Electrical Power and Energy Systems 179 (2026) 112020

Contents lists available at ScienceDirect

International Journal of Electrical Power and Energy Systems

journal homepage: www.elsevier.com/locate/ijepes

Energy-efficient quantum-spiking multi-agent reinforcement learning for adaptive energy management in microgrid networks

Lilia Tightiz [a], Hong Nhung Nguyen [b], L. Minh Dang [c,d], Sanjeevikumar Padmanaban [e,*], Hyosik Yang [a,*]

[a] Department of Computer Engineering, Sejong University, 209, Neungdong-ro, Gwangjin-gu, Seoul, 05006, Republic of Korea
[b] Department of AI, FPT university, Viet Nam
[c] The Institute of Research and Development, Duy Tan University, Da Nang, 550000, Viet Nam
[d] Faculty of Information Technology, Duy Tan University, Da Nang, 550000, Viet Nam
[e] Department of Electrical Engineering, IT and Cybernetic, University of South-Eastern Norway – Campus Porsgrunn, 7430, Norway

ARTICLE INFO

Keywords:
Micro-grid energy management
Multi-agent reinforcement learning
Spiking neural networks
Parameterized quantum circuits
Decentralized control

ABSTRACT

Interconnected micro-grids (MG) energy management requires coordinated decision-making under renewable uncertainty, distributed operation, and network-level trading constraints. Conventional centralized reinforcement learning (RL) approaches often suffer from scalability limitations, large communication overhead, and high computational cost for decentralized edge-level deployment. Existing multi-agent reinforcement learning (MARL) methods also commonly rely on dense artificial neural network (ANN) models and do not explicitly enforce global trading balance during decentralized execution. This paper presents a quantum-spiking-MARL (QS-MARL) framework for adaptive energy management in interconnected MG networks. The proposed method combines decentralized event-driven spiking neural network (SNN) policies for low-power local control with a centralized parameterized quantum circuit (PQC) critic that estimates the global value function under the centralized training and decentralized execution (CTDE) paradigm. Each MG employs a low-power, temporally-aware SNN policy, while a centralized PQC critic captures inter-agent trading dependencies during training under a CTDE paradigm. The PQC critic uses a qubit-per-agent representation with shallow nearest-neighbor entanglement, which maintains linear scaling with the number of MGs. Trading balance is enforced through a differentiable penalty mechanism without requiring centralized execution. Experiments on CityLearn urban districts and a synthetic Pecan+NREL mixed-use cluster show that QS-MARL reduces average daily operational cost by approximately 9% compared to ANN-based CTDE baselines and by approximately 14% compared to fully decentralized MARL, while remaining within 6%–7% of a centralized perfect-foresight oracle. QS-MARL also achieves significant emissions reductions (10%–25%), maintains the narrowest cost distribution under renewable and load uncertainty among all baselines, and outperforms multi-agent proximal policy optimization (MAPPO) and Q-value mixing network (QMIX) in both cost efficiency and constraint satisfaction. Ablation and sensitivity analyses confirm that both the SNN policies and the PQC critic are important to the observed performance gains. These results demonstrate that QS-MARL provides a computationally efficient and coordinated decentralized control method for distributed MG energy management.

1. Introduction

Modern distribution networks increasingly rely on interconnected microgrids (MGs) to integrate renewable generation, improve reliability, and support flexible demand response. In such systems, energy management must coordinate local generation, storage, and power exchange while minimize operating cost and emissions under uncertainty in load and renewable supply.

MG operation presents several technical challenges. The increasing penetration of inverter-based renewable sources reduces system inertia and degrades frequency and voltage stability [1]. The behavior of converter-dominated grids depends on control design rather than physical dynamics, which increases sensitivity to disturbances and faults [2]. In addition, actuator faults and saturation constraints degrade control performance and affect reliable operation of islanded MGs [3]. Furthermore, stable operation requires coordinated control across distributed

* Corresponding authors.
E-mail addresses: sanjeev.padma@usn.no (S. Padmanaban), hsyang@sejong.edu (H. Yang).

https://doi.org/10.1016/j.ijepes.2026.112020
Received 23 March 2026; Received in revised form 13 June 2026; Accepted 17 June 2026
Available online 27 June 2026

units, where centralized and distributed strategies must balance communication overhead, scalability, and reliability [4]. These challenges indicate that advanced control and coordination methods are required for practical MG energy management.

Reinforcement learning (RL) has emerged as a promising approach for adaptive, data-driven energy management in MGs [5]. By learning control policies directly from interaction data, deep reinforcement learning (DRL) can optimize dispatch and trading decisions without requiring explicit system models. However, applying DRL to interconnected MG networks remains challenging in practice.

First, interconnected MGs form a naturally multi-agent system: each MG acts on local information, yet its decisions are coupled to others through shared trading relationships and network-level constraints. Conventional centralized RL frameworks scale poorly and may not respect privacy or decentralized execution requirements. Although multi-agent reinforcement learning (MARL) approaches have been explored [6,7], enforcing global coordination constraints such as trading balance remains difficult.

Second, energy management systems (EMS) are increasingly deployed on resource-constrained edge platforms, where conventional dense neural networks may impose nontrivial computational and energy overhead. Spiking neural networks (SNNs) provide an event-driven alternative with favorable energy-efficiency properties for neuromorphic hardware [8,9]. Their temporal dynamics match sequential MG observations such as demand, renewable generation, and price trajectories.

At the same time, PQCs have attracted attention as compact function approximators for reinforcement learning [10]. Through shallow entangling circuit structures, PQCs capture cross-variable dependencies while remaining compatible with NISQ devices [11]. However, their use in practical multi-MG energy management with decentralized policy learning remains limited.

Existing approaches do not jointly address energy-efficient policy execution, structured modeling of inter-agent dependencies, and explicit enforcement of trading constraints. Most MARL methods rely on dense neural networks and do not consider low-power execution at edge devices. In addition, coordination among MGs is often handled indirectly through price signals, without explicit constraint enforcement.

The proposed method is not a simple combination of SNN and PQC. The SNN is used as a control policy for decentralized decision-making. The PQC is used as a centralized value estimator to capture inter-agent dependencies under the centralized training and decentralized execution (CTDE) paradigm. Trading constraints are explicitly enforced through a differentiable penalty. This design distinguishes the proposed method from prior hybrid MARL approaches.

To address these challenges, this paper proposes an energy-efficient QS-MARL approach for adaptive energy management in interconnected MG networks. The proposed approach combines decentralized SNN-based actors for low-power local control with a centralized PQC-based critic for coordinated value estimation under the CTDE paradigm. The main contributions are as follows:

- We formulate interconnected MG energy management as a cooperative multi-agent Markov decision process (MDP) with explicit trading constraints, enabling coordinated decision-making under decentralized execution.
- We propose event-driven SNN-based policies as decentralized MG control policies, providing temporally-aware and energy-efficient decision-making suitable for edge deployment.
- We introduce a centralized PQC critic as a global value estimator, using qubit-per-agent encoding and shallow entangling structures to capture inter-agent dependencies under CTDE.
- We develop a unified CTDE learning framework that integrates surrogate-gradient training for SNN actors, parameter-shift optimization for the PQC critic, and differentiable penalty mechanisms for enforcing trading and network constraints.
- We demonstrate, through extensive simulations on CityLearn and Pecan+NREL scenarios, that the proposed framework improves cost efficiency, emissions, robustness, and scalability compared to state-of-the-art MARL baselines.

2. Related work

MG control and stability have been studied using model-based and control-theoretic approaches. Robust control methods have been proposed to address low inertia and disturbance sensitivity in inverter-fed power systems [1]. Distributed and fault-tolerant control strategies have been developed to ensure voltage and frequency regulation under actuator faults and saturation constraints [3]. However, these approaches rely on predefined models and do not provide adaptive decision-making under uncertainty and dynamic environments.

Reinforcement learning (RL) has become a promising approach for energy management in MG networks, enabling adaptive scheduling under uncertainty. Existing studies can be broadly grouped into four directions: single-MG EMS optimization, multi-agent coordination in interconnected MGs, constraint-aware and scalable RL architectures, and emerging learning paradigms such as spiking and quantum-enhanced models.

For single-MG EMS, Tightiz et al. [12] developed a double deep-Q network (DDQN)-based framework for Korean net-zero residential MGs, incorporating business model constraints to satisfy regulatory requirements while reducing cost. In follow-up work [13], they applied soft actor–critic (SAC) to islanded MGs with demand response participation to reduce diesel generation. Sun et al. [14] proposed an enhanced D3QN method with mixed penalty terms for low-carbon economic operation. Although effective for local scheduling, these studies do not address coordination across multiple MGs coupled through trading and shared network constraints.

To capture interactions among interconnected MGs, MARL methods have been introduced. Li et al. [6] proposed a MADRL framework robust to missing measurements in multi-MG EMS. Zhang et al. [7] studied RL-based coordination of networked MGs under real-time pricing within a CTDE setting. Wang et al. [15], Zhao et al. [16], and Wu et al. [17] further demonstrated the potential of distributed and model-free RL for energy dispatch and trading in interconnected MG systems. However, these approaches generally rely on conventional dense neural networks and typically treat trading feasibility through indirect price responses or soft penalties rather than explicitly coupling coordination quality with a structured global value estimator.

