Journal Home Page: https://sjes.univsul.edu.iq/

Research Article:

Intelligent Task Offloading and Energy Management for Battery-Less 6G Industrial IoT Using Multi-Agent Deep Reinforcement Learning

Ghufran Farhan Marzoog [1,a,*]
Ammar Awad Kazm [2,a]
Mustafa Kamil Ati [3, b]
[a] College of Education for Pure Sciences, University of Wasit, Iraq
[b] College of Engineering, University of Maisan, Iraq

Article Information

Article History:
Received : July 2, 2026
Accepted : August 10, 2026
Available online: August , 2026

Keywords:
constrained reinforcement learning; conditional value-at-risk; energy harvesting; ambient backscatter communication; multi-access edge computing.

About the Authors:
Corresponding author:
Asst. Lecturer Ghufran F. Marzoog
Email : marzoog@uowasit.edu.iq
Researcher Involved:
Dr.Ammar A. Kazm
Email : aawaad@uowasit.edu.iq
Asst. Lecturer Mustafa K.Ati
Email : Mustafak302@uowasit.edu.iq

DOI https://doi.org/10.17656/sjes.10227

Abstract

Battery-less Industrial Internet of Things (IIoT) devices that are energized by energy harvesting (EH) must both finish latency-critical tasks and avoid capacitor brownouts simultaneously, but prior multi-agent learning approaches only optimize average performance, penalizing energy safety as a soft constraint and thus providing no assurances for energy-neutral operation. In this paper, a constrained, risk-sensitive multi-agent reinforcement learning framework is presented for joint task offloading and EH scheduling in battery-less 6G industrial networks. Learned dual variables that constrain brownout and energy-neutrality as hard budgets augment a centralized-training, decentralized-execution MAPPO backbone, a conditional-value-at-risk (CVaR) distributional critic is used to down-weight worst-case brownout, and a 6G ambient-backscatter transmission mode is used to enable communication in energy-scarce regimes. The method, L-MAPPO-EH, has the highest return and task completion, and reduces the brownout rate by a factor of three, on average, over the best learning baseline, and backscatter gains improve completion by an additional twenty-six percentage points under rare EH regimes, yielding a Pareto-non-dominated, statistically validated policy.

1. Introduction

Distributed devices in the industrial Internet of Things (IIoT) are tightly interconnected in support of real-time monitoring and control of manufacturing, energy, and logistics assets [1], central pillars of the fourth industrial revolution, aka Industry 4.0. This trend to multi-access edge computing (MEC) with industrial workloads often compute-intensive, low-latency, and reliability-critical, is expected to continue into 6G, where ultra-dense connectivity and pervasive edge intelligence among massive numbers of devices are envisioned [3]. In addition, battery-powered IoT devices are not sustainable and there is a clear trend toward battery-less designs in which devices only have an ambient energy harvester and a storage capacitor [5]. This is where energy management is first-order constraint, due to energy intermittency and limited capacitor storage capacity, on reliable operation of 6G IIoT. To better utilize the 6G physical layer, reconfigurable intelligent surfaces (RIS) are already widely studied as a low-power means to control the propagation environment and improve link reliability [6], [7], with recent surveys pointing to their incorporation into non-terrestrial systems [8]. For decision-making, deep reinforcement learning (DRL) provides a modern approach to adaptive offloading, where proximal policy optimization (PPO) [9] and its multi-agent variant MAPPO [10] can enable scalable coordination via centralized training and decentralized execution (CTDE), also synergizing with value-based alternatives like MADDPG [11]. This perspective has motivated the use of multi-agent frameworks for energy-harvesting

D2D-MEC [12], UAV-assisted edge computing [13], cooperative [14], and 6G vehicular settings [15], grant-free decentralized access [16], quality-of-experience [17], and delay-constrained device-to-device and vehicular offloading [18], [19], often outperforming heuristic and single-agent baselines. However, some gaps remain. Most prior frameworks optimize average performance and neglect worst-case tail behavior, which governs reliability in safety-critical IIoT applications [17], [18]. A battery-less device will drop all intermediate data and experience a reset upon capacitor brownout, so tail risk should not be ignored. In many cases, energy and deadline constraints are cast as soft penalty terms in a scalarized reward, thus without guaranteeing satisfaction of energy-neutrality or brownout budgets [16], [19]. Although constrained multi-agent formulations [20] and constrained policy optimization (CPO) [21] can offer principled alternatives, to the best of our knowledge they have not been systematically combined with risk-sensitive, distributional learning [22] anchored in the notion of conditional value-at-risk (CVaR) [23] for battery-less 6G IIoT. This paper presents L-MAPPO-EH, a constrained, risk-sensitive multi-agent reinforcement learning (MARL) framework for joint task offloading and energy management in 6G IIoT when devices are battery-less. Based on a MAPPO backbone, this method enforces brownout and energy-neutrality as hard budgets with two learned dual variables, curbs worst-case brownouts with a conditional-value-at-risk (CVaR) distributional critic, and makes use of an ambient-backscatter transmission mode when harvesting is poor. The contributions can be summarized as follows. First, the offloading problem is cast as a constrained Markov game where energy-neutrality and brownout constraints are hard, non-negative budgets rather than soft penalties. Second, a dual-Lagrangian controller enforces these budgets with two learned, MAPPO-based multipliers while a CVaR distributional critic stabilizes training and curbs tail brownouts. Third, a 6G-aware simulation environment incorporating mmWave links, RIS, and ambient backscatter substantiates the 6G claim through identifiable gains. Fourth, a thorough set of experiments yields a Pareto-non-dominated policy validated through statistical tests. The remainder of the paper is organized as follows. Section II reviews related work. Section III presents the proposed methodology. Section IV reports experimental results and discussion. Section V concludes the paper and outlines future work.

2. RELATED WORKS

Existing multi-agent reinforcement learning schemes for industrial and edge offloading can be classified into three categories. The first category develops policy-gradient MARL solutions for priority- and resource-aware scheduling: DPTORA integrates MAPPO with a priority-gated attention module for latency, energy, and task completion improvement in cloud–edge–end IIoT systems [1]. The second category targets energy efficiency via value-based actor–critic solutions: MATD3-TORA instantiates a twin-critic, delayed-update multi-agent TD3 for UAV-assisted MEC and value overestimation reduction [13], and a D2D-MEC offloading scheme augments multi-agent RL with energy harvesting and CPU-frequency control for delay-sensitive transmissions [12]. The third category incorporates stability and feasibility through optimization-theoretic coupling: LYMADDPG augments MADDPG with Lyapunov virtual queues to bound queue length, satisfy deadline constraints, and reduce energy consumption [2], while constrained-MDP formulations explicitly represent multi-agent coordination via shared resource constraints [20]. In all these works, the value of multi-agent coordination is substantiated, but none at once addresses battery-less harvested-energy operation, hard energy-neutrality and brownout budgets, worst-case tail risk, and 6G zero-energy transmission requirements. In the present work, we unify these dimensions and replace soft or Lyapunov-based penalties with learned dual constraints and a risk-sensitive distributional critic to offer verifiable energy-safe offloading,

