Causally Aligned Multiagent Reinforcement Learning for Coordinated Control of Heterogeneous Home Energy Devices

Xiao Du, Fengji Luo, *Senior Member, IEEE*, Juntao Hu, Wei Zhou, *Member, IEEE*, and Junhao Wen, *Member, IEEE*

***Abstract*—Deep reinforcement learning (DRL) has emerged as a promising paradigm for home energy management systems (HEMSs) due to its model-free nature and ability to handle complex dynamics. However, existing DRL-based approaches typically employ a unified reward function that aggregates multiple objectives into a single scalar, failing to account for the heterogeneous roles of controllable energy devices (CEDs) and their distinct causal relationships with control objectives. This leads to reward misattribution, where CEDs with simpler constraints dominate the optimization process while critical components such as battery storage remain underutilized. To address this challenge, we propose a causally aligned multiagent reinforcement learning (MARL) framework that explicitly models CED-objective causal pathways using a structural causal model (SCM). A causal surgery procedure decomposes shared objectives into CED-specific variants, enabling individualized reward signals aligned with each CED's causal responsibility. The proposed heterogeneous reward-aware multiagent soft actor–critic (HR-MASAC) algorithm features a multihead centralized critic for learning vectorized Q-values and agent-specific entropy coefficients for heterogeneous exploration. Experiments across diverse home scenarios demonstrate that our method achieves 40.1% cost reduction and 57.1% comfort improvement over unified-reward baselines, with robust performance under sensor noise and household heterogeneity.**

***Index Terms*—Causal alignment, home energy management, multiagent reinforcement learning (MARL), reward attribution.**

I. Introduction

THE increasing penetration of renewable energy sources and the growing diversity in residential energy demand have posed new challenges to the design of efficient and adaptive home energy management systems (HEMS). Traditional model-based or rule-based control (RBC) methods often struggle to operate effectively under such complex and uncertain environments, especially when faced with the dynamic characteristics of distributed energy resources and user-driven objectives [1].

Received 6 August 2025; revised 2 January 2026; accepted 23 February 2026. Date of publication 27 February 2026; date of current version 11 May 2026. This work was supported in part by the Coal-Major Project under Grant 2025ZD1700800 and in part by the Australian Research Council through Discovery Project under Grant DP220103881. *(Corresponding author: Junhao Wen.)*

Xiao Du, Juntao Hu, Wei Zhou, and Junhao Wen are with the School of Bigdata and Software Engineering, Chongqing University, Chongqing 400044, China (e-mail: jhwen@cqu.edu.cn).

Fengji Luo is with the School of Civil Engineering, University of Sydney, Sydney, NSW 2008, Australia.

Digital Object Identifier 10.1109/JIOT.2026.3668848

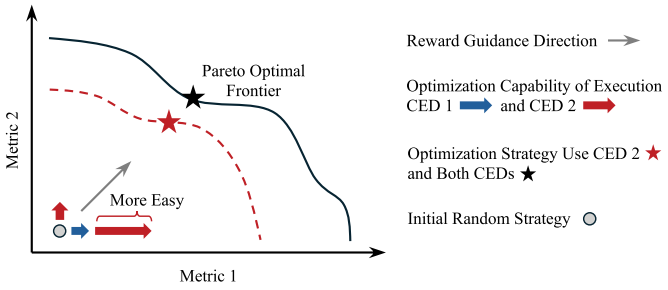

Fig. 1. Illustration of the optimization bias under a unified reward setting in DRL-based HEMS.

In recent years, deep reinforcement learning (DRL) has emerged as a promising paradigm for constructing intelligent HEMS, due to its ability to perform model-free optimization and capture nonlinear patterns in high-dimensional decision spaces [2]. Numerous studies have applied DRL algorithms to optimize residential energy usage, improve demand-side flexibility, and balance cost, comfort, and sustainability objectives [3]. However, most DRL-based HEMS approaches—whether single-agent or multiagent—adopt a unified reward function that aggregates multiple objectives into a single scalar signal.

While straightforward, this design suffers from a fundamental flaw: it does not distinguish the roles and causal responsibilities of heterogeneous controllable energy devices (CEDs). In realistic HEMS settings, different CEDs [e.g., air conditioners (ACs), battery energy storage systems (BESSs), and washing machines (WMs)] contribute differently toward system objectives and vary in controllability, constraints, and action space complexity. As illustrated in Fig. 1, when two devices jointly influence multiple objectives, some devices offer more accessible exploration paths than others. Under a unified reward signal, the learning agent tends to over-exploit the easier-to-optimize devices, converging toward suboptimal strategies. Empirical evidence from our experiments (Section V) confirms this: under unified rewards, agents controlling AC inadvertently dominate cost optimization that should be handled by BESS, leading to 45–80% higher energy costs compared to causally aligned policies.

The core challenge is **reward misattribution**: unified scalar rewards do not clarify which CED contributes to which objective, causing the policy optimizer to favor dominant control pathways while underutilizing strategically impor-

tant CEDs. Several techniques in multiagent reinforcement learning (MARL) address credit assignment: value decomposition methods [4], [5] factorize the joint value function into individual utilities, but assume fully cooperative tasks with a single shared objective; counterfactual methods such as COMA [6] and difference rewards [7] estimate each agent's marginal contribution, though the latter requires additional environment rollouts. However, these methods perform credit assignment implicitly through learning, without prior knowledge of which agent should be responsible for which objective. In multiobjective HEMS where multiple CEDs can influence the same objective through different mechanisms, such implicit approaches may attribute rewards to unintended CEDs. An orthogonal question thus remains: how to systematically determine which CED causally affects which objective based on the physical structure of the system, and accordingly construct heterogeneous rewards that align each CED with its intended role.

To address this, we propose a causally aligned MARL framework for coordinated control of heterogeneous home CEDs, with the following key contributions below.

1) We introduce causal reward attribution for HEMS with heterogeneous CEDs by constructing a structural causal model (SCM) [8] that explicitly encodes CED-to-objective causal pathways based on physical system knowledge.
2) We propose a causal surgery procedure that decomposes shared objectives into CED-specific variants, enabling individualized reward signals aligned with each CED's causal role.
3) We develop heterogeneous reward-aware multiagent soft actor–critic (HR-MASAC), a multiagent actor–critic algorithm with a multihead centralized critic and agent-specific entropy coefficients, supporting heterogeneous reward optimization while maintaining interagent coordination. The vectorized rewards guide each agent's policy update through dedicated critic heads, enabling CED-specific optimization without sacrificing coordination.

II. Related Work

A. DRL for HEMS

DRL provides a data-driven alternative to traditional rule-based or model-predictive control methods by learning from sequential interactions with the environment. The integration of deep neural networks enables DRL to capture the nonlinear dynamics and stochasticity inherent in HEMS [9]. DRL has been applied to a variety of HEMS objectives, including cost minimization [10], load shifting [11], user comfort control [2], and renewable energy integration [12].

1) Centralized DRL With Single-Agent Control: Early DRL applications in HEMS often adopt a single-agent framework, where a centralized agent observes the global state and jointly controls multiple CEDs. Representative methods include value-based DQN [2], as well as policy-gradient approaches such as SAC [13], PPO [14], and TD3 [15]. For example, Salari et al. [11] incorporated fuzzy rules into DQN to reduce the complexity of the action space, enabling efficient control of continuous energy loads while preserving BESS health. Similarly, Tan et al. [16] proposed a SAC-based HEMS strategy that dynamically adjusts appliance operations based on real-time electricity prices, achieving both cost savings and thermal comfort optimization.

However, beyond the reward misattribution and suboptimal convergence issues highlighted in this work, such a centralized single-agent control paradigm also suffers from several limitations: poor scalability as the joint action space grows exponentially with the number of CEDs, lack of modularity since system changes require complete retraining, and low robustness due to single-point-of-failure risks.

2) Multiagent DRL for Distributed CED Control: To overcome the scalability and modularity limitations of centralized control, recent studies have introduced MARL into HEMS. In this setting, each CED is modeled as an autonomous agent, which interacts with its local environment while contributing to system-wide objectives. Typical frameworks follow the centralized training with decentralized execution (CTDE) paradigm, employing shared critics such as MADDPG [17] or value decomposition methods like QMIX [4] and MAPPO [18] to enable decentralized yet cooperative policy learning. For instance, to address privacy concerns, [19] modeled the environment as a partially observable Markov decision process (POMDP) and developed a vectorized Advantage Actor–Critic (A2C) algorithm that enables community-level HEMS to learn control policies without accessing sensitive data such as electric vehicle (EV) arrival times or indoor temperatures. Abishu et al. [20] proposed a hierarchical MARL framework with dynamically reconfigurable agent hierarchies, where low-level agents are prioritized for decision making to prevent system overload. This method outperformed fixed-hierarchy and nonhierarchical baselines, although abstraction may lead to information loss. In another study, Deng et al. [21] designed a MADDPG-based strategy to handle PV generation uncertainty by training on randomized solar outputs and applying joint chance constraint reformulation, significantly improving local renewable energy utilization.

Existing studies primarily leverage the advantages of MARL in decentralized coordination and heterogeneous optimization across differing observation or action spaces. As the number of CEDs grows, mean-field control (MFC) [22] offers a scalable alternative by approximating agent interactions through population statistics, with recent extensions addressing heterogeneous agents [23] and constrained settings [24]. Despite these advances, most MARL methods for HEMS rely on unified reward signals or implicit credit assignment through value decomposition, leaving the question of how to align heterogeneous CEDs with their causally relevant objectives largely unaddressed.

B. Credit Assignment and CRL

Credit assignment—determining how much each agent contributes to the team reward—is a fundamental challenge in cooperative MARL [3]. Existing methods can be broadly categorized into value decomposition and counterfactual approaches.

Value decomposition methods such as QMIX [4] and QTRAN [5] factorize the joint action-value function into individual utilities under structural constraints (e.g., monotonicity), performing implicit credit assignment through end-to-end learning. Counterfactual methods take an alternative approach by estimating each agent's marginal contribution. Difference rewards [7] compute an agent's contribution by comparing outcomes with and without its action, requiring additional environment rollouts. COMA [6] avoids extra rollouts by using a centralized critic that conditions on the joint action to compute counterfactual baselines analytically; it marginalizes over each agent's action space to estimate what would have happened under alternative actions, providing a gradient signal that reflects individual contributions.

While effective in homogeneous cooperative games, these methods face limitations in heterogeneous HEMS. They assume a unified reward signal, yet HEMS involves multiple distinct objectives (cost, comfort, and task completion) that different CEDs should prioritize differently. Moreover, they perform credit assignment implicitly through learning, without explicitly specifying which agent should be responsible for which objective.

Causal inference [25] provides a principled framework for reasoning about cause–effect relationships through SCMs. An SCM represents a system as a directed acyclic graph (DAG) where nodes denote variables and edges encode causal mechanisms, enabling interventional reasoning—predicting outcomes under hypothetical actions—beyond mere correlational analysis. Causal reinforcement learning (CRL) [8] integrates these tools into *RL* to address challenges where standard methods struggle. Existing CRL applications span several domains: improving sample efficiency through causal world models [26], enabling robust policy transfer by identifying causal invariances [27], enhancing off-policy evaluation via importance weighting with causal structure [28], and learning interpretable policies by discovering causal features [29]. These applications primarily focus on single-agent settings with a single objective.

Despite the natural fit between causal reasoning and multiagent credit assignment—where the core question is "which agent's action *caused* which outcome"—this connection remains underexplored in DRL-based energy management. This work bridges this gap by leveraging the physical causal structure inherent in energy systems: CED capabilities determine which observables they affect, and observables determine which objectives they influence. By encoding these relationships in an SCM, we transform the implicit credit assignment problem into explicit causal attribution. Furthermore, the SCM framework provides a principled basis for decomposing confounded objectives into CED-specific variants by severing inappropriate causal pathways—enabling systematic reward decomposition aligned with each CED's intended role.

