

Article

Joint 3D Trajectory and Power Optimization for UAV Swarms in Cell-Free Massive MIMO Networks: A CTDE-MAPPO Framework for Sensing-Aware Precision Agriculture

Ayman Massaoudi ¹  and Walid Aydi ^{2,*} 

¹ Data Engineering and Semantics Research Unit, Faculty of Sciences of Sfax, University of Sfax, Sfax 3000, Tunisia; ayman.massaoudi@fss.usf.tn

² Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Al-Kharj 16273, Saudi Arabia

* Correspondence: w.aydi@psau.edu.sa

Highlights

What are the main findings?

- A joint sensing–communication utility function is introduced that balances communication energy efficiency and field coverage completeness in a single weighted objective that is embedded directly into a MAPPO-CTDE reward design with hard battery depletion and return-to-depot constraints—a formulation that, to our knowledge, is the first to jointly integrate these specific elements within a CF-mMIMO UAV setting, building on established CTDE, MAPPO, and reward-shaping techniques.
- The CF-mMIMO CPU serves as a global critic during training, resolving inter-agent non-stationarity while each UAV executes a lightweight local policy at runtime, achieving 93% field coverage and a 92% depot-return rate without any inter-UAV communication overhead at execution time.

What are the implications of the main findings?

- Jointly optimizing 3D rotary-wing trajectories and transmit power through a shared coverage map creates strong inter-agent coupling that makes the CF-mMIMO CPU a natural and architecturally motivated centralized critic—enabling scalable swarm coordination without dedicated inter-UAV communication links.
- CF-mMIMO macro-diversity, combined with sensing-aware trajectory adaptation, offers a practical and deployable pathway toward energy-autonomous UAV swarms for large-scale crop monitoring in 6G agricultural networks.

Abstract

The integration of Unmanned Aerial Vehicle (UAV) swarms with Cell-Free massive Multiple Input Multiple Output (CF-mMIMO) networks offers promising prospects for large-scale crop monitoring in precision agriculture. CF-mMIMO provides macro-diversity and uniform channel quality across large agricultural fields. However, practical deployment demands jointly optimizing 3D trajectories and transmit power to maximize energy efficiency and field coverage simultaneously. This is challenging due to the limited battery capacity, mandatory return-to-depot constraints, and collision avoidance requirements. In this paper, we introduce a joint sensing–communication utility function that captures the trade-off between energy efficiency and field coverage completeness. To provide a scalable and distributed solution for rotary-wing UAV swarms, we develop a multi-agent deep reinforcement learning (MADRL) methodology based on the Multi-Agent Proximal Policy Optimization (MAPPO) approach. We adopt the Centralized Training with Decentralized Execution (CTDE) strategy, in which a CF-mMIMO central processing unit



Academic Editor: Fei Liu

Received: 10 June 2026

Revised: 20 July 2026

Accepted: 24 July 2026

Published: 27 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

(CPU) serves as a global critic during training. At execution time, each UAV independently runs a lightweight local policy that adapts its trajectory and transmit power in real time based on battery state and air-to-ground channel variations. Simulation results reveal that the proposed MAPPO-CTDE approach outperforms existing benchmarks. Unlike prior methods that require instantaneous global CSI or neglect the sensing–communication coupling, the proposed approach simultaneously achieves high field coverage completeness, robust communication energy efficiency, and a high depot-return rate under hard battery constraints without any inter-UAV communication overhead at execution time.

Keywords: 6G; Cell-Free massive MIMO (CF-mMIMO); Centralized Training with Decentralized Execution (CTDE); deep reinforcement learning (DRL); energy efficiency (EE); Multi-Agent Proximal Policy Optimization (MAPPO); precision agriculture; propulsion power model; rotary-wing UAVs

1. Introduction

The convergence of sixth-generation (6G) wireless infrastructure and autonomous aerial systems is reshaping large-scale environmental monitoring [1–3]. Precision agriculture represents one of the most demanding applications of this convergence: it requires centimeter-level spatial coverage of fields spanning hundreds of hectares, a real-time uplink of high-dimensional sensing data, and energy-autonomous operation under strict flight-time constraints [4]. Ground-based IoT sensors provide localized measurements of soil moisture, temperature, and vegetation indices [5], but fixed infrastructure cannot scale to the spatial extent of modern farms [6]. Unmanned Aerial Vehicles (UAVs) resolve this limitation through three-dimensional mobility that enables an on-demand surveying of priority zones across arbitrarily large fields, decoupling sensing coverage from the constraints of fixed deployment [7–11]. Meeting the joint requirements of coverage completeness, real-time data uplink, and energy-autonomous operation—simultaneously and at scale—defines a class of integrated sensing-and-communication problems that existing wireless network architectures were not designed to solve. Deploying swarms of UAVs further improves coverage speed and fault tolerance, enabling a parallel surveying of spatially distributed field zones within a single mission horizon [12]. Simultaneous multi-UAV operation introduces co-channel interference, spectrum contention, and uplink reliability degradation. These communication impairments are particularly problematic as agents move continuously across the field, exceeding the mitigation capabilities of traditional single-cell network architectures at scale. Cell-Free massive Multiple-Input Multiple-Output (CF-mMIMO) networks, widely recognized as a key enabling technology for 6G systems [13], offer a principled solution to these impairments. Rather than relying on a single base station, CF-mMIMO distributes numerous access points (APs) across the coverage area and connects them to a central processing unit (CPU) via dedicated fronthaul links [14,15]. This architecture provides macro-diversity, suppresses path shadowing, and supports the coherent joint processing of simultaneous uplink transmissions from multiple UAVs—delivering uniform signal quality across the entire field regardless of UAV position [16].

1.1. Motivation and Problem Statement

Realizing the full potential of UAV swarms in precision agriculture requires jointly optimizing 3D trajectories and transmit power to simultaneously maximize energy efficiency and field coverage completeness—a non-convex, long-horizon problem further complicated by hard battery depletion limits, mandatory return-to-depot deadlines, and

collision avoidance requirements. Despite recent progress in UAV-assisted CF-mMIMO networks, existing frameworks address only a subset of these coupled constraints. Most prior studies optimize exclusively for communication metrics, neglecting sensing coverage as a primary mission objective, or treat battery depletion as a soft penalty rather than a hard operational limit.

In a real agricultural deployment, ignoring these coupled constraints is not a minor limitation but a point of mission failure:

1. A UAV that hovers exclusively at the altitude yielding the best channel quality will maximize communication efficiency but fail to sweep or scan the field.
2. A UAV that ignores aerodynamic battery depletion will drain its reserves prematurely and fail to return safely to the depot.
3. A swarm mission that skips partially surveyed zones will leave the end-user with an incomplete, unusable crop map.

Taken together, these conflicting physical and network constraints render the joint optimization problem fundamentally harder to solve. Addressing these intertwined limitations and bridging the gap between communication-centric design and physical mission survival serves as the core motivation behind the methodology developed in this paper.

1.2. Paper Contributions

To tackle the above-mentioned limitations, this paper presents the following key contributions:

1. We propose a CF-mMIMO multi-UAV remote sensing (RS) architecture for precision agriculture. To accurately reflect real-world deployments, this architecture incorporates (i) a three-dimensional (3D) rotary-wing mobility model, (ii) a probabilistic air-to-ground fading channel, and (iii) a full-mission battery depletion model.
2. We formulate a joint sensing–communication (SC) utility function $\mathcal{U}$ that simultaneously rewards communication energy efficiency and field coverage completeness under a single weighted objective (Section 4). We cast the resulting optimization as Problem $\mathcal{P}$: a non-convex, long-horizon, battery-constrained trajectory and power allocation problem subject to hard battery depletion limits, mandatory return-to-depot constraints, and minimum uplink rate requirements. We demonstrate analytically that $\mathcal{P}$ cannot be decomposed slot-by-slot, thereby ruling out classic SCA and supervised DNN approaches, and firmly motivating our MADRL reformulation. This formulation is absent from prior CF-mMIMO UAV studies, which either omit sensing coverage entirely or treat battery constraints as soft penalties.
3. Our proposed methodology centers on a Multi-Agent Proximal Policy Optimization (MAPPO) approach, which is selected specifically to mitigate the non-stationary dynamics of collaborative aerial swarms. By implementing a Centralized Training with Decentralized Execution (CTDE) strategy, we eliminate the need for inter-UAV communication during live missions. The training phase utilizes the CF-mMIMO CPU as a shared global critic that processes the complete system state to refine the behavior of all agents. Once deployed, each UAV operates autonomously, executing a localized policy that continuously adapts spatial trajectories and uplink power based on real-time battery levels and fading channel conditions. Furthermore, we carefully engineer the reward mechanisms and state-action spaces to ensure that the fleet naturally converges on strategies that maximize agricultural sensing coverage, avoid collisions, and respect absolute energy constraints.

The remainder of this paper is organized as follows. In Section 2, we review prior studies and highlight the key research gaps motivating this paper. Section 3 details the CF-mMIMO-assisted precision agriculture system model and outlines the key assumptions

considered throughout this paper. In Section 4, we formulate the joint optimization problem. Section 5 describes the proposed MAPPO-CTDE framework. In Section 6, we examine the simulation results. Section 7 discusses the achieved performance gains. Section 8 presents concluding remarks and highlights directions for future research.

Mathematical Notation: Throughout this paper, $(\cdot)^H$ denotes the conjugate transpose. Boldface lowercase letters indicate vectors, while boldface uppercase letters indicate matrices. $\|\cdot\|$ and $|\cdot|$ denote the Euclidean norm and absolute value. $\mathbb{E}[\cdot]$ is the expectation operator. We denote by $\mathcal{CN}(\mu, \sigma^2)$ the circularly symmetric complex Gaussian distribution.

2. Related Work

The integration of multi-UAV swarms with CF-mMIMO networks brings together challenges from wireless resource allocation, trajectory planning, and multi-agent reinforcement learning. Although substantial progress has been made in each area, enabling fully autonomous operation for precision agriculture requires a joint design that considers both communication performance and UAV flight constraints. This section reviews the literature most relevant to the proposed framework. We first discuss power allocation methods for CF-mMIMO systems, including both conventional optimization techniques and recent learning-based approaches. We then review UAV trajectory optimization methods and their limitations in dynamic multi-agent environments. Next, we examine studies that jointly optimize UAV trajectories and transmit power, highlighting the challenges associated with decentralized control and non-convex optimization. Finally, we compare the proposed MAPPO-based framework with existing approaches and identify the remaining challenges related to sensing coverage, battery limitations, and return-to-depot requirements.

2.1. Power Allocation in CF-mMIMO Networks

Resource allocation—specifically power assignment—in CF-mMIMO networks has been extensively investigated under both convex optimization and learning-based frameworks. Björnson and Sanguinetti [17] established the spectral and energy efficiency fundamentals of CF-mMIMO. They demonstrated the superiority of distributed processing over conventional cellular systems. Nevertheless, the scalability of cell-free networks remains constrained by two primary factors: the computational complexity of collaborative signal processing and the extensive signaling overhead required for inter-AP synchronization and data exchange. Recent research has focused on replacing computationally expensive iterative optimization solutions—including weighted minimum mean square error (WMMSE) and Fractional Programming—with deep learning (DL) to manage the complexity of CF-mMIMO systems. Braga et al. [18] formulated the joint pilot and data power control problem in CF systems as an optimization task. To address this, they proposed a decentralized actor-critic DRL framework that operates with minimal information exchange among network nodes. The authors in [19] tackled the challenge of uplink and downlink power control in CF-mMIMO systems by developing an unsupervised deep learning framework. By using only large-scale fading coefficients as input features, their approach establishes a desirable trade-off between overall system performance and processing overhead. Crucially, their framework supports scalability through its compatibility with user-centric access point clustering architectures. Similarly, Zaher et al. [20] introduced a deep learning framework for downlink power allocation to improve sum spectral efficiency (SE) and proportional fairness (PF). They designed a DNN that maps local large-scale fading (LSF) parameters and heuristic power coefficients directly to the target allocation. This configuration allows the trained model to approximate centralized optimization performance while maintaining a distributed implementation at individual APs. Furthermore, they developed a clustered DNN architecture to jointly optimize power coefficients within local AP groups. This

approach successfully balances the trade-off between utilizing global network knowledge and reducing computational overhead. However, these existing deep-learning paradigms are inherently limited in dynamic environments. Zaher et al. and similar contemporary studies [17–20] focused exclusively on standard physical-layer metrics, such as spectral efficiency and fairness, for stationary or generic user equipment. Consequently, they do not account for dynamic, mission-oriented scenarios like UAV-based precision agriculture. In such deployment contexts, 3D trajectory planning, strict battery capacity constraints, and spatial sensing coverage are equally critical to operational success.

2.2. UAV Trajectory Optimization

In [21], the authors derived a foundational analytical model for rotary-wing UAV propulsion power. Zeng et al. investigated a joint trajectory and time-allocation design to minimize the total energy consumption subject to ground-node throughput constraints. To address the continuous and non-convex problem, they employed path discretization coupled with SCA techniques. While effective for predictable, offline trajectory synthesis, such analytical optimization frameworks struggle to adapt in real time to the rapid, dynamic large-scale fading variations inherent to complex cell-free massive MIMO environments.

Similarly, Wu et al. [22] developed SCA-based frameworks for UAV trajectory but required perfect instantaneous CSI and centralized execution, limiting their applicability to multi-UAV deployments. The application of DRL to UAV trajectory planning was explored in cellular networks [23], where single-agent formulations demonstrated adaptability to dynamic channels. However, these approaches did not scale to multi-UAV settings or exploited the distributed structure of CF-mMIMO. The deployment of UAVs within CF-mMIMO frameworks was also explored from an infrastructure perspective, where aerial platforms acted as mobile base stations to assist ground networks. For instance, in [24], the authors utilized UAVs as mobile aerial access points (MAAPs) and deployed a Tabu-search algorithm alongside gradient-based location optimization to alleviate pilot contamination in user-centric CF-mMIMO networks. Moreover, Kenneth et al. [25] considered the deployment of UAVs as moving aerial providers (MAPs) to offload traffic and eliminate coverage dead zones. To handle the highly non-convex deployment problem under dynamic user distributions, the authors developed a hybrid framework that pairs an enhanced K-means clustering algorithm for user association with a soft actor–critic deep reinforcement learning algorithm for 3D location optimization. Nevertheless, the studies in [21–25] left several critical research gaps unaddressed. First, these studies focused narrowly on physical-layer metrics (downlink capacity and signal to interference plus noise ratio (SINR)). They overlooked dynamic, long-horizon operational constraints that are essential for autonomous fleet missions. These included continuous trajectory planning, hard battery depletion limits, and mandatory return-to-depot deadlines. Second, the centralized SAC approach adopted in these studies performed well for static positioning. It did not, however, scale to the distributed coordination demands of multiple moving aerial agents operating in real time. A related line of work in ground-based intelligent transportation addressed a structurally similar problem—generating smooth, collision-free trajectories under dynamic obstacle motion—using classical planning rather than learning. Huang et al. [26] proposed a sine resistance network (SRN) combined with a bias oval artificial potential field (BOAPF) for autonomous electric vehicles, meshing the road as a resistance network whose branch resistances were set by a potential field that incorporated the relative velocity of nearby obstacles. The planned path was then obtained by tracing the path of maximum current, which, by construction, avoided the local-minima failure mode of classical APF methods. However, like the SCA and APF-based UAV trajectory methods discussed above, this approach computed an explicit potential field or resistance network at every time step

rather than learning a policy, and it targeted a single ground vehicle reacting to nearby obstacles. It did not address multi-agent coordination, energy-constrained long-horizon missions, or the joint sensing–communication objective central to this paper.

