energies

MDPI

Article

A Spatio-Temporal Attention-Based Multi-Agent Deep Reinforcement Learning Approach for Collaborative Community Energy Trading

Sheng Chen [1], Yong Yan [2], Jiahua Hu [2] and Changsen Feng [1,*]

[1] College of Information Engineering, Zhejiang University of Technology, Hangzhou 310026, China; cs583905618@163.com
[2] State Grid Zhejiang Electric Power Research Institute Co., Ltd., Hangzhou 310026, China; yio1986@gmail.com (Y.Y.); dreaminghu2009@163.com (J.H.)
* Correspondence: fcs@tongji.edu.cn

Abstract

The high penetration of distributed energy resources (DERs) poses numerous challenges to community energy management, including intense source-load stochasticity, synchronized load surges triggered by multi-agent gaming, and potential privacy breaches. To tackle these issues, this paper proposes a coordinated energy trading framework driven by an intermediate market-rate pricing mechanism. Within this framework, a novel Multi-Agent Transformer Proximal Policy Optimization (MATPPO) algorithm is developed, adopting an LSTM–Transformer hybrid architecture and the centralized training with decentralized execution (CTDE) paradigm. During centralized training, an LSTM network extracts temporal evolution features from source-load data to handle environmental uncertainty, while a Transformer-based self-attention mechanism reconstructs the dynamic agent topology to capture spatial correlations. In the decentralized execution phase, prosumers make independent decisions using only local observations. This eliminates the need to upload internal device states, significantly enhancing the privacy of sensitive local information during the online execution phase. Additionally, a parameter-sharing mechanism enables agents to share policy networks, significantly enhancing algorithmic scalability. Simulation results demonstrate that MATPPO effectively mitigates power peaks and reduces the transformer capacity pressure at the main grid interface. Furthermore, it significantly lowers total community electricity costs while maintaining high computational efficiency in large-scale scenarios.

Keywords: community energy management; distributed energy resources; multi-agent deep reinforcement learning

Check for updates

Academic Editor: Stéphane Grieu

Received: 2 March 2026
Revised: 19 March 2026
Accepted: 27 March 2026
Published: 1 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

With the large-scale integration of distributed energy resources (DERs), distribution networks are undergoing a profound paradigm shift from unidirectional power supply to interactive source-load dynamics. This transition has given rise to a multitude of prosumers who integrate generation, consumption, and storage capabilities. No longer passive end-users, these prosumers have evolved into autonomous decision-making entities capable of active energy management and market participation. In this context, the community energy market has emerged as a pivotal mechanism designed to promote local energy absorption and deeply exploit the flexibility inherent in load and storage resources [1,2].

Existing optimization methods for community energy markets are primarily categorized into model-based and data-driven approaches. Model-based methods encompass auction theory [3,4], game theory [5,6], and constrained optimization [7–9]. For instance, reference [4] constructs a trading market based on a double auction mechanism tailored for commercial buildings, specifically facilitating demand response via controllable devices. Reference [10] proposes a peer-to-peer (P2P) trading model based on non-cooperative game theory, aiming to maximize social welfare while ensuring rational revenue distribution among energy communities. To further enhance market efficiency, researchers have also employed constrained optimization techniques, such as the Alternating Direction Method of Multipliers (ADMM) [11], to formulate energy trading schemes targeting social welfare maximization. Concurrently, the recent literature has deeply explored the structural design of local electricity trading and optimal prosumer scheduling. For example, Evdokimov et al. [12] proposed a decentralized, blockchain-based architecture for short-term electricity trading and yield derivatives, highlighting the structural evolution of programmable local energy markets. From the perspective of operational planning, Blinov et al. [13] formulated a comprehensive mathematical model for prosumer scheduling, optimizing energy purchase and sale schedules while strategically managing energy storage under generation uncertainties. However, these traditional model-based approaches face a dual dilemma in practical applications. First, they rely heavily on the precision of physical models and the accuracy of source-load forecasting; given the high stochasticity of DERs, prediction errors often lead to significant deviations between scheduled plans and actual operations. Second, as community scale expands, the complexity of game-theoretical interactions causes the modeling difficulty to rise exponentially, making it challenging to meet the requirements of real-time scheduling [14]. Consequently, exploring intelligent scheduling methods that are resilient to strong source-load fluctuations and independent of precise system models has become an urgent research imperative.

In recent years, data-driven approaches represented by deep reinforcement learning (DRL) have offered a novel perspective for addressing these challenges due to their model-free nature. In DRL, agents learn through trial-and-error interactions with the environment; once trained offline, these agents can be deployed for online decision-making, demonstrating robust adaptability to source-load uncertainties [15–18]. Reference [19] employs a Deep Q-Network (DQN) to learn strategies for energy management amidst uncertainties in load demand, renewable generation, and real-time electricity prices. Similarly, Reference [20] utilizes the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm to optimize electric vehicle (EV) charging behaviors on the load side, assuming continuous power adjustability for single EVs. Nevertheless, community energy management is inherently a multi-agent problem involving competition and cooperation. Single-agent approaches struggle to effectively handle the complex interplay and conflicting interests among multiple prosumers.

Multi-Agent Deep Reinforcement Learning (MADRL) facilitates autonomous learning for each prosumer as an independent agent while enabling the cooperative resolution of problems through agent interactions [21]. This approach has been applied to P2P energy trading in multi-source microgrids [22] and distributed collaborative optimization of energy storage systems [23]. However, existing MADRL architectures typically employ simple fully connected or convolutional neural networks as policy extractors, which fail to effectively capture the long-term temporal dependencies of source-load data [24]. More critically, current methods often rely on simplistic global information concatenation or mean-field strategies [25]. These conventional structures struggle to effectively capture the long-term temporal dependencies of source-load data or extract key spatial features from complex

neighbor relationships, thereby trapping agents in local optima and impeding efficient energy absorption within the community.

Despite these advancements, a critical research gap remains explicitly unresolved in the current literature: existing frameworks fail to jointly address the highly dynamic spatiotemporal coupling of renewable energy systems, alongside the stringent requirements for scalability and communication efficiency during online operations in large-scale communities. As the prosumer scale expands, traditional MADRL methods that train networks independently for each agent suffer severely from the curse of dimensionality and environmental non-stationarity. Furthermore, their reliance on continuous, explicit global state sharing among peers not only restricts real-time engineering scalability but also heightens the risk of sensitive device data exposure during the online execution phase.

In response to the above challenges, this paper proposes a coordinated community energy trading framework based on the Mid-Market Rate (MMR) pricing mechanism, and develops a novel Multi-Agent Transformer Proximal Policy Optimization (MATPPO) algorithm. The main contributions of this work are summarized as follows:

(1) A novel MATPPO algorithm integrating a spatiotemporal architecture: Under the Centralized Training with Decentralized Execution (CTDE) paradigm, an LSTM network is employed to extract long-term temporal features to address source-load stochasticity, while a Transformer's self-attention mechanism is leveraged to dynamically capture the complex spatial coupling among prosumers.
(2) Enhanced scalability and privacy preservation: A parameter-sharing mechanism is introduced to overcome the curse of dimensionality and reduce computational overhead. Furthermore, during the execution phase, agents make independent decisions using solely local observations, mitigating the risk of online privacy breaches without requiring explicit communication of sensitive device states during the execution phase.
(3) Superior economic and security performance: Simulation results demonstrate that MATPPO effectively adapts to spatiotemporal uncertainties. It not only significantly reduces the comprehensive community electricity cost but also mitigates active power peaks and the risk of transformer capacity limit violations under a simplified community network model.

2. Framework for Community Energy Collaborative Optimization and Management

The community energy collaborative optimization management framework proposed in this paper is illustrated in Figure 1.

Prosumers: The community comprises N prosumers, each equipped with a set of DERs, including distributed photovoltaics (PV), energy storage (ES), non-controllable loads (NL), and controllable loads. Specifically, controllable loads are categorized into shiftable loads (SL) and adjustable loads (AL). Due to the heterogeneity of DER parameters, each prosumer deploys an independent energy management agent. Based on real-time DER operational states and community market price signals, this agent formulates scheduling strategies aimed at prioritizing participation in internal community energy trading, thereby maximizing local renewable energy absorption and minimizing individual electricity costs.

Cloud Trading Platform: This is typically a digital system endowed with advanced algorithms and data processing capabilities. It is responsible for facilitating local energy trading, determining community market prices, and allocating costs and revenues among market participants.

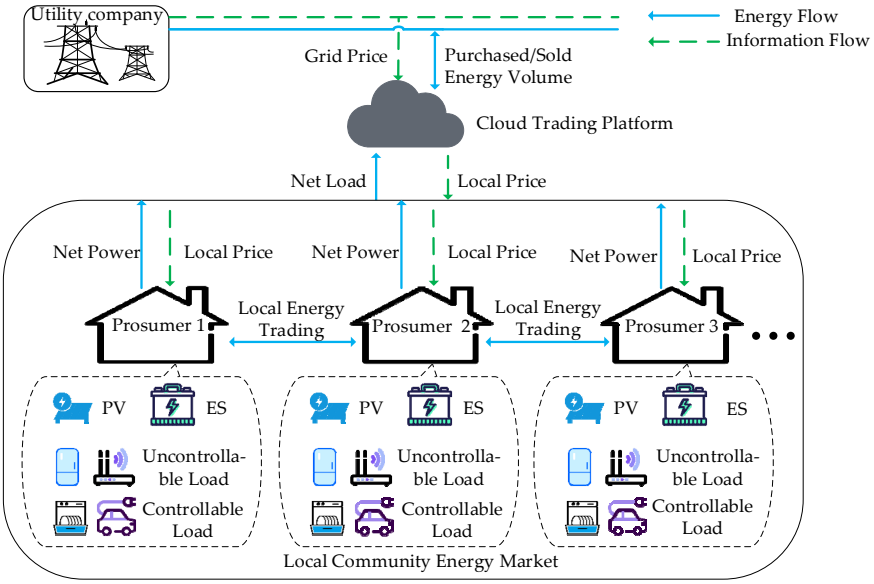


Figure 1. The Community Energy Collaborative Optimization Management Framework.