III. Problem Formulation

A. System Description

Consider a HEMS with n CEDs $\mathcal{D} = \{d_1, \ldots, d_n\}$, a set of observable variables $\mathcal{O} = \{o_1, \ldots, o_p\}$, and m optimization objectives $\mathcal{J} = \{\mathcal{J}_1, \ldots, \mathcal{J}_m\}$. The system operates in discrete time steps $t \in \{1, 2, \ldots, \tau\}$. At each step, a control policy π observes the system state and determines actions for all CEDs.

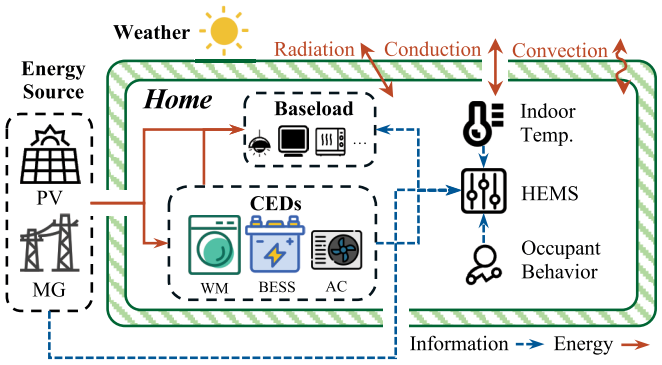

Fig. 2. Home energy management environment.

Due to CED heterogeneity, there exist latent causal dependencies that can be represented as an SCM with the chain: CED → Observable → Objective → Reward → Policy. Let $\mathcal{E}_{do} \subseteq \mathcal{D} \times \mathcal{O}$ and $\mathcal{E}_{ol} \subseteq \mathcal{O} \times \mathcal{J}$ denote the causal relations from CEDs to observables and from observables to objectives, respectively. A unified reward $r = f(\mathcal{J}_1, \ldots, \mathcal{J}_m)$ connects all objectives to all policies regardless of these causal pathways, leading to ambiguous credit assignment [Fig. 3(a)].

B. Problem Statement

Problem (Causally Aligned Multiagent Control): Given the CED set $\mathcal{D}$, observable set $\mathcal{O}$, and objective set $\mathcal{J}$.

1) *Reward Attribution*: Discover $\mathcal{E}_{do}$ and $\mathcal{E}_{ol}$, and construct a causally aligned SCM where each CED d_i has a dedicated reward node r^{d_i} connecting only to objectives it can influence. For objectives influenced by multiple CEDs (confounded), decompose into CED-local variants $\mathcal{J}_j^{d_i}$ when the causal pathway is separable. The output is an attributed objective set $\tilde{\mathcal{J}}^{d_i}$ and observable subset $\mathcal{O}^{d_i}$ for each CED.
2) *Policy Learning*: Design individualized reward functions $r^{d_i} = g(\tilde{\mathcal{J}}^{d_i}, \mathcal{O}^{d_i})$ based on attributed objectives and observables, where the specific form of g incorporates domain knowledge. Then, learn coordinated policies $\{\pi^{d_i}\}_{i=1}^{n}$ that jointly optimize the original objectives $\mathcal{J}$.

C. HEMS Environment

As illustrated in Fig. 2, the household is connected to the utility main grid (MG) and a rooftop photovoltaic (PV) system. The CED set includes a BESS, an AC, and a WM, representing the main categories of residential loads—flexible storage, thermal comfort, and deferrable tasks—commonly analyzed in demand-side management [30]. Other appliances are treated as uncontrollable base load P_t^{BL} (kW). The observable set $\mathcal{O}$ is detailed in Section IV-B. The control interval is Δt (hours).

The optimization objectives $\mathcal{J} = \{\mathcal{J}_c, \mathcal{J}_t, \mathcal{J}_l\}$ correspond to electricity cost, thermal comfort violation, and laundry delay

$$\min_{\pi} \ \mathcal{J}_c = \frac{1}{\tau} \sum_{t=1}^{\tau} \kappa_t P_t^{\mathrm{G}} \Delta t \tag{1a}$$

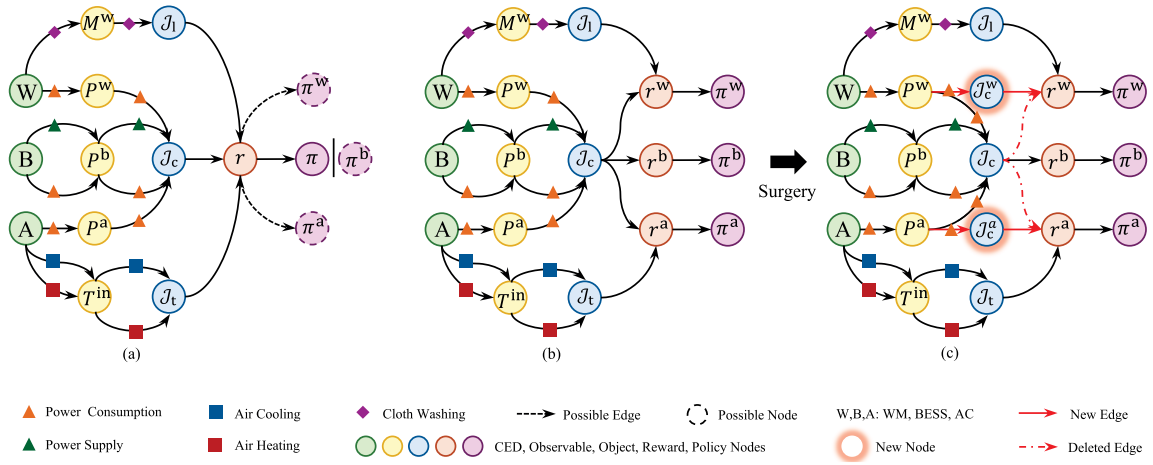

Fig. 3. Comparison of causal structures for reward attribution in the considered HEMS environment. (a) SCM of unified reward (single or multiagent). (b) SCM of heterogeneous reward based on our casual discovery method. (c) SCM of heterogeneous reward after surgery.

$$\min_{\pi} \ \mathcal{J}_{\mathrm{t}} = \frac{1}{\tau}\sum_{t=1}^{\tau} \mathbb{1}\left(\left|T_t^{\mathrm{in}} - T_t^{\mathrm{set}}\right| > T^{\mathrm{band}}\right) \tag{1b}$$

$$\min_{\pi} \ \mathcal{J}_{\mathrm{l}} = \frac{1}{n^{\mathrm{lau}}}\sum_{i=1}^{n^{\mathrm{lau}}} \mathbb{1}\left(\sum_{t=t_i^{\mathrm{lau}}}^{t_i^{\mathrm{lau}}+z_i^{\mathrm{lau}}} M_t^{\mathrm{w}} \geq 1\right) \tag{1c}$$

where τ is the episode length, $\mathbb{1}(\cdot)$ is the indicator function (returns 1 if the condition is satisfied and 0 otherwise), κ_t (/kWh) and P_t^{G} (kW) denote electricity price and grid power exchange (positive for import); T_t^{in}, T_t^{set}, and T^{band} (°C) are indoor temperature, setpoint, and comfort tolerance; n^{lau} is the number of laundry requests, each defined by request time t^{lau} (hours), maximum delay z^{lau} (hours), and actual start time t^{start} (hours).

1) Energy Device Models: The devices follow physics-based dynamics. The BESS is modeled with stored energy E_t (kWh), rated capacity E_{rat} (kWh), state of charge $\mathrm{soc}_t = E_t/E_{\mathrm{rat}}$, self-discharge rate SD (h^{-1}), and efficiency η^{b}

$$E_{t+1} = (1 - \mathrm{SD}\cdot\Delta t)\cdot\begin{cases} E_t + \eta^{\mathrm{b}} P_t^{\mathrm{b}}\Delta t, & M_t^{\mathrm{b}} = 1 \\ E_t - \dfrac{P_t^{\mathrm{b}}\Delta t}{\eta^{\mathrm{b}}}, & M_t^{\mathrm{b}} = -1 \\ E_t, & M_t^{\mathrm{b}} = 0 \end{cases} \tag{2a}$$

$$P_t^{\mathrm{b}} \in \left[0, P_{\mathrm{sup}}^{\mathrm{b}}\right] \tag{2b}$$

$$P_{\mathrm{sup}}^{\mathrm{b}} = \begin{cases} \min\left\{P_{\mathrm{rat}}^{\mathrm{b}}, \frac{E_{\mathrm{rat}}-E_t}{\eta^{\mathrm{b}}\Delta t}\right\}, & M_t^{\mathrm{b}} = 1 \\ \min\left\{P_{\mathrm{rat}}^{\mathrm{b}}, \frac{\eta^{\mathrm{b}}E_t}{\Delta t}\right\}, & M_t^{\mathrm{b}} = -1 \\ 0, & M_t^{\mathrm{b}} = 0 \end{cases} \tag{2c}$$

where $M_t^{\mathrm{b}} \in \{-1, 0, 1\}$ denotes the operating mode (charging/standby/discharging), P_t^{b} (kW) is bounded by $P_{\mathrm{sup}}^{\mathrm{b}}$ (kW) that respects both rated power $P_{\mathrm{rat}}^{\mathrm{b}}$ (kW) and capacity limits.

The AC converts electrical power P_t^{a} (kW) to thermal output P_t^{heat} (kW) with coefficient of performance (COP) dependent on operating mode, indoor temperature T_t^{in} (°C), and outdoor temperature T_t^{out} (°C)

$$P_t^{\mathrm{heat}} = \mathrm{cop}_t P_t^{\mathrm{a}} M_t^{\mathrm{a}}, \quad P_t^{\mathrm{a}} \in \left[P_{\mathrm{sta}}^{\mathrm{a}}, P_{\mathrm{rat}}^{\mathrm{a}}\right] \tag{3a}$$

$$\mathrm{cop}_t = \begin{cases} \overline{\mathrm{cop}}_t, & 0 < \overline{\mathrm{cop}}_t < \mathrm{cop}_{\mathrm{rat}} \\ \mathrm{cop}_{\mathrm{rat}}, & \text{otherwise} \end{cases} \tag{3b}$$

$$\overline{\mathrm{cop}}_t = \begin{cases} \dfrac{\eta^{\mathrm{a}}\left(T_t^{\mathrm{in}}\right)_{\mathrm{K}}}{T_t^{\mathrm{in}} - T_t^{\mathrm{out}}}, & M_t^{\mathrm{a}} = 1 \\ \dfrac{\eta^{\mathrm{a}}\left(T_t^{\mathrm{in}}\right)_{\mathrm{K}}}{T_t^{\mathrm{out}} - T_t^{\mathrm{in}}}, & M_t^{\mathrm{a}} = -1 \\ 0, & M_t^{\mathrm{a}} = 0 \end{cases} \tag{3c}$$

where $M_t^{\mathrm{a}} \in \{-1, 0, 1\}$ indicates cooling/standby/heating mode, $P_{\mathrm{sta}}^{\mathrm{a}}$ and $P_{\mathrm{rat}}^{\mathrm{a}}$ (kW) are standby and rated power, $(\cdot)_{\mathrm{K}}$ denotes Kelvin conversion, $\mathrm{cop}_{\mathrm{rat}}$ is the upper COP limit, and η^{a} is the thermodynamic efficiency factor.

The WM operates as a noninterruptible task with power consumption P_t^{w} (kW) and accumulated run time z_t^{w} (hours)

$$P_t^{\mathrm{w}} = M_t^{\mathrm{w}} P_{\mathrm{rat}}^{\mathrm{w}} \tag{4a}$$

$$z_t^{\mathrm{w}} = \begin{cases} \min\left\{z_{t-1}^{\mathrm{w}} + \Delta t, \ z_{\mathrm{task}}^{\mathrm{w}}\right\}, & M_t^{\mathrm{w}} = 1 \\ 0, & M_t^{\mathrm{w}} = 0 \end{cases} \tag{4b}$$

$$M_t^{\mathrm{w}} = \begin{cases} 1, & \left(0 < z_{t-1}^{\mathrm{w}} < z_{\mathrm{task}}^{\mathrm{w}}\right) \ \vee \ A_t^{\mathrm{w}} = 1 \\ 0, & \text{otherwise} \end{cases} \tag{4c}$$

where $M_t^{\mathrm{w}} \in \{0, 1\}$ is the operating mode (running/idle), $P_{\mathrm{rat}}^{\mathrm{w}}$ (kW) is rated power, $z_{\mathrm{task}}^{\mathrm{w}}$ (hours) is task duration, and $A_t^{\mathrm{w}} \in \{0, 1\}$ is the start signal (allowed only when idle).

