due to the hybrid nature of the action space (discrete selection and continuous power), the policy $\boldsymbol{\pi}$ is decomposed into a hierarchical structure $\boldsymbol{\pi} = \{\pi^{upper}, \pi^{low}\}$.

- Upper-Level Policy: Node Selection ($\pi^{upper}$)
  The upper-level policy, typically governed by a discrete RL agent, maps the current state $\boldsymbol{s}_t$ to the discrete jammer selection action $a_t^D$:

$$\pi^{upper}(\boldsymbol{s}_t) \rightarrow \mathbf{z}_t, \quad \mathbf{z}_t \in \{0,1\}^K. \tag{18}$$

  This policy focuses on the long-term topological security of the network. It identifies which idle untrusted relays $R_k$ are strategically positioned to provide the most effective jamming for the current set of active hops while ensuring they are not wasted on hops that already have high intrinsic security.
- Lower-Level Policy: Power Allocation ($\pi^{low}$)
  Given the state $\boldsymbol{s}_t$ and the discrete action $\mathbf{z}_t$ chosen by the upper level, the lower-level policy determines the continuous power levels:

$$\pi^{low}(\boldsymbol{s}_t, \mathbf{z}_t) \rightarrow \mathbf{p}_t, \quad \mathbf{p}_t = [P_u(t), P_{j,1}(t), \ldots, P_{j,|\mathcal{J}|}(t)]. \tag{19}$$

### *3.5. Reward and Cost Functions ($r, c$)*

The objective of the agent is not merely to maximize SC, but to do so within the safe operating region of the network.

- Objective Reward ($r_t$)
  The primary reward is the sum of the end-to-end SC across all Bobs. It incentivizes the agent to maximize the gap between the bottleneck legitimate rate and the peak leakage rate:

$$r_t = \sum_{n=1}^{N} C_{S,n}(t). \tag{20}$$

- Security Cost ($\boldsymbol{c}_t$)
  To enforce the constraint $C_{S,n}(t) > 0$ and to satisfy the physical transmission constraints of nodes, we define a set of cost functions consisting of three independent dimensions. This multi-dimensional cost mechanism is designed to apply precise quantitative penalties to any state–action pairs that induce information eavesdropping or exceed the power limits of transmitters and jammers, thereby enforcing the security and physical stability of the system within the algorithmic framework.
  Specifically, this multi-dimensional cost function set includes the following three core dimensions:
  Security Outage Cost ($c_{1,t}$): Specifically used to monitor and penalize decisions that lead to information eavesdropping. For any destination node $D_n$, this cost is defined using the mathematical indicator function $\mathbb{I}(\cdot)$ as:

$$c_{1,t}^{(n)} = \mathbb{I}(C_{S,n}(t) \leq 0). \tag{21}$$

  Transmitter Power Limit Cost ($c_{2,t}$): Used to constrain the transmission power of the active transmitter node $u$ so that it does not exceed its hardware physical limit $P_{\max,u}$. It is defined as the truncation amount exceeding the maximum power threshold:

$$c_{2,t} = \max(0, P_u - P_{\max,u}). \tag{22}$$

Jammer Power Limit Cost ($c_{3,t}$): Similarly, used to constrain the transmission power of the cooperative jammer $j$ so that it does not exceed its hardware upper limit $P_{\max,j}$. It is defined as:

$$c_{3,t} = \max(0, P_j - P_{\max,j}). \tag{23}$$

## 4. Constrained Hierarchical Reinforcement Learning Framework

Building upon the CMDP established in the previous section, this section proposes a novel CHRL framework to address the multi-objective optimization problem in the proposed secure cooperative networks.

### *4.1. Framework Overview*

To address the joint optimization of node selection and power control in a multi-hop, multi-destination network, the CHRL framework is proposed, which is illustrated in Figure 4. This framework decomposes the original mixed-integer non-convex optimization problem into a two-level hierarchy on two distinct timescales. The upper-level manager, acting as a meta-controller, makes macro-decisions regarding node selection every $n$ time steps. The lower-level executor, functioning as a controller, operates at each time step to make micro-decisions on precise power allocation based on the upper level's macro-instruction and the current channel state. This decomposition effectively reduces the combinatorial complexity of the action space and enables more efficient exploration.

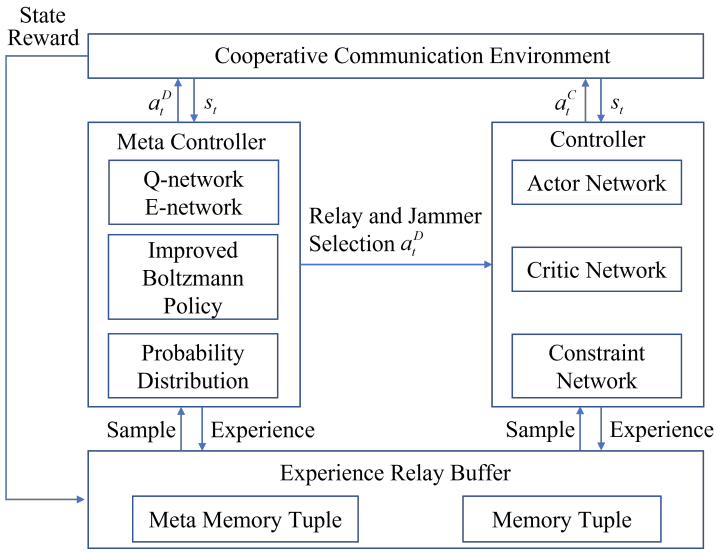


**Figure 4.** The proposed constrained hierarchical reinforcement learning (CHRL) framework for node selection and power allocation.

### *4.2. Upper Level: Risk-Aware Node Selection Policy*

The upper-level controller functions as a strategic coordinator, responsible for selecting the optimal subset of cooperative jammers $\mathcal{J}(t)$ from the pool of idle untrusted relays $\mathcal{R}$.

