Mode-Aware Anti-Jamming for Dynamic Spectrum Sharing UAV Networks: Joint Spectrum and Trajectory Optimization

***Abstract*—**
***Index Terms*—**

I. Introduction

Unmanned aerial vehicles (UAVs) have become a key enabler of low-altitude wireless networks, providing flexible and on-demand communication services such as emergency coverage, temporary hotspots, and aerial relaying. Due to the scarcity of dedicated spectrum, UAV systems commonly operate under dynamic spectrum sharing (DSS) frameworks, where secondary UAV links coexist with incumbent primary users (PUs). Although DSS improves spectrum efficiency, it also exposes UAV communications to highly dynamic interference, especially in adversarial environments with intentional jamming. In practice, jammers rarely employ a fixed strategy; instead, they switch among multiple jamming patterns, such as constant, sweep, and random jamming, to increase uncertainty and hinder adaptive communication. This pattern-switching behavior results in abrupt and non-stationary interference dynamics, leading to frequent quality-of-service (QoS) degradation and communication outages. For low-altitude UAV missions, rapid recovery of communication performance following interference transitions is often more critical than optimizing long-term average throughput.

To address spectrum scarcity and interference, extensive research has investigated joint spectrum allocation and UAV trajectory optimization. By exploiting the UAV's mobility and flexible spectrum access, learning-based approaches can adapt resource allocation and flight trajectories in dynamic environments. In particular, reinforcement learning (RL) has been widely applied to handle the coupled discrete–continuous optimization inherent in spectrum assignment and UAV movement. However, most existing methods treat interference as an exogenous disturbance and directly include raw interference measurements in the state space, elying on the learning agent to implicitly infer jammer behavior from long-term reward feedback. Under pattern-switching jamming, this implicit learning results in an effectively non-stationary environment, where the learned policy tends to converge to conservative average-case behaviors and exhibits slow adaptation when the jamming pattern changes.

This observation reveals a key limitation of existing designs: the lack of explicit awareness of jamming patterns prevents fast reaction to abrupt interference transitions. In contrast to throughput-oriented formulations, anti-jamming UAV communications require rapid frequency agility and timely decision updates to restore link quality with minimal delay. Motivated by this challenge, we focus on a pattern-switching jamming scenario and investigate how explicit jamming cognition can be leveraged to improve the responsiveness and robustness of DSS-enabled UAV communications.

In this paper, we propose a mode-aware anti-jamming framework that explicitly decouples jamming cognition from control. The jammer behavior is modeled as a finite-mode pattern-switching process. Using short-window spectrum observations, the UAV rapidly estimates the posterior distribution of the current jamming mode, referred to as a belief state. This belief state is then incorporated into the decision-making process to condition both spectrum allocation and trajectory control, enabling instantaneous policy adaptation upon jamming mode transitions. Moreover, to capture practical frequency agility constraints, we explicitly introduce a frequency switching cost into the optimization objective, balancing fast reaction to interference and stable spectrum usage.

To solve the resulting problem, we develop a hybrid learning framework that jointly optimizes discrete spectrum allocation and continuous UAV trajectory control. A Proximal Policy Optimization (PPO) agent is employed to learn the spectrum allocation policy, while a continuous-control agent, such as TD3 or SAC, optimizes the UAV trajectory. Both agents are conditioned on the jamming mode belief and trained using a reliability-oriented reward that penalizes QoS outages while ensuring PU protection.

The main contributions of this paper are summarized as follows:

-
-
-

The remainder of this paper is organized as follows. Section II introduces the system model, including the dynamic spectrum sharing UAV network and the pattern-switching jamming scenario. Section III presents the proposed mode-aware anti-jamming framework, including jamming mode recognition and the joint spectrum allocation and trajectory control scheme. Simulation results and performance evaluations are provided in Section IV. Finally, the paper concludes with Section V.