The PV generates output power P_t^{PV} (kW) based on solar irradiance I_t (W/m^2)

$$P_t^{\mathrm{PV}} = \begin{cases} 0, & I_t < I_{\min} \\ \min\left(\eta^{\mathrm{PV}} I_t, P_{\mathrm{rat}}^{\mathrm{PV}}\right) A^{\mathrm{pan}}, & \text{otherwise} \end{cases} \tag{5}$$

where $I_{\min}$ (W/m^2) is the minimum irradiance threshold, η^{PV} is conversion efficiency, $P_{\mathrm{rat}}^{\mathrm{PV}}$ (kW/m^2) is rated output per unit area, and A^{pan} (m^2) is panel area.

2) Building Thermal Dynamics: Indoor temperature T_t^{in} (°C) evolves via heat balance

$$T_{t+1}^{\mathrm{in}} = T_t^{\mathrm{in}} + \frac{Q_t^{\mathrm{con}} + Q_t^{\mathrm{ven}} + Q_t^{\mathrm{solar}} + P_t^{\mathrm{heat}}\Delta t}{c^{\mathrm{air}}\rho^{\mathrm{air}}V^{\mathrm{hou}}} \tag{6a}$$

$$Q_t^{\mathrm{con}} = \lambda A^{\mathrm{suf}}\left(T_t^{\mathrm{out}} - T_t^{\mathrm{in}}\right)\Delta t \tag{6b}$$

$$Q_t^{\text{ven}} = \text{ACH}\, c^{\text{air}} \rho^{\text{air}} V^{\text{hou}} \left(T_t^{\text{out}} - T_t^{\text{in}}\right) \Delta t \tag{6c}$$

$$Q_t^{\text{solar}} = \text{SHGC}\, I_t A^{\text{fen}} \Delta t \tag{6d}$$

where Q_t^{con}, Q_t^{ven}, and Q_t^{solar} (kWh) represent conduction, ventilation, and solar heat gains; λ (kW/(m^2°C)) is thermal conductivity, A^{suf} (m^2) and A^{fen} (m^2) are envelope and window areas, ACH (h^{-1}) is air change rate, SHGC is solar heat gain coefficient, c^{air} (kWh/(kg°C)) is specific heat capacity, ρ^{air} (kg/m^3) is air density, and V^{hou} (m^3) is indoor volume.

3) Power Balance and Stochastic Inputs: The household's power balance requires

$$P_t^{\text{G}} = \max\left(P_t^{\text{D}}, 0\right) \tag{7a}$$

$$P_t^{\text{D}} = P_t^{\text{a}} + P_t^{\text{b}} + P_t^{\text{w}} + P_t^{\text{BL}} - P_t^{\text{PV}} \tag{7b}$$

where P_t^{D} (kW) is net demand and excess generation is curtailed.

Daily occupancy $\text{occ}_{d,t} \in \{0, 1\}$ (1: at home, 0: away) on day d follows stochastic patterns:

$$\text{occ}_{d,t} = 1 - \mathbb{1}\left(t_d^{\text{dep}} \le t < t_d^{\text{arr}}\right) \tag{8a}$$

$$t_d^{\{\text{dep,arr}\}} \sim \mathcal{N}\left(\mu^{\{\text{dep,arr}\}}, \left(\sigma^{\{\text{dep,arr}\}}\right)^2\right) \tag{8b}$$

where t_d^{dep} and t_d^{arr} (hours) are departure and arrival times, and μ, σ are day-type-dependent Gaussian parameters.

Laundry request indicator $\text{lau}_{d,t} \in \{0, 1\}$ is generated as

$$\text{lau}_{d,t} = \text{req}_d \cdot \mathbb{1}\left(t_d^{\text{lau}} \le t \le t_d^{\text{lau}} + z_d^{\text{lau}}\right) \tag{9a}$$

$$\text{req}_d \sim \text{Bernoulli}\left(p_{\text{dow}}^{\text{lau}}\right) \tag{9b}$$

where $\text{req}_d \in \{0, 1\}$ indicates whether a request occurs on day d with day-of-week-dependent probability $p_{\text{dow}}^{\text{lau}}$, t_d^{lau} (hours) is request time sampled from occupied slots, and z_d^{lau} (hours) is maximum tolerable delay. Exogenous variables—I_t, T_t^{out}, P_t^{BL}, κ_t—are derived from real-world datasets [31], [32].

IV. Methodology

This section presents our methodology for addressing the causally aligned multiagent control problem. We first describe the general causal reward attribution framework, then formalize the MDP with vectorized rewards, and finally present the HR-MASAC algorithm for coordinated policy learning.

A. Causal Reward Attribution

We leverage SCMs [8] to systematically attribute rewards based on causal relationships. An SCM represents the system as a DAG encoding the causal chain: CED → Observable → Objective → Reward → Policy. Let $\mathcal{C}(d_i)$ denote the capability set of CED d_i, $\mathcal{C}(\mathcal{J}_j)$ denote the capabilities required to influence objective $\mathcal{J}_j$, and $\mathcal{O}(c)$ denote the observable associated with capability c. Our framework operates through three phases.

1) Phase 1 (Causal Discovery): Construct the SCM by identifying CED-objective causal links via capability matching

$$d_i \rightsquigarrow \mathcal{J}_j \iff \mathcal{C}(d_i) \cap \mathcal{C}\left(\mathcal{J}_j\right) \neq \emptyset. \tag{10}$$

For each matched pair, add edges $d_i \to o \to \mathcal{J}_j$ where $o \in \mathcal{O}(\mathcal{C}(d_i) \cap \mathcal{C}(\mathcal{J}_j))$. Then, connect objectives to CED-specific rewards: $\mathcal{J}_j \to r^{d_i}$ for all $\mathcal{J}_j$ that d_i can influence, and $r^{d_i} \to \pi^{d_i}$. This ensures each CED receives reward signals only for objectives it can causally affect.

2) Phase 2 (Causal Surgery): Causal discovery alone is insufficient when multiple CEDs share influence over the same objective [Fig. 3(b)]. Surgery decomposes such confounded objectives into CED-local variants when appropriate. The key insight is that CEDs with a single capability affecting a shared objective typically have a more important primary task—for example, AC 's primary task is thermal comfort, not cost minimization. We create local objectives so these CEDs avoid resource waste while fulfilling primary tasks. Formally, surgery applies when: 1) $|\mathcal{C}(d_i) \cap \mathcal{C}(\mathcal{J}_j)| = 1$ (single-capability influence) and 2) $|\mathcal{J}^{d_i}| > 1$ (distinct primary task exists). When both hold, create $\mathcal{J}_j^{d_i}$ and redirect $\mathcal{J}_j \to r^{d_i}$ to $\mathcal{J}_j^{d_i} \to r^{d_i}$; otherwise, retain the global objective.

3) Phase 3 (Backward Extraction): Extract attributed objectives and observables by tracing backward from each reward node

$$\tilde{\mathcal{J}}^{d_i} = \text{Pa}\left(r^{d_i}\right) \cap \mathcal{J}^*, \quad \mathcal{O}^{d_i} = \bigcup_{\mathcal{J}_j \in \tilde{\mathcal{J}}^{d_i}} \text{Pa}\left(\mathcal{J}_j\right) \cap \mathcal{O} \tag{11}$$

where $\text{Pa}(\cdot)$ denotes parent nodes in the SCM, and $\mathcal{J}^* = \mathcal{J} \cup \{\mathcal{J}_j^{d_i}\}$ includes both global and local objectives. Importantly, $\mathcal{O}^{d_i}$ provides necessary guidance for reward design rather than a complete specification—supplementary variables may be incorporated based on domain knowledge.

The complete procedure is formalized in Algorithm 1, which outputs the SCM $\mathcal{G}$ and attributed pairs $\{(\tilde{\mathcal{J}}^{d_i}, \mathcal{O}^{d_i})\}_{i=1}^n$ for constructing individualized rewards $r^{d_i} = g(\tilde{\mathcal{J}}^{d_i}, \mathcal{O}^{d_i})$.

Fig. 3 illustrates the framework applied to the HEMS environment. Fig. 3(a) shows the conventional unified reward approach where all objectives aggregate into a single r, intermixing gradient signals. Fig. 3(b) shows the SCM after Phase 1: each CED has reward node r^d connecting only to objectives it can influence, but $\mathcal{J}_{\text{c}}$ remains connected to all rewards (objective entanglement). Fig. 3(c) shows the SCM after Phase 2: AC and WM satisfy the surgery conditions (single-capability influence on $\mathcal{J}_{\text{c}}$ with primary tasks $\mathcal{J}_{\text{t}}$ and $\mathcal{J}_{\text{l}}$), so local objectives $\mathcal{J}_{\text{c}}^{\text{a}}$ and $\mathcal{J}_{\text{c}}^{\text{w}}$ are created; BESS retains global $\mathcal{J}_{\text{c}}$ as it affects cost through both charging and discharging. The resulting attributed sets are: $\tilde{\mathcal{J}}^{\text{a}} = \{\mathcal{J}_{\text{t}}, \mathcal{J}_{\text{c}}^{\text{a}}\}$, $\tilde{\mathcal{J}}^{\text{b}} = \{\mathcal{J}_{\text{c}}\}$, $\tilde{\mathcal{J}}^{\text{w}} = \{\mathcal{J}_{\text{l}}, \mathcal{J}_{\text{c}}^{\text{w}}\}$. By (11), the causally derived observables are: $\mathcal{O}^{\text{a}} = \{T_t^{\text{in}}, P_t^{\text{a}}\}$; $\mathcal{O}^{\text{b}} = \{P_t^{\text{a}}, P_t^{\text{w}}, P_t^{\text{b}}\}$ [aggregating to P_t^{D} via (7)]; $\mathcal{O}^{\text{w}} = \{M_t^{\text{w}}, P_t^{\text{w}}\}$.