The upper-level policy $\pi^{upper}$ focuses on the macro-management of the network topology. It selects the jammer set $\mathcal{J}$ to maximize the expected long-term return $V^{upper}(s)$. The policy is defined as:

$$y_t^{Q^{upper}} = \sum_{i=0}^{n} \gamma^i r_t + \gamma^n \max_{a_{t+n}^{D'}} Q^{upper}(\boldsymbol{s}_{t+n}, a_{t+n}^{D'}; \theta^{Q^{upper'}}), \tag{24}$$

where the discrete action $a_t^D$ represents the selection of relays and jammers. A Q-value function $Q^{upper}(\boldsymbol{s}_t, a_t^D)$ estimates the long-term expected SC when taking a particular node

selection action in a given state. It is updated using a temporal difference (TD) error based on the aggregate reward $r_t$. The target value $y_t^Q$, the loss function $L(\theta^Q)$, and the update of network parameter $\theta_Q^{upper}$ are defined as:

$$y_t^{Q^{upper}} = \sum_{i=0}^{n} \gamma^i r_{t+i} + \gamma^n \max_{a_{t+n}^{D'}} Q^{upper}(\boldsymbol{s}_{t+n}, a_{t+n}^{D'}; \theta^{Q^{upper'}}), \tag{25}$$

$$L(\theta^{Q^{upper}}) = \mathbb{E}\Big[(y_t^{Q^{upper}} - Q^{upper}(\boldsymbol{s}_t, a_t^D; \theta^{Q^{upper}}))^2\Big], \tag{26}$$

$$\theta^{Q^{upper}} \leftarrow \theta^{Q^{upper}} - \alpha \nabla_{\theta^{Q^{upper}}} L(\theta^{Q^{upper}}). \tag{27}$$

$\gamma^n$ is the discounted factor over $n$ time steps, reflecting the macro-timescale of the meta-controller. $\theta^{Q^{upper'}}$ is the target Q-network used to stabilize training. $\alpha$ is the network learning rate.

Complementarily, an E-value function $E^{upper}(\boldsymbol{s}_t, a_t^D)$ estimates the long-term risk exposure, specifically quantifying the probability of violating the security constraint ($C_{S,n}(t) \leq 0$) over time. The E-values are updated based on the $n$ future risk values as:

$$y_t^{E^{upper}} = \sum_{i=1}^{n} \gamma^i c_{t+i} + \gamma^n \max_{a_{t+n}^{D'}} E^{upper}(\boldsymbol{s}_{t+n}, a_{t+n}^{D'}; \theta^{E^{upper'}}), \tag{28}$$

$$L(\theta^E) = \mathbb{E}\Big[(y_t^E - E^{upper}(\boldsymbol{s}_t, a_t^D; \theta^{E^{upper}}))^2\Big], \tag{29}$$

$$\theta^{E^{upper}} \leftarrow \theta^{E^{upper}} - \alpha \nabla_{\theta^{E^{upper}}} L(\theta^{E^{upper}}). \tag{30}$$

An improved Boltzmann policy integrates both estimates to prioritize actions with high reward and low risk. To select the discrete action $a_t^D$, the outputs of both networks are fused into a risk-aware probability distribution. The probability $P(a_i)$ for a specific jammer combination $a_i$ is calculated as:

$$P(a_i^D) = \frac{\exp\left(\frac{Q^{upper}(\boldsymbol{s}_t, a_i^D) - \sum \lambda_n E^{upper}(\boldsymbol{s}_t, a_i^D)}{\tau}\right)}{\sum_{a_j^D \in a_i^D} \exp\left(\frac{Q^{upper}(\boldsymbol{s}_t, a_j^D) - \sum \lambda_n E^{upper}(\boldsymbol{s}_t, a_j^D)}{\tau}\right)}, \tag{31}$$

where $\tau > 0$ serves as a temperature parameter that regulates the exploration–exploitation trade-off.

### *4.3. Lower Level: Constrained Power Allocation Policy*

The lower-level power allocation problem is inherently a competitive scenario, where the friendly jammer's signal acts as a double-edged sword: it degrades the Eve's channel while simultaneously causing potential interference to the legitimate receiver. To address this challenge, we model the power allocation as a constrained zero-sum game between the communication system and a virtual adversary that represents the worst-case channel conditions. This formulation aligns with the M3DDPG framework, whose core objective is to maximize the expected return under the most pessimistic assumptions about the environment or the opponent's behavior.

The original constrained optimization problem Equation (11a) is reformulated using Lagrange multipliers $\boldsymbol{\lambda}$ to penalize constraint violations, leading to the Lagrangian dual problem:

$$\min_{\boldsymbol{\lambda} \geq 0} \max_{\pi} \mathcal{L}(\pi, \boldsymbol{\lambda}), \tag{32}$$

where $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \lambda_3]^T$, and the Lagrangian $\mathcal{L}(\pi, \boldsymbol{\lambda})$ is given by:

$$\mathcal{L}(\pi, \boldsymbol{\lambda}) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{T} \gamma^t \left( \underbrace{r_t}_{\text{Reward}} - \sum_{i=1}^{C} \lambda_i \underbrace{c_{i,t}}_{\text{Cost}_i} \right)\right]. \tag{33}$$

The number of constraints is set to $C = 3$ in this paper, following Equations (21)–(23). This formulation aligns with the M3DDPG's MiniMax objective Equation (11a) by interpreting the penalized costs as adversarial contributions to the reward function, effectively training the agent against worst-case constraint violations.

The constrained M3DDPG architecture employs an Actor network $\mu(\boldsymbol{s}_t|\theta^{\mu})$ to generate power allocation actions [29] and a Critic network $Q(\boldsymbol{s}_t, a_t^C|\theta^Q)$ that estimates state–action values using TD learning. To align with the MiniMax objective, the target value $y_t$ is calculated by minimizing over a virtual adversary's perturbation $\psi$ to simulate worst-case conditions:

$$y_t = \left(r_t - \sum_{i=1}^{C} \lambda_i c_{i,t}\right) + \gamma \min_{\psi} Q'\left(\boldsymbol{s}_{t+1}, \mu'(\boldsymbol{s}_{t+1}|\theta^{\mu'}) + \psi \mid \theta^{Q'}\right). \tag{34}$$

The network parameters $\theta^Q$ are updated by minimizing the loss:

$$L(\theta^Q) = \mathbb{E}\left[\left(y_t - Q(\boldsymbol{s}_t, a_t^C|\theta^Q)\right)^2\right], \tag{35}$$

$$\theta^Q \leftarrow \theta^Q - \alpha \nabla_{\theta^Q} L(\theta^Q). \tag{36}$$

Three separate Constraint networks $C_i(\boldsymbol{s}_t, a_t^C|\theta^{C_i})$ estimate the expected cumulative future cost for each constraint $i \in \{1, 2, 3\}$. The target value $y_{c,i}$ is:

$$y_{c,i} = c_{i,t} + \gamma C_i'\left(\boldsymbol{s}_t, \mu'(\boldsymbol{s}_t|\theta^{\mu'}) \mid \theta^{C_i'}\right). \tag{37}$$

