buildings

MDPI

Article

Coordinated Scheduling of BESS–ASHP Systems in Zero-Energy Houses Using Multi-Agent Reinforcement Learning

Jing Li [1,2], Yang Xu [3,*], Yunqin Lu [4] and Weijun Gao [1,2,5]

[1] Faculty of Environmental Engineering, The University of Kitakyushu, Kitakyushu 808-0135, Japan; li_jing@qut.edu.cn (J.L.); gaoweijun@me.com (W.G.)
[2] Innovation Institute for Sustainable Maritime Architecture Research and Technology, Qingdao University of Technology, Fushun Road 11, Qingdao 266033, China
[3] Faculty of Information and Control Engineering, Qingdao University of Technology, 777 Jialingjiang Road, Qingdao 266520, China
[4] Faculty of Engineering, Osaka Institute of Technology, Osaka 535-8585, Japan; yunqin.lu@oit.ac.jp
[5] Department of Architecture, Institut Teknologi Sepuluh Nopember, Kampus ITS, Sukolilo, Surabaya 60111, Indonesia
* Correspondence: 901020170087@qut.edu.cn

Abstract

This paper addresses the critical challenge of multi-objective optimization in residential Home Energy Management Systems (HEMS) by proposing a novel framework based on an Improved Multi-Agent Proximal Policy Optimization (MAPPO) algorithm. The study specifically targets the low convergence efficiency of Multi-Agent Deep Reinforcement Learning (MADRL) for coupled Battery Energy Storage System (BESS) and Air Source Heat Pump (ASHP) operation. The framework synergistically integrates an action constraint projection mechanism with an economic-performance-driven dynamic learning rate modulation strategy, thereby significantly enhancing learning stability. Simulation results demonstrate that the algorithm improves training convergence speed by 35–45% compared to standard MAPPO. Economically, it delivers a cumulative cost reduction of 15.77% against rule-based baselines, outperforming both Independent Proximal Policy Optimization (IPPO) and standard MAPPO benchmarks. Furthermore, the method maximizes renewable energy utilization, achieving nearly 100% photovoltaic self-consumption under favorable conditions while ensuring robustness in extreme scenarios. Temporal analysis reveals the agents' capacity for anticipatory decision-making, effectively learning correlations among generation, pricing, and demand to achieve seamless seasonal adaptability. These findings validate the superior performance of the proposed centralized training architecture, providing a robust solution for complex residential energy management.

Keywords: multi-agent reinforcement learning; home energy management systems; MAPPO; zero-energy houses

Check for updates

Academic Editor: Antonio Caggiano

Received: 17 December 2025
Revised: 4 January 2026
Accepted: 6 January 2026
Published: 8 January 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

1.1. Background

Amidst accelerating global urbanization, the built environment has emerged as a pivotal domain for addressing worldwide energy demand and environmental challenges [1]. Buildings currently account for approximately 40% of global energy consumption while contributing substantially to carbon emissions, establishing them as critical elements in climate change mitigation strategies [2]. This reality has accelerated the transition toward

smart buildings—advanced structures that integrate renewable energy systems such as photovoltaics (PV) with energy storage units to establish self-sustaining microgrids [3]. Within this context, the Japanese government has introduced comprehensive policies to promote the development of Zero Energy Houses (ZEH) in accordance with its Strategic Energy Plan, which targets net-zero emissions for all newly constructed residential buildings by 2030 [4]. Consequently, the ZEH concept has gained increasing momentum, with integrated energy systems managed through advanced Building Energy Management Systems (BEMS) to optimize operational efficiency and economic performance [5].

The widespread integration of distributed energy resources (DERs) in ZEHs has transformed buildings from passive energy consumers into active energy producers. This transformation is further facilitated by the deployment of advanced metering infrastructure and smart meters, enabling real-time energy monitoring, control, and optimization [6]. Through BEMS, buildings can now engage in bidirectional communication with the power grid, and consequently, the transition toward net-zero energy buildings has substantially accelerated the integration of renewable energy technologies, smart appliances, and energy storage systems [7]. While this transition offers significant environmental and economic benefits, it also introduces new operational challenges. A primary concern lies in the inherent intermittency of renewable energy sources, particularly PV and wind power, which are highly dependent on weather conditions. Furthermore, the stochastic nature of occupant behavior and energy consumption patterns complicates demand forecasting and system optimization [8]. These factors collectively intensify the difficulty of maintaining real-time balance between energy supply and demand, posing significant challenges to conventional building energy management approaches [9].

1.2. Literature Review

In the early stages of BEMS development, control strategies primarily relied on rule-based methods and model predictive control (MPC) approaches. However, rule-based methods are constrained by their dependence on expert knowledge and exhibit limited adaptability to complex environmental variations [10]. Meanwhile, MPC approaches are highly dependent on the accuracy of predictive models and require substantial effort during commissioning, calibration, and maintenance phases [11]. With the advancement of artificial intelligence (AI) technologies, deep reinforcement learning (DRL) has emerged as a flexible, model-free approach for optimizing building energy system operations, effectively overcoming the limitations of conventional model-based control methods [12]. Moreover, while MPC serves as a theoretical benchmark, its deployment is often constrained by high online computational demands and reliance on precise forecasting. Conversely, the DRL approach shifts computation to offline training, enabling low-latency online inference and inherent adaptability to prediction errors.

DRL-based approaches effectively overcome the limitations of traditional optimization methods through their inherently data-driven and model-free methodology [13]. In DRL-based frameworks, the decision-making process is typically formulated as a Markov Decision Process (MDP), wherein one or more agents interact with a simulated environment through trial-and-error learning to maximize long-term cumulative rewards [14]. Recent research has increasingly focused on applying DRL to BEMS to enhance operational intelligence, with subsequent studies further demonstrating its effectiveness in improving system performance and adaptability. Li et al. [15] developed a DRL-based framework for optimizing space heating systems in zero-energy buildings, achieving a remarkable 30% reduction in operational costs while effectively balancing thermal comfort maintenance and energy efficiency enhancement. A model-free DRL-based controller was proposed by Qin et al. [16] for residential HVAC systems to optimize the trade-off between thermal comfort and en-

ergy costs, with experimental results demonstrating a 15.3% improvement compared to conventional rule-based control strategies. Authors in [17] developed a twin-delayed depth deterministic policy gradient (TD3) based DRL approach for a home microgrid, which achieved a 17.82% annual energy cost reduction while requiring only limited training data. Ren et al. [18] employed the Soft Actor-Critic (SAC) algorithm for optimal decision-making and integrated an ensemble learning-based thermal comfort model, with the proposed method effectively addressing high-dimensional decision problems under uncertainty and achieving a 17.7% reduction in household electricity costs. Langer and Volling [19] employed a deep deterministic policy gradient (DDPG) model to optimize control of smart buildings integrated with renewable energy and energy storage systems, achieving 75% energy self-sufficiency while maintaining minimal comfort violations, outperforming rule-based benchmarks. Recent studies have also employed imitation learning strategies that mimic mixed-integer linear programming solvers to significantly accelerate training speed and achieve near-optimal operation costs [20]. Despite these advancements, single-agent DRL methods face inherent limitations arising from a fundamental mismatch between their centralized decision-making paradigm and the inherently distributed, high-dimensional, and multi-objective nature of building energy systems, posing significant challenges when scaling to address the growing complexity of large-scale BEMS.

In response to these limitations, recent research has increasingly shifted toward MADRL and distributed optimization frameworks. MADRL mitigates the complexity of BEMS by decomposing the system into multiple coordinated agents, each responsible for managing a specific subsystem or equipment component. Through inter-agent communication and cooperation, these agents collectively pursue global optimization objectives in a manner aligned with the physical and operational hierarchies of building energy systems. Furthermore, by decentralizing decision-making and learning processes, MADRL exhibits enhanced robustness and adaptability in addressing the non-stationary and dynamic characteristics inherent to integrated BEMS environments. For instance, Xu et al. [21] developed a three-stage scheduling framework for virtual power plants employing a model-assisted MADRL approach, effectively addressing modeling inaccuracies and high-dimensional uncertainties in demand-side resource aggregation. Wang et al. [22] introduced an imitation learning-integrated Multi-Agent Proximal Policy Optimization (PPO) algorithm to optimize the operation of hybrid building energy systems, with results demonstrating that the proposed framework not only improves PV self-consumption rates but also achieves significantly enhanced convergence performance. Deng et al. [23] developed an improved MADRL approach for smart building energy management by incorporating safety violation metrics into the reward function, with simulation results demonstrating a 15.3% reduction in total energy costs compared to the original MADDPG baseline. Zhang et al. [24] developed a MADRL method utilizing interior-point policy optimization for household energy scheduling, ensuring operational safety through constraint satisfaction mechanisms. The proposed method achieved near-zero constraint violations and demonstrated superior performance in balancing cost efficiency with safety requirements compared to benchmark approaches. Kumari et al. [25] proposed a decentralized BEMS framework integrating DRL with blockchain technology, utilizing a multi-agent architecture to achieve optimal scheduling of diversified loads and secure data exchange in energy transactions. Deng et al. [26] proposed a MADRL-based energy management strategy that effectively addresses PV generation uncertainty by decomposing smart buildings into multiple local energy networks, achieving up to 11.5% reduction in total costs within a single scheduling cycle. Yang et al. [27] proposed a multi-stage coordinated dispatch framework for electricity-hydrogen integrated energy systems based on the MAPPO algorithm; by employing a multi-agent approach for intraday corrections, the method reduces economic costs by approximately

7%, effectively balancing economic efficiency with operational safety. Yuan et al. [28] proposed an improved MAPPO framework integrating electricity price feature extraction with a reward shaping mechanism, effectively addressing multiple uncertainties in demand-side management while preventing the formation of new charging and discharging peaks.

1.3. Motivation and Contributions

The aforementioned studies have made notable contributions to the modeling and optimization of uncertainty scheduling in BEMS. However, the application of MADRL in BEMS still faces several challenges. First, since state transition probabilities in MADRL environments depend on the joint actions of multiple agents, the integration of numerous intelligent agents introduces substantial computational and coordination complexity. Previous studies have demonstrated that achieving convergence in multi-agent algorithms often requires extensive training data, and in some cases, large-scale simulator-generated datasets have been employed to approximate real-world operating conditions. Consequently, enabling efficient convergence of multi-agent algorithms with limited real-world data has become a critical challenge for their practical deployment in building energy systems (BES). Second, conventional DRL algorithms typically incorporate soft constraint penalty mechanisms (such as adding penalty terms to the reward function) to ensure safe device operation in BES. This approach tends to cause exploratory actions to oscillate between reward maximization and constraint satisfaction, while the transition from unsafe to safe actions is not explicitly reinforced, leading to slow or unstable convergence. Furthermore, existing studies have not effectively addressed the inherent trade-offs among economic efficiency, operational safety, and start–stop costs of BES devices. These unresolved conflicts may give rise to a range of operational issues, including microgrid imbalance, system overload, excessive operating costs, and scheduling delays.

