Federated Dual Proximal Regularization for Energy-Efficient Decentralized UAV Navigation

Huy Dang Mac[a], Thanh Phung Truong[a], Kiet Nguyen Tuan Tran[a], Tung Son Do[a], Manh Cuong Ho[a], and Sungrae Cho[a,*]

[a]*School of Computer Science and Engineering, Chung-Ang University, Seoul 06974, South Korea*

Abstract

Federated reinforcement learning (FRL) for unmanned aerial vehicles (UAVs) in mobile edge computing (MEC) typically relies on fixed-strength proximal constraints or Kullback–Leibler (KL) divergence penalties in distribution space, both of which break down when UAVs in a fleet have different rotor disc areas, coverage radii, and maximum accelerations: fixed coefficients cannot adapt across training phases, and KL estimates become noisy when heterogeneous UAVs explore different state regions. FedDPR (Federated Dual Proximal Regularization) replaces the KL penalty with dual L2-norm proximal terms in parameter space for state-independent gradients, and anneals the global coefficient over FL rounds to transition from fleet-wide knowledge sharing to per-UAV hardware specialization. FedDPR outperforms FedProx, FMARL, and FedKL in energy efficiency, average waiting time, and cumulative reward.

Keywords: Federated reinforcement learning, UAVs-enabled MEC, Proximal regularization, Age of information, Hardware heterogeneity

Table 1: Comparison of federated RL approaches for UAV networks.

Method	Constraint	Space	Adaptive	HW Het.
FedAvg [8]	None	–	–	×
FedProx [9]	Fixed proximal	Param.	×	×
FMARL	Grad. decay	Gradient	×	×
FedKL [10]	Dual KL	Distrib.	✓	×
FedDPR	Dual proximal	Param.	✓	✓

1. Introduction

UAV-enabled mobile edge computing (MEC) dynamically deploys computational resources and exploits line-of-sight (LoS) links [1], but UAVs carry limited batteries and physically distinct hardware. Two metrics are central in this setting: energy efficiency (EE), the data processed per joule, and age of information (AoI) [2], the time each user waits before being re-served. Recent AoI minimization has been studied across heterogeneous MEC, energy-harvesting D2D, UAV-assisted WPCNs, and wireless-powered IoT settings [3, 4, 5, 6]; yet jointly optimizing EE and AoI across a hardware-heterogeneous UAV fleet [7] via parameter-space regularization remains unexplored.

Deep reinforcement learning (DRL) is preferred over the convex relaxations in [11] because the joint EE–AoI problem in (4)–(9) is sequential and non-convex under hardware heterogeneity, and a learned policy is amortized across unseen obstacle layouts. The energy and kinematics models follow the air-ground MEC tradition [12, 13, 14].

Training a separate DRL agent per UAV is sample-inefficient and ignores fleet-wide navigation knowledge (obstacle avoidance, sweeping). Federated learning (FL) [8] addresses this by aggregating local models without sharing raw trajectories, and recent federated reinforcement learning (FRL) work includes momentum-based aggregation across heterogeneous environments [15]; how the local–global coupling should evolve over training under hardware heterogeneity, however, remains open.

1.1. Related Work

FedAvg [8] drifts under heterogeneous distributions; FedProx [9] adds a fixed proximal term whose constant coefficient cannot adapt across training phases. Chen et al. [16] jointly optimized FL accuracy and communication efficiency. FMARL applies gradient-level decay but no explicit FL constraint, while FedKL [10] uses dual KL penalties in distribution space that become unstable when heterogeneous UAVs visit different state regions. As Table 1 summarizes, no existing method combines dual parameter-space regularization, adaptive constraint strength, and hardware-heterogeneity support.

1.2. Contributions and Organization

FedDPR (Federated Dual Proximal Regularization) is proposed to tackle hardware heterogeneity in FRL. The main contributions are summarized as follows:

*Corresponding author

Email addresses: hdmac@uclab.re.kr (Huy Dang Mac), tptruong@uclab.re.kr (Thanh Phung Truong), knttran@uclab.re.kr (Kiet Nguyen Tuan Tran), tsdo@uclab.re.kr (Tung Son Do), hmcuong@uclab.re.kr (Manh Cuong Ho), srcho@cau.ac.kr (and Sungrae Cho)