The parameters $\theta^{C_i}$ are updated by minimizing the mean squared error:

$$L(\theta^{C_i}) = \mathbb{E}\left[\left(y_{c,i} - C_i(\boldsymbol{s}_t, a_t^C|\theta^{C_i})\right)^2\right], \tag{38}$$

$$\theta^{C_i} \leftarrow \theta^{C_i} - \alpha \nabla_{\theta^{C_i}} L(\theta^{C_i}). \tag{39}$$

The Actor network $\mu(\boldsymbol{s}_t|\theta^{\mu})$ determines the continuous power allocation vector $a_t^C$. The update rule performs gradient ascent on the Lagrangian objective, balancing the maximization of $Q$ and the minimization of constraint costs $C_i$:

$$\nabla_{\theta^{\mu}} J \approx \frac{1}{B} \sum_{j=1}^{B} \nabla_{\theta^{\mu}} \mu(\boldsymbol{s}_j) \cdot \left[\nabla_a Q(\boldsymbol{s}_j, a) - \sum_{i=1}^{C} \lambda_i \nabla_a C_i(\boldsymbol{s}_j, a)\right]\Bigg|_{a=\mu(S_j)}, \tag{40}$$

$$\theta^{\mu} \leftarrow \theta^{\mu} + \alpha \nabla_{\theta^{\mu}} J, \tag{41}$$

where $B$ is the batch size. In M3DDPG, the network is not updated using just the single most recent experience. Instead, a random set of transitions is sampled from the replay buffer $\mathcal{D}_{low}$ to break the temporal correlation between consecutive steps. $\boldsymbol{s}_j$ is the state component of the $j$-th transition in this random batch.

The multipliers are updated to enforce the constraints. If the expected violation estimated by the Constraint network is positive, the penalty weight $\lambda_i$ is updated by:

$$\lambda_i \leftarrow \Gamma_{\Lambda}[\lambda_i + \eta_{\lambda} \cdot \mathbb{E}[C_i(\boldsymbol{s}_t, \mu(\boldsymbol{s}_t))]], \tag{42}$$

where $\Gamma_\Lambda[x] = \max(0, x)$ ensures non-negativity.

### *4.4. Joint Learning Algorithm*

The complete hierarchical learning algorithm integrates both levels through coordinated updates and experience sharing, as summarized in Algorithm 1.

**Algorithm 1** Joint CHRL for Secure Cooperative Communication

**Require:** Environment, Network Architectures, Hyperparameters $(\gamma, \tau, \eta_\lambda, B, \beta)$

**Initialize:**
Upper-level networks: $Q^{upper}, E^{upper}$ with weights $\theta^{Q^{upper}}, \theta^{E^{upper}}$
Lower-level networks: Actor $\mu$, Critic $Q$, Constraint Critics $C_1, C_2, C_3$ with weights $\theta^\mu, \theta^Q, \theta^{C_i}$
Target networks: $\theta^{\mu'}, \theta^{Q'}, \theta^{C_i'} \leftarrow \theta^\mu, \theta^Q, \theta^{C_i}$
Replay buffers $\mathcal{D}_{upper}, \mathcal{D}_{low}$
Lagrange multipliers $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \lambda_3]^T \leftarrow \mathbf{0}$
**for** each iteration **do**
  Initialize environment and observe initial state $S_1$
  **for** $t = 1$ to $T$ **do**
    **if** $t \mod n == 1$ **then**
      Observe upper state $\boldsymbol{s}_t$
      Select discrete joint action $a_t^D$ by Equation (31)
    **end if**
    Construct lower state $\boldsymbol{s}_t$
    Select continuous power action $a_t^C$
    Execute joint action $(a_t^D, a_t^C)$
    Observe reward $r_t$, constraint costs $\mathbf{c}_t = [c_1, c_2, c_3]$, and next state $\boldsymbol{s}_{t+1}$
    Store transition $(\boldsymbol{s}_t, a_t^C, r_t, c_t, \boldsymbol{s}_{t+1})$ in $\mathcal{D}_{low}$
    **if** samples sufficient in $\mathcal{D}_{low}$ **then**
      Sample random minibatch of $B$ transitions from $\mathcal{D}_{low}$
      Update $\theta^Q$ by Equation (36)
      Update $\theta^{C_i}$ by Equation (39)
      Update $\theta^\mu$ by Equation (41)
      Update Multipliers by Equation (42)
      Soft Update Targets: $\theta' \leftarrow \rho\theta + (1-\rho)\theta'$ for all networks
    **end if**
    **if** $t \mod n == 0$ **then**
      Store transition in $\mathcal{D}_{upper}$
      Update $Q^{upper}$ and $E^{upper}$ networks using stored experiences
    **end if**
  **end for**
**end for**

### *4.5. Computational Complexity Analysis*

To demonstrate the practical feasibility of the proposed CHRL framework for real-time implementation in 6G cooperative networks, we quantitatively analyze its computational complexity from the perspectives of space dimensionality and time execution overhead [32] and compare its computational performance with traditional HRL methods [25].

Dimensionality Analysis: By employing hierarchical decoupling, the proposed framework successfully mitigates this combinatorial explosion. The upper-level discrete action space is reduced to $\mathcal{O}(K)$, and the lower-level continuous action space is reduced to $\mathcal{O}(1 + |\mathcal{J}|)$. Furthermore, the dimension of the comprehensive state space $\boldsymbol{s}_t$, denoted as

$D_s$, is defined by the concatenation of $\mathbf{G}_{leg}(t)$, $\mathbf{G}_{int}(t)$, $\mathbf{G}_{ext}(t)$, $\mathbf{P}_{max}$, and $\mathbf{\Phi}(t)$. The total dimension evaluates to $D_s = \sum_{n=1}^{N} M_n + 2K + N - 2 + (|\mathcal{J}| + 1)Q$.

Time Complexity Analysis: Assuming that all employed neural networks consist of $L$ hidden layers with $H$ neurons per layer, the computational time complexity must be evaluated separately for the offline training and online execution phases due to the nature of the Actor–Critic architecture.

During the offline training phase, the algorithm updates all seven networks (upper $Q$, upper $E$, lower Actor $\mu$, lower Critic $Q$, and three Constraint networks $C_i$) via backpropagation using minibatches of size $B$. The computational overhead per training step is approximately $\mathcal{O}(B \times [D_s H + LH^2 + H(K + |\mathcal{J}|)])$. This intensive computation is designated to be executed on centralized base stations or cloud servers a priori.