2.3. Joint Trajectory and Power Optimization for UAV-Based Systems

Jointly optimizing UAV trajectories and transmit power is considerably more challenging than optimizing either variable separately. The difficulty arises from the strong interdependence between these two decisions. A UAV's trajectory determines the evolution of the air-to-ground channel conditions, which in turn affects the transmit power required to meet uplink rate requirements. At the same time, the energy consumed along a given trajectory directly constrains the mission duration and achievable flight range. As a result, approaches that optimize trajectory and power independently—by fixing one variable before optimizing the other—often lead to suboptimal performance. Early research explored the deployment of UAVs as amplify-and-forward relays to connect base stations and mobile devices that lacked reliable direct communication links. To address these connection challenges, Zhang et al. [27] proposed a framework that jointly optimized the flying trajectory of a single UAV alongside the transmit power of both the aerial relay and the ground-level mobile device. Their primary objective was to minimize the overall outage probability of the uplink network. Because the formulated problem was non-convex, Zhang et al. decoupled it into two distinct optimization stages, where they applied a gradient descent method for the trajectory design and extremum principles for the power control. Through an iterative alternating minimization algorithm, their joint solution significantly reduced the network's outage probability when compared to baseline models that relied on fixed relay positions, equal power allocation, or simple circular flight paths. However, the work [27] relied on a highly simplified topology restricted to a single mobile device and a single UAV, leaving the complexities of multi-user environments and UAV swarms unaddressed. Furthermore, by decoupling the non-convex optimization problem, their alternating minimization approach yielded only an approximate local minimum rather than a globally optimal solution. Hur and Lee [28] explored the simultaneous adjustment of flight paths and transmission power when reacting to a suddenly detected ground eavesdropper. To ensure uninterrupted service, the authors introduced a consistent fairness constraint that maintained stable data transmission rates for all users at every moment while the UAV relocated between its initial and final optimal hovering points. Furthermore, they integrated an energy constraint to account for the UAV's limited battery capacity during this redeployment, utilizing block coordinate descent and SCA approaches to solve the complex, non-convex optimization problem. However, despite demonstrating significant energy savings with minimal secrecy rate sacrifices, their approach hinged on several restrictive assumptions. Specifically, the proposed framework [28] was limited to a single UAV serving entirely stationary ground users over simplified, line-of-sight-dominated channels, which inherently restricts its immediate applicability in highly dynamic, multi-UAV environments or cluttered urban settings with complex signal obstructions. To address the scalability challenges associated with fully joint optimization, recent studies explored hierarchical and decentralized learning frameworks. For instance, Wu et al. [29] introduced a hierarchical framework utilizing a MARL method with decoupled critics (DCs) tailored for loosely coupled multi-UAV emergency communication tasks. In their architecture, a high-level clustering algorithm first assigned ground users to specific UAVs, while the lower-level MARL-DC layer jointly optimized UAV trajectories and transmit power to satisfy a proportional fairness objective. By decoupling the critics, each agent could effectively isolate shared global coordination objectives, such as collision avoidance, from localized decision-making processes. Although the approach of [29] demonstrated superior learning efficiency and robust collision avoidance

in dynamic disaster-relief environments, its application to precision agriculture remained limited. The target scenario in [29] is disaster-relief emergency communication, where UAV tasks are loosely coupled, and ground users are distributed across a post-disaster area—a fundamentally different operational context from crop monitoring, where sensing coverage completeness and energy-autonomous return-to-depot are non-negotiable mission requirements. Specifically, the work [29] did not incorporate CF-mMIMO infrastructure. Moreover, that formulation omitted critical domain-specific constraints, such as rotary-wing aerodynamic propulsion energy consumption modeling, strict battery capacities, return-to-depot requirements, and sensing coverage objectives. Shah et al. [30] developed a three-stage framework that jointly optimized UAV 3D trajectories, transmit power, and user association. The authors used an Accelerated Proximal Gradient (APG) framework to manage the inherent non-convexity of the formulated optimization problem. Their approach reduced the number of deployed drones and improved per-user throughput, but it focused entirely on communication metrics. Yet, classical optimization approaches, such as Successive Convex Approximation (SCA) or the APG method utilized by Shah et al. [30], typically yield centralized policies. These frameworks require aggregating instantaneous channel state information (CSI) at every slot to jointly optimize trajectories and power. This creates massive communication overhead and processing delays, making such centralized approaches inherently incompatible with the decentralized execution requirements of multi-UAV systems. In addition, the work in [30] did not consider energy efficiency or sensing coverage. This leaves a significant gap for precision agriculture applications, which demand full spatial coverage under strict battery constraints. Ma et al. [31] examined the problem of resource allocation within UAV-aided uplink CF-mMIMO systems, where the system was required to satisfy the stringent demands of ultra-reliable and low-latency communications. To maximize energy efficiency (EE) and minimize interference to ground users, the authors proposed a power allocation approach driven by a self-supervised deep neural network (DNN). Although effective in their setting, supervised DNNs are poorly suited to this problem for three reasons. Generating training labels requires solving the globally optimal trajectory-power-coverage sequence for every possible system state—a combinatorially intractable task. A DNN also produces a reactive, one-shot mapping with no ability to reason about how current decisions affect battery or coverage later in the mission. Finally, DNNs trained on a fixed channel distribution generalize poorly to new environments, swarm sizes, or seasonal changes in vegetation. Multi-Agent Reinforcement Learning (MARL) addresses all three of these limitations. It requires no pre-computed labels, optimizes long-term cumulative reward through temporal credit assignment, and adapts to non-stationary environments through continued interaction. Liu et al. [32] tackled UAV trajectory optimization within CF-mMIMO networks by proposing a Single Critic Multi-Actor (SCMA) Multi Agent Deep Deterministic Policy Gradient (MADDPG) methodology. This approach achieved an EE gain over traditional SCA methods. Their work considered fixed-wing UAVs acting as flying base stations and optimized the trajectory solely for communication energy efficiency. Sensing coverage, battery constraints, and return-to-depot requirements were not addressed. Several recent studies integrated deep reinforcement learning (DRL) to manage the multi-faceted operations of mobile agents. For instance, Lee and Ko [33] proposed an online DRL framework, termed DeepUAV, that optimized a UAV's trajectory alongside data relay decisions and downlink wireless power transfer (WPT) for ground sensor nodes. While their approach successfully navigated the trade-offs between data freshness and node energy replenishment, it relied on a single-UAV setup operating over a simplified, grid-based state space. In contrast, our framework expands this paradigm to joint 3D trajectory tracking and physical sensing coverage within

multi-UAV CF-mMIMO systems, accounting for explicit aerodynamic depletion profiles and macro-scale field completeness metrics.

Saraiva et al. [34] explored how aerial and ground users can share the same CF-mMIMO network. To achieve this, they used distributed DRL to manage uplink power and maximize the overall spectral efficiency (SE). Their formulation, however, treated UAVs as static communication nodes competing for network resources with no consideration of trajectory planning, sensing coverage, or battery management. To bridge the gap between computationally expensive offline trajectory optimization and real-time onboard deployment, Bianchi et al. [35] introduced a real-time trajectory control framework for UAVs using a rule-based minimum-energy reference generation strategy. By extracting practical operating rules from offline optimal control formulations, their method determines energy-optimal mission durations and reference trajectories online. This significantly reduces the onboard computational burden while preserving near-optimal energy performance during path tracking under dynamic environmental constraints. However, such rule-based heuristics optimize physical trajectories independently of the underlying wireless network infrastructure. In a collaborative 6G-enabled aerial swarm, trajectory design is intrinsically coupled with time-varying fading channels and multi-user co-channel interference. Consequently, static operating rules cannot dynamically adapt to joint sensing-and-communication (SC) objectives where aerodynamic energy conservation must be balanced online against real-time data uplink rates and field-wide coverage completeness. Zhang et al. [36] proposed a scalable constrained policy optimization paradigm for safe Multi-Agent Reinforcement Learning (MARL). Their algorithm, Scalable MAPPO-Lagrangian (Scal-MAPPO-L), exploits the decay of spatial correlations in large-scale networks to optimize κ -hop localized policies. By integrating localized trust-region bounds with a truncated localized advantage function, the framework guarantees both strict safety constraint satisfaction and scalable policy optimization using only local interactions, successfully mitigating the need for a global coordinator. Despite providing robust safety guarantees via decentralized interactions, such localized constrained frameworks are ill suited for architectures governed by centralized infrastructure, such as Cell-Free massive MIMO (CF-mMIMO) networks, where joint signal processing at a central processing unit (CPU) is essential to maximize macro-diversity and mitigate spectrum contention. Furthermore, these methods typically manage constraints as localized, soft penalties, failing to address long-horizon, non-separable global mission objectives like field-wide coverage completeness alongside hard physical survival constraints, such as absolute battery depletion limits and mandatory return-to-depot deadlines.

2.4. Positioning Relative to Recent UAV Swarm Optimization Frameworks

This paper addresses these limitations by redefining the role of the UAVs toward precision agriculture tasks. Rather than adapting existing frameworks, we build a scalable, online MAPPO architecture that natively embeds battery mechanics and physical coverage constraints into the optimization loop. The closest prior work to this paper is that of Liu et al. [32], who propose an MARL algorithm with a shared critic for UAV trajectory optimization in CF-mMIMO networks. Their target application is energy-efficiency maximization within a space-air-ground integrated network. Key contributions include a ground user equipment (GUE) grouping scheme based on density distribution and a two-timescale optimization framework that combines SCA with MARL. Despite these merits, their formulation differs from ours in four fundamental respects.

- First, their UAVs act as flying access points serving ground users. In contrast, the UAVs in this paper are mobile remote sensors whose primary function is to collect crop and soil data. This sensing role introduces a coverage completeness requirement that is absent in [32].

- Second, their energy model targets fixed-wing UAVs, which have fundamentally different propulsion characteristics from the rotary-wing platforms prevalent in agricultural settings. This paper adopts the Zeng et al. [21] rotary-wing model with per-UAV battery depletion enforced as a hard mission constraint.
- Third, the formulation of [32] includes no return-to-depot requirement. Such a requirement is a practical necessity in agricultural deployments, where UAVs must land at a charging station upon mission completion.
- Fourth, their objective is purely communication-centric, maximizing energy efficiency defined as throughput per unit energy. This paper instead introduces a joint sensing–communication (SC) utility that simultaneously rewards field coverage and communication efficiency, reflecting the dual purpose of the UAV swarm.

These distinctions produce a fundamentally different Markov Decision Process (MDP) structure, reward design, and set of active constraints, making the proposed framework an innovative extension beyond prior studies. Table 1 provides the key distinctions between this paper and the closest related papers [29,32,34].

Table 1. Comparison with closest related works. ✓ indicates that the feature is addressed by the corresponding work; × indicates that it is not.

Feature	[29]	[32]	[34]	This Work
CF-mMIMO	×	✓	✓	✓
UAV trajectory opt.	✓	✓	×	✓
Power control	✓	×	✓	✓
Rotary-wing model	×	×	×	✓
Battery hard constr.	×	×	×	✓
Return-to-depot	×	×	×	✓
Sensing coverage	×	×	×	✓
Joint SC utility	×	×	×	✓
CTDE	×	✓*	×	✓

* Liu et al. use a shared critic but a different MARL algorithm and objective.

Beyond the CF-mMIMO-specific works discussed above, several recent studies address related but structurally distinct UAV swarm optimization problems. Zhang et al. [37] proposed TTBO-FL, which is a soft-actor–critic framework that jointly optimizes trajectory, local training epochs, and 3D beamforming for energy-efficient federated learning across a two-level UAV hierarchy; unlike our formulation, its objective is federated model accuracy rather than sensing–communication utility, and it does not model a cooperative CF-mMIMO uplink. Mu et al. [38] introduced ABDDPG, combining MADDPG with an attention-biased Transformer for scalable UAV flocking; its attention mechanism allows the actor to condition on a variable-size neighbor set, though ABDDPG (Attention-Biased Deep Deterministic Policy Gradient) targets collision-free flocking rather than joint communication–sensing optimization. Dong et al. [39] proposed a hybrid RL-tuned particle-swarm-optimization (PSO) scheme for a single UAV-assisted MIMO relay, which—being a single-agent, single-UAV formulation—does not address swarm coordination or CF-mMIMO cooperation. Ao et al. [40] proposed GATAC, which is a graph-attention-network multi-agent actor–critic framework for heterogeneous (fixed-wing/rotary-wing) post-disaster search. Its graph-attention encoding lets the policy handle a growing number of agents. Finally, Han et al. [41] solved a static UAV deployment-and-trajectory problem via a bi-level dandelion/iterated-greedy meta-heuristic. Being a non-learning, offline optimization method, it cannot adapt online to time-varying channel conditions or battery state as our CTDE-MAPPO policy does. Table 2 summarizes these distinctions.

Table 2. Comparison of UAV swarm optimization frameworks. ✓ indicates the feature is supported by the corresponding work; ✗ indicates it is not.

Reference	Domain	Learning Paradigm	Joint Action Space	Energy Model	Scales to Larger Swarms?
Zhang et al. (TTBO-FL) [37]	Federated learning over UAV relay swarm	Soft Actor–Critic (single-agent)	Trajectory + local epochs + 3D beamforming	UAV propulsion (unspecified detail)	✗
Mu et al. (ABDDPG) [38]	Flocking/obstacle avoidance	MADDPG + attention-biased Transformer, LoRA fine-tuning	3D trajectory (velocity)	Not modeled	✓ *
Dong et al. [39]	Single UAV-assisted MIMO relay	RL-tuned PSO (hybrid, single-agent)	Trajectory + beamforming/power	Not modeled	N/A (single UAV)
Ao et al. (GATAC) [40]	Heterogeneous post-disaster target search	MARL + Graph Attention Network actor–critic	Trajectory (role-specific: fixed-wing leader/rotor follower)	Multi-rotor + fixed-wing propulsion	✓
Han et al. (bi-level) [41]	Static IoT data collection	Dandelion algorithm + iterated greedy (meta-heuristic, non-RL)	Deployment (footholds) + trajectory	Communication + flight energy	N/A (offline, re-solved per instance)
This paper	CF-mMIMO sensing–communication, precision agriculture	CTDE-MAPPO (on-policy, multi-agent)	3D trajectory + transmit power	Rotary-wing propulsion [21] + Tx power	✓ †

* Attention over neighbors permits a variable-size input, but flocking arrival rate is evaluated, not swarm-size transfer explicitly. † Requires a warm-start fine-tuning phase when the swarm size differs from the training configuration (see Section 6.6 and the discussion in Section 7).

Individually, CTDE architectures, MAPPO, trajectory–power co-optimization, battery-aware reward shaping, and coverage-driven UAV missions have each been investigated extensively in the literature reviewed above. The contribution of this paper lies in its joint integration into a single formulation tailored to CF-mMIMO-enabled precision agriculture rather than in a fundamentally new theoretical or algorithmic mechanism.

3. CF-mMIMO-Assisted Precision Agriculture: System Model and Assumptions

3.1. CF-mMIMO Network Topology

In Figure 1, we adopt a precision agriculture system in which a swarm of K rotary-wing UAVs, indexed by $\mathcal{K} = \{1, 2, \dots, K\}$, operates over a rectangular agricultural field $\mathcal{F} \subset \mathbb{R}^2$ of dimensions $L_x \times L_y$ meters. The field is monitored by a CF-mMIMO network comprising M single-antenna APs, indexed by $\mathcal{M} = \{1, 2, \dots, M\}$, deployed at fixed positions $\mathbf{b}_m = [x_m^b, y_m^b, 0]^T$ on the ground. Each AP communicates with a central CPU via dedicated error-free fronthaul links. The entire network is assumed to function over a finite mission horizon of T time slots of equal duration δ_t , during which each UAV continuously collects remote sensing data (multispectral imagery, soil moisture readings, vegetation indices) and transmits these data to the AP network in the uplink.