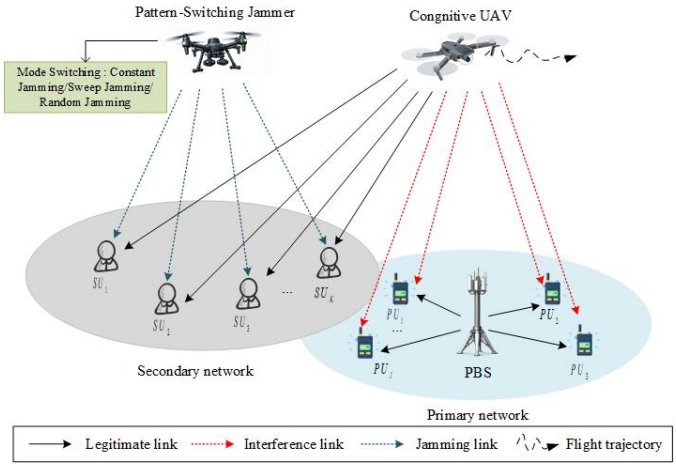

Fig. 1. Dynamic spectrum sharing UAV network under pattern-switching jamming, where the C-UAV performs jamming mode awareness and jointly optimizes spectrum allocation, trajectory, and transmit power to ensure reliable communication.

II. System Model

A. Scenario Description

We consider a UAV-enabled dynamic spectrum sharing (DSS) network under intentional pattern-switching jamming, as illustrated in Fig. 1. The network consists of a primary network and a cognitive UAV-assisted secondary network, which coexist over a set of licensed spectrum bands.

The primary network includes a primary base station and J primary users (PUs). The PUs are indexed by $j \in \mathcal{J} \triangleq \{1, 2, \ldots, J\}$. In order to serve more secondary users (SUs) and provide better services, a wideband spectrum is divided into M sub-carrier spectrum bands. The spectrum bands are indexed by $m \in \mathcal{M} \triangleq \{1, 2, \ldots, M\}$. To eliminate mutual interference among PUs, each PU is assigned an exclusive licensed spectrum band. The secondary network comprises a cognitive UAV serving K ground SUs. The SUs are indexed by $k \in \mathcal{K} \triangleq \{1, 2, \ldots, K\}$. The C-UAV operates as an aerial secondary base station, providing downlink communications to SUs by opportunistically accessing the licensed spectrum while ensuring that the interference inflicted on the primary network remains within acceptable limits.

A three-dimensional Cartesian coordinate system is adopted. The horizontal locations of the PBS, the j-th PU, the k-th SU, the C-UAV and the J-UAV at time slot n are denoted by $\mathbf{w}_b = (x_b, y_b)$, $\mathbf{w}_{p,j} = (x_{p,j}, y_{p,j})$, $\mathbf{w}_{s,k} = (x_{s,k}, y_{s,k})$, $\mathbf{q}_c = (x_c, y_c)$ and $\mathbf{q}_{jam} = (x_{jam}, y_{jam})$ respectively. These UAVs fly at a fixed altitude H_c. The total communication duration T is divided into N equal-length time slots, each with duration $\Delta t = T/N$. Due to the short duration of each slot, the C-UAV is assumed to be quasi-static within one slot. Let $n \in \mathcal{N} \triangleq \{1, 2, \ldots, N\}$ denote the set of time slots. The dynamic positions of the SUs can be formulated as

$$\begin{aligned} x_{s,k}[n+1] &= x_{s,k}[n] + v_s \cos\big(\phi_{s,k}[n]\big), \\ y_{s,k}[n+1] &= y_{s,k}[n] + v_s \sin\big(\phi_{s,k}[n]\big), \end{aligned} \tag{1}$$

where $\phi_{s,k}[n] \in [-\pi, \pi]$ denotes the direction of the k-th SU at time slot n. The SUs move at a fixed speed v_s. Similarly, the dynamic position of the C-UAV can be formulated as