For the considered HEMS environment, the individualized reward functions instantiated from Algorithm 1 are

$$r_t^{\text{a}} = \underbrace{\text{occ}_t \cdot \exp\left(\min\left(0, T^{\text{band}} - \left|T_t^{\text{in}} - T_t^{\text{set}}\right|\right)\right)}_{\text{from } \mathcal{J}_{\text{t}}, \text{ denoted } r_t^{\text{t}}} - \underbrace{k^{\text{a}} \kappa_t P_t^{\text{a}}}_{\text{from } \mathcal{J}_{\text{c}}^{\text{a}}} \tag{12a}$$

Algorithm 1 Causal Discovery and Surgery for Reward Attribution

Require: CED set $\mathcal{D}$, Objective set $\mathcal{J}$, Capability function $\mathcal{C}(\cdot)$
Ensure: SCM $\mathcal{G}$, attributed objectives $\{\tilde{\mathcal{J}}^{d_i}\}$, observables $\{\mathcal{O}^{d_i}\}$

```
# Phase 1: Causal Discovery
Initialize SCM G with nodes: D, O, J
for each CED d_i ∈ D do
    Add reward node r^{d_i} and policy node π^{d_i} to G
    J^{d_i} ← ∅
    for each objective J_j ∈ J do
        C_match ← C(d_i) ∩ C(J_j)
        if C_match ≠ ∅ then
            J^{d_i} ← J^{d_i} ∪ {J_j}
            Add edges: d_i → o_j → J_j for o_j ∈ O(C_match)
        end if
    end for
    Add edges: J_j → r^{d_i} for all J_j ∈ J^{d_i}, r^{d_i} → π^{d_i}
end for
# Phase 2: Causal Surgery
for each confounded J_j (i.e., |{d_i : J_j ∈ J^{d_i}}| > 1) do
    for each d_i where J_j ∈ J^{d_i} do
        C_match ← C(d_i) ∩ C(J_j)
        if |C_match| = 1 and |J^{d_i}| > 1 then
            Create local objective J_j^{d_i}; add to G
            Redirect: remove J_j → r^{d_i}, add J_j^{d_i} → r^{d_i}
            Update J^{d_i}: replace J_j with J_j^{d_i}
        end if
    end for
end for
# Phase 3: Backward Extraction
for each CED d_i do
    J̃^{d_i} ← Pa(r^{d_i}) ∩ J*
    O^{d_i} ← ⋃_{J_j ∈ J̃^{d_i}} Pa(J_j) ∩ O
end for
return G, {(J̃^{d_i}, O^{d_i})}_{i=1}^n
```

$$r_t^{\mathrm{b}} = \underbrace{\exp\left(-\left|P_t^{\mathrm{D}}\right| \cdot \kappa_t\right)}_{\text{from } \mathcal{J}_{\mathrm{c}}\text{, denoted } r_t^{\mathrm{c}}} \tag{12b}$$

$$r_t^{\mathrm{w}} = \underbrace{\exp\left(-\Delta t_t^{\mathrm{w}}\right)}_{\text{from } \mathcal{J}_{\mathrm{l}}\text{, denoted } r_t^{\mathrm{l}}} \underbrace{-k^{\mathrm{w}}\kappa_t P_t^{\mathrm{w}}}_{\text{from } \mathcal{J}_{\mathrm{c}}^{\mathrm{w}}} \tag{12c}$$

where $\Delta t_t^{\mathrm{w}} = |t - t^{\mathrm{lau}} - z^{\mathrm{lau}}|$ if $\mathrm{lau}_t = 1, t > (t^{\mathrm{lau}} + z^{\mathrm{lau}})$ and $\sum_{\tau=t^{\mathrm{lau}}}^{t} M_\tau^{\mathrm{w}} = 0$ (i.e., laundry requested but not yet started), and $\Delta t_t^{\mathrm{w}} = 0$ otherwise. Weights k^{a}, k^{w} balance primary tasks against local cost objectives. In practical deployments, these weights can be preconfigured by system integrators or mapped to user-selectable preference modes (e.g., "comfort-priority" versus "cost-priority"), rather than requiring manual tuning by end-users. Since preference-aware policy adaptation is orthogonal to our focus on causal reward attribution, we leave this extension to future work. The conventional unified reward aggregates all objectives into a single scalar [following Fig. 3(a)]:

$$r_t^{\mathrm{uni}} = w_{\mathrm{c}} \cdot r_t^{\mathrm{c}} + w_{\mathrm{t}} \cdot r_t^{\mathrm{t}} + w_{\mathrm{l}} \cdot r_t^{\mathrm{l}}. \tag{13}$$

The weights $w_{\mathrm{c}}, w_{\mathrm{t}}, w_{\mathrm{l}}$ balance objectives but provide no CED-specific guidance. For a fair comparison, we set $w_{\mathrm{c}} = w_{\mathrm{t}} = w_{\mathrm{l}} = 1$.

B. HR-MASAC: MDP Formulation and Algorithm

A Markov decision process (MDP) formally defines the agent–environment interaction in reinforcement learning. At each time step t, the agent observes state o_t, selects action a_t according to policy $\pi(a_t|o_t)$, and receives reward r_t as the environment transitions to o_{t+1}. Conventional multiagent MDPs use a shared scalar reward $r_t \in \mathbb{R}$, which does not distinguish individual contributions.

To support heterogeneous reward attribution, we extend the formulation to a multiagent MDP with vector-valued rewards: $\mathcal{M} = \langle \mathcal{D}, \mathcal{S}, \{\mathcal{A}^i\}, \mathcal{P}, \mathcal{O}, \mathbf{R}, \gamma \rangle$. Each CED $d_i \in \mathcal{D}$ is controlled by a dedicated agent i with action space $\mathcal{A}^i$ and policy π^i. For notational simplicity, we use superscript i to index agents in the following DRL formulations, with the understanding that agent i controls CED d_i. The key extension is the reward function $\mathbf{R} : \mathcal{S} \times \prod_i \mathcal{A}^i \to \mathbb{R}^n$, which outputs $\mathbf{r}_t = (r_t^1, \ldots, r_t^n)$ where each r_t^i is the individualized reward designed according to the causal attribution framework (Section IV-A). The discount factor $\gamma \in (0, 1]$ determines the importance of future rewards, and $\mathcal{P}$ denotes the transition dynamics instantiated through the environment model in Section III.

1) Observation and Action Spaces: The state $s_t \in \mathcal{S}$ comprises temporal features U_t, weather conditions W_t, CED status D_t, electrical status G_t, and user behavior B_t. The observation $o_t \in \mathcal{O}$ excludes unobservable variables (e.g., thermal conductivity). For the HEMS environment

$$o_t = \{U_t, W_t, D_t, G_t, B_t\} \tag{14}$$

where U_t includes time-of-day encoding, $W_t = \{T_t^{\mathrm{out}}, T_t^{\mathrm{in}}, I_t\}$, $D_t = \{M_t^d, P_t^d, \mathrm{soc}_t\}_{d \in \mathcal{D}}$, $G_t = \{P_t^{\mathrm{PV}}, P_t^{\mathrm{BL}}, P_t^{\mathrm{D}}, P_t^{\mathrm{G}}, \kappa_t\}$, and $B_t = \{\mathrm{occ}_t, T_t^{\mathrm{set}}, \mathrm{lau}_t, z_t^{\mathrm{w}}\}$. All agents share this common observation.

The action space combines discrete mode selection and continuous setpoints

$$a_t = \left\{M_t^d, \bar{P}_t^d\right\}_{d \in \{\mathrm{a,b}\}} \cup \left\{A_t^{\mathrm{w}}\right\} \tag{15}$$

where $M_t^d \in \{-1, 0, 1\}$ specifies operating mode, $\bar{P}_t^d \in [0, P_{\mathrm{rat}}^d]$ specifies target power for AC and BESS, and $A_t^{\mathrm{w}} \in \{0, 1\}$ is the WM start signal. Continuous actions are sampled via reparameterized Gaussian distributions, and discrete actions via categorical distributions.

2) Algorithm Design: To enable each agent to learn an optimal policy under causally attributed heterogeneous rewards while maintaining coordination, we propose **HR-MASAC**. The algorithm follows the CTDE paradigm (Fig. 4), with two core innovations: a multihead centralized critic and agent-specific entropy temperatures.

The multihead critic outputs agent-specific Q-values

$$\mathbf{q}_t = Q\left(o_t, \mathbf{a}_t; \mathbf{w}\right) = \left\{Q^i\left(o_t, \mathbf{a}_t; w^i\right)\right\}_{i=1}^n \tag{16}$$

where all heads share a common encoder and w^i parameterizes the ith head. We adopt twin Q-networks to reduce overestimation [13].

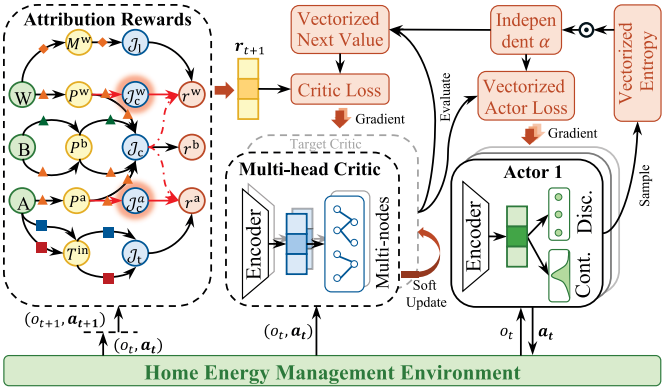

Fig. 4. HR-MASAC framework with CTDE. The multihead critic outputs agent-specific Q-values during training.

HR-MASAC is a maximum entropy *RL* method, whose optimization objective includes both the expected return and the entropy of the policy. The soft value function is

$$\mathbf{v}_t = \mathbb{E}_{\mathbf{a}_{t+1}\sim\{\pi^i\}}\left[Q\left(o_{t+1}, \mathbf{a}_{t+1}; \mathbf{w}'\right) - \boldsymbol{\alpha} \odot \left\{\log \pi^i\left(a^i_{t+1}|o_{t+1}\right)\right\}_{i=1}^{n}\right] \tag{17}$$

where $\mathbf{w}'$ denotes the parameters of the target critic, and $\boldsymbol{\alpha} = \{\alpha^i\}$ represents the vector of agent-specific temperature coefficients. $\odot$ denotes element-wise multiplication.

The critic network is trained by minimizing the soft Bellman temporal-difference (TD) error over transitions sampled from replay buffer $\mathcal{B}$

$$\mathcal{L}_{\text{critic}} = \mathbb{E}_{(o_t,\mathbf{a}_t,\mathbf{r}_{t+1},o_{t+1})\sim\mathcal{B}}\left[\left\|Q\left(o_t, \mathbf{a}_t; \mathbf{w}\right) - \left(\mathbf{r}_{t+1} + \gamma \mathbf{v}_t\right)\right\|^2\right]. \tag{18}$$

Each actor is optimized independently by maximizing the entropy-augmented expected return, leading to the loss

$$\mathcal{L}^i_{\text{actor}} = \mathbb{E}_{o_t\sim\mathcal{B}, a^i_t\sim\pi^i}\left[\alpha^i \log \pi^i\left(a^i_t|o_t\right) - Q^i\left(o_t, \mathbf{a}_t; w^i\right)\right]. \tag{19}$$

Each agent maintains a learnable temperature α^i, which is adjusted to match a given target entropy $\mathcal{H}^i_{\text{target}}$

$$\mathcal{L}_\alpha = \mathbb{E}_{a^i_t\sim\pi^i}\left[-\alpha^i \cdot \left(\log \pi^i\left(a^i_t|o_t\right) + \mathcal{H}^i_{\text{target}}\right)\right]. \tag{20}$$

Following standard practice in SAC [13], the target entropy is set as $\mathcal{H}^i_{\text{target}} = -\dim(\mathcal{A}^i)$, where $\dim(\mathcal{A}^i)$ is the dimensionality of agent i's continuous action space. This choice ensures sufficient exploration while remaining computationally tractable. The agent-specific entropy coefficients α^i are automatically tuned via gradient descent during training, eliminating the need for manual per-agent calibration.

Finally, the parameters of the target critic are updated using Polyak averaging

$$\mathbf{w}' \leftarrow \rho \mathbf{w} + (1-\rho)\,\mathbf{w}', \quad \rho \in (0,1). \tag{21}$$

HR-MASAC's training loop is shown in Algorithm 2.

Algorithm 2 HR-MASAC Training Procedure

Require: Actor networks $\{\pi^i_{\theta^i}\}_{i=1}^n$, critic $Q_{\mathbf{w}}$, target critic $Q_{\mathbf{w}'}$, temperatures $\{\alpha^i\}_{i=1}^n$, replay buffer $\mathcal{B}$
Ensure: Trained policies $\{\pi^i_{\theta^i}\}_{i=1}^n$

for each training iteration **do**
  for each environment step **do**
    Observe o_t from environment
    Sample actions: $a^i_t \sim \pi^i(o_t; \theta^i)$ for all $i = 1, \ldots, n$
    Execute joint action $\mathbf{a}_t$ in environment
    Receive next observation o_{t+1} and heterogeneous rewards $\mathbf{r}_{t+1}$
    Store transition $(o_t, \mathbf{a}_t, \mathbf{r}_{t+1}, o_{t+1})$ into $\mathcal{B}$
  end for
  for each gradient step **do**
    Sample mini-batch $\{(o_t, \mathbf{a}_t, \mathbf{r}_{t+1}, o_{t+1})\}$ from $\mathcal{B}$
    # Critic update
    Sample next actions $a^i_{t+1} \sim \pi^i(o_{t+1})$ for all i
    Compute soft target values according to (17)
    Minimize critic loss according to (18)
    # Actor and temperature update
    Sample $a^i_t \sim \pi^i(o_t)$ for all i, compute $\log \pi^i(a^i_t|o_t)$
    for each agent $i = 1$ to n **do**
      Evaluate $Q^i(o_t, \mathbf{a}_t; w^i)$
      Minimize actor loss according to (19)
      Minimize temperature loss according to (20)
    end for
    Target critic soft update according to (21)
  end for
end for

V. Experiment

This section presents comprehensive experiments to evaluate the proposed causal reward attribution framework and the HR-MASAC algorithm. We first describe the experimental setup and baseline methods. We then present the main results, including overall performance comparison across diverse home scenarios and agent behavior analysis. Finally, we provide in-depth analysis and discussion covering: ablation studies quantifying individual component contributions, generalizability and scalability analysis from three perspectives (MARL architecture, device configuration, and computational overhead), and robustness evaluation under exogenous data perturbations and household heterogeneity.