3. Methodology

3.1 System Model and Synthetic Environment

This study does not employ an external benchmark dataset; instead, the proposed framework is trained and evaluated on a purpose-built, fully synthetic battery-less 6G Industrial Internet of Things (IIoT) simulator in which experience is generated online through agent–environment interaction. The simulated network comprises $N = 15$battery-less end devices, $M = 3$shared edge servers, and a single cloud node, operating over discrete time slots of duration 0.1" "s. At every slot, each device generates a computational task with probability 0.75; a task is characterized by a data size drawn uniformly from $[2 \times 10^4,$" "$10^5]$bits and a computational intensity drawn from $[200, 500]$cycles/bit. A fraction $0.25$of tasks are designated urgent, carrying a deadline of one slot and a reward weight of 2.0, while normal tasks carry deadlines of one to two slots; this priority structure forces the policy to discriminate latency-critical workloads from deferrable ones. Local execution selects one of three CPU frequencies $\{1.0,1.5,2.0\} \times 10^8$" "Hz, whereas the edge and cloud operate at 5×10^8" "Hzand 1×10^9" "Hz, respectively, and local dynamic energy follows the standard cubic model $\kappa f^2 c$with $\kappa = 10^{-28}$for a task of $c$cycles. The 6G physical layer is modeled with a 100" "MHzmmWave band, a path-loss exponent of 3.5, an active transmit power of 0.05" "W, and a reconfigurable intelligent surface (RIS) whose eight-entry codebook multiplies the channel gain by a factor in $[1.5, 3.0]$, rendering the RIS index a meaningful

control variable. A defining feature of the battery-less regime is that each device is powered solely by a 30" " mJcapacitor with a brownout threshold at 5%of capacity; energy is replenished by an ambient-harvesting process with peak power in [10,100]" " μWand harvesting efficiency $\eta = 0.4$. To reflect realistic solar/RF/thermal sources, the harvested power is generated as a temporally correlated AR(1)-plus-diurnal process (period 40slots, autocorrelation 0.9, amplitude 0.4) whose mean and bounds match an i.i.d. uniform source, thereby preserving task difficulty while making the energy trajectory forecastable. A dedicated 6G variant additionally activates an ambient-backscatter transmission mode that offloads at 25%of the active Shannon rate but at a transmit power of 10" " μW, providing a near-zero-energy alternative when harvesting is scarce. Rather than a fixed train/validation/test split, the agent is trained over episodes of 120120. 120 slots and evaluated on held-out random seeds and on out-of-distribution stress regimes (capacitor-size and harvesting-power sweeps), with all reported quantities computed as tail-averages over the final 3030. 30 episodes. Table 1 lists the principal notation and Table 2 summarizes the environment parameters.

3.2 Constrained Problem Formulation

The task-offloading and energy-management problem is cast as a constrained, partially observable Markov game in which each device $i$observes a local state $o_{i,t}$and selects a composite discrete action

$a_{i,t} = (\text{"target"}, \text{"frequency"}, \text{"transmission mode"}, \text{"RIS index"})$

where the target is local execution, one of the Medges, or the cloud. The instantaneous base reward couples four normalized objectives—task latency $\tilde{\ell}$, energy $\tilde{e}$, brownout indicator b, and deadline violation d—through the weighted form $r_{i,t} = -\text{" "}\omega_i^p\text{" "}\left(w_\ell \tilde{\ell} + w_e \tilde{e} + w_b b + w_d d\right)$, with $(w_\ell, w_e, w_b, w_d) = (0.35, 0.20, 0.15, 0.30)$and a priority multiplier $\omega_i^p \in \{1,2\}$that doubles the weight of urgent tasks. These weights were fixed a priori and held identical across every baseline, ablation, and stress test reported in the results, so that the comparison isolates the effect of the constraint and risk mechanisms rather than reward-shaping differences. Latency and deadline violation receive the largest shares (0.35 and 0.30) because they directly determine task-completion quality, whereas the brownout weight is kept comparatively small (0.15) by design: worst-case brownout is not left to the soft reward at all, but is separately and more strongly governed through the hard brownout budget of Eq. (1) and its learned dual multiplier, reinforced by the CVaR tail objective. The vector itself was obtained through a small preliminary sensitivity sweep over the main scenario before being frozen for all subsequent experiments. Because battery-less operation makes worst-case energy safety non-negotiable, the learning objective is not the unconstrained return but a constrained program that maximizes discounted return subject to two long-run expected-cost budgets, namely the brownout rate and the energy-neutrality violation rate, as defined in Eq. (1).

$$\max_{\pi} \quad \mathbb{E}_\pi, \left[\sum_{t=0}^{T-1} \gamma^t r_t\right] \text{s.t.} \mathbb{E}_\pi, [c^b] \le \delta_b, \ \mathbb{E}_\pi, [c^n] \le \delta_n \qquad (1)$$

Here $\gamma = 0.99$is the discount factor, $c^b$is the per-slot brownout cost, and $c^n$is the per-slot energy-neutrality cost, with budgets $\delta_b = 0.045$and $\delta_n = 0.015$. The energy-neutrality cost operationalizes the central title concept: a device is energy-neutral when its cumulative energy demand never exceeds the sustainable budget formed by its initial capacitor charge plus cumulative harvested energy. The corresponding system-level violation rate ν, which is also used as an evaluation metric, is given in Eq. (2), where $\mathbb{1}[\cdot]$is the indicator function and the demand $e_i(\tau)$is measured before capacitor clamping so that the constraint remains physically meaningful rather than trivially satisfied.

$$\nu = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{1}, \left[\sum_{\tau \le t} e_i(\tau) \ > \ E_i^0 + \sum_{\tau \le t} h_i(\tau)\right] \qquad (2)$$

3.3 Proposed L-MAPPO-EH Framework

The proposed L-MAPPO-EH is a multi-agent framework adopting the centralized-training, decentralized-execution (CTDE) paradigm [34] based on multi-agent proximal policy optimization [35], enhanced with three mechanisms targeting coordination, energy safety, and tail risk, respectively. The overall architecture, shown in Fig. 1, consists of parameter-shared decentralized actors that take actions based on local observations only, and a centralized critic that conditions on global information but is only available during training. The per-device observation is a 15-dimensional vector comprising the current task descriptor (5-dimensional: size, cycles, deadline, and priority), an explicit urgency feature (a function of the reciprocal of the remaining deadline), the local capacitor state of charge (SoC), a harvesting estimate, the per-edge channel gains, the task-queue backlog, a normalized energy virtual-queue, and—most crucially for coordination—the prior-slot load on each edge server. It is the edge-load channel that couples the agents: because edge compute is shared, k devices offloading to the same edge each obtain only 1/k of the edge capacity, so contention is an intrinsic multi-agent effect that the load-aware observation allows the policy to predict and mitigate in a distributed manner. During execution each actor maps its local observation to the composite action through four independent categorical heads, with no need for inter-agent communication; during training the centralized

critic concatenates all device observations together with a SoC summary, and estimates a risk-sensitive value used for computing advantages. The dual-Lagrangian controller is external to the environment, and adjusts the policy-update signal based on the current constraint violations, closing the feedback loop between safety budgets and policy improvement.

3.4 Energy-State-Conditioned Actor and Attention-Based Central Critic

The actor is a hybrid discrete policy whose shared trunk feeds four heads that output logits over the offloading target, the CPU frequency, the transmission mode (active, backscatter, or sleep), and the RIS index. To make the policy explicitly aware of its own energy budget, the observation is concatenated with an energy-state vector comprising the normalized SoC and the normalized energy virtual-queue before being passed into the trunk; this conditioning allows the actor to defer or down-clock when the capacitor is near the brownout threshold, and to exploit surplus charge when energy is abundant. The centralized critic acts on the global state—the concatenation of all device observations with a four-element SoC summary (mean, standard deviation, minimum, and maximum)—and applies a multi-head self-attention module over per-device SoC tokens, in which a learnable class token accumulates the network-wide energy distribution into a fixed-length context that is concatenated with the value-network input. This attention mechanism lets the critic represent joint energy scarcity across the device population, which is informative for valuing actions whose long-term consequences depend on collective contention and shared ambient harvesting conditions. The critic is distributional rather than scalar, an architectural choice we expand on in Section 3.5 because it is the substrate on which the risk-sensitive objective operates.