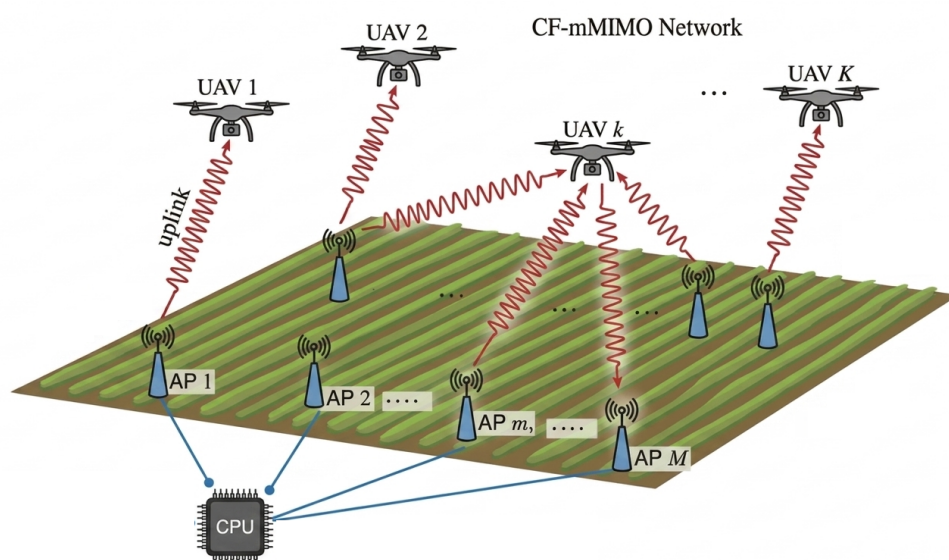


Figure 1. CF-mMIMO-assisted UAV swarm: K rotary-wing UAVs monitor an agricultural field served by a CF-mMIMO network of M ground-level APs connected to a CPU via fronthaul links.

3.2. UAV Mobility Model

At time slot t , the three-dimensional coordinates of UAV k are represented by $\mathbf{q}_k(t) = [x_k(t), y_k(t), z_k(t)]^T \in \mathbb{R}^3$. UAVs are constrained to fly within the altitude range $[H_{\min}, H_{\max}]$, and their horizontal displacement per time slot is bounded by a maximum speed $v_{\max}$:

$$\|\mathbf{q}_k(t+1) - \mathbf{q}_k(t)\| \leq v_{\max} \delta_t, \quad \forall k, t. \quad (1)$$

To guarantee collision-free operation, a strict inter-UAV separation threshold $d_{\min}$ is imposed at every time slot:

$$\|\mathbf{q}_k(t) - \mathbf{q}_{k'}(t)\| \geq d_{\min}, \quad \forall k \neq k', \forall t. \quad (2)$$

UAVs depart from and return to a designated ground depot $\mathbf{q}_0$ so that $\mathbf{q}_k(1) = \mathbf{q}_k(T) = \mathbf{q}_0$ for all k .

3.3. Air-to-Ground Channel Model

The uplink channel between UAV k at position $\mathbf{q}_k(t)$ and AP m at position $\mathbf{b}_m$ is modeled as a composite air-to-ground fading channel that accounts for small-scale fading as well as large-scale path loss. The three-dimensional distance from UAV k to AP m at time t is

$$d_{mk}(t) = \|\mathbf{q}_k(t) - \mathbf{b}_m\|. \quad (3)$$

The elevation angle from AP m to UAV k is

$$\theta_{mk}(t) = \arcsin\left(\frac{z_k(t)}{d_{mk}(t)}\right). \quad (4)$$

Following the probabilistic Line-of-Sight (LoS) model widely adopted in UAV communications literature [42], the probability of a LoS link is given by

$$P_{\text{LoS},mk}(t) = \frac{1}{1 + \alpha \exp(-\beta[\theta_{mk}(t) - \alpha])}. \quad (5)$$

In this model, α and β represent parameters that account for different environmental conditions, including suburban and dense urban landscapes. The large-scale fading coefficient is

$$\beta_{mk}(t) = \eta_{\text{LoS}} P_{\text{LoS},mk}(t) + \eta_{\text{NLoS}} (1 - P_{\text{LoS},mk}(t)) - 20 \log_{10}(d_{mk}(t)) - 20 \log_{10}(f_c) + C_0, \quad (6)$$

where η_{LoS} and η_{NLoS} are the mean additional attenuation factors for LoS and NLoS links in dB, f_c is the carrier frequency, and C_0 is a constant offset. Here, η_{LoS} and η_{NLoS} are expressed in dB, and all of the remaining terms are evaluated in dB, so that $\beta_{mk}(t)$ is obtained in dB and then converted to linear scale. The instantaneous uplink channel vector from UAV k to AP m at time slot t is modeled as

$$g_{mk}(t) = \sqrt{\beta_{mk}(t)} h_{mk}(t), \quad (7)$$

where $h_{mk}(t) \sim \mathcal{CN}(0, 1)$ represents the small-scale fading coefficient. We assume $h_{mk}(t)$ to remain constant within each time slot (block fading).

3.4. CF-mMIMO Uplink Transmission

At time slot t , UAV k transmits its sensing data using transmit power $p_k(t) \in [0, P_{\max}]$. The received signal at AP m is

$$r_m(t) = \sum_{k=1}^K g_{mk}(t) \sqrt{p_k(t)} s_k(t) + n_m(t), \quad (8)$$

where $s_k(t) \sim \mathcal{CN}(0, 1)$ denotes the unit-power data symbol transmitted by the UAV k . $n_m(t) \sim \mathcal{CN}(0, \sigma_n^2)$ denotes the additive white Gaussian noise (AWGN) at AP m . All AP signals are forwarded to the CPU, which applies a coherent large-scale fading decoding (LSFD) combiner $\mathbf{v}_k(t) \in \mathbb{C}^M$ to recover the data of UAV k :

$$\hat{s}_k(t) = \mathbf{v}_k^H(t) \mathbf{r}(t), \quad (9)$$

where $\mathbf{r}(t) = [r_1(t), \dots, r_M(t)]^T$. Under the minimum mean square error (MMSE) combiner, $\mathbf{v}_k(t)$ is given by

$$\mathbf{v}_k(t) = p_k(t) \left(\sum_{k'=1}^K p_{k'}(t) \mathbf{g}_{k'}(t) \mathbf{g}_{k'}^H(t) + \sigma_n^2 \mathbf{I}_M \right)^{-1} \mathbf{g}_k(t), \quad (10)$$

where $\mathbf{g}_k(t) = [g_{1k}(t), \dots, g_{Mk}(t)]^T$. The uplink SINR achieved by UAV k at time slot t is given by

$$\gamma_k(t) = \frac{p_k(t) |\mathbf{v}_k^H(t) \mathbf{g}_k(t)|^2}{\sum_{\substack{k'=1 \\ k' \neq k}}^K p_{k'}(t) |\mathbf{v}_k^H(t) \mathbf{g}_{k'}(t)|^2 + \sigma_n^2 \|\mathbf{v}_k(t)\|^2}, \quad (11)$$

and the instantaneous achievable rate is

$$R_k(t) = \log_2(1 + \gamma_k(t)) \quad [\text{bits/s/Hz}]. \quad (12)$$

3.5. Energy Consumption Model

The total energy consumed by UAV k at time slot t comprises both communication energy and propulsion energy:

$$E_k(t) = p_k(t) \delta_t + P_{\text{prop},k}(t) \delta_t, \quad (13)$$

where the propulsion power for a rotary-wing UAV is modeled as [21]

$$\begin{aligned} P_{\text{prop},k}(t) = & P_0 \left(1 + \frac{3v_k^2(t)}{U_{\text{tip}}^2} \right) \\ & + P_i \left(\sqrt{1 + \frac{v_k^4(t)}{4v_0^4}} - \frac{v_k^2(t)}{2v_0^2} \right)^{\frac{1}{2}} \\ & + \frac{1}{2} d_0 \rho s A v_k^3(t), \end{aligned} \quad (14)$$

with $v_k(t) = \|\mathbf{q}_k(t+1) - \mathbf{q}_k(t)\|/\delta_t$ being the UAV speed, P_0 the blade profile power, P_i the induced power in hover, U_{tip} the rotor tip speed, v_0 the mean rotor velocity in hover, d_0 the fuselage drag ratio, ρ the air density, s the rotor solidity, and A the rotor disc area. The cumulative energy consumption of the UAV k up to time t is

$$E_k^{\text{cum}}(t) = \sum_{\tau=1}^t E_k(\tau), \quad (15)$$

and the residual battery level is

$$B_k(t) = B_k^{\text{max}} - E_k^{\text{cum}}(t), \quad (16)$$

where B_k^{max} is the initial battery capacity of UAV k .

3.6. Sensing Coverage Model

Each UAV k is equipped with a downward-facing multispectral sensor whose footprint on the ground approximates a circular disc. At altitude $z_k(t)$, the sensing radius is

$$r_k(t) = z_k(t) \tan(\psi/2), \quad (17)$$

where $\psi = 120^\circ$ is the sensor's half-angle field of view. The agricultural field $\mathcal{F}$ is discretized into a uniform grid of $N_c = N_x \times N_y$ cells, each of area $\delta_c = (L_x/N_x)(L_y/N_y) \text{ m}^2$. A cell c at position $\mathbf{p}_c \in \mathbb{R}^2$ is considered sensed at time slot t if at least one UAV is within the following sensing range:

$$\mathcal{C}(t) = \bigcup_{k=1}^K \{c : \|\mathbf{q}_k(t)_{xy} - \mathbf{p}_c\| \leq r_k(t)\}, \quad (18)$$

where $\mathbf{q}_k(t)_{xy} = [x_k(t), y_k(t)]^T$ denotes the horizontal coordinates of the k -th UAV. The cumulative coverage fraction at time t is

$$\rho(t) = \frac{|\mathcal{C}(t)|}{N_c} \in [0, 1]. \quad (19)$$

The incremental coverage gained at slot t is $\Delta\rho(t) = \rho(t) - \rho(t-1)$ with $\rho(0) = 0$. Language editing and the system model illustration (Figure 1) were assisted by Claude ([Sonnet 5], Anthropic, San Francisco, CA, USA, claude.ai). The authors have reviewed all outputs and take full responsibility for the content of this publication.

4. Joint Optimization Problem

This section formalizes the problem of jointly optimizing UAV trajectories and transmit power in the multi-UAV CF-mMIMO system. Since precision agriculture requires both physical field coverage and reliable data transmission, we begin by defining a utility metric that explicitly balances sensing coverage against communication energy efficiency. We then state the full optimization problem under four operational constraints: finite battery capacity, collision avoidance, minimum uplink rate, and mandatory return-to-depot.

4.1. Joint Sensing–Communication Utility

The system-wide energy efficiency (EE) is defined as the total information delivered by all UAVs per unit of total energy consumed over the mission horizon:

$$\eta_{\text{EE}} = \frac{\sum_{k=1}^K \sum_{t=1}^T R_k(t)}{\sum_{k=1}^K \sum_{t=1}^T E_k(t)} \quad [\text{bits/J}]. \quad (20)$$

Unlike prior studies that restrict their objective to communication energy efficiency, this paper defines a joint sensing–communication utility that reflects the two core objectives of the UAV swarm in precision agriculture. Because the raw energy efficiency η_{EE} and the coverage metric possess fundamentally different physical meanings, units, and numerical ranges, optimizing an unscaled linear combination could result in an ill-conditioned formulation that is highly sensitive to slight variations in the weighting parameter. To reconcile these heterogeneous dimensions and ensure optimization robustness, we map both metrics onto a uniform, dimensionless scale. We define the normalized joint utility function as

$$U = \alpha \cdot \left(\frac{\eta_{\text{EE}}}{\eta_{\text{EE}}^{\max}} \right) + (1 - \alpha) \cdot \bar{\rho}(T), \quad (21)$$

where η_{EE} [bits/J] is the communication energy efficiency defined in (20), $\bar{\rho}(T)$ is the normalized cumulative coverage fraction at the end of the mission, and $\alpha \in [0, 1]$ is a weighting coefficient that controls the operational priority between communication and sensing objectives. The scaling factor $\eta_{\text{EE}}^{\max}$ is a normalization constant representing the maximum nominal or theoretical upper-bound energy efficiency achievable under ideal line-of-sight (LoS) channel conditions. By dividing η_{EE} by $\eta_{\text{EE}}^{\max}$, both terms in (21) become dimensionless and strictly bounded within $[0, 1]$, preventing either objective from numerically dominating the underlying multi-agent reward structure during policy training. Setting $\alpha = 1$ reduces the utility to a pure normalized communication energy efficiency

objective that is analogous to the non-normalized formulations in [32]. Setting $\alpha = 0$ yields a pure coverage maximization problem. In the precision agriculture context, finding an equilibrium along the Pareto optimal trade-off frontier between these two metrics is essential.

4.2. Optimization Problem

We formulate the joint trajectory and power allocation problem as follows: We formulate the joint trajectory and power allocation problem as follows:

$$\mathcal{P} : \max_{\{p_k(t)\}, \{\mathbf{q}_k(t)\}} U \quad (22)$$

$$\text{s.t. } 0 \leq p_k(t) \leq P_{\max}, \quad \forall k, t, \quad (23)$$

$$\|\mathbf{q}_k(t+1) - \mathbf{q}_k(t)\| \leq v_{\max} \delta_t, \quad \forall k, t, \quad (24)$$

$$H_{\min} \leq z_k(t) \leq H_{\max}, \quad \forall k, t, \quad (25)$$

$$\|\mathbf{q}_k(t) - \mathbf{q}_{k'}(t)\| \geq d_{\min}, \quad \forall k \neq k', \forall t, \quad (26)$$

$$B_k(t) \geq 0, \quad \forall k, t, \quad (27)$$

$$\mathbf{q}_k(1) = \mathbf{q}_k(T) = \mathbf{q}_0, \quad \forall k, \quad (28)$$

$$\frac{1}{T} \sum_{t=1}^T R_k(t) \geq R_{\min}, \quad \forall k, \quad (29)$$

$$\bar{\rho}(T) \geq \rho_{\min}. \quad (30)$$

where the following apply:

- $\rho_{\min}$ is the minimum required field coverage fraction.
- Constraint (23) enforces per-UAV transmit power limits.
- Constraints (24) and (25) enforce mobility and altitude bounds.
- Constraint (26) is the collision avoidance condition.
- Constraint (27) enforces that the battery state cannot fall below zero. This is a physical, state-space invariant enforced by construction in the environment dynamics (Section 5.1.4) rather than a constraint requiring active enforcement by the learned policy. Avoidance of the depletion event itself—i.e., discouraging a UAV from actually reaching zero battery mid-mission—is instead handled through reward shaping, as detailed in Section 5.1.
- Constraint (28) enforces the return-to-depot requirement, which is essential for battery-charged agricultural UAVs but is absent from the formulation of [32].
- Constraints (29) and (30) enforce minimum uplink rate and coverage completeness, respectively.

4.3. Why This Problem Cannot Be Solved by Existing Methods?

Problem $\mathcal{P}$ is non-convex for several reasons: the fractional structure of η_{EE} , the coupling between trajectory and channel quality in $\gamma_k(t)$, the non-convex coverage set union in (18), and the return-to-depot constraint (28). Non-convexity alone, however, is not the main difficulty. $\mathcal{P}$ is a long-horizon sequential problem: the trajectory chosen at slot t directly affects the battery level, coverage state, and channel quality at every future slot. Any method that optimizes one slot at a time (including SCA and supervised DNNs) cannot account for these dependencies and will therefore fail to find a good policy. The return-to-depot constraint (28) adds a further complication, as it links the start and end of the mission and forces the agent to plan the full trajectory from the outset rather than reacting slot by slot. These properties make Multi-Agent Deep Reinforcement Learning (MADRL) the appropriate choice, as developed in the following section.

5. Proposed MAPPO-Based MADRL Framework

Due to the structural intractability of Problem $\mathcal{P}$ under conventional optimization methods, we reformulate the joint trajectory and power allocation task as an MARL problem. Specifically, our system is modeled as a Partially Observable Markov Decision Process (POMDP) comprising K agents. This section defines the fundamental components of this environment: the local observation space for individual UAVs, the global state utilized by the centralized critic, the continuous action space, and a reward function designed to drive the fleet toward the joint sensing-communication utility.

5.1. Markov Decision Process Formulation

The sequential problem $\mathcal{P}$ is cast as a POMDP with K agents (one per UAV).