Grid Company: The utility company owns the associated generation, transmission, and distribution infrastructure. It interacts with prosumers by purchasing surplus energy or supplying deficits at retail electricity prices, thus serving as a supplementary guarantee for the market supply–demand balance.

2.1. Modeling of Operational Characteristics of Shiftable Loads

The core characteristic of controllable loads lies in the flexibility of their operation timing or power consumption. Subject to usage constraints, their operating periods can be shifted, or their power levels can be adjusted in response to dispatch signals, making them key flexibility resources in community energy optimization management.

To provide a unified description of the operational characteristics and corresponding energy management actions for various controllable DER devices, this paper employs a state triplet ψ_t to characterize the device status at time t.

$$\psi_t = (\chi_t, \rho_t, \zeta_t) \tag{1}$$

Shiftable loads allow their operation time to be shifted within a specific interval but cannot perform load curtailment or power modulation. Typical examples of such loads include dehumidifiers, washing machines, and dishwashers.

Unlike simple constant-power devices, such appliances, once initiated, must continuously complete a preset operation cycle, during which the power consumption may vary across different steps. As illustrated in Figure 2, the user-defined allowable operation window for an SL is $[t_\alpha^{SL}, t_\beta^{SL}]$; the load cannot operate outside this interval. The total duration of the SL operation cycle is T^{SL}, consisting of M sequential operation steps $\tau = 1, \ldots, M$, each corresponding to a fixed power level in the sequence $\left\{P_1^{sl}, \ldots P_M^{sl}\right\}$. The state triplet ψ_t^{SL} for an SL at time t is defined as follows:

$$(\chi_t^{SL}, \rho_t^{SL}, \zeta_t^{SL}) = \begin{cases} (1, \rho_{t-1}^{SL} + \frac{u_t^{SL}}{M}, t_\beta^{SL} - t), & t \in [t_\alpha^{SL}, t_\beta^{SL}] \\ (0,0,0), & t \notin [t_\alpha^{SL}, t_\beta^{SL}] \end{cases} \tag{2}$$

where χ_t^{SL} is the availability indicator (equal to 1 within the allowable window, and 0 otherwise); $\rho_t^{SL} \in [0,1]$ represents the task completion rate (the ratio of completed steps to total steps); and ζ_t^{SL} denotes the remaining time within the operation window.

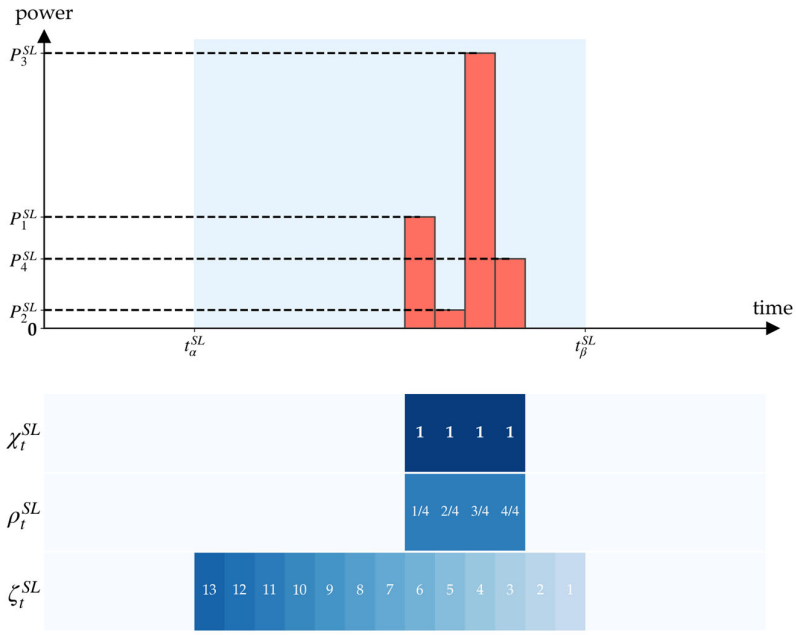


Figure 2. Operational timing diagram of shiftable loads.

Given that SL operation is uninterruptible, its control variable $u_t^{SL} \in \{0,1\}$ (where 1 indicates execution and 0 indicates stoppage) must satisfy the following continuity constraints:

$$u_t^{SL} = \begin{cases} 1, & if\ u_{t-1}^{SL} = 1 \text{ and } 0 < \rho_t^{SL} < 1 \\ 1, & if\ \rho_t^{SL} = 0\ and\ t = t_\beta^{SL} - M \\ 0, & if\ t \notin [t_\alpha^{SL}, t_\beta^{SL}] \end{cases} \tag{3}$$

The three constraints in Equation (3) enforce the following logic: (a) once a task begins, it must be executed continuously until completion; (b) the task must be completed before the deadline t_β^{SL}, arrives; and (c) the SL cannot operate outside the allowable time window $[t_\alpha^{SL}, t_\beta^{SL}]$. Given u_t^{SL}, let $m = \text{round}(M * \rho_t) + 1$ denote the current sequential operation step of the appliance (where $m \in \{1, 2 \cdots M\}$), and the load power of the SL can be described by Equation (4):

$$P_t^{SL} = \begin{cases} P_m^{SL}, & \text{if } u_t^{SL} = 1 \\ 0, & \text{if } u_t^{SL} = 0 \end{cases} \tag{4}$$

where P_m^{SL} is the predefined power consumption level for the m-th sequential step.

2.2. Modeling of Operational Characteristics of Adjustable Loads

Adjustable loads refer to flexible resources capable of continuously modulating their operating power within a specified range. This study primarily focuses on ES systems and EVs. Given the high similarity in their charge–discharge response characteristics, they are modeled under a unified framework.

Leveraging Vehicle-to-Grid (V2G) and Vehicle-to-Home (V2H) technologies, EVs possess bidirectional energy exchange capabilities and can be regarded as highly dispatchable mobile energy storage units. However, due to their primary function as transportation

tools, EVs' daily charge–discharge demands and grid connection/disconnection timing are highly dependent on user travel behavior. Consequently, they exhibit significant stochasticity and intermittency. Let $[t_\alpha^{EV}, t_\beta^{EV}]$ denote the effective grid-connection time window determined by the user's travel patterns. To characterize the grid-connected status of an EV at time t, a state triplet ψ_t^{EV} is defined as follows:

$$(\chi_t^{EV}, \rho_t^{EV}, \zeta_t^{EV}) = \begin{cases} (1, \varphi_t^{EV}, t), & t \in [t_\alpha^{EV}, t_\beta^{EV}] \\ (0,0,0), & t \notin [t_\alpha^{EV}, t_\beta^{EV}] \end{cases} \tag{5}$$

where χ_t^{EV} serves as the EV grid-connection indicator, which is 1 within the operable window and 0 otherwise; ρ_t^{EV} represents the task completion progress corresponding to the EV's State of Charge (SoC), defined as the ratio of current capacity to rated capacity; and ζ_t^{EV} is a temporal index used to anchor the specific grid-connection time t of the EV within the multi-agent system.

The net load power of the EV at time t is denoted by P_t^{EV}, where a positive value indicates charging and a negative value indicates discharging. Aside from the maximum charging and discharging power limits, the feasible range of P_t^{EV} is also constrained by the EV's current SoC and available battery capacity, as formulated in Equations (6) and (7):

$$-\overline{P}_{disc}^{EV} \le P_t^{EV} \le \overline{P}_{chr}^{EV} \tag{6}$$

$$P_t^{EV} = \begin{cases} \min(\overline{P}_{chr}^{EV}\chi_t^{EV}, \frac{(\overline{\omega}^{EV} - \omega_t^{EV})E^{EV}}{\eta_{chr}^{EV}\Delta t}), & P_t^{EV} \ge 0 \\ \min(-\overline{P}_{disc}^{EV}\chi_t^{EV}, \frac{(\omega_t^{EV} - \underline{\omega}^{EV})E^{EV}\eta_{disc}^{EV}}{\Delta t}), & P_t^{EV} < 0 \end{cases} \tag{7}$$

where $\overline{P}_{chr}^{EV}$ and $\overline{P}_{disc}^{EV}$ represent the maximum charging and discharging power of the EV, respectively; E^{EV} denotes the EV battery capacity; η_{chr}^{EV} and η_{disc}^{EV} are the charging and discharging efficiencies, respectively; $\overline{\omega}^{EV}$ and $\underline{\omega}^{EV}$ represent the maximum and minimum allowable SoC limits for the EV.

The SoC of the EV battery at time $t + 1$ can be described by Equation (8):

$$\omega_{t+1}^{EV} = \begin{cases} \omega_t^{EV} + \frac{\eta_{chr}^{EV} P_t^{EV} \Delta t}{E^{EV}}, & P_t^{EV} \ge 0 \\ \omega_t^{EV} + \frac{P_t^{EV}\Delta t}{\eta_{disc}^{EV} E^{EV}}, & P_t^{EV} < 0 \end{cases} \tag{8}$$

The fundamental distinction between ESs and EVs lies in the fact that the charging and discharging availability of ESs covers the entire day. Since their other characteristics align with those of EVs, their operational state can be modeled using the same modeling methodology described above.

By integrating the aforementioned load models and aggregating the power of non-controllable loads, energy storage, and distributed PV within each prosumer's system, the net load $P_{n,t}^{load}$ of prosumer n at time t is derived as Equation (9):

$$P_{n,t}^{load} = P_{n,t}^{NL} + P_{n,t}^{SL} + P_{n,t}^{EV} + P_{n,t}^{ES} - P_{n,t}^{PV} \tag{9}$$

where $P_{n,t}^{NL}$, $P_{n,t}^{ES}$ and $P_{n,t}^{PV}$ represent the power of non-controllable loads, the charging/discharging power of the stationary energy storage, and the PV generation power, respectively.