Beyond MG-specific EMS studies, cooperative MARL algorithms such as QMIX, MAPPO, and MADDPG have also been applied to related energy coordination problems. Zhang et al. [18] formulated privacy-preserving demand-side management for multi-MG users as a Dec-POMDP and developed a CTDE-based MARL framework with an encryptor network. Jiang et al. [19] proposed a Double Hypernetwork QMIX method for cooperative energy management and trading among electric vehicle charging stations. These works highlight the strength of cooperative MARL in modeling inter-agent coupling, but they remain based on conventional neural networks, focus primarily on profit or privacy objectives, and do not consider energy-efficient neuromorphic policies or quantum-enhanced critics for MG trading under explicit balance constraints.

Constraint-aware RL has also received increasing attention. Li et al. [20] introduced a safe RL framework with short-horizon forecasts to improve robustness under uncertainty. Chen et al. [21] proposed multi-agent safe policy learning for risk-constrained power management in networked MGs. Zhou et al. [22] developed a two-step diffusion-policy DRL approach for low-carbon multi-energy MG EMS. While these studies improve safety and feasibility, they continue to depend on conventional deep neural networks and do not target low-power decentralized execution.

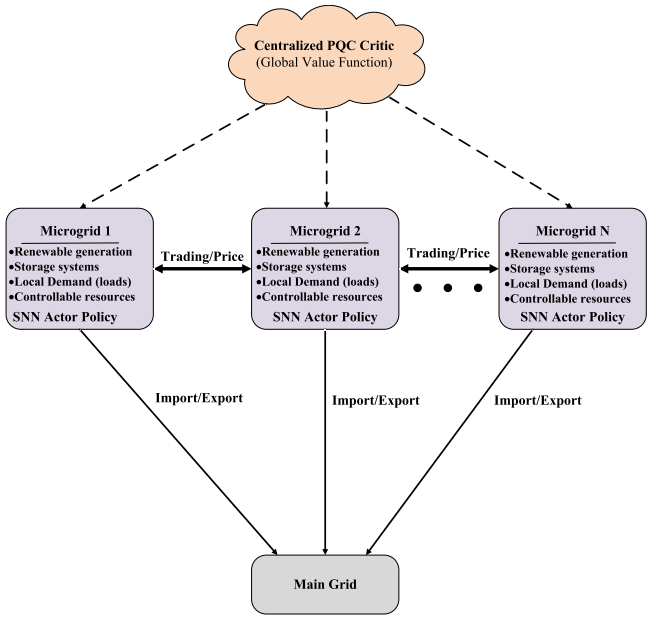


Fig. 1. System architecture for QS-MARL: interconnected MGs equipped with local SNN-based policy agents and a centralized PQC-based value critic coordinating joint energy management.

Scalable and modular RL architectures have likewise been explored for practical EMS deployment. Sharma et al. [23] proposed a DRL framework for IIoT-enabled MG EMS with modular communication-aware design. Ahmed et al. [24] developed a modular DRL-based EMS architecture to improve scalability. Lin et al. [25] studied DDQN-based distributed battery storage operation under uncertainty, and Kim et al. [26] used DRL for coordinated battery swapping in islanded MGs. Although these approaches improve scalability, they do not explicitly address the joint challenge of low-power policy execution and global coordination under trading constraints.

Outside mainstream ANN-based RL, SNNs have shown promise for energy-efficient, event-driven computation on neuromorphic hardware. Zhang et al. [27] demonstrated the applicability of spiking neural models to smart-grid fault diagnosis. However, the use of SNNs as decentralized control policies for multi-MG EMS remains largely unexplored.

In parallel, quantum RL has emerged as a mechanism for capturing high-dimensional correlations through parameterized quantum circuits. Zhang et al. [28] demonstrated the use of PQCs for value approximation in RL, showing their potential to represent complex dependencies compactly. Yet these studies remain limited to relatively small-scale control settings and do not address practical multi-MG EMS with trading constraints, decentralized execution, and topology-aware coordination.

In contrast to the above literature, the proposed QS-MARL approach is distinguished by its role-specific integration of SNN and PQC within CTDE. The SNN is not used only as a low-power neural model, but also as the decentralized control policy of each MG. The PQC is used not only as a compact function approximator but also as the centralized critic that estimates the global value from compressed multi-agent information. In addition, trading-balance and line-capacity constraints are not treated only as post-processing rules, but are included in the learning objective through differentiable penalties. Therefore, the novelty lies in the joint design of low-power local control, quantum-structured global value estimation, and explicit constraint-aware coordination for interconnected MG energy management.

3. System model and problem formulation

We consider a distribution-level network comprising N interconnected MGs, each acting as an autonomous agent in a multi-agent system. The MGs share energy via local trading and can also exchange power with the main grid. The goal is adaptive and energy-efficient management under uncertainty in load and renewable generation.

The overall system architecture and interaction structure are illustrated in Fig. 1. Each MG operates as a decentralized agent with local observations and control actions, while coordination is achieved through a centralized value function during training under the CTDE paradigm. The figure highlights bilateral energy trading among MGs, interactions with the main grid, and the role of the centralized PQC critic in capturing global system dependencies.

3.1. MG network model

Each MG $i \in \{1, 2, \dots, N\}$ consists of: Local generation units (renewable and dispatchable), Energy storage systems (ESS), Flexible and inflexible loads, and Power trading interfaces with neighboring MGs and the main grid.

The active-power balance at time t is modeled as

$$P_{i,t}^{\text{gen}} + P_{i,t}^{\text{import}} + P_{i,t}^{\text{trade}} = P_{i,t}^{\text{load}} + P_{i,t}^{\text{storage}}. \tag{1}$$

This representation follows the standard "aggregated-bus" abstraction widely adopted in multi-agent energy management studies [17–19]. Here, each MG is treated as an equivalent active-power node, and reactive power flows, voltage magnitudes, and branch loss models are omitted. This modeling choice isolates the impact of the learning architecture from detailed AC network physics, allowing the proposed QS-MARL framework to focus on coordination, temporal decision-making, and constraint-aware control.

In practical distribution systems, power flow, voltage limits, and line thermal constraints are indeed important. However, incorporating full AC-OPF dynamics would introduce additional continuous variables, coupled nonlinearities, and nonconvex feasibility regions. These would obscure the algorithmic contributions of the quantum-inspired critic and the event-driven SNN actors. For this reason, we adopt the commonly used active-power formulation and later introduce a more topology-aware validation scenario with line capacity limits in Section 5.1, which partially captures congestion phenomena while preserving tractability.

The storage state of charge (SoC) evolves according to

$$\text{SoC}_{i,t+1} = \text{SoC}_{i,t} + \eta_i P_{i,t}^{\text{storage}} \Delta t, \tag{2}$$

subject to

$$\text{SoC}_i^{\min} \le \text{SoC}_{i,t} \le \text{SoC}_i^{\max}. \tag{3}$$

The storage dynamics in (2)–(3) directly constrain the feasible action space of each agent, particularly the charging and discharging decision $P_{i,t}^{\text{storage}}$ defined in (9). These constraints ensure physically valid SoC evolution over time and are enforced implicitly through environment transitions and explicitly through the reward formulation with penalty terms described in Section 4.5.

3.2. Energy management objectives

The system aims to minimize expected operational cost while satisfying network-wide coupling. The cost function for MG i over horizon T is

$$J_i = \mathbb{E}\left[\sum_{t=0}^{T}\left(C_{i,t}^{\text{gen}} + C_{i,t}^{\text{grid}} + C_{i,t}^{\text{trade}} + C_{i,t}^{\text{emission}}\right)\right]. \tag{4}$$

Generation, grid, trading, and emission costs are defined as:

$$C_{i,t}^{\text{gen}} = c_i^{\text{gen}} P_{i,t}^{\text{gen}}, \qquad C_{i,t}^{\text{grid}} = \lambda_t P_{i,t}^{\text{import}}, \tag{5}$$

$$C_{i,t}^{\text{trade}} = \sum_{j \in \mathcal{N}_i} \pi_{ij,t} P_{ij,t}^{\text{trade}}, \qquad C_{i,t}^{\text{emission}} = e_i P_{i,t}^{\text{gen}}. \tag{6}$$

3.3. Multi-agent MDP formulation

We formulate the adaptive EMS problem as a cooperative Markov game with a shared team reward under the CTDE paradigm, following standard MARL formulations [29]. In this formulation, each agent optimizes a decentralized policy based on local information, while a shared reward aligns all agents toward a common objective.

$$\mathcal{M} = (S, \mathcal{A}, P, R, \gamma), \tag{7}$$

where S denotes the joint state space, $\mathcal{A}$ the joint action space, P the transition kernel capturing uncertainty in demand and renewable generation, R the team reward, and $\gamma \in (0, 1)$ the discount factor.

