algorithms

MDPI

Article

A Community Multi-Building Energy Management Method Based on Multi-Head Attention-Enhanced Multi-Agent Proximal Policy Optimization

Xiaoyuan Fu [1], Li Huang [2,*], Weiwei Du [1,3] and Yuqi Jin [1]

[1] North Automatic Control Technology Institute, Taiyuan 030006, China; ysfxy0819@163.com (X.F.); wwd09291026@163.com (W.D.); sxiict2021@163.com (Y.J.)
[2] School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China
[3] Key Laboratory of Information and Control Technology of Shanxi Province, Taiyuan 030006, China
* Correspondence: l_huang@sxu.edu.cn

Abstract

Community multi-building energy management is a key approach for reducing carbon emissions from the building sector and alleviating peak grid pressure. However, load coupling among buildings and coordinated energy-storage operation make control-policy design highly challenging. To address the limitation of the standard multi-agent proximal policy optimization (MAPPO) algorithm, in which the centralized critic simply concatenates building observations and therefore struggles to model inter-building interactions, this paper proposes an improved MAPPO algorithm with a multi-head-attention-enhanced centralized critic, referred to as a multi-head-attention MAPPO (MHA-MAPPO). Without changing the decentralized execution framework, the proposed method improves the critic network in three aspects. First, a dual-branch gated embedding module is designed to adaptively fuse local building observations and global interaction information. Second, an interaction-attention path is constructed to explicitly capture pairwise dependencies among buildings through multi-head attention. Third, a context-attention path is introduced to extract high-level community-level global features by means of learnable query vectors. These improvements enable the critic to estimate the joint-state value more accurately and provide more reliable advantage estimates for all agents. Experiments in the CityLearn environment show that, compared with the original MAPPO, MHA-MAPPO improves the mean evaluation reward by approximately 19.2%, reduces the reward standard deviation by one order of magnitude, and decreases peak net load and total net load by approximately 15.4% and 35.5%, respectively. The results verify the effectiveness of multi-head attention for coordinated multi-building scheduling. The proposed method provides a useful reference for improving multi-agent reinforcement learning algorithms in community energy management.

Keywords: multi-agent reinforcement learning; multi-agent proximal policy optimization (MAPPO); multi-head attention; building energy management; peak-load reduction

Check for updates

Academic Editor: Leijiao Ge

Received: 2 June 2026
Revised: 21 June 2026
Accepted: 22 June 2026
Published: 25 June 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

The building sector occupies a substantial share of global energy consumption and carbon emissions. Building operation accounts for approximately 30–40% of global final energy use, and its share of carbon dioxide emissions is at a similar level [1]. In 2024, operational carbon emissions from civil buildings in China alone reached 2.47 billion tons

of CO_2, accounting for 22.1% of national energy consumption [2]. Caballero-Guereca et al. conducted a systematic review of 212 studies on global building and real-estate decarbonisation and found that, although the ambition of decarbonisation efforts has increased since 2015, most studies remain fragmented, narrow in scope, and small in scale, with insufficient research pathways that are compatible with the temperature-control targets of the Paris Agreement [3]. As global building stock is expected to double by 2060 and the construction sector already contributes 37% of energy-related carbon emissions, rapid decarbonisation of the building sector has become a central issue in global climate-change mitigation [4].

Meanwhile, the wide deployment of distributed energy resources, such as photovoltaic (PV) generation, energy-storage systems, and heat pumps, creates new opportunities for peak shaving, valley filling, and demand response in community multi-building energy management, while also substantially increasing system-control complexity [5]. At the community microgrid level, the intermittency of renewable generation and load-side uncertainty are jointly reshaping coordinated energy management across multiple buildings [6]. Demand response mechanisms for building clusters have made considerable progress in recent years. Energy-flexibility-based load-clustering methods can support customized and predictive renewable-energy accommodation through demand response; nevertheless, the volatility of cluster members under different price signals remains an important challenge [7]. PEDF systems, namely, photovoltaic-storage–direct-current–flexible-load systems, provide an integrated pathway for low-carbon smart-building energy management. However, how to coordinate energy flows among heterogeneous buildings effectively at the community scale remains a key scientific problem [8].

Extensive research has been conducted on the optimal control for building energy systems. Traditional rule-based control strategies, such as start–stop control, threshold-based control, and time schedules, remain widely used in engineering practice because they are simple to implement and easy to interpret [9]. However, Wang et al. showed through comparative experiments that fixed-rule control methods often deviate from optimal operating conditions when facing dynamic load demand and renewable-generation fluctuations, thereby causing unnecessary energy waste [10]. Model predictive control (MPC), as an advanced control method, can optimize system operation trajectories in a rolling manner under constraints and has achieved good performance in scenarios such as heating, ventilation, and air conditioning (HVAC) control and energy-storage dispatch [11]. Liu pointed out that the large-scale deployment of MPC for building thermal management faces two major challenges: equipment heterogeneity and limited operational data under comfort constraints. Although data-driven predictive control can extract system behaviour from historical data, a clear gap remains between open-loop accuracy and closed-loop performance [12]. In broader community-building coordination scenarios, centralized convex MPC can jointly optimize HVAC and battery-storage systems across multiple buildings, but its computational burden increases significantly with system scale, while model calibration is costly and online optimization is difficult [13]. Experiments on machine-learning-enhanced hybrid MPC for demand flexibility in small commercial buildings show that daily HVAC energy costs can be reduced by 8.7%, peak-price-period load can be shifted by 41.7%, and peak demand can be reduced by 29.2%; however, the performance of such methods still depends heavily on prediction-model accuracy and environmental matching [14]. Its practical deployment is still limited by data quality, BAS communication, and standardization issues [15]. Although MPC and its variants have achieved many advances in building energy management, high model-calibration costs, substantial computational requirements, and reliance on expert knowledge limit their large-scale deployment in complex multi-building scenarios [16].

Reinforcement learning (RL), as a trial-and-error sequential decision-making method driven by interaction, has received increasing attention in building energy optimization in recent years [17]. Unlike traditional control methods that require an accurate environmental model, RL agents learn optimal policies through continuous interaction with the environment, allowing them to adapt to dynamic system changes and handle nonlinear characteristics [18]. A review of deep reinforcement learning (DRL) for grid-interactive efficient buildings indicates that DRL has great potential for autonomous energy management, balancing renewable-energy fluctuations, and supporting grid services in the context of building Internet of Things and smart-grid integration, although it still faces challenges such as high algorithmic complexity, strong data dependence, and difficulties in integration with existing infrastructure [19]. In DRL-based operational strategies for residential community microgrid systems, proximal policy optimization (PPO)-based energy management achieved stable real-time convergence in dynamic non-stationary environments, reduced total microgrid costs by 31–58% compared with fixed scheduling, and obtained the highest occupant-comfort reward [20]. A systematic review by Khushalani et al. on intelligent HVAC control in residential buildings showed that DRL, as a model-free alternative, can reduce energy costs by about 15% and reduce comfort violations by up to 98% [21]. Despite the great potential of DRL in building control, Ji et al. warned against excessive hype in this field, noting that most RL studies are still limited to simulation environments, short time scales, and idealized assumptions, while long-term multi-season deployment in real buildings remains rare [22].

Most of the above studies focus on energy management for a single building, whereas coordinated control at the community scale faces additional challenges. If a single-agent RL method is used by concatenating the states of all buildings and controlling them as a whole, the state-action space expands exponentially with the number of buildings, making training much more difficult and limiting scalability to communities of different sizes [23]. If single-agent algorithms are deployed independently for each building without considering coordination among agents, the potential for energy-storage sharing and load complementarity among buildings is ignored, making community-level global optimization difficult [24]. Multi-agent systems have been widely regarded as an effective solution to such complexity because of their decentralization, autonomy, and scalability [25]. The multi-agent deep deterministic policy gradient algorithm proposed by Lowe et al. introduced the centralized-training-and-decentralized-execution (CTDE) framework, in which agents can access global information during training to learn cooperative policies, while relying only on local observations during execution, thereby balancing coordination needs and communication constraints [26]. In a buildings-to-grid generalized energy-storage framework, researchers developed a multi-agent decomposed RL algorithm for delayed rewards in storage systems. By using a dual-critic architecture to redistribute immediate rewards and construct delayed rewards, the method improved building profits by 93.74–104.06% and reduced grid costs by 0.94% [27]. Yu et al. extended proximal policy optimization to multi-agent settings and proposed multi-agent proximal policy optimization (MAPPO). Their systematic evaluation in benchmark tasks such as multi-agent particle environments and the StarCraft Multi-Agent Challenge showed that MAPPO can achieve comparable or even better performance than state-of-the-art algorithms in multiple tasks, with relatively low hyperparameter sensitivity [28]. In energy systems, MAPPO has recently been extended to coordinated adaptive control of photovoltaic inverter clusters in active distribution networks, reducing voltage deviation by more than 40%, network loss by about 25%, and photovoltaic curtailment by nearly 50% compared with centralized optimization methods [29]. In mixed residential and public-building communities, MAPPO-based distributed energy-storage management can cope effectively with supply-demand uncertainty and

reduce total operating cost and peak load at the community level [30]. Nevertheless, the centralized critic in standard MAPPO usually directly concatenates all agents' observation vectors and feeds them into a multilayer perceptron. This simple concatenation mixes all building features in one vector and cannot explicitly model inter-building interaction dependencies, which limits performance in strongly coupled tasks such as multi-building energy management.

To address these issues, this paper proposes an improved MAPPO algorithm with a multi-head-attention-enhanced centralized critic, denoted as MHA-MAPPO. The proposed method first constructs local embeddings and global interaction embeddings for each building and uses a gating mechanism to adaptively fuse the two types of features, enabling the network to dynamically balance building-specific information and neighbourhood-interaction information according to the current state. On this basis, a dual-path attention encoding structure is used to model inter-building dependencies and global contextual information separately. In this structure, the interaction-attention path directly calculates pairwise influence relationships among buildings through multi-head attention, while the context-attention path uses learnable query vectors to extract high-level global community features. The design aims to improve the critic's value-estimation capability for coordinated community load regulation, thereby guiding building agents to learn better energy-storage scheduling policies. Experiments are conducted in the CityLearn environment. Training convergence curves, evaluation reward, peak net load, total net load, load smoothing, and other indicators are used to evaluate the improved algorithm, and ablation experiments are designed to analyze the contribution of each module. This study is expected to provide an effective reference for improving multi-agent reinforcement learning in community energy management.

In this setting, the central research problem is how to learn coordinated storage-control policies for heterogeneous buildings when local loads, PV output, electricity prices, and storage states are coupled in time and across buildings. The core research question addressed in this paper is whether a centralized critic that explicitly models inter-building dependencies can provide more accurate value estimates than the simple feature concatenation-based critic used in standard MAPPO. The contribution of this work is therefore not a new execution architecture, but a critic design that maintains the decentralized execution while making the training signal more aware of community-level energy interactions.

The significance and main contributions of this research to the field of the built environment can be summarized as follows:

- This study addresses community-scale building energy management, which is a key problem for low-carbon built environments, peak-load reduction, and grid-interactive efficient buildings.
- The proposed MHA-MAPPO method improves multi-agent reinforcement learning for heterogeneous building clusters by explicitly modelling inter-building energy interactions while maintaining decentralized execution.
- The multi-head-attention-enhanced centralized critic provides a scalable way to capture load coupling, photovoltaic uncertainty, and coordinated energy-storage behaviour among multiple buildings.
- Experimental results show practical benefits for built-environment operation, including improved evaluation reward, reduced peak net load, lower total net load, and smoother community-level load profiles.
- The ablation analysis clarifies the roles of gated feature fusion, interaction attention, and dual-path encoding, offering methodological guidance for future AI-based control strategies in smart communities and building-to-grid applications.

2. Data Construction and Methodology

This section describes the methodological pipeline used in the study. It begins with the CityLearn data and observation construction, then reviews the standard MAPPO framework, identifies the limitations of its centralized critic in multi-building control, and finally provides the basis for the MHA-MAPPO design developed in Section 3.

2.1. CityLearn Environment and Dataset

The experiments use the CityLearn dataset, which contains real load and photovoltaic-generation data for multiple buildings. The dataset is collected from actual building operation records. The load curves exhibit typical daily and seasonal patterns, while photovoltaic output is significantly affected by weather conditions and sunshine duration. In the dataset, the mean building load is approximately 45.2 kW, the standard deviation is 38.6 kW, and the maximum peak value is 312.5 kW. The mean photovoltaic output is approximately 32.7 kW; because of intermittent weather effects, its standard deviation reaches 51.3 kW, which is higher than the load fluctuation amplitude, and the maximum value is approximately 287.4 kW.

During data preprocessing, building load, meteorological variables, electricity prices, and carbon-emission-intensity series are first read in a unified manner, and time-step alignment is established to ensure consistency in sampling interval, start and end time, and building index. This avoids misalignment between training samples and environmental states. Next, missing values, infinite values, and outliers in observation vectors are checked and replaced with finite floating-point values, after which all observations are converted to float32 format. For features with substantially different units and scales, such as load, photovoltaic output, temperature, electricity price, and carbon intensity, standardization or normalization is performed so that different features are placed on comparable numerical scales, thereby reducing the dominance of large-scale variables in neural-network gradient updates.