A. Experiment Setup

To ensure experimental realism and reliability, we instantiate the modeling approach described in Section III-C1 into five distinct smart home energy management simulation environments. These environments differ significantly in terms of physical structure, demand distribution, and appliance configurations. The physical and statistical parameters are generated using large language model [33] to align with real-world scenarios. Historical weather and electricity consumption data are sourced from the NOAA and PecanStreet datasets [31], [32], specifically from residential communities in San Diego, California. The control interval is set to $\Delta t = 5$ min, and

each episode spans a full day ($\tau = 288$ time steps). For each month, two days are selected as validation and test sets, while the remaining days are used for training.

In HR-MASAC, both the critic and actor networks adopt a symmetric MLP backbone with hidden layers sized [256, 128, 64, 64, 128, 256]. All baseline algorithms use the same backbone architecture and hyperparameters, with only output heads differing according to their algorithmic requirements, ensuring fair comparisons. Reward weights are set to $k^{\mathrm{a}} = k^{\mathrm{w}} = 0.02$.

All experiments are conducted on a workstation with Ubuntu 20.04, Python 3.10, PyTorch 2.0, an Intel i7-14600K CPU, and an NVIDIA RTX 4090 GPU.

B. Baselines

We compare our method against several categories of baselines as follows.

1) *RBC*: A heuristic baseline implementing standard ON/OFF control strategies following expert-defined rules [34].
2) *Single-Agent RL*: Single-agent actor–critic baselines with a unified reward (13), including SAC [13], DDPG [35], PPO [36], and TD3 [37]. The global state and joint action spaces are concatenated as input.
3) *Unified-Reward Centralized Critic MARL*: Multiagent baselines with a unified reward (13) and a single shared critic, including MASAC [38] and MADDPG [39]. Each agent maintains a decentralized actor while sharing the centralized critic during training.
4) *Value Decomposition (QMIX-SAC)*: Integrates QMIX's monotonic value decomposition [4] into the SAC framework. A mixing network combines individual Q-values Q^i into a global Q_{tot} under the constraint $(\partial Q_{\mathrm{tot}})/(\partial Q^i) \geq 0$, enabling decentralized execution while maintaining centralized training with the unified reward.
5) *Counterfactual Credit Assignment (COMA-SAC)*: Integrates COMA with SAC. The centralized critic computes a counterfactual baseline $b^i(o_t, \mathbf{a}_t^{-i}) = \sum_{a^i} \pi^i(a^i|o_t) Q(o_t, \mathbf{a}_t^{-i}, a^i)$ for each agent, and the advantage $A^i = Q(o_t, \mathbf{a}_t) - b^i$ is used for policy updates with the unified reward.
6) *Heterogeneous-Reward Independent Critics MARL*: Each agent is assigned a causally attributed reward (12) and paired with an independent critic, including I-MASAC and I-MADDPG. This decouples agent learning but lacks the shared critic structure for coordination.
7) Ablation variants:
   a) *w/o surgery*: A variant without Phase 2 (causal surgery) of Algorithm 1. Each agent receives rewards from all objectives it can influence [as in Fig. 3(b)], without decomposing shared objectives like $\mathcal{J}_{\mathrm{c}}$ into CED-local variants;
   b) *w/o ind. α*: A variant with a shared entropy coefficient across agents.

All experiments are conducted with five independent random seeds, and we report means with 95% confidence intervals (CI) computed via bootstrapping.

C. Performance Comparison

Fig. 5 illustrates the overall reward trends, variations in different reward components ($r^{\mathrm{a}}, r^{\mathrm{b}}, r^{\mathrm{w}}$), and entropy coefficient evolution during training. Note that while unified-reward baselines (e.g., SAC and COMA-SAC) use (13) for training and causal attribution methods (e.g., HR-MASAC) use (12), all methods are evaluated using the same causally attributed rewards, ensuring comparable episode returns. As shown in Figs. 5(a) and (b), multiagent algorithms generally achieve higher overall rewards than single-agent algorithms. Among single-agent baselines, only SAC and PPO demonstrate effective learning, indicating that independent actor parameters are better suited for heterogeneous CED control.

As shown in Fig. 5 and Table I, baselines based on maximum entropy reinforcement learning (SAC, MASAC, I-MASAC, and HR-MASAC) significantly outperform other algorithms, suggesting that learnable entropy terms effectively enhance exploration. Among these, HR-MASAC and I-MASAC, which perceive heterogeneous attributional rewards, perform better, demonstrating that vectorized rewards more effectively guide CED-specific optimization.

As shown in Fig. 5(d)–(f), the convergence patterns of reward components differ markedly. All baselines except MADDPG quickly reach optimal values for r^{w} with minimal variance, while r^{a} and r^{b} converge slower and more rugged—especially r^{b}. This correlates with CED-specific physical constraints and task complexity, validating the core motivation that different CEDs follow distinct optimization paths. Performance differences among baselines mainly stem from their ability to optimize the challenging r^{b}: HR-MASAC and I-MASAC achieve significantly higher convergence values than other baselines. Moreover, HR-MASAC outperforms I-MASAC in both r^{b} and r^{a} during later training stages, as the shared encoder in its critic facilitates interagent coordination required for BESS optimization.

As shown in Fig. 5(c), the learnable entropy coefficients α in HR-MASAC rapidly decrease and then diverge distinctly among agents, indicating that each agent requires an independent exploration–exploitation balance. Using a shared entropy coefficient (w/o ind. α) significantly degrades performance with noticeable drops in r^{a} and r^{b} convergence, highlighting the importance of independent α for heterogeneous optimization goals.

Table I shows that the performance of each baseline across five households is consistent with their reward convergence behaviors in Fig. 5. HR-MASAC consistently performs the best across all homes and objectives, achieving satisfactory comfort and meeting user laundry demands with lower energy costs. The narrow 95% confidence intervals indicate stable performance across random seeds.

Notably, QMIX-SAC and COMA-SAC, despite incorporating implicit credit assignment mechanisms, exhibit distinct failure modes. QMIX-SAC performs poorly on both thermal comfort ($\mathcal{J}_{\mathrm{t}} \approx 9.5$) and electricity cost ($\mathcal{J}_{\mathrm{c}} \approx 12.3$), likely due to its monotonicity constraint ($(\partial Q_{\mathrm{tot}})/(\partial Q^i) \geq 0$) preventing proper penalty signals. COMA-SAC achieves adequate

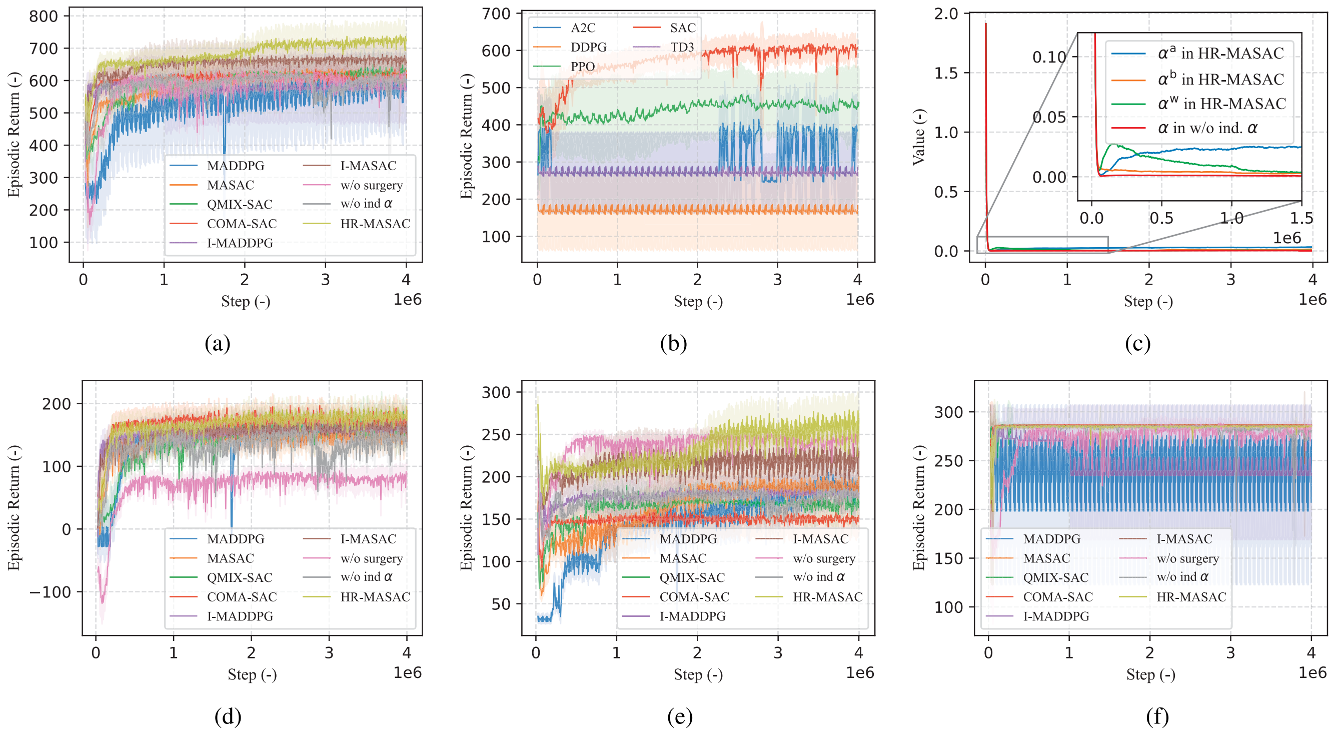

Fig. 5. Training dynamics over 4 000 000 steps. Shaded regions indicate 95% CI across five seeds. (a) MARL baselines. (b) SARL baselines. (c) Entropy coefficient α in HR-MASAC. (d) r^{a}. (e) r^{b}. (f) r^{w}.