At time step t, each MG agent $i \in \{1, \ldots, N\}$ receives a local observation

$$o_{i,t} = \left(\text{SoC}_{i,t}, P_{i,t}^{\text{load}}, P_{i,t}^{\text{renew}}, \lambda_t, \xi_{i,t}\right), \tag{8}$$

where $SoC_{i,t}$ is the storage SoC when storage is available (and zero otherwise), $P_{i,t}^{\text{load}}$ is the local demand, $P_{i,t}^{\text{renew}}$ is the renewable generation associated with the agent when available (and zero otherwise), λ_t is the grid electricity price when available or externally defined pricing signal used for training, and $\xi_{i,t}$ denotes optional environment-specific auxiliary inputs such as local trading-price information or short-horizon demand and renewable forecasts when available in the dataset. This formulation accommodates both datasets used in this study after preprocessing and feature selection, where a compact, task-relevant subset of variables (e.g., demand, renewable generation, storage state, and price signals) is extracted from the original dataset and used as the agent's local input. The local observation is derived from the global system state and reflects partial information available to each agent, consistent with partially observable MARL settings [30].

Each agent selects a local action

$$a_{i,t} = \left(P_{i,t}^{\text{gen}}, P_{i,t}^{\text{storage}}, P_{i,t}^{\text{trade}}\right), \tag{9}$$

where $P_{i,t}^{\text{gen}}$ represents controllable generation when available, or is interpreted as net grid import/export in environments without dispatchable units, $P_{i,t}^{\text{storage}}$ denotes battery charging/discharging power when storage is available, and $P_{i,t}^{\text{trade}}$ represents net peer-to-peer energy trading power. The action space is defined in a flexible and environment-adaptive manner, ensuring consistency across heterogeneous datasets with varying levels of controllable resources.

The base instantaneous reward of agent i is defined as

$$r_{i,t}^{\text{base}} = -\left(C_{i,t}^{\text{gen}} + C_{i,t}^{\text{grid}} + C_{i,t}^{\text{trade}} + C_{i,t}^{\text{emission}}\right), \tag{10}$$

which represents the negative local operating cost.

The cooperative team reward is then defined as

$$R_t = \sum_{i=1}^{N} r_{i,t}, \tag{11}$$

where $r_{i,t}$ denotes the reward actually used during training, i.e., the base reward in (10) augmented with constraint-violation penalties as described in Section 4.5.

The global performance objective is to minimize the cumulative discounted system cost,

$$\min_{\{\pi_i\}} \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t \sum_{i=1}^{N} \left(C_{i,t}^{\text{gen}} + C_{i,t}^{\text{grid}} + C_{i,t}^{\text{trade}} + C_{i,t}^{\text{emission}}\right)\right]. \tag{12}$$

This corresponds to a cooperative Markov game with a shared team reward. All agents share an aligned global objective, and no adversarial interactions are considered. Therefore, unlike general-sum stochastic games, the learning target is a coordinated joint policy that optimizes system-wide performance under decentralized execution.

Remark (*Quantum-Structured Value Approximation*)**.** The EMS environment remains a classical cooperative Markov game with standard Markovian dynamics. The quantum component appears only in the centralized critic. During training, the joint multi-agent state is compressed into a qubit-compatible representation and processed by a PQC to estimate the global value function. Thus, the proposed framework preserves standard MDP semantics while introducing a quantum-structured value approximator for capturing inter-agent dependencies.

3.4. Agent interaction and constraints

MGs interact through bilateral energy trading and dynamic price exchanges. The aggregate trading balance must satisfy

$$\sum_{i=1}^{N} P_{i,t}^{\text{trade}} = 0, \tag{13}$$

which ensures a closed local trading market and captures a key global coupling among decentralized decision makers. In the extended-topology scenario (Section 5.1), we additionally incorporate line-capacity constraints across interconnecting feeders, enabling the evaluation of congestion-aware behavior under realistic network limitations.

Despite operating on local observations and executing decentralized actions, all agents optimize the shared global objective in (12), confirming the cooperative nature of the problem. This structure motivates a CTDE MARL framework in which a joint value function, implemented here via a PQC critic (Section 4.3), coordinates distributed SNN-based policies (Section 4.4) while respecting trading and network-level constraints.

4. Proposed method: QS-MARL

We propose the QS-MARL framework for adaptive energy management in interconnected MG networks. Building on the system model in Section 3, QS-MARL integrates decentralized SNN policies at each MG with a centralized PQC critic under the CTDE paradigm, as illustrated in Fig. 1.

Each MG agent i implements a local policy $\pi_i(a_{i,t} \mid o_{i,t}; \theta_i)$ using an event-driven SNN. The SNN processes local observation features, such as load forecasts, renewable generation predictions, storage SoC, and price signals. These inputs are encoded into spike trains for energy-efficient neuromorphic processing on edge devices.

The local observation is derived from the global system state and provides partial information available to each agent, consistent with partially observable MARL formulations [30].

4.1. Spiking neuron model and encoding

Local policies are modeled as discrete-time leaky integrate-and-fire (LIF) SNNs. For neuron k at time t, the membrane potential $u_{k,t}$ evolves according to:

$$u_{k,t} = \lambda u_{k,t-1} + W_k x_t + b_k, \tag{14}$$

where λ is the leak factor, W_k is the synaptic weight vector, x_t is the encoded input vector, and b_k is a bias term. A spike is emitted when the membrane potential crosses a threshold θ:

$$z_{k,t} = H(u_{k,t} - \theta), \tag{15}$$

where $H(\cdot)$ is the Heaviside step function.

Continuous-valued local observation features $o_{i,t}$ are encoded into spike trains using rate-based or temporal coding. Load forecasts are transformed into Poisson-distributed spike trains proportional to normalized demand levels. Price signals and SoC trajectories are encoded using rate or time-to-first-spike schemes. This encoding preserves temporal patterns required for demand prediction, storage scheduling, and trading decisions.

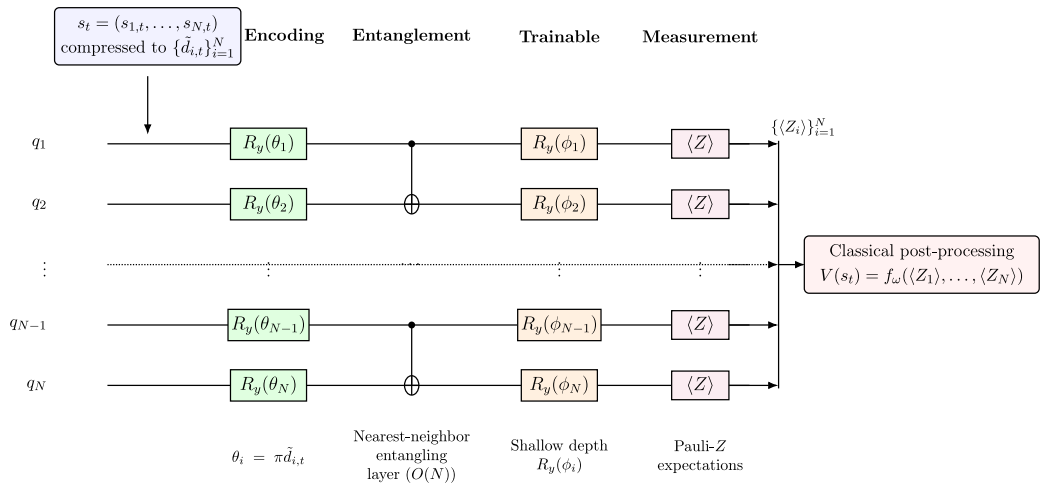


Fig. 2. Proposed PQC critic architecture for QS-MARL. The joint state is compressed into one scalar per agent and encoded through single-qubit R_y rotations.

4.2. Surrogate gradient training

To enable gradient-based learning despite the non-differentiability of $H(\cdot)$, we employ surrogate gradients during backpropagation through time (BPTT):

$$\frac{\partial H(u)}{\partial u} \approx \frac{1}{1+\alpha|u-\theta|}, \tag{16}$$

where α controls the sharpness of the surrogate gradient around the firing threshold. This approximation enables stable temporal credit assignment in spiking neural networks while preserving the discrete spiking mechanism.

The surrogate gradient mechanism allows efficient training of event-driven policies over long decision horizons, which is essential for EMSs with temporally correlated dynamics.

4.3. Global value estimation with PQC

While policy execution remains decentralized, training requires a centralized value function to capture inter-agent dependencies induced by shared grid constraints and energy trading. In QS-MARL, this value function is represented by a PQC, which provides a structured representation of the joint system state within the CTDE paradigm. The overall PQC critic architecture is shown in Fig. 2. The circuit is defined for the general N-agent case, making the qubit-per-agent scaling explicit.

The PQC critic adopts a qubit-per-agent encoding, assigning one qubit to each MG. For a system with N agents, the critic therefore operates on N qubits, which yields linear scaling in quantum resources. This design matches the physical multi-MG structure and avoids the exponential growth associated with direct representations of the full joint state.