For state construction, each building is treated as an agent and its local observation is retained as the actor input. At the same time, the observations of all buildings are concatenated in a fixed order into a joint state, which is used as the input of the centralized critic. This design preserves local information required for decentralized execution while allowing the critic to learn inter-building coupling relationships during training. Each episode uses a time window of the same length, and the evaluation stage uses the same time-step setting. For action processing, the policy network outputs standardized actions, which are clipped to the interval $[-1, 1]$ and then mapped to the real control space according to the lower and upper bounds provided by the environment, ensuring that energy-storage charging and discharging actions satisfy physical constraints.

Figure 1 shows the temporal variation in raw series, including building-side load, photovoltaic output, cooling-load demand, outdoor temperature, electricity price, and carbon-emission intensity. Several key characteristics can be observed. First, the building load exhibits clear daily periodic fluctuations, with distinct morning and evening peaks and night-time valleys, while cooling-load demand increases significantly in summer and is positively correlated with temperature. Second, photovoltaic output exhibits a daily pattern with high values during the day and zero output at night, but its fluctuation amplitude is strongly affected by weather conditions and therefore shows clear intermittency and uncertainty. Third, electricity-price signals and carbon-emission intensity show temporal correlation, and peak-price periods largely coincide with load peaks, providing explicit economic incentives for energy-storage scheduling. The alignment check of the multi-source time series confirms temporal consistency among variables and lays a data foundation for subsequent model training.

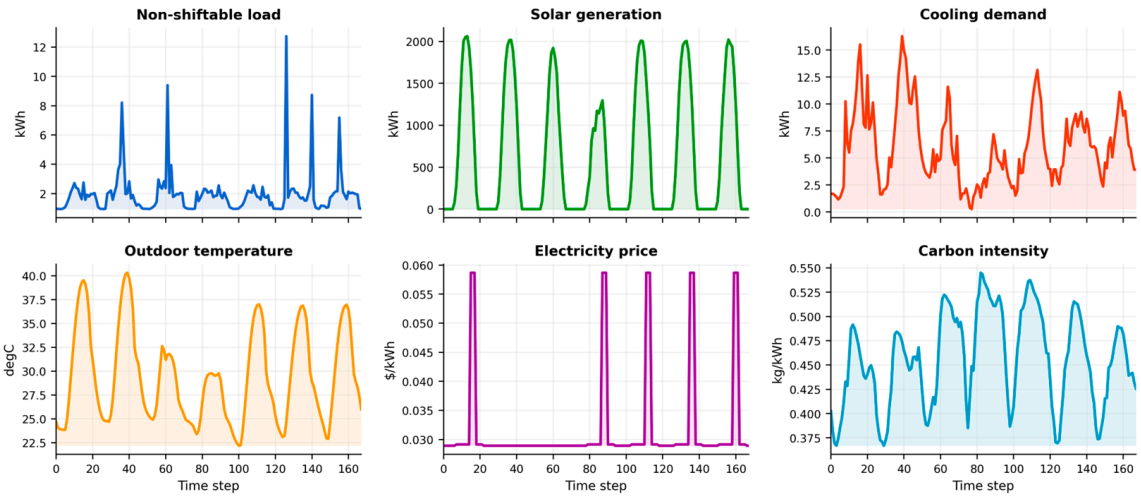

Figure 1. Multi-source time-series inspection.

However, Figure 1 also shows that the numerical ranges of different features vary greatly. Load can reach several hundred kilowatts, electricity price is generally around 0–0.5 USD/kWh, and temperature ranges from −10 to 40 degrees Celsius. Without processing, such scale differences cause gradient updates during neural-network training to be dominated by large-scale variables and weaken the contribution of small-scale features. Therefore, before multi-source observations are fed into the model, it is necessary to standardize different feature types so that they are on similar scales. Figure 2 uses representative features to compare distributions before and after standardization. The left side shows the original distributions, where electricity price, load, and temperature differ substantially in magnitude. The right side shows the distributions after standardization, in which all features are compressed to comparable scales, with a mean of 0 and a standard deviation of about 1. This processing effectively removes scale effects, improves training stability, accelerates convergence, and prevents the model from relying excessively on a single type of feature.

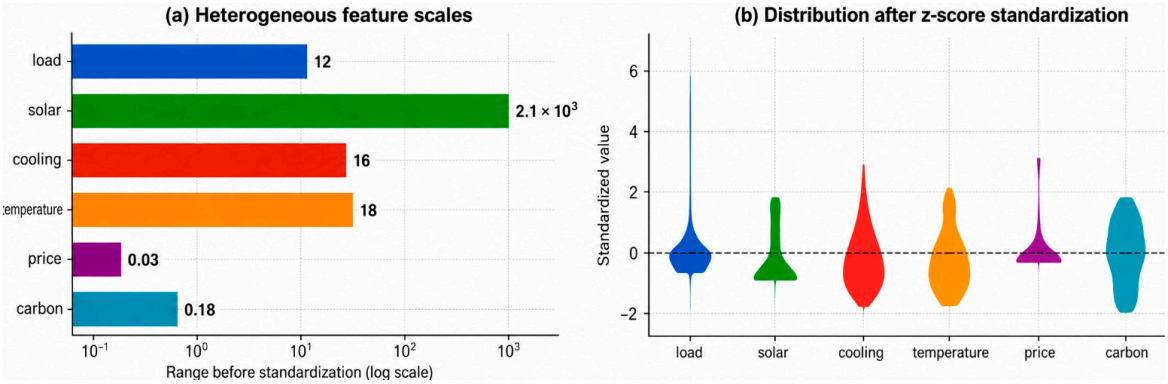

Figure 2. Illustration of feature-scale normalization.

2.2. Standard MAPPO Algorithm Framework

The proposed MHA-MAPPO algorithm is based on the classical MAPPO framework and strictly follows the CTDE paradigm commonly used in multi-agent cooperative control. This paradigm contains two stages. During training, a centralized value network is optimized using the community joint state formed by concatenating all building states. During execution, each building agent makes decentralized energy-storage charging and discharging decisions only from its own local observation, without the real-time exchange of global information. MAPPO consists mainly of two types of networks. The first is the

distributed actor network. Each building has an independent policy network that takes only the building's local observation as input and outputs the policy-distribution parameters for energy-storage control actions. The actor parameters of different buildings are independent so as to adapt to heterogeneity in load level, storage capacity, and photovoltaic configuration. The second is the centralized critic network. All buildings share a global value network that takes the community joint state as input and outputs the value estimate of the current global state. The critic is used only during centralized training for advantage estimation and policy updating.

For policy optimization, MAPPO introduces the proximal policy clipping mechanism. For the i-th building agent, the optimization objective of its actor network is the clipped surrogate loss:

$$L_{clip}(\theta_i) = E_t[\min(r_{i,t}(\theta_i) \cdot A_t, clip(r_{i,t}(\theta_i), 1-\varepsilon, 1+\varepsilon) \cdot A_t)] \quad (1)$$

where $r_{i,t}(\theta_i) = \frac{\pi_{\theta_i}(a_{i,t}|o_{i,t})}{\pi_{\theta_{i,old}}(a_{i,t}|o_{i,t})}$ is the action probability ratio of the new and old policies of the ith building agent, which is used to measure the range of the strategy update. ε is the clipping threshold, which limits the maximum step size of the policy update to ensure that the difference between the new policy and the old policy is kept in a controllable range. $clip(\cdot)$ is the clipping function. When the probability ratio exceeds the interval $[1-\varepsilon, 1+\varepsilon]$, it is clipped to the threshold boundary to avoid the training instability caused by excessive update of the strategy. A_t is the global Advantage function, which is used to measure the degree of the current joint action relative to the global average level. Based on the value estimation of the centralized critic network, and calculated by the Generalized Advantage Estimation (GAE) method, the calculation formula is as follows: the expectation is taken over sampled trajectories; subscript i denotes the building agent, and the clipping threshold is set to 0.2 in the experiments.

$$A_t = \delta_t + (\gamma\lambda)\delta_{t+1} + (\gamma\lambda)^2\delta_{t+2} + \ldots + (\gamma\lambda)^{T-t-1}\delta_{T-1} \quad (2)$$

where $\delta_t = r_t + \gamma V_\phi(s_{t+1}) - V_\phi(s_t)$ is the temporal difference error and $\lambda \in [0,1]$ is the smoothing parameter of GAE, which is used to balance the bias and variance of the advantage estimation. The temporal index t denotes the current control step, and the discount factor and GAE coefficient are set to 0.99 and 0.95, respectively.

The centralized critic network is optimized using a mean-square-error loss function. By minimizing the deviation between the value estimate and the target return, the accuracy of value estimation is improved. The loss function is defined as:

$$L_{critic}(\phi) = E_t[(V_\phi(s_t) - \hat{R}_t)^2] \quad (3)$$

where $\hat{R}_t$ is the cumulative discounted return of the trajectory. The final joint loss function of MAPPO is the weighted sum of the clipped objective function, the value loss and the entropy regularization term: the return and the value estimate are scalar community-level quantities at each time step.

$$L(\theta, \phi) = L^{CLIP}(\theta) - c_1 L^{VF}(\phi) + c_2 S(\pi_\theta) \quad (4)$$

where $S(\pi_\theta)$ is the policy entropy regularization term, which is used to improve the policy exploration ability. The loss weights balance policy improvement, value fitting, and exploration during each MAPPO update.

2.3. Limitations of the Centralized Critic in MAPPO

Although MAPPO performs well in many multi-agent tasks, its centralized critic has three key problems when directly applied to community multi-building energy management. First, the centralized critic usually feeds a simple concatenation of building observations into a multilayer perceptron, treating mixed features uniformly and making it difficult to explicitly model dynamic coupling and load complementarity among buildings. Second, buildings in a community are heterogeneous, such as residential buildings, offices, and commercial buildings; their observation spaces and operating characteristics differ significantly, and simple concatenation cannot adaptively fuse heterogeneous features. Third, the relationship between global value estimation and the marginal contribution of each building is unclear, leading to inaccurate advantage estimation and reduced policy-update efficiency. These problems limit the performance of standard MAPPO in coordinated multi-building scheduling. Therefore, while maintaining full compatibility with the original MAPPO training framework, the next section improves the centralized critic and proposes MHA-MAPPO.

3. Multi-Building Energy Management Method Based on MHA-MAPPO

To address the three key problems identified in Section 2.3, namely, difficulty in modelling inter-building coupling, insufficient fusion of heterogeneous features, and unclear global credit assignment, this section proposes the MHA-MAPPO algorithm. The algorithm keeps the training framework, actor structure, and policy-update method of standard MAPPO unchanged, and only reconstructs the centralized critic. The improved critic is called the multi-head-attention-enhanced critic, or MHA-critic, and the overall algorithm is therefore named MHA-MAPPO. MHA-critic solves the above problems through three modular structural innovations, and its overall architecture is shown in Figure 3.

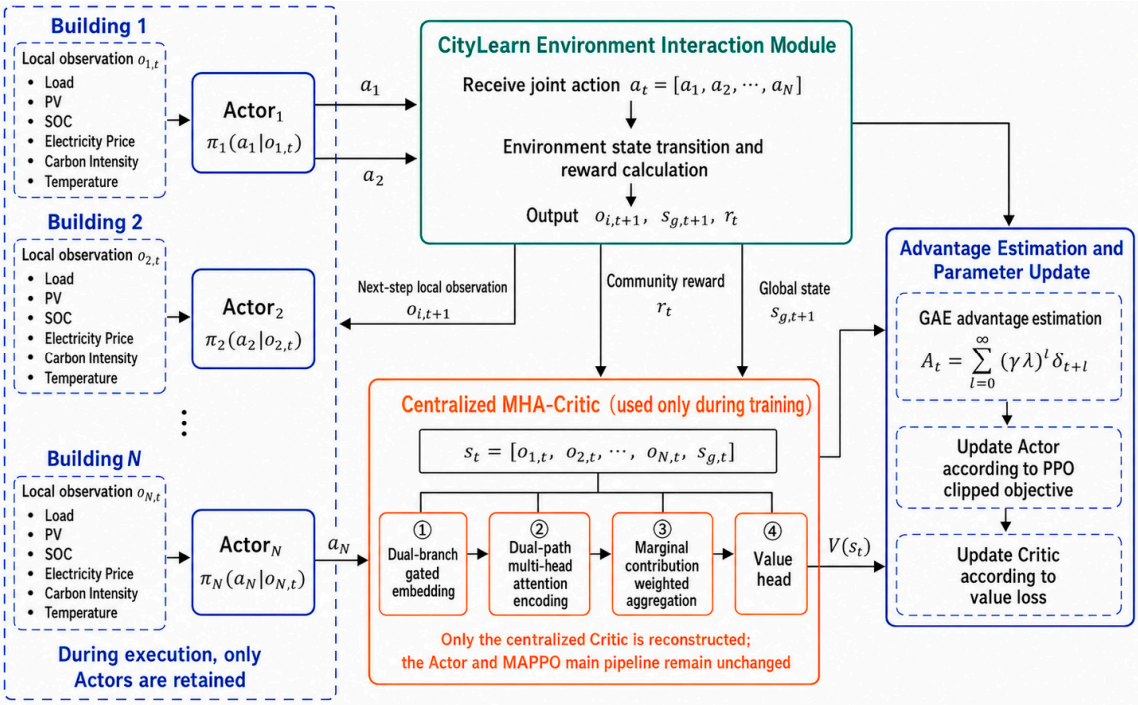

Figure 3. Overall workflow of the MHA-MAPPO algorithm.