$$\begin{aligned} x_c[n+1] &= x_c[n] + v_c \cos\big(\phi_c[n]\big), \\ y_c[n+1] &= y_c[n] + v_c \sin\big(\phi_c[n]\big), \end{aligned} \tag{2}$$

where v_c is the constant flight speed, and $\phi_c[n] \in [-\pi, \pi]$ denotes the direction of the C-UAV. The distance between the C-UAV and the k-th SU, and that between the PBS and the j-th PU at time slot n are respectively given as,

$$\begin{aligned} d_{c,k}[n] &= \sqrt{\big\|\mathbf{q}_c[n] - \mathbf{w}_{s,k}[n]\big\|^2 + H_c^2}, \\ d_{p,j}[n] &= \sqrt{\big\|\mathbf{w}_b[n] - \mathbf{w}_{p,j}[n]\big\|^2}, \end{aligned} \tag{3}$$

The wireless channel between the UAV and the ground users is dominated by the line-of-sight (LoS) link. Let β_{ref} represent the channel power gain at the reference distance of 1 m. Thus, the channel power gain from the C-UAV to the k-th SU can be expressed as

$$h_{C,k,n}^{\text{LoS}} = \beta_{\text{ref}} d_{c,k}[n]^{-2}, \tag{4}$$

where $d_{c,k}[n]$ denotes the distance between the C-UAV and the k-th SU at time slot n. The channel model between the PBS and the ground users is different from that between the UAV and the ground users.It is required to consider both the distance-dependent path loss with exponent $\varphi \geq 2$ and small-scale Rayleigh fading. Thus, the channel power gains from the PBS to the j-th PU can be given as

$$h_{P,j,n}^{\text{NLoS}} = \beta_{\text{ref}} d_{p,j}[n]^{-\varphi} \zeta_j, \tag{5}$$

where $d_{p,j}[n]$ is the distance between the PBS and the j-th PU at time slot n. The random variable ζ_j follows an exponential distribution with unit mean, which accounts for the Rayleigh fading.

Unlike conventional fixed or random jamming models, the J-UAV is assumed to adopt a pattern-switching strategy. Let $z[n] \in \mathcal{Z} \triangleq \{\text{Constant}, \text{Sweep}, \text{Random}\}$ denote the jamming mode at time slot n, corresponding to constant jamming, sweep jamming, and random jamming, respectively. The jamming mode evolves over time according to a finite-state stochastic process. In particular, the mode transition is modeled as a first-order Markov chain, i.e.,

$$\Pr(z[n+1] = z' \mid z[n] = z) = \Pi_{z,z'}, \quad z, z' \in \mathcal{Z}. \tag{6}$$

where $\mathbf{\Pi}$ denotes the mode transition probability matrix. The transition probabilities are not known to the C-UAV, and the instantaneous jamming mode cannot be directly observed. The pattern-switching behavior of the jammer directly translates into time-varying and non-stationary interference across the spectrum.

Considering the sub-carrier allocation, the binary variables $\rho_{k,n}[m]$ and $\rho_{J,n}[m]$ are introduced to characterize the spectrum usage of the SUs and the J-UAV, respectively.Specifically, $\rho_{k,n}[m] = 1$ indicates that the m-th sub-carrier is assigned to the k-th SU at time slot n, and $\rho_{k,n}[m] = 0$ otherwise. Each

PU operates on a preassigned orthogonal sub-carrier with a fixed transmit power. The J-UAV dynamically occupies sub-carriers according to its current jamming mode, which is captured by the jamming indicator $\rho_{J,n}[m]$. When $\rho_{J,n}[m] = 1$, the m-th sub-carrier is subject to intentional interference from the jammer at time slot n. Based on the sub-carrier availability and the time-varying jamming condition, the C-UAV aims to dynamically adapt the spectrum allocation for the SUs, such that the communication quality can be rapidly restored after jamming pattern changes, while opportunistically accessing the licensed spectrum and guaranteeing the quality-of-service requirements of the primary network.