To enable hardware-efficient encoding, the global system state is compressed into one representative scalar per agent defined by net demand:

$$d_{i,t}^{\text{net}} = P_{i,t}^{\text{load}} - P_{i,t}^{\text{renew}}. \tag{17}$$

It is noted that this compression is performed on the global state representation available during centralized training. The compressed variable is normalized to a bounded interval and mapped to a rotation angle:

$$\theta_i = \pi \tilde{d}_{i,t}, \qquad \tilde{d}_{i,t} \in [-1, 1], \tag{18}$$

where $\tilde{d}_{i,t}$ denotes the normalized net demand of agent i. Each qubit is initialized in $|0\rangle$ and encoded through

$$R_y(\theta_i) = \exp\left(-i\frac{\theta_i}{2}Y\right). \tag{19}$$

After state encoding, the PQC applies trainable single-qubit rotations and an entangling layer. A single circuit layer is defined as

$$U(\theta, \phi) = \left(\bigotimes_{i=1}^{N} R_y(\phi_i)\right)\left(\prod_{i=1}^{N-1} \text{CNOT}_{i,i+1}\right)\left(\bigotimes_{i=1}^{N} R_y(\theta_i)\right), \tag{20}$$

where ϕ_i denotes the trainable rotation parameter of qubit i.

The entangling layer follows a nearest-neighbor CNOT chain. This is a deliberate design choice: it preserves linear gate complexity, supports shallow circuit depth, and remains compatible with NISQ-oriented execution. A fully connected entanglement pattern would increase gate count quadratically with the number of agents and would weaken scalability for larger multi-MG systems. In the present implementation, the PQC critic is evaluated with up to 16 qubits while maintaining shallow circuit structure.

After the circuit is applied, the expectation value of the Pauli-Z observable is measured on each qubit:

$$\langle Z_i \rangle = \langle \psi(\theta, \phi) | Z_i | \psi(\theta, \phi) \rangle. \tag{21}$$

These measurements form a compact representation of the global state and are processed by a lightweight classical function to produce the scalar value estimate:

$$V(s_t; \phi, \omega) = f_\omega\big(\langle Z_1 \rangle, \ldots, \langle Z_N \rangle\big). \tag{22}$$

The resulting circuit depth is kept shallow, and the gate count scales linearly with the number of agents, requiring $O(N)$ single-qubit rotations and $O(N)$ entangling gates. This keeps the PQC critic compatible with near-term quantum hardware while supporting larger multi-agent systems. Overall, the PQC critic provides a structured and scalable mechanism for modeling global coordination through quantum correlations within the CTDE framework.

The PQC critic is used as a compact value approximator for inter-agent dependency estimation under centralized training. Unlike deeper classical critics, the PQC critic maintains a shallow circuit structure with linear qubit scaling as the number of MGs increases. To examine whether the performance improvement originates only from critic capacity, an additional comparison with a capacity-matched classical critic is presented in Section 5.2.

Scalability is supported through both the decentralized SNN actors and the qubit-per-agent PQC critic structure. During execution, each MG uses only local observations and does not require access to the full joint state, which limits communication overhead as the number of agents increases. In the centralized critic, the number of qubits and entangling gates scales linearly with the number of MGs. This avoids the exponential complexity associated with full joint-state representations and supports extension to larger interconnected MG networks.

4.4. Combined learning procedure

The proposed QS-MARL framework follows a cooperative actor–critic architecture under the CTDE paradigm. Each MG executes an independent local SNN-based policy, while a centralized PQC critic estimates the global value from a compressed representation of the joint state. Since all agents optimize a shared team reward, the setting is formulated as a cooperative Markov game and does not involve competitive or Nash-equilibrium learning.

- **Actors:** Decentralized SNN policies $\pi_i(a_{i,t} \mid o_{i,t}; \theta_i)$ are executed independently by each MG using only local observations.
- **Critic:** A centralized PQC value function $V_\phi(\bar{s}_t)$ that receives a compressed representation $\bar{s}_t$ of the joint state and outputs a scalar estimate of the global value.

During training, the agents interact with the environment in parallel. At each time step, local rewards are first computed from the base operating-cost form in (10) and then augmented with the balance and line-capacity penalties defined in Section 4.5. These penalized local rewards are aggregated into the team reward

$$R_t = \sum_{i=1}^{N} r_{i,t}, \tag{23}$$

which is used for critic training and advantage estimation.

Because the PQC critic operates with one qubit per agent, the joint state is mapped to a compact critic input. Specifically, the critic uses the centralized training information to construct one normalized scalar $\tilde{d}_{i,t}$ per agent, which represents its net operational condition. The critic's input is then written as

$$\bar{s}_t = (\tilde{d}_{1,t}, \tilde{d}_{2,t}, \ldots, \tilde{d}_{N,t}), \tag{24}$$

which is encoded into the PQC as described in Section 4.3.

We adopt a one-step temporal-difference advantage estimate,

$$\hat{A}_t = R_t + \gamma V_\phi(\bar{s}_{t+1}) - V_\phi(\bar{s}_t), \tag{25}$$

where R_t denotes the global team reward at time step t. Each actor is updated using the shared advantage signal:

$$\theta_i \leftarrow \theta_i + \alpha_\theta \, \mathbb{E}\left[\sum_{t=0}^{T} \nabla_{\theta_i} \log \pi_i(a_{i,t} \mid o_{i,t}) \, \hat{A}_t\right]. \tag{26}$$

The centralized PQC critic is trained by minimizing the temporal-difference loss

$$\mathcal{L}_{\text{critic}} = \left(V_\phi(\bar{s}_t) - \hat{R}_t\right)^2, \tag{27}$$

with bootstrapped target

$$\hat{R}_t = R_t + \gamma V_\phi(\bar{s}_{t+1}). \tag{28}$$

The critic parameters ϕ are updated using gradient-based optimization over this temporal-difference (TD) objective.

Overall, the environment, rewards, and transitions remain fully classical, while the centralized value function is represented by a quantum-structured critic to capture inter-agent dependencies during training.

4.5. Constraint handling

To encourage satisfaction of the energy trading balance constraint

$$\sum_{i=1}^{N} P_{i,t}^{\text{trade}} = 0, \tag{29}$$

we incorporate a differentiable quadratic penalty into the training reward. The penalized per-agent reward is defined as

$$r_{i,t} = -\left(C_{i,t}^{\text{gen}} + C_{i,t}^{\text{grid}} + C_{i,t}^{\text{trade}} + C_{i,t}^{\text{emission}}\right) - \frac{\beta_{\text{bal}}}{N}\left(\sum_{j=1}^{N} P_{j,t}^{\text{trade}}\right)^2, \tag{30}$$

where $\beta_{\text{bal}} > 0$ controls the strength of balance enforcement.

In the extended-topology scenario, line-capacity constraints are incorporated through an additional penalty term

$$\mathcal{P}_{\text{line},t} = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \left(\max\{0, |P_{\ell,t}| - P_\ell^{\max}\}\right)^2. \tag{31}$$

The corresponding penalized reward becomes

$$r_{i,t} = -\left(C_{i,t}^{\text{gen}} + C_{i,t}^{\text{grid}} + C_{i,t}^{\text{trade}} + C_{i,t}^{\text{emission}}\right) - \frac{\beta_{\text{bal}}}{N}\left(\sum_{j=1}^{N} P_{j,t}^{\text{trade}}\right)^2 - \beta_{\text{line}} \mathcal{P}_{\text{line},t}, \tag{32}$$

where $\beta_{\text{line}} > 0$ determines the penalty strength for feeder-capacity violations. In addition, the SoC dynamics and constraints in (2)–(3) restrict the feasible action space through $P_{i,t}^{\text{storage}}$ and are enforced via environment transitions. This ensures physically consistent storage behavior and prevents infeasible SoC evolution.

As shown in Algorithm 1, these penalty terms are incorporated directly into the reward computation at each time step so that constraint violations are discouraged during learning. The coefficients β_{bal} and β_{line} are selected to balance effective constraint enforcement and stable training dynamics.

During training, this penalty structure encourages constraint-aware behavior. In practice, the average trading imbalance converges close to zero, and the frequency of line-capacity violations is reduced relative to baseline methods, as reported in Section 5.2.

To improve feasibility at deployment, a lightweight post-processing step can be applied after action selection. The trading decisions are adjusted as

$$P_{i,t}^{\text{trade}} \leftarrow P_{i,t}^{\text{trade}} - \frac{1}{N}\sum_{j=1}^{N} P_{j,t}^{\text{trade}}, \tag{33}$$

which enforces exact balance. In the topology-aware setting, line flows may additionally be limited through a simple clipping-based approximation,

$$P_{\ell,t} \leftarrow \text{clip}\left(P_{\ell,t}, -P_\ell^{\max}, P_\ell^{\max}\right). \tag{34}$$

These mechanisms are summarized in Algorithm 1, which integrates decentralized SNN-based policies, a centralized PQC critic, and constraint-aware learning dynamics.

A key distinction of the proposed framework is that, unlike prior uses of SNNs in power-system applications that often focus on analysis or event-detection tasks, it employs SNNs as online control policies within a cooperative MARL setting.