First, to address insufficient heterogeneous-feature fusion, a dual-branch gated embedding module is designed, as described in Section 3.1. This module decomposes building observations into general state features and type-specific functional features, and adap-

tively fuses the two kinds of information through learnable gating weights. As a result, heterogeneous features of different building types are aligned in a unified semantic space, avoiding the semantic confusion and dimensional dilution caused by simple concatenation.

Second, to address the difficulty of modelling inter-building coupling, a dual-path multi-head attention encoding structure is designed, as described in Section 3.2. The structure contains two parallel attention paths. The interaction-attention path explicitly captures pairwise dependencies among buildings through multi-head attention and quantifies the mutual influence of different buildings in community energy coordination. The context-attention path uses learnable query vectors to extract high-level global community features. The outputs of the two paths are fused to form an enhanced representation that considers both local interactions and the global situation.

Third, to address unclear global credit assignment, a marginal-contribution weighted aggregation module is designed, as described in Section 3.3. Based on the weight matrix output by the interaction-attention path, the module quantifies each building's marginal contribution to the global community value and uses contribution scores as weights to aggregate encoded features. Buildings that have greater influence on the global value therefore receive higher weights in the final state representation, enabling more reasonable credit assignment.

These three improvements allow MHA-critic to estimate the joint-state value more accurately and provide more reliable advantage estimates for all agents. At the same time, the method remains fully compatible with the original MAPPO training framework. The global state-value estimate output by MHA-critic can be directly used in MAPPO advantage calculation and value-loss optimization without modifying the actor network, the main algorithmic process, or the hyperparameter settings. The detailed design of the three improvements, the reward function, and the training procedure are introduced below.

3.1. Dual-Branch Gated Embedding Network

In heterogeneous multi-building scenarios, the raw observation spaces of different buildings differ significantly. Feeding them directly into the attention layer causes inconsistent feature distributions and affects the learning of attention weights. Therefore, based on building-type-specific embeddings, this paper designs a general-specific dual-branch gated embedding network. The module removes feature-dimensional differences while retaining building-type information and operational attributes.

For the i-th building agent, the input feature vector consists of four parts: local observation features, such as real-time environmental and operational information obtained from building sensors, including load, photovoltaic output, outdoor temperature, and electricity price; building-type one-hot encoding, which represents the fixed category of the building, such as residential, office, educational, or commercial; operation-mode encoding, which represents the current community operating context or temporal attributes, such as weekday/holiday and peak/off-peak periods; and dynamic attribute features, which represent real-time building states, including energy-storage state of charge, renewable-energy utilization, net load, and action-related states.

The above input features are further decomposed into two types of sub-features. The first type consists of general state features shared by all buildings, including cross-building common attributes such as net load, photovoltaic output, electricity price, carbon intensity, and storage state. Their dimensionality is fixed and independent of building category. The second type consists of building-type-specific functional features, such as the more pronounced evening load fluctuations of residential buildings and the more regular working-hour energy-use patterns of office buildings. The dimensions and semantics of these type-specific features can differ greatly across building categories.

Traditional fully connected embedding layers apply an undifferentiated unified mapping to heterogeneous features, which can bury low-dimensional features and cause semantic confusion. This paper therefore designs a dual-branch gated embedding structure to encode the two types of features separately and fuse them adaptively. The overall structure is shown in Figure 4.

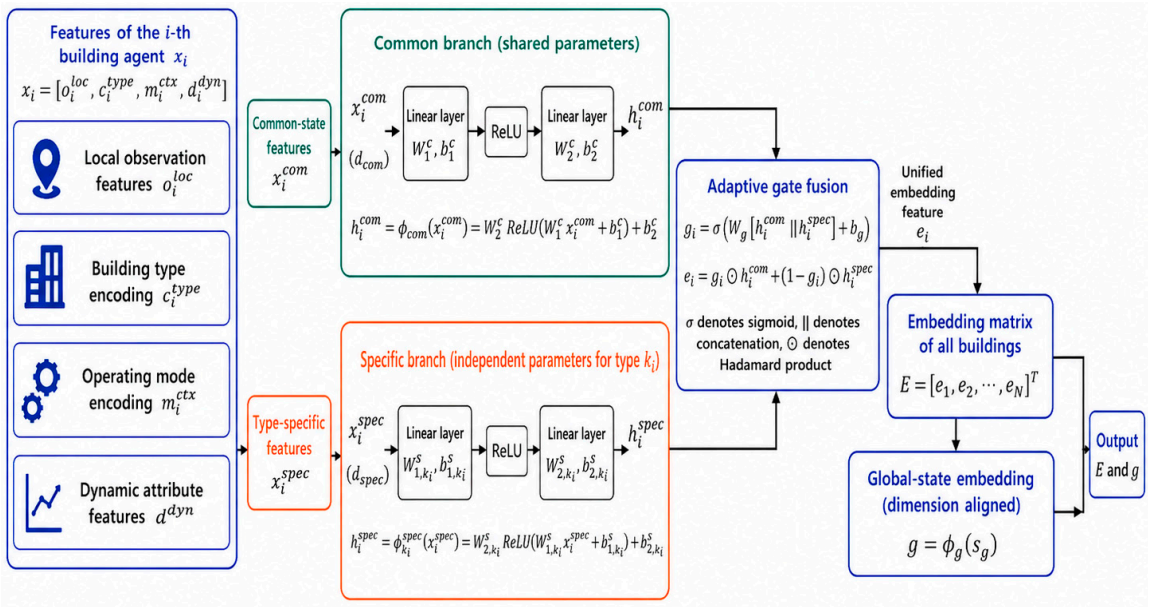

Figure 4. Workflow of the dual-branch gated embedding design.

The complete mathematical implementation of the module is as follows.

First, the general-feature branch shares the same mapping parameters across all buildings and learns general attribute representations across building categories, thereby ensuring alignment among different building types in the basic semantic space:

$$f_{i,t}^{common} = LayerNorm(\mathrm{ReLU}(W_{common} x_{i,t}^{common} + b_{common})) \quad (5)$$

where W_{common} and b_{common} are learnable parameters shared by the whole network, and the output features are uniformly mapped to the embedding space of dimension d_{emb}, independent of the functional category of the agent. Here, the general feature vector contains variables shared by all building types, and its mapped embedding has dimension h.

Next, the type-specific feature branch assigns an independent two-layer fully connected embedding network to each building type. This branch learns type-specific operating characteristics and avoids semantic conflicts and dimensional mismatch when heterogeneous features are mapped into a unified space:

$$f_{i,t}^{spec} = LayerNorm(\mathrm{ReLU}(W_{k_i}^{spec} x_{i,t}^{spec} + b_{k_i}^{spec})) \quad (6)$$

where $k_i \in \{1, 2, \ldots, K\}$ is the building type number to which the ith building belongs, and K is the total number of building types in the community, which can be flexibly configured according to specific scenes. $W_{k_i}^{spec}$ and $b_{k_i}^{spec}$ are the learnable parameters of the exclusive branch of the k_i-type building. The ReLU activation function is used to map the exclusive features of different dimensions and different semantics into the same embedding space. The superscript for this branch denotes type-specific features, and the category index selects the corresponding embedding network.

Then, adaptive gated fusion uses learnable gating weights to balance the proportions of general features and type-specific features in the final representation. This achieves the

dynamic and precise alignment of heterogeneous features and avoids representation bias caused by fixed fusion weights:

$$g_{i,t} = \text{Sigmoid}(W_G \times \text{Concat}(f_{i,t}^{common}, f_{i,t}^{spec}) + b_g) \tag{7}$$

$$f_{i,t} = g_{i,t} \odot f_{i,t}^{common} + (1 - g_{i,t}) \odot f_{i,t}^{spec} \tag{8}$$

where $g_{i,t} \in [0, 1]$ is the gating weight and $\odot$ is the Hadamard product. The gating mechanism can dynamically adjust the fusion weights of the two types of features according to the building type and the current running state. For example, in the high-temperature and high-load period, the weight of the specific features of the buildings with a high air-conditioning load will be automatically increased. In the noon period, when the photovoltaic output is sufficient, the relevant features of the building with strong PV-storage coordination capabilities will also obtain higher weights. The fused embedding is computed element by element, so the general and type-specific embeddings must have the same dimension.

After feature embedding is completed for all agents, a unified embedding feature matrix is obtained:

$$F_t = [f_{1,t}, f_{2,t}, \ldots, f_{N,t}]^T \in \mathbb{R}^{N \times d_{emb}} \tag{9}$$

where N is the total number of agents. At the same time, the global environment state s_t is mapped to the global context feature $f_t^{global} \in \mathbb{R}^{N \times d_{emb}}$ with the same dimension as the agent feature through the independent global state embedding network to ensure the consistency of the feature space. The embedding matrix stacks the N building embeddings in a fixed order, while the global state represents the community-level environment at the same time step.

3.2. Dual-Path Multi-Head Attention Mechanism

The core advantage of multi-head self-attention is that multiple attention heads can capture feature dependencies from different representation subspaces in parallel and adaptively learn the association weights among input elements. This is well-suited to dynamic interaction modelling in cooperative scenarios with heterogeneous agents. Traditional single-path self-attention can model only one type of feature dependency and cannot simultaneously consider global cooperative relationships and the importance of individual features in multi-agent scenarios. Therefore, this paper designs a dual-path parallel attention encoding structure. The two paths share the same input feature matrix and encode information from different dimensions. The complete structure is shown in Figure 5.

The global-situation feature and the embedding features of all agents are first concatenated to form the input matrix of the multi-head attention layer, ensuring that global-situation information and individual-agent features are input simultaneously:

$$X_t = [f_t^{global}; F_t] \in \mathbb{R}^{(N+1) \times d_{emb}} \tag{10}$$

The first row of the matrix contains the global context features, and the subsequent row N corresponds to the embedded features of the N agent.

The dual-path attention module uses scaled dot-product attention as its basic computational unit. Its strict mathematical form is

$$Attention(Q, K, V, M) = softmax(\frac{Q_k K_k^T}{\sqrt{d_k}} + M)V \tag{11}$$

where Q, K, V are the query matrix, key matrix and value matrix, respectively, and d_k is the key dimension of single-head attention. $\sqrt{d_k}$ is the scaling factor used to avoid the disappearance of the gradient of the softmax function due to the large value of the inner

product. M is the mask matrix, which can mask invalid dependencies through assignment and realize the filtering of irrelevant information. The query, key, and value matrices are obtained from the same input representation through learnable linear projections.

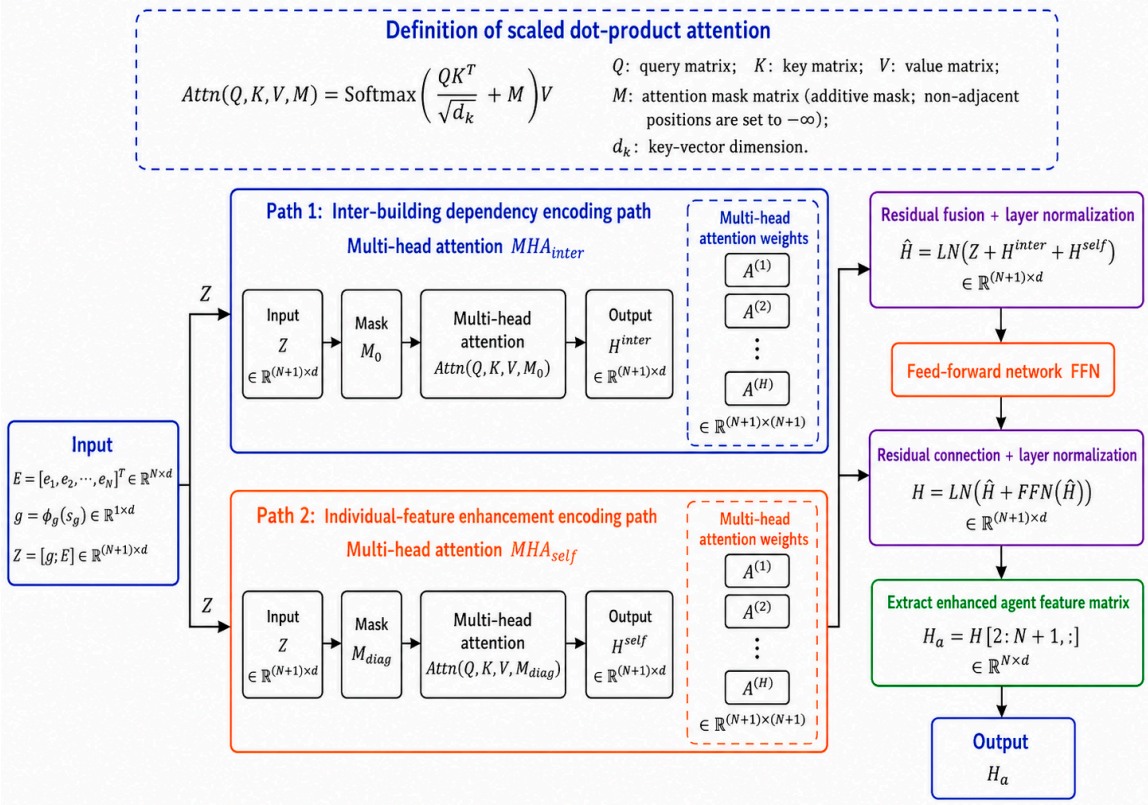

Figure 5. Workflow of the dual-path multi-head attention mechanism.

The total number of attention heads can be adjusted according to the number of buildings and the complexity of the cooperative pattern. Each attention head independently learns one type of interaction dependency, enabling the decoupled modelling of multi-dimensional building cooperation and adaptation to complex load complementarity and storage coordination among heterogeneous buildings.