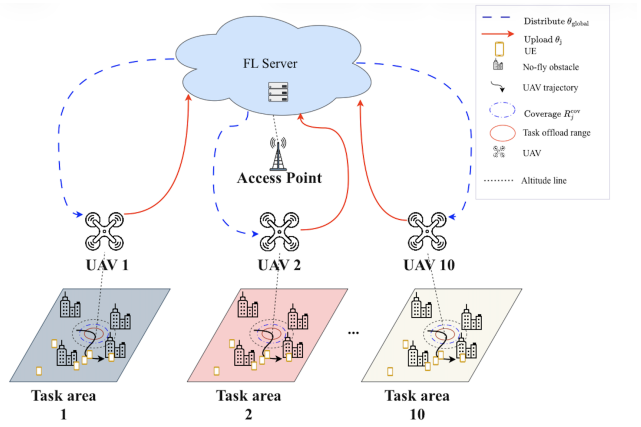


Figure 1: Decentralized UAV-MEC network architecture. Each UAV j navigates its task area, serves stationary UEs, and uploads local model updates to the FL server.

- **Dual L2-norm proximal regularization.** A local per-iteration term $(\mu_{\text{local}}/2 \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{old}}\|^2)$ ensures update stability, and a global per-round term $(\mu_{\text{global}}(r)/2 \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{global}}\|^2)$ enforces knowledge sharing. Both are defined in parameter space, yielding gradients independent of the state regions a UAV happens to visit to a stability advantage over distribution-space KL penalties.

- **Linear annealing schedule.** $\mu_{\text{global}}(r)$ decreases linearly from 5×10^{-3} to 10^{-4} over $R_{\text{anneal}} = 150$ rounds, transitioning from fleet-wide sharing in early training to hardware-specific specialization later that not addressed by prior FRL for hardware heterogeneity.

2. System Model and Problem Formulation

2.1. UAV-MEC Network Model

A decentralized UAVs-enabled MEC network is considered, consisting of $J = 10$ heterogeneous UAVs deployed over a $L \times L$ m^2 area ($L = 300$ m), as illustrated in Fig. 1. Each UAV $j \in \{1, 2, \ldots, J\}$ is assigned a dedicated task area containing $N_{\text{UE}} = 50$ stationary ground UEs that generate computational tasks. The UAVs navigate autonomously to serve UEs within their coverage area while avoiding $N_{\text{obs}} \in [10, 20]$ randomly placed no-fly zone obstacles.

2.1.1. Hardware heterogeneity

Each UAV j has physically distinct hardware parameters controlled by a heterogeneity coefficient α:

$$A_j = 0.5 + \alpha \cdot j, \tag{1}$$

$$R_j^{\text{cov}} = 50 + \tfrac{50\alpha}{3} \cdot j, \tag{2}$$

$$a_j^{\max} = 5 + \tfrac{25\alpha}{3} \cdot j, \tag{3}$$

where A_j is the rotor disc area (m^2), R_j^{cov} is the coverage radius (m), and $a_j^{\max}$ is the maximum acceleration (m/s^2). Base values follow the rotary-wing model of [12] and typical UAV-MEC settings; at $\alpha = 0.03$, UAV 10 has approximately 51% larger rotor area 9% wider coverage, and 43% higher maximum acceleration than UAV 1, spanning the dominant physical drivers of UAV-MEC behavior in flight power (via the parasite term in (6)), UE-serving capacity, and trajectory dynamics.

2.1.2. UAV kinematics

At each discrete time step t with duration $\Delta t = 1$ s, UAV j selects an acceleration action $\boldsymbol{a}_j(t) \in [-1, 1]^2$, which is scaled by its maximum acceleration:

$$\boldsymbol{v}_j(t{+}1) = \text{clip}\big(\boldsymbol{v}_j(t) + a_j^{\max} \cdot \boldsymbol{a}_j(t) \cdot \Delta t,\ -V_{\max}, V_{\max}\big)\,, \tag{4}$$

$$\boldsymbol{p}_j(t{+}1) = \text{clip}(\boldsymbol{p}_j(t) + \boldsymbol{v}_j(t{+}1) \cdot \Delta t,\ 0, L)\,, \tag{5}$$

where $V_{\max} = 20$ m/s is the maximum flight speed. If the new position falls within an obstacle, the UAV remains stationary, and its velocity is reset to zero.