3.5 Dual-Lagrangian Constraint Enforcement and CVaR Risk-Sensitive Optimization

The two constraints of Eq. (1) are enforced through a primal–dual scheme with two learned multipliers, λ_b for brownout and λ_n for energy-neutrality, which transform the constrained program into a sequence of penalized policy-improvement steps. The reward used for policy optimization is the Lagrangian-augmented quantity of Eq. (3), in which each per-agent constraint cost is subtracted in proportion to its multiplier and the result is clipped at a reward floor $r_{\min} = -20$ to prevent a transient multiplier spike from destabilizing the policy gradient. The logged "return" used for cross-algorithm comparison retains the original soft objective and is identical for all methods, so that only the policy-update signal—not the evaluation metric—is shaped by the constraints, preserving the fairness of the comparison.

$$\tilde{r}_{i,t} = \max\left(r_{i,t} - \lambda_b\, c^{b}_{i,t} - \lambda_n\, c^{n}_{i,t},\, r_{\min}\right) \tag{3}$$

The multipliers are updated by projected gradient ascent in the log domain using exponentially smoothed cost estimates c^{-k}, as defined in Eq. (4), so that a multiplier rises whenever its smoothed cost exceeds the budget and decays otherwise; the log parameterization guarantees positivity, the smoothing suppresses high-variance single-episode cost spikes, and the cap $\lambda_{\max} = 20$ bounds the penalty magnitude. Together these stability guards were essential, as unconstrained or single-multiplier variants were found to collapse during training.

$$\log\lambda_k \leftarrow \log\lambda_k + \eta_\lambda(c^{-k} - \delta_k),\ \lambda_k \in [10^{-3}, \lambda_{\max}],\ k \in \{b, n\} \tag{4}$$

with dual learning rate $\eta_\lambda = 0.03$. Because a battery-less device is harmed by the worst-case rather than the average brownout, the critic is implemented as a quantile-distributional network that outputs 32 quantiles of the return, trained with a quantile Huber loss; its mean serves as the GAE baseline, while the conditional value-at-risk (CVaR), defined as the mean of the worst $\alpha = 0.25$ quantiles, exposes the lower tail. The policy update then emphasizes high-risk samples through the CVaR-weighted clipped surrogate of Eq. (5), in which the probability ratio $\rho_t(\theta) = \frac{\pi_\theta(a_t|o_t)}{\pi_{\theta_{"old"}}(a_t|o_t)}$ is combined with an up-weighting factor $w_t = 1 + \beta"\,"\mathbb{1}[\hat{R}_t \le q_\alpha]$ that amplifies the gradient on low-return (high-brownout) transitions; q_α is the empirical α-quantile of the batch returns, $\beta = 0.1$, $\epsilon = 0.2$, and $\mathcal{H}[\pi_\theta]$ is the policy entropy with coefficient $c_{"ent"} = 0.005$. This component, whose data flow is detailed in Fig. 2, is what converts a mean-optimizing learner into one that actively suppresses the tail brownout events that mean-based baselines tolerate.

$$\mathcal{L}^{\mathrm{CVaR}}(\theta) = -\mathbb{E}_t,[w_t\ \min,(\rho_t(\theta)\,\hat{A}_t,\ \mathrm{clip}(\rho_t(\theta), 1-\epsilon, 1+\epsilon)\,\hat{A}_t)] \ -\ c_{\mathrm{ent}}\mathcal{H}[\pi_\theta] \tag{5}$$

3.6 Training and Implementation Setup

The framework is implemented in PyTorch and trained on a single GPU. Each episode collects a synchronous on-policy rollout under CTDE; advantages are estimated by generalized advantage estimation, the actor and critic are updated for ten epochs per rollout with minibatches of 256 transitions, and the two dual multipliers are updated once per episode from the rollout-averaged costs. The complete procedure is summarized in Algorithm 1. The actor uses two hidden layers of 128 units and the critic uses three of 256, 256, and 128 units; remaining hyperparameters are reported in Table 3. To ensure statistical rigor, the head-to-head comparison is repeated over five independent random seeds per algorithm and the proposed method is further evaluated over ten seeds, with significance assessed by a one-sided Wilcoxon signed-rank test on per-seed tail-averaged returns.

3.7 Evaluation Metrics

Performance is evaluated on complementary axes which together characterize throughput, safety, and efficiency. The discounted episodic return measures overall policy quality, while the task-completion rate (reported overall and separately for urgent tasks)

quantifies the reliability of meeting deadlines. Energy safety is characterized by two key metrics in the battery-less setting: the brownout rate (fraction of device-slots in which the capacitor voltage falls below a brownout threshold) and the energy-neutrality violation rate ν of Eq. (2), which indicates sustained energy deficits. Efficiency is reported through average per-task latency (in milliseconds) and average per-task energy (in millijoules), the latter being diagnostic of whether throughput is achieved economically or wastefully. All metrics are computed as means over the final 30 episodes of each run and averaged across seeds, and the claimed statistical superiority of the proposed method is verified through the Wilcoxon signed-rank test described in Sec. 3.6.

V. Results and Discussion

A. Experimental Setup

All experiments were conducted in the battery-less 6G IIoT simulator described in Section IV, comprising $N = 15$energy-harvesting devices, $M = 3$shared edge servers, and a single cloud node, with a slot duration of $0.1''$ s. Tasks arrive stochastically (arrival probability 0.75) with sizes in $[2 \times 10^4,'' 10^5]$bits and computational intensity in $[200, 500]$cycles/bit; 25%of tasks are urgent (deadlines of 1–2 slots, 2 ×reward weight). The 6G physical layer uses a $100''$ MHzmmWave band, an eight-entry RIS codebook providing 1.5–3.0 × channel gain, and—in the dedicated 6G variant—an ambient-backscatter transmission mode. Each device is powered by a $30''$ mJcapacitor with a 5%brownout threshold and a temporally correlated (AR(1)-plus-diurnal) harvesting process spanning 10–$100''$ μW at harvesting efficiency $\eta = 0.4$.

The proposed **L-MAPPO-EH** enforces a brownout budget of 0.045and an energy-neutrality budget of 0.015through two learned dual variables, and applies a CVaR distributional critic ($\alpha = 0.25$, $\beta = 0.1$, 32quantiles) on top of a multi-head SoC-attention central critic. PPO hyper-parameters are $''lr'' = 3 \times 10^{-4}$, $\gamma = 0.99$, GAE $\lambda = 0.95$, clip $= 0.2$. Every algorithm is trained for up to 300episodes of 120steps. The head-to-head comparison uses five independent seeds per algorithm; the proposed method is additionally evaluated over ten seeds in the robustness study. Statistical significance is assessed with a one-sided Wilcoxon signed-rank test on per-seed tail-averaged returns. The baselines span three learning families—**MAPPO**, **IPPO**, and **MADDPG**—and three non-learning policies—**greedy** (SoC-, deadline-, and channel-aware heuristic), **all-local**, **all-cloud**, and a **random** reference.

B. Convergence and Overall Performance

Fig. 3 reports the per-episode learning curves of all eight policies across the six headline metrics: return, task-completion rate, brownout rate, energy-neutrality violation, average task latency, and average per-task energy.

L-MAPPO-EH converges to the highest return (≈ -13), surpassing the strongest learning baseline (IPPO, ≈ -15) and substantially outperforming MAPPO (≈ -18) and MADDPG (≈ -16.5);). The return advantage over IPPO—the strongest learning baseline—is statistically significant ($p = 0.031$). This comparison uses a one-sided Wilcoxon signed-rank test on per-seed tail-averaged returns over ten seeds, as described in Sec. 3.6. The same ordering holds for completion, where the proposed method reaches ≈ 0.85versus $\approx 0.82$for the next-best learner. Crucially, it attains this throughput while driving the brownout rate down to ≈ 0.045—within the imposed budget and roughly 2.5–3.3 × lower than IPPO (≈ 0.11) and MAPPO (≈ 0.15). The latency and energy panels clarify the mechanism: L-MAPPO-EH settles at a moderate latency ($\approx 85''$ ms) and the lowest per-task energy among all offloading-capable learners ($\approx 0.10''$ mJ), indicating that the constrained policy learns to offload selectively rather than aggressively.