The SINR of the k-th SU and the j-th PU in the m-th sub-band can be respectively expressed as

$$\gamma^{s}_{m,k,n} = \frac{\rho_{k,n}[m]P_C[n]h^{\mathrm{LoS}}_{C,k,n}}{\sigma^2 + P_B h^{\mathrm{NLoS}}_{P,k,n} + \rho_{J,n}[m]P_J h^{\mathrm{LoS}}_{J,k,n}}, \tag{7a}$$

$$\gamma^{p}_{m,j,n} = \frac{P_B h^{\mathrm{NLoS}}_{P,j,n}}{\sigma^2 + \sum_{k=1}^{K} \rho_{k,n}[m]P_C[n]h^{\mathrm{LoS}}_{C,j,n} + \rho_{J,n}[m]P_J h^{\mathrm{LoS}}_{J,j,n}}, \tag{7b}$$

where P_B and P_J denote the transmit power of the PBS and the J-UAV, respectively. $P_C[n] \in [P_C^{\min}, P_C^{\max}]$ denotes the transmit power of the C-UAV at time slot n. Moreover, σ^2 is the noise power. Then, the achievable data transmission rate of the k-th SU and the j-th PU can be respectively expressed as

$$R^{s}_{m,k,n} = B\log_2\big(1+\gamma^{s}_{m,k,n}\big), \tag{8a}$$

$$R^{p}_{m,j,n} = B\log_2\big(1+\gamma^{p}_{m,j,n}\big), \tag{8b}$$

where B denotes the bandwidth for each sub-band.

To capture communication reliability under pattern-switching jamming, we introduce a QoS threshold R_{th}. An outage event is declared when the achievable rate of a SU falls below R_{th}. Accordingly, the outage indicator is defined as

$$\mathbb{I}^{\mathrm{out}}_{k,n} \triangleq \mathbf{1}\big\{R^{s}_{k,n} < R_{\mathrm{th}}\big\}, \quad \forall k \in \mathcal{K},\ \forall n \in \mathcal{N}, \tag{9}$$

where $\mathbf{1}\{\cdot\}$ equals 1 if the condition holds and 0 otherwise. This definition penalizes each outage time slot; thus, longer outage durations yield larger cumulative penalties, which encourages fast recovery after jamming pattern transitions.

B. Problem Formulation

In this paper, we address a joint spectrum allocation, UAV trajectory control, and transmit power adaptation problem in a dynamic spectrum sharing UAV network under pattern-switching jamming, aiming to achieve fast recovery and reliable communication for secondary users while satisfying the primary user protection constraints. Accordingly, the optimization problem is formulated as follows

$$\mathbf{P1}: \max_{\boldsymbol{\rho},\boldsymbol{\phi},\mathbf{P}} \quad \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{m=1}^{M} R^{\mathrm{s}}_{m,k,n} - \frac{\lambda_{\mathrm{out}}}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{I}^{\mathrm{out}}_{k,n} \tag{10a}$$

$$\text{s.t.} \quad \sum_{m=1}^{M} R^{p}_{m,j,n} \geq R^{\min}_{j}, \quad \forall j \in \mathcal{J},\ \forall n \in \mathcal{N}, \tag{10b}$$

$$\rho_{k,n}[m] \in \{0,1\}, \quad \forall k \in \mathcal{K},\ \forall m \in \mathcal{M},\ \forall n \in \mathcal{N}, \tag{10c}$$

$$\sum_{k=1}^{K} \rho_{k,n}[m] \leq 1, \quad \forall m \in \mathcal{M},\ \forall n \in \mathcal{N}, \tag{10d}$$

$$|\phi_c[n]| \leq \pi, \quad \forall n \in \mathcal{N}, \tag{10e}$$

$$P_C^{\min} \leq P_C[n] \leq P_C^{\max}, \quad \forall n \in \mathcal{N}. \tag{10f}$$