5. Experimental evaluation

5.1. Experimental setup

We evaluate the proposed QS-MARL method in realistic multi-MG environments consistent with the model in Section 3. Each MG agent manages local generation, storage, demand, and energy trade under renewable uncertainty, while the trade balance constraint is satisfied. In addition to aggregated-bus environments, an extended topology case with feeder capacity limits is used to capture network-induced coupling.

We use two complementary environments to evaluate generalization across different data sources and temporal resolutions:

(i) CityLearn v1.3 [31]: A benchmark multi-agent environment is used. It consists of a district of interconnected buildings, where each building is treated as one MG agent. Each agent includes controllable storage components and is associated with hourly time-series data. These data include electricity demand, on-site renewable generation, weather variables, and price signals. A reduced set of local state features is extracted per agent and compressed into one normalized scalar for the centralized PQC critic.

Algorithm 1: QS-MARL

Input: SNN actor parameters $\{\theta_i\}_{i=1}^{N}$, PQC critic parameters ϕ, learning rates $\alpha_\theta, \alpha_\phi$
Output: Optimized decentralized policies $\{\pi_i\}_{i=1}^{N}$ and centralized critic V_ϕ
for *each episode* **do**
  Initialize environment state s_0;
  for $t = 0$ *to* T **do**
    for *each agent* i *in parallel* **do**
      Receive local observation $o_{i,t}$ derived from s_t;
      Sample action $a_{i,t} \sim \pi_i(a_{i,t} \mid o_{i,t}; \theta_i)$;
    Execute joint action $a_t = (a_{1,t}, \ldots, a_{N,t})$;
    Environment transitions to s_{t+1};
    Compute penalized per-agent rewards $r_{i,t}$ using the base operating cost and constraint penalties;
    Form team reward $R_t = \sum_{i=1}^{N} r_{i,t}$;
    Compress the joint state into critic input $\bar{s}_t = (\tilde{d}_{1,t}, \ldots, \tilde{d}_{N,t})$;
    Compress the next joint state into $\bar{s}_{t+1}$;
    Store $(s_t, a_t, R_t, \bar{s}_t, \bar{s}_{t+1}, s_{t+1})$;
  Sample a minibatch of transitions;
  Compute TD targets $\hat{R}_t = R_t + \gamma V_\phi(\bar{s}_{t+1})$;
  Compute advantages $\hat{A}_t = \hat{R}_t - V_\phi(\bar{s}_t)$;
  for *each agent* i **do**
    Update actor parameters θ_i using (26);
  Update critic parameters ϕ by minimizing (27);
Deployment step: Trading actions are projected to satisfy the power balance constraint. In topology-aware settings, clipping-based flow limits are applied to enforce line capacity constraints.

The raw data consist of building-level demand and renewable generation time series, together with exogenous price and weather signals. The electricity demand corresponds to $P_{i,t}^{\text{load}}$, the renewable generation corresponds to $P_{i,t}^{\text{renew}}$, the electricity price corresponds to λ_t, while meteorological variables, including temperature, humidity, and radiation components, form the auxiliary input vector $\xi_{i,t}$ in Eq. (8).

(ii) Synthetic Pecan+NREL/OpenEI Cluster [32,33]: A constructed 6-agent mixed-use network is used. Residential load profiles are obtained from Pecan Street Dataport. Commercial load profiles and weather variables are obtained from the NREL/OpenEI End-Use Load Profiles (EULP) repository. The dataset uses a 15-minute temporal resolution. Each agent is represented by a compact set of local state features, which are also compressed into one scalar for qubit-per-agent encoding in the PQC critic.

The residential demand traces are derived from measured household electricity consumption under Texas operating conditions. The commercial demand traces are derived from calibrated end-use load profiles representing Texas building characteristics. Weather variables are obtained from meteorological data and include temperature, humidity, and radiation components derived from meteorological measurements, specifically global horizontal radiation, direct normal radiation, and diffuse horizontal radiation. All time series are aligned by timestamp, resampled to a common resolution, and normalized to a bounded range before use in the model.

In this setting, renewable generation is available only in the CityLearn environment, while it is not explicitly modeled in the synthetic Pecan+NREL/OpenEI cluster.

Table 1 defines the relationship between the dataset variables and the MDP formulation in Section 3.

Several EMS-related variables, including storage states and inter-MG trading actions, are not directly available in the raw datasets and are therefore generated dynamically within the simulation environment according to the EMS transition and control equations described in Section 3.

The electricity price signal λ_t in Eq. (8) is constructed using a tariff-informed approximation under Texas conditions. The resulting price profile reflects demand-dependent cost variations during summer operation.

Representative raw data samples and temporal profiles are included to illustrate the statistical characteristics of the input variables.

To evaluate performance under network constraints, we construct a synthetic radial distribution topology in which the MGs are interconnected via feeders $\mathcal{L}$. Each feeder $\ell \in \mathcal{L}$ is assigned a capacity limit $P_{\ell}^{\max}$ in the range of 10–25 kW, representative of low-voltage distribution constraints. The corresponding line flows $P_{\ell,t}$ are derived from inter-agent energy trade actions. Congestion is modeled through a quadratic penalty (Section 4.5). While this formulation does not solve a full AC optimal power flow problem, it provides a tractable and topology-aware approximation for evaluation under network constraints.

5.1.1. Implementation details

Experiments are implemented in Python 3.9 using PyTorch for SNN policies and Qiskit for PQC simulation. The SNN actors employ LIF neurons trained via surrogate gradients, while the PQC critic follows the architecture described in Section 4.3 and illustrated in Fig. 2, including R_y encoding, nearest-neighbor entanglement, and Pauli-Z expectation measurements.

Training is conducted for 300 episodes using a centralized replay buffer under the CTDE paradigm. SNN components are executed on an NVIDIA RTX 3090 GPU, while PQC evaluation is performed on CPU-based simulators. Experiments consider systems with up to 9 qubits (CityLearn) and 6 qubits (Pecan+NREL), with additional scalability tests up to 16 qubits.

5.1.2. Hyperparameter settings

Hyperparameters are selected through grid-search validation experiments under identical training and uncertainty conditions. The tuning procedure is performed separately for optimization stability, constraint satisfaction, and final operational cost. Learning rates, penalty coefficients, SNN parameters, and PQC dimensions are varied within predefined candidate ranges, and the final configuration is selected according to the lowest validation cost with stable convergence behavior.

For the actor and critic optimizers, learning rates in the range $[10^{-4}, 10^{-3}]$ are evaluated. The penalty coefficients β_{bal} and β_{line} are tuned within the ranges $[1, 20]$ and $[1, 10]$, respectively, to balance constraint enforcement and training stability. The surrogate gradient slope parameter α, described in (16), is evaluated within the range $[1, 10]$ and is finally set to $\alpha = 5$ to maintain stable gradient propagation near the firing threshold. The SNN leak factor λ, described in (14), is selected from the range $[0.7, 0.99]$ to preserve temporal dependency information while avoiding unstable neuron dynamics.

Table 2 summarizes the final hyperparameter configuration used in all experiments.

5.1.3. Datasets and environments

We compare QS-MARL with representative MARL baselines (MAPPO, QMIX, ANN-based CTDE, and fully decentralized MARL) under identical observation spaces, action definitions, training horizons, and parameter budgets (see Table 3).

5.2. Results and discussion

We evaluate QS-MARL on the environments described in Section 5.1. All baselines share identical system models, datasets, observation and action spaces, training horizons, and comparable parameter budgets.

Fig. 3(a) compares training performance across methods. All learning-based approaches reduce episodic cost over time, while the rule-based EMS remains unchanged. QS-MARL converges within approximately 200–250 episodes and achieves the lowest final cost. MAPPO shows fast initial improvement but plateaus above QS-MARL, whereas

Table 1
Mapping between dataset variables and MDP variables.

MDP variable	Dataset variable description	Role
$P_{i,t}^{\text{load}}$	Electricity demand time series from residential and commercial profiles	Observed input
$P_{i,t}^{\text{renew}}$	On-site renewable generation when available, and zero otherwise	Observed input
λ_t	Electricity price signal from available tariff data or tariff-informed approximation	Observed input
$\xi_{i,t}$	Auxiliary variables such as weather features, local price information, or short-horizon demand and renewable forecasts when available	Auxiliary input
$SoC_{i,t}$	Internal battery state maintained by the EMS through the storage dynamics in Eq. (2), rather than directly provided by the raw dataset	Internal state
$P_{i,t}^{\text{gen}}$	Controllable generation when available, or net grid import/export representation in environments without dispatchable units	Control action
$P_{i,t}^{\text{storage}}$	Battery charging or discharging power when storage is available	Control action
$P_{i,t}^{\text{trade}}$	Inter-MG energy exchange decision generated by the MARL policy during environment interaction	Control action

Table 2
Hyperparameter configuration of the proposed QS-MARL framework.