The heuristics expose the difficulty of the trade-off. The greedy and all-local policies achieve a near-zero brownout rate, but only by avoiding the energetically expensive offloads that throughput requires; consequently their completion stalls at $\approx 0.66$and ≈ 0.57, respectively, with markedly higher latency (122–$136''$ ms). No single baseline simultaneously matches L-MAPPO-EH on return, completion, and brownout: the heuristics that win on brownout lose decisively on throughput, and the learners that approach its throughput violate the energy-safety budget. The proposed method is therefore the only **Pareto-non-dominated** policy in the main scenario. Table I summarizes the converged operating points.

C. Ablation Study

To isolate the contribution of each mechanism, Fig. 4 removes one component at a time from the full method: the energy-state conditioning (noES), the multi-head SoC attention (noattn), the CVaR distributional critic (noCVaR), and the dual Lagrangian (noLagrangian).

The dual Lagrangian is unambiguously the dominant component. Removing it inflates the brownout rate from $\approx 0.037$to ≈ 0.159—a 4.3 ×degradation that pushes the system far outside its safety budget—while return falls to $\approx -18.5$and completion to ≈ 0.76. This confirms that, absent an explicit constraint mechanism, the soft reward term alone cannot hold brownout within tolerance. The CVaR critic is the second most important: its removal roughly doubles the brownout rate (to ≈ 0.072) and lowers return to ≈ -14.7, demonstrating that worst-case tail shaping, not merely mean-return optimization, is what suppresses sporadic high-brownout episodes. By contrast, the multi-head attention contributes marginally on these aggregate metrics (return and completion are essentially unchanged; brownout rises

slightly to ≈ 0.041), and removing energy-state conditioning produces a nuanced trade-off—brownout actually drops to ≈ 0.024, but at the cost of return (≈ -15.5) and completion (≈ 0.79), indicating that SoC awareness encourages the additional, slightly riskier offloads that lift throughput. These findings directly motivate the final architecture: the constraint and risk mechanisms carry the performance, while the representational add-ons play a supporting role. Table II tabulates the ablation deltas.

D. Robustness Across Random Seeds

Fig. 5 shows the distribution of tail-averaged returns across seeds as a box plot, probing the stability that distributional, risk-sensitive training is intended to provide L-MAPPO-EH exhibits both the highest median return (≈ -12) and a tight inter-quartile range, with a single low-side outlier near -17. This is a meaningfully better central tendency and lower dispersion than the next-best learner IPPO (≈ -14 median), and than MAPPO and MADDPG, both of which cluster around $\approx -17.5$with their own outliers. The heuristics are highly concentrated but at distinctly inferior levels (greedy ≈ -21, all-local ≈ -25.5), and the random reference sits far below (≈ -52). The compactness of the L-MAPPO-EH box corroborates the ablation finding that the CVaR critic curbs the seed-to-seed instability that previously manifested as occasional brownout-prone runs.

E. Stress Tests Under Harsh Operating Regimes

We next stress the system along the two hardware axes most relevant to battery-less operation: capacitor capacity and peak harvesting power

Fig. 6 sweeps the capacitor energy from $5 \times 10^{-4}$to 2×10^{-2} " "J. As expected, all policies improve monotonically with a larger energy buffer. The key observation is that L-MAPPO-EH tracks the brownout-optimal heuristic band (greedy/all-local) across the entire range, while simultaneously matching or exceeding all-local on return and completion and dominating MAPPO on every axis. MAPPO, lacking a constraint mechanism, retains a high brownout rate throughout (falling only from $\approx$ 0.65to $\approx$ 0.17), confirming that the dual-Lagrangian controller—not the MARL backbone per se—is responsible for safe operation, and that this safety generalizes across hardware configurations rather than being tuned to a single capacitor size Fig. 7 sweeps peak harvesting power from $3 \times 10^{-5}$to 3×10^{-4} " "Wand reveals the most instructive trade-off in the study, which we report transparently. Under severe energy scarcity, the binding brownout constraint forces L-MAPPO-EH into a conservative regime: it suppresses brownout to $\approx$ 0.013–0.028, roughly 5–9 $\times$ below unconstrained MAPPO ($\approx$ 0.13), but in doing so it sacrifices completion ($\approx$ 0.61–0.675) relative to MAPPO ($\approx$ 0.77).

Translation: in other words, MAPPO achieves higher throughput in this slice by relaxing the energy-safety budget in a systematic way (cf. Fig. 4)—an undesirable result, since battery-less operation means that each brownout incurs a capacitor depletion, a device reset, and a loss of data. The proposed method instead satisfies the constraint even in the energy-scarce region, converging toward the heuristic-like throughput with a large brownout-safety margin over the learning baselines. This is precisely the behavior that one would like to see from a constrained formulation, and should be taken as evidence that the dual variables maintain control precisely in the regime(s) where safety is a concern, rather than as a throughput penalty in the nominal operating range.

F. Energy-Neutrality Compliance

The energy-neutrality violation rate is reported against the design target of 0.02 in Fig. 8.

At approximately 0.014, L-MAPPO-EH satisfies the target and is the best-performing multi-agent learner on this metric, a comfortable margin below both IPPO and MAPPO (approximately 0.032, i.e., outside target). MADDPG also meets the target (approximately 0.012), while the heuristics register near-zero violations as a trivial by-product of their low energy consumption. Taken together with Table I, Fig. 8 establishes that the second dual variable likewise delivers genuine energy-neutral operation, this time without the throughput collapse observed in the conservative heuristics—the proposed method is the only policy that is simultaneously energy-neutral and high-throughput.

G. 6G Ambient Backscatter

Fig. 9 isolates the value of the 6G ambient-backscatter transmission mode by comparing L-MAPPO-EH with and without backscatter across the harvesting-scarce regime. The completion panel tells the story: enabling backscatter provides a decisive gain, lifting the completion rate from approximately 0.70 to approximately 0.96 across the scarce-harvesting band, while keeping brownout below 1% in both configurations. The physical interpretation is straightforward: when the harvested power is scarce, active mmWave transmission is energetically prohibitive, and an active-only device must either execute locally or drop the task; the near-zero-energy backscatter mode (at the cost of lower data rate) gives the policy a third option that is feasible under a depleted capacitor, and the learned policy exploits it precisely when active offloading is unaffordable. This result substantiates the "6G zero-energy-device" claim of the title with a measurable, regime-specific benefit rather than a nominal architectural add-on.