To jointly account for throughput performance and communication reliability, a reliability-aware objective function is adopted. Specifically, the achievable rate of the secondary users is maximized, while a penalty term is introduced to capture QoS outages. This outage penalty discourages prolonged service interruptions caused by dynamic jamming and implicitly promotes fast recovery of communication quality after jamming pattern transitions. In problem (P1), $R_j^{\min}$ denotes the minimum transmission rate requirement of the j-th PU, and $\boldsymbol{\rho} = \{\rho_{k,n}[m]\}_{k\in\mathcal{K},m\in\mathcal{M},n\in\mathcal{N}}$ denotes the spectrum allocation policy for the SUs. The variable $\boldsymbol{\phi} = \{\phi[n]\}_{n\in\mathcal{N}}$ denotes the UAV heading angle control sequence, while $\mathbf{P} = \{P_C[n]\}_{n\in\mathcal{N}}$ denotes the transmit power adaptation policy of the cognitive UAV. Constraint (10b) guarantees the QoS requirements of the PUs in the dynamic spectrum sharing environment. Constraints (10c) and (10d) ensure orthogonal sub-carrier allocation among SUs to avoid intra-secondary interference. Constraint (10e) enforces feasible UAV mobility, while constraint (10f) restricts the transmit power of the C-UAV within practical limits.

III. Our Proposed Mode-Aware Anti-Jamming Framework

A. Jamming Mode Recognition

The jammer switches among finite modes $z[n] \in \mathcal{Z}$ following the Markov model in (6). The C-UAV performs spectrum sensing and collects the per-slot interference-energy vector

$$\mathbf{I}[n] = [I_1[n], \ldots, I_M[n]] \in \mathbb{R}^M, \tag{11}$$

and forms a length-L short window

$$\mathcal{I}[n] = \{\mathbf{I}[n-L+1], \ldots, \mathbf{I}[n]\} \in \mathbb{R}^{L\times M}. \tag{12}$$

A lightweight recognizer $f_\theta(\cdot)$ maps the input $\mathcal{I}[n]$ to a mode probability vector

$$\boldsymbol{\pi}[n] = f_\theta(\mathcal{I}[n]) \in \mathbb{R}^{|\mathcal{Z}|}, \quad \sum_{c=1}^{|\mathcal{Z}|}\pi_c[n] = 1, \tag{13}$$

which provides semantic information about the current jamming behavior. We adopt a CNN–GRU classifier, where a 1D CNN encodes the spectral characteristics of each $\mathbf{I}[\cdot]$, and a GRU captures short-term temporal dynamics over L time slots. The model is trained offline using the cross-entropy loss

$$\mathcal{L}_{\text{cls}}(\theta) = -\sum_{c=1}^{|\mathcal{Z}|} z_c[n] \log\big(\pi_c[n]\big), \tag{14}$$

where $z_c[n]$ denotes the one-hot encoded ground-truth jamming mode at time slot n. During online operation, the model is used solely for inference to output $\boldsymbol{\pi}[n]$.

B. Semantic-Augmented MDP Formulation

Following the MDP formulation in the reference work [1], we incorporate the mode probability vector $\boldsymbol{\pi}[n]$ as a semantic descriptor of the interference environment to facilitate fast adaptation under mode-switching jamming.

State: The system state at slot n is defined as

$$\mathbf{s}[n] = \big(\mathbf{a}[n-1],\, \mathbf{q}_c[n-1],\, \mathbf{g}[n],\, \boldsymbol{\pi}[n]\big), \tag{15}$$

where $\mathbf{a}[n-1]$ denotes the previous action at slot $n-1$, $\mathbf{q}_c[n-1]$ denotes the previous UAV position at slot $n-1$, $\mathbf{g}[n]$ denotes channel and geometry-related features for rate evaluation, and $\boldsymbol{\pi}[n]$ encodes the high-level semantic belief of the current jamming behavior.