Path 1 is the inter-building interaction-dependency encoding path. It models dynamic cooperative relationships among multiple buildings and quantifies the contribution of different buildings to community energy management, thereby providing a basis for subsequent credit assignment. Its multi-head attention calculation is

$$MultiHead_{coop}(X_t) = Concat(head_1, head_2, \ldots, head_h)W^{coop} \tag{12}$$

$$head_m = Attention(X_tW_m^Q, X_tW_m^K, X_tW_m^V, M_0), m = 1,2,\ldots,h \tag{13}$$

where W_m^Q, W_m^K, $W_m^V \in \mathbb{R}^{d_{emb}\times d_k}$ is the learnable linear projection parameter of the h-the attention head, and $d_k = \frac{d_{emb}}{h}$ is the feature dimension of the single head. This setting ensures that the total parameter quantity of the multi-head attention is consistent with the single-head attention. M_0 is a zero mask matrix, which does not pre-restrict the dependencies between buildings, so that the model can adaptively learn the dynamic interactive structure of multiple buildings that is difficult to capture using the fully connected layer. The attention weight matrix output by this path can directly represent the dependence strength between buildings, and the elements of the matrix correspond to the degree of collaboration between two buildings. The larger the weight value, the stronger the influence

of the corresponding building on the global decision of the community. The head index ranges from 1 to H, where H is the number of attention heads used in the critic.

Path 2 is the individual-feature enhancement path. It models the importance of each building's own feature dimensions, adaptively enhances key operating features that have greater influence on the global value, filters noise and redundant information in partially observable scenarios, and improves the robustness of value estimation. Its calculation is

$$MultiHead_{self}(X_t) = Concat(head_1, head_2, \ldots, head_h)W^{self} \tag{14}$$

$$head_m = Attention(X_t W_m^{Qs}, X_t W_m^{Ks}, X_t W_m^{Vs}, M_{diag}), m = 1, 2, \ldots, h \tag{15}$$

In the formula, $d_k = \frac{d_{emb}}{h}$ is the diagonal mask matrix, and only the elements on the diagonal of the matrix are retained, forcing the attention to only focus on the feature dimension of the building itself, and realizing the adaptive screening and enhancement of the importance of individual features. For example, peak hour net load, energy-storage SOC or PV utilization-related features can be automatically enhanced to suppress the interference caused by noisy observations on value estimation. The diagonal mask keeps the self-feature relations in this path and suppresses unrelated cross-feature terms.

The output features of the two paths are fused by element-wise addition. A residual connection and layer normalization are then introduced to alleviate gradient disappearance in deep-network training and avoid feature-distribution shifts. The final encoded matrix fuses global cooperative dependencies and individual key features:

$$X_t^{att} = LayerNorm(X_t + MultiHead_{coop}(X_t) + MultiHead_{self}(X_t)) \tag{16}$$

A two-layer feed-forward neural network (FFN) is then used to perform further nonlinear transformation and mine the deeper operating relationships among features. Residual connection and layer normalization are introduced again to obtain the final enhanced feature matrix:

$$X_t^{ffn} = FFN(X_t^{att}) = W_2 \cdot \mathrm{Re}LU(W_1 \cdot X_t^{att} + b_1) + b_2 \tag{17}$$

$$F_t^{att} = LayerNorm(X_t^{ffn}[1:N,:] + X_t^{att}[1:N,:]) \tag{18}$$

In the formula, the dimension of the middle layer of $W_1 \in \mathbb{R}^{d_{\mathrm{model}} \times 4d_{\mathrm{model}}}$, $W_2 \in \mathbb{R}^{4d_{\mathrm{model}} \times d_{\mathrm{model}}}$ is set to d_{model}, which conforms to the classical configuration of the mainstream Transformer architecture and ensures the feature expression ability of the network. $F_t^{att} \in \mathbb{R}^{N \times d_{emb}}$ is the enhanced feature matrix of the agent after eliminating the global context feature. The feed-forward network is applied to each encoded building representation before the global feature is separated.

3.3. Marginal-Contribution Weighted Aggregation Strategy

Traditional critic networks generally use average pooling or max pooling to aggregate building encoding features globally. Average pooling treats all features equally and can dilute key information from high-contribution buildings, leading to inaccurate credit assignment. Max pooling preserves only extreme feature values and loses substantial cooperative information, which can bias global value estimation. In multi-building energy management, this limitation directly weakens the representation of key building cooperation and affects community-level global optimization.

To address this problem, this paper designs an attention-weight-based marginal-contribution weighted aggregation strategy. The method uses the attention weights output by the cooperative-dependency encoding path to quantify each agent's marginal contri-

bution to the global value and uses these weights to aggregate features. The complete structure is shown in Figure 6.

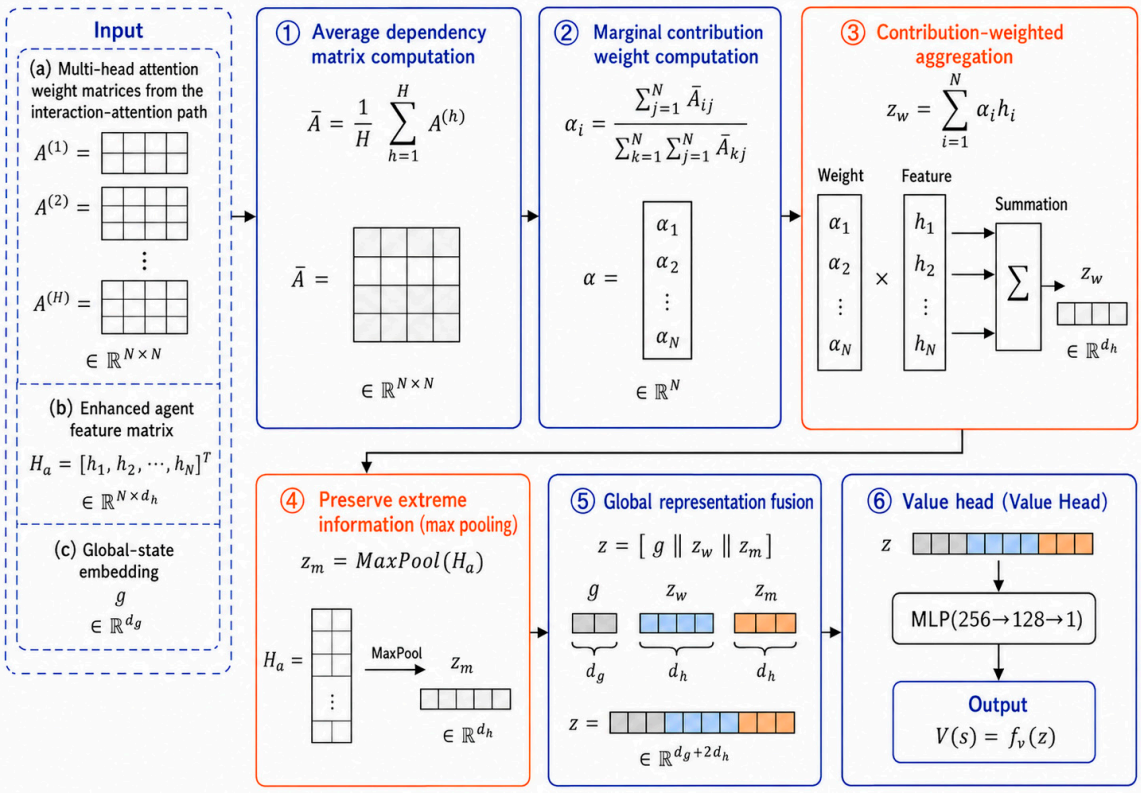

Figure 6. Workflow of marginal-contribution weighted aggregation.

During weight calculation, the attention-weight matrices of all heads in the global cooperative-dependency path are first averaged to obtain the mean dependency matrix among agents. Each row of the matrix is then summed and normalized to obtain the marginal-contribution weight of the corresponding agent to the global value:

$$\overline{A}_t = \frac{1}{h}\sum_{m=1}^{h} softmax\left(\frac{F_t W_m^Q (F_t W_m^K)^T}{\sqrt{d_k}}\right) \tag{19}$$

$$\alpha_{i,t} = \frac{\sum_{j=1}^{N} \overline{A}_t[i,j]}{\sum_{i=1}^{N}\sum_{j=1}^{N} \overline{A}_t[i,j]}, \sum_{i=1}^{N} \alpha_{i,t} = 1 \tag{20}$$

where $\alpha_{i,t} \in [0,1]$ is the normalized marginal contribution weight of the ith building. The larger the weight value is, the greater the impact of the building on the global value of the community is, which allows for the quantitative assessment of the contribution degree of heterogeneous buildings. In a multi-building energy management scenario, this weight can adaptively reflect the change in the importance of different buildings at different times and under different operating conditions. The contribution weights are normalized across buildings before aggregation, so their relative magnitudes indicate how strongly each building affects the global value estimate.

Contribution-weighted feature aggregation uses marginal-contribution weights to aggregate the encoded feature matrix. Max pooling is also combined to preserve extreme information about the global situation. After fusing global-situation features, the final global-state representation is obtained:

$$f_{weighted} = \sum_{i=1}^{N} \alpha_{i,t} \cdot F_t^{att}[i,:] f_{\max} = \max_{i=1,\ldots,N} F_t^{att}[i,:] \tag{21}$$

$$f_{final} = Concat(f_{weighted}, f_{\max}, f_t^{global}) \tag{22}$$

This aggregation strategy assigns higher weights in the final global representation to buildings that contribute more to the global value. It preserves community-coordination information and enables more reasonable credit assignment. For compatibility with the value-output head and the MAPPO framework, the aggregated global feature is fed into a two-layer, fully connected value head to output the final global state-value estimate:

$$f_{hidden} = LayerNorm(\mathrm{ReLU}(W_1 f_{finnal} + b_1)) \tag{23}$$

$$V_\phi(s_t) = W_2 f_{hibben} + b_2 \tag{24}$$

where W_1, W_2, b_1 and b_2 are the learnable parameters of the value output head, the dimensions of the two hidden layers are 256 and 128 respectively, the ReLU activation function is used, and the final output dimension is 1, which corresponds to the scalar value estimation of the global state. The MHA–critic network designed in this paper is fully compatible with the original training framework of MAPPO algorithm, and there is no need to modify the actor network structure, the main process of the algorithm and the hyperparameter settings. The global value estimate $V_\phi(s_t)$ of the network output is directly connected to the advantage function calculation and value loss optimization of the MAPPO algorithm, and its value loss function is consistent with the base MAPPO. The parameters of the value head are trained jointly with the attention critic, and the final scalar output is used as the community value estimate.

3.4. Reward Function Design

In this experiment, the amount of electricity purchased from the grid is used as a penalty term. When net electricity consumption is positive, a negative reward is given, and a larger net consumption leads to a stronger penalty. When net electricity consumption is less than or equal to zero, the reward is zero. The default reward of the i-th building at time t is defined as

$$r_t^i = -\max(e_t^i, 0)^\alpha \tag{25}$$

where e_t^i represents the net power consumption of the ith building at time t, and α is the reward index, which is set to 1 by default. The idea behind the reward function is to encourage the agent to use local PV generation to meet load demand whenever possible, and to store excess power in storage rather than sending it back to the grid. Since this paper uses a multi-agent setting, the environment will return the reward of each building, and in this experiment, the reward of each building is averaged as a scalar reward at the community level for MAPPO update: a positive net power denotes grid purchase after local PV and storage actions are considered.

$$r_t = \frac{1}{N}\sum_{i=1}^{n} r_t^i \tag{26}$$

where N is the number of buildings. The episode reward is defined as the cumulative value $R_{episode} = \sum$ of the average community reward over all time steps in an episode, and the episode reward in the training curve is the cumulative value. The averaging step keeps the reward scale comparable when the same policy framework is applied to communities with different numbers of buildings.

4. Results and Analysis

4.1. Experimental Settings and Evaluation Metrics

The experimental environment runs on Windows 11 with an NVIDIA RTX4070 GPU and an Intel i7-14700HX processor. The hyperparameters are set as follows: discount factor 0.99, GAE smoothing coefficient 0.95, PPO clipping coefficient 0.2, and a unified learning rate of 3×10^{-4}, ensuring fair comparison.

The experiments use three different random seeds, and each seed is trained for 200 episodes. Each episode corresponds to a full-year simulation period containing 8760 continuous time steps. Multiple random seeds are used to eliminate the influence of initialisation differences and improve the statistical reliability of the conclusions. During training, the current policy is evaluated on a validation set at fixed episode intervals, and evaluation reward and load-related indicators are recorded. In addition to reward, this paper reports peak net load, total net load, ramping amount, and load factor to evaluate policy performance from the perspective of grid operation. The reward reflects the comprehensive environmental penalty, whereas the load indicators more directly reflect peak reduction, energy-consumption scale, and smoothness. Using three seeds is a practical compromise for reinforcement-learning training cost; it captures major initialization effects, but the resulting standard deviations should still be read as indicative rather than as a full statistical characterization.

$$P_{peak} = max_t\ L_t \tag{27}$$

$$E_{total} = \sum_t L_t \tag{28}$$

$$R_{ramp} = \sum_t \left| L_t - L_{(t-1)} \right| \tag{29}$$

$$LF = \frac{mean(L_t)}{max(L_t)} \tag{30}$$

The average reward reflects the mean cumulative reward obtained by the algorithm over the evaluation period. Because the default CityLearn reward is negative, an average reward closer to zero means that the agents purchase less electricity from the grid and achieve better local photovoltaic utilization and storage scheduling. Peak net load is defined as the maximum community total net load during the evaluation period and is used to measure peak-shaving capability; a lower peak indicates that the algorithm can shift load more effectively from peak periods to valley periods. Total net load reflects the total community net load during evaluation and measures the impact of the algorithm on the overall electricity-consumption scale of the community. A lower total net load indicates a higher self-consumption ratio of photovoltaic energy and a reduction in grid electricity purchase. Total ramping amount is defined as the sum of absolute load changes between adjacent time steps and measures the fluctuation degree of the load curve; a lower value indicates smoother load variation and more grid-friendly operation. Load factor is defined as the ratio of average load to peak load and describes the degree of load balance.