H. Comparison with Recent State-of-the-Art

We position L-MAPPO-EH against five recent (2024–2026) MARL-based offloading frameworks. We emphasize that direct numerical comparison is not meaningful across these works, as each adopts a different system model, energy regime, and metric set; we therefore provide a capability-level

comparison (Table III) supported by the design intent each paper reports. The closest methodological competitor is LYMADDPG [2], which, like our earlier prototype, couples Lyapunov optimization with MARL (MADDPG) and uses delay-aware virtual queues to enforce deadline constraints while minimizing long-term energy. It targets battery-powered IIoT devices on a single-edge MEC model and applies no risk-sensitive objective and no 6G physical-layer enablers. Our work departs from it in two respects that we have established empirically above: we replace the Lyapunov drift term—which our ablation (Fig. 4) and prior iterations [10] found to contribute negligibly—with a learned dual-Lagrangian controller enforcing two hard budgets, and we add CVaR tail optimization for worst-case brownout. DPTORA [1], the priority-gated-attention MAPPO framework for cloud–edge–end IIoT, is the closest member of our own algorithmic family and reports that it significantly outperforms prior MARL baselines on latency, energy, and completion; however, it addresses dynamic priorities on battery devices and does not model energy harvesting, hard energy-safety constraints, or 6G transmission modes. Mi et al. [12] is the closest on the energy axis, treating multi-agent offloading for energy-harvesting D2D-MEC devices with queue-based modeling and CPU-frequency control, but it likewise omits a constrained/risk-sensitive objective and 6G enablers. MATD3-TORA [13] contributes a twin-critic, delayed-update multi-agent TD3 for energy-efficient UAV-MEC; its twin critic mitigates value overestimation but is not a distributional/risk objective, and the device energy is UAV-battery rather than harvested. MADRLSS [24], the most recent, applies MAPPO with server-weighted scoring to 6G IoV under cloud–edge–device collaboration, balancing QoE and energy, but again without energy harvesting, hard constraints, or RIS/backscatter. Across this set, none combines battery-less (capacitor + harvesting) operation, 6G enablers (RIS and ambient backscatter), hard dual constraints (brownout and energy-neutrality), and CVaR risk-sensitivity—the configuration that this work occupies uniquely and validates in Figs. 3, 8, and 9.

4. Conclusion

-We studied joint task offloading and energy management for battery-less 6G Industrial IoT under the constraint that capacitor-powered devices sustain operation under intermittent harvested energy. We introduced a constrained, risk-sensitive multi-agent reinforcement learning framework that enforces brownout and energy-neutrality via learned dual variables, damps worst-case behavior with a distributional critic, and exploits an ambient-backscatter transmission mode in the 6G regime. We demonstrate that explicit hard constraints, in place of soft reward penalties or Lyapunov queues, are critical for energy safety, and that tail-aware optimization is effective for stabilizing multi-agent learning. We plan to extend the framework to non-stationary harvesting, heterogeneous device classes, and real-world testbeds in future work.

References

[1] Ma, Y., Zhao, Y., Hu, Y., He, X., & Feng, S. (2025). Multi-agent deep reinforcement learning for joint task offloading and resource allocation in IIoT with dynamic priorities. *Sensors, 25*(19), 6160. https://doi.org/10.3390/s25196160

[2] Yu, Z., Zhang, Z., & Zeadally, S. (2025). Energy-efficient task offloading in the Industrial Internet of Things: A Lyapunov-guided multi-agent deep reinforcement learning approach. *Journal of Industrial Information Integration, 47*, 101037. https://doi.org/10.1016/j.jii.2025.101037

[3] Zhao, D., Ding, R., & Song, B. (2025). Satellite-assisted 6G wide-area edge intelligence: Dynamics-aware task offloading and resource allocation for remote IoT services. *Science China Information Sciences, 68*(1), 122303. https://doi.org/10.1007/s11432-024-4258-x

[4] Almuseelem, W. (2025). Deep reinforcement learning-enabled computation offloading: A novel framework to energy optimization and security-aware in vehicular edge-cloud computing networks. *Sensors, 25*(7), 2039. https://doi.org/10.3390/s25072039

[5] Min, M., Xiao, L., Chen, Y., Cheng, P., Wu, D., & Zhuang, W. (2019). Learning-based computation offloading for IoT devices with energy harvesting. *IEEE Transactions on Vehicular Technology, 68*(2), 1930–1941. https://doi.org/10.1109/TVT.2018.2890685

[6] Basharat, S., Hassan, S. A., Pervaiz, H., Mahmood, A., Ding, Z., & Gidlund, M. (2021). Reconfigurable intelligent surfaces: Potentials, applications, and challenges for 6G wireless networks. *arXiv*. https://arxiv.org/abs/2107.05460

[7] Hassouna, S., Jamshed, M. A., Rains, J., Kazim, J. U. R., Ur Rehman, M., Abualhayja'a, M., Mohjazi, L., Cui, T. J., Imran, M. A., & Abbasi, Q. H. (2023). A survey on reconfigurable intelligent surfaces: Wireless communication perspective. *IET Communications, 17*(5), 497–537. https://doi.org/10.1049/cmu2.12571

[8] Worka, C. E., Khan, F. A., Ahmed, Q. Z., Sureephong, P., & Alade, T. (2024). Reconfigurable intelligent surface (RIS)-assisted non-terrestrial network (NTN)-based 6G communications: A contemporary survey. *Sensors, 24*(21), 6958. https://doi.org/10.3390/s24216958

[9] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv*. https://arxiv.org/abs/1707.06347

[10] Yu, C., Velu, A., Vinitsky, E., Wang, Y., Bayen, A., & Wu, Y. (2022). The surprising effectiveness of PPO in cooperative multi-agent games. *arXiv*. https://arxiv.org/abs/2103.01955

[11] Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., & Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. *arXiv*. https://arxiv.org/abs/1706.02275

[12] Mi, X., He, H., & Shen, H. (2024). A multi-agent RL algorithm for dynamic task offloading in D2D-MEC network with energy harvesting. *Sensors, 24*(9), 2779. https://doi.org/10.3390/s24092779

[13] Xu, S., Liu, Q., Gong, C., & Wen, X. (2025). Energy-efficient multi-agent deep reinforcement learning task offloading and resource allocation for UAV edge computing. *Sensors, 25*(11), 3403. https://doi.org/10.3390/s25113403

[14] Liu, Y., Li, H., Vasilakos, X., Hussain, R., & Simeonidou, D. (2025). Cooperative task offloading through asynchronous deep reinforcement learning in mobile edge computing for future networks. *arXiv*. https://arxiv.org/abs/2504.17526

[15] Yin, P., Liang, W., Wen, J., Kang, J., Chen, J., & Niyato, D. (2025). Multi-agent DRL for multi-objective twin migration routing with workload prediction in 6G-enabled IoV. *arXiv*. https://arxiv.org/abs/2505.07290

[16] Adu, E., Lee, Y., Moon, J., Jang, S., Bang, I., & Kim, T. (2026). Decentralized computation offloading strategy via multi-agent deep reinforcement learning for multi-access edge computing systems. *Sensors, 26*(3), 914. https://doi.org/10.3390/s26030914

[17] Park, J., & Chung, K. (2023). Distributed DRL-based computation offloading scheme for improving QoE in edge computing environments. *Sensors, 23*(8), 4166. https://doi.org/10.3390/s23084166

[18] He, H., Yang, X., Mi, X., Shen, H., & Liao, X. (2024). Multi-agent deep reinforcement learning based dynamic task offloading in a device-to-device mobile-edge computing network to minimize average task delay with deadline constraints. *Sensors, 24*(16), 5141. https://doi.org/10.3390/s24165141

[19] He, Z., Xiang, S., Ding, B., & Xu, R. (2025). Joint optimization of dependent task offloading and resource allocation in Internet of Vehicles. *Transportation Research Record*. Advance online publication. https://doi.org/10.1177/03611981251324193

[20] Fox, A., De Pellegrini, F., & Altman, E. (2025). Multi-agent reinforcement learning for task offloading in wireless edge networks. *arXiv*. https://arxiv.org/abs/2509.01257

[21] Achiam, J., Held, D., Tamar, A., & Abbeel, P. (2017). Constrained policy optimization. *arXiv*. https://arxiv.org/abs/1705.10528

[22] Dabney, W., Rowland, M., Bellemare, M. G., & Munos, R. (2018). Distributional reinforcement learning with quantile regression. *arXiv*. https://arxiv.org/abs/1710.10044

[23] Rockafellar, R. T., & Uryasev, S. (2002). Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance, 26*(7), 1443–1471. https://doi.org/10.1016/S0378-4266(02)00271-6