Action: The action consists of a discrete spectrum decision and a continuous control vector,

$$\mathbf{a}[n] = \big(\mathbf{a}_d[n],\, \mathbf{a}_c[n]\big), \tag{16}$$

The discrete action $\mathbf{a}_d[n]$ represents a joint subcarrier allocation for all K SUs at slot n, i.e.,

$$\mathbf{a}_d[n] = \big(a_1[n],\, a_2[n],\, \ldots,\, a_K[n]\big), \tag{17}$$

where $a_k[n] \in \{1, 2, \ldots, M\}$ denotes the selected subcarrier index for the k-th SU. Equivalently, $\mathbf{a}_d[n]$ can be mapped to the binary allocation variables $\{\rho_{k,n}[m]\}$ satisfying $\sum_{m=1}^{M} \rho_{k,n}[m] = 1$ for each k. The continuous action $\mathbf{a}_c[n]$ is defined as

$$\mathbf{a}_c[n] = \big(\phi_c[n],\, P_C[n]\big), \tag{18}$$

where $\phi_c[n] \in [-\pi, \pi]$ and $P_C[n] \in [P_C^{\min},\, P_C^{\max}]$.

Reward: Consistent with (10), we adopt a reliability-aware reward that jointly captures the secondary throughput, outage reliability, and PU protection, defined as

$$r[n] = \alpha \sum_{k=1}^{K} R_k^s[n] + \beta \sum_{j=1}^{J} \delta_j[n] - \gamma \sum_{k=1}^{K} \mathbb{I}(R_k^s[n] < R_{\text{th}}), \tag{19}$$

where $\alpha, \beta, \gamma > 0$ are constant coefficients, $\delta_j[n]$ is a penalty term when the interference to the PUs exceeds the tolerated threshold and formulated as

$$\delta_j[n] = \begin{cases} 0, & R_j^p[n] \geq R_j^{\min}, \\ R_j^p[n] - R_j^{\min}, & 0 \leq R_j^p[n] < R_j^{\min}, \end{cases} \tag{20}$$

where $R_j^p[n]$ is the achievable rate of the j-th PU and $R_j^{\min}$ is its minimum rate requirement.

C. Hybrid PPO-TD3 Framework

Problem (10) involves discrete spectrum allocation and continuous joint control of UAV mobility and transmit power. Thus, we employ a hybrid solver, where PPO is adopted for the discrete spectrum decision $\mathbf{a}_d[n]$ and TD3 is used for the continuous control vector $\mathbf{a}_c[n]$, both conditioned on the semantic-augmented state $\mathbf{s}[n]$.

PPO for Spectrum Allocation: Let $\pi_\theta(\mathbf{a}_d|\mathbf{s})$ denote the stochastic policy for spectrum allocation parameterized by θ, where $\mathbf{a}_d[n]$ is the discrete joint subcarrier allocation decision at time slot n. Different from the multi-agent A2C design in [1], we adopt a centralized PPO-based spectrum allocator. To avoid the exponential growth of the joint discrete action space, the policy is implemented in a factorized multi-discrete form as

$$\pi_\theta(\mathbf{a}_d|\mathbf{s}) = \prod_{k=1}^{K} \pi_\theta^{(k)}\big(a_k|\mathbf{s}\big), \tag{21}$$

where each $\pi_\theta^{(k)}(\cdot|\mathbf{s})$ is a categorical distribution over the M candidate subcarriers. The objective of PPO is to maximize the expected discounted return while preventing excessively large policy updates. Specifically, PPO optimizes the clipped surrogate objective

$$\begin{aligned} \mathcal{L}^{\text{CLIP}}(\theta) = \mathbb{E}_n\Big[\min\Big\{ & r_n(\theta)\hat{A}_n, \\ & \text{clip}\big(r_n(\theta),\, 1-\epsilon,\, 1+\epsilon\big)\, \hat{A}_n \Big\}\Big], \end{aligned} \tag{22}$$