To address these challenges, this study proposes a novel BES optimization method based on the MADRL framework. By integrating an action constraint projection mechanism and a perception-based dynamic learning rate into the conventional MAPPO algorithm, the proposed approach effectively enhances the training efficiency of multi-agent learning in dynamic and complex environments while ensuring safe, efficient, and reliable BEMS operation. The main contributions of this study are summarized as follows:

- An improved MAPPO algorithm is proposed. By incorporating an action constraint projection mechanism that constrains deviations from the safe operating range during action execution, all BESS operations are strictly maintained within SOC boundaries, enabling the MADRL framework to formulate policies ensuring safe agent behavior. Additionally, an economic-performance-driven dynamic learning rate modulation strategy is introduced in the policy network update process, which adaptively adjusts the update intensity, achieving approximately 35–45% improvement in convergence speed compared to standard MAPPO and effectively enhancing the stability and robustness of multi-agent learning.
- A novel multi-objective reward function is proposed within the MADRL framework. By integrating expert knowledge into the reward design process, the proposed function simultaneously considers economic efficiency and renewable energy utilization, successfully guiding agents to learn the inherent correlations among PV generation windows, electricity price structures, and load demand cycles, thereby enabling seasonally-adaptive energy management strategies without explicit mode switching.
- This study employs real measurement data collected from an operational zero-energy house in Kitakyushu, Japan, to train and validate the proposed algorithm. Experimental results demonstrate that the proposed method achieves a cumulative cost reduction of 15.77%, with PV self-consumption rates consistently outperforming the

baseline rule-based control strategy across all three representative test months. The proposed framework maintains robust performance and scalability under varying seasonal conditions, thereby validating its practical applicability and potential for real-world deployment in BEMS.

2. BESS–ASHP Coupled Energy System Model

This study develops a multi-objective optimization framework for an intelligent HEMS deployed in a ZEH located in Kitakyushu, Japan. As illustrated in Figure 1, the proposed system is designed to coordinate the operation of distributed renewable generation units and multi-energy storage devices within a dynamic residential environment, with the dual objectives of minimizing average daily household operating costs and maximizing PV self-consumption rates. Specifically, the system comprises four major components:

- A rooftop PV generation system that provides renewable electricity for household consumption;
- An ASHP serving as the primary heating device, with its coefficient of performance (COP) varying dynamically with ambient temperature;
- A BESS that enables temporal balancing of electrical energy and mitigates the intermittency of PV generation;
- A thermal storage tank (TSK) that buffers and stores thermal energy produced by the ASHP, enabling temporal decoupling between heat production and consumption.

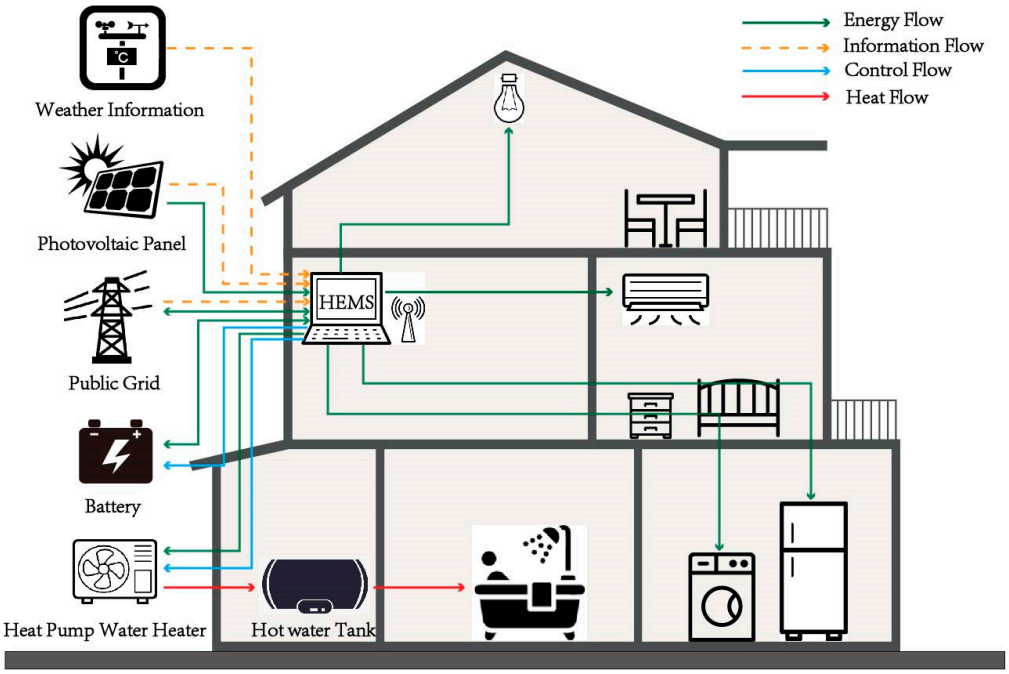

Figure 1. Schematic diagram of the HEMS architecture in the ZEH.

2.1. Optimization Objectives

The optimization problem is formulated to coordinate energy flows among the PV system, BESS, ASHP, and household electrical loads. The first objective aims to minimize the household's average daily operating cost, thereby characterizing the economic performance of the system. This objective is mathematically defined as:

$$\mathrm{m}inC_{avg} = \frac{1}{N}\sum_{t\in T}\left(E_{gird_buy}(t)\cdot P_{price}(t) - E_{sell}(t)\cdot P_{fit}(t)\right) \quad (1)$$

where $N = max(|T|, 1)$, C_{avg} denotes the daily average operational cost, $E_{\text{grid_buy}}(t)$ and $E_{\text{sell}}(t)$ represent the purchased and exported electricity at time t, respectively; $P_{\text{price}}(t)$ is the real-time electricity price, and $P_{\text{fit}}(t)$ denotes the fixed feed-in tariff. In parallel, the model also seeks to maximize the self-consumption rate of PV generation, which is defined as:

$$\text{m}axr_{pv} = \frac{\sum_{t\in T}\left(E_{pv}(t) - E_{sell}(t)\right)}{\sum_{t\in T}\left(E_{pv}(t) + \varepsilon\right)} \times 100\% \quad (2)$$

where r_{pv} represents the proportion of locally consumed PV electricity relative to the total PV generation, ε is a small constant used to prevent division by zero.

2.2. Multi-Energy Flow Balance Constraints

To ensure operational stability and power supply reliability, both electrical and thermal energy flows must remain balanced at each time step. These balance requirements constitute the fundamental physical constraints of the optimization model. The electrical energy balance is expressed in Equation (3):

$$E_{demand}(t) = E_{pv_demand}(t) + E_{battery_discharge}(t) + E_{gird_bug}(t) \quad (3)$$

The generated PV output is allocated across multiple utilization pathways, as expressed below:

$$E_{pv}(t) = E_{pv_demand}(t) + E_{pv_battery}(t) + E_{pv_gird}(t) + E_{pv_hp}(t) \quad (4)$$

Similarly, the thermal energy balance among the ASHP, the TSK, and the heat demand can be formulated as Equation (5):

$$P_{hp_thermal}(t) + P_{tank_charge}(t) + P_{tank_discharge}(t) = d_{thermal}(t) \quad (5)$$

where $d_{\text{thermal}}(t)$ denotes the instantaneous thermal demand. The integration of the ASHP with the TSK enables temporal decoupling of heat production from consumption, allowing the system to store thermal energy during periods of low electricity prices or high PV generation.

2.3. Dynamic Models and Operational Constraints

The BESS serves two primary functions: enabling temporal energy shifting and mitigating the intermittency of PV generation. The state of charge (SOC) of the BESS evolves over time as governed by the following equation:

$$SOC(t+1) = (1-\sigma)\cdot SOC(t) + (\eta_{ch}\cdot E_{ch}(t) - \frac{E_{dch}(t)}{\eta_{dch}})/E^{cap} \quad (6)$$

where σ represents the self-discharge rate, while η_{ch} and η_{dch} denote charging and discharging efficiencies. E^{cap} is the battery capacity. Operational constraints are defined as Equation (7):

$$SOC_{min} \le SOC_t \le SOC_{max}, 0 \le E_{ch}(t) \le E_{ch_max}, 0 \le E_{dch}(t) \le E_{dch_max} \quad (7)$$

The performance of the ASHP is influenced by the ambient temperature. As characterized in Equation (8), the COP incorporates the empirical model proposed by Authors in [6],

which is derived from the polynomial fitting of measured scatter data regarding ASHP power consumption versus ambient temperature:

$$COP_{hp}(t) = \frac{3.41}{1 - 0.014 \cdot Temp(t) - 3.5 \times 10^{-6} \cdot Temp(t)^2} \quad (8)$$

The heat output from the ASHP is defined as Equation (9):

$$P_{hp_thermal}(t) = COP_{hp}(t) \cdot E_{hp}(t) \quad (9)$$

The dynamic energy content in the TSK is updated as:

$$E_{tank}(t+1) = (1-\varepsilon) \cdot E_{tank}(t) + \eta_{ch_tank} \cdot P_{tank_ch}(t) - \frac{P_{tank_dch}(t)}{\eta_{dch_tank}} \quad (10)$$

where ε is the thermal loss rate, and $\eta_{\text{tank_ch}}$, $\eta_{\text{tank_dch}}$ are the charging and discharging efficiencies. The ASHP thermal output is constrained by Equation (11):

$$P_{thermal_min} \leq P_{hp_thermal}(t) \leq P_{thermal_max} \quad (11)$$

2.4. Baseline Rule-Based Control Strategy

The target building currently operates under a rule-based control strategy. Specifically, the BESS and ASHP operate at their rated power according to the energy balance relationships defined in Equations (3)–(5), while complying with the operational constraints specified in Equations (7)–(11). Regarding thermal management, the ASHP performs full-power operation from 04:00 to 07:00 daily to charge the TSK, whereas during the remaining hours it operates on a demand-driven basis to satisfy real-time heating requirements and maintain indoor thermal comfort.

The system interacts with the electricity market through a dual-pricing mechanism: electricity is purchased at the real-time dynamic price $P_{price}(t)$, while surplus renewable generation is sold at a fixed feed-in tariff $P_{fit}(t)$. To prevent energy arbitrage, the model prohibits direct bidirectional energy exchange between the BESS and the grid. Consequently, the BESS is permitted to charge exclusively from PV generation.

This baseline strategy serves as the reference benchmark for this study, enabling quantification of performance improvements achieved by subsequent optimization strategies and providing a consistent basis for comparison across different control schemes.

3. HEMS Optimization Framework Based on MADRL

To address the multi-objective control requirements inherent in the BESS–ASHP coupled energy system within ZEHs, this study develops an operational optimization framework based on the MAPPO algorithm, as illustrated in Figure 2. Within this framework, the integrated BESS–ASHP energy coupling model established in the preceding section is first reformulated as a multi-agent Markov game under the DRL paradigm. Building upon this formulation, the MAPPO algorithm is further integrated with a constraint projection mechanism and a dynamic learning rate modulation strategy to enable coordinated optimal scheduling of BESS and ASHP operations.