3. Community Energy Collaborative Optimization Management Framework Based on Multi-Agent Deep Reinforcement Learning

The proposed approach adopts the Multi-Agent Markov Decision Process (MAMDP) as the underlying framework for MADRL. An MAMDP can be formally defined as a tuple

$\langle N, \{S^n\}, \{A^n\}, P, R, \gamma \rangle$. Here, N denotes the number of agents. S^n represents the state set of agent n, where $s_{n,t} \in S^n$ signifies the state of agent n at time step t. The aggregation of the states of all agents constitutes the joint state space $\boldsymbol{S} = S^1 \times S^2 \times \ldots S^N$, with $s_t \in \boldsymbol{S}$. Similarly, A^n denotes the action set of agent n, where $a_{n,t} \in A^n$ represents the action selected by agent n at time step t. The joint action space is formed by the aggregation of all individual actions, denoted as $\boldsymbol{A} = A^1 \times A^2 \times \ldots A^N$, with $a_t \in \boldsymbol{A}$. $P : \boldsymbol{S} \times \boldsymbol{A} \times \boldsymbol{S} \to [0, 1]$ is the state transition probability function, representing the probability of transitioning to the next state s_{t+1} when the joint action a_t is executed in state s_t. R is the reward function, indicating the reward received by the agents from the environment after executing action a_t in state s_t, where $r_t = R(s_t, a_t)$. Additionally, $\gamma \in [0, 1)$ represents the reward discount factor. At each time step t, each agent executes an action $a_{n,t}$ based on its observation, which jointly acts upon the environment. The agents then receive reward feedback r_t from the environment, and the environmental state transitions to the next state s_{t+1}. Through this cyclical interaction with the environment, the agents utilize the generated data to optimize their policies—specifically, the mapping from states to actions—to ultimately achieve the maximization of cumulative returns.

3.1. State Space

The state space consists of the environmental information perceived by the agents. In this study, the observation space $o_{n,t}$ for agent n is configured to include the grid purchase price $\lambda^{grid,buy}_{t:(t+5)}$ and grid selling price $\lambda^{grid,sell}_{t:(t+5)}$ at half-hourly intervals from the current moment t for the upcoming 3 h, the PV generation power $P^{PV}_{n,t:(t+5)}$, the non-controllable load power $P^{NL}_{n,t:(t+5)}$, the state triplets of all DER devices owned by the user at time t, and the user's unique identity ID e_n, and the current time index t. The specific formulation is given in Equation (10). This specific 3 h look-ahead horizon serves as a pragmatic engineering trade-off: it is sufficiently long to encompass typical residential peak durations for effective load shifting, yet sufficiently constrained to prevent the LSTM from being overwhelmed by the com-pounding uncertainties of distant PV and load forecasts.

$$o_{n,t} = \left\{ \lambda^{grid,buy}_{t:(t+5)}, \lambda^{grid,sell}_{t:(t+5)}, P^{PV}_{n,t:(t+5)}, P^{NL}_{n,t:(t+5)}, \psi_{n,t}, e_n, t \right\} \tag{10}$$

3.2. Action Space

The action $a_{n,t}$ for agent n is defined as Equation (11). The continuous action variable $a^{EV}_{n,t}$ takes values in the range $[-1, 1]$, representing the ratio of the charging/discharging power of ES and EV to their respective maximum rated powers; a positive value indicates charging, while a negative value indicates discharging. The component $a^{SL}_{n,t} = [u^{SL}_{1,t}, \ldots, u^{SL}_{G,t}]$ represents the discrete control actions for G SLs. At the control policy level, this paper proposes a hybrid continuous-discrete policy function. For discrete actions, a Bernoulli distribution $B(p)$ is adopted, where p denotes the probability of executing an SL operation step. For continuous actions, a Gaussian distribution $N(\mu, \sigma^2)$ is utilized, where μ and σ^2 are used to characterize the mean and variance of the continuous action, respectively.

$$a_{n,t} = \left\{ a^{ES}_{n,t}, a^{EV}_{n,t}, a^{SL}_{n,t} \right\} \tag{11}$$

3.3. Reward Function

Given the geographic proximity of devices sharing a single distribution feeder, transmission losses and complex internal AC power flows are neglected. The community is modeled as a single unified node, subject only to power balance and transformer capacity constraints at the main grid interface. While this copper-plate abstraction is acceptable for algorithm-oriented energy scheduling research, it inherently limits the prac-

tical interpretability of the results regarding complex distribution network constraints (e.g., voltage deviations, reactive power flows, and internal line congestion). Furthermore, since individual economic interests are deeply coupled with the community's overall supply–demand state, we formulate the problem as a cooperative MAMDP. To align individual cost-minimization behaviors with the global objective of maximizing social welfare, the comprehensive reward function is designed with three components: the basic electricity cost r_t^{cost}, the EV travel satisfaction penalty r_t^{trip}, and the distribution network capacity violation penalty r_t^{cap}.

3.3.1. Basic Electricity Cost

To minimize electricity costs, prosumers tend to prioritize energy sharing with peers in the community energy market through negotiated pricing, rather than trading directly with the utility grid. Existing pricing mechanisms used in community markets include Bill Sharing (BS), Supply–Demand Ratio (SDR), and Mid-Market Rate (MMR) [26]. Under the BS mechanism, the local electricity price may drop below the grid retail price, which could undermine the economic interests of prosumers with abundant PV generation and reduce their willingness to participate in the market. In contrast, the local buying and selling prices determined by the MMR and SDR mechanisms lie between the grid retail price and the feed-in tariff. This structure is capable of encouraging prosumers to prioritize participation in local energy trading, ensuring reasonable returns for various types of prosumers while enhancing the efficiency of community energy sharing. Therefore, this paper selects the MMR mechanism as the pricing scheme for the community energy market.

Let the set of prosumers be $\mathcal{N} = \{1, \ldots N\}$. Let the community net generation power is denoted by $P_{com,t}^{gen}$ and the net load be $P_{com,t}^{load}$. The community net power by $P_{com,t}$. A positive value indicates that supply fails to meet demand, while a negative value indicates that supply exceeds demand, as shown in Equations (12) and (13):

$$P_{com,t} = P_{com,t}^{gen} + P_{com,t}^{load} \tag{12}$$

$$\begin{cases} P_{com,t}^{gen} = \sum\limits_{n \in \mathcal{N}^{gen}} P_{n,t}^{load} \\ P_{com,t}^{load} = \sum\limits_{n \in \mathcal{N}^{load}} P_{n,t}^{load} \end{cases} \tag{13}$$

where $\mathcal{N}^{gen} = \left\{ n \in \mathcal{N} : P_{n,t}^{load} \leq 0 \right\}$ represents the set of producers, and $\mathcal{N}^{load} = \left\{ n \in \mathcal{N} : P_{n,t}^{load} > 0 \right\}$ represents the set of consumers.

The MMR mechanism first establishes the average of the utility grid's retail purchase and selling prices as a baseline rate, denoted as λ_t^{bla} which is calculated as:

$$\lambda_t^{bla} = \frac{\lambda_t^{grid,buy} + \lambda_t^{grid,sell}}{2} \tag{14}$$

Based on the community's real-time net power state $P_{com,t}$, the mechanism dynamically implements differentiated price adjustments. The specific community internal buying price λ_t^{buy} and λ_t^{sell} are formulated in Equations (15) and (16):

$$\lambda_t^{buy} = \begin{cases} \frac{\lambda_t^{bla} |P_{com,t}^{gen}| + \lambda_t^{grid,buy} P_{com,t}}{P_{com,t}^{load}}, & P_{com,t} > 0 \\ \lambda_t^{bla}, & P_{com,t} \leq 0 \end{cases} \tag{15}$$

$$\lambda_t^{sell} = \begin{cases} \lambda_t^{bla}, & P_{com,t} \geq 0 \\ \frac{\lambda_t^{bla} P_{com,t}^{load} + \lambda_t^{grid,sell} |P_{com,t}|}{|P_{com,t}^{gen}|}, & P_{com,t} < 0 \end{cases} \tag{16}$$

When there is no power exchange between the community and the main grid (i.e., $P_{com,t} = 0$), both the buying and selling prices in the community market default to λ_t^{bla}. Conversely, when power exchange occurs (i.e., $P_{com,t} \neq 0$), any resulting surplus or deficit from external trading is reallocated among users via the trading platform. Specifically, in the event of a power shortage ($P_{com,t} > 0$), electricity consumption is settled at λ_t^{buy}, whereas in the event of a surplus ($P_{com,t} < 0$), generation revenue is settled at λ_t^{sell}. These additional costs or revenues are uniformly apportioned among prosumers, weighted by their real-time net load $P_{n,t}^{load}$. Consequently, the basic electricity cost $r_{n,t}^{cost}$ for prosumer n under the MMR pricing mechanism is defined as follows:

$$r_{n,t}^{cost} = -(\lambda_t^{buy} \max\{P_{n,t}^{load}, 0\} + \lambda_t^{sell} \min\{P_{n,t}^{load}, 0\}) \tag{17}$$

3.3.2. EV Travel Satisfaction

To ensure EV users have sufficient energy for their scheduled trips and avoid travel disruptions caused by insufficient battery charge, a penalty term reflecting violations of EV user travel satisfaction requirements is incorporated into the basic reward function, as expressed in Equation (18).

$$r_{n,t}^{trip} = \kappa_1 \max(0, E_{n,trip} - \omega_{n,t}^{EV} E_n^{EV}), \qquad if\ t = t_{trip} \tag{18}$$

where κ_1 is the penalty factor, t_{trip} denotes the scheduled departure time for prosumer n, and $E_{n,trip}$ represents the energy required for the trip. When the EV departs at time t_{trip}, if the stored energy in the EV battery $\omega_{n,t}^{EV} E_n^{EV}$ is less than the required energy $E_{n,trip}$, a penalty proportional to the energy deficit is imposed.

3.3.3. Distribution Network Capacity Penalty

Motivated by cost minimization, prosumers tend to cluster DER device regulation during periods of extreme electricity prices. This behavior tends to cause load rebound or generation peaks, which threatens the safe and stable operation of the distribution network. To mitigate such issues, a distribution network capacity threshold penalty is introduced, as defined in Equation (19).

$$r_{n,t}^{cap} = \begin{cases} -\kappa_2 \left(\dfrac{P_{n,t}^{load}}{\sum\limits_{\in \mathcal{N}^{load}} P_{n,t}^{load}} \right), & P_{com,t} > P^{cap} \\ -\kappa_2 \left(\dfrac{P_{n,t}^{load}}{\sum\limits_{\in \mathcal{N}^{gen}} P_{n,t}^{load}} \right), & P_{com,t} < -P^{cap} \end{cases} \tag{19}$$

where κ_2 serves as the penalty factor. When the absolute value of the community net power $P_{com,t}$ exceeds the grid capacity threshold P^{cap}, all prosumers incur a threshold violation penalty $r_{n,t}^{cap}$. And these penalty factors are deliberately scaled to be significantly higher than standard retail electricity prices. This ensures that the reinforcement learning agent mathematically prioritizes strict adherence to physical safety limits and operational reliability over marginal economic gains, effectively acting as soft-barriers for hard physical constraints.