Parameter	Final value
Episodes (training)	300
Batch size (samples)	64
Learning rate (Actor)	5×10^{-4}
Learning rate (Critic)	1×10^{-3}
Discount factor γ	0.99
Validation criterion	Lowest validation cost with stable convergence
Penalty coefficient β_{bal}	10
Penalty coefficient β_{line}	5
SNN neurons per layer	64
Surrogate gradient slope α	5
SNN leak factor λ	0.9
Qubits (PQC)	6–9 (dataset-dependent), up to 16 (scalability study)
Feature compression	One feature per agent
Entanglement topology	Linear nearest-neighbor CNOT chain
Optimizer	Adam

Table 3
Architectures used for baseline MARL methods. All ANN-based models use ReLU activations and comparable parameter budgets for fair evaluation.

Baseline method	Policy/Critic architecture
Classical CTDE (ANN)	Policy: 2-layer MLP (64 units per layer, ReLU) Critic: 2-layer MLP (64 units) with full joint-state input Training: Centralized critic with decentralized execution
MAPPO (Cooperative Actor–Critic)	Shared PPO-style actor per agent (2-layer MLP, 64 units) Centralized value function with joint-state input Clipped PPO objective with GAE for coordinated updates
QMIX (Value Factorization)	Decentralized per-agent Q-networks (2-layer MLP, 64 units) Mixer network with hypernetwork-generated weights Monotonicity constraint for joint value decomposition
Classical NN Critic (PQC Replacement)	Policy: Same SNN actors as QS-MARL Critic: ANN with capacity matched to PQC critic Input: Compressed joint-state (1 feature per agent)
Independent MARL (Decentralized SNN)	Policy: Same SNN as QS-MARL (LIF neurons, surrogate gradients) No centralized critic; trained using local rewards only Fully decentralized learning without coordination

QMIX converges more slowly and exhibits higher residual cost. Classical CTDE and Independent MARL perform consistently worse, highlighting the importance of coordinated learning under coupled energy trading.

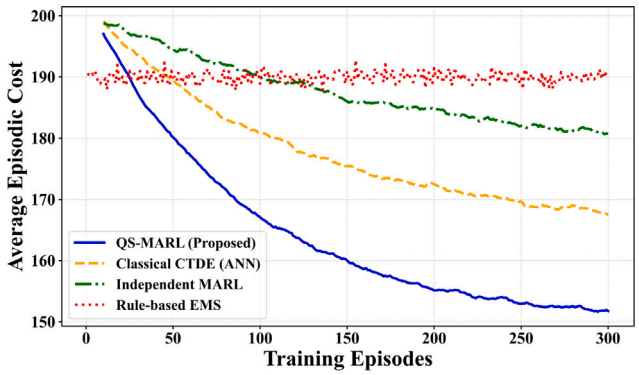
Fig. 3(b) shows the evolution of trading-balance violations. QS-MARL reduces imbalance to near zero during training, whereas MAPPO and QMIX stabilize at non-zero plateaus (approximately 1.2–1.5 kW). Classical CTDE remains around 2 kW, and Independent MARL exhibits the largest persistent imbalance because it lacks a centralized coordination signal. These results indicate that the PQC critic, combined with the penalty-based reward in Section 4.5, improves learning of trading-feasible policies. The values reported in Fig. 3(b) correspond to raw outputs during training; exact balance can be enforced at deployment using the projection step in Section 4.5. Note that the "Classical CTDE (ANN)" baseline corresponds to the PQC replacement variant used in Tables 2 and 3.

Fig. 4 and Table 4 summarize operational cost and emissions. QS-MARL achieves the lowest total cost among all learning-based methods, mainly through lower grid-import and generation costs enabled by better coordination of storage, renewable utilization, and bilateral trading. Relative to Classical CTDE and MAPPO, grid-import cost is reduced by approximately 15%–20%. Emissions follow the same trend because the reward penalizes carbon-intensive generation. Compared with ANN-based baselines, QS-MARL reduces emissions by approximately 15%–30%.

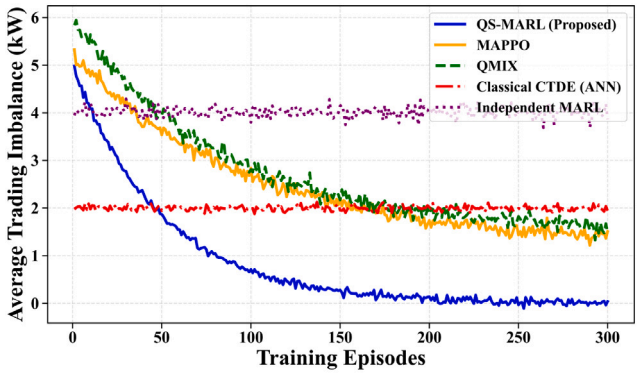
Table 4 also compares all methods against a centralized oracle with perfect foresight. Under the aggregated-bus model, the oracle provides a practical lower bound of 145 USD/day. QS-MARL achieves

Table 4
Average daily cost and emissions comparison.

Method	Total cost ($/day)	CO_2 emissions (tons/day)
Centralized Oracle (perfect foresight)	145	0.95
QS-MARL (Proposed)	155	1.00
MAPPO	167	1.18
QMIX	172	1.22
Classical NN Critic	165	1.15
Classical CTDE (ANN)	170	1.20
Independent MARL	180	1.30
Rule-Based EMS	190	1.40



(a) Average episodic cost over training episodes.



(b) Average trading imbalance over training episodes.

Fig. 3. Training performance comparison (mean over multiple runs) of QS-MARL and baseline methods.

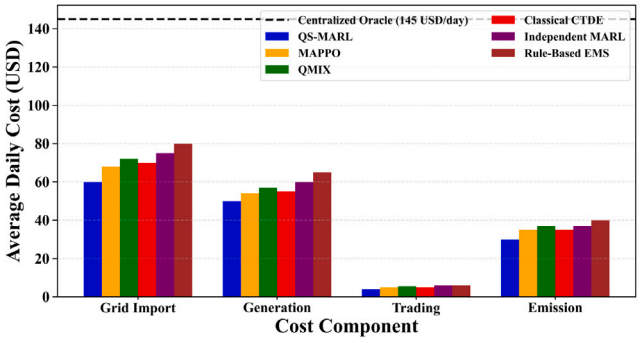


Fig. 4. Average daily cost breakdown for QS-MARL and baselines.

155 USD/day, remaining within approximately 6%–7% of this bound while preserving decentralized execution. MAPPO and QMIX exhibit

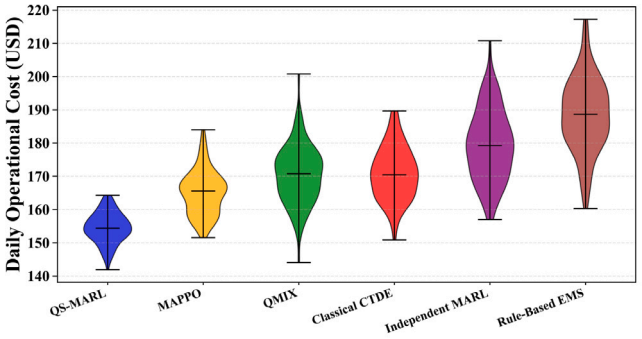


Fig. 5. Distribution of daily operational costs under forecast uncertainty scenarios.

larger gaps of approximately 15% and 19%, respectively, and Classical CTDE and Independent MARL perform substantially worse. This indicates that QS-MARL approaches near-oracle coordination more closely than the other learning-based baselines.

Robustness under uncertainty is shown in Fig. 5, which reports daily cost distributions evaluated over 100 independently sampled uncertainty realizations for each trained policy. No significant performance variance across random seeds was observed, indicating stable convergence behavior. QS-MARL achieves both the lowest median cost and the narrowest spread, indicating more stable performance under renewable variability, load uncertainty, and forecast errors. MAPPO and QMIX show broader distributions and higher medians, while Classical CTDE, Independent MARL, and Rule-Based EMS are more sensitive to uncertainty. Relative to Classical CTDE, the median cost reduction of QS-MARL is approximately 9%–12%; the improvement is about 5%–10% over MAPPO and QMIX, and about 18%–20% over Independent MARL.

Fig. 6 evaluates scalability on networks of 4, 8, and 16 interconnected MGs. QS-MARL maintains the lowest cost per MG across all scales, and its curve remains nearly flat as the number of agents increases. In contrast, Classical CTDE degrades moderately with scale, while Independent MARL shows a clearer loss of coordination. These results are consistent with the qubit-per-agent PQC design and the decentralized SNN policies, which together support larger multi-agent systems without a sharp increase in per-agent cost.

The nearly flat cost trend across increasing MG counts indicates that the proposed architecture maintains coordination quality while preserving scalable decentralized execution.

To further justify the use of the PQC critic, an additional comparison is conducted against a capacity-matched classical critic. The compared variant, denoted QS-MARL-MLP, uses the same SNN actors, reward function, CTDE procedure, and training settings as QS-MARL. Only the critic is replaced by a compact two-layer MLP with a comparable number of trainable parameters. As shown in Table 5, the PQC critic achieves lower average cost and fewer constraint violations across all tested MG counts. The performance gap is modest for 4 MGs, but it becomes clearer for 8 and 16 MGs. This result indicates that the PQC critic provides better coordination efficiency than a compact classical critic with similar capacity.