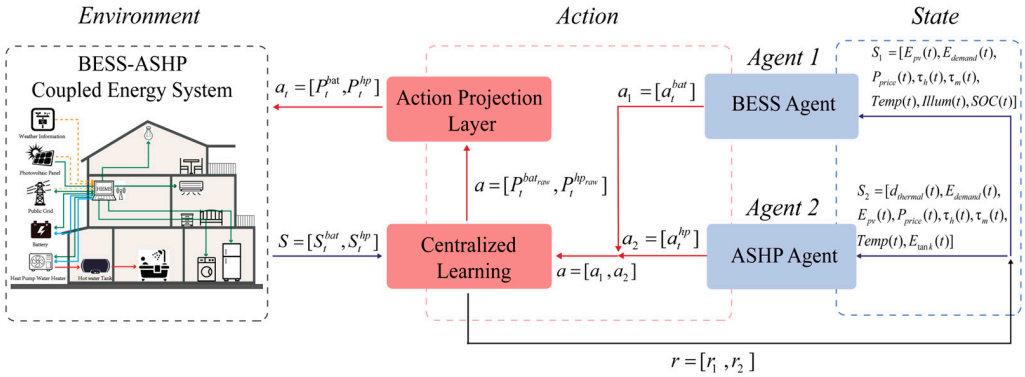

Figure 2. DRL-Based Multi-Agent Solution Framework for HEMS Optimization.

3.1. Formulation of the MDP

The multi-agent environment is formulated as a MDP, defined by the tuple ($\mathcal{S}$, $\mathcal{A}$, R, γ, $\mathcal{P}$), where $\mathcal{S}$ denotes the state space; $\mathcal{A} = \prod_{i=1}^{n} \mathcal{A}_i$ represents the joint action space of all agents; $\mathcal{P}$ is the state transition probability distribution; R is the reward function; and γ is the discount factor.

3.1.1. State Space

It can be seen from Figure 2 that each agent receives a local observation that corresponds to a subset of the global state. These observations include both environmental variables and internal device states, as defined in Equations (12) and (13):

$$\mathcal{S}_t^{\text{bat}} = \left[E_{pv}(t), E_{demand}(t), P_{price}(t), \tau_h(t), \tau_m(t), Temp(t), Illum(t), SOC(t)\right] \tag{12}$$

$$\mathcal{S}_t^{\text{hp}} = \left[d_{thermal}(t), E_{demand}(t), E_{pv}(t), P_{price}(t), \tau_h(t), \tau_m(t), Temp(t), E_{tank}(t)\right] \tag{13}$$

where $\tau_h(t)$ denotes the hour of the day (encoded using a sinusoidal representation), $\tau_m(t)$ denotes the month (encoded as an integer), $Temp(t)$ is the outdoor temperature, and $Illum(t)$ represents the solar irradiance.

3.1.2. Action Space and Action Projection Mechanism

The system incorporates two cooperative agents within the MADRL framework: a BESS agent, which determines the charging/discharging command a_t^{bat}, and a ASHP agent, which outputs the ASHP operating command a_t^{hp}. The resulting joint action at time step t is expressed as:

$$a_t = \left\{a_t^{bat}, a_t^{hp}\right\} \tag{14}$$

The variable a_t^{bat} ranges from -1 to 1, where negative values indicate charging and positive values indicate discharging. The variable a_t^{hp} represents the ASHP control factor, which varies between 0 and 1. These normalized actions are linearly scaled to obtain the actual power outputs of the BESS and the ASHP, which can be expressed as follows:

$$P_t^{bat_raw} = a_t^{bat} \cdot P_{max}^{bat} \tag{15}$$

$$P_t^{hp_raw} = a_t^{hp} \cdot P_{max}^{hp} \tag{16}$$

However, due to the physical and operational constraints of the BESS and ASHP subsystems, the raw actions generated by the policy network may not be feasible. Therefore, an

action projection layer is applied after the agent outputs but before environment execution, ensuring safe and physically consistent operation. The projection process is formulated as:

$$\widetilde{a}_t = \Pi_{\mathcal{A}}(a_t) \tag{17}$$

where $\Pi_{\mathcal{A}}(\cdot)$ denotes the projection operator onto the feasible action set $\mathcal{A}$. To ensure that the policy outputs comply with the battery's physical constraints (Equation (7)) and the SOC boundary conditions (Equation (6)), an action-projection mechanism is introduced. This mechanism projects the raw power commands onto the intersection of the static power limits and the SOC-dependent dynamic feasible region, as defined in Equation (18):

$$P_t^{bat} = clip\left(P_t^{bat_raw}, P_{max,t}^{ch}, P_{max,t}^{dch}\right) \tag{18}$$

Here, $P_{max,t}^{ch}$ denotes the maximum feasible charging power at time t, and $P_{max,t}^{dch}$ denotes the maximum feasible discharging power at time t. Since the ASHP is constrained by its COP (Equation (8)), the TSK capacity (Equation (9)), and the real-time heat demand, an action-projection mechanism is applied, which is defined as follows:

$$P_t^{hp} = clip\left(P_t^{hp_raw}, P_{min,t}^{hp}, P_{max,t}^{hp}\right) \tag{19}$$

Here, $P_{max,t}^{hp}$ represents the maximum feasible operating power of the ASHP at time t, and $P_{min,t}^{hp}$ represents the minimum feasible operating power at time t. The incorporation of the aforementioned action constraint layer enables agents to explore freely within a normalized continuous action space. This layer projects the raw outputs of agents into final executable actions in real time, thereby ensuring strict compliance with all physical and operational constraints of the energy system. Despite the additional computational overhead introduced by this layer, it not only effectively eliminates the occurrence of unsafe actions but has also been empirically demonstrated to significantly enhance training stability and accelerate convergence.

3.1.3. Reward Function

Figure 2 shows that two agents—the BESS agent and the ASHP agent—jointly determine the system operation. For any given state s_i, each agent seeks an optimal action a_i that maximizes its immediate reward r_i. A centralized training and decentralized execution (CTDE) paradigm are adopted: during training, global states and joint actions are utilized to capture systemwide couplings among renewable generation, loads, and storage units; during execution, each agent makes decisions independently based on local observations, ensuring deployability and robustness in real-world operation. To simultaneously evaluate operational economy and PV utilization efficiency, the total reward is formulated as follows:

$$R(t) = \alpha R_{eco}(t) + \beta R_{pv}(t) \tag{20}$$

Here, $R_{\mathrm{eco}}(t)$ denotes the grid cost reward, and $R_{\mathrm{pv}}(t)$ represents the PV utilization reward. The coefficients α and β regulate the relative importance of the two components. To ensure consistent reward scaling, the $R_{\mathrm{eco}}(t)$ is formulated based on the average electricity cost over the entire scheduling horizon, as expressed in Equation (21):

$$R_{eco} = -\left(\frac{1}{N}\sum\nolimits_{t\in T}\left(E_{gird_buy}(t)\cdot P_{price}(t) - E_{sell}(t)\cdot P_{fit}(t)\right)\right) \tag{21}$$

The negative sign ensures that higher electricity costs lead to lower rewards, guiding agents to learn economically optimal strategies. The $R_{pv}(t)$ encourages effective utilization

of on-site renewable energy. It is defined based on the incremental PV usage relative to a baseline:

$$R_{pv}(t) = r_{pv}(t) - r_{pv}^{baseline}(t) \tag{22}$$

$r_{\mathrm{pv}}(t)$ denotes the PV self-consumption rate achieved by the MADRL agents at time t, while $r_{\mathrm{pv}}^{\mathrm{Baseline}}(t)$ represents the corresponding baseline ratio. Their computation is defined as shown in Equation (23):

$$r_{pv}(t) = \frac{E_{pv}(t) - E_{sell}(t)}{E_{pv}(t) + \varepsilon} \tag{23}$$

where ε is a small constant used to prevent division by zero. The proposed reward function effectively guides the agents' learning process by jointly optimizing economic efficiency and PV utilization. Under the CTDE paradigm, this approach leads to policies that are not only performant but also practical for real-world deployment.

3.2. The Selection of MADRL Framework

In MADRL, the information-processing schemes employed during training and execution phases often differ. When applied to HEMS, various MADRL frameworks demonstrate distinct advantages in terms of coordination capability, scalability, and deployment flexibility. Based on information availability and the degree of inter-agent cooperation, existing approaches can be classified into three principal paradigms:

Decentralized Training and Decentralized Execution (DTDE)

Under this paradigm, each subsystem or device within the building energy system (e.g., ASHP, BESS, or thermoelectric coupling equipment) is modeled as an independent agent. Both training and execution processes rely exclusively on local observations, thereby offering favorable scalability and suitability for loosely coupled scenarios. However, the absence of centralized coordination mechanisms may result in action inconsistency among agents, consequently impeding the achievement of global energy management objectives. IPPO represents a typical method within this category.

Centralized Training and Centralized Execution (CTCE)

Under this paradigm, both training and action selection depend on global observations, with a unified policy governing system-wide operations. Although this approach ensures action consistency at the system level, its deployment in complex environments is constrained by substantial data requirements, limited sample efficiency, and communication bottlenecks, thereby compromising scalability.

Centralized Training and Decentralized Execution (CTDE)

The CTDE paradigm incorporates global information during training to facilitate accurate value estimation and effective credit assignment, while preserving operational flexibility through decentralized execution. The centralized critic network assists in mitigating the inherent instability associated with multi-device coordination, rendering this paradigm particularly well-suited for strongly coupled HEMS scenarios, such as coordinated scheduling of source–load–storage resources and multi-energy complementary operations. The MAPPO algorithm adheres to this paradigm and is therefore adopted as the target framework in this study.

3.3. The Training Mechanism of the MAPPO

This section develops the MAPPO-based training mechanism for HEMS Optimization, as illustrated in Figure 3. The MAPPO framework consists of a policy network and a value network: the policy network generates the control action for agent i given its local state s_t^i, while the value network evaluates the corresponding long-term value. As we can see, the BESS agent and the ASHP agent generate their respective actions a_t^i during decentralized execution based on local observations o_t^i. During training, however, centralized learning

is performed using the global state s_t to enable coordinated policy updates. The training procedure is composed of three main stages: training sample generation, value network training, and policy network optimization.