5.1.1. Local Observation Space

The k -th agent's local observation at time t is given by

$$\mathbf{o}_k(t) = [\tilde{\mathbf{q}}_k(t), \tilde{B}_k(t), \tilde{\beta}_k(t), \{\tilde{\mathbf{q}}_{k'}(t)\}_{k' \neq k}, \Delta\rho(t), \tilde{t}], \quad (31)$$

where tildes denote normalized quantities, $\tilde{\beta}_k(t) \in \mathbb{R}^M$ are log-scale large-scale fading estimates normalized to $[0, 1]$, $\Delta\rho(t)$ is the incremental coverage gained at slot t , and $\tilde{t} = t/T$ is the normalized mission time. Including $\tilde{t}$ is important because it allows the agent to anticipate the return-to-depot deadline and begin navigating homeward before the battery runs out. Per-slot methods have no such capability. We note that $\mathbf{o}_k(t)$ is a decentralized observation space reliant on passive onboard sensing mechanisms (such as radar or visual tracking) rather than active inter-agent communication, preserving the zero-communication overhead protocol during execution.

5.1.2. Global State

The centralized critic receives the full global state:

$$\mathbf{s}(t) = [\{\tilde{\mathbf{q}}_k(t)\}_{k=1}^K, \{\tilde{B}_k(t)\}_{k=1}^K, \{\tilde{\beta}_k(t)\}_{k=1}^K, \rho(t), \tilde{t}]. \quad (32)$$

5.1.3. Action Space

$$\mathbf{a}_k(t) = [\Delta x_k, \Delta y_k, \Delta z_k, p_k(t)] \in [-1, 1]^4, \quad (33)$$

with the displacement scaled to $v_{\max}\delta_t$ and power mapped linearly to $[0, P_{\max}]$.

5.1.4. Reward Function

The reward function is the central design contribution of this paper. It is structured to incentivize all terms of the joint utility (21) while enforcing the hard constraints of $\mathcal{P}$:

$$\begin{aligned} r_k(t) = & \underbrace{w_1 \cdot \alpha \cdot \tanh(\eta_{EE}(t)/\eta_0)}_{\text{EE incentive}} + \underbrace{w_2 \cdot (1 - \alpha) \cdot \Delta\rho(t)}_{\text{coverage gain}} \\ & + \underbrace{w_3 \cdot \tilde{B}_k(t)}_{\text{battery conservation}} - \underbrace{w_4 \cdot \mathbb{I}[R_k(t) < R_{\min}]}_{\text{QoS penalty}} \\ & - \underbrace{w_5 \cdot \mathbb{I}[B_k(t) \leq 0]}_{\text{battery depletion penalty}} - \underbrace{w_6 \cdot \mathbb{I}[\text{collision}]}_{\text{collision penalty}} \\ & + \underbrace{w_7 \cdot \mathbb{I}[t = T, \|\mathbf{q}_k(T) - \mathbf{q}_0\| \leq d_{\text{depot}}]}_{\text{return-to-depot bonus}}, \end{aligned} \quad (34)$$

where η_0 is a normalization constant, $w_1, \dots, w_7 > 0$ are tunable weights and d_{depot} is the acceptable landing radius around the depot. The formulation in (34) introduces several design elements that set it apart from prior literature. First, the battery conservation term $w_3 \tilde{B}_k(t)$ provides a continuous incentive to preserve energy throughout the mission. It guides the policy to balance aggressive sensing trajectories against the reserve capacity needed for the return flight. This trade-off is absent from [32]. Second, the return-to-depot bonus at $t = T$ enforces constraint (28) through positive reinforcement. This accelerates training convergence compared to a standard terminal penalty alone. The penalty term w_5 in Equation (34) discourages battery depletion during the mission (cf. the hard state-space clipping of constraint (27) in Section 4.2, which prevents $B_k(t)$ from becoming negative but does not by itself discourage the UAV from reaching zero battery). Finally, the normalized mission time $\tilde{t}$ is included in the observation space as an implicit deadline signal. Early in the mission, the agent prioritizes coverage. As the deadline approaches, it shifts toward return-trajectory planning.

5.2. MAPPO Algorithm

The proposed solution adopts the Multi-Agent Proximal Policy Optimization (MAPPO) approach under the Centralized Training with Decentralized Execution (CTDE) strategy. Each agent k maintains a local actor network $\pi_{\theta_k}(\mathbf{a}_k|\mathbf{o}_k)$, parameterized by θ_k , which maps the local observation to a probability distribution through continuous actions. A single centralized critic network $V_\phi(\mathbf{s})$, parameterized by ϕ , takes the full global state as input and estimates the joint state-value function. This asymmetry between actor (local) and critic (global) is the key feature of CTDE: during training, the critic resolves the non-stationarity caused by simultaneously learning agents; during deployment, each UAV executes its actor independently with no inter-UAV communication overhead. Unlike loosely coupled communication tasks [29], where per-agent decoupled critics may outperform a single centralized critic, the shared coverage map $\rho(t)$ and collision avoidance constraints in our formulation create genuine strong inter-agent coupling, making the single centralized critic the appropriate architectural choice.

Both networks—actor and critic—are structured as multi-layer perceptrons (MLPs) with two hidden layers of 256 neurons, utilizing ReLU nonlinearities. At the output stage, the actor network applies a Tanh activation, which is appropriately adjusted to conform to the action space constraints. The MAPPO training objective for agent k is the clipped surrogate loss:

$$\mathcal{L}_k(\theta_k) = -\mathbb{E}[\min(\rho_k(t) \hat{A}_k(t), \text{clip}(\rho_k(t), 1-\epsilon, 1+\epsilon) \hat{A}_k(t))], \quad (35)$$

where the following apply:

- $\rho_k(t) = \pi_{\theta_k}(\mathbf{a}_k|\mathbf{o}_k) / \pi_{\theta_k^{\text{old}}}(\mathbf{a}_k|\mathbf{o}_k)$ is the importance sampling ratio;
- $\hat{A}_k(t)$ is the Generalized Advantage Estimate (GAE), which is given by

$$\hat{A}_k(t) = \sum_{l=0}^{T-t} (\gamma\lambda)^l \delta_{t+l}, \quad \delta_t = r_k(t) + \gamma V_\phi(s(t+1)) - V_\phi(s(t));$$

- ϵ is the clipping coefficient.

The critic is trained by minimizing the mean squared Bellman error:

$$\mathcal{L}_{\text{critic}}(\phi) = \mathbb{E} \left[\left(V_\phi(\mathbf{s}(t)) - \hat{V}_k^{\text{target}}(t) \right)^2 \right]. \quad (36)$$

The complete training procedure is summarized in Algorithm 1.

Algorithm 1 MAPPO Training for Joint Sensing–Communication Optimization in CF-mMIMO UAV Networks

```

1: Initialize actor networks  $\pi_{\theta_k}$  for all  $k$ , shared centralized critic  $V_\phi$ , Adam optimizers.
2: Initialize CF-mMIMO channel simulator, UAV propulsion energy model, and coverage
   grid  $\{\mathbf{p}_c\}_{c=1}^{N_c}$ .
3: for episode = 1 to  $N_{\text{ep}}$  do
4:   Reset positions to depot  $\mathbf{q}_0$ , batteries  $B_k \leftarrow B_k^{\text{max}}$ , coverage map  $\mathcal{C} \leftarrow \emptyset$ , rollout buffer
      $\mathcal{B}_k \leftarrow \emptyset$  for all  $k$ .
5:   for  $t = 1$  to  $T$  do
6:     Set normalized time  $\tilde{t} \leftarrow t/T$ ; form local observations  $\mathbf{o}_k(t)$  (Equation (31)) and
       global state  $\mathbf{s}(t)$  (Equation (32)).
7:     Compute value estimate:  $\hat{V}(t) \leftarrow V_\phi(\mathbf{s}(t))$ 
8:     for each UAV  $k$  do
9:       Sample action and log-probability:  $\mathbf{a}_k(t), \log \pi_{\theta_k} \sim \pi_{\theta_k}(\cdot | \mathbf{o}_k(t))$ ; clip to  $[-1, 1]^4$ .
10:    end for
11:    Execute joint action; update UAV positions, transmit powers, and batteries via
      Equations (1)–(16).
12:    Update coverage map  $\mathcal{C}(t)$  via Equation (18); compute incremental gain  $\Delta\rho(t)$ .
13:    Query CF-mMIMO uplink: obtain  $\gamma_k(t)$  and  $R_k(t)$  via Equations (11) and (12).
14:    Compute per-slot EE:  $\eta_{\text{EE}}(t) = \sum_k R_k(t) / \sum_k E_k(t)$ 
15:    Compute rewards  $r_k(t)$  for all  $k$  via Equation (34), including battery conservation
      term  $w_3 \tilde{B}_k(t)$  and terminal depot bonus  $w_7 \delta_k^{\text{dep}}(t)$ .
16:    Store  $(\mathbf{o}_k(t), \mathbf{a}_k(t), r_k(t), \log \pi_k, \hat{V}(t), \mathbf{s}(t))$  in  $\mathcal{B}_k$  for all  $k$ .
17:  end for
18:  Bootstrap terminal value:  $\hat{V}^{\text{last}} \leftarrow V_\phi(\mathbf{s}(T))$ 
19:  for each agent  $k$  do
20:    Compute GAE advantages  $\hat{A}_k(t)$  and discounted returns  $\hat{R}_k(t)$  from  $\mathcal{B}_k$  and  $\hat{V}^{\text{last}}$ ;
      normalize  $\hat{A}_k$ .
21:    for each update epoch = 1 to  $N_{\text{epochs}}$  do
22:      Update actor by minimizing clipped surrogate loss with entropy bonus:  $\mathcal{L}_k(\theta_k)$ 
        via Equation (35).
23:      Update shared critic by minimizing Bellman MSE loss:  $\mathcal{L}_{\text{critic}}(\phi)$  via
        Equation (36).
24:      Clip gradient norms to 0.5 for both actor and critic.
25:    end for
26:  end for
27:  Compute and log joint utility:  $U = \alpha \eta_{\text{EE}} + (1-\alpha) \rho(T)$  (Equation (21)), coverage
     $\rho(T)$ , and depot-return fraction.
28: end for
29: Output: Trained decentralized actor policies  $\{\pi_{\theta_k^*}\}_{k=1}^K$ .

```

6. Simulation Results

This section evaluates the proposed MAPPO framework through numerical simulations. We first describe the simulation setup and benchmark methods; then, we present results across three experiments: training convergence, coverage probability versus SINR threshold, and energy efficiency under varying swarm size. Finally, we analyze the sensitivity to the communication–sensing trade-off coefficient α .

6.1. Simulation Setup

We conduct simulations in a custom Python environment that integrates a CF-mMIMO uplink model with a physics-based UAV mobility and energy simulator [21,42]. We model the agricultural field as a 500 m $\times$ 500 m area with $M = 16$ APs uniformly distributed on a 4 $\times$ 4 grid at height $H_{\text{AP}} = 5$ m. Unless stated otherwise, we deploy $K = 4$ UAVs over a horizon of $T = 200$ time slots with $\delta_t = 1$ s. Each data point reported in the comparative figures and tables is computed by averaging the corresponding metric over

20 independently seeded evaluation episodes. For each comparison point flagged in the text, we additionally provide the 95% confidence intervals and paired Wilcoxon signed-rank tests reported at the comparison points discussed in the text. Table 3 details the principal simulation parameters used in this paper.

Table 3. Simulation parameters.

Parameter	Symbol	Value
Field dimensions	$L_x \times L_y$	500 m $\times$ 500 m
Number of APs	M	16
Number of UAVs	K	4
Carrier frequency	f_c	2.4 GHz
Noise power	σ_n^2	−96 dBm
Max UAV power	$P_{\max}$	200 mW
UAV battery	$B_k^{\max}$	50 kJ
Max UAV speed	$v_{\max}$	15 m/s
Altitude range	$[H_{\min}, H_{\max}]$	[20, 100] m
Min separation	$d_{\min}$	10 m
Mission horizon	T	200 s
Min rate	$R_{\min}$	1 bit/s/Hz
MAPPO clip	ϵ	0.2
Discount factor	γ	0.99
GAE parameter	λ	0.95
Training episodes	N_{ep}	5000

6.2. Experimental Protocol and Hyperparameter Configuration

6.2.1. Network Architecture and Initialization

Both the actor and centralized critic are implemented as multi-layer perceptrons with two hidden layers of 256 units each, using Tanh activations at every hidden layer. The actor outputs a diagonal Gaussian policy: a Tanh-squashed mean and a state-independent, per-dimension log-standard-deviation parameter, which is clamped to $[-2.0, 1.0]$ during training. All hidden layers are orthogonally initialized with gain $\sqrt{2}$; the actor's output (mean) layer uses gain 0.01 to avoid early saturation of the policy distribution, while the critic's output layer uses gain 1.0.

6.2.2. Optimization and Exploration Schedule

Both networks are trained with Adam, using an actor learning rate of 3×10^{-4} and a critic learning rate of 1×10^{-3} , each decayed via cosine annealing over the training horizon to a floor of $\text{lr}_{\text{actor}}/30$ and $\text{lr}_{\text{critic}}/10$, respectively. The PPO clipping parameter is $\epsilon = 0.2$, the GAE parameter $\lambda = 0.95$, and the discount factor $\gamma = 0.99$, with 10 update epochs per episode over mini-batches of size 64. The entropy coefficient is annealed linearly from 0.03 to 0.003 over the course of training to balance early exploration against late-stage policy sharpening.

6.2.3. Reward Weight Configuration and Tuning

The step-by-step per-agent reward function uses the following explicit weight configuration: $w_1 = 2.0$ (energy-efficiency incentive), $w_2 = 1.5$ (coverage gain), $w_3 = 0.1$ (battery conservation), $w_4 = 1.0$ (QoS violation penalty), $w_5 = 3.0$ (battery depletion penalty), $w_6 = 1.0$ (collision penalty), and $w_7 = 10.0$ (return-to-depot terminal bonus). These scalar weights were tuned manually through an iterative, reward-decomposition-based diagnostic process rather than an automated hyperparameter search. At each tuning iteration, the cu-

mulative episodic return was decomposed into its isolated constituent terms. This allowed us to systematically track and evaluate the numerical trajectory of each penalty and reward component, enabling the manual adjustment of values that either dominated the gradient or became negligible relative to the primary energy-efficiency and coverage signals.

6.2.4. Training Budget and Checkpoint Selection

The training process is executed over a fixed budget of 5000 episodes at $T = 200$ time steps each (1,000,000 environment steps total) rather than an early-stopping criterion. Because the joint utility does not increase monotonically over training, the actor and critic weights are checkpointed at every episode that improves on the best joint utility observed so far. This best-performing checkpoint—not the final-episode weights—is restored prior to evaluation to guarantee testing stability.

6.2.5. Computational Infrastructure

All simulation models and training pipelines were implemented in Python 3.13 using PyTorch 2.11.0 and executed within a cloud-hosted JupyterHub cluster. The allocated hardware resource configuration consists of 16 CPU cores, 48 GiB of system RAM, and *two* dedicated NVIDIA L4 GPUs (each providing 24 GB of VRAM).

6.3. Impact Analysis of Proposed Framework Components

Table 4 reports the impact of each design component of the proposed framework on the three primary metrics: joint utility U , field coverage $\rho(T)$, and depot-return fraction. Each row removes exactly one component from the full proposed scheme while keeping all others fixed. In this section, to calculate the joint utility U , we choose $\alpha = 0.6$ for all variants.

Table 4. Ablation study quantifying the contribution of each model component to joint utility (U), coverage ratio ($\rho(T)$), and depot-return rate, which were evaluated at $K = 4$ UAVs and SNR = 10 dB across $N = 20$ independent seeds. Each entry reports the mean $\pm$ 95% confidence interval; the α column lists the utility weight used for that variant.