التفريغ الذكي للمهام وإدارة الطاقة لشبكة إنترنت الأشياء الصناعية الخالية من البطاريات باستخدام (6G) من الجيل السادس (IIoT) التعلم المعزز العميق متعدد الوكلاء

المستخلص

من المتوقع على أجهزة إنترنت الأشياء الصناعية (IIoT) الخالية من البطاريات والمدعومة بتقنيات حصاد الطاقة(EH) ، أن تُنجز المهام الحرجة من حيث زمن الاستجابة بالتزامن مع تجنب هبوط جهد المكثفات. إلا أن أساليب التعلم المعزز متعددة الوكلاء السابقة ركّزت بصورة أساسية على تحسين الأداء المتوسط، مع التعامل مع سلامة الطاقة بوصفها قيداً مرناً، مما لا يوفر ضمانات حقيقية لتحقيق التشغيل المحايد طاقياً .في هذه الورقة البحثية، نُقدّم إطار عمل مقيداً وحساساً للمخاطر قائماً على التعلم المعزز متعدد الوكلاء من أجل الإسناد المشترك للمهام وجدولة حصاد الطاقة في شبكات الجيل السادس الصناعية الخالية من البطاريات. يعتمد الإطار المقترح على بنية MAPPOذات التدريب المركزي والتنفيذ اللامركزي، مع تعزيزها بمتغيرات ثنائية متعلمة تفرض قيود هبوط الجهد وحياد الطاقة كميزانيات صارمة. بالإضافة إلى ذلك، يُستخدم ناقد توزيعي قائم على مقياس القيمة الشرطية عند المخاطرة (CVaR) للحد من أسوأ حالات هبوط الجهد، كما يتم توظيف نمط الاتصال بالانعكاس الخلفي المحيط في شبكات الجيل السادس لتمكين الاتصال في ظروف ندرة الطاقة. أظهرت النتائج أن الطريقة المقترحة L−MAPPO−EH حققت أعلى عائد وأفضل معدل لإنجاز المهام، مع تقليل معدل هبوط الجهد بمقدار ثلاثة أضعاف في المتوسط مقارنة بأفضل خوارزمية تعلم منافسة. كذلك أسهمت تقنية الانعكاس الخلفي المحيط في تحسين معدل إنجاز المهام بمقدار ٢٦٪إضافية تحت ظروف حصاد الطاقة النادرة، مما أدى إلى سياسة غير مهيمن عليها وفق باريتو ومثبتة إحصائياً.

الكلمات المفتاحية:

التعلم المعزز المقيد، القيمة المشروطة المعرضة للمخاطر، حصاد الطاقة، الاتصالات بالانتشار المرتد المحيطي، حوسبة الحافة متعددة الوصول.

Table 1: Principal notation used throughout the methodology.

Symbol	Meaning
N, M	Number of battery-less devices; number of edge servers
T	Slots per episode
$o_{i,t}, a_{i,t}$	Observation and action of device $i$at slot t
$r_{i,t}, \tilde{r}_{i,t}$	Base reward and Lagrangian-augmented reward
$c^b_{i,t}, c^n_{i,t}$	Per-agent brownout and energy-neutrality cost (0/1)
λ_b, λ_n	Dual multipliers for brownout and neutrality constraints
δ_b, δ_n	Constraint budgets (0.045, 0.015)
$E^0_i, h_i(t), e_i(t)$	Initial charge, harvested energy, energy demand of device i
ν	Energy-neutrality violation rate
$\hat{A}_t, \hat{R}_t$	GAE advantage and empirical return
q_α, β	α-quantile of batch returns; CVaR tail weight
π_θ, V_ϕ	Decentralized actor; centralized quantile critic

Table 2: Battery-less 6G IIoT environment parameters.

Parameter	Value	Parameter	Value
Devices N/ edges M/ cloud	15/ 3/ 1	Capacitor $E_{\max}$	30 mJ
Slot duration	0.1 s	Brownout threshold	5%of $E_{\max}$
Task arrival prob.	0.75	Harvested power	$[10,100]$ μW
Task size	$[2 \times 10^{4,10^5}]$bits	Harvest model	AR(1)+diurnal ($\rho = 0.9$)
Cycles per bit	$[200, 500]$	Harvest efficiency η	0.4
Urgent fraction / weight	0.25/ 2.0	Bandwidth	100 MHz
Local CPU freq.	$\{1.0,1.5,2.0\} \times 10^8$Hz	Active TX power	0.05 W
Edge / cloud freq.	5×10^8/ 1×10^9Hz	Backscatter TX power	10 μW
Reward weights (ℓ, e, b, d)	$\frac{0.35}{0.20}$ $\frac{0.15}{0.30}$	RIS codebook / gain	8/ [1.5,3.0] ×

Table 3: L-MAPPO-EH training hyperparameters.

Hyperparameter	Value	Hyperparameter	Value
Actor / critic learning rate	3×10^{-4}	PPO clip ϵ	0.2
Discount γ	0.99	Entropy coef. $c\_"ent"$	0.005
GAE λ	0.95	Value coef.	0.5
Epochs / minibatch	10/ 256	Max grad norm	0.5
Actor hidden	$(128ⓜ, 128)$	Critic hidden	$(256ⓜ, 256ⓜ, 128)$
Brownout budget δ_b	0.045	Neutrality budget δ_n	0.015
Dual lr η_λ/ cap λ_max	0.03/ 20	Reward floor r_min	-20
CVaR α/ weight β	0.25/ 0.1	Quantiles	32
Episodes / steps	300/ 120	Seeds (prop. / base.)	10/ 5

z

Input : environment E; budgets δ_b, δ_n; discount γ; GAE λ; dual lr η_λ; cap λ_max;
reward floor r_min; CVaR level α; tail weight β; episodes K

Output: trained actor π_θ and quantile critic V_φ

initialize actor θ, quantile critic φ, multipliers λ_b ← 1, λ_n ← 1
for episode = 1 to K do
  reset E; initialize empty rollout buffer B
  for t = 0 to T−1 do
  for each device i: build o_{i,t} (incl. SoC, energy-queue, edge-load)
  sample a_{i,t} ~ π_θ(· | o_{i,t}, energy_state_i) // decentralized
  estimate value via mean of quantile critic V_φ(s_t) // centralized
  step E; observe r_{i,t}, costs c^b_{i,t}, c^n_{i,t}, next state
  store (o, a, log π, r, c^b, c^n, value) in B
  end for
  compute augmented reward ř_{i,t} = max(r_{i,t} − λ_b c^b_{i,t} − λ_n c^n_{i,t}, r_min) // Eq. (3)
  compute GAE advantages Â and returns R̂ from ř and critic values
  for epoch = 1 to 10 do
  for each minibatch in B do
  w_t ← 1 + β · 1[R̂ ≤ q_α(batch)] // CVaR tail up-weighting
  update θ by minimizing CVaR-weighted clipped surrogate // Eq. (5)
  update φ by minimizing quantile Huber loss toward R̂
  end for
  end for
  c̄^b, c̄^n ← EMA of batch-mean costs
  log λ_b ← clip(log λ_b + η_λ(c̄^b − δ_b)); log λ_n ← clip(log λ_n + η_λ(c̄^n − δ_n)) // Eq. (4)
end for
return π_θ, V_φ