where $r_n(\theta) = \frac{\pi_\theta(\mathbf{a}_d[n]|\mathbf{s}[n])}{\pi_{\theta_{\text{old}}}(\mathbf{a}_d[n]|\mathbf{s}[n])}$ is the probability ratio between the new and old policies, ϵ is the clipping parameter, and $\hat{A}_n$ denotes the advantage estimate. To reduce variance and improve sample efficiency, the advantage is computed using Generalized Advantage Estimation (GAE),

$$\hat{A}_n = \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{n+l}, \quad \delta_n = r[n] + \gamma V(\mathbf{s}[n+1]) - V(\mathbf{s}[n]), \tag{23}$$

where $V(\cdot)$ is the value function, γ is the discount factor, and λ controls the bias–variance tradeoff.

TD3 for Trajectory Control: Let $\mu_\varphi(\mathbf{s})$ denote the deterministic actor that outputs the continuous control vector

$$\mathbf{a}_c[n] = \mu_\varphi(\mathbf{s}[n]) = \big(\phi_c[n],\, P_C[n]\big). \tag{24}$$

To stabilize learning in the continuous action space, TD3 employs two critic networks $Q_{\omega_1}(\mathbf{s}, \mathbf{a}_c)$ and $Q_{\omega_2}(\mathbf{s}, \mathbf{a}_c)$ to mitigate overestimation bias.

Given a transition $(\mathbf{s}[n], \mathbf{a}_c[n], r[n], \mathbf{s}[n+1])$, the target value is constructed as

$$y[n] = r[n] + \gamma \min_{i=1,2} Q_{\bar{\omega}_i}\big(\mathbf{s}[n+1], \tilde{\mathbf{a}}_c[n+1]\big), \tag{25}$$

where

$$\tilde{\mathbf{a}}_c[n+1] = \mu_{\bar{\varphi}}\big(\mathbf{s}[n+1]\big) + \boldsymbol{\eta}, \quad \boldsymbol{\eta} \sim \text{clip}\big(\mathcal{N}(0, \sigma^2), -c, c\big), \tag{26}$$

and $(\cdot)^-$ denotes the corresponding target networks. The critic networks are updated by minimizing the squared temporal-difference error

$$L_c(\omega_i) = \mathbb{E}\Big[\big(Q_{\omega_i}(\mathbf{s}[n], \mathbf{a}_c[n]) - y[n]\big)^2\Big], \quad i \in \{1, 2\}. \quad (27)$$

The actor is updated in a delayed manner by maximizing the Q-value estimated by the first critic,

$$\nabla_\varphi J(\varphi) = \mathbb{E}\Big[\nabla_\varphi \mu_\varphi(\mathbf{s}) \nabla_{\mathbf{a}_c} Q_{\omega_1}(\mathbf{s}, \mathbf{a}_c)\big|_{\mathbf{a}_c = \mu_\varphi(\mathbf{s})}\Big], \quad (28)$$

while the target networks are softly updated using Polyak averaging.

At each time slot, the C-UAV first obtains the jamming mode probability vector $\boldsymbol{\pi}[n]$ from the recognition module, constructs the semantic-augmented state $\mathbf{s}[n]$, and then outputs the discrete spectrum allocation $\mathbf{a}_d[n]$ via PPO and the continuous joint control action $\mathbf{a}_c[n] = (\phi_c[n], P_C[n])$ via TD3. PPO is updated using on-policy trajectories, whereas TD3 is trained off-policy with an experience replay buffer, enabling stable learning under pattern-switching jamming.

IV. SIMULATION RESULTS

V. CONCLUSION

REFERENCES

[1] R. Ding, F. Zhou, Y. Qu, C. Dong, Q. Wu, and T. Q. S. Quek, "Novel online-offline ma2c-ddpg for efficient spectrum allocation and trajectory optimization in dynamic spectrum sharing uav networks," in *2023 IEEE/CIC International Conference on Communications in China (ICCC)*, 2023, pp. 1–6.