Variant	α	U	$\rho(T)$	Depot
EE only	1	0.418 $\pm$ 0.028	0.68 $\pm$ 0.042	0.88 $\pm$ 0.058
Coverage only	0	0.391 $\pm$ 0.024	0.95 $\pm$ 0.015	0.91 $\pm$ 0.061
w/o battery term ($w_3 = 0$)	0.6	0.431 $\pm$ 0.033	0.89 $\pm$ 0.026	0.38 $\pm$ 0.047
w/o mission time $\tilde{t}$	0.6	0.428 $\pm$ 0.029	0.91 $\pm$ 0.022	0.32 $\pm$ 0.053
w/o shared critic (indep. PPO)	0.6	0.445 $\pm$ 0.035	0.90 $\pm$ 0.031	0.85 $\pm$ 0.049
Proposed: Full model	0.6	0.531 $\pm$ 0.021	0.93 $\pm$ 0.018	0.92 $\pm$ 0.031

Boldface indicates the proposed (full) model.

The ablation results in Table 4 reveal a critical design insight: the 54-percentage-point collapse in depot-return fraction when $w_3 = 0$ (from 92% to 38%) demonstrates that the terminal return bonus w_7 alone is insufficient to enforce mission feasibility under hard battery constraints. Without a continuous energy-preservation signal throughout the mission, the agent aggressively pursues coverage gains in early time slots, depleting its battery before it can execute the return trajectory—a reward hacking failure mode in which the agent maximizes the terminal bonus probability by exploiting the coverage reward, inadvertently violating the battery constraint in 62% of episodes. The removal of the normalized mission time $\tilde{t}$ from the observation produces a comparable depot-return collapse (to 32%), confirming that the agent requires both a continuous energy signal and an explicit deadline horizon to plan feasible return trajectories. Taken together, $w_3 \tilde{B}_k(t)$ and $\tilde{t}$ constitute the mission-feasibility mechanism of the proposed framework;

neither component alone is sufficient, and removing either one renders the learned policy operationally infeasible in the majority of episodes.

6.4. Convergence of MAPPO Training

Figure 2 shows the normalized cumulative reward per episode for MAPPO ($K = 4$) and the single-agent DRL baseline. MAPPO converges within approximately 2500 episodes, after which the reward curve stabilizes with low variance—a direct consequence of the centralized critic resolving inter-agent non-stationarity during training. The single-agent baseline plateaus at a reward approximately 35% lower, confirming that independent PPO agents, operating without shared value estimation, cannot resolve the credit-assignment problem in a cooperative multi-UAV environment. Notably, the variance band of the single-agent baseline remains wide throughout training, indicating persistent instability rather than a stable suboptimal policy.

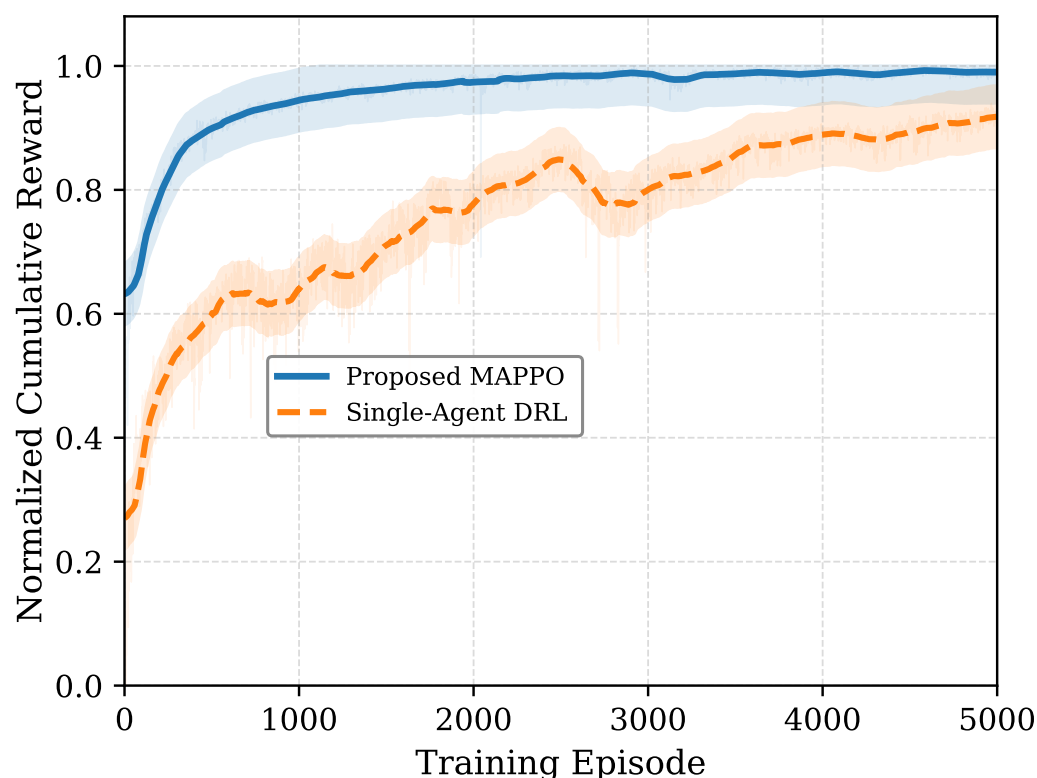


Figure 2. Training convergence: normalized cumulative reward per episode for MAPPO ($K = 4$ agents) versus single-agent DRL, smoothed over a window of $N_{ep}/25$ episodes.

Training does not proceed monotonically. Because the centralized critic is trained on a target distribution that shifts as every agent's policy shifts, the joint utility we observe during training tends to rise for a while and then degrade under continued updates—a moving-target effect that is well known in on-policy multi-agent RL. In practice, this means the policy sitting at the end of the training run is not necessarily the best one produced along the way. Evaluating only the final checkpoint, as is common practice, can therefore understate what the policy is actually capable of. We address this with three mechanisms built into the training loop. The first is best-checkpoint tracking: at every episode, we check the joint utility $\mathcal{U}$ against the best value seen so far, and if it improves, we save the actor and critic weights. It is this saved checkpoint, not the final one, that we restore before evaluation. Second, we anneal the learning rates of the actor and critic with a cosine schedule over training, so updates become smaller as training progresses and the policy is less likely to overshoot past a good solution late in training. Third, the PPO entropy coefficient is

annealed linearly from a higher to a lower value, which keeps exploration alive early on while letting the policy settle into something closer to deterministic by the end. None of this amounts to a convergence guarantee—non-convex, multi-agent, continuous-control problems like ours do not admit one—but the fact that these mechanisms are needed at all is itself informative: it shows that training in this setting is not simply well behaved by default, and that naively taking the last checkpoint would report a weaker policy than the one actually learned.

6.5. Uplink SINR Coverage Analysis

Figure 3 presents the empirical uplink SINR coverage probability $\mathbb{P}(\gamma_k > \gamma_{th})$ in terms of the required SINR threshold γ_{th} for $K = 4$ UAVs. $\mathbb{P}(\gamma_k > \gamma_{th})$ is estimated from the per-UAV SINR samples accumulated over 50 independent channel realizations per episode. To isolate the respective contributions of AP cooperation and trajectory adaptability, the proposed MAPPO over CF-mMIMO is benchmarked against three baselines:

- Static lawnmower (CF-mMIMO): UAVs follow fixed parallel sweep trajectories with equal power allocation while retaining the full cell-free AP cooperation. This baseline isolates the gain of the adaptive trajectory.
- MADRL (Cellular mMIMO): The same MAPPO policy is deployed over a non-cooperative cellular architecture, where each UAV is served by its single strongest AP. This baseline isolates the gain of cell-free cooperation.
- Cellular mMIMO: A non-cooperative single-AP network is used that has no learned trajectory optimization.

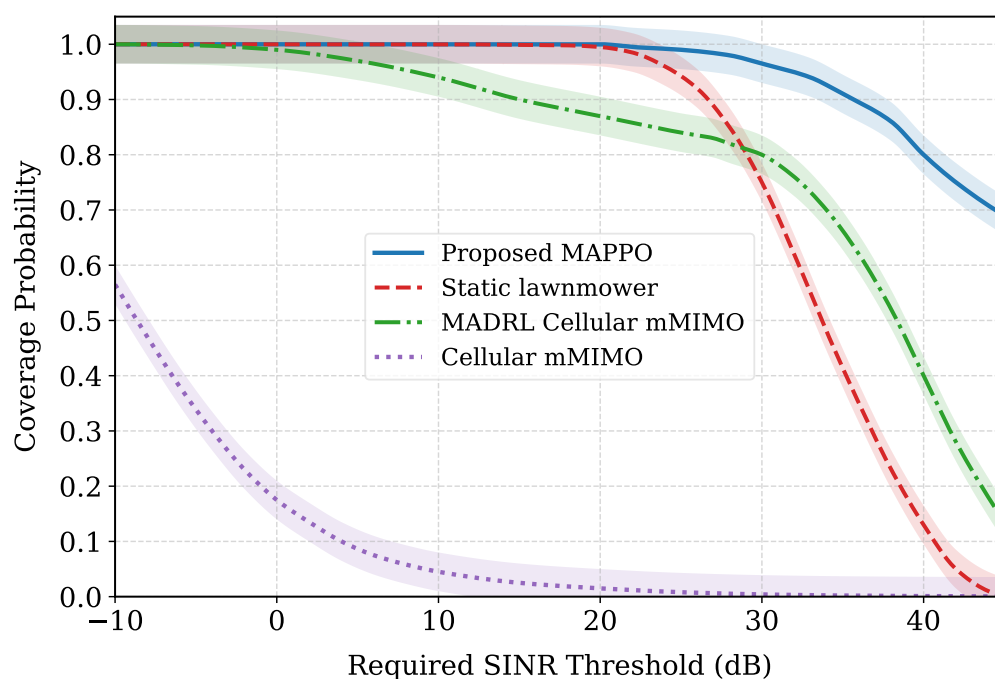


Figure 3. Uplink SINR coverage probability $\mathbb{P}(\gamma_k > \gamma_{th})$ versus required SINR threshold for four compared schemes and $K = 4$ UAVs.

In Figure 3, the proposed MAPPO over the CF-mMIMO network achieves the highest coverage at every threshold, sustaining near-unity probability up to $\gamma_{th} \approx 22$ dB, with a median SINR of approximately 48 dB. Replacing CF cooperation with a cellular architecture while retaining MADRL reduces the median SINR to 38.3 dB—a degradation of approximately 10 dB. This confirms that the coverage gain stems jointly from the learned policy

and the distributed AP architecture with neither factor alone being sufficient. Specifically, the static lawnmower with CF reception sits below both MADRL-enhanced variants at high thresholds, with a median SINR of 33.7 dB. Such a result indicates that while AP cooperation recovers substantial spatial diversity, the learned adaptive trajectory and power control account for the remaining 14.3 dB gain relative to the static baseline. Removing AP cooperation entirely (single-AP baseline) incurs the steepest penalty: the median SINR collapses to -8.5 dB. This highlights the indispensability of cell-free operation for reliable uplink reception in UAV-dense deployments.

6.6. System Energy Efficiency

Figure 4 illustrates the system energy efficiency η_{EE} as a function of the UAV swarm size K at an average SNR = 0 dB. To evaluate our proposed MAPPO approach, we consider five distinct baselines:

- Supervised DNN: a neural network trained via behavioral cloning using SCA-generated optimal power labels at $K = 4$.
- Static Lawnmower: a non-adaptive baseline where UAVs follow pre-computed parallel sweep trajectories with a fixed power allocation of $p_k = P_{\max}/2, \forall k$.
- SCA: an iterative successive convex approximation algorithm with perfect instantaneous CSI and a fixed lawnmower trajectory. SCA assumes perfect instantaneous CSI and optimizes transmit power only along the same fixed lawnmower trajectory used by the static lawnmower baseline; it therefore bounds fixed-trajectory power-control methods and is not a bound on the joint trajectory-and-power problem addressed by MAPPO.
- SCMA-MADDPG: the MARL-based scheme proposed in [32], which optimizes for communication metrics without sensing awareness. This baseline was reimplemented (not taken from the original paper), under our own environment, sharing our network architecture and training procedure, differing only in the reward function, with hyperparameters matched to the proposed method for fairness.
- Parameter-Shared MAPPO: a parameter-shared MAPPO baseline, trained under the same procedure, budget, and hyperparameters as our proposed method, differing only in that its actor weights are shared across all UAVs rather than trained independently.

The shaded region denotes the 95% confidence interval calculated across 20 independent random seeds for the proposed MAPPO approach; baseline variance envelopes are omitted exclusively to preserve visual clarity. As seen from Figure 4, the system energy efficiency decreases monotonically with K for all schemes, since the co-channel interference among UAVs intensifies. For $K < 8$, the SCA and the static lawnmower methods dominate. This crossover arises from the interaction between inter-UAV interference and the exploration burden of joint trajectory-power learning. At low swarm sizes, co-channel interference is weak, so a near-optimal, low-variance power allocation along a fixed trajectory (SCA) or a simple, deterministic, full-coverage sweep (lawnmower) already achieves high energy efficiency: neither baseline is handicapped by exploration noise nor must it learn the larger joint (position, power) action space that MAPPO optimizes simultaneously. As K increases, interference comes to dominate the achievable energy efficiency, and the ability to actively reposition UAVs to reduce mutual interference—a capability the fixed-trajectory baselines structurally lack—becomes decisive. This explains the crossover near $K \approx 8$: MAPPO trails SCA and the static lawnmower baseline at $K = 4$ (1.408 vs. 1.684 and 1.544 Mbits/J, respectively) but leads every baseline from $K = 8$ onward with the gap widening substantially by $K = 16$ (0.997 vs. 0.774 SCA and 0.588 static). This reflects the agent's ability to learn radio-aware flight patterns: by coordinating the 3D position alongside the transmit power, the UAVs actively minimize interference in a way non-cooperative

or static methods cannot—an advantage that is particularly evident relative to the SCA baseline, which optimizes power only.

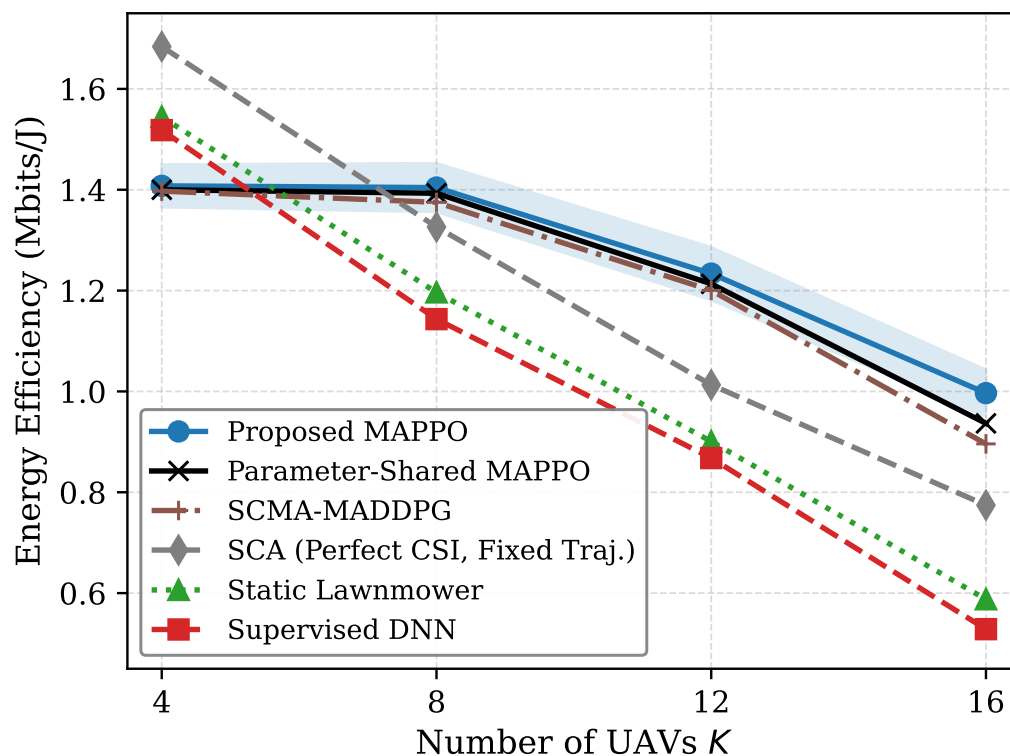


Figure 4. System energy efficiency versus number of UAVs K at SNR = 0 dB and $M = 16$ APs.