TABLE I

PERFORMANCE COMPARISON ACROSS HOMES (MEAN ± 95% CI OVER 5 SEEDS). THE UNITS OF $\mathcal{J}_{\text{C}}$, $\mathcal{J}_{\text{T}}$, AND $\mathcal{J}_{\text{L}}$ ARE (/H), (%), AND (%), RESPECTIVELY. LOWER VALUES INDICATE BETTER PERFORMANCE FOR ALL METRICS. BEST RESULTS IN **BOLD**

	Home 1			Home 2			Home 3			Home 4			Home 5		
	$\mathcal{J}_\text{c}$	$\mathcal{J}_\text{t}$	$\mathcal{J}_\text{l}$	$\mathcal{J}_\text{c}$	$\mathcal{J}_\text{t}$	$\mathcal{J}_\text{l}$	$\mathcal{J}_\text{c}$	$\mathcal{J}_\text{t}$	$\mathcal{J}_\text{l}$	$\mathcal{J}_\text{c}$	$\mathcal{J}_\text{t}$	$\mathcal{J}_\text{l}$	$\mathcal{J}_\text{c}$	$\mathcal{J}_\text{t}$	$\mathcal{J}_\text{l}$
RBC	$14.25_{\pm0.31}$	$12.10_{\pm0.28}$	$\mathbf{0.00}_{\pm0.00}$	$18.73_{\pm0.38}$	$10.24_{\pm0.24}$	$\mathbf{0.00}_{\pm0.00}$	$11.82_{\pm0.26}$	$15.67_{\pm0.35}$	$\mathbf{0.00}_{\pm0.00}$	$16.94_{\pm0.34}$	$13.45_{\pm0.30}$	$\mathbf{0.00}_{\pm0.00}$	$15.38_{\pm0.32}$	$11.23_{\pm0.26}$	$\mathbf{0.00}_{\pm0.00}$
Single-Agent RL															
SAC	$8.85_{\pm0.42}$	$7.10_{\pm0.38}$	$4.19_{\pm0.31}$	$12.34_{\pm0.51}$	$5.82_{\pm0.29}$	$2.86_{\pm0.22}$	$7.23_{\pm0.35}$	$9.45_{\pm0.47}$	$5.52_{\pm0.38}$	$10.57_{\pm0.48}$	$8.31_{\pm0.41}$	$3.61_{\pm0.27}$	$9.42_{\pm0.44}$	$6.17_{\pm0.33}$	$4.87_{\pm0.35}$
Unified-Reward MARL															
MADDPG	$11.21_{\pm0.58}$	$7.94_{\pm0.43}$	$5.16_{\pm0.39}$	$14.87_{\pm0.67}$	$6.58_{\pm0.35}$	$3.57_{\pm0.28}$	$9.43_{\pm0.49}$	$10.82_{\pm0.54}$	$6.90_{\pm0.45}$	$12.95_{\pm0.61}$	$8.67_{\pm0.45}$	$4.17_{\pm0.32}$	$11.78_{\pm0.55}$	$7.35_{\pm0.39}$	$5.90_{\pm0.41}$
MASAC	$9.78_{\pm0.47}$	$5.02_{\pm0.28}$	$5.48_{\pm0.37}$	$13.46_{\pm0.59}$	$4.31_{\pm0.24}$	$3.81_{\pm0.29}$	$8.15_{\pm0.41}$	$6.89_{\pm0.36}$	$7.24_{\pm0.48}$	$11.82_{\pm0.54}$	$5.74_{\pm0.31}$	$4.44_{\pm0.33}$	$10.34_{\pm0.49}$	$4.58_{\pm0.26}$	$6.15_{\pm0.42}$
Value Decomposition & Credit Assignment															
QMIX-SAC	$12.31_{\pm0.58}$	$9.52_{\pm0.51}$	$2.53_{\pm0.21}$	$15.87_{\pm0.72}$	$8.24_{\pm0.44}$	$1.85_{\pm0.16}$	$10.45_{\pm0.52}$	$11.73_{\pm0.59}$	$3.28_{\pm0.25}$	$14.23_{\pm0.66}$	$10.15_{\pm0.53}$	$2.17_{\pm0.19}$	$13.08_{\pm0.61}$	$8.87_{\pm0.47}$	$2.74_{\pm0.22}$
COMA-SAC	$10.23_{\pm0.52}$	$2.64_{\pm0.19}$	$2.87_{\pm0.23}$	$13.58_{\pm0.63}$	$2.31_{\pm0.17}$	$2.15_{\pm0.18}$	$8.47_{\pm0.44}$	$3.42_{\pm0.24}$	$3.56_{\pm0.27}$	$12.15_{\pm0.57}$	$2.89_{\pm0.20}$	$2.48_{\pm0.20}$	$11.02_{\pm0.53}$	$2.52_{\pm0.18}$	$3.04_{\pm0.24}$
Independent MARL with Heterogeneous Rewards															
I-MADDPG	$9.35_{\pm0.48}$	$5.90_{\pm0.34}$	$0.97_{\pm0.12}$	$12.78_{\pm0.57}$	$4.83_{\pm0.28}$	$0.48_{\pm0.08}$	$7.92_{\pm0.43}$	$7.45_{\pm0.41}$	$1.38_{\pm0.15}$	$11.23_{\pm0.53}$	$6.57_{\pm0.37}$	$0.83_{\pm0.11}$	$9.87_{\pm0.47}$	$5.24_{\pm0.31}$	$1.13_{\pm0.13}$
I-MASAC	$7.35_{\pm0.38}$	$2.45_{\pm0.18}$	$1.29_{\pm0.14}$	$10.23_{\pm0.49}$	$2.08_{\pm0.15}$	$0.71_{\pm0.10}$	$5.87_{\pm0.33}$	$3.62_{\pm0.24}$	$1.38_{\pm0.15}$	$8.96_{\pm0.44}$	$2.91_{\pm0.20}$	$1.94_{\pm0.18}$	$7.68_{\pm0.40}$	$2.17_{\pm0.16}$	$1.54_{\pm0.16}$
Ablation Variants															
w/o surgery	$6.73_{\pm0.36}$	$13.82_{\pm0.68}$	$5.83_{\pm0.41}$	$9.45_{\pm0.47}$	$11.54_{\pm0.57}$	$4.27_{\pm0.32}$	$5.52_{\pm0.31}$	$16.35_{\pm0.79}$	$7.14_{\pm0.48}$	$8.17_{\pm0.42}$	$14.28_{\pm0.69}$	$5.42_{\pm0.38}$	$7.08_{\pm0.38}$	$12.67_{\pm0.62}$	$6.18_{\pm0.43}$
w/o ind. α	$8.17_{\pm0.43}$	$6.14_{\pm0.35}$	$2.26_{\pm0.19}$	$11.45_{\pm0.52}$	$5.23_{\pm0.30}$	$1.43_{\pm0.14}$	$6.78_{\pm0.37}$	$8.37_{\pm0.44}$	$3.10_{\pm0.24}$	$9.84_{\pm0.47}$	$7.05_{\pm0.39}$	$2.22_{\pm0.19}$	$8.52_{\pm0.43}$	$5.68_{\pm0.33}$	$2.56_{\pm0.21}$
HR-MASAC	$\mathbf{5.81}_{\pm0.32}$	$\mathbf{2.15}_{\pm0.16}$	$1.29_{\pm0.14}$	$\mathbf{8.34}_{\pm0.43}$	$\mathbf{1.86}_{\pm0.14}$	$0.95_{\pm0.11}$	$\mathbf{4.52}_{\pm0.27}$	$\mathbf{2.94}_{\pm0.21}$	$1.72_{\pm0.16}$	$\mathbf{7.26}_{\pm0.38}$	$\mathbf{2.53}_{\pm0.18}$	$1.67_{\pm0.16}$	$\mathbf{6.13}_{\pm0.34}$	$\mathbf{1.92}_{\pm0.15}$	$1.54_{\pm0.15}$

comfort ($\mathcal{J}_\text{t} \approx 2.6$) but elevated costs ($\mathcal{J}_\text{c} \approx 10.2$), suggesting that while counterfactual baselines handle single-path causal influences (e.g., AC $\rightarrow \mathcal{J}_\text{t}$), they struggle with multipath objectives like $\mathcal{J}_\text{c}$ where multiple CEDs contribute through different mechanisms. The detailed behavioral analysis is provided in Section V-D.

These results confirm that implicit credit assignment methods designed for single-objective cooperative tasks are inadequate for HEMS, where agents must balance cooperation and competition across multiple objectives—a finding consistent with QMIX's original scope of fully cooperative settings [4]. HR-MASAC achieves 43.2% lower cost than COMA-SAC and 52.8% lower cost than QMIX-SAC on average, demonstrating the necessity of explicit causal reward attribution. Furthermore, compared to RBC, all DRL-based methods capable of effective learning achieve better overall performance, demonstrating the advantage of learned policies over heuristic rules.

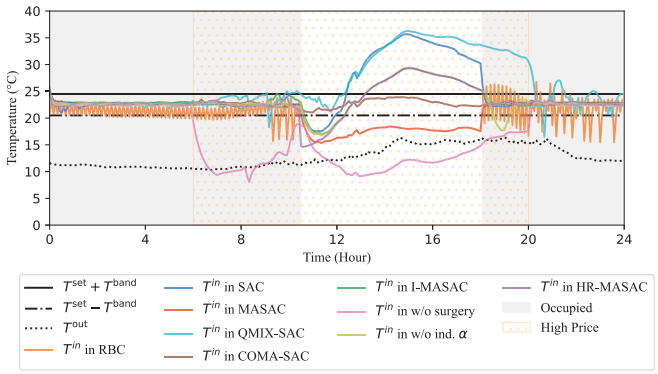

Fig. 6. Temperature curves of a typical day in Home 1.

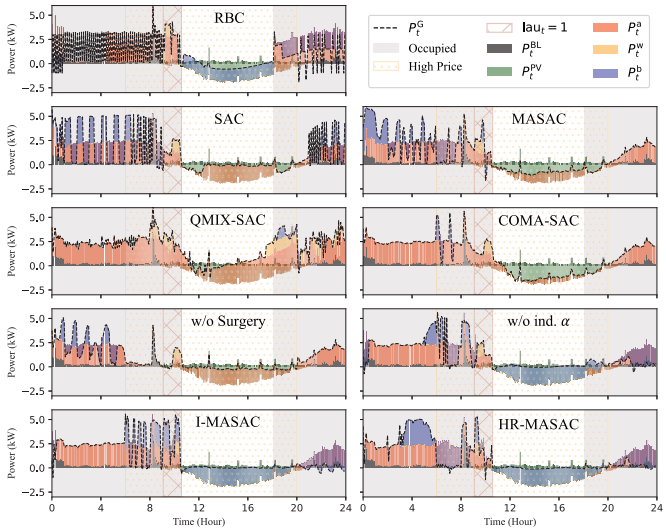

Fig. 7. Power profiles of a typical day in Home 1.

D. Agent Behavior Analysis

To understand why HR-MASAC achieves superior performance, we analyze the behavioral patterns of individual agents through temperature curve, power profile, and mutual information (MI) metrics.

From Fig. 6, we observe that maximum entropy-based baselines adjust indoor temperatures to comfortable ranges when the house is occupied. Unlike the threshold-triggered control used in RBC, DRL-based policies directly output appropriate AC power levels, resulting in smoother temperature curves. A key observation is that in SAC and MASAC with unified rewards, the AC remains active even when the house is unoccupied—SAC opts for heating while MASAC for cooling. Since the comfort reward r^{a} becomes negative in unoccupied states due to the electricity cost component, this suggests r^{a} is not the dominant signal in these cases. Further inspection of the power profiles in Fig. 7 shows that AC is activated to balance surplus PV energy, indicating that r^{b} is the dominant reward. This highlights the failure of unified rewards to properly attribute rewards in heterogeneous agent settings, causing agents to follow suboptimal, easier optimization paths—reinforcing the core motivation shown in Fig. 1.

The behavioral patterns of QMIX-SAC, COMA-SAC, and the ablation variant w/o surgery reveal distinct failure modes. QMIX-SAC exhibits erratic AC control, failing to reduce power output promptly and allowing indoor temperatures to deviate significantly from the comfort zone. As shown in Fig. 7, both QMIX-SAC and COMA-SAC barely utilize the BESS throughout the day. QMIX-SAC relies on AC to balance household power during PV surplus periods, similar to unified-reward methods (SAC and MASAC). COMA-SAC demonstrates more reasonable primary task control: it correctly activates the AC to maintain thermal comfort and only operates the WM when laundry is requested. However, COMA-SAC leaves much of the PV generation unutilized during midday periods, resulting in elevated electricity costs. These observations indicate that implicit credit assignment methods fail to provide effective gradient signals for high-exploration-difficulty agents (BESS) when facing entangled multiobjective rewards, causing the shared cost objective to be preferentially satisfied by agents with simpler constraints (i.e., AC).

The w/o surgery variant presents a particularly instructive case. Without causal surgery, the AC agent receives both r_t^{t} and the global cost reward r_t^{c} [see Fig. 3(b)]. As shown in Fig. 6, this causes the AC to over-prioritize cost minimization during high-price periods, leading to insufficient cooling/heating. Moreover, because the AC has simpler physical constraints and responds faster, it preemptively consumes surplus PV energy to balance household demand, leaving the BESS underutilized. This demonstrates that under equal reward weights, shared objectives tend to be preferentially satisfied by CEDs with simpler constraints and lower exploration difficulty, and that without surgical decomposition, agents with competing reward components cannot properly prioritize their designated primary tasks.