2.2. Energy Consumption Model

Following [12], the rotary-wing UAV power model is adopted to capture the relationship between flight speed V and instantaneous power:

$$P(V) = \underbrace{P_b\left(1 + \frac{3V^2}{U_{\text{tip}}^2}\right)}_{\text{blade profile}} + \underbrace{P_i\sqrt{\sqrt{1 + \frac{V^4}{4v_0^4}} - \frac{V^2}{2v_0^2}}}_{\text{induced}} + \underbrace{\frac{1}{2}d_0\rho s A_j V^3}_{\text{parasite}}, \tag{6}$$

where $P_b = 79$ W, $P_i = 89$ W, $U_{\text{tip}} = 120$ m/s, $v_0 = 4.05$ m/s, $d_0 = 0.6$, $\rho = 1.225$ kg/m^3, $s = 0.05$, and A_j is from (1). The energy consumed is $E_j(t) = P(\|\boldsymbol{v}_j(t)\|) \cdot \Delta t$. Since the parasite term depends on A_j, UAVs have inherently different energy profiles which are the key motivation for FedDPR's annealing.

2.3. Task Offloading and Age of Information

At each time step, each UE generates 10 KB of new data. UAV j can offload tasks from up to $K_{\max} = 10$ UEs simultaneously within coverage radius R_j^{cov}. Among eligible UEs, those with the highest AoI are prioritized, and each served UE has 50 KB of data processed. The AoI of UE n served by UAV j evolves as:

$$\tau_n^{(j)}(t+1) = \begin{cases} 0, & \text{if UE } n \text{ is served at time } t, \\ \tau_n^{(j)}(t) + \Delta t, & \text{otherwise.} \end{cases} \tag{7}$$

The average AoI is defined over all served UEs: $\bar{\tau}(t) = \frac{1}{|\mathcal{S}(t)|}\sum_{n \in \mathcal{S}(t)} \tau_n(t)$, where $\mathcal{S}(t)$ is the set of UEs served at least once up to time t.

2.4. Problem Formulation

Energy efficiency is defined as the ratio of total processed data (in Kbits) to total energy consumed (in Joules) across all UAVs over one episode of $T = 150$ time steps:

$$\eta_{\text{EE}} = \frac{\sum_{j=1}^{J}\sum_{t=1}^{T} D_j(t) \times 8}{\sum_{j=1}^{J}\sum_{t=1}^{T} E_j(t)}\ [\text{Kbits/J}], \tag{8}$$

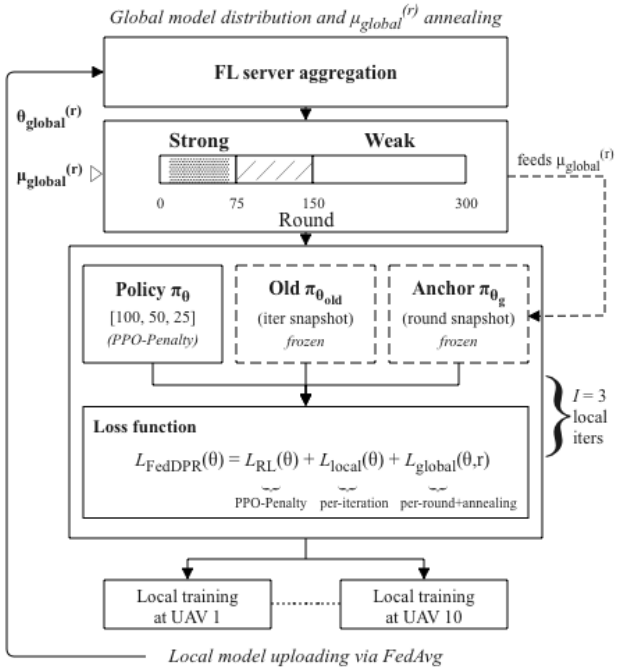


Figure 2: Overview of the proposed FedDPR framework. Each UAV client performs PPO-Penalty-based local training with dual proximal regularization. The global proximal coefficient $\mu_{\text{global}}^{(r)}$ anneals from strong (knowledge sharing) to weak (specialization) over FL rounds.

where $D_j(t)$ is the data processed by UAV j at time t (in KB). The objective is to find decentralized policies $\{\pi_j\}_{j=1}^{J}$ that jointly maximize cumulative reward:

$$\max_{\{\pi_j\}} \sum_{j=1}^{J} \mathbb{E}_{\pi_j}\left[\sum_{t=0}^{T-1} \gamma^t r_j(t)\right], \tag{9}$$

where $\gamma = 0.99$ and the per-step reward balances four objectives:

$$r_j(t) = \underbrace{10 \cdot n_j^{\text{task}}(t)}_{\text{data service}} - \underbrace{0.01 \cdot E_j(t)}_{\text{energy cost}} - \underbrace{1.0 \cdot \bar{\tau}_j(t)}_{\text{AoI penalty}} - \underbrace{100 \cdot \mathbb{I}_j^{\text{col}}(t)}_{\text{collision}}. \tag{10}$$