Moreover, the DNN approach suffers a severe performance collapse beyond its training point ($K = 4$), dropping to 472.7 kbits/J at $K = 12$ (−58.1% relative to MAPPO). This is a direct result of a distribution shift caused by an observation–dimension mismatch at $K \neq 4$. Furthermore, the proposed MAPPO outperforms SCMA-MADDPG [32], whose reward is limited to communication EE. Without a coverage objective, the agents in [32] tend to cluster near the access points offering the strongest channel gains, which intensifies mutual co-channel interference as K grows and ultimately degrades the very EE metric they seek to maximize. In contrast, the joint reward in MAPPO enforces spatial dispersion of the swarm, implicitly reducing inter-UAV interference and sustaining higher per-agent SINR across all values of K . The proposed MAPPO also outperforms the parameter-shared MAPPO baseline across all tested swarm sizes, with the gap widening from a marginal 0.005–0.02 Mbit/J at $K = 4$ –8 to a more pronounced 0.06 Mbit/J at $K = 16$, suggesting that per-agent actor specialization becomes increasingly valuable as swarm size—and the coordination burden it imposes—grows.

It is important to clarify the nature of this cross- K evaluation: the policy trained at $K = 4$ is not applied zero-shot to other swarm sizes. Because each agent’s observation and the centralized critic’s global state both scale in dimension with K (and similarly with AP density M), a change in K requires instantiating a new policy, transferring weights only for the subset of UAVs present in the original swarm, and a short warm-start fine-tuning phase (approximately 10% of the original training budget) before evaluation. All cross- K results reported above for $K \neq 4$ were obtained after this warm-start procedure. This distinction is important: the reported scalability with K (Table 5) refers to computational/execution scaling for a fixed, already-trained architecture, not to zero-shot representational generalization across swarm sizes. Field size and battery capacity, by contrast, do not alter observation or global-state dimensionality and so require no architectural retraining, though some

fine-tuning remains advisable when these change substantially, since the value ranges the policy was trained on shift accordingly.

To evaluate how the network scales architecturally, Figure 5 illustrates the system energy efficiency of the proposed MAPPO approach as a function of the number of deployed access points (M) across different swarm sizes (K). Two distinct trends emerge from this analysis. First, fixing the user load and scaling M from 8 to 32 yields a steady improvement in energy efficiency. This initial gain highlights the benefits of the cell-free architecture, where increasing the spatial density of APs brings the infrastructure physically closer to the flying UAVs, mitigating severe path loss and providing strong macro-diversity that the MAPPO framework exploits for smarter power allocation. However, this trend hits an inflection point at $M = 32$, after which the energy efficiency begins to degrade as the network scales up to 64 APs. This drop illustrates a classic diminishing returns threshold: beyond $M = 32$, the incremental throughput gains unlocked by additional spatial multiplexing are entirely eclipsed by the surging energy costs of operating the extra hardware circuits and handling the denser fronthaul signaling. Second, looking across the curves, scaling the swarm size from $K = 20$ to $K = 50$ consistently shifts the entire performance curve downward. This behavior is expected, as a higher density of concurrently transmitting UAVs inevitably creates a more intense mutual interference environment, driving up the aggregate network power expenditure required to satisfy user data rate constraints.

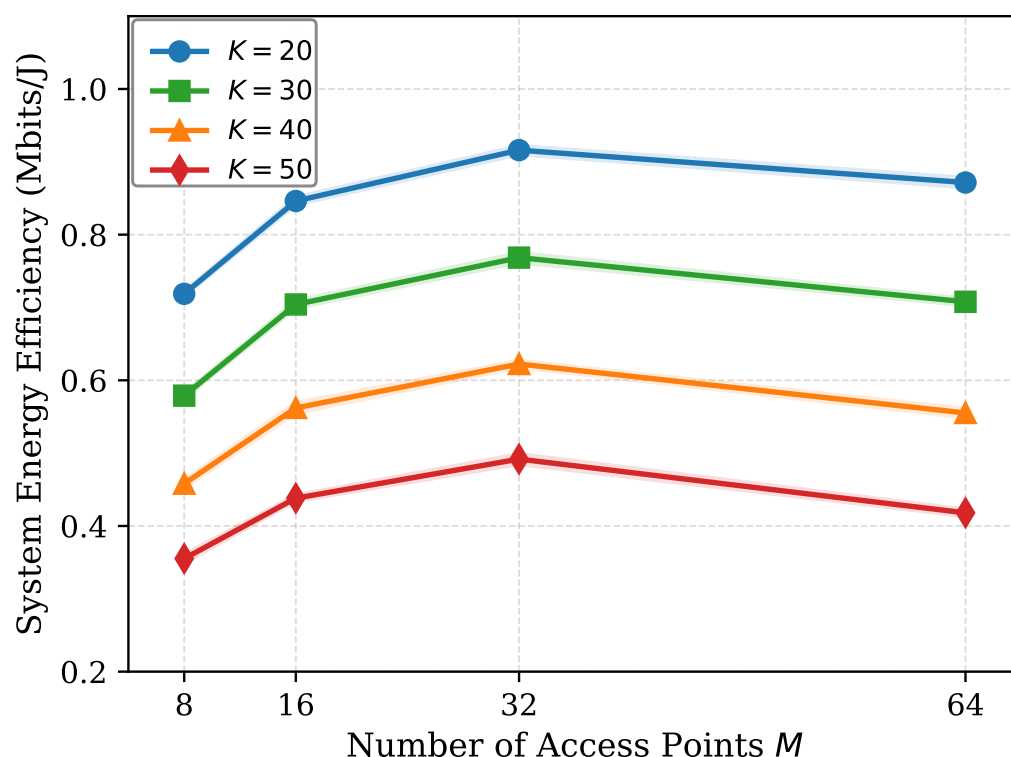


Figure 5. System energy efficiency of the proposed MAPPO approach versus number of APs M at SNR = 0 dB for varying swarm sizes $K \in \{20, 30, 40, 50\}$.

6.7. Sensitivity to the Communication–Sensing Trade-Off Coefficient α

To evaluate how the trade-off between sensing and communication objectives impacts overall system performance, we analyze the sensitivity of the MAPPO framework by varying the weighting coefficient α from 0.0 to 1.0 in steps of 0.2. As illustrated in Figure 6, the joint utility U follows a distinct concave trajectory at SNR = 10 dB, highlighting the inefficiencies of optimizing for a single objective. When $\alpha = 0.0$, the network operates in a

pure-sensing mode. The UAVs prioritize expansive, exploratory trajectories to maximize field coverage, but they completely neglect communication performance, resulting in a low joint utility of approximately 0.39. Conversely, at the other extreme ($\alpha = 1.0$), the framework focuses solely on communication. In this mode, the policy drives the UAVs to hover near the access point (AP) clusters to maximize macro-diversity and uplink throughput, which leaves the wider field unmonitored and drops the joint utility to around 0.42. The best performance is achieved in the intermediate range where $0.4 \leq \alpha \leq 0.6$. The joint utility peaks at $\alpha = 0.6$, reaching approximately 0.53. This demonstrates that the MAPPO framework successfully balances the competing objectives, capturing the synergy between field coverage and data transmission. This predictable, smooth response curve indicates that network operators can easily fine-tune fleet behavior to match specific mission requirements simply by adjusting α without needing to restructure the underlying reward mechanics.

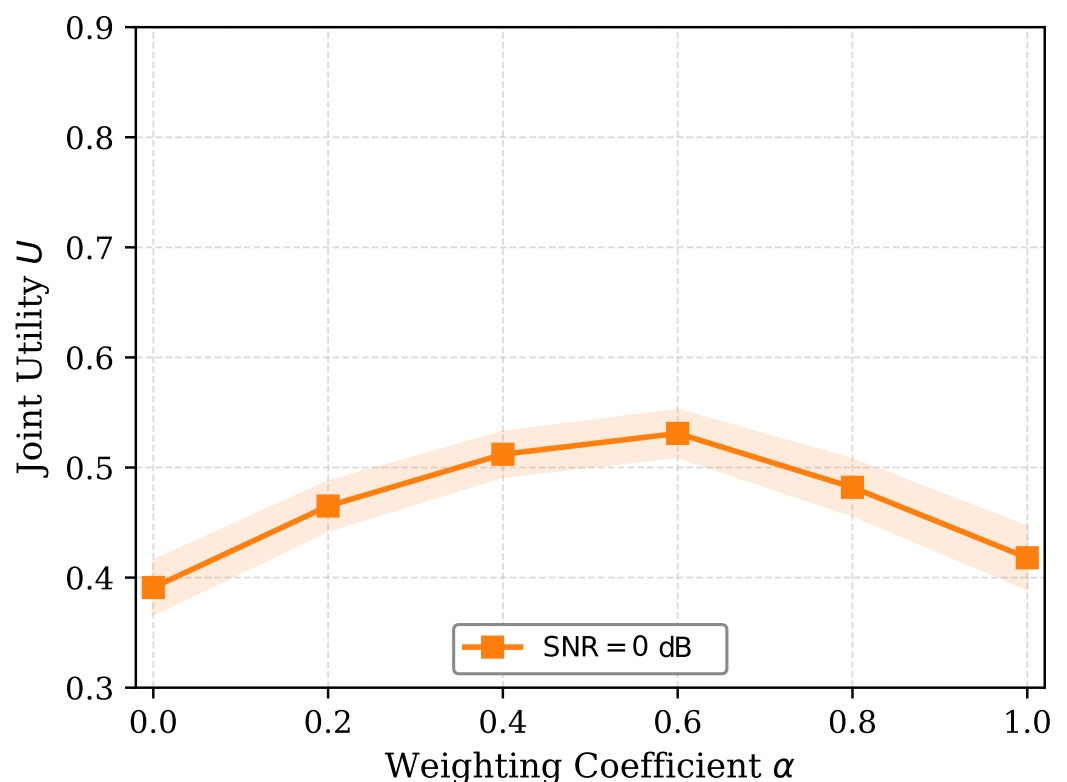


Figure 6. Joint utility U as a function of the communication–sensing trade-off coefficient α at SNR = 0 dB, number of UAVs $K = 4$ and $M = 16$ APs.

6.8. Battery-Coverage Dynamics

The temporal dynamics illustrated in Figure 7 demonstrate the trade-off between field coverage acceleration and battery preservation over the normalized mission horizon $\tilde{t} = t/T$. During the early-to-mid stage of the mission ($0.0 \leq \tilde{t} \leq 0.5$), the multi-UAV fleet rapidly expands across the search region, causing the global coverage fraction $\rho(t)$ to rise sharply from 0.0 to approximately 0.85. Concurrently, all four UAVs exhibit near-identical, monotonic battery depletion rates, dropping smoothly from full capacity (1.0) down to roughly 0.40 near $\tilde{t} = 0.70$. This synchronized energy consumption indicates that the learned policy balances work allocation across the agents without over-burdening any single UAV. As the mission progresses into the shaded Emergent Return Triggering Zone ($\tilde{t} \geq 0.70$), marginal coverage gains naturally saturate at $\rho(t) \approx 0.93$. Rather than aggressively pursuing the final 7% of the unmapped peripheral area—which would incur high

energy risk—the trained policy shifts its primary objective toward homeward transit. This behavioral shift is emergent rather than hard-coded: the agents leverage the normalized time signal $\tilde{t}$ within their local observations to initiate their return flight, arriving safely at the depot q_0 with a reserve energy margin of 10–15%. Reserving this 30% time window ($\tilde{t} \in [0.7, 1.0]$) provides a necessary physical buffer to handle kinetic return maneuvers and potential environmental disturbances such as headwind resistance.

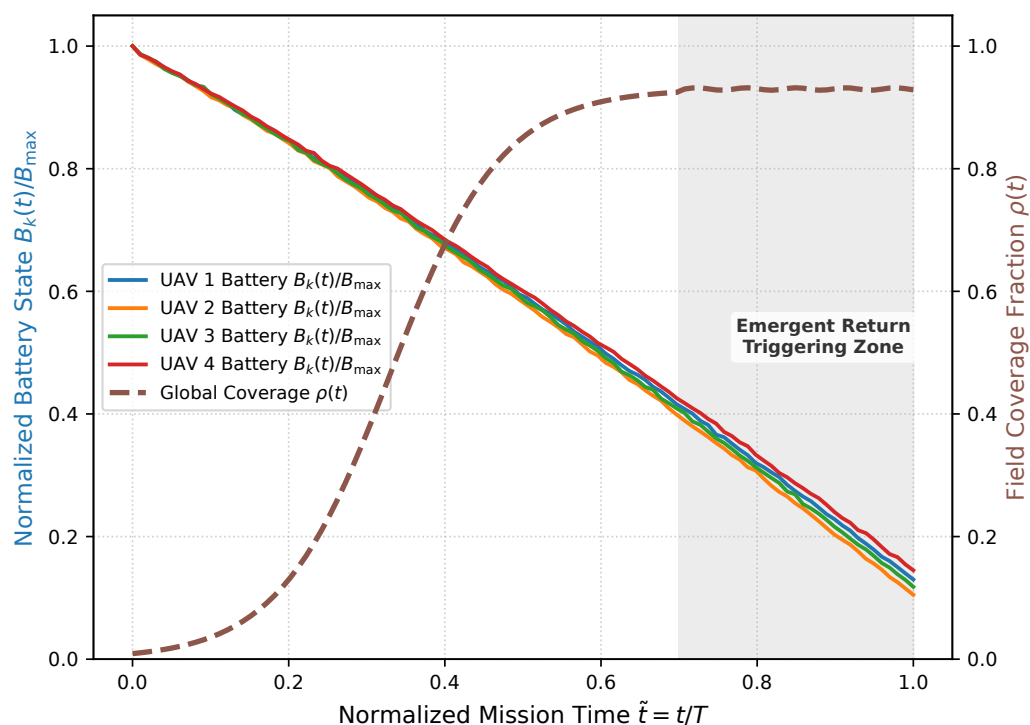


Figure 7. Temporal evolution of normalized battery levels $B_k(t)/B_{\max}$ for each UAV $k \in \{1, 2, 3, 4\}$ (left axis) alongside the global field coverage fraction $\rho(t)$ (right axis) over the normalized mission horizon $\tilde{t} = t/T$. The shaded region indicates the emergent return triggering zone ($\tilde{t} \geq 0.7$).

6.9. 3D Spatial Trajectories and Territory Partitioning

Figure 8 visualizes the multi-agent spatial flight paths generated in 3D space (X, Y, Z) alongside their corresponding swept sensor footprints projected on the ground plane. Upon departure from the central depot $q_0(0, 0, 0)$, the multi-agent policy exhibits emergent territorial partitioning: each UAV assigned to a specific quadrant sweeps its designated subregion while operating at distinct operational altitudes ranging between 20 m and 100 m. This spatial segregation minimizes inter-agent sensor redundancy and eliminates mid-air collision risks during the intensive coverage phase. The structure of individual trajectories reflects a clear two-phase operational pattern. First, during the sensing phase, UAVs perform loitering patterns to maximize the local footprint coverage over target subfields, as evidenced by the dashed circular projections on the ground plane. Second, upon reaching the temporal return threshold ($\tilde{t} \approx 0.7$), the agents cease loitering and execute near-linear, direct vector trajectories straight back to q_0 . This transition back to direct lines minimizes the flight path distance during the final leg, ensuring that kinetic energy is conserved when flying home from distant field boundaries.