4.2. Training Results

To verify the effectiveness of MHA-MAPPO, this section compares the original MAPPO and MHA-MAPPO from four perspectives: training convergence, evaluation-reward distribution, cumulative-reward curve, and net-load curve. All results are based on three random seeds and 200 training episodes, and the evaluation stage uses a fixed validation set.

Figure 7a compares the training convergence curves of the original MAPPO and MHA-MAPPO. The original MAPPO has a mean reward of −7144.82 and shows large fluctuations.

The curves of the three seeds are widely dispersed, and some seeds converge to poor results. By contrast, MHA-MAPPO has a mean reward of −5772.60 and a standard deviation of only 580.66, with the three seed curves closely distributed around the mean. This indicates that introducing multi-head attention and the gating structure makes training convergence smoother and greatly reduces the sensitivity of the algorithm to random seeds.

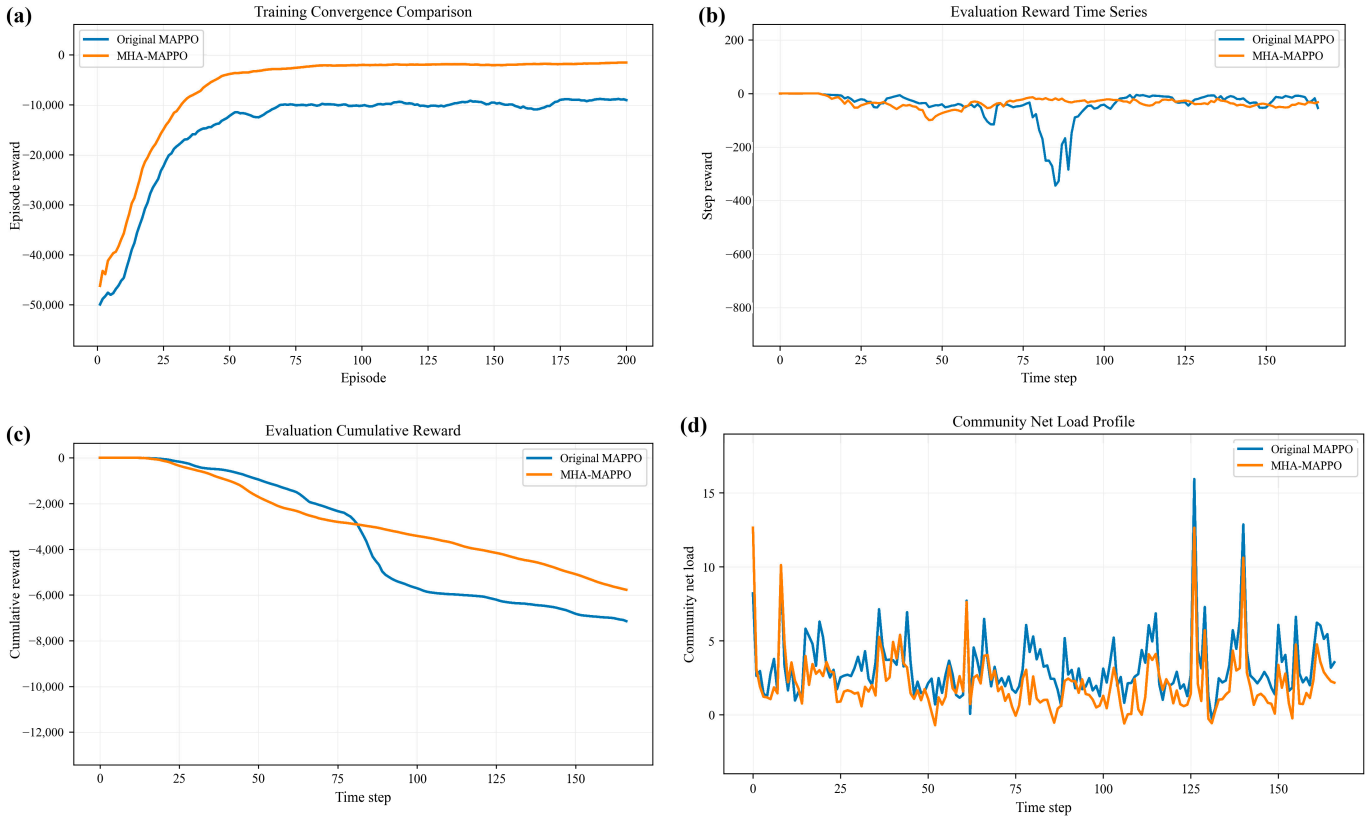

Figure 7. (**a**) Training convergence curves of MAPPO and MHA-MAPPO; (**b**) evaluation reward distributions over random seeds; (**c**) cumulative reward during evaluation; (**d**) mean and fluctuation range of the net-load curve.

Figure 7b shows the distribution of evaluation rewards for different methods over random seeds. The figure presents reward variation from the perspective of the evaluation time series. Because the reward is a net-grid-purchase penalty, a curve closer to zero is better. MHA-MAPPO has a smaller overall penalty and its curve lies clearly above that of the original MAPPO, indicating that it coordinates storage actions and load demand more effectively. In particular, during load-peak periods, the reward drop in MHA-MAPPO is much smaller than that of the original MAPPO, demonstrating the advantage of the attention mechanism in sequential decision making.

Figure 7c shows the cumulative reward curve during evaluation. A higher cumulative-reward curve indicates a smaller accumulated penalty, and a slower decline indicates more stable control performance. MHA-MAPPO declines more slowly, suggesting that it can maintain a lower operating penalty throughout the evaluation period without sudden performance degradation. This further verifies the stability and robustness of the MHA-MAPPO control policy.

Figure 7d shows the mean and fluctuation range of the net-load curves of different methods. The figure intuitively presents the distribution of net load at each time step during the evaluation period. MHA-MAPPO achieves an average peak net load of 13.49, compared with 15.95 for the original MAPPO. From the curve shape, the improved model shows a more obvious peak-shaving effect during high-load periods, and the peak of the

net-load curve is effectively reduced. The fluctuation range, represented by the shaded area, is also smaller for MHA-MAPPO, indicating more stable control performance.

From an energy-flow perspective, the lower peaks in MHA-MAPPO indicate that storage is charged more conservatively before high-demand periods and discharged more consistently when building loads and electricity prices rise. This reduces the occurrence of simultaneous grid purchase among multiple buildings and makes the community net-load curve less sensitive to the load peaks in individual building peaks.

Figure 8 shows the load-shaping effect relative to the uncontrolled baseline. The controlled net load is compared with the uncontrolled baseline to determine whether the policy truly changes the load profile. MHA-MAPPO makes the net-load curve flatter in multiple periods, especially during periods that originally had a high load, where the net load is significantly reduced. This demonstrates the shaping effect of storage scheduling on community load and shows that MHA-MAPPO can effectively use energy-storage systems to shift load over time.

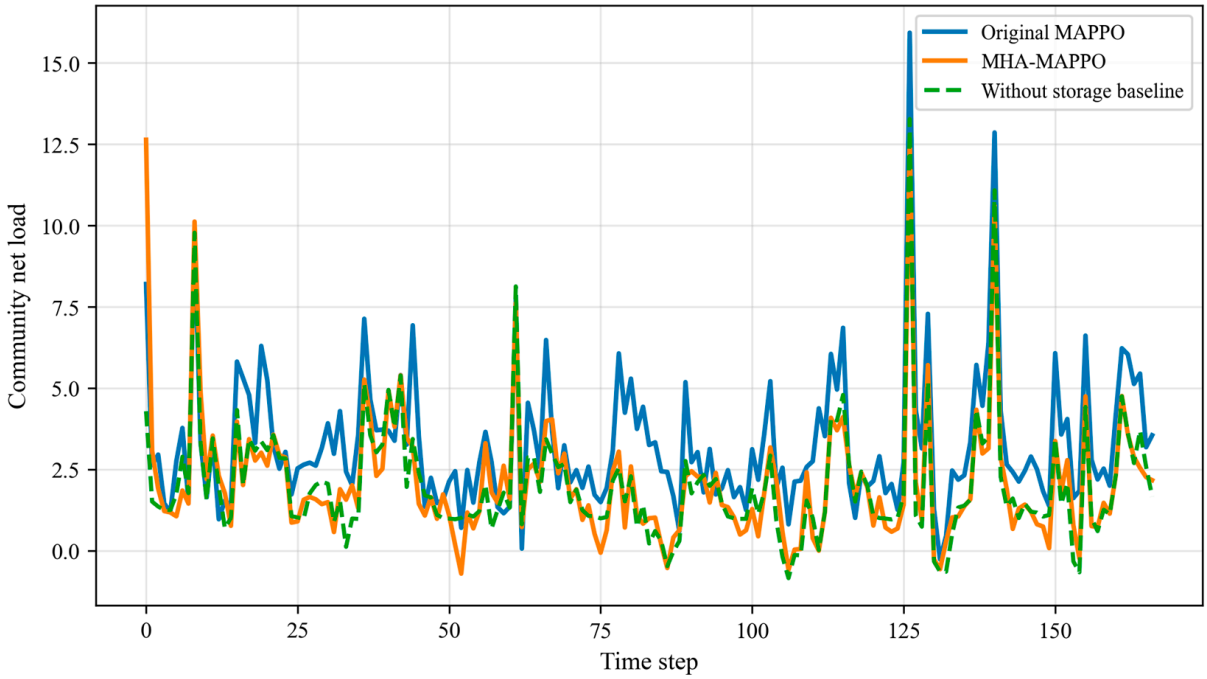

Figure 8. Load-shaping effect relative to the uncontrolled baseline.

4.3. Comparative Experiments

The training convergence curves show that MHA-MAPPO performs more stably under multiple random seeds and achieves a better later-stage average reward than the original MAPPO. This result indicates that introducing inter-building attention modelling into the centralized critic enables the algorithm to use multi-building cooperative information more effectively.

Table 1 presents the comparative results between the proposed method and existing models, including the random policy, rule-based control, single-agent PPO, independent multi-agent PPO (IPPO), centralized PPO, and the original MAPPO. In terms of evaluation reward, single-agent PPO achieves the highest value, −2957.09, and significantly outperforms all multi-agent methods. This phenomenon requires careful interpretation. PPO concatenates the states of all buildings and controls them as a whole, so it can theoretically learn a globally optimal policy; however, its action space grows linearly with the number of buildings and is difficult to scale to larger communities. More importantly, although PPO obtains the highest reward, its peak net load is 14.79 and its total net load is 408.69,

both higher than the MHA-MAPPO values of 13.49 and 352.31. This shows that reward indicators and load indicators are not fully consistent. PPO may obtain a higher reward through certain strategies, but MHA-MAPPO has clearer advantages in peak reduction and total energy consumption, which are more relevant to grid operation. Rule-based control achieves the lowest total net load, 231.29, among all methods, and its peak net load of 13.25 is comparable to that of MHA-MAPPO; however, its reward is −14,794.89, far lower than those of learning-based methods. This indicates that rule-based control can achieve low total grid purchase, but its purchase behaviour is poorly matched with the reward function and fails to avoid penalties effectively during high-price or concentrated-load periods. The random policy obtains a reward of −55,067.03, much lower than all learning-based methods, confirming the effectiveness of reinforcement-learning control policies. IPPO and centralized PPO both perform worse than the original MAPPO and MHA-MAPPO, indicating that the MAPPO multi-agent framework is well suited to the CityLearn task.

Table 1. Comparative experimental results with existing models.

Method	Evaluation Reward	Peak Net Load (kW)	Total Net Load (kWh)
Random	−55,067.03 ± 1063.98	17.85 ± 4.16	943.83 ± 3.71
Rule-based	−14,794.89 ± 0.00	13.25 ± 0.00	231.29 ± 0.00
PPO	−2957.09 ± 172.68	14.79 ± 1.31	408.69 ± 6.18
IPPO	−6648.75 ± 3620.18	15.58 ± 3.86	425.30 ± 81.57
Centralized PPO	−17,418.26 ± 20,107.54	15.26 ± 2.02	717.72 ± 193.26
Original MAPPO	−7144.82 ± 5646.64	15.95 ± 1.97	546.34 ± 173.09
MHA-MAPPO	−5772.60 ± 580.66	13.49 ± 0.26	352.31 ± 22.07

Figure 9a shows the evaluation-reward comparison with existing models. It compares random, rule-based, PPO, IPPO, centralized PPO, original MAPPO, and MHA-MAPPO over three random seeds. PPO achieves the highest reward, but its boxplot span is large, indicating substantial variation among seeds. The reward distribution of MHA-MAPPO is concentrated and better than those of the original MAPPO and IPPO, showing that the attention-enhanced critic improves community-reward performance compared with the original multi-agent baseline. Centralized PPO has the worst reward performance and extremely large fluctuations, suggesting that simply centralizing all agents is not a good choice in multi-building scenarios.