The four weights are scale-balanced across the per-step dynamic ranges of their terms ($n_j^{\text{task}} \in [0, 10]$, $E_j \in [125, 220]$ J, $\bar{\tau}_j \in [0, 150]$ s) so that data service, energy, and AoI contribute on comparable orders, while the rare collision indicator (weight 100) acts as a dominant safety penalty. The challenge lies in the hardware heterogeneity: UAVs with larger rotor areas consume more parasite power, while those with larger coverage radii can serve more UEs per step. A single global policy cannot be optimal for all UAVs, yet fully independent training wastes shared navigation knowledge. To address this, a novel federated learning framework is proposed in the next section.

3. Proposed Method: FedDPR

3.1. Federated Reinforcement Learning Framework

FedDPR operates in the standard FL paradigm with $R = 300$ communication rounds. In each round r: (i) the FL server distributes the global policy parameters $\boldsymbol{\theta}_{\text{global}}^{(r)}$ to all J clients; (ii) each client j performs $I = 3$ local training iterations on trajectories collected from its own environment; and (iii) the server aggregates local updates using weighted averaging:

$$\boldsymbol{\theta}_{\text{global}}^{(r+1)} = \sum_{j=1}^{J} \frac{n_j}{\sum_{k=1}^{J} n_k} \boldsymbol{\theta}_j^{(r)}, \tag{11}$$

where n_j is the number of trajectory samples collected by client j. This aggregation is identical to FedAvg [8]; the novelty of FedDPR lies in the local training objective.

Algorithm 1 FedDPR: Federated Dual Proximal Regularization

Require: J UAV clients, R FL rounds, I local iterations, annealing parameters $(\mu_{\text{init}}, \mu_{\text{final}}, R_{\text{anneal}})$
Ensure: Trained policies $\{\pi_{\boldsymbol{\theta}_j}\}_{j=1}^{J}$

```
Initialize global parameters θ_global^(0)
for round r = 0, 1, ..., R − 1 do
    Server distributes θ_global^(r) to all clients
    for each client j = 1, ..., J in parallel do
        θ_j ← θ_global^(r); θ_anchor ← θ_j
        Update μ_global(r) via Eq. (16)
        for local iteration i = 1, ..., I do
            θ_old ← θ_j; collect trajectories using π_θj; compute GAE
            Update θ_j by minimizing L(θ) in (17)
        end for
        Upload (n_j, θ_j) to server
    end for
    θ_global^(r+1) ← Σ_j (n_j / Σ_k n_k) θ_j  {FedAvg}
end for
```

3.2. PPO-Penalty-based Local Policy Optimization

Each UAV client maintains three policy-network copies sharing the same architecture like current $\pi_{\boldsymbol{\theta}}$, old $\pi_{\boldsymbol{\theta}_{\text{old}}}$ (per-iteration snapshot), and anchor $\pi_{\boldsymbol{\theta}_{\text{global}}}$ (per-round snapshot) with a 3-layer multi-layer perceptron (MLP) with hidden dimensions $[100, 50, 25]$ and tanh activations, outputting the mean of a diagonal Gaussian over 2D acceleration actions. This architecture matches that of FedKL [10] so the comparison isolates the effect of the FL constraint rather than network capacity, and the resulting 122,254-parameter model fits within onboard UAV flight-controller memory. Since AoI is part of the state, the policy is approximately Markovian and feedforward networks suffice; larger architectures and recurrent (LSTM) variants are reserved for our extended journal version.

The proximal policy optimization (PPO) with Adaptive KL Penalty [17] surrogate objective with importance sampling is:

$$L_{\text{surr}}(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{\pi_{\boldsymbol{\theta}}(a|s)}{\pi_{\boldsymbol{\theta}_{\text{old}}}(a|s)} \hat{A}(s,a)\right], \tag{12}$$

where $\hat{A}(s,a)$ is the generalized advantage estimate (GAE) [18] computed using a separate 2-layer $[64, 64]$ value network with linear output. A local KL divergence penalty maintains trust-region stability:

$$L_{\text{KL}}(\boldsymbol{\theta}) = \beta \cdot \mathbb{E}[D_{\text{KL}}(\pi_{\boldsymbol{\theta}_{\text{old}}} \,\|\, \pi_{\boldsymbol{\theta}})], \tag{13}$$

where β is adaptively adjusted: β increases when $D_{\mathrm{KL}} > 1.1 \times \mathrm{KL}_{\mathrm{targ}}$ and decreases when below $\mathrm{KL}_{\mathrm{targ}}/1.1$.