In contrast, baselines with proper heterogeneous reward attribution (I-MASAC, HR-MASAC) correctly use the BESS to balance excess PV power, allowing heterogeneous agents to operate collaboratively. Additionally, Fig. 7 shows that HR-MASAC more effectively utilizes the BESS for power balancing and energy arbitrage, which explains its superior performance.

To further verify proper reward attribution, we introduce MI to measure the influence of different CED agents' actions a^d on the reward component r^{b}

$$\mathrm{MI}^d = \iint p\left(a^d, r^{\mathrm{b}}\right) \log\left(\frac{p\left(a^d, r^{\mathrm{b}}\right)}{p\left(a^d\right) p\left(r^{\mathrm{b}}\right)}\right) da^d \, dr^{\mathrm{b}} \tag{22}$$

where $d \in \{\mathrm{a}, \mathrm{b}, \mathrm{w}\}$, $p(\cdot, \cdot)$ denotes the joint probability density function, and $p(\cdot)$ represents the marginal density function. We adopt the Kraskov–Stögbauer–Grassberger (KSG) algorithm [40] based on k-nearest neighbors (k-NN) to estimate MI in practice.

As shown in Fig. 8, the MI distributions reveal distinct reward attribution patterns. Among methods with heterogeneous reward attribution (I-MASAC, HR-MASAC), MI^{b} dominates (45%–78%), confirming that BESS actions

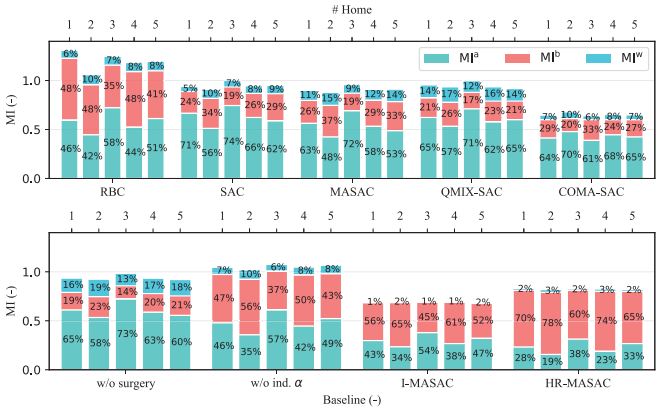

Fig. 8. MI between CED agent actions and r^b across different homes.

primarily drive r^b optimization. In contrast, unified-reward methods (SAC and MASAC) and implicit credit assignment methods (QMIX-SAC and COMA-SAC) exhibit elevated MI^a proportions (48%–74%), reflecting their tendency to use AC rather than BESS for power balancing. The w/o surgery variant shows high MI^a (58%–73%) but low MI^b (14%–23%), confirming that the AC agent dominates at the expense of proper BESS utilization when causal surgery is absent. Note that while absolute MI values do not directly reflect policy performance, comparing relative MI values across agents within the same baseline provides insight into reward attribution.

E. Ablation Study

To quantify the individual contribution of causal surgery and agent-specific entropy adaptation, we analyze the two ablation variants defined in Section V-B. As shown in Table I, removing causal surgery (w/o surgery) leads to dramatically degraded performance on CED-specific objectives (averaged across five homes): $\mathcal{J}_t$ increases from 2.28 to 13.73 (502.2% worse) and $\mathcal{J}_l$ increases from 1.43 to 5.77 (302.2% worse), while $\mathcal{J}_c$ shows only moderate degradation from 6.41 to 7.39 (15.3% worse). This asymmetric degradation—where cost remains acceptable but comfort and laundry suffer severely—confirms that without causal surgery, agents sacrifice their primary tasks to optimize shared objectives. The behavioral patterns underlying this phenomenon are analyzed in Section V-D.

Removing agent-specific entropy coefficients (w/o ind. α) results in more uniform degradation across all objectives: $\mathcal{J}_c$ increases by 39.6%, $\mathcal{J}_t$ by 184.6%, and $\mathcal{J}_l$ by 61.4%. The particularly large impact on thermal comfort suggests that the AC agent requires a distinct exploration–exploitation balance compared to other CEDs. Overall, these results demonstrate complementary roles: causal surgery ensures agents prioritize their designated tasks, while independent entropy adaptation enables efficient exploration within each agent's heterogeneous objective space.

F. Generalizability and Scalability

A key contribution of this work is the causal reward attribution framework (Algorithm 1), which is designed to be general-purpose rather than tailored to specific algorithm variants or CED configurations. We evaluate generalizability from three perspectives: 1) applicability across different MARL base algorithms; 2) adaptability to different CED configurations; and 3) computational scalability as the number of CEDs increases.

TABLE II

PERFORMANCE IMPROVEMENT FROM CAUSAL REWARD ATTRIBUTION ACROSS DIFFERENT MARL BASE ALGORITHMS (AVERAGED ACROSS FIVE HOMES). ↑ INDICATES IMPROVEMENT

	Unified Reward		+ Causal Attribution	
Transition	$\mathcal{J}_c$	$\mathcal{J}_t$	$\mathcal{J}_c$ (↑%)	$\mathcal{J}_t$ (↑%)
MADDPG → I-MADDPG	12.05	8.27	10.23 (15.1%)	6.00 (27.5%)
MASAC → I-MASAC	10.71	5.31	8.02 (25.1%)	2.65 (50.1%)
MASAC → HR-MASAC	10.71	5.31	6.41 (40.1%)	2.28 (57.1%)

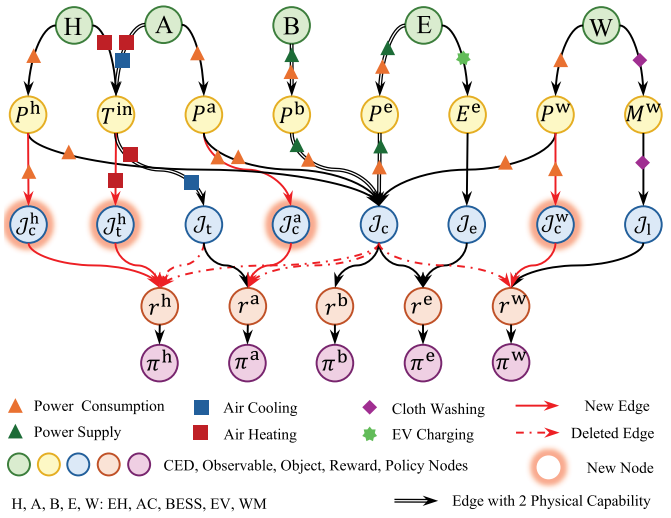

Fig. 9. Extended SCM graph incorporating the EH and EV, obtained by applying Algorithm 1.

1) Generalizability Across MARL Architectures: To demonstrate the framework's generalizability across MARL algorithms, we apply causal reward attribution to different base algorithms and measure performance improvements in Table II.

The results show that causal reward attribution consistently improves performance across all tested MARL algorithms. Even the simplest application (MADDPG with independent critics) achieves 15.1% cost reduction and 27.5% comfort improvement. The improvement is more pronounced for SAC-based methods due to their better exploration capabilities. Notably, the transition from MASAC to HR-MASAC (which adds both causal attribution and multihead critic coordination) yields the largest gains (40.1% and 57.1%), confirming that the centralized multihead architecture effectively leverages causal attribution while maintaining interagent coordination.

2) Generalizability Across CED Configurations: Beyond the AC -BESS-WM configuration studied in the main experiments, the proposed causal reward attribution framework naturally extends to other CED types. Fig. 9 illustrates an extended causal graph incorporating an electric heater (EH)

TABLE III
COMPUTATIONAL COST COMPARISON ACROSS DIFFERENT ARCHITECTURES AND CED SCALES

	3 CEDs		6 CEDs		12 CEDs	
Method	Params	Time/Ep (s)	Params	Time/Ep (s)	Params	Time/Ep (s)
SAC	347.93K	$2.15_{\pm 0.05}$	434.61K	$3.36_{\pm 0.04}$	607.98K	$3.81_{\pm 0.06}$
MASAC	521.76K	$2.92_{\pm 0.04}$	869.17K	$5.40_{\pm 0.28}$	1.56M	$8.30_{\pm 0.31}$
MADDPG	520.99K	$3.51_{\pm 0.08}$	867.63K	$5.63_{\pm 0.15}$	1.56M	$7.86_{\pm 0.30}$
QMIX-SAC	1.04M	$6.63_{\pm 0.04}$	2.43M	$15.10_{\pm 0.29}$	6.22M	$29.53_{\pm 0.21}$
COMA-SAC	521.76K	$6.78_{\pm 0.12}$	869.17K	$10.82_{\pm 0.53}$	1.56M	$21.15_{\pm 0.67}$
I-MASAC	984.98K	$6.89_{\pm 0.05}$	2.31M	$15.07_{\pm 0.16}$	6.00M	$30.23_{\pm 0.39}$
HR-MASAC	522.78K	$3.77_{\pm 0.02}$	871.74K	$6.37_{\pm 0.11}$	1.57M	$10.12_{\pm 0.32}$

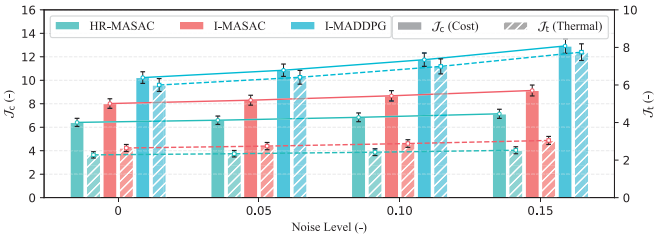

Fig. 10. Robustness under exogenous data perturbations (averaged across five homes). Performance degradation across noise intensity levels β for $\mathcal{J}_c$ and $\mathcal{J}_t$.

with power P^h and an EV with power P^e and battery energy E^e. The EH, like the AC, can regulate indoor temperature and consume electricity, but is limited to heating only. The EV, like the BESS, can both consume and supply electricity, but additionally has an EV charging capability to meet departure requirements. Applying Algorithm 1, the framework outputs: 1) for EH, since it affects thermal comfort and cost through single capabilities, respectively (heating and power consumption), causal surgery creates local objectives, yielding attributed objectives $\tilde{\mathcal{J}}^h = \{\mathcal{J}_t^h, \mathcal{J}_c^h\}$ with observables $\mathcal{O}^h = \{P^h, T^{in}\}$ and 2) for EV, since it affects cost through two capabilities (charging and discharging), the global cost objective $\mathcal{J}_c$ is retained, yielding attributed objectives $\tilde{\mathcal{J}}^e = \{\mathcal{J}_c, \mathcal{J}_e\}$ with observables $\mathcal{O}^e = \{P^h, P^a, P^b, P^e, P^w, E^e\}$. The specific reward functions can then be designed based on these attributed components.

3) Computational Scalability: Extending to more CEDs raises the question of computational overhead. Table III compares representative architectures as the CED count scales from 3 to 12. The results reveal distinct scaling behaviors: methods with per-agent components (I-MASAC and QMIX-SAC) exhibit super-linear parameter growth—QMIX-SAC increases from 1.04M to 6.22M (6×) due to its mixing network, while I-MASAC grows from 0.98M to 6.00M (6.1×) due to independent critics. In contrast, methods with shared architectures (MASAC, COMA-SAC, and HR-MASAC) scale more favorably, with parameters roughly tripling as CEDs quadruple.

HR-MASAC achieves the best tradeoff between efficiency and performance. Compared to I-MASAC, it reduces parameters by 47% at 3 CEDs and 74% at 12 CEDs, while training 1.8× and 3.0× faster, respectively. This efficiency stems from the shared encoder that amortizes feature extraction across agents. Notably, HR-MASAC's overhead relative to MASAC remains modest: only 0.6% more parameters and 22% longer training time at 12 CEDs, yet it achieves substantially better performance through proper credit assignment (Section V-C).