Figure 9b shows the peak-net-load comparison with existing models. A lower peak net load indicates stronger peak-shaving capability. MHA-MAPPO achieves a peak net load of 13.49, lower than PPO at 14.79 and the original MAPPO at 15.95, and comparable to rule-based control at 13.25. This indicates that the proposed method has a clear advantage in suppressing community load peaks and can achieve a peak-shaving performance that is comparable to rule-based control while outperforming it on other indicators.

Figure 9c shows the total-net-load comparison with existing models. Total net load reflects the community net-load scale during the evaluation period. MHA-MAPPO obtains a total net load of 352.31, lower than PPO at 408.69 and the original MAPPO at 546.34, but higher than rule-based control at 231.29. This indicates that MHA-MAPPO outperforms other learning-based methods in reducing total net load, although a gap remains relative to carefully designed rule-based control. It also shows that rule-based control remains a strong baseline in the CityLearn task.

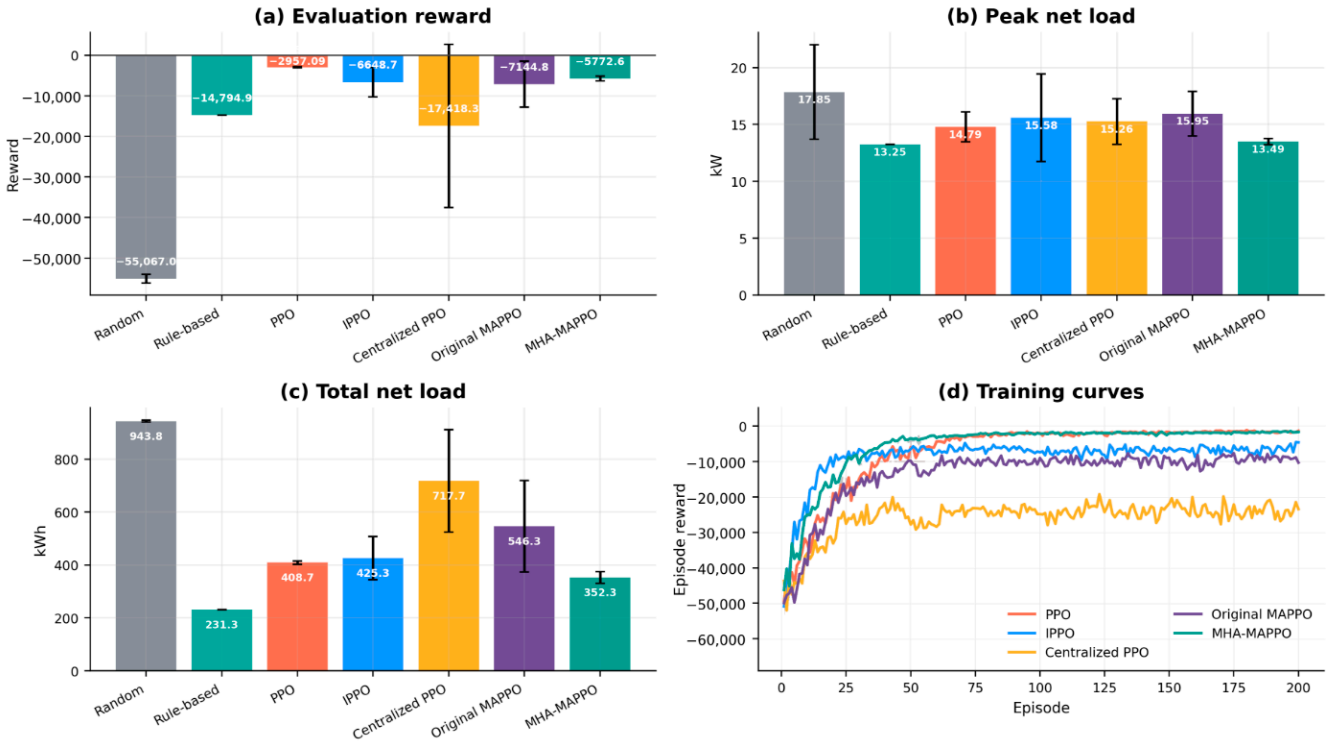

Figure 9. Comparison with existing models.

Figure 9d shows the training convergence curves of existing learning-based models. The figure presents the training processes of PPO, IPPO, centralized PPO, original MAPPO, and MHA-MAPPO. PPO converges fastest and achieves the highest final reward, but its curve fluctuates substantially. IPPO and centralized PPO show poor convergence, with low curve positions and severe fluctuations. The original MAPPO lies in the middle, whereas MHA-MAPPO stabilizes at a higher level in the later stage and fluctuates much less than the original MAPPO. The differences among the curves reflect training-stability differences in coordinated multi-building control, and MHA-MAPPO is more stable than the original MAPPO.

4.4. Ablation Experiments

The ablation experiments systematically verify the independent contributions and interaction effects of the multi-head attention module (A), the dual-branch gated embedding module (G), and the dual-path attention encoding structure (D). Seven ablation combinations are designed, covering single-module, dual-module, and complete-model settings. The original MAPPO is used as the baseline, and evaluation reward, peak net load, total net load, and training curves are compared. All ablation models use the same hyperparameter settings as the original MAPPO to ensure fairness.

Table 2 presents the ablation results of eight models, including the original MAPPO. Among the single-module combinations, introducing G alone achieves the highest evaluation reward of −4532.68, an improvement of about 36.6% over the original MAPPO at −7144.82, but its peak net load is slightly higher than the baseline. This indicates that the gating mechanism is effective for optimizing the reward objective but contributes less to peak shaving. Introducing +A alone gives an evaluation reward of −5760.64, reduces peak net load to 14.77, and reduces total net load to 347.06. All three indicators improve over the original MAPPO, showing that the attention-based modelling of inter-building interactions is critical for improving overall algorithmic performance. Notably, introducing D alone results in an evaluation reward of −8082.81, lower than the original MAPPO, but

its total net load decreases to 300.28, the lowest among all single-module combinations. This suggests that the dual-path structure has a unique advantage in reducing total energy consumption, but when used alone it may have compatibility issues with other modules and thus reduce the reward indicator.

Table 2. Ablation results of eight module combinations.

Model	Evaluation Reward	Peak Net Load (kW)	Total Net Load (kWh)
Original MAPPO	-7144.82 ± 5646.64	15.95 ± 1.97	546.34 ± 173.09
+A	-5760.64 ± 3058.49	14.77 ± 2.82	347.06 ± 51.34
+G	-4532.68 ± 998.03	16.21 ± 2.47	365.68 ± 27.33
+D	-8082.81 ± 1467.43	15.82 ± 4.03	300.28 ± 35.40
+A+G	-6218.09 ± 492.33	14.80 ± 2.93	339.16 ± 38.15
+A+D	-6580.83 ± 249.45	16.39 ± 3.31	330.48 ± 9.97
+G+D	-4851.82 ± 964.87	13.66 ± 0.22	360.05 ± 9.75
+A+G+D	-5772.60 ± 580.66	13.49 ± 0.26	352.31 ± 22.07

Among dual-module combinations, +G+D achieves an evaluation reward of -4851.82 and reduces peak net load to 13.66, outperforming the single-module combinations on both indicators. This indicates a positive synergy between gated embedding and dual-path encoding. The +A+G combination obtains an evaluation reward of -6218.09 with a standard deviation of only 492.33, the best training stability among all dual-module combinations, showing that attention and gating jointly reduce sensitivity to random seeds. The +A+D combination gives an evaluation reward of -6580.83, with a standard deviation of 249.45, which is the most stable, but its peak net load of 16.39 is the highest among all combinations. This suggests that without the gating mechanism, the combination of attention and dual-path encoding may overemphasize interaction information and insufficiently optimize the peak-shaving objective.

The complete model (+A+G+D), namely, MHA-MAPPO, outperforms all ablation combinations in peak net load (13.49) and total net load (352.31). Although its evaluation reward (-5772.60) is lower than that of +G, its standard deviation (580.66) is much smaller than that of +G (998.03), indicating better training stability and generalization. Compared with the original MAPPO, the complete model improves evaluation reward by about 19.2%, reduces peak net load by about 15.4%, and reduces total net load by about 35.5%. These results show that the three modules form an effective cooperative mechanism in the complete model: gated embedding adaptively fuses local and global features, multi-head attention models dynamic dependencies among buildings, and dual-path encoding extracts high-level global contextual information. Together, they improve comprehensive performance in multi-building energy management.

Figure 10 compares the ablation results of the eight models from four dimensions: evaluation reward, peak net load, total net load, and training curve. In the reward distribution shown in Figure 9a, the +G combination has the highest mean reward but a large boxplot span, indicating strong sensitivity to random seeds. The complete +A+G+D model has the second-highest mean reward but a more compact boxplot, showing better stability. The original MAPPO has the most dispersed reward distribution, with a large gap between the minimum and maximum values, further confirming its sensitivity to initialisation. In the peak-net-load comparison, the complete model and the +G+D combination are at the lowest level, while +A+D has the highest peak, indicating that gated embedding plays a key role in peak shaving. In the total-net-load comparison, the +D combination achieves

the lowest total net load and the complete model ranks second; both are much lower than the original MAPPO, showing the importance of dual-path encoding for reducing total energy consumption. In the training convergence curves, the complete model remains at a relatively high level with the smallest fluctuations, whereas the original MAPPO fluctuates strongly and some seeds show obvious performance regression. Overall, the complete model achieves the best balance among reward, peak shaving, energy saving, and stability, confirming the necessity of the synergistic action of the three modules.

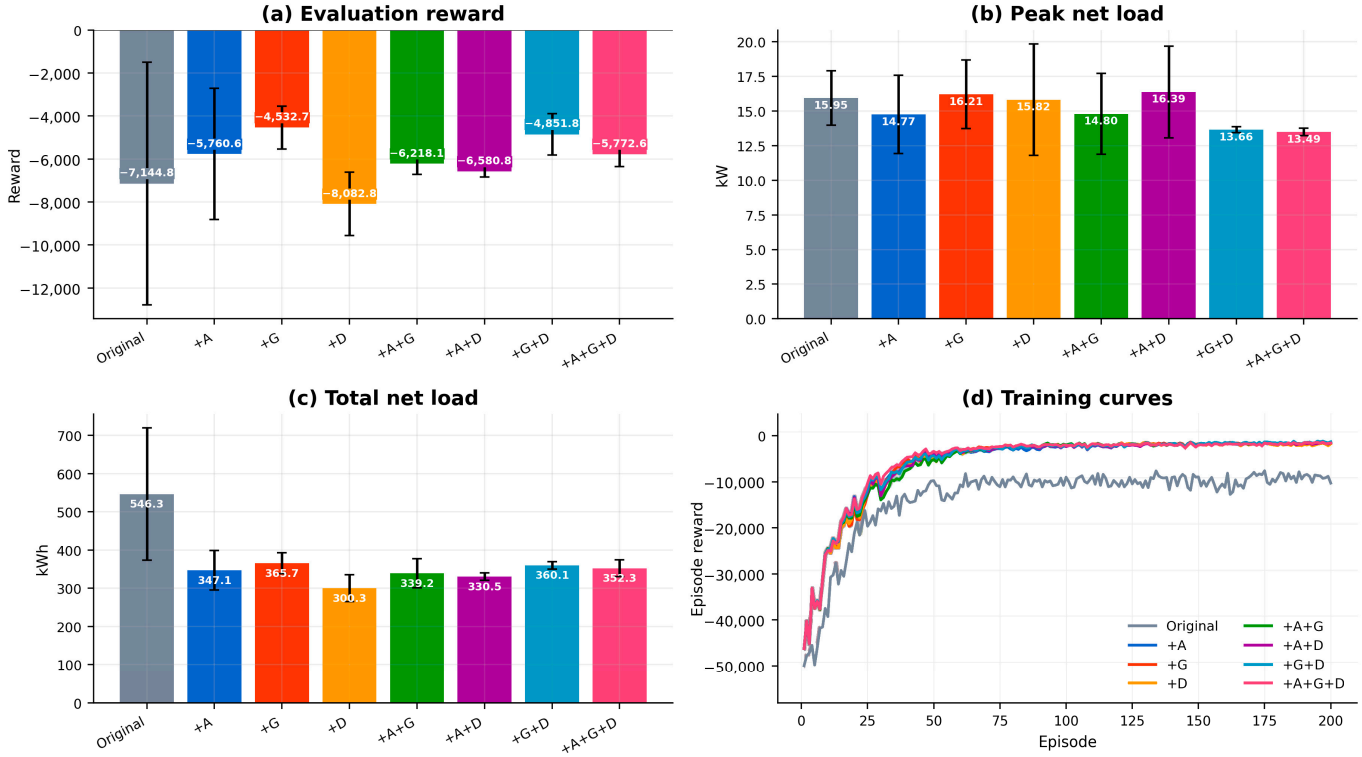

Figure 10. Ablation results of module combinations.

Figure 11 shows the mean community net-load curves of the eight ablation models during evaluation, allowing for direct comparison of their effects on load shape. The complete model (+A+G+D) suppresses the peak most strongly during high-load periods and produces a smoother overall curve. The gated-embedding-only model (+G) performs well in valley periods, but its peak-control effect is weaker than that of the complete model. The attention-only model (+A) shows some peak-shaving effect in medium- and high-load intervals, but obvious residual peaks remain in several periods. The dual-path-only model (+D) produces lower net load in some periods, but its overall fluctuation is larger and peak control is unstable. Among dual-module combinations, +G+D has a curve shape closest to that of the complete model, further confirming the synergy between gating and dual-path encoding. Overall, the net-load curves are consistent with the quantitative indicators in Table 2 and Figure 10, and the complete model performs best in both peak shaving and smoothness.