In summary, the total reward for prosumer n at time step t is expressed as:

$$r_{n,t} = r_{n,t}^{cost} + r_{n,t}^{cap} + r_{n,t}^{trip} \tag{20}$$

And the global reward of the community at time slot t is given by:

$$r_t^{global} = \sum_{n=1}^{n=N} r_{n,t} \tag{21}$$

3.4. Deep Reinforcement Learning Framework Based on Spatiotemporal Attention Mechanism

To address the challenges of partial observability and the privacy protection requirements among multi-agents in community energy management, this paper adopts a CTDE architecture and proposes a MATPPO algorithm that integrates LSTM and Transformer. The network framework of this algorithm is illustrated in Figure 3.

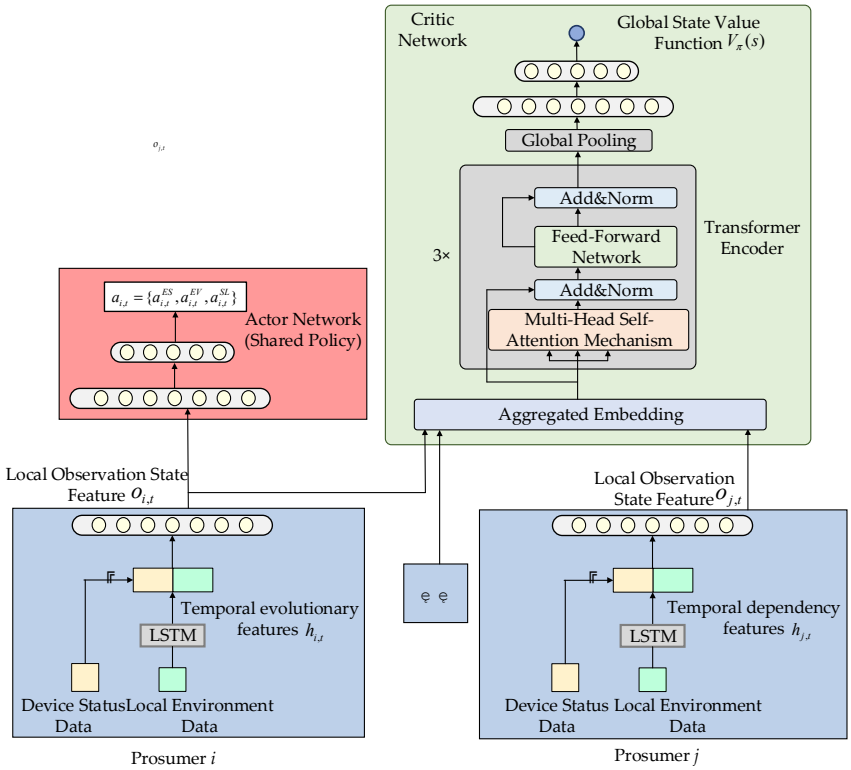


Figure 3. Network Architecture of the MATPPO Algorithm Integrating LSTM and Transformer.

During the distributed execution phase, to accommodate the strong stochasticity of source-load variations and preserve data privacy, each agent relies solely on local observations for independent decision-making. Given that PV output, load demand, and dynamic prices exhibit strong sequential dependencies over time, simple fully connected networks cannot effectively capture their upcoming evolutionary trends. Consequently, this study integrates an LSTM network into the Actor as a state encoder to process the observational sequences over the upcoming three-hour forecast window and extract long-term temporal features $h_{n,t}$. Compared to relying directly on discrete point prediction values, the high-dimensional hidden states extracted by the LSTM preserve the probability distribution and temporal trend characteristics of the upcoming source-load fluctuations. This feature representation empowers agents with a forward-looking perspective, enabling them to make robust scheduling decisions even when faced with source-load uncertainties. Furthermore, considering the homogeneity of prosumer devices within the community and to avoid network parameter explosion as the number of agents increases, a parameter-sharing mechanism is adopted in the Actor network. Specifically, all agents share the same set of Actor network weights for policy learning, while performing decentralized inference based on their respective local observations. Finally, the temporal dependency features $h_{n,t}$

are concatenated with the device internal states represented by the triplet $\psi_{n,t}$, and undergo linear mapping to construct the final local observation state feature $o_{n,t}$ for agent n. Based on $o_{n,t}$, the shared Actor network outputs a hybrid action distribution.

In the centralized training phase, to tackle the dual shortcomings of traditional multi-agent algorithms—namely, the input dimensionality explosion of the value network caused by the crude concatenation of global states, and the inability of mean-field methods to capture complex game relationships—this paper introduces the Transformer encoder into the centralized value network (Critic) to establish a spatial attention mechanism. Under the CTDE architecture, the single Critic network is deployed on the aforementioned Cloud Trading Platform during the training phase, and its input is reconstructed as a feature sequence $O_t = [o_{1,t}, \ldots o_{N,t}, S_{global,t}]$. $S_{global,t}$ denotes the global aggregation feature that characterizes the collective trading behaviors of the community. In the community energy system, the interaction relationships among prosumers change dynamically in real-time according to supply and demand states. For instance, when prosumer i is in a power deficit state, it will prioritize attention toward neighbor j who is in a surplus state. The Transformer utilizes a Multi-Head Self-Attention mechanism to accurately model this dynamic coupling at both physical and economic levels.

Specifically, for each agent i, the network maps it to a Query vector (Q), Key vector (K), and Value vector (V) using learnable weight matrices W^Q, W^K and W^V:

$$Q_i = o_{i,t}W^Q, K_i = o_{i,t}W^K, V_i = o_{i,t}W^V \tag{22}$$

Subsequently, the attention coefficient $\xi_{i,j}$ between agent i and other agents j is calculated, which characterizes the importance and coupling strength of agent j to the decision-making of agent i, as shown below:

$$\xi_{i,j} = \text{softmax}(\frac{Q_i K_j^T}{\sqrt{d_k}}) \tag{23}$$

where K_j^T is the transpose of the key vector for agent j, and d_k represents the feature dimension of the key vector. The denominator $\sqrt{d_k}$ acts as a scaling factor to normalize the dot product. The softmax function is applied over all neighbor agents j (i.e., along the sequence dimension).

By leveraging this attention mechanism, the Critic network automatically filters redundant information and adaptively reconstructs the current critical interaction topology from the global joint state embedding. Consequently, it generates a feature vector $z_{i,t}$ which integrates global collaborative information, as calculated in Equation (24). This feature vector not only alleviates the limitations of a single agent's restricted field of view but also provides a value benchmark with a global perspective for computing the advantage function in the PPO algorithm.

$$z_{i,t} = \sum_{j=1}^{N} \xi_{i,j} V_j \tag{24}$$

Given that community energy management is intrinsically a control problem within a continuous action space, the PPO algorithm based on the Actor–Critic framework is well suited to effectively address challenges associated with high-dimensional state spaces and stochasticity. Thus, this paper adopts the PPO algorithm for policy optimization. The specific execution and parameter update process of the MATPPO algorithm is illustrated in Figure 4. The architecture adopts a parameter-separated design. For the Actor network, each agent utilizes only the local observation state feature $o_{n,t}$ as input, constructing an independent decision-making structure for the hybrid action space. This design ensures

that agents require no communication with others during the execution phase, relying solely on local information to make decisions, preserving operational privacy during the execution phase and ensuring response speed. In contrast, the Critic network fully leverages the global information available at the training center, takes the sequence of context vectors $[z_{1,t}, z_{2,t} \cdots z_{n,t}]$ output by the Transformer, aggregates them via a global pooling layer, and utilizes the resulting feature to fit the global state value function. The theoretical definition of the state value function is presented in Equation (25). This function characterizes the expected cumulative return of an agent executing the hybrid policy under the global state. The estimated value output by the value network is used not only to calculate the advantage function for assessing the quality of current actions but also to drive the joint update of the Transformer attention weights and Critic fully connected layer parameters by minimizing prediction error. We denote the parameter set of the entire network as $\theta = (\theta_a, \theta_c)$, where θ_a represents the shared Actor parameters, and θ_c includes the Transformer and Critic parameters.

$$V_\pi(s) = E_\pi(\sum_{k=0}^{\infty} \gamma^k r_{t+k}^{global} | s_t = s) \tag{25}$$

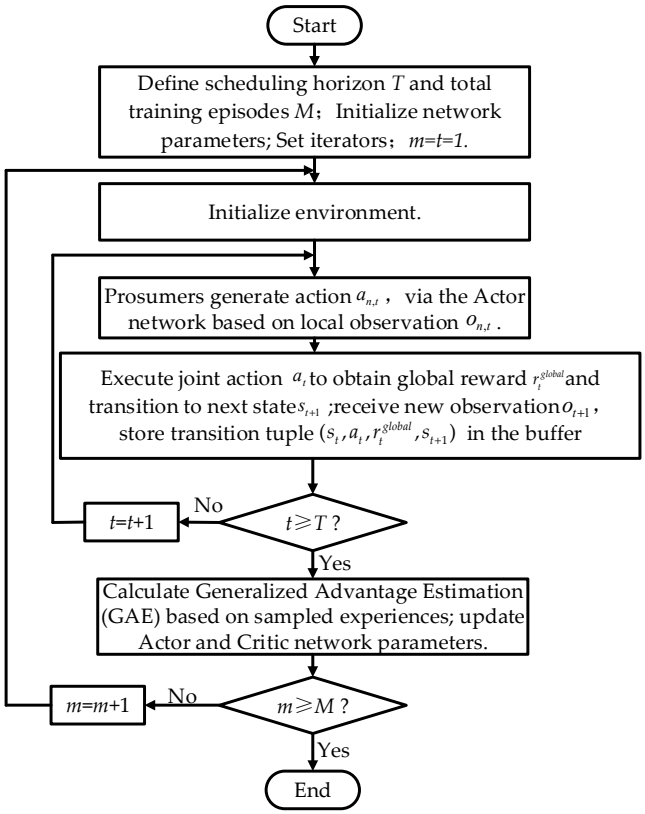


Figure 4. Flowchart of MATPPO Algorithm Training and Execution.