3.3. Adaptive Dual Proximal Regularization

FedDPR replaces FedKL's distribution-space penalties with dual proximal terms in parameter space; since $\|\boldsymbol{\theta}_A - \boldsymbol{\theta}_B\|^2$ is state-independent, gradients remain stable when heterogeneous UAVs visit very different state regions.

3.3.1. Local proximal term

The local proximal term penalizes deviation from the pre-iteration parameters:

$$L_{\mathrm{local}}(\boldsymbol{\theta}) = \frac{\mu_{\mathrm{local}}}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathrm{old}}\|^2, \tag{14}$$

where $\mu_{\mathrm{local}} = 10^{-3}$ is a fixed coefficient. Unlike L_{KL} that operates on output distributions, this L2 penalty directly constrains the parameter trajectory, providing smoother updates regardless of policy parameterization.

3.3.2. Global proximal term with annealing

The global proximal term constrains local parameters toward the global model:

$$L_{\mathrm{global}}(\boldsymbol{\theta}, r) = \frac{\mu_{\mathrm{global}}(r)}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathrm{global}}\|^2, \tag{15}$$

where the coefficient $\mu_{\mathrm{global}}(r)$ anneals linearly over FL rounds:

$$\mu_{\mathrm{global}}(r) = \mu_{\mathrm{init}} + \min\left(\frac{r}{R_{\mathrm{anneal}}}, 1\right) \cdot (\mu_{\mathrm{final}} - \mu_{\mathrm{init}}), \tag{16}$$

with $\mu_{\mathrm{init}} = 5 \times 10^{-3}$, $\mu_{\mathrm{final}} = 10^{-4}$, and $R_{\mathrm{anneal}} = 150$ rounds.

Early in training, a large μ_{global} helps all UAVs share general navigation skills (obstacle avoidance, UE discovery). Past round 150, hardware differences dominate and each UAV needs freedom to specialize, so μ_{global} decays to 10^{-4}. The linear schedule with $R_{\mathrm{anneal}} = 150$ is a deterministic choice prioritizing predictability and reproducibility; an adaptive variant in which μ_{global} tracks the empirical variance of local updates is developed in our extended journal version, together with a convergence-rate analysis under time-varying $\mu_{\mathrm{global}}(r)$.

3.3.3. Total training objective

The complete FedDPR loss for local training is:

$$\mathcal{L}(\boldsymbol{\theta}) = -(L_{\mathrm{surr}}(\boldsymbol{\theta}) - L_{\mathrm{KL}}(\boldsymbol{\theta})) + L_{\mathrm{local}}(\boldsymbol{\theta}) + L_{\mathrm{global}}(\boldsymbol{\theta}, r). \tag{17}$$

The complete procedure is given in Algorithm 1 and the framework is illustrated in Fig. 2.

4. Simulation Results

4.1. Simulation Setup

All algorithms are implemented in TensorFlow 2.x with TFv1 compatibility mode on an Apple M1 Pro platform. Table 2 summarizes the key simulation parameters.

FedDPR is benchmarked against three baselines that span the three principal axes of FL-constraint design: (i) **FedKL** [10] distribution-space penalty, PPO-Penalty with dual KL penalties ($\sigma = 0.5$, norm target $= 0.1$) and adaptive σ; (ii) **FedProx** [9] parameter-space penalty, PPO-Penalty with a fixed proximal term ($\mu = 10^{-3}$); (iii) **FMARL** gradient-level constraint, PPO-Clip [17] with decay-based gradient descent ($\lambda_{\mathrm{dc}} = 0.98$). All three baselines share identical FedAvg aggregation (11) the same 3-layer policy network architecture and the same training pipeline, so performance differences relative to FedDPR are attributable to the FL-constraint design rather than to other pipeline differences.