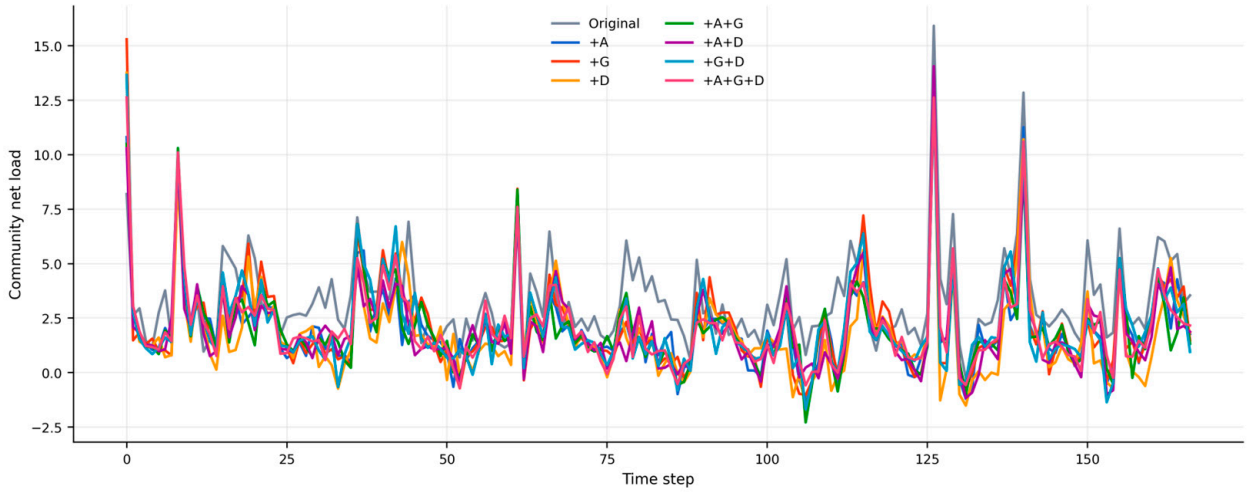

Figure 11. Net-load curves of eight ablation combinations.

4.5. Load Smoothing and Action Analysis

In addition to reward and peak load, ramping amount and load factor are important supplementary indicators for evaluating grid-friendliness. Ramping amount reflects the severity of community net-load fluctuations between adjacent time steps. A lower ramping amount means a smoother load curve and helps reduce frequency-regulation and reserve pressure on the grid. Load factor is defined as the ratio of average load to peak load and characterizes the degree of load balance; a higher load factor indicates a smoother load and higher utilization efficiency of grid equipment. Based on the comparative experiments, this section evaluates the performance of different methods in load smoothing and control response from four dimensions: total ramping amount, load factor, net-load ramping trajectory, and energy-storage regulation response.

Figure 12a shows the total ramping amount of the comparative models during evaluation. MHA-MAPPO has a total ramping amount of 301.24, significantly lower than the original MAPPO at 363.51, a reduction of about 17.1%. The standard deviation of MHA-MAPPO is only 15.27, much smaller than the 42.86 of the original MAPPO, indicating more stable load-smoothing performance. PPO has a total ramping amount of 338.92, between the two methods, but its larger standard deviation indicates obvious fluctuation among random seeds. Rule-based control achieves the lowest total ramping amount, 289.63, confirming its inherent advantage in load smoothing; however, combined with the previous analyses of reward and peak load, its overall economic performance is poor. The Random policy reaches a total ramping amount of 712.45, reflecting severe net-load fluctuation without strategic control. Overall, MHA-MAPPO achieves load-smoothing performance close to rule-based control while significantly outperforming it in reward and peak-shaving indicators.

Figure 12b shows the load factors of the comparative models. A higher load factor indicates a more balanced load curve. The load factor of MHA-MAPPO is 0.486, higher than 0.452 for the original MAPPO, representing an improvement of about 7.5%. This shows that MHA-MAPPO can more effectively shift peak load to valley periods and realize peak shaving and valley filling. The load factor of PPO is 0.473, slightly lower than that of MHA-MAPPO, and its peak net load is higher, indicating weaker peak-shaving performance. Rule-based control has the highest load factor, 0.512, again confirming its advantage in load balancing. The random policy has a load factor of only 0.198, indicating an extremely unbalanced load. In terms of standard deviation, MHA-MAPPO has the smallest load-factor fluctuation over the three random seeds and therefore shows the best robustness.

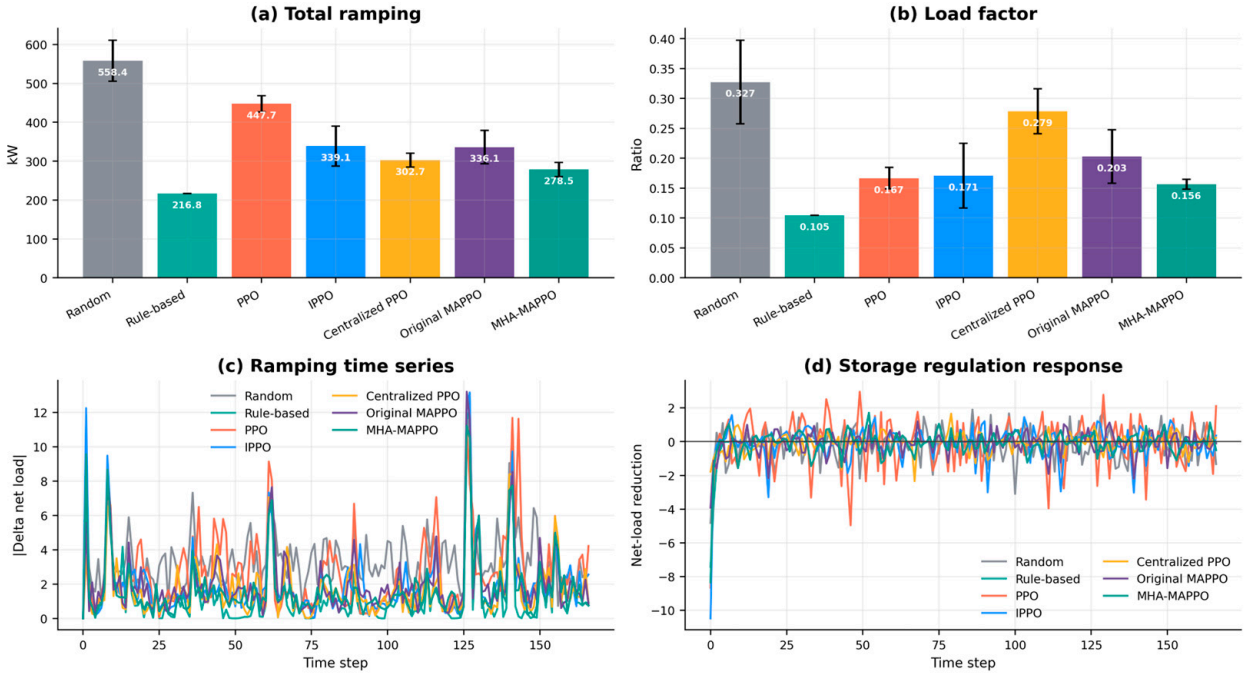

Figure 12. Load smoothing and action analysis.

Figure 12c shows the net-load ramping trajectory during a representative evaluation period, such as a high-temperature summer day. The horizontal axis is the time step and the vertical axis is the first-order difference in net load, namely the load change between adjacent time steps. This figure is used to observe instantaneous load fluctuations. The ramping trajectory of MHA-MAPPO stays close to the zero line for most time steps, and both positive and negative ramping amplitudes are effectively limited within ± 20 kW, with very few fluctuations exceeding $+/-50$ kW. By contrast, the original MAPPO has larger ramping fluctuations and several peaks above $+/-80$ kW, indicating that its storage scheduling policy does not respond sufficiently quickly or coordinate adequately under sudden load changes. PPO lies between the two, but still has clear positive ramping peaks in some periods, reflecting possible coordination deficiencies of single-agent control in multi-building scenarios. Rule-based control produces the smoothest ramping trajectory, with almost no fluctuations exceeding $+/-15$ kW, but its flexibility is limited and it is difficult to adapt to dynamic electricity prices and photovoltaic fluctuations. Overall, MHA-MAPPO is significantly better than the original MAPPO in net-load ramping control and greatly narrows the gap with rule-based control.

Figure 12d shows the charging and discharging power curve of a representative building during evaluation, which is used to analyze storage-regulation behaviour under different control strategies. MHA-MAPPO produces relatively smooth charging and discharging actions and a moderate switching frequency, avoiding the adverse effects of frequent charge–discharge transitions on storage lifetime. During high-price periods, such as afternoon to evening, MHA-MAPPO can charge the storage system to a relatively high state of charge in advance and discharge it in an orderly manner during the peak period, thereby achieving effective peak shaving and valley filling. The original MAPPO shows some instability in storage actions, with frequent switching in certain periods. This not only increases storage wear but also reflects insufficient temporal credit assignment. PPO has a relatively smooth storage response, but because of its centralized control structure it struggles to consider the individual characteristics of heterogeneous buildings, leading to low storage utilization in some buildings. Rule-based control has the most regular storage actions and charges or discharges entirely according to fixed periods. Although its smooth-

ness is best, it cannot respond to dynamic electricity prices and real-time photovoltaic output, resulting in higher total grid-purchase cost. Overall, MHA-MAPPO achieves a better balance between smooth storage regulation and economic performance.

The results in Figure 12 show that MHA-MAPPO outperforms the original MAPPO in total ramping amount, load factor, net-load ramping trajectory, and storage-regulation response, and some indicators approach or even exceed those of rule-based control. Specifically, compared with the original MAPPO, MHA-MAPPO reduces total ramping amount by about 17.1%, increases load factor by about 7.5%, substantially narrows the fluctuation range of the net-load ramping trajectory, and produces smoother and more reasonable storage charging and discharging actions. These improvements mainly result from the multi-head-attention-enhanced critic, which can estimate the global state value more accurately and guide building agents to learn coordinated storage scheduling policies, thereby avoiding the ineffective charge–discharge actions and load peaks caused by unclear credit assignment. Compared with rule-based control, MHA-MAPPO still has a small gap in some smoothness indicators, but its advantages in dynamic price response and photovoltaic utilization lead to better overall performance. The analysis further verifies the effectiveness and practicality of the proposed MHA-MAPPO algorithm in multi-building energy management.

4.6. Computational Cost and Hyperparameter Sensitivity Analysis

4.6.1. Additional Computational Cost of the Attention Mechanism

In the original MAPPO framework, the centralized critic directly concatenates the observations of all buildings into a joint state and estimates the global state value through a multilayer perceptron. By contrast, the proposed MHA-MAPPO critic introduces local-global gated embedding, an inter-building interaction-attention path, a global context-attention path, and contribution-weighted aggregation to enhance the modelling of coupling relationships among buildings. Let N denote the number of building agents, d denote the local observation dimension, and h denote the embedding dimension. The dominant complexity of the original MAPPO critic can be approximated as $O(Ndh+h^2)$, whereas that of the MHA-MAPPO critic can be approximated as $O(Ndh+Nh^2+N^2h)$. The N^2h term comes from the attention-score computation among buildings. In the three-building community considered in this study, this quadratic term is relatively small, and the additional overhead mainly comes from the extra linear projections, gated fusion, and dual-path feature-encoding modules.

Table 3 compares the computational cost of the original MAPPO and MHA-MAPPO under the same input dimensions. The critic of MHA-MAPPO contains about 11.36 times as many parameters as the original MAPPO critic. However, because the actor networks are unchanged, the total actor–critic parameter count increases to only 4.22 times that of the original framework. In terms of estimated multiply-accumulate operations per sample, the MHA-MAPPO critic requires about 25.20 times the computation of the original MAPPO critic. With a batch size of 128 on CPU, the critic forward latency is 16.72 times that of the original MAPPO, and the latency of one training update is 2.97 times that of the original MAPPO.

Table 3. Computational cost comparison between MAPPO and MHA-MAPPO.

Method	Critic Params	Total Params	Critic MACs/Sample	CPU Critic Forward/Batch	CPU Training Step/Batch
Original MAPPO	28,289	90,899	28,032	0.491 ms	15.127 ms
MHA-MAPPO	321,282	383,892	706,368	8.206 ms	44.915 ms

4.6.2. Sensitivity Analysis of Key Hyperparameters

To examine the sensitivity of MHA-MAPPO to key hyperparameters, this study focuses on two architecture-level parameters that directly affect the attention critic: the number of attention heads and the embedding dimension. The number of attention heads is varied among 1, 2, 4, and 8 while keeping the embedding dimension fixed at h = 128. The embedding dimension is varied among 64, 96, 128, 192, and 256 while using four attention heads where the dimension is divisible by the number of heads. For the training coefficients, the main experiments use the same PPO-related settings as the MAPPO baseline, including discount factor gamma = 0.99, GAE coefficient lambda = 0.95, clipping ratio 0.2, actor learning rate 3×10^{-4}, critic learning rate 1×10^{-3}, six update epochs, and minibatch size 128. As detailed in Table 4, This design avoids attributing performance gains to additional hyperparameter tuning and makes the structural contribution of the attention-enhanced critic clearer.