During the interaction phase, agents execute decentralized actions based on the current shared policy. The generated transition data—comprising local observations, individual actions, global states and global reward is stored in a rollout buffer. Once a sufficient batch of trajectories is collected, the algorithm utilizes this data to update the parameters of both the Actor and the Transformer-based Critic through the following two key steps:

(1) Generalized Advantage Estimation

To effectively evaluate the quality of the current action a_t while balancing bias and variance in policy gradient estimation, we employ the Generalized Advantage Estimation

(GAE) technique. Using the state value $V(s)$ output by the Critic network, the GAE advantage function $\widehat{\mathrm{A}}_t(s)$ at time step t is computed as follows:

$$\delta_t = r_t^{global} + \gamma V(s_{t+1}) - V(s_t) \tag{26}$$

$$\widehat{\mathrm{A}}_t = \sum_{k=0}^{T-t-1} (\vartheta\gamma)^k \delta_{t+k} \tag{27}$$

where r_t^{global} is the community global immediate reward; δ_t represents the Temporal-Difference (TD) Error, reflecting the deviation between the immediate reward and the expected state value; T is the time horizon of the sampled trajectory; and ϑ is the GAE smoothing factor used to regulate the trade-off between variance and bias.

(2) Optimization Targets and Parameter Updates

To implement the CTDE architecture and prevent training collapse caused by excessive policy update step sizes, we construct independent optimization targets for the Actor and Critic, respectively. For the Actor network, we define a policy objective function $J^{actor}(\theta_a)$ that incorporates the PPO clipping mechanism and an entropy regularization term. For the Critic network, we define a value loss function $L^{critic}(\theta_c)$ based on the Mean Squared Error (MSE), as detailed below:

$$J^{actor}(\theta_a) = \frac{1}{N}\sum_{n=1}^{N} \widehat{\mathbb{E}}_t\left[J_{n,t}^{clip}(\theta_a) + \upsilon S[\pi_{\theta_a}](o_{n,t})\right] \tag{28}$$

$$L^{critic}(\theta_c) = \widehat{\mathbb{E}}_t\left[(V_{\theta_c}(S_t) - V_{target})^2\right] \tag{29}$$

where $o_{n,t}$ serves as the local observation input to the decentralized Actor for agent n; while S_t represents the global joint state fed into the centralized Critic. To encourage exploration and prevent premature convergence, the policy entropy $S[\pi_\theta]$ is introduced, scaled by the regularization coefficient υ. Furthermore, V_{target} denotes the target state value. Equation (30) formulates the clipped surrogate objective $J_{n,t}^{clip}(\theta_a)$. By bounding the probability ratio $l_{n,t}(\theta_a)$ (defined in Equation (30)) between the current and previous policies, this clipping mechanism effectively prevents destructively large policy updates, thereby guaranteeing monotonic improvement and training stability.

$$J_{n,t}^{clip} = \min(l_{n,t}(\theta_a)\widehat{\mathrm{A}}_t, clip(l_{n,t}(\theta_a), 1-\varepsilon, 1+\varepsilon)\widehat{\mathrm{A}}_t) \tag{30}$$

$$l_{n,t}(\theta_a) = \frac{\pi_\theta(a_{n,t}|o_{n,t})}{\pi_{\theta,old}(a_{n,t}|o_{n,t})} \tag{31}$$

Through the above parameter sharing and joint gradient ascent mechanism, the shared Actor network can absorb and generalize the exploration experiences of N agents under different local states simultaneously in one parameter backpropagation. This effectively avoids the curse of dimensionality and parameter explosion caused by the expansion of node scale, and greatly improves the convergence efficiency of the algorithm in large-scale community scenarios.

4. Results and Discussion

4.1. Parameter Settings

This study utilizes real-world data provided by the Australian distribution network operator Ausgrid [27] as the simulation scenario to validate the effectiveness of the proposed MATPPO algorithm. The dataset includes data on prosumers' uncontrollable loads and PV generation from 1 July 2011, to 30 June 2012, with a data collection cycle and control interval of 30 min. The test set comprises 52 days, selected by randomly choosing one day

from each week of the year, while the remaining data constitutes the training set. The grid operator in this study employs a Time-of-Use pricing scheme, defined as follows: the peak period is from 14:00 to 20:00; the flat periods are from 07:00 to 14:00 and 20:00 to 22:00; and the remaining hours constitute the valley period. Specific pricing details are provided in Appendix A Table A1.

Each prosumer is equipped with one EV, one ES unit, and three types of SLs: a dehumidifier, a washing machine, and a dishwasher. Furthermore, to capture the diversity in DER usage among different prosumers, a truncated normal distribution is employed to simulate the daily arrival and departure times of EVs, as well as the earliest start and latest completion times of SLs for each prosumer. While the device types are uniform across prosumers in this base case, this continuous stochastic sampling of operational windows, combined with the inherently highly volatile real-world Ausgrid data, ensures that each agent faces a distinctly dynamic local environment during the training and execution phases. Specific parameters for these distributions are listed in Appendix A Tables A2–A4. And the training parameters are provided in Appendix B. Additionally, the grid capacity threshold for the low-voltage transformer connecting the community is set at 600 kW.

4.2. Model Training Convergence Analysis

Figure 5 illustrates the training reward curve for 100 prosumers, evolving through three primary phases. In the initial phase (0–1000 episodes), agents rapidly learn basic strategies, causing a swift reward increase. Notably, an intermediate plateau occurs between 1000 and 2000 episodes, accompanied by large downward variance in raw samples. This stagnation reflects a bottleneck in the exploration-exploitation trade-off: agents test aggressive cost-minimization strategies that frequently violate distribution capacity and user satisfaction constraints. Consequently, heavy penalty terms temporarily dominate the reward. As network parameters further optimize, agents learn to navigate these complex constraint boundaries. The curve subsequently breaks out of the plateau, converging to a high-level steady state after 4000 episodes. This demonstrates MATPPO's robustness in balancing economic benefits and system security constraints.



Figure 5. Global Reward during MATPPO Training Process.

4.3. Analysis of Community Energy Trading in a 100-Prosumer Scenario

To further substantiate the effectiveness and superiority of the proposed method, the following section presents a comparative analysis based on the same 100-prosumer experimental scenario. The proposed method is compared against the baseline MAPPO algorithm and an improved MAPPO algorithm integrated with an LSTM neural network.

To mitigate the impact of stochasticity on experimental results, this study generated 10 independent random seeds for the DRL-based methods (MAPPO, MAPPO + LSTM, and MATPPO). Each learning algorithm completed 10,000 training episodes based on each random seed, where each training round corresponds to one random day from the training set, containing 48 time steps. During the training process, the performance of each algorithm was quantitatively evaluated using the test set every 100 episodes. In Figure 6, the solid lines and shaded areas represent the mean and standard deviation, respectively, of the community's daily electricity costs calculated from the test data under different random seeds. The final convergence values for these learning algorithms, alongside the static evaluation result of the deterministic, non-learning Rule-Based Control (RBC) baseline evaluated on the exact same test set, are presented in Table 1.

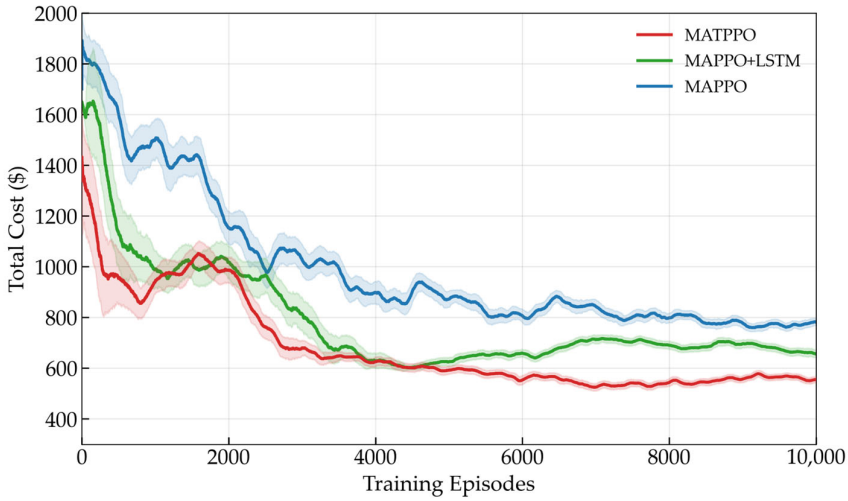


Figure 6. Average Daily Community Electricity Costs under Different Methods.

Table 1. Mean and Standard Deviation of Average Daily Community Electricity Costs at Convergence under Different Methods.

Method	Cost/$
MAPPO	732.78 ± 18.46
MAPPO + LSTM	694.91 ± 14.85
MATPPO	564.60 ± 8.61
Rule-Based Control (RBC)	1127.94

The training results indicate that the operating costs associated with the MAPPO method exhibit substantial fluctuations, reflecting suboptimal control performance. Although the improved MAPPO algorithm incorporating LSTM enhances adaptability to source-load uncertainty to a certain extent, resulting in reduced operating costs, the margin for optimization remains limited. In contrast, the proposed MATPPO method demonstrates significant advantages in terms of cost convergence value, convergence speed, and stability. Its average daily operating cost is reduced by 22.88% and 18.75% compared to the aforementioned methods, respectively.

To illustrate the efficacy of the proposed method in promoting local renewable energy consumption, balancing local supply and demand, and explicitly avoiding catastrophic synchronized load surges, Figure 7 compares the variations in community net power curves under the three learning-based methods. Furthermore, Figure 8 depicts the detailed aggregated power profiles of all controllable DER devices within the community under the four evaluated methods. To provide a more comprehensive and explicit quantitative

evaluation beyond the average daily cost, we further analyzed the multi-dimensional performance metrics of the evaluated algorithms, as detailed in Table 2.