4.2. Performance Evaluation

Fig. 3 illustrates the learning curves over 300 FL rounds. FedDPR initially lags behind FedProx due to the stronger global constraint ($\mu_{\mathrm{global}} = 5 \times 10^{-3}$), but surpasses it after approximately round 100 as the annealing allows hardware-specific optimization. In Fig. 3(a), FedDPR and FedProx reach approximately 11.7 Kbits/J, while FMARL converges to 10.0 Kbits/J and FedKL remains below 7.5 Kbits/J. In Fig. 3(b), FedKL nears 60 s waiting time, while FedDPR and FedProx converge to around 30 s. In summary, FedDPR achieves the best EE and AoI, with the largest improvement over FedKL (+61% EE, −52% AoI), validating that parameter-space L2 regularization is more stable than distribution-space KL penalties for hardware-heterogeneous UAV fleets.

System performance is then evaluated under different environmental scenarios. Varying $K_{\max} \in \{4, 6, 8, 10, 12, 14\}$ (Fig. 4), FedDPR achieves comparable or higher EE than FedProx across all task loads, while FedKL's waiting time exceeds 64 s more than double FedDPR's (<29 s). Varying $L \in \{290, 300, 310, 320, 330\}$ m (Fig. 5), FedDPR maintains the highest EE up to 320 m and stays below 30 s AoI at all sizes, confirming that L2 penalties generalize better than KL penalties.

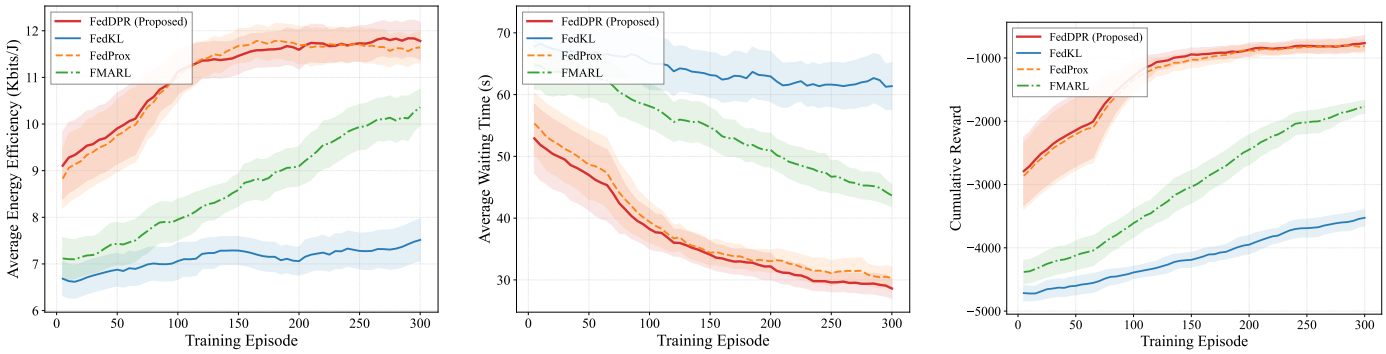
A limitation of the present setup is that ground UEs are stationary. Since FedDPR's regularization operates on policy parameters, mobile UEs change only the observation, not the regularization itself; the hardware-heterogeneity advantage over fixed-constraint baselines is expected to carry over. A systematic mobile-UE study is included in our extended journal version.

4.3. Computational Analysis

All four methods share the same MLP $[100, 50, 25]$ policy and MLP $[64, 64]$ value network, totaling 122,254 trainable

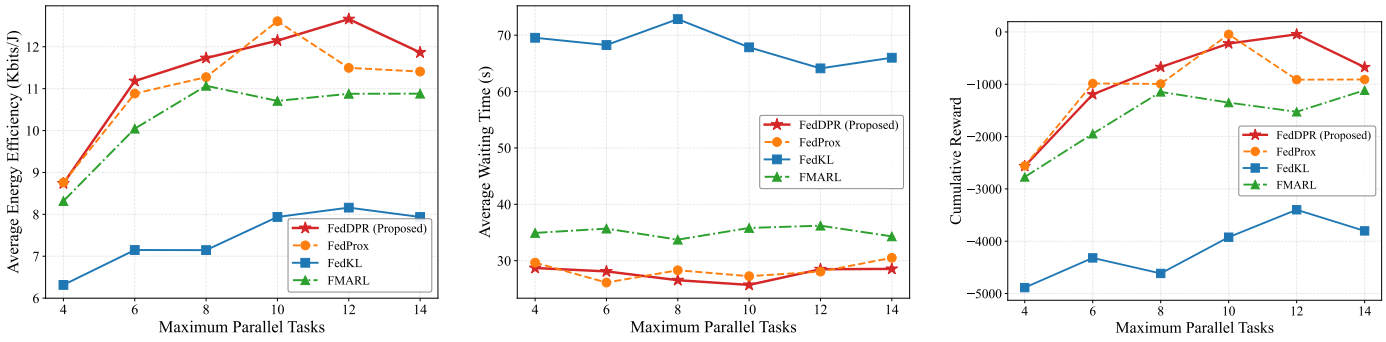
Table 2: Simulation parameters.