Table 4. Hyperparameter sensitivity settings and observations.

Sensitivity Item	Tested Values	Main Observation	Recommended Setting
Attention heads	1, 2, 4, 8	Parameter count is unchanged; latency varies mildly because Q/K/V projection size is fixed.	4 heads
Embedding dimension	64, 96, 128, 192, 256	Parameters and latency increase clearly with h; h = 128 balances cost and capacity.	128
Training coefficients	gamma = 0.99, lambda = 0.95, clip = 0.2	Same coefficients as MAPPO isolate the effect of the attention critic.	Keep baseline coefficients

As shown on the left side of Figure 13, when the embedding dimension is fixed at 128, the number of attention heads is not the main source of computational overhead. The parameter count remains unchanged for different head numbers, and the variation range of CPU forward latency is about 10.2%. This is because changing the number of heads mainly partitions the same embedding dimension into different subspaces without changing the overall size of the Q/K/V projection matrices. Therefore, using four attention heads provides a balanced setting between multi-subspace relation modelling and computational stability.

In contrast, the embedding dimension has a more obvious influence on computational cost. As shown on the right side of Figure 13, when h increases from 64 to 128, the critic parameter count increases from 117,890 to 321,282; when h further increases to 256, the parameter count reaches 1,084,418. The forward latency also increases with the embedding dimension. This suggests that a very small embedding dimension may limit the representation of inter-building interactions, whereas an excessively large embedding dimension substantially increases computational burden. Considering representation capacity, training stability, and computational cost together, h = 128 is selected as the default setting in this study.

Overall, MHA-MAPPO is not highly sensitive to the number of attention heads in terms of computational overhead, but it is more sensitive to the embedding dimension. With the same training coefficients as the original MAPPO baseline, the proposed method still achieves higher mean evaluation reward and lower reward standard deviation over three random seeds. This indicates that the performance improvement does not rely on

special hyperparameter tuning, but mainly comes from the ability of the attention-enhanced critic to model inter-building interactions more effectively.

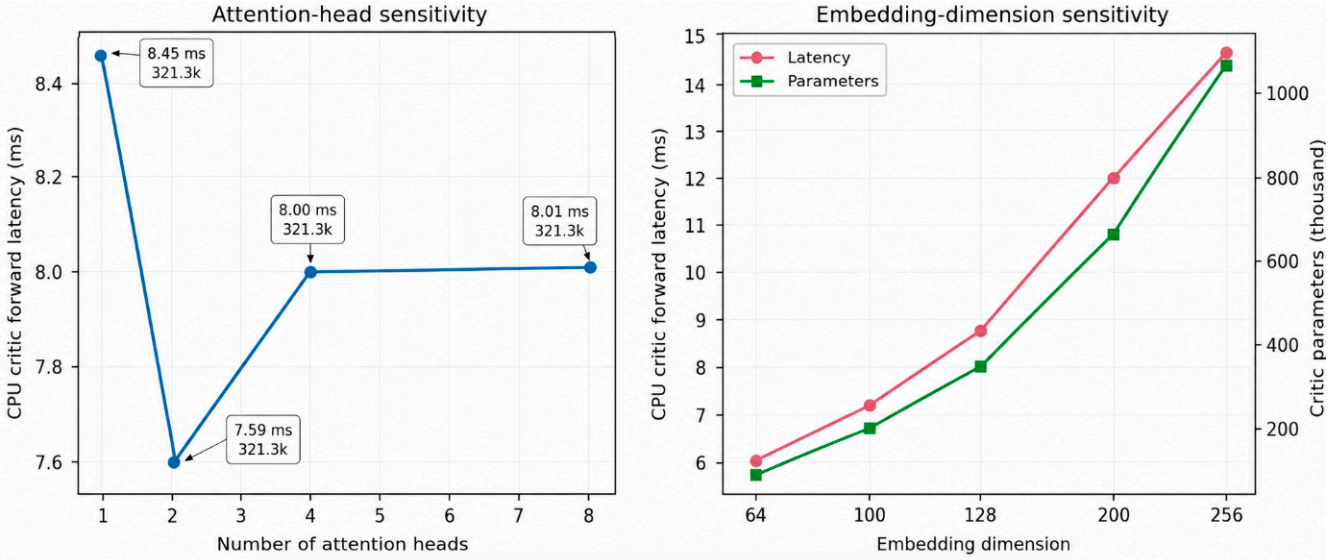

Figure 13. Hyperparameter sensitivity of the MHA critic.

5. Conclusions

This paper addresses the problem of multi-agent coordinated control in community multi-building energy management and proposes an improved MAPPO algorithm based on a multi-head-attention-enhanced centralized critic, namely, MHA-MAPPO. Without changing the decentralized execution framework, the algorithm uses a dual-branch gated embedding module to adaptively fuse local building observations and global interaction information. It also uses a dual-path multi-head attention encoding structure to model pairwise dependencies among buildings and global contextual features at the community level, thereby significantly improving the value-estimation accuracy of the centralized critic for the joint state. Experiments in the CityLearn environment show that the mean evaluation reward of MHA-MAPPO is -5772.60, better than the -7144.82 obtained for the original MAPPO, corresponding to a relative improvement of about 19.2%. The reward standard deviation decreases sharply from 5646.64 to 580.66, indicating significantly improved training stability. In terms of load indicators, peak net load decreases by about 15.4%, total net load decreases by about 35.5%, and total ramping amount decreases by about 17.1%. The proposed method therefore achieves better comprehensive performance in reward, peak shaving, energy saving, and load smoothing. Ablation experiments further verify that the multi-head attention module, the gated-fusion mechanism, and the dual-path encoding structure all contribute to the performance improvement, with multi-head attention playing a key role in modelling inter-building interactions. This study provides an effective reference for improving multi-agent reinforcement learning algorithms in community energy management. Future work may further introduce attention mechanisms into the actor network and verify scalability in larger community scenarios.

Several limitations remain. The experiments are carried out in the CityLearn environment and use three random seeds, so the results mainly reflect the tested community configuration and simulation assumptions rather than the full range of real building operation. The conclusions should therefore be further checked with more building types, longer weather and price records, and, when data access allows, real operational traces. Future work will extend the validation to larger and more diverse communities, couple the control

policy with electricity-market signals, and test whether the learned coordination behaviour remains stable under different climates and building-use patterns.

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

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. International Energy Agency. *Global Status Report for Buildings and Construction 2024*; IEA: Paris, France, 2024.
2. China Association of Building Energy Efficiency. *China Urban and Rural Construction Carbon Emission Research Report 2025*; CABEE: Beijing, China, 2026.
3. Caballero-Güereca, C.E.; Vogel, J.; Alaux, N.; Ouellet-Plamondon, C.M.; Santana, J.S.; Foliente, G.; Lützkendorf, T.; Passer, A. Towards net zero: Sectoral ambitions and global trends in building decarbonisation. *Build. Cities* **2026**, *7*, 396–424. [CrossRef]
4. Pokhrel, D.; Dulal, S.; Liu, G. Community microgrid scheduling considering building thermal dynamics using a deep learning approach. *Electronics* **2026**, *15*, 1719. [CrossRef]
5. O'Dwyer, E.; Pan, I.; Acha, S.; Shah, N. Smart energy systems for sustainable smart cities: Current developments, trends and future directions. *Appl. Energy* **2019**, *237*, 581–597. [CrossRef]
6. Amer, A.A.; Bayhan, S.; Abu-Rub, H.; Massoud, A.; Ehsani, M. A review of deep reinforcement learning algorithms for grid services in grid-interactive efficient buildings. *Energy Rep.* **2026**, *15*, 108900. [CrossRef]
7. Jafarinejad, T.; Erfani, A.; Saelens, D. Consensus clustering of residential districts under various DR regimes: Stable flexibility oriented building clusters. *Energy Build.* **2026**, *351*, 116694. [CrossRef]
8. Jin, Q.; Lu, W. Toward low-carbon buildings: Breakthroughs and challenges in PV–storage–DC–flexibility system. *Energies* **2026**, *19*, 197. [CrossRef]
9. Shaikh, P.H.; Nor, N.B.M.; Nallagownden, P.; Elamvazuthi, I.; Ibrahim, T. A review on optimized control systems for building energy and comfort management of smart sustainable buildings. *Renew. Sustain. Energy Rev.* **2014**, *34*, 409–429. [CrossRef]
10. Wang, Z.; Hong, T.; Piette, M.A. Building thermal load prediction through shallow machine learning and deep learning. *Appl. Energy* **2020**, *263*, 114683. [CrossRef]
11. Serale, G.; Fiorentini, M.; Capozzoli, A.; Bernardini, D.; Bemporad, A. Model predictive control for heating and cooling of buildings: A review. *Energy Build.* **2018**, *168*, 416–430. [CrossRef]
12. Liu, Y.; Kergus, P.; Claveau, F.; Chevrel, P.; Lacarrière, B. A hierarchical model predictive control framework for building thermal management: A comparison of grey-box and black-box models under limited data. *J. Build. Eng.* **2025**, *110*, 112936. [CrossRef]
13. Zhang, S.; Lei, X.; Shan, X.; Li, T.; Wu, W. Optimization of an MPC controller based on a hybrid cooling load prediction model and experimental validation in HVAC systems. *Buildings* **2026**, *16*, 1269. [CrossRef]
14. Ham, S.W.; Kim, D.; Paul, L.; Casillas, A.; Prakash, A.K.; Brown, R.; Pritoni, M. Machine learning-enhanced MPC for demand flexibility in small commercial buildings: An experimental study. *Energy Build.* **2026**, *360*, 117324. [CrossRef]
15. Saloux, E.; Zhang, K.; Candanedo, J.A. A critical perspective on current research trends in building operation: Pressing challenges and promising opportunities. *Buildings* **2023**, *13*, 2566. [CrossRef]
16. Drgoňa, J.; Arroyo, J.; Figueroa, I.C.; Blum, D.; Arendt, K.; Kim, D.; Ollé, E.P.; Oravec, J.; Wetter, M.; Vrabie, D.L.; et al. All you need to know about model predictive control for buildings. *Annu. Rev. Control* **2020**, *50*, 190–232. [CrossRef]
17. Arulkumaran, K.; Deisenroth, M.P.; Brundage, M.; Bharath, A.A. A brief survey of deep reinforcement learning. *IEEE Signal Process. Mag.* **2017**, *34*, 26–38. [CrossRef]
18. Li, Y. Deep reinforcement learning: An overview. *arXiv* **2017**, arXiv:1701.07274. [CrossRef]
19. Vamvoudakis, K.G.; Lewis, F.L. Online actor–critic algorithm to solve the continuous-time infinite horizon optimal control problem. *Automatica* **2010**, *46*, 878–888. [CrossRef]
20. Zhang, J.; Jiang, Z.; Wang, W.; Ju, Y. Deep reinforcement learning-driven operational strategy for community microgrid systems: Balancing energy efficiency and occupant comfort in residential buildings. *Energy Build.* **2025**, *352*, 116749. [CrossRef]
21. Félez Val, R.; Félez Mindán, J. Intelligent HVAC control in residential buildings: A review of advanced techniques and AI applications. *Appl. Sci.* **2026**, *16*, 2006. [CrossRef]

22. Haq, M.S.U.; Ji, W.; Geng, Y.; Lin, B. Can reinforcement learning move beyond hype in building control? *Innov. Energy* **2026**, *3*, 100150. [CrossRef]
23. Hernandez-Leal, P.; Kartal, B.; Taylor, M.E. A survey and critique of multiagent deep reinforcement learning. *Auton. Agents Multi-Agent Syst.* **2019**, *33*, 750–797. [CrossRef]
24. Gronauer, S.; Diepold, K. Multi-agent deep reinforcement learning: A survey. *Artif. Intell. Rev.* **2022**, *55*, 895–943. [CrossRef]
25. Abdalla, A.N.; Ji, R.; Awad, O.I.; Wang, A.; Tang, Y.; Liu, Y.; Mahdi, M.A. Distributed control and optimization in building energy systems: A technical survey of multi-agent architectures and algorithms. *Energy Build.* **2026**, *350*, 116644. [CrossRef]
26. Lowe, R.; Wu, Y.; Tamar, A.; Harb, J.; Abbeel, P.; Mordatch, I. Multi-agent actor-critic for mixed cooperative-competitive environments. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6379–6390.
27. Jin, J.; Sun, G.; Chen, S.; Li, Y.; Zhu, H.; Mao, W.; Ji, W. Buildings-to-grid with generalized energy storage: A multi-agent decomposed deep reinforcement learning approach for delayed rewards. *Appl. Energy* **2026**, *404*, 127181. [CrossRef]
28. Yu, C.; Velu, A.; Vinitsky, E.; Gao, J.; Wang, Y.; Bayen, A.M.; Wu, Y. The surprising effectiveness of PPO in cooperative multi-agent games. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24611–24624. [CrossRef]
29. Wang, G.; Sun, S.; Cheng, Y.; Yu, P.; Wang, S.; Zhao, X. Multi-agent proximal policy optimization for coordinated adaptive control of photovoltaic inverter clusters in active distribution networks. *Energies* **2026**, *19*, 978. [CrossRef]
30. Chen, B.; Cai, Z.; Bergés, M. Gnu-RL: A practical and scalable reinforcement learning framework for building energy management. *J. Build. Eng.* **2021**, *43*, 103119. [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.