Conversely, during the online execution phase, the system only requires forward propagation to make real-time decisions. Crucially, the lower Critic and Constraint networks are strictly bypassed, as they solely serve as evaluators during training. The active networks during inference are limited to the upper macro-controllers and the lower Actor network. Consequently, the online execution complexity is drastically condensed to $\mathcal{O}(D_s H + LH^2 + H(K + |\mathcal{J}|))$. This polynomial-time complexity firmly establishes that the proposed CHRL framework imposes minimal computational delay during online deployment, guaranteeing rapid adaptation to the highly dynamic time-varying Rayleigh fading channels without compromising the strict PLS requirements.

Using the same method, we analyze the computational complexity of conventional HRL methods [25]. Existing HRL approaches critically rely on the discretization of the continuous action space to formulate their lower-level decision-making. Assuming the transmit power is uniformly discretized into $M$ discrete levels, the combinatorial action space for one active transmitter and $|\mathcal{J}|$ jammers scales exponentially as $\mathcal{O}(M^{1+|\mathcal{J}|})$. Consequently, the online execution complexity of conventional HRL severely expands to $\mathcal{O}(D_s H + LH^2 + H \cdot M^{1+|\mathcal{J}|})$. This exponential growth introduces a dilemma: employing coarse discretization (a small $M$) inevitably leads to sub-optimal power allocation and massive performance degradation, whereas fine-grained discretization (a large $M$) triggers the curse of dimensionality, rendering real-time computation entirely infeasible. By deploying an Actor network to directly output continuous actions linearly scaled at $\mathcal{O}(1 + |\mathcal{J}|)$, our proposed CHRL elegantly sidesteps this discrete combinatorial explosion, guaranteeing both theoretically optimal power allocation and strict computational feasibility for real-time 6G deployments.

## 5. Numerical Results

In this section, we evaluate the performance of the proposed CHRL framework. The setting of simulation parameters is shown in the Table 1. The curves presented from Figures 5–8 are based on actual training data obtained from a single training run, and a moving average with a window size of 50 iterations is applied to the raw training data.

First of all, the learning rate of the optimizer updating network parameters should be set to an appropriate value. As shown in Figure 5, if the learning rate is too large (e.g., 0.1), it may cause the model to converge to a local optimum. Conversely, if the learning rate is too small, the convergence process will be significantly slower. For instance, with a learning rate of 0.0001, the model only converges around the 5000th iteration. We ultimately set the learning rate to 0.001 for the subsequent simulations.

**Table 1.** Experimental parameter settings.

| Parameter | Symbol | Value |
|---|---|---|
| Candidate intermediate nodes | $K$ | 100 |
| Number of Bobs | $N$ | 4 |
| Number of Eves | $Q$ | 4 |
| SDE channel paths | - | 100 |
| SDE drift and initial parameters | $B, I_0, Q_0$ | 1, 0, 0 |
| Max transmit power (relays/jammers) | $P_{max,u}, P_{max,j}$ | 20 dBm |
| Normalized noise variance | $\sigma^2$ | 1 |
| Independent channel realizations | - | $10^3$ |
| Communication duration | $T$ | 4s |
| Time scale of upper controller | $n$ | 10 |
| Learning rate of optimizer | $\alpha$ | 0.001 |
| Batch size | $B$ | 256 |
| Temperature coefficient of Boltzmann policy | $\tau$ | 1 |
| Update step size of Lagrange multipliers | $\eta_\lambda$ | 0.01 |
| Hidden layers of neural network | $L$ | 3 |
| Number of neurons in each layer | $H$ | 256 |

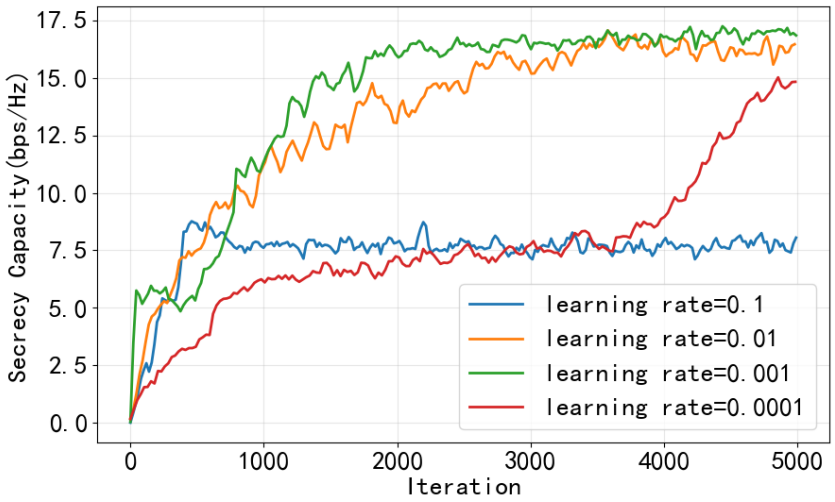


**Figure 5.** Sum secrecy capacity (SC) under different learning rates.

The influence of different batch sizes on the convergence performance during training is investigated in Figure 6. It can be observed that a small batch size fails to utilize all the data stored in the experience buffer and leads to slow convergence. In contrast, a large batch size, for instance, 256, results in the fastest convergence speed, albeit at the cost of increased training time per epoch.

To ensure the reliability of the proposed CHRL framework in practical deployments, it is imperative to investigate the robustness of key parameters, specifically the temperature coefficient $\tau$ in the upper-level Boltzmann policy and the update step size $\eta_\lambda$ for the lower-level Lagrange multipliers. Improper tuning of these parameters can disrupt the balance between reward maximization and risk avoidance, potentially trapping the network in sub-optimal local minima.

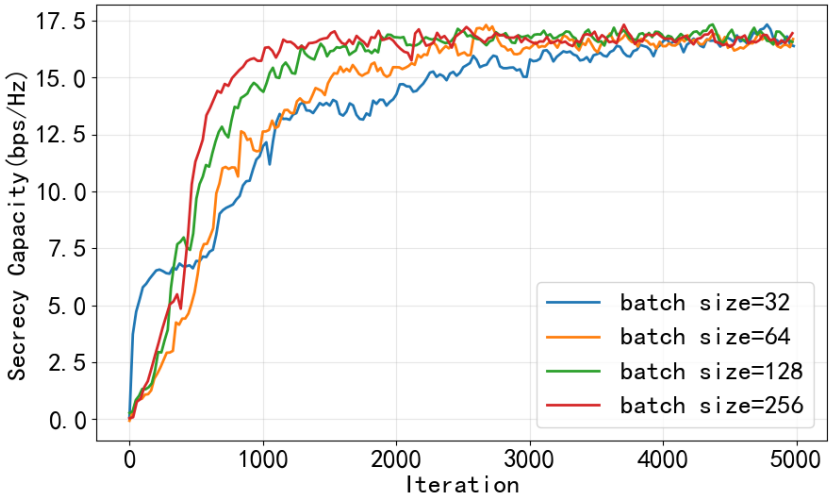


**Figure 6.** Sum SC under different batch sizes.

Figure 7 quantitatively demonstrates the sum SC versus iterations under various parameter settings.