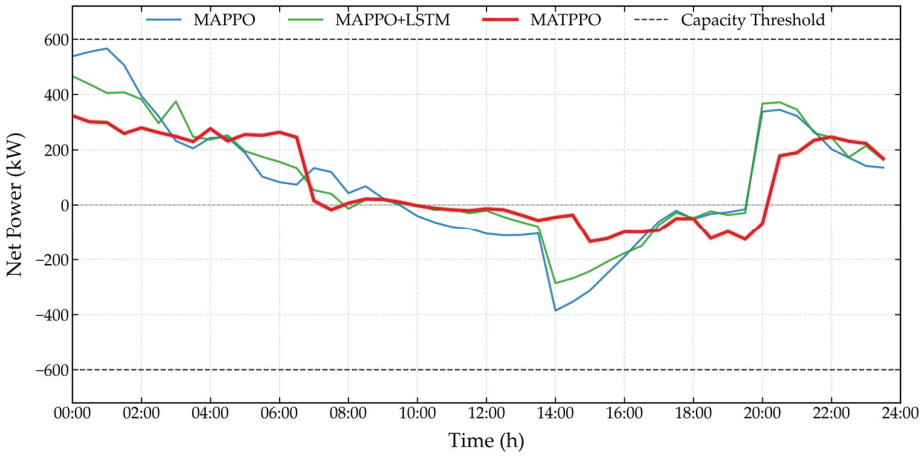


Figure 7. Community Net Power Curves under Different Methods.

Table 2. Quantitative Performance Metrics of Evaluated Algorithms.

Metric\Methods	RBC (Greedy)	MAPPO	MAPPO + LSTM	MATPPO
Total Daily Cost ($)	1127.94	732.78	694.91	564.60
Max Peak Load (kW/% of Limit)	1037.8 (173.0%)	564.3 (94.05%)	442.7 (73.8%)	320.1 (53.3%)
Capacity Violations (Steps/Day)	9	0	0	0
PV Self-Consumption Rate (%)	0	73.2%	81.3%	90.5%
Average Inference Delay (ms/step)	<0.01 ms	1.8 ms	2.6 ms	3.7 ms

As can be seen from Figures 7 and 8, the system operation strategies generally exhibit distinct time-of-use price response characteristics, but differ significantly in device coordinated scheduling strategies. Among the baseline methods, the deterministic RBC approach demonstrates the most extreme and myopic behavior. As clearly illustrated in Figure 8a, driven rigidly by price rules and lacking spatiotemporal coordination capabilities, all 100 agents initiate maximum charging simultaneously at the very beginning of the valley period, triggering a catastrophic synchronized load surge that severely disrupts normal grid operation (with a power peak of 1037.8 kW and 9 violation steps). Furthermore, RBC blindly charges ES units to full capacity at night, leaving almost no storage margin to absorb surplus PV generation at noon. As a result, all PV power is sold to the grid at a low price, yielding an extremely poor PV self-consumption rate of 0%. This fundamental lack of spatial coordination directly leads to a sharp rise in its total daily cost to $1127.94.

While the MADRL-based MAPPO method avoids the rigid extremes of RBC, it still exhibits severe power fluctuations. Constrained by its lack of long-term planning capabilities, Figure 8b reveals that a large number of EVs engage in disordered, clustered charging at the initial moment. Although the MAPPO + LSTM method introduces temporal memory to mitigate some oscillations, its effectiveness in suppressing extreme values remains limited.

In contrast, MATPPO demonstrates superior peak-shaving and valley-filling capabilities. As evidenced in Figure 8d and Table 2, this success is primarily attributed to the global attention mechanism of the Transformer, which successfully caps the peak at 320.1 kW and strictly reduces capacity violations to zero. On the other hand, it precisely plans energy storage actions to efficiently absorb excess PV output at noon, boosting the self-consumption rate to 90.5%. This explicitly proves its practical necessity and superiority in highly penetrated community grids.

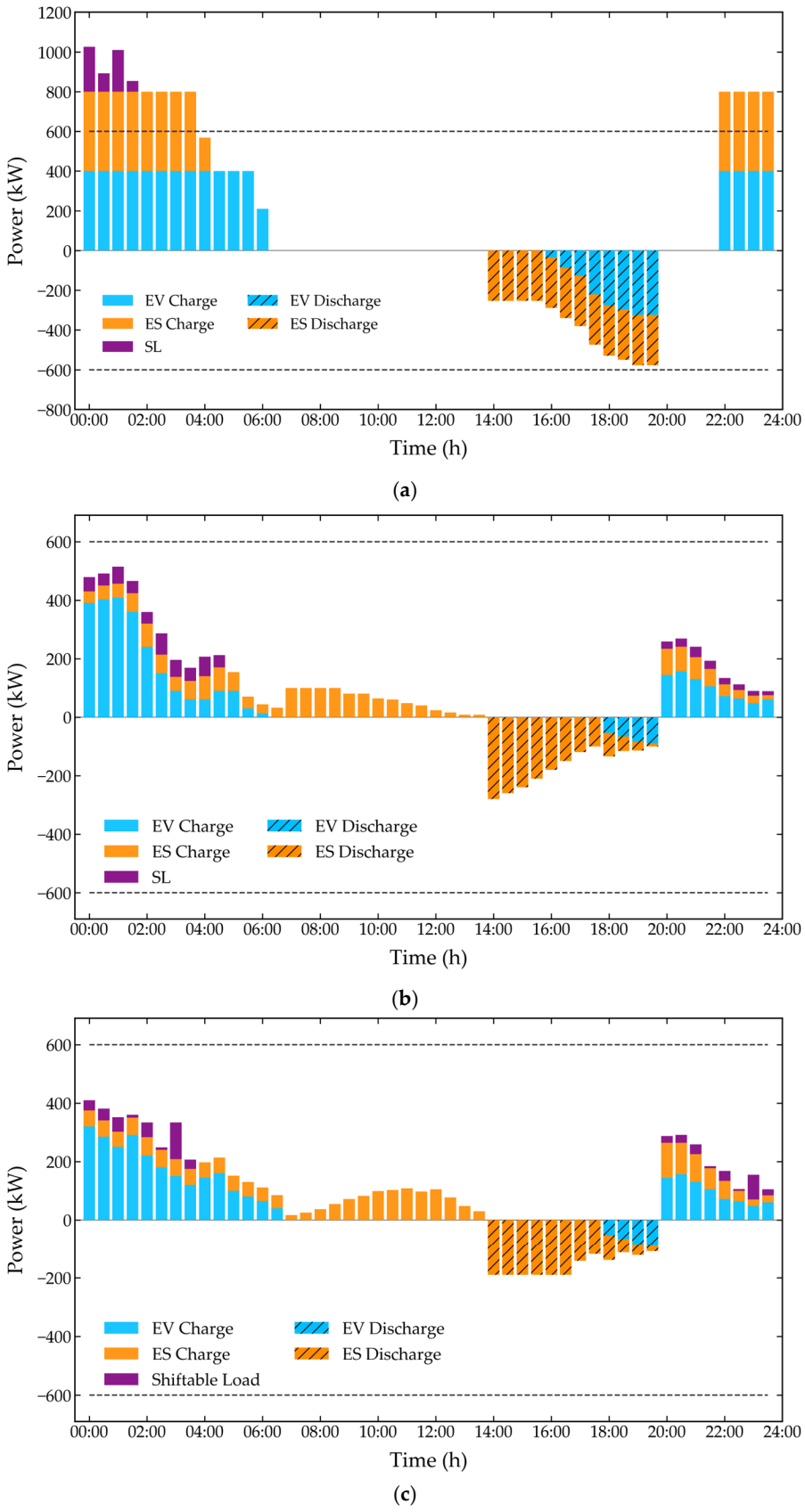
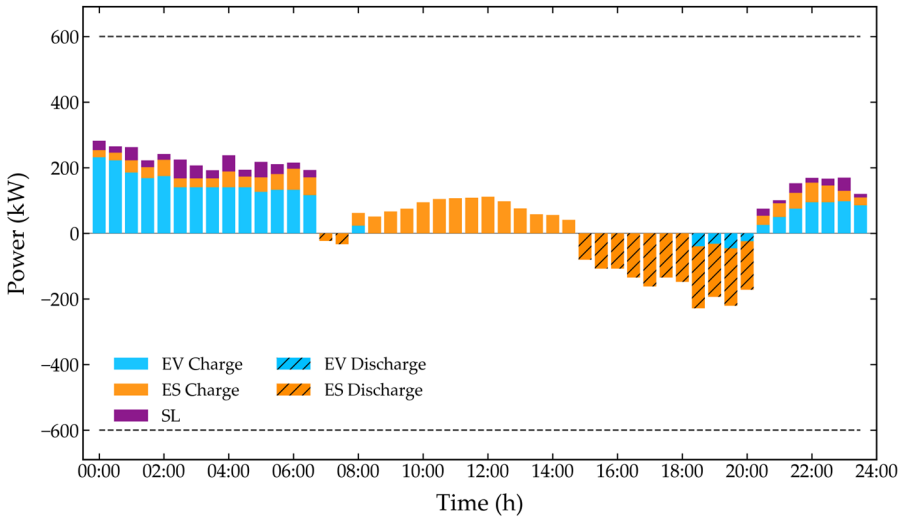


Figure 8. *Cont.*



(**d**)

Figure 8. Aggregated Power of Community DERs under Different Methods. (**a**) Aggregated DER Power under Rule-Based Control. (**b**) Aggregated DER Power under MAPPO Algorithm. (**c**) Aggregated DER Power under MAPPO + LSTM Algorithm. (**d**) Aggregated DER Power under MATPPO Algorithm.

4.4. Policy Rationality Under Alternative Market Rules

To explicitly address whether the learned behavior of the proposed MATPPO algorithm remains desirable and rational under alternative tariff structures, we conducted an additional decoupling experiment. In this scenario, the physical community energy trading framework was retained, but the dynamically adjusted MMR pricing mechanism was replaced with a rigid standard tariff. Specifically, all internal peer-to-peer energy exchanges were uniformly settled using the external grid's Time-of-Use (TOU) retail prices for purchasing, combined with a fixed, low-value Feed-in Tariff (FIT) for surplus generation.