Environment		*Federated learning*		*RL training*		*FedDPR-specific*	
Map size $L \times L$	300×300 m^2	FL rounds R	300	Policy network	MLP [100, 50, 25], tanh	μ_{local}	10^{-3}
UAVs J	10	Clients/round	10 (full)	Value network	MLP [64, 64], linear	μ_{global} (init/final)	5×10^{-3} / 10^{-4}
UEs/area N_{UE}	50 (stationary)	Local iters I	3	Learning rate	5×10^{-4} (SGD)	Anneal rounds	150
Obstacles N_{obs}	10–20 (random)	Eval interval	Every 5 rounds	Batch / Epochs	64 / 5	KL target	3×10^{-3}
Episode length T	150 time steps			Discount γ / λ	0.99 / 0.98		
Max flight speed $V_{\max}$	20 m/s			Gradient clip norm	5.0		
Het. coefficient α	0.03						



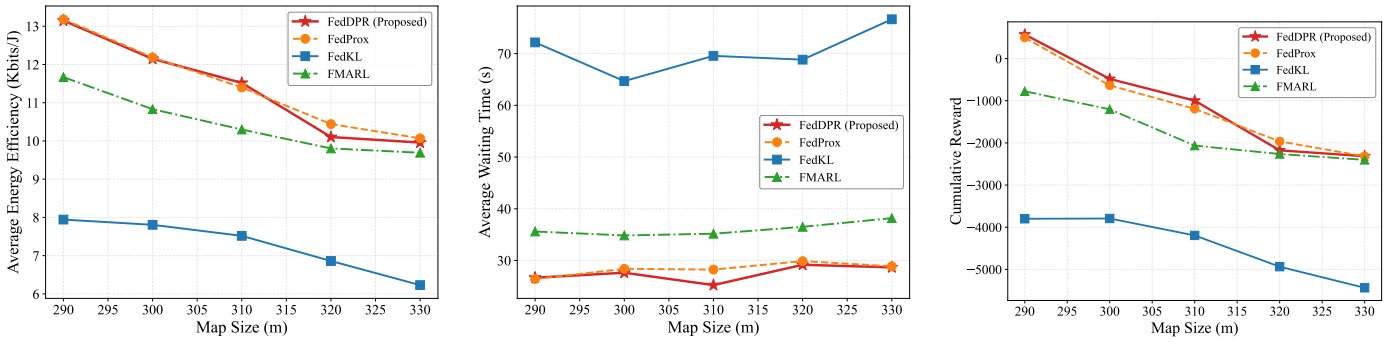
(a) Energy Efficiency (Kbits/J) (b) Average Waiting Time (s) (c) Cumulative Reward

Figure 3: Learning curves of energy efficiency, waiting time, and cumulative reward over 300 FL rounds. All hyperparameters follow Table 2. Results use the same random seed.



(a) Energy Efficiency (b) Average Waiting Time (c) Cumulative Reward

Figure 4: The average energy efficiency, waiting time, and cumulative reward under varying maximum parallel tasks $K_{\max} \in \{4, 6, 8, 10, 12, 14\}$. Each data point represents the average over the last 100 training rounds.



(a) Energy Efficiency (b) Average Waiting Time (c) Cumulative Reward

Figure 5: The average energy efficiency, waiting time, and cumulative reward under varying deployment area size $L \in \{290, 300, 310, 320, 330\}$ m. Each data point represents the average over the last 100 training rounds.

parameters ($\approx$ 489 KB at FP32). Each FL round transmits one model copy per UAV; total federated communication over 300 rounds with 10 UAVs is approximately 2.93 GB. FedDPR adds only two additive L2 terms to the local loss, with no extra trajectory sampling or network evaluation relative to FedProx with the per-iteration overhead is negligible.

5. Conclusion

This paper has investigated the hardware-heterogeneity problem in multi-UAV MEC networks and proposed FedDPR, which combines dual L2-norm proximal constraints in parameter space with a linear annealing schedule. One conclusion from this work is that the coupling strength between local and global models in heterogeneous FRL

should not remain static throughout training. Open challenges include convergence guarantees under time-varying proximal settings (addressed in our extended journal version), variance-adaptive annealing schedules, and validation under intermittent low-altitude connectivity with mobile UEs and real channel measurements.