Table 5
Scalability comparison between the PQC critic and a capacity-matched classical critic.

MG count	Critic	Parameters	Avg. daily cost ($)	Constraint violations
4	PQC	24	102.4	0.012
4	MLP	26	104.1	0.018
8	PQC	48	118.6	0.021
8	MLP	51	123.9	0.037
16	PQC	96	141.3	0.044
16	MLP	101	151.7	0.073

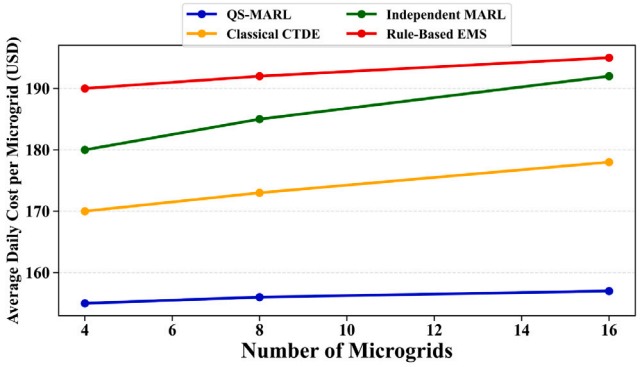


Fig. 6. Average daily operational cost per MG for varying network sizes.

Table 6
Performance under extended radial topology with feeder capacity limits.

Method	Cost (USD/day)	Line violations (per day)
QS-MARL (Proposed)	158	**0.7**
Classical CTDE (ANN)	170	3.4
Independent MARL	183	5.1
Rule-Based EMS	192	6.8

To assess topology-aware coordination, we next consider the extended radial network introduced in Section 5.1. Table 6 reports average daily cost and feeder-capacity violations. QS-MARL again achieves the best performance, with the lowest cost and the fewest violations. Classical CTDE exhibits moderate congestion-related violations, whereas Independent MARL and Rule-Based EMS frequently overuse constrained feeders. These results show that the proposed framework remains effective when network coupling is strengthened beyond the aggregated-bus abstraction.

QS-MARL achieves the lowest cost and the fewest violations among the considered topology-aware baselines. Classical CTDE exhibits moderate congestion-related violations, whereas Independent MARL and Rule-Based EMS frequently exceed feeder-capacity limits. These results indicate that the proposed method remains effective when network coupling is strengthened beyond the aggregated-bus abstraction.

MAPPO and QMIX are not included in Table 6 because their original formulations do not explicitly support feeder-constrained topology-aware coordination under the considered radial MG setting. A fair comparison would require substantial redesign of their coordination and value-decomposition structures beyond their standard implementations. Therefore, the topology-aware comparison is restricted to ANN-based CTDE and decentralized baselines that can be adapted consistently under identical feeder-capacity constraints. This comparison isolates the effect of the PQC critic under topology-aware coordination while avoiding inconsistent baseline reformulations.

Fig. 7 presents the ablation study. Removal of the PQC critic increases the average operational cost by approximately 12%, whereas replacement of the SNN actors with ANN actors increases the cost by approximately 8%. These results indicate that both the PQC critic and the SNN policies contribute to coordinated multi-agent decision quality under the CTDE setting.

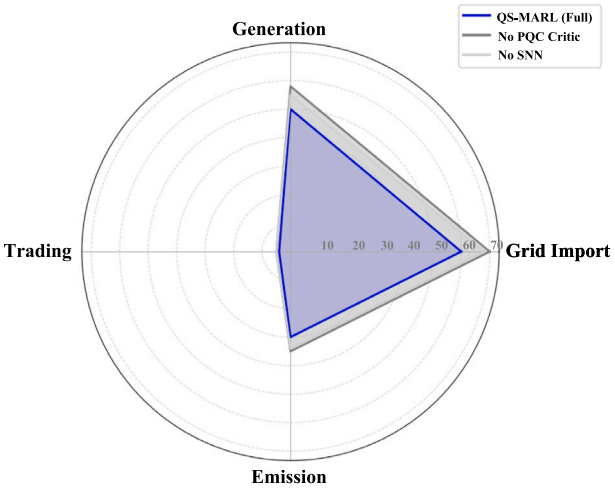


Fig. 7. Ablation study of average daily operational cost.

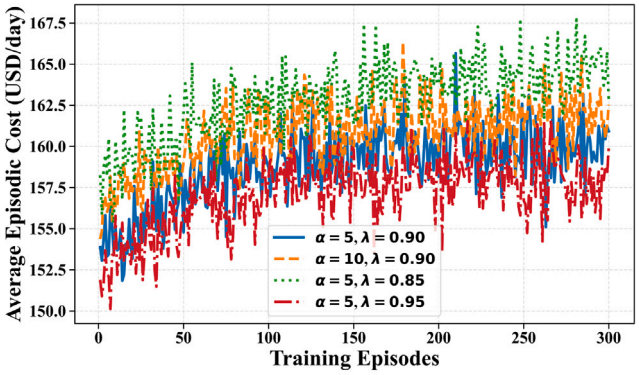


Fig. 8. Sensitivity analysis of surrogate gradient hyperparameters in SNN policy training.

Fig. 8 reports sensitivity to the surrogate-gradient slope α and leak factor λ. Training remains stable across all tested settings, with modest differences in variance and final cost. This suggests that QS-MARL is not overly brittle to moderate hyperparameter changes, although tuning remains beneficial for best performance.

Finally, Fig. 9 evaluates two stress conditions: forecast perturbations with 15% Gaussian noise and partial communication failure with 20% random price-signal drops. QS-MARL shows the smallest relative cost increase (approximately 8%), compared with 14% for Classical CTDE, 18% for Independent MARL, and more than 20% for Rule-Based EMS. This indicates that the proposed framework remains more resilient under both exogenous forecast noise and degraded information exchange.

Overall, the results demonstrate that QS-MARL consistently improves training stability, operating cost, emissions, robustness under uncertainty, scalability, and topology-aware coordination compared to the considered baselines. These improvements arise from the complementary roles of the PQC critic, which captures inter-agent dependencies during centralized training, and the SNN actors, which enable temporally-aware and decentralized decision-making.

Despite these advantages, several limitations remain and point to directions for future work. First, the PQC critic employs a shallow circuit with linear qubit-per-agent scaling, consistent with the formulation in Section 4.3. While this design ensures compatibility with NISQ-era devices, quantum execution latency and hardware overhead for

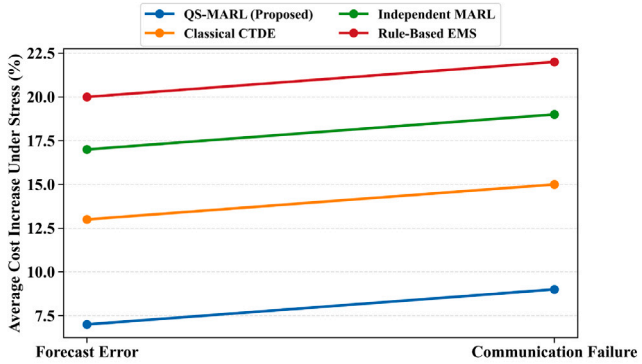


Fig. 9. Average daily cost increase under forecast error and communication failure scenarios.

larger systems (e.g., beyond 16 agents) have not yet been systematically evaluated and require further investigation.

Second, the current system model adopts an aggregated-bus representation with simplified feeder constraints enforced via differentiable penalties (Section 4.5). While this abstraction enables efficient learning of coordinated behavior, it does not capture full AC power flow dynamics, voltage constraints, or reactive power interactions. Extending QS-MARL to more detailed and physically accurate power system models remains an important direction for practical deployment.

Third, although hardware-level measurements are not included, an analytical estimate suggests that SNN-based policies can achieve low inference energy due to their event-driven operation. For a representative LIF-based SNN with 64 neurons per layer and 10 simulation steps, assuming a typical firing rate of approximately 10% and an energy cost of 0.27 μJ per spike (based on Intel Loihi benchmarks [34]), the inference energy is on the order of tens of μJ (approximately 17 μJ under a single-layer approximation). By comparison, ANN-based policies typically require 1–5 mJ per inference on MCU-class processors [35], indicating a potential order-of-magnitude reduction in energy consumption. These results support the feasibility of deploying QS-MARL policies on edge devices, while future work will validate these estimates through direct hardware measurements of energy and latency.

Finally, while this study focuses on CTDE with centralized value estimation under trading constraints, alternative MARL paradigms such as graph-based and federated approaches [36,37] offer complementary directions for scalable coordination and privacy-preserving learning. Integrating these approaches with constraint-aware energy management and quantum-enhanced critics represents a promising direction for future research.