As illustrated in Figure 9, stripped of the cooperative financial margins provided by the MMR, the agents predictably default to purely self-interested individual optimization. Driven solely by the TOU retail price, agents aggressively shift their EV and ES charging strictly to the early valley periods (00:00–02:00), causing the aggregated load to surge to approximately 510 kW. While this greedy behavior is highly rational from an individual cost-minimization perspective—and impressively, the MATPPO algorithm still implicitly coordinates to avoid triggering the 600 kW capacity threshold—it inherently results in a more volatile grid profile. Crucially, the absence of MMR incentives severely diminishes cooperative energy sharing. Because selling surplus energy to peers now yields the same unfavorable FIT rate as exporting to the grid, prosumers exhibit no motivation to stagger their discharges. Instead, during the evening peak price period (14:00–20:00), all agents simultaneously dump their stored energy to strictly offset their own high-cost consumption. As shown in Figure 9, this uncoordinated, self-serving discharging creates a massive reverse power surge (approaching −400 kW), sharply contrasting with the smooth, localized energy circulation achieved under the MMR mechanism (Figure 8d).

Economically, this decoupling experiment allows us to explicitly separate the financial benefits of the control algorithm from those of the pricing mechanism: under the standard TOU structure without any MMR trading bonuses, the MATPPO algorithm yields a daily operating cost of $647.69, which remains lower than the $732.78 incurred by the baseline MAPPO algorithm. This substantial reduction clearly proves that the physical load-shifting and spatiotemporal coordination capabilities of MATPPO are inherently superior, independently generating significant savings without relying on specific market rules. Furthermore, when the proposed MMR pricing mechanism is activated, the daily cost of

MATPPO further decreases from $647.69 to $564.60. This result quantifies the additional economic value brought by community energy trading incentives, and confirms that the proposed framework exhibits desirable and rational scheduling behaviors under other market mechanisms.

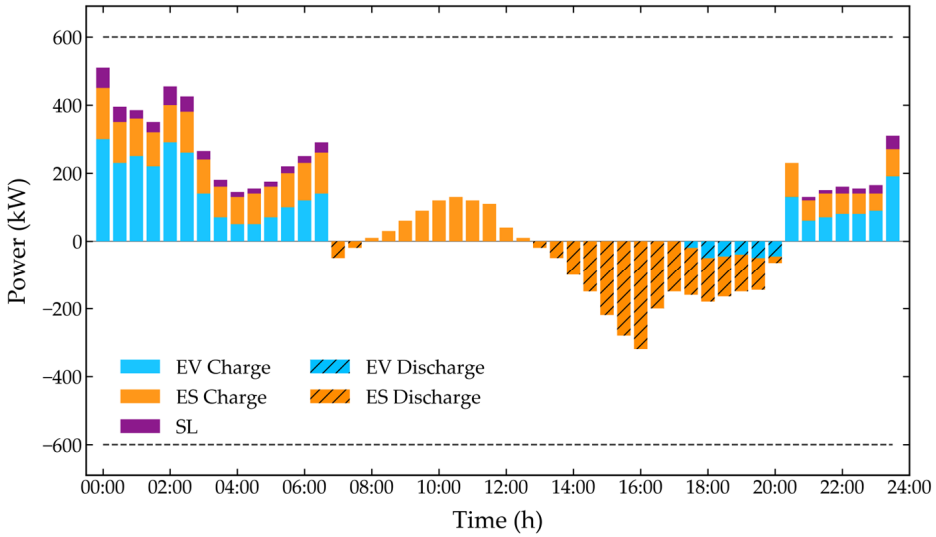


Figure 9. The Aggregated Power Profile Without MMR.

4.5. Scalability Analysis Across Scenarios with Varying Prosumer Populations

This section evaluates the scalability of the algorithms by establishing three experimental scenarios with varying scales of prosumers: 10, 30, and 100. The results are presented in Table 3.

Table 3. Scalability Analysis of Different Algorithms.

Comparison Item	Prosumer Number/N	MAPPO	MAPPO + LSTM	MATPPO
Training Time/h	10	0.62	0.74	0.85
	30	1.94	2.40	2.92
	100	6.52	8.53	10.25
Daily Cost/$	10	66.65	72.49	56.30
	30	219.89	209.42	163.46
	100	732.78	694.91	564.60

Table 3 demonstrates the scalability performance of the algorithms across prosumer scenarios of varying scales. From a formal complexity perspective, the centralized Critic in MATPPO utilizes a standard Transformer self-attention mechanism to aggregate global states. This operation inherently incurs a quadratic computational complexity of $\mathcal{O}(N^2 \cdot d)$ with respect to the number of agents N (where d is the embedding dimension), Consequently, as N scales from 10 to 100, the offline training time of MATPPO increases noticeably, reaching 10.25 h. While the framework remains manageable and yields superior cost reductions at the 100-prosumer scale, extrapolating this theoretical trend reveals that standard self-attention will inevitably become a computational bottleneck in ultra-large-scale systems. Overcoming this offline scaling limitation by introducing sparse attention mechanisms which reduce complexity to $\mathcal{O}(N \log N)$ forms a highly promising direction for our future research.

However, regarding real-time engineering applicability, the proposed CTDE architecture strictly separates execution from training. During the execution phase, each agent independently infers its action based solely on local observations. Crucially, this operation

is mathematically independent of the total community size N, meaning the online execution complexity scales at $\mathcal{O}(1)$ with respect to N. Therefore, despite the quadratically increasing offline training burden, the online inference delay strictly remains at the millisecond level. This provides robust quantitative evidence that, once trained, MATPPO possesses exceptional real-time applicability for practical engineering deployments.

Regarding optimization effectiveness, MATPPO exhibits remarkable robustness against scale expansion. In small-scale scenarios, the cost discrepancies among the three algorithms are relatively minor. However, as the node scale expands to 100, the traditional MAPPO algorithm struggles to handle complex games within high-dimensional state spaces, leading to a significant surge in costs. In contrast, MATPPO, leveraging its global attention mechanism, maintains exceptionally high cooperative efficiency in large-scale scenarios, saving approximately 23% in operating costs compared to MAPPO. This substantiates the robustness of the MATPPO architecture as a foundational baseline scheduling solver for community energy management under the specific tested scenarios, although universal deployability to ultra-large grids requires further validation.

4.6. Discussion on Current Limitations and Future Directions

Despite the demonstrated advantages in economic optimization and safe peak-shaving, this study has several theoretical and practical limitations that warrant explicit discussion.

First, the physical modeling of the distribution system is highly simplified. Treating the community as a unified node neglects complex AC power flow constraints, such as voltage behaviors and reactive power, which limits its immediate applicability to real-world congested networks. Second, while execution is decentralized, centralized training still relies on historical data aggregation, posing theoretical limitations on full-cycle privacy guarantees. Third, the quadratic computational complexity of standard self-attention may lead to training bottlenecks in ultra-large-scale systems. Fourth, the experimental validation relies on a specific historical dataset and an idealized uniform device composition. As such, the current study may not fully capture the complex dynamics of present-day community energy systems characterized by extreme DER penetration and shifting prosumer behaviors. Therefore, the current results should be strictly viewed as a foundational algorithmic benchmark rather than a universally deployable solution. Fifth, while evaluated against conventional rule-based heuristics, the current study lacks a comprehensive benchmarking against classical mathematical optimization frameworks and other distinct state-of-the-art MADRL architectures.

5. Conclusions

To address the challenges of intense source-load uncertainty and the propensity for load peaks triggered by multi-agent gaming in communities with high penetration of DERs, this paper establishes a coordinated community energy trading framework based on the MMR pricing mechanism. Furthermore, a MATPPO algorithm integrating an LSTM–Transformer architecture and parameter sharing is proposed. Based on the case studies, the main conclusions are drawn as follows:

(1) By leveraging spatiotemporal perception capabilities, the MATPPO algorithm not only realizes agent adaptability to dynamic source-load variations but also achieves implicit, ordered coordination among multiple agents without explicit communication. This effectively suppresses the phenomenon of synchronized load rebound induced by price signals.

(2) Compared to traditional MRL algorithms, the proposed method demonstrates superior performance in balancing economic efficiency and security. It significantly reduces the daily operating costs of the community while effectively mitigating the risk of exceeding transformer capacity limits at the point of common coupling.
(3) Benefiting from the CTDE architecture, each prosumer makes independent decisions based solely on local observations, eliminating the need to upload internal load details or device parameters. This decentralized execution mechanism structurally avoids online state sharing, thereby enhancing privacy protection during the operational phase, although historical data aggregation is still required during offline centralized training.

Future work will incorporate distribution network physical constraints by integrating advanced electrical model-free voltage calculation methods [28] to manage nodal AC behaviors without relying on complex power flow equations. Furthermore, we will explore federated learning to ensure offline privacy, investigate sparse attention mechanisms to overcome scaling limitations, test the framework across more heterogeneous modern datasets, and systematically supplement comparative experiments with classical optimization solvers and emerging MADRL algorithms.

Author Contributions: Conceptualization, S.C.; methodology, S.C.; software, S.C.; formal analysis, S.C.; investigation, Y.Y.; writing—original draft preparation, S.C.; writing—review and editing, J.H.; supervision, Y.Y.; Data curation, C.F.; project administration, C.F. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the Science and Technology Projects of State Grid Zhejiang Electric Power Co., Ltd. Ningbo Power Supply Company, which is a subsidiary of State Grid Zhejiang Electric Power Co., Ltd. (5211NB240004).

Data Availability Statement: The raw Ausgrid dataset used in this study is publicly available. To ensure full reproducibility and adhere to open science standards, the core algorithmic implementation code of the proposed MATPPO architecture has been made publicly available and can be accessed at: [https://github.com/csleonsdsd/train.py (accessed on 20 March 2026)].