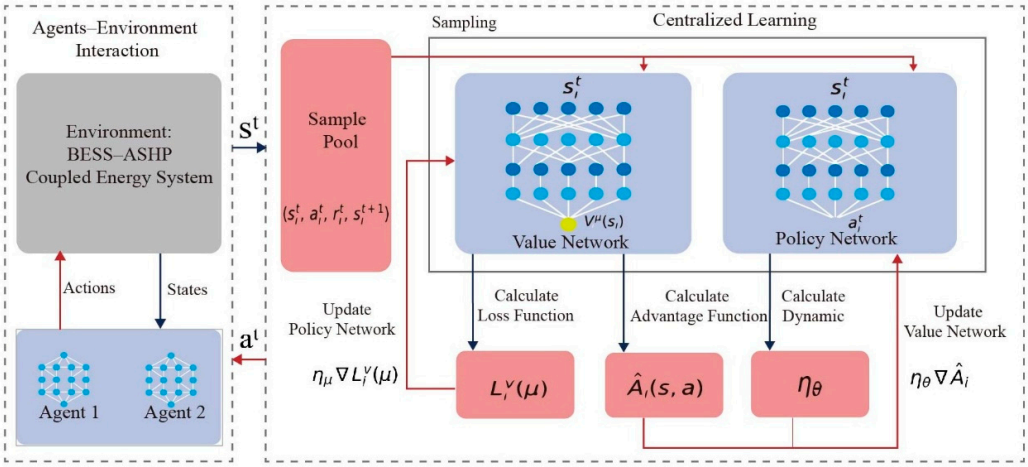

Figure 3. MAPPO-Based Training Workflow for HEMS Optimization.

3.3.1. Training Sample Generation

During this phase, each agent produces the parameters of a Gaussian distribution through its policy network $\pi_{\theta_i}(a_t^i \mid s_t^i)$ based on the local observation s_t^i at each sampling step. This defines a conditional probability distribution over continuous actions, from which the action a_t^i at time t is sampled. The actions of all agents are then aggregated into a joint action vector $a_t = \{a_t^1, \ldots, a_t^N\}$, which is fed into the HEMS environment to obtain the economic reward r_t (as defined in Equations (20)–(22)) and the next global state s_{t+1}. The resulting transition tuple (s_t, a_t, r_t, s_{t+1}) is stored in the replay buffer for subsequent batch training.

3.3.2. Value Network Training

The value network is designed to estimate the long-term return of the current policy under a given local state, thereby providing a reliable basis for advantage computation and policy improvement. During the training phase, each agent first samples a mini-batch of experiences $[s_t^i, a_t^i, r_t, s_{t+1}^i]$ from the replay buffer, and feeds its local observation s_t^i into the value network $V_{\mu_i}(s_t^i)$ to obtain the estimated state value [29]. The target value is defined in Equation (24):

$$V_i^{\text{target}}(s_t^i) = \mathbb{E}\left[\sum_{\tau=t}^{\infty} \gamma^{\tau-t} r_\tau\right] \tag{24}$$

$V_i^{\text{target}}(s_t^i)$ represents the discounted cumulative return under the current policy and serves as the supervision signal for training the value network. The training objective $L_v(\mu_i)$ is formulated as minimizing the mean squared error between the predicted and target values [30]:

$$L_v(\mu_i) = \mathbb{E}\left[\left(V_i^{\text{target}}\left(s_t^i\right) - V_{\mu_i}\left(s_t^i\right)\right)^2\right] \tag{25}$$

In each iteration, the parameters of the value network μ_i are updated through gradient descent so that the network progressively approximates the true long-term economic return, which is defined in Equation (26):

$$\mu_i = \mu_i - \eta_\mu \nabla_{\mu_i} L_v(\mu_i) \tag{26}$$

where η_μ denotes the learning rate of the value network. Through continuous training and iterative optimization, the value network gradually enhances its ability to accurately

estimate long-term returns, thereby providing a stable and trustworthy basis for advantage calculation and subsequent policy optimization.

3.3.3. Policy Network Optimization

The primary objective of policy network training is to increase the probability of selecting actions that yield higher economic performance. To quantify the relative improvement of an action over the baseline policy, the advantage function $\hat{A}_i(s_t^i, a_t^i)$ is incorporated during training, which is formulated as:

$$\hat{A}_t = \sum_{l=0}^{K-1} (\gamma\lambda)^l (r_{t+l} + \gamma V_{\mu_i}(s_{t+l+1}) - V_{\mu_i}(s_{t+l})). \tag{27}$$

where γ is the discount factor and λ is the smoothing parameter that balances bias and variance. This advantage estimator effectively characterizes the performance gain of the current action relative to the value baseline, thereby guiding the update direction of the policy. Next step, the policy network is optimized using the Clipped PPO objective, which constrains the deviation between the updated and previous policies, ensuring stable learning and preventing policy collapse [31]. The clipped surrogate objective is defined as:

$$L_{\text{CLIP}}(\theta_i) = \mathbb{E}_t\left[\min\left(\rho_t(\theta_i)\hat{A}_t,\ \text{clip}(\rho_t(\theta_i), 1-\delta, 1+\delta)\hat{A}_t\right)\right] \tag{28}$$

where $\rho_t(\theta_i)$ denotes the probability ratio between the new and old policies, and δ is the clipping threshold. This design effectively mitigates excessive policy updates and enhances convergence stability [32]. The policy parameters θ_i are updated according to:

$$\theta_i = \theta_i - \eta_\theta \nabla_{\theta_i} L_{\text{CLIP}}(\theta_i) \tag{29}$$

where η_t represents the learning rate of the policy network. Through repeated cycles of environment interaction, value estimation, and policy improvement, the multi-agent system gradually converges to a cooperative energy management strategy that maximizes economic performance.

3.4. Dynamic Learning Rate Modulation Strategy Based on Economic Performance Scoring

As discussed in the preceding sections, the efficiency and stability of multi-agent learning remain critical challenges that hinder practical deployment. To address this issue, this study proposes a dynamic learning rate modulation strategy driven by economic performance scoring. The proposed mechanism utilizes the average grid cost over each scheduling cycle as an economic indicator, which is directly integrated into the MAPPO policy update process [33]. By adaptively adjusting the learning rate in accordance with the actual economic performance of the policy, this method accelerates policy improvement during periods of performance degradation while stabilizing updates when satisfactory performance is achieved, thereby enhancing the overall efficiency and robustness of the multi-agent training process [34].

The integrated economic performance metric J_{eco} is defined as the average cost associated with electricity import and export over a scheduling horizon of length T. Building upon this metric, a compact sigmoid-based adaptive learning rate mechanism is introduced to dynamically regulate the policy update speed according to the economic performance. Specifically, the average economic score J_{eco} is compared against a reference value J_{ref}. The normalized deviation is then mapped through a sigmoid function to produce a bounded scaling factor $f(J_{\text{eco}}) \in [0.5,\ 2]$. This factor adaptively amplifies or suppresses the learning rate depending on whether the economic performance is favorable or suboptimal. During policy parameter updates (as shown in Equation (29)), the learning rate is dynamically

modulated by the scaling factor, ensuring stable convergence when the performance is satisfactory while enabling accelerated improvement when the performance deteriorates. The complete adjustment process is described in Equations (30)–(33).

$$J_{eco} = -\frac{1}{N}\sum_{t\in T}\left(E_{grid}(t)P_{grid}(t) - E_{sell}(t)P_{sell}(t)\right) \tag{30}$$

$$J_{\text{ref}} \leftarrow \alpha_{\text{EMA}}\, J_{\text{eco}} + (1-\alpha_{\text{EMA}})\, J_{\text{ref}} \tag{31}$$

$$f(J_{\text{eco}}) = \kappa_{\min} + (\kappa_{\max} - \kappa_{\min}) \cdot \text{sigmoid}(k\,(J_{\text{eco}} - J_{\text{ref}})) \tag{32}$$

$$\eta_t = \eta_0 \cdot f(J_{eco}) \tag{33}$$

Here, $\alpha_{\text{EMA}} \in (0,1]$ denotes the decay coefficient of the exponential moving average (EMA), which governs the weighting between recent and historical reference values. We set α_{EMA} as 0.05 to strike a balance between responsiveness and numerical smoothness. The parameters $\kappa_{\min} = 0.5$ and $\kappa_{\max} = 2.0$ specify the lower and upper bounds of the learning-rate scaling factor. The coefficient k represents the sensitivity parameter of the sigmoid mapping and is set to 2 to ensure an appropriate responsiveness of the scaling factor to economic deviations. Finally, η_0 denotes the baseline learning rate of the policy network prior to dynamic modulation.

4. Case Study

4.1. Data Source

To evaluate the effectiveness of the proposed optimization framework, this study utilizes measurement data collected from the HEMS of an operational ZEH located in Kitakyushu, Japan. The dataset encompasses energy output and demand records of household energy systems spanning from 1 January 2022 to 30 August 2023, with a sampling interval of 30 min. In addition to the measured data, real-time electricity prices (RTP) were synthesized based on electricity futures published by Kyushu Electric Power Company, while the real-time COP of the ASHP was calculated in accordance with Equation (8).

Figure 4 illustrates the characteristic seasonal and intraday variations in hot water demand, electricity demand, and electricity price. At the seasonal scale, all three variables exhibit a consistent pattern characterized by elevated values in winter, moderate levels in summer, and reduced magnitudes in spring and autumn. In January, both hot water demand (0.501 kWh) and electricity demand (1.158 kWh) reach their annual peaks, accompanied by the highest electricity price (17.647 Yen/kWh), attributable to increased heating loads and tighter supply–demand conditions. During summer (August), hot water demand declines to its annual minimum (0.251 kWh), whereas cooling loads maintain electricity demand and prices at the second-highest levels. Spring (April) and autumn (October) demonstrate notably lower demand and price levels, reflecting operational characteristics under mild climatic conditions. At the intraday scale, all three variables display pronounced periodicity. Hot water demand consistently exhibits morning (6:00–9:00) and evening (18:00–21:00) peaks across all seasons. Electricity demand follows a typical residential load profile characterized by distinct dual peaks and a pronounced nighttime valley, with winter presenting elevated overall levels. Electricity prices maintain a stable time-of-use pattern throughout the year, featuring off-peak pricing during 0:00–6:00, mid-peak pricing during 9:00–16:00, and prominent peak pricing during 18:00–21:00. In summary, the seasonal contrasts and intraday cyclic structures depicted in Figure 4 establish essential contextual features that enable DRL-based strategies to perform effective multi-timescale demand response optimization.

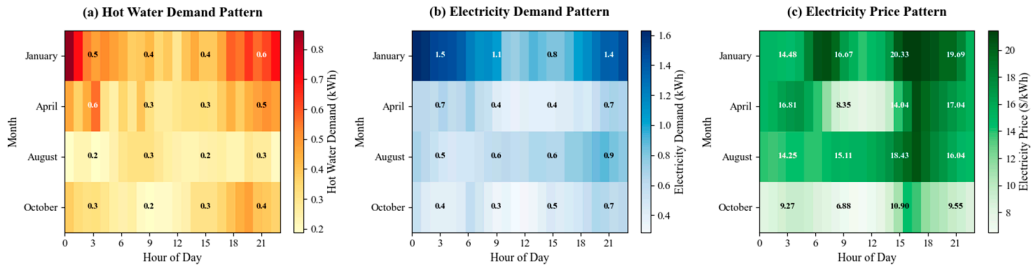

Figure 4. Seasonal and intraday variations in hot water demand, electricity demand, and electricity price.