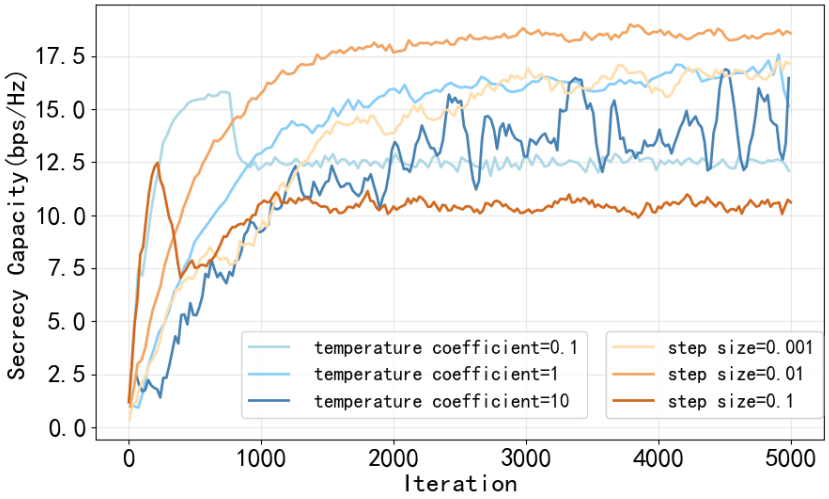


**Figure 7.** Sum SC versus the temperature coefficients and update step sizes of the Lagrange multiplier.

The blue curves in Figure 7 demonstrate that the parameter $\tau$ orchestrates the trade-off between maximizing the expected reward and minimizing the constraint violation risk. An excessively large temperature ($\tau = 10$) causes the agent to over-explore, resulting in high variance and severe oscillations throughout the training process. Conversely, if $\tau$ is too small ($\tau = 0.1$), the policy becomes overly greedy. In this scenario, the agent is severely inhibited from exploring potentially high-reward node combinations if the E-network marginally overestimates their risk during early iterations, causing a premature convergence to a sub-optimal topology, approximately 12.5 bps/Hz. The fine-tuned value of $\tau = 1$ optimally balances initial thorough exploration with subsequent stable exploitation.

The orange curves in Figure 7 indicate that the step size $\eta_\lambda$ dictates the sensitivity of the continuous power allocation policy to security constraint violations. When $\eta_\lambda$ is excessively large ($\eta_\lambda = 0.1$), a single security outage causes the penalty term to surge drastically. This massive penalty overrides the primary objective of maximizing SC, forcing the agent into a state of excessive inhibition. The agent learns to apply extreme, conservative continuous actions to avoid any conceivable risk, causing the sum SC to plummet sharply early in the training and lock into a sub-optimal baseline 10.5 bps/Hz. Alternatively, a very small step size ($\eta_\lambda = 0.001$) causes the penalty to update too slowly, dragging down the convergence

speed. The empirically chosen value of $\eta_\lambda = 0.01$ provides the optimal robustness, allowing the agent to gracefully learn the safety boundaries while achieving the highest stable sum SC 18.5 bps/Hz.

We then compare the proposed method against seven baselines: (1) HRL [25], (2) H-Q/E networks [23], (3) Flat M3DDPG [29], (4) fixed penalty with a small weight $\lambda = 1$, (5) fixed penalty with a large weight $\lambda = 10$, (6) game theory [6] and (7) random. As shown in Figure 8, in the conventional fixed penalty baseline, a static penalty weight $\lambda$ is directly incorporated into the reward function to penalize security constraint violations instead of using Lagrange multipliers. The reshaped reward $R_t$ is formulated as $R_t = r_t - \lambda \cdot \sum_{i=1}^{C} c_{i,t}$, where $r_t$ is the objective reward Equation (20) and $c_{i,t}$ is the immediate security cost Equations (21)–(23).

The simulation results in Figure 8 demonstrate that the proposed method consistently achieves the highest cumulative scores, stabilizing around 18.5 bps/Hz after approximately 1000 iterations. H-RL and H-Q/E networks follow closely, plateauing at 17.5 bps/Hz and 15.8 bps/Hz, respectively. The non-hierarchical M3DDPG struggles severely with the combinatorial explosion of the hybrid action space, converging extremely slowly to a suboptimal 12.4 bps/Hz, which firmly validates the necessity of stratification. Furthermore, replacing our dynamic Lagrange method with static penalty terms proves highly ineffective. The fixed penalty with small $\lambda$ baseline exhibits severe oscillations. It prioritizes short-term rewards but suffers from frequent security outages, failing to maintain a stable positive SC. Conversely, the fixed penalty with large $\lambda$ baseline induces excessive inhibition, forcing the agent into overly conservative power allocation strategies that lock the performance at a severely degraded 10.0 bps/Hz. Other conventional methods, such as the game theory and random baseline, show gradual improvement but remain significantly lower. Overall, the demonstrated figure shows the superior convergence and secrecy performance of the proposed approach in this iterative task.