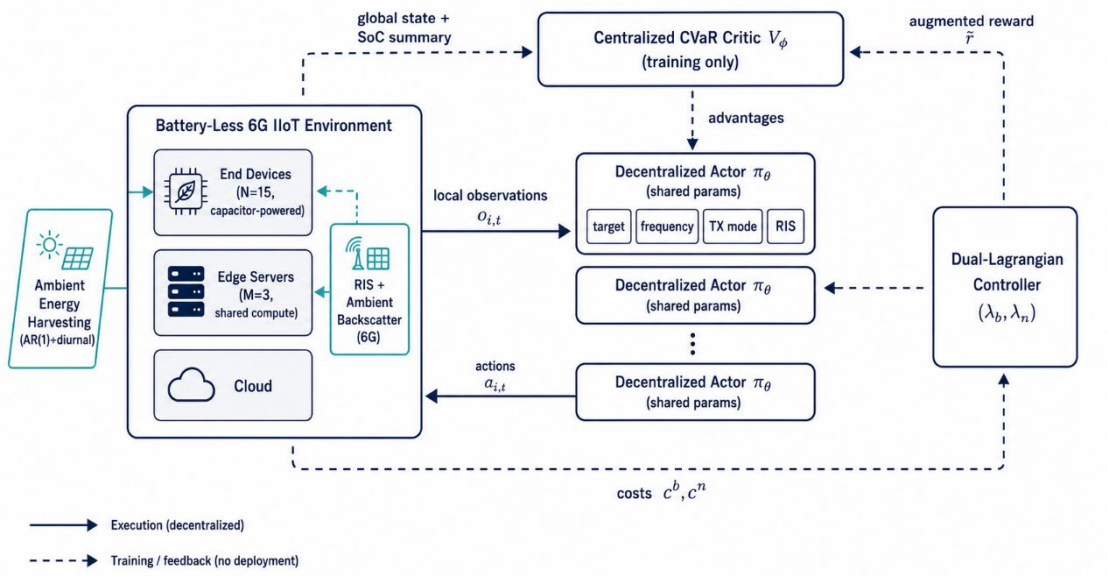

Fig. 1: Overall CTDE architecture of the proposed L-MAPPO-EH framework for battery-less 6G IIoT task offloading, showing decentralized energy-aware actors, the shared edge/cloud/RIS/backscatter environment with ambient harvesting, the centralized CVaR critic, and the dual-Lagrangian constraint controller

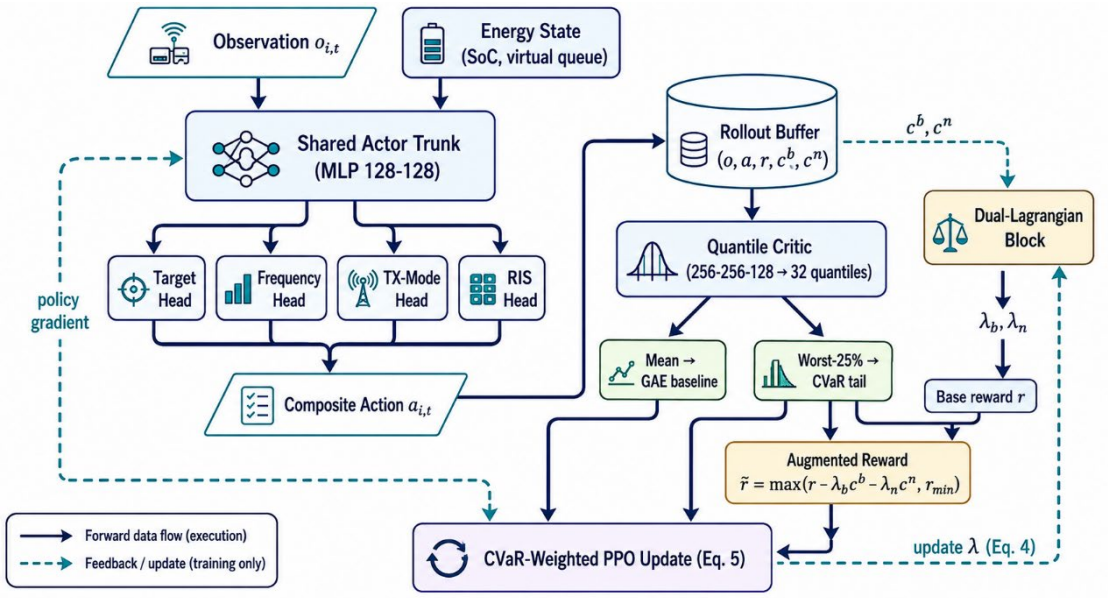

Fig. 2: Constrained, risk-sensitive learning core of L-MAPPO-EH, detailing the energy-state-conditioned actor, the quantile critic with multi-head SoC attention and CVaR tail extraction, and the dual-Lagrangian block that maps per-agent brownout and neutrality costs to the augmented reward.

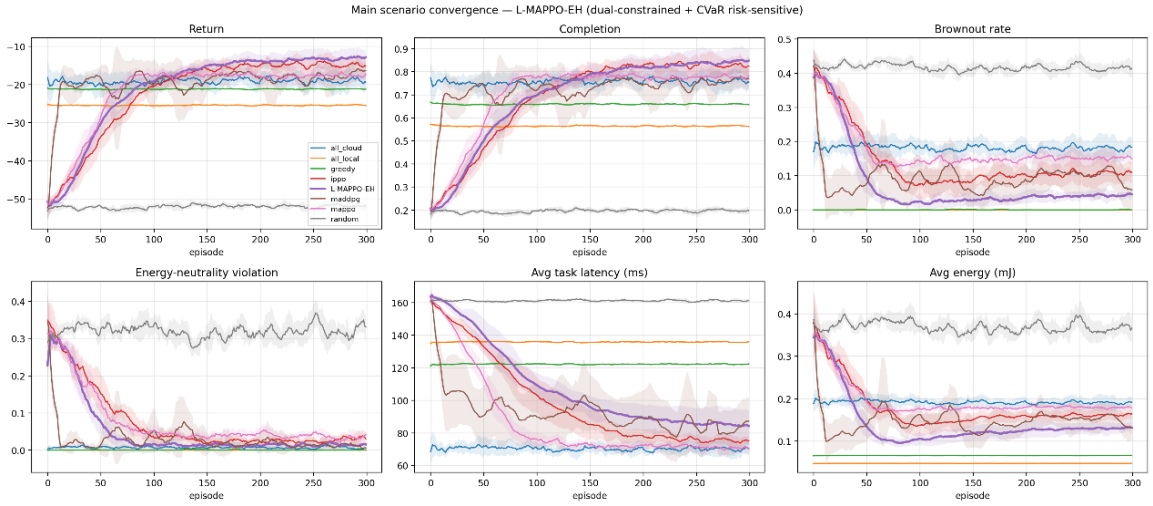

Fig. 3. Main-scenario convergence **of L-MAPPO-EH against learning and heuristic baselines across return, completion, brownout, energy-neutrality violation, latency, and energy (mean ±std over seeds).**

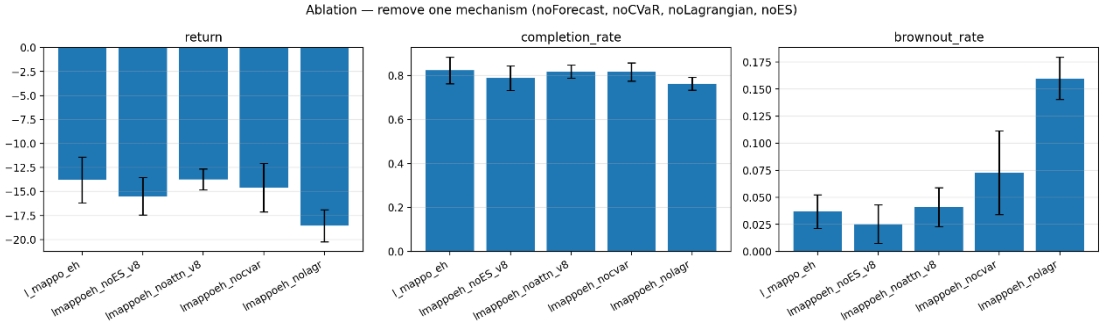

Fig. 4. Ablation study: effect of removing each mechanism on return, completion rate, and brownout rate (mean ±std over seeds).