Conflicts of Interest: Author Yong Yan was employed by the company State Grid Zhejiang Electric Power Research Institute Co., Ltd. Author Jiahua Hu was employed by the company State Grid Zhejiang Electric Power Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The authors declare that this study received funding from State Grid Zhejiang Electric Power Co., Ltd. and Ningbo Power Supply Company. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

Appendix A

Table A1. Retail Electricity Price Schedule.

Transaction Mode	Time Period	Price ($/kWh)
Purchase	Peak (14:00–20:00)	0.40
	Flat (07:00–14:00 and 20:00–22:00)	0.20
	Valley (22:00–07:00 next day)	0.12
Sale	All Periods	0.04

Table A2. EV Operational Parameters.

Parameter	Value
E^{EV}/(kWh)	30
E_0^{EV}/(kWh)	6
E_{trip}/(kWh)	24
$\overline{\omega}^{EV}$,$\underline{\omega}^{EV}$	1, 0.2
$\overline{P}_{chr}^{EV}$,$\overline{P}_{disc}^{EV}$/kW	4, 4
η_{chr}^{EV},η_{disc}^{EV}	0.95
t_{dep}	[6, 9]
t_{arr}	[17, 19]

Table A3. ES Operational Parameters.

Parameter	Value
E^{ES}/(kWh)	20
E_0^{ES}/(kWh)	4
$\overline{\omega}^{ES}$,$\underline{\omega}^{ES}$	1, 0.2
$\overline{P}_{chr}^{ES}$,$\overline{P}_{disc}^{ES}$/kW	4, 4
η_{chr}^{ES},η_{disc}^{ES}	0.95

Table A4. SL Operational Parameters.

SL Type	SL Stage Power (kW)				SL Operation Window
Dehumidifier	0.85	0.14	2.1	0.53	[0, 12] ∪ [40, 47]
Dishwasher	0.58	0.61	—	—	
Washing Machine	0.83	0.17	—	—	

Appendix B

Table A5. Core Algorithm Hyperparameters.

Category	Parameter	Value
Network Dimensions	LSTM Extracted Feature Dimension	96
	Fused State Embedding Dimension	128
	Actor MLP Architecture	2 layers, 128 units each
	Transformer Encoder Layers	3
	Transformer Attention Heads	4
PPO Training Protocol	Mini-batch Size	256
	PPO Clipping Coefficient	0.2
	Discount Factor	0.99
	GAE Smoothing Factor	0.95
	Entropy Coefficient	0.05
Optimization and Process	Actor Learning Rate	1×10^{-4}
	Critic Learning Rate	3×10^{-4}
	Optimizer	Adam

References

1. Qiu, D.; Ye, Y.; Papadaskalopoulos, D. Exploring the effects of local energy markets on electricity retailers and customers. *Electr. Power Syst. Res.* **2020**, *189*, 106761. [CrossRef]
2. Li, Z.; Hilber, P.; Li, Z.; Laneryd, T.; Ivanell, S. Temporally Coordinated Operation of Green Multi-Energy Airport Microgrids with Climatic Correlations and Flexible Loads via Decomposed Stochastic Programming. *IEEE Trans. Sustain. Energy* **2025**, *17*, 1909–1922. [CrossRef]
3. Guerrero, J.; Chapman, A.C.; Verbic, G. Decentralized P2P energy trading under network constraints in a low-voltage network. *IEEE Trans. Smart Grid* **2019**, *10*, 5163–5173. [CrossRef]
4. Hao, H.; Corbin, C.D.; Kalsi, K.; Pratt, R.G. Transactive control of commercial buildings for demand response. *IEEE Trans. Power Syst.* **2017**, *32*, 774–783. [CrossRef]

5. Tushar, W.; Saha, T.K.; Yuen, C.; Morstyn, T.; Al-Masood, N.; Poor, H.V. Grid influenced peer-to-peer energy trading. *IEEE Trans. Smart Grid* **2019**, *11*, 1407–1418. [CrossRef]
6. Hua, H.; Zhou, Y.; Qadrdan, M.; Wu, J.; Jenkins, N. Blockchain enabled decentralized local electricity markets with flexibility from heating sources. *IEEE Trans. Smart Grid* **2023**, *14*, 1607–1620. [CrossRef]
7. Lüth, A.; Zepter, J.M.; Granado, P.C.; Egging, R. Local electricity market designs for peer-to-peer trading: The role of battery flexibility. *Appl. Energy* **2018**, *229*, 1233–1243. [CrossRef]
8. Chen, J.; Wu, P.; Chen, W.; Guerrero, J.M.; Niu, Z.; Li, Z. Two-layer coordinated operation of multi-energy system considering carbon-oriented collaborative pricing mechanism via two-stage stochastic programming approach. *Appl. Energy* **2026**, *406*, 127298. [CrossRef]
9. Feng, C.; Huang, Z.; Lin, J.; Wang, L.; Zhang, Y.; Wen, F. Aggregation Model and Market Mechanism for Virtual Power Plant Participation in Inertia and Primary Frequency Response. *IEEE Trans. Power Syst.* 2025, *in press*. [CrossRef]
10. Jing, R.; Xie, M.; Wang, F.; Chen, L.X. Fair P2P energy trading between residential and commercial multi-energy systems enabling integrated demand-side management. *Appl. Energy* **2020**, *262*, 114551. [CrossRef]
11. Guo, Z.; Pinson, P.; Chen, S.; Yang, Q.; Yang, Z. Chance-constrained peer-to-peer joint energy and reserve market considering renewable generation uncertainty. *IEEE Trans. Smart Grid* **2021**, *12*, 798–809. [CrossRef]
12. Evdokimov, V.; Kudin, A.; Chikhladze, V.; Artemchuk, V. A Blockchain Architecture for Hourly Electricity Rights and Yield Derivatives. *FinTech* **2025**, *5*, 2. [CrossRef]
13. Blinov, I.V.; Parus, Y.V.; Artemchuk, V.O. Prosumer Operation Planning Model in the Retail Electricity Market. *Tech. Tech. Electrodyn.* **2026**, *1*, 50–61. [CrossRef]
14. Crespo-Vazquez, J.L.; AlSkaif, T.; González-Rueda, Á.M.; Gibescu, M. A Community-Based Energy Market Design Using Decentralized Decision-Making Under Uncertainty. *IEEE Trans. Smart Grid* **2021**, *12*, 1782–1793. [CrossRef]
15. Hua, H.; Qin, Y.; Hao, C.; Cao, J. Optimal energy management strategies for energy Internet via deep reinforcement learning approach. *Appl. Energy* **2019**, *239*, 598–609. [CrossRef]
16. Nie, H.; Zhang, J.; Chen, Y.; Xiao, T. Real-time Economic Dispatch of Community Integrated Energy System Based on a Double-layer Reinforcement Learning Method. *Power Syst. Technol.* **2021**, *45*, 1330–1336.
17. Lin, L.; Guan, X.; Peng, Y.; Wang, N.; Maharjan, S.; Ohtsuki, T. Deep Reinforcement Learning for Economic Dispatch of Virtual Power Plant in Internet of Energy. *IEEE Internet Things J.* **2020**, *7*, 6288–6301. [CrossRef]
18. Feng, J.; Ren, Z.; Li, C.; Li, W. A Benders-Combined Safe Reinforcement Learning Framework for Risk-Averse Dispatch Considering Frequency Security Constraints. *IEEE Trans. Circuits Syst. II Express Briefs* **2025**, *72*, 1063–1067. [CrossRef]
19. Liu, J.; Chen, J.; Wang, X.; Zeng, J.; Huang, Q. Energy Management and Optimization of Multi-energy Grid Based on Deep Reinforcement Learning. *Power Syst. Technol.* **2020**, *40*, 3794–3803.
20. Zhao, X.; Hu, J. Deep Reinforcement Learning Based Optimization for Charging of Aggregated Electric Vehicles. *Power Syst. Technol.* **2021**, *45*, 2319–2327.
21. Nguyen, T.T.; Dm Nguyen, N.; Nahavandi, S. Deep Reinforcement Learning for Multiagent Systems: A Review of Challenges, Solutions, and Applications. *IEEE Trans. Cybern.* **2020**, *50*, 3826–3839. [CrossRef]
22. Chen, T.; Bu, S.; Liu, X.; Kang, J.; Yu, F.R.; Han, Z. Peer-to-Peer Energy Trading and Energy Conversion in Interconnected Multi-Energy Microgrids Using Multi-Agent Deep Reinforcement Learning. *IEEE Trans. Smart Grid* **2022**, *13*, 715–727. [CrossRef]
23. Zhu, F.; Yang, Z.; Lin, F.; Xin, Y. Decentralized Cooperative Control of Multiple Energy Storage Systems in Urban Railway Based on Multiagent Deep Reinforcement Learning. *IEEE Trans. Power Electron.* **2020**, *35*, 9368–9379. [CrossRef]
24. Feng, C.; Li, H.; Tang, F.; Wen, F.; Zhang, Y. Deep reinforcement learning method for voltage control consideringtopology change of distribution system. *Electr. Power Autom. Equip.* **2025**, *45*, 156–163.
25. Zhang, J.; Pu, T.; Li, Y.; Wang, Y.; Zhou, X. Multi-agent Deep Reinforcement Learning Based Optimal Dispatch of Distributed Generators. *Power Syst. Syst. Technol.* **2021**, *46*, 3496–3504. [CrossRef]
26. Hadiya, N.; Teotia, F.; Bhakar, R.; Mathuria, P.; Datta, A. A comparative analysis of pricing mechanisms to enable P2P energy sharing of rooftop solar energy. In *Proceedings of the 2020 IEEE International Conference*; IEEE: New York, NY, USA, 2020; pp. 1–6.
27. Ausgrid. Time of Use Pricing [EB/OL]. Available online: https://www.ausgrid.com.au/Your-energy-use/Meters/Time-of-use-pricing (accessed on 27 April 2020).
28. Wang, Z.; Gu, Z.; Guerrero, J.; Shen, Y.; Guo, Z.; Deng, Z.; Huang, C.; Li, Z. Two-Period Two-Layer Electrical Model-Free Voltage Calculation for Active Distribution Network via Improved Broad Learning System. *IEEE Trans. Power Syst.* 2026, *in press*. [CrossRef]

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