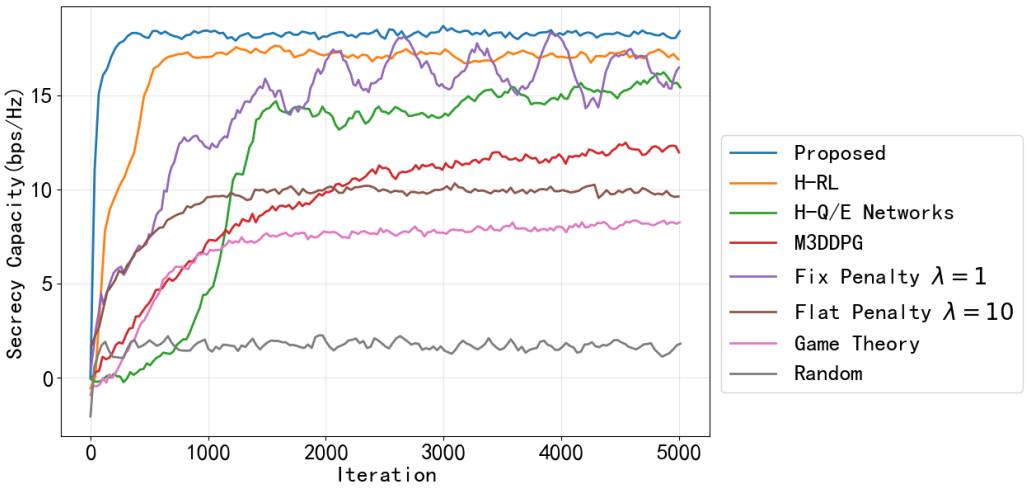


**Figure 8.** Sum SC using different methods.

Next, we investigate the impact of varying the number of Bobs and Eves. As shown in Figure 9, as the number of Bobs and Eves increases simultaneously from 1 to 10, the sum SC exhibits a monotonically increasing trend for all schemes, driven by the spatial multiplexing gains of serving more users. The proposed CHRL framework consistently outperforms the baselines, rising from approximately 10 to 23 bps/Hz and maintaining a significant performance gap over H-RL, H-Q/E networks, M3DDPG, and the game theory approach. Notably, the growth rate for the game learning-based methods gradually slows down as the network becomes denser.

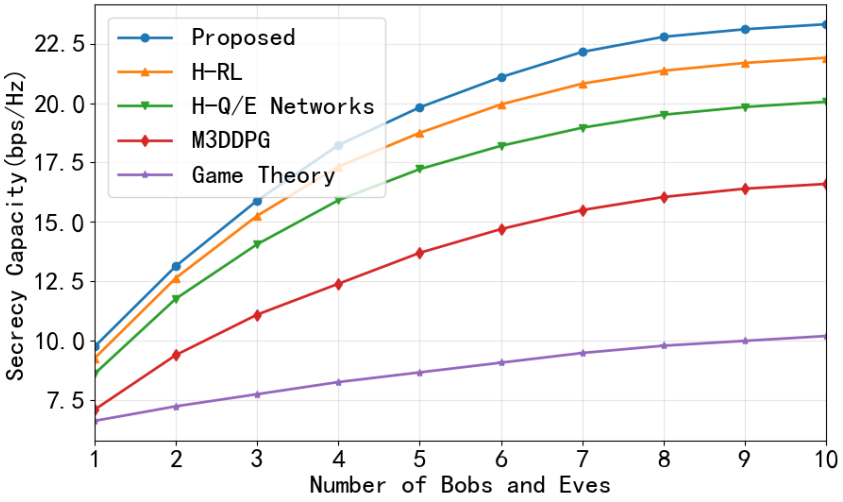


**Figure 9.** Sum SC versus number of Bobs and Eves.

To analyze the influence of node transmit power, we vary the power levels of relays serving Bobs and jamming nodes targeting. As shown in Figure 10, increasing the legitimate transmission power improves the legitimate channel capacity, thereby enhancing SC. The proposed CHRL framework consistently outperforms the baselines, exhibiting a steeper growth curve that reaches approximately 26.5 bps/Hz at 30 dBm, compared to roughly 24 bps/Hz for H-RL, 22 bps/Hz for H-Q/E networks, and 17 bps/Hz for M3DDPG.

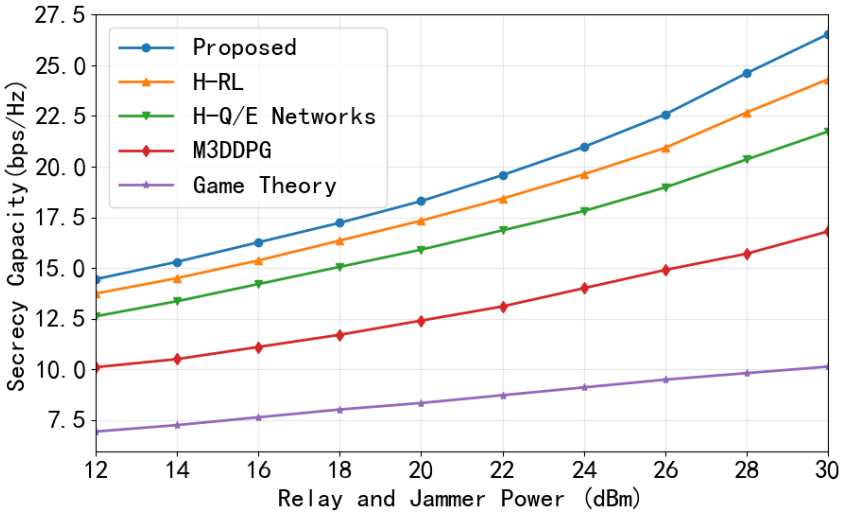


**Figure 10.** Sum SC versus relay and jammer transmit power.

To evaluate the efficacy of the proposed CHRL framework in satisfying security requirements, Figure 11 illustrates the instantaneous SC over time across various methods under identical channel realizations. All four methods exhibit similar trends, but only the proposed method and H-Q/E, with its added constraint, maintain strictly positive SC throughout the transmission. In contrast, M3DDPG and game theory methods intermittently fall below zero, failing to guarantee persistent security. Notably, the game-theoretic approach proves inadequate in this highly dynamic environment. The lack of an adaptive learning mechanism prevents it from performing rapid, real-time adjustments in response to the fast-fading channel coefficients modeled by the SDEs.

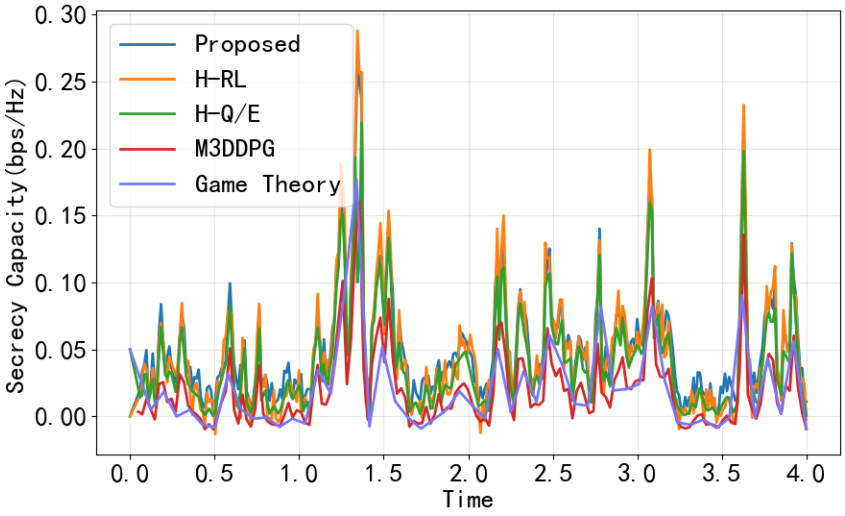


**Figure 11.** SC versus time under different methods.

The boxplot based on Figure 11 is shown in Figure 12, which provides a statistical summary of the instantaneous SC distribution for each method. The proposed method exhibits the highest median value, demonstrating superior average performance compared to the baselines. Moreover, the distribution of the proposed method is strictly non-negative, with the lower whisker staying at or above zero, which validates the framework's ability to strictly enforce the positive SC constraint. In contrast, H-RL, M3DDPG, and game theory methods show lower whiskers extending below zero, indicating frequent security outages where the SC becomes negative.