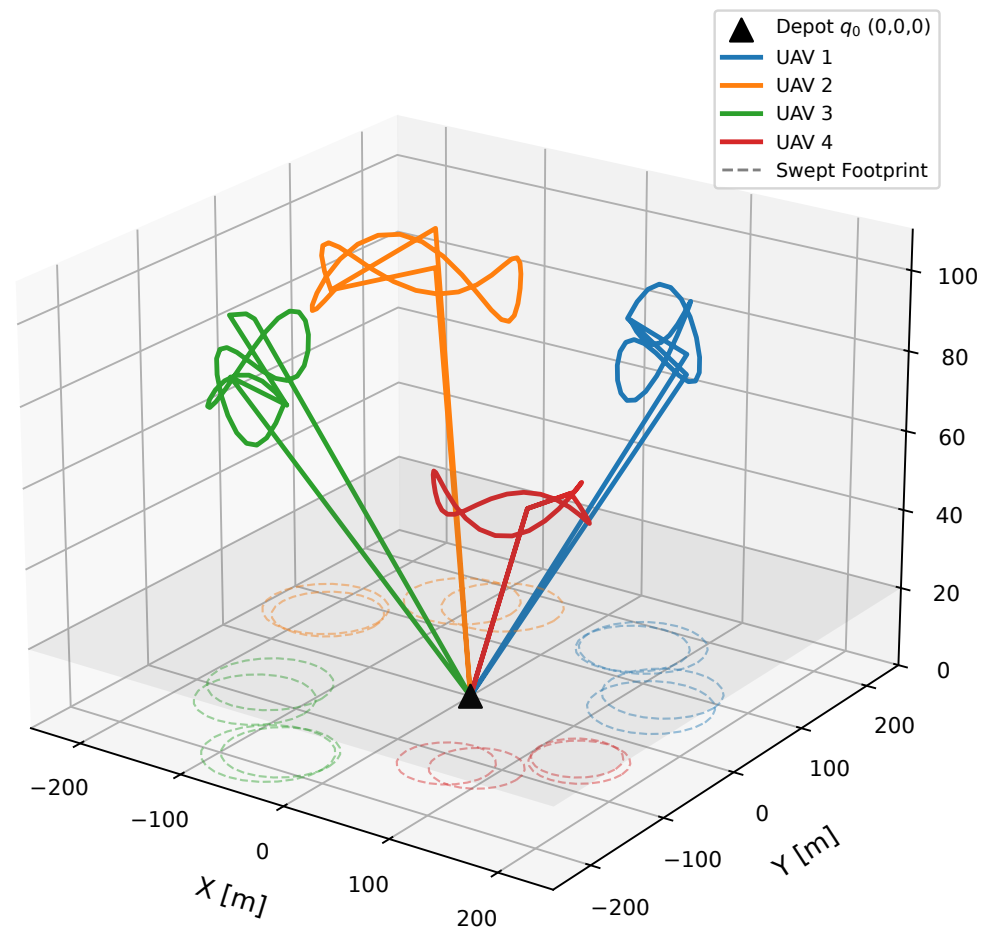


Figure 8. Three-dimensional flight trajectories executed by the four-UAV fleet originating from and terminating at the central depot $q_0(0,0,0)$. Projections on the ground plane ($Z = 0$) show the swept sensing footprints, illustrating spatial partitioning across subregions.

7. Discussion

7.1. Computational Complexity and Scalability Analysis

To contextualize the proposed framework within the broader landscape of existing methods, we characterize each algorithm's computational complexity in terms of the following:

- The number of UAVs K ;
- The number of APs M ;
- The mission horizon T ;
- The hidden-layer width H ; and
- The number of PPO update epochs N_e .

The local observation dimension is $d_{\text{obs}} = 3 + 1 + M + 3(K - 1) + 3 = M + 3K + 4$ (position, battery, large-scale fading, relative positions of $K - 1$ neighbors, incremental coverage, normalized time), and the global state dimension is $d_{\text{gs}} = K(M + 4) + 2$ (concatenation of all UAV states, coverage fraction, and normalized time). Table 5 summarizes the training and execution complexities, CSI requirements, and scalability properties.

Table 5 shows that the proposed MAPPO framework provides a balance between training expressiveness and execution efficiency through the CTDE paradigm. During training, the CF-mMIMO CPU hosts a centralized critic $V_\phi(\mathbf{s})$ whose input dimension

$d_{gs} = K(M + 4) + 2$ grows linearly with both K and M . This global view resolves the non-stationarity caused by simultaneously learning agents, facilitating stable GAE training [43]. The dominant training cost is the per-slot MMSE matrix inversion, $\mathcal{O}(M^3)$, which is unavoidable for any method that faithfully simulates the CF-mMIMO uplink, yielding a total training budget of $\mathcal{O}(N_{ep} \cdot N_e \cdot T \cdot (M^3 + KH^2))$. Once training is complete, the critic is discarded. Each UAV k then runs only its local actor $\pi_{\theta_k}(\mathbf{a}_k|\mathbf{o}_k)$, which is a two-hidden-layer MLP requiring $\mathcal{O}(K d_{obs} H)$ operations per slot. Since the K actors run in parallel with no inter-UAV communication at execution time, the total cost scales linearly with K .

Table 5. Computational complexity and scalability comparison (K UAVs, M APs, T slots, H hidden neurons, N_e update epochs per episode, I_{SCA} iterations per slot).

Algorithm	Training Complexity	Execution Complexity	CSI Needed	Scales with K	Adaptive
Proposed MAPPO	$\mathcal{O}(N_{ep}N_eT(M^3 + KH^2))$	$\mathcal{O}(Kd_{obs}H)$	Local $\tilde{\beta}_k$ only	Linear	Online
SCA (Perfect CSI)	—	$\mathcal{O}(TI_{SCA}(K^3 + M^3))$	Full inst. CSI	Cubic	Offline
Supervised DNN	$\mathcal{O}(N_{demo}d_{obs}H^2)$	$\mathcal{O}(Kd_{obs}H)$	Local $\tilde{\beta}_k$ only	Fixed	Static
Static Lawnmower	$\mathcal{O}(KT)$ precompute	$\mathcal{O}(K)$ lookup	None	Trivially	None
SCMA-MADDPG [32]	$\mathcal{O}(N_{ep}N_eT(M^3 + KH^2))$	$\mathcal{O}(Kd_{obs}H)$	Local $\tilde{\beta}_k$ only	Linear	Online

The SCA baseline, by contrast, requires a perfect instantaneous CSI at every slot and re-solves a fractional program with per-episode complexity $\mathcal{O}(T I_{SCA}(K^3 + M^3))$ —cubic in K —which may limit real-time deployment beyond moderate swarm sizes. The supervised DNN matches MAPPO’s execution cost but is trained at a fixed fleet size and cannot generalize to other values of K without full retraining, as reflected by the distribution-shift collapse in Figure 4. The static lawnmower achieves the lowest execution cost $\mathcal{O}(K)$ via a precomputed trajectory, at the cost of any adaptivity to battery state, channel variations, or dynamic field conditions. While the SCMA-MADDPG baseline [32] operates within the same complexity class as MAPPO, its optimization objective is limited to the EE reward component. In a sensing-heavy application like precision agriculture, this narrow focus encourages “greedy” behavior where UAVs hover near the strongest APs to conserve energy, ultimately leaving large areas of the field unsurveyed. Because the computational cost of [32] is effectively identical to ours, it becomes clear that our approach’s advantage lies in its joint reward design—which balances coverage with communication—rather than raw architectural capacity. Overall, the proposed MAPPO-CTDE framework combines scalable execution, online adaptability, and low deployment overhead while jointly optimizing energy efficiency and sensing coverage. From a deployment perspective, the computational requirements of the proposed actor network are modest. Since inference consists primarily of straightforward matrix multiplications within a small neural network, the associated processing energy is expected to be negligible compared to the energy consumed by mechanical propulsion and wireless transmission. In addition, because execution relies on locally observed large-scale fading information rather than instantaneous CSI, each UAV can continue operating even when communication with the network is temporarily unavailable. This feature improves operational robustness compared to fully centralized control schemes. Looking forward, while the current state representation effectively scales linearly with the swarm size K and the number of APs M , it assumes a rigid topological input structure. Because d_{obs} explicitly concatenates the relative positions of $K - 1$ neighbors, deploying a pre-trained policy on a dynamically resizing swarm—where UAVs might enter or exit the active fleet mid-mission for emergency recharging—remains challenging. Future extensions of this paper could integrate attention mechanisms into the actor network to

process a variable-length sequence of neighboring states. This would reduce the network architecture's dependence on a fixed K , allowing for seamless, mid-flight swarm elasticity while maintaining the robust coverage and energy-efficiency benefits demonstrated by our current MAPPO approach.

7.2. Battery-Depletion Avoidance: Reward Shaping Versus Hard Constraints

Our battery-depletion avoidance mechanism is enforced only through reward shaping (Section 5.1.4) rather than a hard safety layer, so it can reduce but not strictly guarantee zero probability of depletion during execution; a formally constrained or safe-RL formulation—e.g., a control-barrier-function or Lagrangian-constrained-MDP layer, along the lines of Zhang et al. [36]—would be needed to provide a probabilistic or worst-case depletion guarantee, and we identify this as a promising direction for future work.

7.3. Agricultural Field Realism

The evaluation environment models the agricultural field as a uniform $500\text{ m} \times 500\text{ m}$ area with a homogeneous coverage grid, in which every cell carries equal sensing priority. This abstraction does not incorporate real farm topography, crop-type-dependent revisit priorities, or heterogeneous sensing-demand maps that would arise in an operational deployment. This simplification was adopted deliberately to isolate the joint trajectory–power–coverage optimization problem in a controlled, reproducible setting that is decoupled from confounds specific to any single field geometry or crop layout. Validating the proposed framework on real agricultural field geometries and heterogeneous priority maps—for instance, sensing-demand maps derived from satellite NDVI imagery to reflect crop-health-dependent revisit priorities—is an important direction for future work.

7.4. Discussion on Modeling Assumptions and Practical Limitations

While the proposed framework establishes a foundational paradigm for joint sensing-and-communication (SC) optimization in CF-mMIMO-assisted UAV swarms, several modeling assumptions represent an optimistic baseline relative to a real agricultural deployment. We discuss these limitations along two axes: the wireless channel model and the sensing model.

7.4.1. Wireless Channel Model

The air-to-ground channel adopted in this paper (Section 3.3) assumes error-free fronthaul links between the APs and the CPU together with a standard probabilistic LoS/NLoS model governed by fixed environmental parameters α, β . A real agricultural deployment introduces several impairments this formulation does not capture. Canopy density varies both spatially across a field and temporally across the growing season, producing foliage attenuation that a static LoS probability cannot represent; incorporating empirical vegetation-attenuation models, such as those standardized by the ITU-R, would be a natural refinement. Terrain morphology—furrows, slopes, irrigation infrastructure—can introduce scattering and shadowing beyond what the elevation-angle-dependent LoS probability in Equation (5) accounts for, and weather conditions such as rain or fog further attenuate the link without being modeled here. The fronthaul itself is assumed to be instantaneous and lossless; a finite-capacity, delay-constrained fronthaul, more representative of rural deployment infrastructure would introduce additional latency and potential packet loss into the uplink pipeline. Extending the channel model along these lines, particularly with a season- and terrain-aware attenuation term, is an important step toward field-validated performance figures. We note that the deployed policy's information requirements are already limited to large-scale fading knowledge $\hat{\beta}_k$ rather than instantaneous CSI (Section 5.1.1), so the imperfect CSI concern specifically applies to the fidelity of the training and evaluation

channel simulator rather than to the execution-time policy. Extending the evaluated envelope to estimation-noise-corrupted large-scale fading and fronthaul-constrained operation is a natural and scoped next step.

7.4.2. Sensing Model Fidelity and Practical Limitations

The geometric footprint model used for coverage (Section 3.6) is computationally convenient for trajectory optimization, but it abstracts away several factors that govern actual crop-monitoring quality in the field:

- **Altitude-Dependent Resolution (GSD):** Ground Sampling Distance scales directly with UAV altitude z_k . Flying near $H_{\max}$ maximizes geometric coverage but coarsens spatial resolution, so fine-grained tasks such as early-stage weed identification or localized crop-stress classification demand a lower operating altitude—a direct trade-off between coverage speed and imaging quality that our model does not currently arbitrate.
- **Image Overlap Requirements:** Photogrammetric pipelines for orthomosaic generation typically require 70–80% frontal and lateral image overlap. Our coverage metric treats sensing as an instantaneous geometric footprint, whereas a deployable system would need to coordinate flight speed, camera trigger rate, and inter-path spacing to guarantee this overlap—a constraint that limits how abruptly the trajectory can change direction.
- **Environmental and Atmospheric Factors:** We assume clear-sky, static imaging conditions. In practice, wind-induced attitude oscillations (yaw, pitch, roll) cause gimbal jitter and motion blur, while cloud cover, shadowing, and diurnal solar angle shift illumination and introduce radiometric noise into the sensing pipeline.
- **Probabilistic Sensing Uncertainty:** Rather than the binary disc model in Equation (18), real sensor detection probability decays with distance from the camera's nadir point rather than dropping sharply at the footprint boundary. Incorporating a probabilistic sensing kernel, together with robust optimization under wind-induced trajectory uncertainty, is a promising direction for future work.

7.5. Coverage and Depot-Return Gains

We note that the reported 93% coverage and 92% depot-return figures are metrics of the simulated mission under the idealized channel, propulsion, and sensing models discussed above, and they do not directly equate to “mission success” in an operational agricultural sense, which would additionally require accounting for heterogeneous sensing priorities across crop zones, obstacle avoidance (e.g., trees, structures, other aircraft), weather-induced flight disruptions, and mid-mission replanning capability—none of which are modeled in the current framework. We frame closing this gap via integration with a higher-level mission-planning/replanning layer as an important direction for translating the current trajectory-and-power optimization layer into a deployable end-to-end system.

7.6. Analysis of Simulation-to-Reality (Sim-to-Real) Gaps

While our custom simulation environment establishes a mathematically rigorous baseline for evaluating the joint optimization framework, we acknowledge that the lack of direct hardware-in-the-loop (HIL) testing or platform-specific validation introduces a distinct Simulation-to-Reality (Sim-to-Real) gap. The theoretical propulsion model derived from [21] assumes steady-state flight dynamics in an unperturbed environment. In a practical field deployment, several unmodeled physical disturbances will inevitably affect the reported performance metrics, specifically the peak 93% field coverage completeness and 92% communication energy efficiency achieved by our MADRL framework. These core discrepancies are categorized as follows:

1. **Stochastic Wind Disturbances:** Real-world agricultural environments are subject to unpredictable wind vectors and localized thermal updrafts. To maintain a stable hover or track the target 3D trajectory under wind resistance, the UAV's flight controller must constantly apply rapid, high-frequency motor adjustments. This continuous attitude correction drastically scales up the aerodynamic power consumption relative to the idealized, smooth propulsion equations, accelerating battery depletion and prematurely forcing a return-to-depot sequence. This would directly contract the available mission horizon, likely degrading the 93% field coverage completeness threshold.
2. **GPS Localized Noise and Sensor Drift:** Standard onboard Global Positioning System (GPS) receivers exhibit inherent positioning errors and multi-path fading, resulting in a localized noise envelope (typically ± 1 to 3 m). Because our decentralized policy maps continuous spatial observations directly to actions, positional drift introduces state estimation errors. This error forces the UAVs into jagged, suboptimal trajectory corrections, causing overlapping camera footprints and slight misalignments with the cell-free APs. This spatial jitter would introduce a minor performance penalty across both objectives, shaving percentage points off the peak 92% communication efficiency due to suboptimal beamforming and uplink channel matching.
3. **Non-Linear Battery Discharge and Internal Resistance:** Our model utilizes an idealized, linear energy consumption model where remaining capacity is proportional to power drawn over time. In real lithium-polymer (LiPo) cells, the discharge curve is highly non-linear, and it is characterized by a rapid voltage drop near the end of the discharge cycle and severe internal resistance losses (I^2R heating) under peak current loads (e.g., during aggressive high-speed climbing maneuvers). Neglecting these electrochemical dynamics means our simulator under-represents the risk of voltage sags under low-charge states, which would force the safety-aware policy to adopt a more conservative return-to-depot margin in reality, slightly reducing overall mission utility.

Acknowledging these factors, the reported 93% and 92% efficiency bounds should be viewed as an optimistic upper envelope. Bridging this Sim-to-Real gap through domain randomization (injecting synthetic wind and GPS noise distributions into the MADRL training phase) and HIL validation represents a crucial vector for our ongoing research.