Acknowledgment

This work was supported by the Chung-Ang University Young Scientist Scholarship in 2024.

References

[1] Y. Tang, G. Zhu, W. Xu, M. H. Cheung, T.-M. Lok, S. Cui, Integrated sensing, computation, and communication for uav-assisted federated edge learning, IEEE Transactions on Wireless Communications 24 (4) (2025) 2647–2662.

[2] S. Kaul, R. Yates, M. Gruteser, Real-time status: How often should one update?, in: 2012 Proceedings IEEE INFOCOM, IEEE, 2012, pp. 2731–2735.

[3] X. Liu, J. Zheng, K. Zheng, J. Liu, T. Taleb, N. Shiratori, Aoi minimization in heterogeneous mec networks: A federated learning-assisted hybrid drl and convex approach, IEEE Transactions on Mobile Computing (2026).

[4] X. Liu, J. Xu, K. Zheng, G. Zhang, J. Liu, N. Shiratori, Throughput maximization with an aoi constraint in energy harvesting d2d-enabled cellular networks: An msra-td3 approach, IEEE Transactions on Wireless Communications 24 (2) (2024) 1448–1466.

[5] X. Liu, H. Liu, K. Zheng, J. Liu, T. Taleb, N. Shiratori, Aoi-minimal clustering, transmission and trajectory co-design for uav-assisted wpcns, IEEE Transactions on Vehicular Technology 74 (1) (2024) 1035–1051.

[6] K. Zheng, R. Luo, X. Liu, J. Qiu, J. Liu, Distributed ddpg-based resource allocation for age of information minimization in mobile wireless-powered internet of things, IEEE Internet of Things Journal 11 (17) (2024) 29102–29115.

[7] O. A. Amodu, C. Jarray, R. A. R. Mahmood, H. Althumali, U. A. Bukar, R. Nordin, N. F. Abdullah, N. C. Luong, Deep reinforcement learning for aoi minimization in uav-aided data collection for wsn and iot applications: A survey, IEEE Access 12 (2024) 108000–108040.

[8] B. McMahan, E. Moore, D. Ramage, S. Hampson, B. A. y Arcas, Communication-efficient learning of deep networks from decentralized data, in: Artificial intelligence and statistics, Pmlr, 2017, pp. 1273–1282.

[9] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, V. Smith, Federated optimization in heterogeneous networks, Proceedings of Machine learning and systems 2 (2020) 429–450.

[10] Z. Xie, S. Song, Fedkl: Tackling data heterogeneity in federated reinforcement learning by penalizing kl divergence, IEEE Journal on Selected Areas in Communications 41 (4) (2023) 1227–1242.

[11] C. H. Liu, Z. Chen, J. Tang, J. Xu, C. Piao, Energy-efficient uav control for effective and fair communication coverage: A deep reinforcement learning approach, IEEE Journal on Selected Areas in Communications 36 (9) (2018) 2059–2070.

[12] Y. Zeng, J. Xu, R. Zhang, Energy minimization for wireless communication with rotary-wing uav, IEEE transactions on wireless communications 18 (4) (2019) 2329–2345.

[13] P. Hou, X. Jiang, Z. Wang, S. Liu, Z. Lu, Federated deep reinforcement learning-based intelligent dynamic services in uav-assisted mec, IEEE Internet of Things Journal 10 (23) (2023) 20415–20428.

[14] X. Song, M. Cheng, L. Lei, Y. Yang, Multitask and multiobjective joint resource optimization for uav-assisted air-ground integrated networks under emergency scenarios, IEEE Internet of Things Journal 10 (23) (2023) 20342–20357.

[15] H. Wang, S. He, Z. Zhang, F. M. Miao, J. Anderson, Momentum for the win: collaborative federated reinforcement learning across heterogeneous environments, in: Proceedings of the 41st International Conference on Machine Learning, 2024, pp. 50530–50560.

[16] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, S. Cui, A joint learning and communications framework for federated learning over wireless networks, IEEE transactions on wireless communications 20 (1) (2020) 269–283.

[17] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, arXiv preprint arXiv:1707.06347 (2017).

[18] J. Schulman, P. Moritz, S. Levine, M. Jordan, P. Abbeel, High-dimensional continuous control using generalized advantage estimation, arXiv preprint arXiv:1506.02438 (2015).