4.2. Experimental Setup

Based on the aforementioned dataset analysis, the division of training and testing data in this study explicitly accounts for seasonal variability and characteristic operational patterns of annual energy consumption. The training set comprises continuous half-hourly measurements spanning from 1 January to 31 December 2022, thereby capturing the complete range of annual fluctuations in demand, real-time electricity prices, and climatic conditions. In the training process, a single episode is defined as a 24-h dispatch cycle, consisting of 48 time steps. The experimental framework was implemented using Python (v3.9, Python Software Foundation, USA), with PyTorch (v2.0, Meta Platforms, Inc., USA) and OpenAI Gym (v0.21.0, OpenAI, USA). This comprehensive dataset enables the MADRL framework to learn multi-timescale load response behaviors across diverse seasonal regimes, meteorological conditions, and user activity patterns, ultimately enhancing the robustness of the control strategy. The test set encompasses data from January, April, and August 2023, representing typical winter, spring, and summer operating conditions, respectively. This experimental design is intended to evaluate the generalization capability of the model under seasonal transitions and to assess whether the proposed strategy can maintain stable scheduling and optimization performance when subjected to varying external driving factors.

To demonstrate the advantages of the proposed methodology, four simulation cases are designed. All MADRL-based methods employ a unified set of hyperparameters, with detailed configurations presented in Table 1. The four simulation cases are defined as follows:

Table 1. Hyperparameter configurations for MADRL models.

Parameter	Value	Parameter	Value
Learning rate (actor)	0.0003	Batch size	256
Learning rate (critic)	0.0005	Max gradient norm	0.5
Discount factor γ	0.99	Replay buffer size	50,000
GAE λ	0.95	Hidden layer size	256×256
Clip ratio ε	0.2	Activation function	ReLU

Case 1: Baseline Scenario

The baseline scenario represents the actual operational behavior of the system in the absence of any optimization measures, with system dynamics strictly adhering to the modeling assumptions outlined in Section 2.4. This scenario serves as a reference benchmark for quantifying the performance improvements achieved by subsequent optimization strategies and provides a consistent basis for comparative evaluation across different control schemes.

Case 2: IPPO-Based Optimization Scenario

In this case, the IPPO framework is applied to optimize the BESS–ASHP coupled energy system established in Section 2. This framework corresponds to the Decentralized Training and Decentralized Execution (DTDE) architecture described in Section 3.2. The objective of this scenario is to evaluate the optimization capability of IPPO and to establish a comparative basis for quantifying the performance gains achieved by MAPPO over IPPO.

Case 3: Traditional MAPPO Scenario

This case employs the standard MAPPO algorithm for joint scheduling optimization of the BESS–ASHP system, serving as a benchmark for evaluating the proposed methodological improvements with respect to learning stability and optimization performance.

Case 4: Improved MAPPO Scenario

The final scenario implements the enhanced MAPPO approach proposed in Section 3, which incorporates both a constraint projection mechanism and a dynamic learning rate modulation strategy. This case is designed to validate the effectiveness of the proposed improvements and to demonstrate their potential for enhancing multi-agent coordination in complex coupled energy systems.

Furthermore, the interaction environment for multi-agent algorithms is constructed based on the integrated BESS–ASHP energy system model developed in Section 2. Specifically, the weighting coefficients α and β were determined through heuristic tuning to optimally balance economic efficiency and PV self-consumption. All key system parameters are rigorously configured in accordance with the technical specifications of the actual equipment installed in the target building, as summarized in Table 2.

Table 2. Model parameters and operational settings for the BESS–ASHP coupled energy system.

Category	Symbol	Description	Value
BESS	σ	Battery self-discharge coefficient	0.0001
	η_{ch}	Charging efficiency	0.95
	η_{dch}	Discharging efficiency	0.95
	E^{cap}	Battery energy capacity	6.0 kWh
	E_{ch_max}	Maximum charging power	0.75 kW
	E_{dch_max}	Maximum discharging power	0.75 kW
	SOC_{min}	Minimum allowable SOC	0.2
	SOC_{max}	Maximum allowable SOC	0.9
ASHP	$P_{thermal_min}$	Minimum thermal output	0
	$P_{thermal_max}$	Maximum thermal output	3.0 kW
	ε	Thermal energy loss factor	0.0005
TSK	η_{ch_tank}	Tank charging efficiency	0.95
	η_{dch_tank}	Tank discharging efficiency	0.95
	E_{tank_cap}	Thermal storage capacity	20.0 kWh
Optimization	α	Weighting coefficient for economic cost	0.6
	β	Weighting coefficient for PV self-consumption	0.4

5. Result and Discussion

5.1. Training Process Analysis

Figure 5 illustrates the evolution of training rewards for the three MADRL frameworks. The solid lines and shaded areas in Figure 5 represent the mean rewards and standard deviations obtained across five independent runs with different random seeds. Overall, all methods demonstrate continuous performance improvement and eventual convergence; however, considerable differences are observed in their convergence rates and final reward levels. IPPO exhibits the slowest learning progress, requiring approximately

250–350 episodes to reach a stable region, with its final reward consistently remaining lower than those of the other two approaches. This inferior performance can be attributed to the fully independent policy update mechanism employed by IPPO, which limits its capacity to capture inter-agent interactions and coordinated behaviors, thereby substantially diminishing its ability to address the coupled dynamics inherent in building energy systems. In contrast, MAPPO demonstrates a noticeably faster reward increase during the early training phase and achieves convergence within approximately 180–250 episodes. By leveraging a centralized value function to share global state information, MAPPO facilitates more effective learning of system-level couplings, consequently enhancing cooperative control performance. Nevertheless, its performance remains constrained by the fixed learning rate scheme, which results in larger reward fluctuations during the initial training stages and limits the achievable policy optimality in certain cases. The proposed Improved-MAPPO demonstrates superior performance in terms of both convergence speed and final reward magnitude. Benefiting from the economic-performance-guided dynamic learning rate mechanism, the policy update step size is adaptively adjusted based on real-time scheduling quality. Furthermore, the incorporation of action projection constraints effectively reduces invalid or inefficient exploratory actions. As depicted in Figure 5, Improved-MAPPO attains near-optimal reward levels within only 100–150 episodes, representing the fastest convergence among the three methods. Moreover, the notably narrow confidence intervals demonstrate the algorithm's high stability and reproducibility against random initialization, effectively mitigating the sensitivity issues regarding hyperparameters and seed settings common in traditional MADRL.

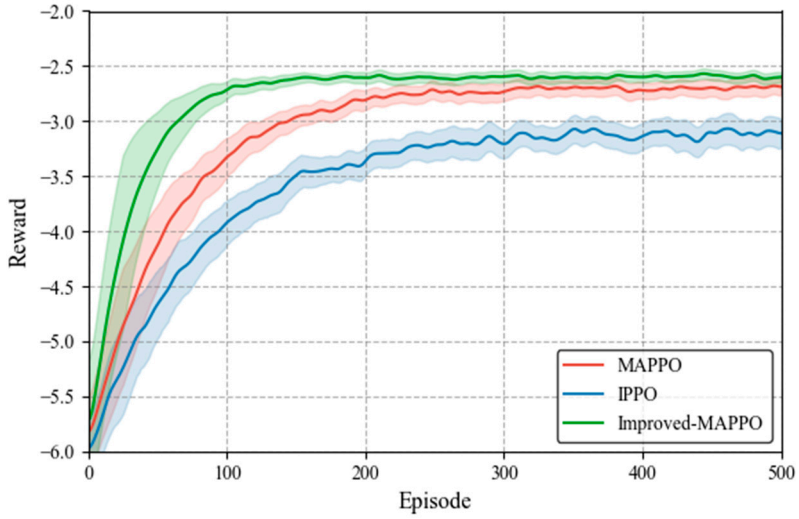

Figure 5. Training reward curves of the three MADRL algorithms.

5.2. Comparative Performance Analysis

Table 3 presents a comparative analysis of operational costs across different control strategies for three representative months (January, April, and August). In general, all MADRL-based approaches demonstrate superior performance compared to the baseline rule-based strategy; however, the magnitude of improvement varies substantially depending on the coordination mechanisms employed by each algorithm.

Table 3. Cost performance comparison of different MADRL-based control strategies.

Model	Metric	January	April	August	Cumulative Total
Baseline	Cost (Yen)	23003.61	14565.58	10201.38	47770.57
	vs. Baseline	-	-	-	-
IPPO	Cost (Yen)	20992.92	12485.34	8678.74	42157
	vs. Baseline	−8.74%	−14.28%	−14.93%	−11.75%
MAPPO	Cost (Yen)	20498.71	12240.92	8017.83	40757.46
	vs. Baseline	−10.89%	−15.96%	−21.40%	−14.68%
Improved-MAPPO	Cost (Yen)	20032.79	12023.48	8178.62	40234.89
	vs. Baseline	−12.91%	−17.45%	−19.83%	−15.77%

IPPO achieves only marginal cost reductions, which reflects the inherent limitations of independent policy learning in multi-agent coordination tasks. The independent policy update mechanism and absence of global information sharing among agents prevent IPPO from effectively capturing the coupled dynamics between the BESS and ASHP subsystems, thereby constraining its optimization capability in complex coupled energy systems. In contrast, MAPPO exhibits markedly superior performance. By leveraging centralized value estimation that facilitates effective inter-agent coordination, this approach achieves substantially larger cost reductions across all three test months, with its cumulative improvement (14.68%) considerably exceeding that of IPPO. These findings corroborate the advantages of centralized training architectures for addressing complex coupled energy systems. The proposed Improved-MAPPO further augments multi-agent cooperative capability and training robustness through the integration of an economic-performance-driven adaptive learning rate mechanism and an action constraint projection scheme. Experimental results demonstrate that this approach consistently achieves optimal performance across all representative months, yielding the highest cumulative cost reduction (15.77%). Notably, the incremental improvement of Improved-MAPPO over standard MAPPO (1.09 percentage points) indicates that the proposed dynamic learning rate modulation strategy and action constraint mechanism effectively enhance both convergence efficiency and policy quality of multi-agent learning in dynamic and complex environments.

To further elucidate the performance disparities among the MADRL methods and their underlying mechanisms, the following subsections provide a systematic evaluation and in-depth analysis of the three MADRL-based control approaches across different test months, with particular focus on three key performance metrics: daily operating cost, grid electricity procurement, and PV self-consumption rate.

5.2.1. Performance Evaluation Under Winter Conditions

Figure 6 presents the performance differences among the four control strategies under typical winter conditions in January. As illustrated in Figure 4, January represents the month with the highest annual energy demand, during which both hot water demand and electricity demand reach their annual peaks, accompanied by the highest electricity price levels. These conditions pose significant challenges to the optimization capability of energy management strategies.