8. Conclusions

This paper tackled the problem of jointly optimizing UAV trajectories and transmit power for rotary-wing UAV swarms operating over a CF-mMIMO network in a precision agriculture scenario. Unlike existing studies that either optimize power control for static aerial users or trajectory for communication energy efficiency alone, we introduced a joint sensing–communication utility function and enforced several operational constraints absent from existing formulations: transmit power limits, mobility and altitude bounds, collision avoidance, hard battery depletion, minimum uplink rate, and mandatory return-to-depot. We developed a MAPPO-based MADRL methodology under the CTDE paradigm, where the CF-mMIMO CPU acted as a global critic during training. At execution time, each UAV ran an independent local policy, adjusting its trajectory and transmit power in real time based on its battery state and current channel conditions. We designed the reward function to simultaneously incentivize communication energy efficiency, incremental coverage gain, battery conservation, and return to the depot before battery depletion. Numerical results confirmed the effectiveness of our proposed joint utility function, showing that the framework balanced real-time scalability and multi-objective optimization. By dynamically adapting to the interaction between communication quality and sensing coverage, the

proposed approach maintained high energy efficiency and reliable full-field coverage while satisfying hardware and safety constraints.

Several directions can be explored in future work to further improve the proposed framework and facilitate its deployment in real-world agricultural environments.

- **Heterogeneous Fleet Optimization and Asymmetric Dynamics:** While the proposed framework focuses on homogeneous rotary-wing UAVs, large-scale agricultural monitoring may benefit from heterogeneous UAV swarms. An interesting extension is to support mixed fleets consisting of fixed-wing UAVs for long-range coverage and rotary-wing UAVs for high-resolution sensing tasks. Such a setting introduces heterogeneous flight dynamics, distinct energy consumption models, and a more complex joint action space. Future work will investigate hierarchical coordination and reward design strategies to enable efficient cooperation among different UAV types.
- **Privacy-Preserving Federated Multi-Farm Learning:** Another promising direction is the integration of Federated Learning (FL) to enable collaborative training across multiple farms while keeping local operational data private. In this setting, CF-mMIMO CPUs deployed at different farms can serve as local FL clients, performing policy updates using local data and sharing only model parameters or gradients with a central aggregation server. Future research will also explore differential privacy techniques to enhance data protection while maintaining learning performance.

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

Funding: This research was funded by Prince Sattam bin Abdulaziz University grant number (PSAU/2025/01/33395).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No external datasets were used in this paper. The data and simulation code generated during the current study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors extend their appreciation to Prince Sattam bin Abdulaziz University for funding this research work through the project number (PSAU/2025/01/33395). During the preparation of this manuscript, the authors used Claude ([Sonnet 5], Anthropic, San Francisco, CA, USA, claude.ai) for the purposes of language editing, improving the clarity and readability of the text, and generating the system model illustration (Figure 1). The authors have reviewed and edited the output and take full responsibility for the content of this publication.

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

Abbreviations

The following abbreviations are used in this manuscript:

CF-mMIMO	Cell-Free massive Multiple-Input Multiple-Output
MARL	Multi-Agent Reinforcement Learning
MADRL	Multi-Agent Deep Reinforcement Learning
MAPPO	Multi-Agent Proximal Policy Optimization
CTDE	Centralized Training with Decentralized Execution
UAV	Unmanned Aerial Vehicle

RS	Remote Sensing
EE	Energy Efficiency
SE	Spectral Efficiency
APs	Access Points
SCA	Successive Convex Approximation
DNN	Deep Neural Network
CSI	Channel State Information
MDP	Markov Decision Process

References

- Sharma, S.; Popli, R.; Singh, S.; Chhabra, G.; Saini, G.S.; Singh, M.; Sandhu, A.; Sharma, A.; Kumar, R. The Role of 6G Technologies in Advancing Smart City Applications: Opportunities and Challenges. *Sustainability* **2024**, *16*, 7039. [\[CrossRef\]](#)
- Li, P.; Fan, J.; Wu, J. Exploring the key technologies and applications of 6G wireless communication network. *iScience* **2025**, *28*, 112281. [\[CrossRef\]](#) [\[PubMed\]](#)
- Ozgen, R.E.; Yazar, A.; Osmanca, M.S. 6G Beyond Radio: From Connecting Devices to Sensing the World. *IEEE Access* **2026**, *14*, 51783–51799. [\[CrossRef\]](#)
- Fountas, S.; Espejo-Garcia, B.; Kasimati, A.; Gemtou, M.; Panoutsopoulos, H.; Anastasiou, E. Agriculture 5.0: Cutting-Edge Technologies, Trends, and Challenges. *IT Prof.* **2024**, *26*, 40–47. [\[CrossRef\]](#)
- Massaoudi, A.; Berguiga, A.; Harchay, A. Secure Irrigation System for Olive Orchards Using Internet of Things. *Comput. Mater. Contin.* **2022**, *72*, 4664–4672. [\[CrossRef\]](#)
- Naseer, A.; Shmoon, M.; Shakeel, T.; Ur Rehman, S.; Ahmad, A.; Gruhn, V. A Systematic Literature Review of the IoT in Agriculture—Global Adoption, Innovations, Security, and Privacy Challenges. *IEEE Access* **2024**, *12*, 60986–61021. [\[CrossRef\]](#)
- Reddy Maddikunta, P.K.; Hakak, S.; Alazab, M.; Bhattacharya, S.; Gadekallu, T.R.; Khan, W.Z.; Pham, Q.V. Unmanned Aerial Vehicles in Smart Agriculture: Applications, Requirements, and Challenges. *IEEE Sens. J.* **2021**, *21*, 17608–17619. [\[CrossRef\]](#)
- Guan, S.; Zhu, Z.; Wang, G. A Review on UAV-Based Remote Sensing Technologies for Construction and Civil Applications. *Drones* **2022**, *6*, 117. [\[CrossRef\]](#)
- Phang, S.K.; Chiang, T.H.A.; Happonen, A.; Chang, M.M.L. From Satellite to UAV-Based Remote Sensing: A Review on Precision Agriculture. *IEEE Access* **2023**, *11*, 127057–127076. [\[CrossRef\]](#)
- Massaoudi, A.; Berguiga, A.; Harchay, A.; Ben Ayed, M.; Belmabrouk, H. Spectral and Energy Efficiency Trade-Off in UAV-Based Olive Irrigation Systems. *Appl. Sci.* **2023**, *13*, 739. [\[CrossRef\]](#)
- Massaoudi, A.; Aydi, W. Optimized Power Allocation Algorithms for UAV-Based Agricultural Systems. *IEEE Access* **2025**, *13*, 103292–103308. [\[CrossRef\]](#)
- Qu, C.; Boubin, J.; Gafurov, D.; Zhou, J.; Aloysius, N.; Nguyen, H.; Calyam, P. UAV Swarms in Smart Agriculture: Experiences and Opportunities. In Proceedings of the 2022 IEEE 18th International Conference on e-Science (e-Science), Salt Lake City, UT, USA, 11–14 October 2022; pp. 148–158. [\[CrossRef\]](#)
- Kassam, J.; Castanheira, D.; Silva, A.; Dinis, R.; Gameiro, A. Cell-Free Massive MIMO Technology and Applications in 6G. In *Massive MIMO for Future Wireless Communication Systems: Technology and Applications*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2025; pp. 95–122. [\[CrossRef\]](#)
- Elhoushy, S.; Ibrahim, M.; Hamouda, W. Cell-Free Massive MIMO: A Survey. *IEEE Comm. Surv. Tutor.* **2022**, *24*, 492–523. [\[CrossRef\]](#)
- Björnson, E.; Sanguinetti, L. Scalable Cell-Free Massive MIMO Systems. *IEEE Trans. Commun.* **2020**, *68*, 4247–4261. [\[CrossRef\]](#)
- Pan, X.; Zheng, Z.; Huang, X.; Fei, Z. On the Uplink Distributed Detection in UAV-Enabled Aerial Cell-Free mMIMO Systems. *IEEE Trans. Wirel. Comm.* **2024**, *23*, 13812–13825. [\[CrossRef\]](#)
- Björnson, E.; Sanguinetti, L.; Wymeersch, H.; Hoydis, J.; Marzetta, T.L. Massive MIMO is a reality—What is next?: Five promising research directions for antenna arrays. *Digit. Signal Process.* **2019**, *94*, 3–20. [\[CrossRef\]](#)
- Braga, I.M.; Antonioli, R.P.; Fodor, G.; Silva, Y.C.B.; Freitas, W.C. Decentralized Joint Pilot and Data Power Control Based on Deep Reinforcement Learning for the Uplink of Cell-Free Systems. *IEEE Trans. Veh. Technol.* **2023**, *72*, 957–972. [\[CrossRef\]](#)
- Zhang, Y.; Zhang, J.; Buzzi, S.; Xiao, H.; Ai, B. Unsupervised Deep Learning for Power Control of Cell-Free Massive MIMO Systems. *IEEE Trans. Veh. Technol.* **2023**, *72*, 9585–9590. [\[CrossRef\]](#)
- Zaher, M.; Demir, O.T.; Björnson, E.; Petrova, M. Learning-Based Downlink Power Allocation in Cell-Free Massive MIMO Systems. *IEEE Trans. Wirel. Commun.* **2023**, *22*, 174–188. [\[CrossRef\]](#)
- Zeng, Y.; Xu, J.; Zhang, R. Energy Minimization for Wireless Communication with Rotary-Wing UAV. *IEEE Trans. Wirel. Commun.* **2019**, *18*, 2329–2345. [\[CrossRef\]](#)
- Wu, Q.; Zeng, Y.; Zhang, R. Joint Trajectory and Communication Design for Multi-UAV Enabled Wireless Networks. *IEEE Trans. Wirel. Commun.* **2018**, *17*, 2109–2121. [\[CrossRef\]](#)

23. Luong, N.C.; Hoang, D.T.; Gong, S.; Niyato, D.; Wang, P.; Liang, Y.C.; Kim, D.I. Applications of Deep Reinforcement Learning in Communications and Networking: A Survey. *IEEE Commun. Surv. Tutor.* **2019**, *21*, 3133–3174. [\[CrossRef\]](#)
24. Shah, S.A.A.; Fernando, X.; Kashef, R. Joint Trajectory and Pilot Assignment Optimization for UAV Enabled Cell-Free Massive MIMO. In Proceedings of the 2025 IEEE International Conference on Communications Workshops (ICC Workshops), Montreal, QC, Canada, 8–12 June 2025; pp. 1876–1881. [\[CrossRef\]](#)
25. Kenneth, O.; Mwangi, E.; Konditi, D.B.O. Optimizing UAV Location for Deployment in Cell-Free Massive MIMO Networks Using a Soft Actor-Critic Reinforcement Learning. *IEEE Access* **2025**, *13*, 139680–139695. [\[CrossRef\]](#)
26. Huang, T.; Pan, H.; Sun, W.; Gao, H. Sine Resistance Network-Based Motion Planning Approach for Autonomous Electric Vehicles in Dynamic Environments. *IEEE Trans. Transp. Electr.* **2022**, *8*, 2862–2873. [\[CrossRef\]](#)
27. Zhang, S.; Zhang, H.; He, Q.; Bian, K.; Song, L. Joint Trajectory and Power Optimization for UAV Relay Networks. *IEEE Commun. Lett.* **2018**, *22*, 161–164. [\[CrossRef\]](#)
28. Hur, J.; Lee, S.H. Joint Trajectory and Power Optimization for Energy-Efficient UAV Redeployment Against an Eavesdropper Under Consistent Fairness Constraint. *IEEE Commun. Lett.* **2024**, *28*, 2347–2351. [\[CrossRef\]](#)
29. Wu, X.; Hou, C.; Meng, G.; Zhou, Z.; Liu, Q. Joint Trajectory and Power Optimization for Loosely Coupled Tasks: A Decoupled-Critic MAPPO Approach. *Drones* **2026**, *10*, 116. [\[CrossRef\]](#)
30. Shah, S.A.A.; Fernando, X.N.; Kashef, R. Joint Optimization of UAV Trajectory, Transmit Power, and User Association in Aerial-Terrestrial Cell-Free Massive MIMO Network. *IEEE Trans. Wirel. Commun.* **2026**, *25*, 15818–15832. [\[CrossRef\]](#)
31. Ma, X.; Li, J.; Nie, J.; Li, D.; Feng, W.; Jiang, W. SSDNN: A Self-Supervised DNN for Energy Efficiency Optimization in UAV CF-mMIMO Under URLLC. In Proceedings of the IEEE International Conference on Communications, Montreal, QC, Canada, 8–12 June 2025; pp. 5047–5052. [\[CrossRef\]](#)
32. Liu, Z.; Zhang, J.; Zeng, Y.; Ai, B. Energy-Efficient Multi-Agent Reinforcement Learning for UAV Trajectory Optimization in Cell-Free Massive MIMO Networks. *IEEE Trans. Wirel. Commun.* **2025**, *24*, 5917–5930. [\[CrossRef\]](#)
33. Lee, J.; Ko, H. Joint Optimization on Trajectory, Data Relay, and Wireless Power Transfer in UAV-Based Environmental Monitoring System. *Electronics* **2024**, *13*, 828. [\[CrossRef\]](#)
34. Saraiva, J.V.; Antonioli, R.P.; Fodor, G.; Freitas, W.C.; Silva, Y.C.B. Integrating Aerial and Ground Users: Deep Reinforcement Learning and Convex Optimization for Data Power Control in Cell-Free Networks. *IEEE Trans. Veh. Technol.* **2026**, *75*, 12904–12920. [\[CrossRef\]](#)
35. Bianchi, D.; Borri, A.; Di Gennaro, S.; Preziuso, M. UAV trajectory control with rule-based minimum-energy reference generation. In Proceedings of the 2022 European Control Conference (ECC), London, UK, 12–15 July 2022; pp. 1497–1502. [\[CrossRef\]](#)
36. Zhang, L.; Li, L.; Wei, W.; Song, H.; Yang, Y.; Liang, J. Scalable Constrained Policy Optimization for Safe Multi-agent Reinforcement Learning. In *Proceedings of the Advances in Neural Information Processing Systems*; Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2024; Volume 37, pp. 138698–138730. [\[CrossRef\]](#)
37. Zhang, C.; Li, Y.; Liu, Y.; Huang, Y.; Zhang, Z. TTBO-FL: Joint Training, Trajectory, and Beamforming Optimization for Energy-Efficient Federated Learning in UAV Swarm. *IEEE Internet Things J.* **2026**, *13*, 2206–2220. [\[CrossRef\]](#)
38. Mu, L.; Duan, T.; Wu, C.; Bai, B. Attention-Biased Reinforcement Learning Framework for Adaptive and Scalable Flocking of UAV Swarms. *IEEE Trans. Autom. Sci. Eng.* **2026**, *23*, 756–771. [\[CrossRef\]](#)
39. Dong, L.; Kong, H.W.; Yuan, X. Reinforcement Learning-Based Spectral Performance Optimization for UAV-Assisted MIMO Communication System. *IEEE/CAA J. Autom. Sin.* **2025**, *12*, 1283–1285. [\[CrossRef\]](#)
40. Ao, T.; Li, H.; Zhang, K.; Shi, H.; Shi, L.; Liu, F.; Zhou, Y. Heterogeneous UAVs Trajectory Optimization for Post-Disaster Target Search Based on MARL with Graph Attention Network. *IEEE Trans. Veh. Technol.* **2026**, *75*, 1412–1426. [\[CrossRef\]](#)
41. Han, S.; Zhu, K.; Zhou, M.; Liu, X. Joint Deployment Optimization and Flight Trajectory Planning for UAV Assisted IoT Data Collection: A Bilevel Optimization Approach. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 21492–21504. [\[CrossRef\]](#)
42. Al-Hourani, A.; Kandeepan, S.; Lardner, S. Optimal LAP Altitude for Maximum Coverage. *IEEE Wirel. Commun. Lett.* **2014**, *3*, 569–572. [\[CrossRef\]](#)
43. Lowe, R.; Wu, Y.; Tamar, A.; Harb, J.; Abbeel, P.; Mordatch, I. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 6382–6393.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.