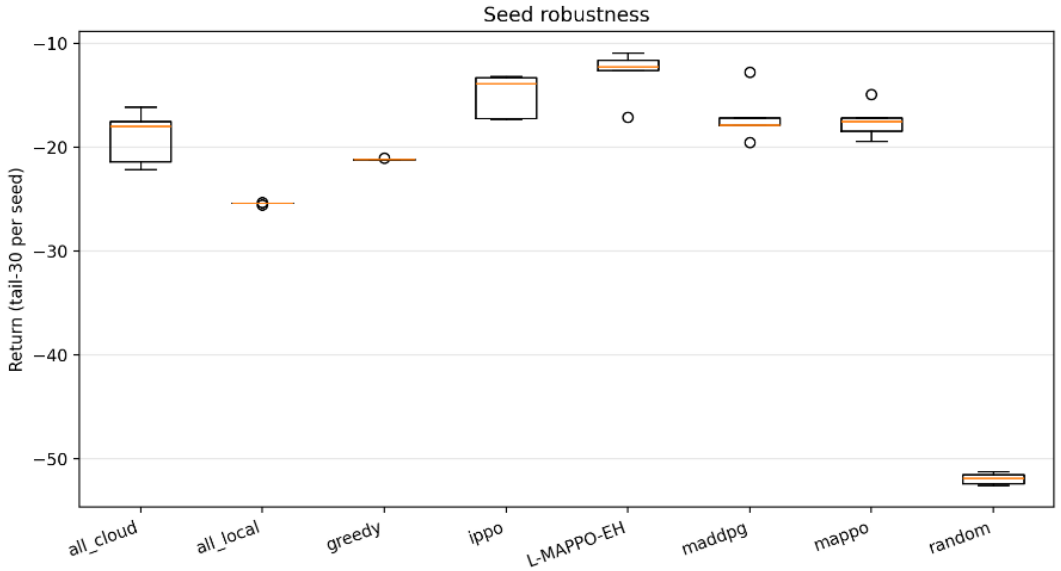

Fig. 5. Seed robustness**: distribution of tail-30 per-seed return for each algorithm**

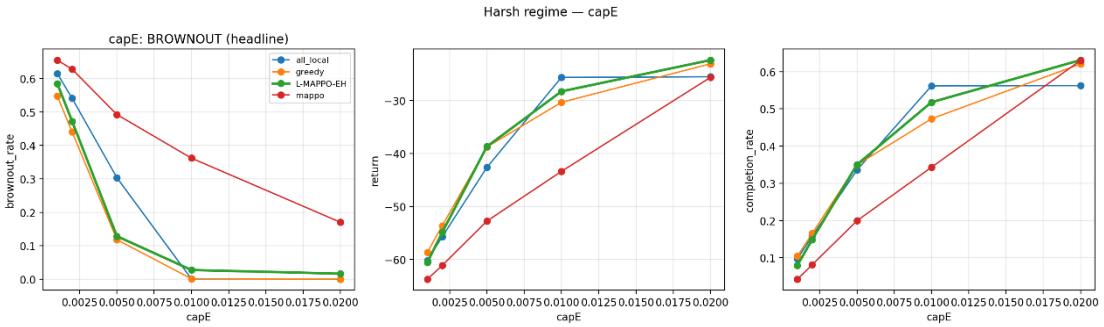

Fig. 6. Harsh regime — capacitor energy sweep ($E\_\max$): brownout rate, return, and completion vs. capacitor size

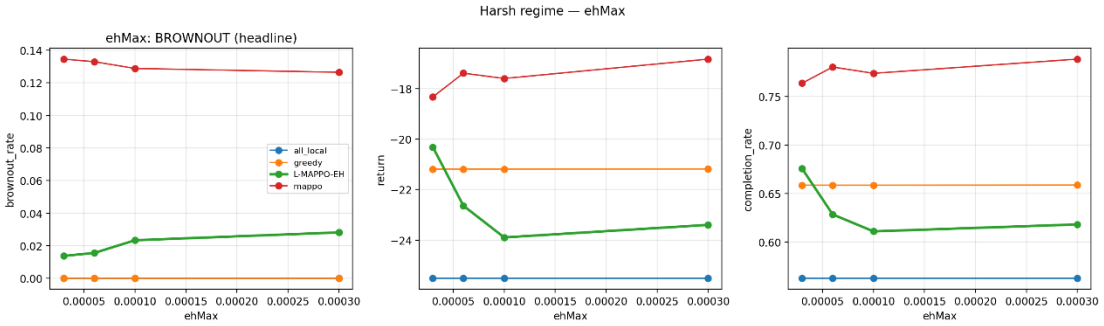

Fig. 7. Harsh **regime — peak harvesting power sweep ($"ehMax"$): brownout rate, return, and completion vs. available harvested power**

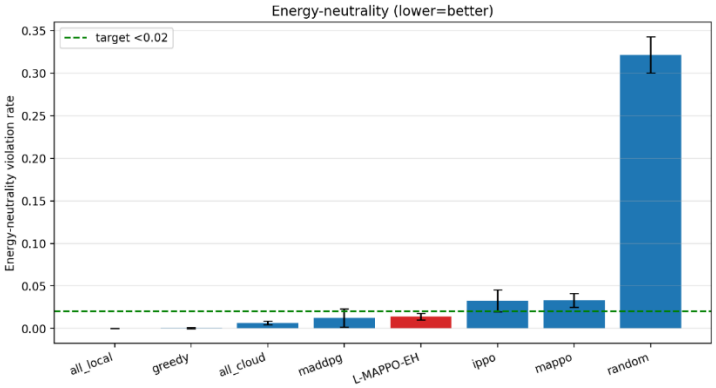

Fig. 8. Energy-neutrality violation rate per algorithm (lower is better); dashed line marks the 0.02 target

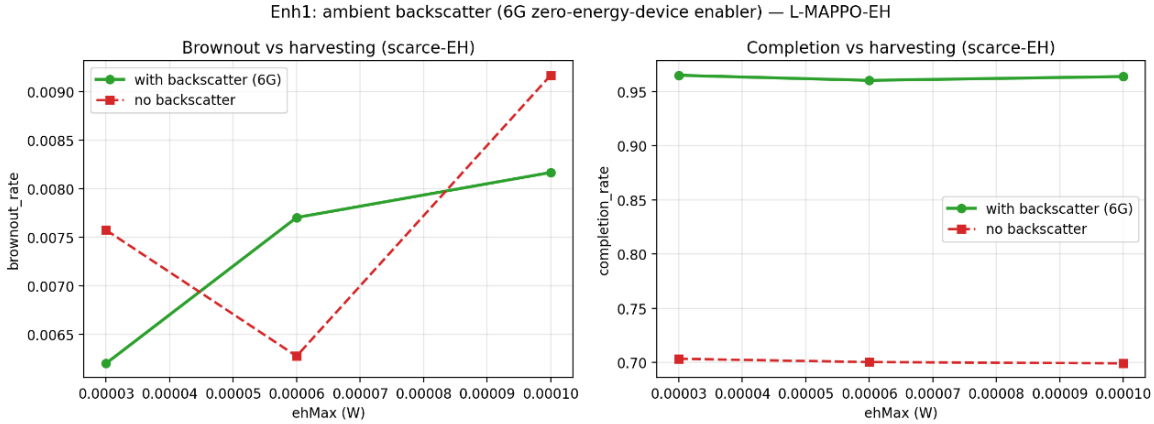

Fig. 9. Effect of **6G ambient backscatter under scarce harvesting: brownout rate and completion rate** vs. ehMax, with vs. without the backscatter transmission mode