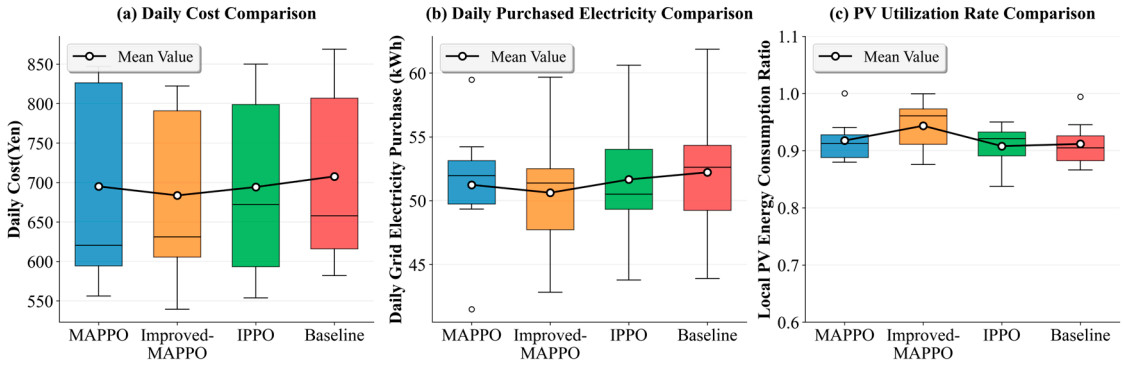

Figure 6. Comparative box plots of daily cost, grid electricity purchase, and PV utilization rate across four control strategies in January.

Analysis of the daily operating cost boxplot morphology reveals that although IPPO successfully reduces part of the unnecessary grid dependence, its independently updated policies restrict cooperative behavior among agents, resulting in higher average daily costs than the two centralized approaches with considerable variability (std = 114.41), and the boxplot displays a notably right-skewed distribution. MAPPO effectively alleviates this issue through centralized value estimation, achieving a more stable cost distribution with significantly narrowed box height. Notably, Improved-MAPPO demonstrates the best overall performance: it attains the lowest average daily cost with the most compact cost distribution, and its median (approximately 672 Yen) closely approximates the mean, indicating that the dynamic learning rate and action-constrained update mechanisms substantially enhance decision stability and operational quality.

The daily electricity purchase comparison further corroborates these observations. The boxplot distribution characteristics reveal that Improved-MAPPO's median closely matches its mean, indicating excellent consistency in scheduling decisions. In contrast, IPPO's median is notably higher than its mean, exhibiting a left-skewed distribution, which suggests that while the algorithm achieves lower electricity purchases on certain operating days, its overall performance lacks stability. Improved-MAPPO achieves the lowest average daily purchase (38.38 kWh), demonstrating its superior capability to exploit the storage system for peak shaving during high-price periods.

The PV self-consumption rate comparison reveals significant differences among strategies in renewable energy integration. Analysis of the boxplot distribution characteristics shows that Improved-MAPPO achieves the highest mean level (94.35%) and an even higher median value (96.07%), with its upper quartile approaching 1.0 and multiple outliers achieving complete absorption, indicating that the algorithm can achieve full local PV consumption on operating days with favorable solar irradiance conditions. Furthermore, Improved-MAPPO's lower box edge (approximately 0.91) is notably higher than that of Baseline (approximately 0.78), demonstrating that even on operating days with poor PV output conditions, the improved algorithm maintains relatively high absorption levels. These results fully demonstrate its robust capacity to sustain high PV utilization across varying weather and load conditions.

5.2.2. Performance Evaluation Under Mid-Seasonal Conditions

Figure 7 presents the box plot comparison results for the four control strategies during April. As illustrated in Figure 4, April represents a transitional season characterized by progressively diminishing heating demands and relatively abundant yet highly variable PV output.

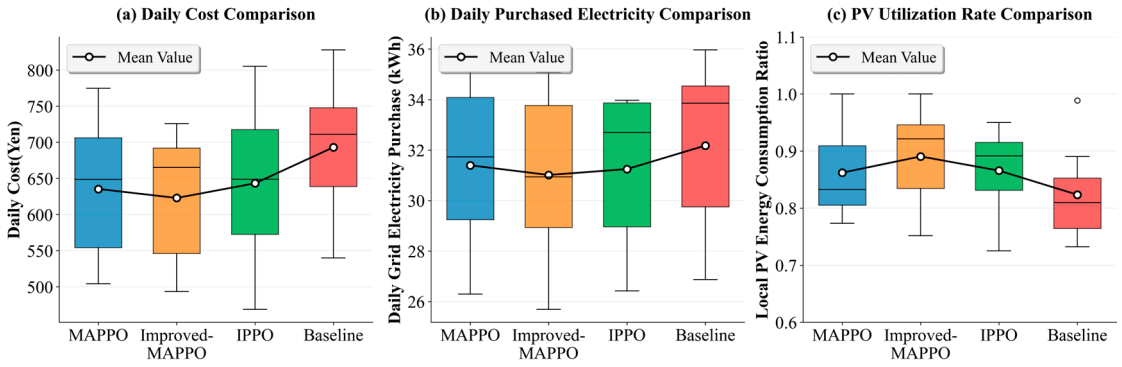

Figure 7. Comparative box plots of daily cost, grid electricity purchase, and PV utilization rate across four control strategies in April.

Analysis of the daily operating cost box plot morphology reveals that Improved-MAPPO exhibits a considerably smaller box height compared to the other three schemes, indicating a more concentrated daily cost distribution with enhanced consistency across different operating days. In contrast, IPPO demonstrates larger whisker spans, suggesting comparatively weaker scheduling decision stability when confronted with the variable PV output conditions characteristic of April.

Regarding daily electricity procurement, all optimization algorithms achieve substantial improvements over the Baseline. The box plot distribution characteristics reveal that Improved-MAPPO's median (30.94 kWh) closely approximates its mean (31.01 kWh), indicating excellent distributional symmetry and superior scheduling stability. Conversely, IPPO's median (32.70 kWh) is considerably higher than its mean (31.23 kWh), exhibiting a right-skewed distribution, which suggests that while the algorithm achieves lower electricity procurement on certain operating days, its overall performance lacks consistency.

The PV utilization rate comparison further corroborates the superiority of Improved-MAPPO in renewable energy integration. The algorithm achieves an average PV utilization rate of 89.06%, significantly outperforming the Baseline (82.35%) and surpassing both MAPPO (86.22%) and IPPO (86.57%). Several salient phenomena can be discerned from the box plot distribution characteristics: firstly, Improved-MAPPO's upper quartile reaches 1.0 with multiple outliers achieving complete absorption, demonstrating that the algorithm can attain full local PV consumption on days with favorable solar irradiance conditions; secondly, Improved-MAPPO's lower box edge (approximately 0.82) is considerably higher than that of the Baseline (approximately 0.73), indicating that even on operating days with suboptimal PV output conditions, the improved algorithm sustains relatively high absorption levels; furthermore, while MAPPO and IPPO exhibit comparable box heights and positions, IPPO's lower whisker extends further downward (minimum 0.73 vs. 0.77), reflecting that independent policy optimization is more susceptible to diminished absorption efficiency under extreme scenarios.

5.2.3. Performance Evaluation Under Summer Conditions

Figure 8 presents the comparison results for the four control strategies during August. As illustrated in Figure 4, the energy consumption characteristics of August are defined by negligible heating demands, dominant cooling loads, and PV output attaining annual peaks with extended daylight hours.

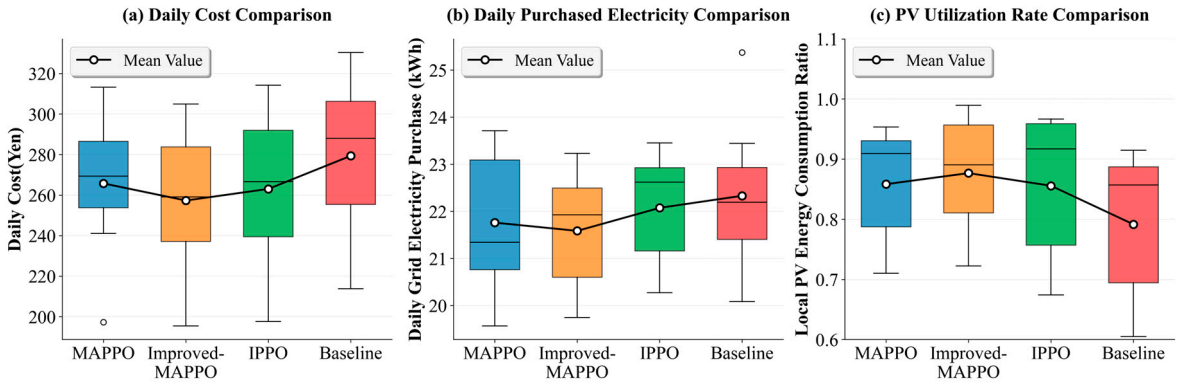

Figure 8. Comparative box plots of daily cost, grid electricity purchase, and PV utilization rate across four control strategies in August.

Compared to April, the daily operating costs across all schemes are substantially lower in August, with narrowed cost differentials among optimization algorithms. This phenomenon is primarily attributable to the absence of heating requirements and abundant PV generation, which consequently diminishes the marginal benefits of multi-agent coordination. Nevertheless, the box plot morphology reveals that Improved-MAPPO occupies the lowest overall position, with its minimum value (195.46 Yen) outperforming all other schemes, demonstrating the improved algorithm's enhanced capability to capture low-cost operating opportunities during summer months.

The daily electricity procurement distribution exhibits patterns distinctly different from April, primarily due to the substantial substitution effect of abundant summer PV generation on grid dependency. Under these conditions, Improved-MAPPO achieves the lowest average daily procurement (21.59 kWh) with the smallest standard deviation (1.35 kWh) among all four schemes, indicating robust scheduling performance under such scenarios.

The PV utilization rate comparison reveals that Improved-MAPPO achieves an average rate of 87.67%, representing an improvement of approximately 10.8% over the Baseline (79.15%). Notably, the overall PV utilization rates in August are marginally lower than those in April, which is attributable to the disappearance of heating loads substantially reducing ASHP operating hours and overall system electricity demand, consequently resulting in PV generation surplus that cannot be fully absorbed during certain periods. Under these circumstances, BESS scheduling strategies become particularly critical. The box plot distribution demonstrates that Improved-MAPPO's upper quartile approaches 0.99 with a maximum value of 0.989, significantly outperforming the Baseline (0.915), indicating that the improved algorithm effectively mitigates summer PV surplus issues through more precise BESS charging and discharging timing control. In contrast, IPPO exhibits the greatest dispersion in PV utilization rate distribution (standard deviation 0.126), with a considerably larger box height than other schemes and a lower whisker extending to 0.67, reflecting the limitations of independent policy optimization in addressing summer PV-load mismatch issues.

5.3. Temporal Energy Flow Analysis

To thoroughly investigate the operational intelligence of the proposed Improved-MAPPO framework, this section presents a detailed analysis of energy flow scheduling patterns during representative weeks across the three test months. Figures 9–11 illustrate the temporal coordination among BESS charging/discharging, ASHP operation, and grid interaction under the Improved-MAPPO control strategy for typical weeks in January, April, and August 2023, respectively.