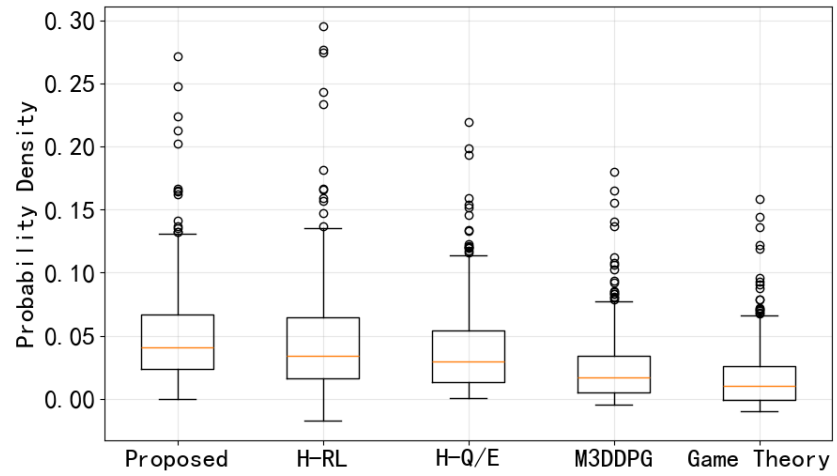


**Figure 12.** Boxplot comparison of SC distributions. The orange line within each box represents the median value, and the dots indicate outliers.

## 6. Conclusions

In this paper, we proposed a novel CHRL framework to address the challenge of secure cooperative communication in highly dynamic 6G networks. By modeling the time-varying wireless channels using SDEs, we established a realistic environment where traditional optimization methods often fail due to the lack of perfect CSI. This study offers several novel insights into PLS. First, we demonstrated that hierarchically decoupling the mixed-integer optimization problem fundamentally resolves the dilemma between the curse of dimensionality and performance degradation caused by action space discretization. Second, we shifted the traditional paradigm of merely maximizing SC to providing strict security guarantees. By utilizing a risk-aware meta-controller that models secrecy violations

as long-term risks, the system ensures a strictly positive SC even under highly dynamic channel variations. Finally, we revealed the double-edged-sword nature of cooperative jamming. By framing continuous power allocation as a constrained zero-sum game and deploying a constrained M3DDPG algorithm against worst-case adversarial perturbations, our framework achieves unprecedented robustness. The simulation results confirm that the proposed method outperforms existing baselines, achieving SC improvements of approximately 6% over H-RL, 17% over H-Q/E networks, and 49% over M3DDPG, while the integration of Lagrangian multipliers rigorously prevents security outages.

**Author Contributions:** Conceptualization, X.T. and Z.W.; methodology, X.T.; software, X.T.; validation, X.T., Z.W. and Y.N.; formal analysis, X.T.; investigation, X.T.; resources, Z.W.; data curation, X.T.; writing—original draft preparation, X.T.; writing—review and editing, X.T., Z.W. and Y.N.; visualization, X.T.; supervision, Y.N.; project administration, Y.N.; funding acquisition, Y.N. and Z.W. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was funded in part by the China Postdoctoral Science Foundation under Grant Number 2024M764088 and in part by the National Natural Science Foundation of China under Grant 61971025, 62331002.

**Data Availability Statement:** Data is contained within the article

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Saad, W.; Bennis, M.; Chen, M. A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems. *IEEE Netw.* **2020**, *34*, 134–142. [CrossRef]
2. Bolgouras, V.; Farao, A.; Xenakis, C. Roadmap to Secure 6G Networks. In Proceedings of the International Workshop on Computer Aided Modeling and Design of Communication Links and Networks, Athens, Greece, 21–23 October 2024; pp. 1–6.
3. Mukherjee, A.; Fakoorian, S.A.A.; Huang, J.; Swindlehurst, A.L. Principles of Physical Layer Security in Multiuser Wireless Networks: A Survey. *IEEE Commun. Surv. Tutor.* **2014**, *16*, 1550–1573. [CrossRef]
4. Wu, Y.; Duong, T.Q.; Swindlehurst, A.L. Safeguarding 5G-and-Beyond Networks with Physical Layer Security. *IEEE Wirel. Commun.* **2019**, *26*, 4–5. [CrossRef]
5. Bloch, M.; Barros, J.; Rodrigues, M.R.D.; McLaughlin, S.W. Wireless Information-Theoretic Security. *IEEE Trans. Inf. Theory* **2008**, *54*, 2515–2534. [CrossRef]
6. Wang, K.; Yuan, L.; Miyazaki, T.; Zeng, D.; Guo, S.; Sun, Y. Strategic Antieavesdropping Game for Physical Layer Security in Wireless Cooperative Networks. *IEEE Trans. Veh. Technol.* **2017**, *66*, 9448–9457. [CrossRef]
7. Wu, H.; Li, M.; Gao, Q.; Wei, Z.; Zhang, N.; Tao, X. Eavesdropping and Anti-Eavesdropping Game in UAV Wiretap System: A Differential Game Approach. *IEEE Trans. Wirel. Commun.* **2022**, *21*, 9906–9920. [CrossRef]
8. Zazo, S.; Valcarcel Macua, S.; Sánchez-Fernández, M.; Zazo, J. Dynamic Potential Games With Constraints: Fundamentals and Applications in Communications. *IEEE Trans. Signal Process* **2016**, *64*, 3806–3821. [CrossRef]
9. Wang, Q.; Zhang, T.; Dong, D. Achievable Secure Degrees of Freedom of MIMO Interference Channel with a Cooperative Jammer. *IEEE Wirel. Commun. Lett.* **2021**, *10*, 320–324. [CrossRef]
10. Qu, J.; Cai, Y.; Lu, J.; Wang, A.; Zheng, J.; Yang, W.; Weng, N. Power allocation based on Stackelberg game in a jammer-assisted secure network. In Proceedings of the International Conference on Cyberspace Technology, Beijing, China, 8–10 November 2014; pp. 347–352.
11. Liu, W.; Zhou, X.; Durrani, S.; Popovski, P. Secure Communication With a Wireless-Powered Friendly Jammer. *IEEE Trans. Wirel. Commun.* **2016**, *15*, 401–415. [CrossRef]
12. Lee, J.H. Optimal Power Allocation for Physical Layer Security in Multi-Hop DF Relay Networks. *IEEE Trans. Wirel. Commun.* **2016**, *15*, 28–38. [CrossRef]
13. Li, H.; Li, J.; Liu, M.; Gong, F. Performance Analysis and Secure Resource Allocation for Relay-Aided MISO-NOMA Systems. *IEEE Syst. J.* **2024**, *18*, 1617–1628. [CrossRef]
14. Charalambous, C.D.; Menemenlis, N. Stochastic models for short-term multipath fading channels: Chi-square and Ornstein-Uhlenbeck processes. In Proceedings of the 38th IEEE Conference on Decision and Control, Phoenix, AZ, USA, 7–10 December 1999; Volume 5, pp. 4959–4964.