G. Robustness Evaluation

Real-world deployments face uncertainties in exogenous inputs due to sensor measurement errors and forecast inaccuracies. To evaluate robustness under such conditions, we inject Gaussian noise into solar irradiance I_t, base load P_t^{BL}, and outdoor temperature T_t^{out}, simulating noise with variance $\sigma_\epsilon^2 = \beta \cdot \sigma_{data}^2$ where $\beta \in \{0.05, 0.10, 0.15\}$. As shown in Fig. 10, HR-MASAC exhibits the smallest performance degradation among methods with heterogeneous rewards: at $\beta = 0.15$, it experiences only 11% degradation in $\mathcal{J}_c$ compared to 14% for I-MASAC and 26% for I-MADDPG. The larger degradation of I-MADDPG reflects the inherent sensitivity of deterministic policies to input perturbations, whereas maximum entropy methods (HR-MASAC and I-MASAC) benefit from stochastic exploration that provides implicit regularization against noise.

Beyond exogenous perturbations, Table I provides additional evidence of robustness. First, HR-MASAC consistently outperforms all baselines across five homes with diverse physical parameters and demand patterns, demonstrating generalization to household heterogeneity. Second, despite stochastic occupancy and laundry patterns (8) and (9) that vary across random seeds, HR-MASAC maintains narrower 95% confidence intervals than baselines (e.g., ±0.32 versus ±0.47 for MASAC in $\mathcal{J}_c$ of Home 1), indicating stable performance under behavioral uncertainty.

VI. CONCLUSION

This work addresses the reward misattribution problem in DRL-based HEMS through a causally aligned MARL framework. By constructing an SCM that encodes CED-objective causal pathways, we introduce a causal surgery procedure to decompose shared objectives into CED-specific variants, enabling individualized reward attribution. The proposed HR-MASAC algorithm combines a multihead centralized critic with agent-specific entropy coefficients to support coordinated yet specialized policy learning. Experiments across diverse home scenarios demonstrate 40.1% cost reduction and 57.1% comfort improvement over unified-reward baselines, with robust performance under sensor noise. Behavioral analysis confirms that our method successfully leverages the BESS for price-aware power balancing—a capability that unified-reward methods fail to exploit. Future work will investigate the integration of causal discovery from data to automate the SCM construction process, reducing reliance on domain expertise, as well as preference-aware policy adaptation that enables dynamic adjustment of reward weights to accommodate diverse user pReferences without retraining.

REFERENCES

[1] Z. Nagy et al., "Ten questions concerning reinforcement learning for building energy management," *Building Environ.*, vol. 241, Aug. 2023, Art. no. 110435.

[2] A. A. Amer, K. Shaban, and A. M. Massoud, "DRL-HEMS: Deep reinforcement learning agent for demand response in home energy management systems considering customers and operators perspectives," *IEEE Trans. Smart Grid*, vol. 14, no. 1, pp. 239–250, Jan. 2023.

[3] L. Yu, S. Qin, M. Zhang, C. Shen, T. Jiang, and X. Guan, "A review of deep reinforcement learning for smart building energy management," *IEEE Internet Things J.*, vol. 8, no. 15, pp. 12046–12063, Aug. 2021.

[4] T. Rashid, M. Samvelyan, C. S. D. Witt, G. Farquhar, J. Foerster, and S. Whiteson, "Monotonic value function factorisation for deep multi-agent reinforcement learning," *J. Mach. Learn. Res.*, vol. 21, no. 178, pp. 1–51, 2020.

[5] K. Son, D. Kim, W. J. Kang, D. E. Hostallero, and Y. Yi, "QTRAN: Learning to factorize with transformation for cooperative multi-agent reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5887–5896.

[6] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2018, pp. 2974–2982.

[7] A. K. Agogino and K. Tumer, "Analyzing and visualizing multiagent rewards in dynamic and stochastic domains," *Auto. Agents Multi-Agent Syst.*, vol. 17, no. 2, pp. 320–338, Oct. 2008.

[8] Y. Zeng, R. Cai, F. Sun, L. Huang, and Z. Hao, "A survey on causal reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 4, pp. 5942–5962, Nov. 2024.

[9] W. Pinthurat, T. Surinkaew, and B. Hredzak, "An overview of reinforcement learning-based approaches for smart home energy management systems with energy storages," *Renew. Sustain. Energy Rev.*, vol. 202, Sep. 2024, Art. no. 114648.

[10] X. Xu, Y. Jia, Y. Xu, Z. Xu, S. Chai, and C. S. Lai, "A multi-agent reinforcement learning-based data-driven method for home energy management," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3201–3211, Jul. 2020.

[11] A. Salari, S. E. Ahmadi, M. Marzband, and M. Zeinali, "Fuzzy Q-learning-based approach for real-time energy management of home microgrids using cooperative multi-agent system," *Sustain. Cities Soc.*, vol. 95, Aug. 2023, Art. no. 104528.

[12] S. T. Spantideas, A. E. Giannopoulos, and P. Trakadas, "Autonomous price-aware energy management system in smart homes via actor-critic learning with predictive capabilities," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 15018–15033, 2025.

[13] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1861–1870.

[14] G. Wei, M. Chi, Z.-W. Liu, M. Ge, C. Li, and X. Liu, "Deep reinforcement learning for real-time energy management in smart home," *IEEE Syst. J.*, vol. 17, no. 2, pp. 2489–2499, Jun. 2023.

[15] Y. Xu, W. Gao, Y. Li, and F. Xiao, "Operational optimization for the grid-connected residential photovoltaic-battery system using model-based reinforcement learning," *J. Building Eng.*, vol. 73, Aug. 2023, Art. no. 106774.

[16] Z. Tan, M. Zeng, B. Liu, S. Feng, and B. Peng, "Integrating deep reinforcement learning into home energy management system based on soft actor-critic framework," in *Proc. ECITech; Int. Conf. Electr., Control Inf. Technol.*, Mar. 2022, pp. 1–6.

[17] D. Qiu, J. Xue, T. Zhang, J. Wang, and M. Sun, "Federated reinforcement learning for smart building joint peer-to-peer energy and carbon allowance trading," *Appl. Energy*, vol. 333, Mar. 2023, Art. no. 120526.

[18] C. Yu et al., "The surprising effectiveness of PPO in cooperative, multi-agent games," in *Proc. Adv. Neural Inf. Process. Syst.*, Nov. 2021, pp. 24611–24624.

[19] Z. Qin, D. Liu, H. Hua, and J. Cao, "Privacy preserving load control of residential microgrid via deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 12, no. 5, pp. 4079–4089, Sep. 2021.

[20] H. N. Abishu, A. Mohammed Seid, A. Erbad, S. Márquez-Sánchez, J. Hernandez Fernandez, and J. Manuel Corchado, "Multiagent DRL-based demand response optimization for IoT-based smart home energy management systems," *IEEE Internet Things J.*, vol. 12, no. 14, pp. 27003–27020, Jul. 2025.

[21] J. Deng, X. Wang, and F. Meng, "Multi-agent deep reinforcement learning for smart building energy management with chance constraints," *Energy Buildings*, vol. 331, Mar. 2025, Art. no. 115408.

[22] R. Carmona and F. Delarue, *Probabilistic Theory of Mean Field Games With Applications I: Mean Field FBSDEs, Control, and Games*. Cham, Switzerland: Springer, 2018.

[23] J. Subramanian, A. Sinha, R. Seraj, and A. Mahajan, "On the approximation of cooperative heterogeneous multi-agent reinforcement learning (MARL) using mean field control (MFC)," *J. Mach. Learn. Res.*, vol. 23, no. 129, pp. 1–46, 2022.

[24] W. U. Mondal, V. Aggarwal, and S. V. Ukkusuri, "Mean-field approximation of cooperative constrained multi-agent reinforcement learning (CMARL)," *J. Mach. Learn. Res.*, pp. 1–35, 2022.

[25] J. Pearl, *Causality: Models, Reasoning and Inference*, 2nd ed., Cambridge, U.K.: Cambridge Univ. Press, 2009.

[26] B. Huang et al., "Action-sufficient state representation learning for control with structural constraints," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 9260–9279.

[27] F. Lv et al., "Causality inspired representation learning for domain generalization," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, Sep. 2022, pp. 8036–8046.

[28] A. Bibaut, I. Malenica, N. Vlassis, and M. J. V. D. Laan, "More efficient off-policy evaluation through regularized targeted learning," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 654–663.

[29] P. Madumal, T. Miller, L. Sonenberg, and F. Vetere, "Explainable reinforcement learning through action masking," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 2626–2633.

[30] S. S. Shuvo and Y. Yilmaz, "Home energy recommendation system (HERS): A deep reinforcement learning method based on residents' feedback and activity," *IEEE Trans. Smart Grid*, vol. 13, no. 4, pp. 2812–2821, Jul. 2022.

[31] *Pecan Street Database*. Accessed: Jun. 2024. [Online]. Available: http://www.pecanstreet.org/

[32] National Oceanic and Atmospheric Administration (NOAA)., *NOAA Data*. Accessed: Jun. 2024. [Online]. Available: https://www.ncdc.noaa.gov/

[33] J. Achiam et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.

[34] F. De Angelis, M. Boaro, D. Fuselli, S. Squartini, F. Piazza, and Q. Wei, "Optimal home energy management under dynamic electrical and thermal constraints," *IEEE Trans. Ind. Informat.*, vol. 9, no. 3, pp. 1518–1527, Aug. 2013.

[35] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 387–395.

[36] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.

[37] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1587–1596.

[38] Y. Pu, S. Wang, R. Yang, X. Yao, and B. Li, "Decomposed soft actor-critic method for cooperative multi-agent reinforcement learning," 2021, *arXiv:2104.06655*.

[39] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proc. NIPS*, 2017, pp. 6379–6390.

[40] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E, Stat., Nonlinear, Soft Matter Phys.*, vol. 69, no. 6, Jun. 2004, Art. no. 066138.

Xiao Du received the B.E. degree in renewable energy science and engineering from Chongqing University, Chongqing, China, in 2019, and the M.E. degree in energy and power engineering from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2023. He is currently pursuing the Ph.D. degree in software engineering with Chongqing University.

His research interests include deep reinforcement learning, smart home energy management, and energy informatics.

Fengji Luo (Senior Member, IEEE) received the bachelor's and master's degrees in software engineering from Chongqing University, Chongqing, China, in 2006 and 2009, respectively, and the Ph.D. degree in electrical engineering from The University of Newcastle, Newcastle, NSW, Australia, in 2014.

He is currently a Senior Lecturer with the School of Civil Engineering, The University of Sydney, Sydney, NSW, Australia. He held a UUKI Rutherford International Researcher with the Future Energy Institute, Brunel University London, Uxbridge, U.K. He has over 200 publications in these fields. His research interests include computational intelligence, energy demand side management, smart grid, and energy informatics.

Juntao Hu received the B.Eng. degree in electronic information science and technology and the M.E. degree in computer science from Chongqing Normal University, Chongqing, China, in 2018 and 2021, respectively. He is currently pursuing the Ph.D. degree in software engineering with Chongqing University, Chongqing.

His research interests include recommender systems, generative models, computer vision, and energy informatics.

Wei Zhou (Member, IEEE) received the Ph.D. degree in computer science from Chongqing University, Chongqing, China, in 2015.

He was a Researcher with the University of Auckland, Auckland, New Zealand, The Hong Kong Polytechnic University, Hong Kong, China, and Shenzhen Institutes of Advanced Technology, CAS, Shenzhen, China, respectively. He is a Professor with the School of Big Data and Software Engineering, Chongqing University. His research interests include recommendation systems, social computing, data mining, and energy informatics.

Junhao Wen (Member, IEEE) received the Ph.D. degree in computer science from Chongqing University, Chongqing, China, in 2005.

He is a Professor and the Dean with the School of Big Data and Software Engineering, Chongqing University. His research interests include recommendation systems, social computing, service computing, and smart energy systems.