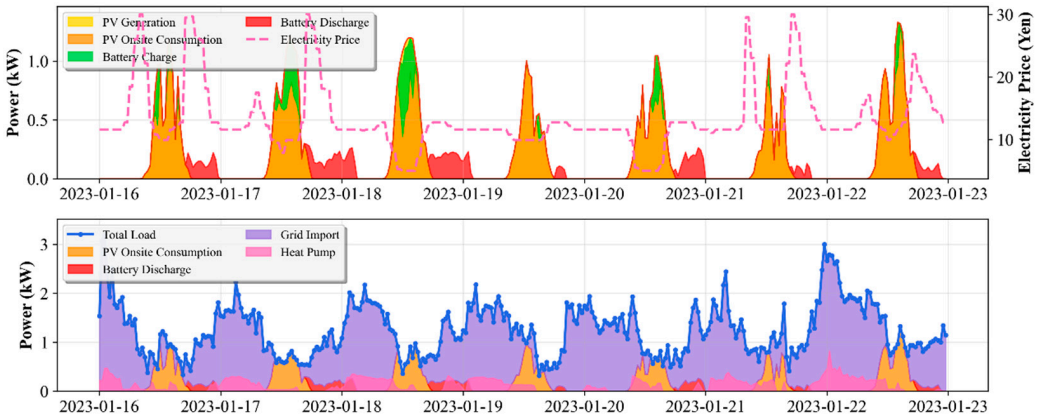

Figure 9. Energy flow scheduling profiles under Improved-MAPPO control strategy during a representative winter week (January 2023).

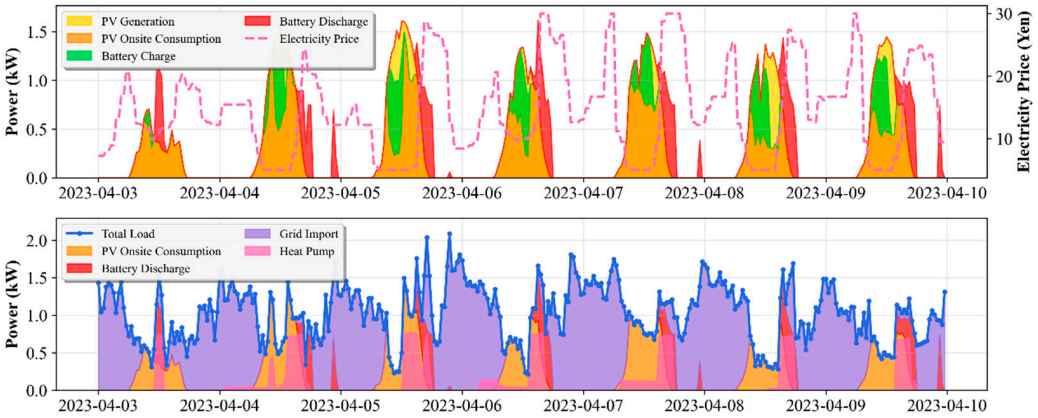

Figure 10. Energy flow scheduling profiles under Improved-MAPPO control strategy during a representative spring week (April 2023).

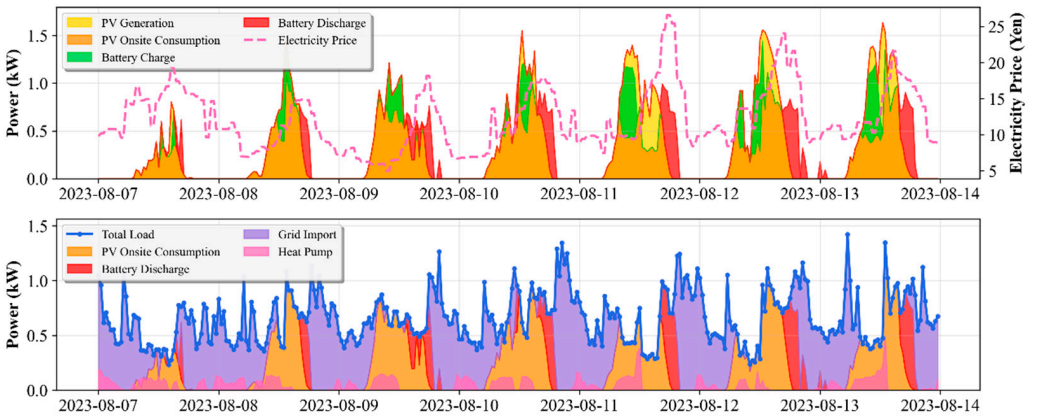

Figure 11. Energy flow scheduling profiles under Improved-MAPPO control strategy during a representative summer week (August 2023).

The winter energy flow profile depicted in Figure 9 demonstrates the agent's sophisticated anticipatory capability in coordinating thermal and electrical loads under challenging conditions characterized by limited PV availability and elevated heating demand. Specifically, the ASHP is strategically scheduled during early morning low-tariff periods (04:00–07:00) to pre-charge the thermal storage tank, thereby establishing sufficient thermal energy reserves prior to the morning hot water demand peak. Regarding BESS operation,

the algorithm exhibits intelligent PV-harvesting behavior by maximizing battery charging during the limited winter daylight hours. The stored energy is subsequently discharged strategically during the evening peak-tariff periods (18:00–21:00), effectively displacing high-cost grid electricity with self-generated renewable energy. These scheduling patterns indicate that the trained agents have successfully learned the temporal correlation between PV generation windows and optimal storage opportunities, while accurately anticipating the evening dual-peak characteristics of electricity prices and household demand. The coordinated operation of BESS and ASHP embodies a holistic energy management strategy that fully exploits the economic value of limited winter PV generation.

The spring energy flow profile illustrated in Figure 10 reveals the algorithm's adaptive response to transitional seasonal characteristics, during which heating demand diminishes substantially while PV output increases significantly but exhibits high day-to-day variability. Under these conditions, the Improved-MAPPO demonstrates enhanced flexibility in coordinating the dual storage systems: compared to winter, the more abundant PV generation enables significant expansion of BESS charging windows, while the reduced heating demand allows for more flexible ASHP operation. The algorithm strategically schedules ASHP operation during peak PV output periods, achieving direct renewable-to-thermal conversion and minimizing grid dependency for heat production.

The summer operational profile presented in Figure 11 exhibits a distinctly different scheduling paradigm, characterized by near-zero heating loads, peak annual PV generation potential, and significantly extended daylight hours. Under these conditions, the primary challenge confronting the Improved-MAPPO lies in formulating refined strategies to maximize renewable energy self-consumption. The BESS plays a pivotal role in addressing the temporal mismatch between PV output and load demand: the algorithm initiates aggressive battery charging during high PV output periods, with the stored energy subsequently discharged during evening hours when PV generation ceases while household demand and electricity prices remain elevated. The minimized ASHP operation reflects the agent's accurate recognition of the seasonal absence of space heating requirements, with thermal storage activated only as necessary to meet residual domestic hot water demand. When ASHP operation is required, it is preferentially scheduled during peak PV output periods to maximize direct renewable energy utilization. This charge-discharge cycle effectively enhances the economic value of self-generated renewable energy through temporal optimization between PV availability and load demand patterns.

Comparative analysis across the three representative weeks reveals consistent behavioral patterns that comprehensively validate the effectiveness of the proposed methodological improvements. Three key conclusions can be drawn from this analysis. First, the action constraint projection mechanism demonstrates stable effectiveness across all seasonal conditions. All BESS operations are strictly maintained within the SOC boundaries (0.2 to 0.9), and ASHP thermal outputs comply with equipment capacity limits, with no instances of thermal storage overflow or deficit observed throughout the evaluation period. Second, the agents exhibit genuine anticipatory intelligence rather than merely reactive responses. The precise temporal alignment between control actions and predictable patterns, including PV generation profiles, electricity price structures, and load demand cycles, indicates that the multi-objective reward function successfully guides the learning process to recognize and exploit these inherent correlations. Third, the economic performance guided dynamic learning rate enables seasonally adaptive strategies without requiring explicit mode switching or seasonal parameter adjustments. The algorithm automatically modulates its operational priorities, shifting from thermal management prioritization in winter to balanced PV harvesting in spring and renewable energy self-consumption maximization in summer. This

seamless seasonal adaptability demonstrates the robust generalization capability of the proposed framework when confronted with substantially different operational contexts.

6. Conclusions

This paper addresses the critical challenges of multi-objective optimization in residential HEMSs by proposing a MADRL framework based on an improved MAPPO algorithm. Through the integration of an action constraint projection mechanism and an economic-performance-driven dynamic learning rate modulation strategy, the proposed framework effectively alleviates the low convergence efficiency issues encountered by conventional MADRL methods in complex coupled BESs. The main conclusions are summarized as follows:

1. The proposed Improved-MAPPO algorithm demonstrates significant advantages in training efficiency. Compared to standard MAPPO, the algorithm achieves a convergence speed improvement of approximately 35–45%, with substantially narrowed confidence intervals during training. These results indicate that the synergistic effect of the dynamic learning rate mechanism and action constraint projection effectively enhances the stability and robustness of multi-agent learning.
2. Regarding operational cost optimization, Improved-MAPPO achieves optimal performance across all three representative test months, delivering a cumulative cost reduction of 15.77% and saving 7535.68 Yen in operating costs compared to the baseline rule-based control strategy. Compared to IPPO (11.75%) and standard MAPPO (14.68%), the cost optimization capability of the proposed method is significantly enhanced, validating the effectiveness of the centralized training architecture and the proposed improvement mechanisms in addressing BESS–ASHP coupled systems.
3. In terms of PV self-consumption rate, Improved-MAPPO achieves the highest renewable energy utilization levels across all seasonal conditions. Boxplot analysis demonstrates that the algorithm achieves nearly 100% complete PV absorption under favorable solar irradiance conditions while maintaining relatively high absorption levels under extreme scenarios, exhibiting robust seasonal adaptability.
4. Temporal energy flow analysis reveals the anticipatory decision-making capability of the Improved-MAPPO agents. The algorithm successfully learns the inherent correlations among PV generation windows, RTP, and load demand cycles, thereby achieving seasonally-adaptive strategies without explicit mode switching.

Building upon the aforementioned findings, future work can be further extended in the following directions: (1) Extending the proposed methodology to multi-building or community-level energy systems to explore hierarchical multi-agent coordination mechanisms based on graph neural networks; (2) Integrating actual electricity market mechanisms to conduct demand response and grid interaction optimization studies, thereby further enhancing the applicability of the algorithm in practical deployment scenarios; (3) Developing scalable agent architectures to adapt to households with different numbers of energy devices without the need to redesign the state space; and (4) Incorporating Automated Fault Detection and Diagnosis (AFDD) modules to enhance system reliability by leveraging recent deep learning advancements in addressing data imbalance and cross-building transferability challenges [35,36].

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

Funding: This work was supported by the Open Fund of the Academic Innovation Center for Coastal Habitat (iSMART), Qingdao University of Technology [Grant Number CK-2024-0062]; and JSPS KAKENHI Grant-in-Aid for Research Activity Start-up [Grant Number 24K23002].