15. Feng, T.; Field, T.R.; Haykin, S. Stochastic Differential Equation Theory Applied to Wireless Channels. *IEEE Trans. Commun.* **2007**, *55*, 1478–1483. [CrossRef]
16. Ben Amar, E.; Ben Rached, N.; Tempone, R.; Alouini, M.S. Stochastic Differential Equations for Performance Analysis of Wireless Communication Systems. *IEEE Trans. Wirel. Commun.* **2025**, *24*, 4040–4054. [CrossRef]
17. Kurniawan, F.; Agustian, H.; Dermawan, D.; Nurdin, R.; Ahmadi, N.; Dinaryanto, O. Hybrid Rule-Based and Reinforcement Learning for Urban Signal Control in Developing Cities: A Systematic Literature Review and Practice Recommendations for Indonesia. *Appl. Sci.* **2025**, *15*, 10761. [CrossRef]
18. Hoang, D.M.T.; Pham, T.V.; Pham, A.T.; Nguyen, C.T. Joint Design of Adaptive Modulation and Precoding for Physical Layer Security in Visible Light Communications Using Reinforcement Learning. *IEEE Access* **2024**, *12*, 82318–82332. [CrossRef]
19. Hoseini, S.A.; Bouhafs, F.; Aboutorab, N.; Sadeghi, P.; Hartog, F.d. Cooperative Jamming for Physical Layer Security Enhancement Using Deep Reinforcement Learning. In Proceedings of the IEEE Globecom Workshops, Kuala Lumpur, Malaysia, 4–8 December 2023; pp. 1838–1843.
20. Gu, Z.; Gao, L.; Ma, H.; Li, S.E.; Zheng, S.; Jing, W.; Chen, J. Safe-State Enhancement Method for Autonomous Driving via Direct Hierarchical Reinforcement Learning. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 9966–9983. [CrossRef]
21. Tran, H.; Pham, V.; Ngo, S. A survey on the applications of machine learning, deep learning, and reinforcement learning in wireless communications. *Comput. Telecommun. Eng.* **2025**, *3*, 3170. [CrossRef]
22. Huang, C.; Chen, G.; Wong, K.K. Multi-Agent Reinforcement Learning-Based Buffer-Aided Relay Selection in IRS-Assisted Secure Cooperative Networks. *IEEE Trans. Inf. Forensics Secur.* **2021**, *16*, 4101–4112. [CrossRef]
23. Lu, X.; Xiao, L.; Niu, G.; Ji, X.; Wang, Q. Safe Exploration in Wireless Security: A Safe Reinforcement Learning Algorithm With Hierarchical Structure. *IEEE Trans. Inf. Forensics Secur.* **2022**, *17*, 732–743. [CrossRef]
24. Li, Y.; Xu, Y.; Xu, Y.; Liu, X.; Wang, X.; Li, W.; Anpalagan, A. Dynamic Spectrum Anti-Jamming in Broadband Communications: A Hierarchical Deep Reinforcement Learning Approach. *IEEE Wirel. Commun. Lett.* **2020**, *9*, 1616–1619. [CrossRef]
25. Geng, Y.; Liu, E.; Wang, R.; Liu, Y. Hierarchical Reinforcement Learning for Relay Selection and Power Optimization in Two-Hop Cooperative Relay Network. *IEEE Trans. Commun.* **2022**, *70*, 171–184. [CrossRef]
26. Anil Akyildiz, H.; Faruk Gemici, O.; Hokelek, I.; Ali Cirpan, H. Hierarchical Reinforcement Learning Based Resource Allocation for RAN Slicing. *IEEE Access* **2024**, *12*, 75818–75831. [CrossRef]
27. Qian, J.; Li, H.; Zhu, P.; Zhou, A.; Liu, S.; Wang, F. Cooperative Jamming and Relay Selection for Covert Communications Based on Reinforcement Learning. *Sensors* **2025**, *25*, 6218. [CrossRef]
28. Vahidian, S.; Aïssa, S.; Hatamnia, S. Relay Selection for Security-Constrained Cooperative Communication in the Presence of Eavesdropper's Overhearing and Interference. *IEEE Wirel. Commun. Lett.* **2015**, *4*, 577–580. [CrossRef]
29. Li, S.; Wu, Y.; Cui, X.; Dong, H.; Russell, S. Robust Multi-Agent Reinforcement Learning via Minimax Deep Deterministic Policy Gradient. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 4213–4220.
30. Wang, A.; Tang, J.; Luo, H.; Wen, H.; Han-Ho, P.; Chang, S.Y. Physical Layer Security Performance Prediction Based on Deep Learning. In *Proceedings of the 2023 2nd International Conference on Computing, Communication, Perception and Quantum Technology (CCPQT)*; IEEE: New York, NY, USA, 2023; pp. 222–228.
31. Deng, D.; Li, X.; Zhao, M.; Rabie, K.M.; Kharel, R. Deep Learning-Based Secure MIMO Communications with Imperfect CSI for Heterogeneous Networks. *Sensors* **2020**, *20*, 1730. [CrossRef] [PubMed]
32. Robert, A.; Pike-Burke, C.; Faisal, A.A. Sample complexity of goal-conditioned hierarchical reinforcement learning. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 62696–62712.