6. Conclusion

This paper proposed QS-MARL, an energy-efficient multi-agent reinforcement learning framework that combines decentralized SNN policies with a centralized PQC critic for coordinated MG energy management. The SNN actors enable low-power, temporally-aware control suitable for edge deployment, while the PQC critic captures inter-agent coupling under the CTDE paradigm. A differentiable penalty mechanism enforces trading balance and supports feasible decentralized execution. Experimental results on CityLearn and Pecan+NREL MG clusters show that QS-MARL outperforms ANN-based CTDE baselines, achieving approximately 9% lower operational cost and about 14% improvement over fully decentralized MARL, while remaining within 6%–7% of a centralized perfect-foresight oracle. The framework also reduces emissions, improves robustness under uncertainty, and achieves better constraint satisfaction compared to MAPPO, QMIX, and other baselines. Ablation results confirm the importance of both SNN actors and the PQC critic. While the approach is currently limited by shallow NISQ-compatible circuits and requires careful tuning, it demonstrates the potential of combining neuromorphic policies with quantum-enhanced value estimation for distributed energy systems. Future work will focus on more scalable circuit designs, heterogeneous MG settings, and hardware-level validation.

CRediT authorship contribution statement

Lilia Tightiz: Writing – review & editing, Writing – original draft, Visualization, Validation, Investigation, Conceptualization. **Hong Nhung Nguyen:** Visualization, Software, Methodology, Conceptualization. **L. Minh Dang:** Visualization, Validation, Supervision, Methodology, Conceptualization. **Sanjeevikumar Padmanaban:** Writing – review & editing, Visualization, Methodology, Investigation, Conceptualization. **Hyosik Yang:** Writing – review & editing, Validation, Methodology, Conceptualization.

Declaration of Generative AI and AI-Assisted Technologies in the Manuscript Preparation Process

During the preparation of this work, the authors used ChatGPT (OpenAI) to assist with grammar correction, language refinement, and improving the readability of the manuscript. After using this tool, the authors carefully reviewed and edited the content as needed and take full responsibility for the content of the published article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors would like to express gratitude to the Clean Energy Transition Partnership (CETP) of the European Commission for funding this project: CETP-FP-2023-00567 “Optimal operation and decentralized control of offshore DC grid with RES/Electrolyzer and hybrid energy storage system” OptiDCG4H2.

Data availability

Data will be made available on request.

References

[1] Xiong L, Huang S, Li P, Wang Z, Khan MW, Niu T. Hidden Markov jump system based robust control of inverter-fed power systems with asynchronous sliding mode observer. IEEE Trans Ind Electron 2025;72(12):13287–99.
[2] Wachter J, Gröll L, Hagenmeyer V. Survey of real-world grid incidents–opportunities, arising challenges and lessons learned for the future converter dominated power system. IEEE Open J Power Electron 2024;5:50–69.
[3] Huang S, Zhou Y, Xiong L, Dong ZY, Huang W, Gao F, Zhou Q, Li X, Zhu L, Li Y. Resilient distributed observer-based fault-tolerant control for islanded AC microgrids subject to actuator fault and saturation constraints. IEEE Trans Consum Electron 2026.
[4] Satapathy AS, Mohanty S, Mohanty A, Rajamony RK, Soudagar ME, Khan TY, Kalam MA, Ali MM, Bashir MN. Emerging technologies, opportunities and challenges for microgrid stability and control. Energy Rep 2024;11:3562–80.
[5] Tightiz L, Yang H. Resilience microgrid as power system integrity protection scheme element with reinforcement learning based management. IEEE Access 2021;9:83963–75.
[6] Li Y, Zhang X, Chen J, Wang Y. Energy management of multiple microgrids considering missing measurements: A novel MADRL approach. IEEE Trans Smart Grid 2023;14(2):1529–40.
[7] Zhang X, Li Y, Chen J. Energy management of networked microgrids with real-time pricing by reinforcement learning. IEEE Trans Smart Grid 2022;13(6):4809–21.
[8] Nunes JD, Carvalho M, Carneiro D, Cardoso JS. Spiking neural networks: A survey. IEEE Access 2022;10:60738–64.

[9] Lobo JL, Del Ser J, Bifet A, Kasabov N. Spiking neural networks and online learning: An overview and perspectives. Neural Netw 2020;121:88–100.
[10] Wilms A, Ohff L, Skolik A, Eisert J, Khatri S, Reiss DA. Quantum reinforcement learning of classical rare dynamics: Enhancement by intrinsic Fourier features. 2025, arXiv preprint arXiv:2504.16258.
[11] Park S, Kim JP, Park C, Jung S, Kim J. Quantum multi-agent reinforcement learning for autonomous mobility cooperation. IEEE Commun Mag 2024;62(6):106–12.
[12] Tightiz L, Yoo J. A novel deep reinforcement learning based business model arrangement for Korean net-zero residential micro-grid considering whole stakeholders' interests. ISA Trans 2023;136:420–33.
[13] Tightiz L, Yoo J. A robust energy management system for Korean green islands project. Sci Rep 2022;12(1):1–14.
[14] Sun S, Wang B, Guo J, Li Y. Low carbon economic energy management method in a microgrid based on enhanced D3QN algorithm with mixed penalty function. Energy 2023;278.
[15] Wang B, Sun S, Guo J, Li Y. Optimized energy dispatch for microgrids with distributed reinforcement learning. IEEE Trans Smart Grid 2022;13(5):4036–46.
[16] Zhao Y, Chen X, Liu J. Reinforcement learning-based energy trading and management of regional interconnected microgrids. Appl Energy 2022;313.
[17] Wu H, Zhao Y, Li X, Wang J. Intelligent multi-microgrid energy management based on deep neural network and model-free reinforcement learning. IEEE Trans Smart Grid 2022;13(4):3219–30.
[18] Zhang F, Yang Q, An D. Privacy preserving demand side management method via multi-agent reinforcement learning. IEEE/CAA J Autom Sin 2023;10(10):1984–99.
[19] Jiang KY, Chiu WY, Tsai YP, Yin L, Xu Z. Profit maximization for electric vehicle charging stations using multiagent reinforcement learning. Sustain Energy,= Grids Networks 2025;102009.
[20] Li Y, Chen J, Zhang X. Robust energy management system with safe reinforcement learning using short-horizon forecasts. IEEE Trans Smart Grid 2022;13(4):3242–53.
[21] Chen J, Zhao Y, Li X. Multi-agent safe policy learning for power management of networked microgrids. Appl Energy 2023;341.
[22] Zhou J, Wang L, Liu Y. Two-step diffusion policy deep reinforcement learning method for low-carbon multi-energy microgrid energy management. Appl Energy 2023;344.
[23] Sharma P, Kumar S, Singh M. A novel deep reinforcement approach for iIoT microgrid energy management systems. Energy 2023;270.
[24] Ahmed M, Khan F, Ali R. Novel architecture of energy management systems based on deep reinforcement learning in microgrid. Energy Rep 2023;9:3105–14.
[25] Lin H, Zhang C, Liu X. Double deep Q-learning-based distributed operation of battery energy storage system considering uncertainties. IEEE Trans Smart Grid 2021;12(4):3201–12.
[26] Kim J, Lee S, Choi H. Renewable energy maximization for pelagic islands network of microgrids through battery swapping using deep reinforcement learning. Appl Energy 2023;348.
[27] Zhang X, Li Y, Chen J. Fault diagnosis of power systems using intuitionistic fuzzy spiking neural p systems. Int J Electr Power Energy Syst 2022;136.
[28] Zhang Z, Wu Y, Li J. Quantum reinforcement learning with parametrized quantum circuits. Npj Quantum Inf 2021;7(1).
[29] Miuccio L, Panno D, Riolo S, Schilirò A. Reliable and energy-efficient MAC protocols in industrial IoT networks via multi-agent reinforcement learning. IEEE Trans Mach Learn Commun Netw 2026;4:677–705.
[30] Miuccio L, Riolo S, Samarakoon S, Bennis M, Panno D. Emerging generalized wireless MAC communication protocols via abstraction. IEEE Open J Commun Soc 2025;6:6842–65.
[31] Vázquez-Canteli Z, Naghibi Z, Henze J, Naghibi Z. CityLearn challenge 2023. 2023, Zenodo. [Online]. Available: https://www.citylearn.net.
[32] Pecan Street Inc. Dataport energy database. 2019, [Online]. Available: https://www.pecanstreet.org/dataport/.
[33] National Renewable Energy Laboratory. Commercial building load profiles. 2020, [Online]. Available: https://data.openei.org/submissions/4520.
[34] Davies M, Wild A, Orchard G, Sandamirskaya Y, Guerra GA, Joshi P, Plank P, Risbud SR. Advancing neuromorphic computing with loihi: A survey of results and outlook. Proc IEEE 2021;109(5):911–34.
[35] Neftci EO, Mostafa H, Zenke F. Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. IEEE Signal Process Mag 2019;36(6):51–63.
[36] Hassouna M, Holzhüter C, Lytaev P, Thomas J, Sick B, Scholz C. Graph reinforcement learning in power grids: A survey. 2024, arXiv e-prints. Jul:arXiv-2407.
[37] Li Y, He S, Li Y, Shi Y, Zeng Z. Federated multiagent deep reinforcement learning approach via physics-informed reward for multimicrogrid energy management. IEEE Trans Neural Networks Learn Syst 2024;35(5):5902–14.