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, and there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled, "Coordinated Scheduling of BESS–ASHP Systems in Zero-Energy Houses Using Multi-Agent Reinforcement Learning".

Nomenclature

Abbreviation	Description	Abbreviation	Description
ASHP	Air Source Heat Pump	MAPPO	Multi-Agent Proximal Policy Optimization
BEMS	Building Energy Management System	MDP	Markov Decision Process
BES	Building Energy System	MPC	Model Predictive Control
BESS	Battery Energy Storage System	PPO	Proximal Policy Optimization
COP	Coefficient of Performance	PV	Photovoltaic
CTCE	Centralized Training and Centralized Execution	RL	Reinforcement Learning
CTDE	Centralized Training and Decentralized Execution	RTP	Real-Time Pricing
DDPG	Deep Deterministic Policy Gradient	SAC	Soft Actor-Critic
DRL	Deep Reinforcement Learning	SOC	State of Charge
DTDE	Decentralized Training and Decentralized Execution	TD3	Twin-Delayed Depth Deterministic Policy Gradient
HEMS	Home Energy Management System	TSK	Thermal Storage Tank
IPPO	Independent Proximal Policy Optimization	ZEH	Zero-Energy House
MADRL	Multi-Agent Deep Reinforcement Learning		

References

1. Tan, J.; Peng, S.; Liu, E. Spatio-temporal distribution and peak prediction of energy consumption and carbon emissions of residential buildings in China. *Appl. Energy* **2024**, *376*, 124330. [CrossRef]
2. Samancioglu, N.; Castaño-Rosa, R.; Väänänen, K.; Komal, M. The role of smart meter feedback in enhancing Inhabitants' energy efficient behaviour in residential buildings. *Energy Build.* **2025**, *351*, 116689. [CrossRef]
3. Zhang, X.; Liu, Y.; Li, Y.; Lv, X.; Xiao, F.; Gao, W. Analyzing variability and coordinated demand management for various flexible integrations of residential distributed energy resources. *Renew. Energy* **2024**, *237*, 121619. [CrossRef]
4. Li, Y.; Jia, Z.; Zhang, X.; Liu, Y.; Xiao, F.; Gao, W.; Xu, Y. Energy flexibility analysis and model predictive control performances of space heating in Japanese zero energy house. *J. Build. Eng.* **2023**, *76*, 107365. [CrossRef]
5. Yip, S.; Athienitis, A.; Lee, B. Analysis of building form influence on energy flexibility in an archetype net zero energy building with building-integrated PV/thermal (BIPV/T) roof for early stage design. *Energy Build.* **2025**, *349*, 116557. [CrossRef]
6. Zhang, X.; Li, Y.; Xiao, F.; Gao, W. Energy efficiency measures towards decarbonizing Japanese residential sector: Techniques, application evidence and future perspectives. *Energy Build.* **2024**, *319*, 114514. [CrossRef]

7. Li, Y.; Zhang, X.; Gao, W.; Xu, W.; Wang, Z. Operational performance and grid-support assessment of distributed flexibility practices among residential prosumers under high PV penetration. *Energy* **2022**, *238*, 121824. [CrossRef]
8. Yu, T.; Chen, X.; Liu, X.; Chen, H.; Tang, S.; Cui, L.; Liu, H.; Niu, K.; Deng, X. Dynamic relationship between renewable energy, economic development, and energy security based on SVAR and ARDL-ECM models: Evidence from China. *Appl. Energy* **2025**, *402*, 126886. [CrossRef]
9. Farheen, M.; Nasreen, S.; Mefteh-Wali, S. The Impact of Renewable and Non-Renewable Energy Aid on Energy Poverty in Developing Asian Countries. *Renew. Energy* **2025**, *257*, 124737. [CrossRef]
10. Ning, W.; Lyu, X.; Liao, P.; Chen, L.; Tao, W.-Q. Performance analysis of a 1-kW PEMFC-CHP system under different rule-based energy management strategies in China. *Renew. Energy* **2025**, *248*, 123110. [CrossRef]
11. Ordoñez, J.G.; Barco-Jiménez, J.; Pantoja, A.; Revelo-Fuelagán, J.; Candelo-Becerra, J.E. Comprehensive analysis of MPC-based energy management strategies for isolated microgrids empowered by storage units and renewable energy sources. *J. Energy Storage* **2024**, *94*, 112127. [CrossRef]
12. Yazar, O.; Coskun, S.; Zhang, F.; Li, L.; Huang, C.; Mei, P.; Karimi, H.R. A novel energy management strategy for hybrid electric vehicles using deep reinforcement incentive learning. *Energy* **2025**, *334*, 137594. [CrossRef]
13. Ahmad, T.; Madonski, R.; Zhang, D.; Huang, C.; Mujeeb, A. Data-driven probabilistic machine learning in sustainable smart energy/smart energy systems: Key developments, challenges, and future research opportunities in the context of smart grid paradigm. *Renew. Sustain. Energy Rev.* **2022**, *160*, 112128. [CrossRef]
14. Xu, Y.; Gao, W.; Li, Y. Cost-effective optimization of the grid-connected residential photovoltaic battery system based on reinforcement learning. *Hum.-Centric Comput. Inf. Sci.* **2024**, *73*, 106774.
15. Li, Y.; Wang, Z.; Xu, W.; Gao, W.; Xu, Y.; Xiao, F. Modeling and energy dynamic control for a ZEH via hybrid model-based deep reinforcement learning. *Energy* **2023**, *277*, 127627. [CrossRef]
16. Qin, H.; Yu, Z.; Li, T.; Liu, X.; Li, L. Energy-efficient heating control for nearly zero energy residential buildings with deep reinforcement learning. *Energy* **2023**, *264*, 126209. [CrossRef]
17. Xu, Y.; Gao, W.; Li, Y.; Xiao, F. Operational optimization for the grid-connected residential photovoltaic-battery system using model-based reinforcement learning. *J. Build. Eng.* **2023**, *73*, 106774. [CrossRef]
18. Ren, K.; Liu, J.; Wu, Z.; Liu, X.; Nie, Y.; Xu, H. A data-driven DRL-based home energy management system optimization framework considering uncertain household parameters. *Appl. Energy* **2024**, *355*, 122258. [CrossRef]
19. Langer, L.; Volling, T. A reinforcement learning approach to home energy management for modulating heat pumps and photovoltaic systems. *Appl. Energy* **2022**, *327*, 120020. [CrossRef]
20. Gao, S.; Xiang, C.; Yu, M.; Tan, K.T.; Lee, T.H. Online Optimal Power Scheduling of a Microgrid via Imitation Learning. *IEEE Trans. Smart Grid* **2022**, *13*, 861–876. [CrossRef]
21. Xu, B.; Luan, W.; Yang, J.; Zhao, B.; Long, C.; Ai, Q.; Xiang, J. Integrated three-stage decentralized scheduling for virtual power plants: A model-assisted multi-agent reinforcement learning method. *Appl. Energy* **2024**, *376*, 123985. [CrossRef]
22. Wang, Z.; Xiao, F.; Ran, Y.; Li, Y.; Xu, Y. Scalable energy management approach of residential hybrid energy system using multi-agent deep reinforcement learning. *Appl. Energy* **2024**, *367*, 123414. [CrossRef]
23. Deng, J.; Wang, X.; Meng, F. A novel safe multi-agent deep reinforcement learning-based method for smart building energy management. *Energy Build.* **2025**, *347*, 116256. [CrossRef]
24. Zhang, Y.; Lin, R.; Mei, Z.; Lyu, M.; Jiang, H.; Xue, Y.; Zhang, J.; Gao, D.W. Interior-point policy optimization based multi-agent deep reinforcement learning method for secure home energy management under various uncertainties. *Appl. Energy* **2024**, *376*, 124155. [CrossRef]
25. Kumari, A.; Kakkar, R.; Tanwar, S.; Garg, D.; Polkowski, Z.; Alqahtani, F.; Tolba, A. Multi-agent-based decentralized residential energy management using Deep Reinforcement Learning. *J. Build. Eng.* **2024**, *87*, 109031. [CrossRef]
26. Deng, J.; Wang, X.; Meng, F. Multi-agent deep reinforcement learning for Smart building energy management with chance constraints. *Energy Build.* **2025**, *331*, 115408. [CrossRef]
27. Yang, Z.; Ren, Z.; Li, H.; Sun, Z.; Feng, J.; Xia, W. A multi-stage stochastic dispatching method for electricity-hydrogen integrated energy systems driven by model and data. *Appl. Energy* **2024**, *371*, 123668. [CrossRef]
28. Yuan, Q. Residential demand response online optimization based on multi-agent deep reinforcement learning. *Electr. Power Syst. Res.* **2024**, *237*, 110987. [CrossRef]
29. Yang, L.; Shi, C.; Sun, Q.; Zhang, N.; Li, Y. Multi-agent deep reinforcement learning-enhanced optimal scheduling for transmission-constrained pelagic island groups considering environmental disturbances. *Electr. Power Syst. Res.* **2025**, *249*, 112053. [CrossRef]
30. Yang, L.; Li, H.; Zhang, H.; Wu, Q.; Cao, X. Stochastic-Distributionally Robust Frequency-Constrained Optimal Planning for an Isolated Microgrid. *IEEE Trans. Sustain. Energy* **2024**, *15*, 2155–2169. [CrossRef]
31. Yang, S.; Lao, K.-W.; Hui, H.; Chen, Y. Secure Distributed Control for Demand Response in Power Systems Against Deception Cyber-Attacks With Arbitrary Patterns. *IEEE Trans. Power Syst.* **2024**, *39*, 7277–7290. [CrossRef]

32. Tian, J.; Jia, H.; Wang, G.; Huang, Q.; Wu, R.; Gao, H.; Liu, C. Optimal scheduling of shared autonomous electric vehicles with multi-agent reinforcement learning: A MAPPO-based approach. *Neurocomputing* **2025**, *622*, 129343. [CrossRef]
33. Cao, J.; Xu, X.; Yao, W.; Xue, F.; Long, C. A distributed dynamic multi-agent reinforcement learning based cooperative game framework for multi-sectoral transaction in electricity markets considering dynamic revenue allocation optimisation and Renewable Energy Certificate trading. *Energy* **2025**, *335*, 138123. [CrossRef]
34. Wang, H.; He, S.; Yan, J.; Han, S.; Liu, Y. Deep reinforcement learning-driven wind farm flow control considering dynamic wind. *Energy Convers. Manag.* **2025**, *337*, 119888. [CrossRef]
35. Wang, S. A hybrid SMOTE and Trans-CWGAN for data imbalance in real operational AHU AFDD: A case study of an auditorium building. *Energy Build.* **2025**, *348*, 116447. [CrossRef]
36. Wang, S. Evaluating cross-building transferability of attention-based automated fault detection and diagnosis for air handling units: Auditorium and hospital case study. *Build. Environ.* **2026**, *287*, 113889. [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.