

Article

A Digital-Twin-Aided Safe Multi-Agent Reinforcement Learning Framework for Renewable-Integrated Residential Energy Management

Ziqi Ren ¹, Minglei You ¹ , Marco Rivera ^{1,*}  and Zigeng Fang ² 

¹ Power Electronics, Machines and Control (PEMC) Research Institute, Department of Electrical and Electronic Engineering, Faculty of Engineering, University of Nottingham, Nottingham NG7 2RD, UK; eexzr2@nottingham.ac.uk (Z.R.); minglei.you@nottingham.ac.uk (M.Y.)

² Department of Civil Engineering, Faculty of Engineering, University of Nottingham, Nottingham NG7 2RD, UK; zigeng.fang@nottingham.ac.uk

* Correspondence: marco.rivera@nottingham.ac.uk

Abstract

The increasing penetration of distributed renewable energy sources and electric vehicles (EVs) introduces significant operational challenges for residential energy management systems (HEMS), including stochastic renewable generation, uncertain load demand, device coupling, and physical safety constraints. This paper proposes a digital-twin-aided safe multi-agent reinforcement learning framework for coordinated energy management in renewable-integrated residential systems. The proposed approach models the battery energy storage system and the EV as independent agents and employs a multi-agent soft actor–critic (MASAC) algorithm with a centralised critic to capture the interactions among distributed energy resources. To improve decision quality under uncertainty, a digital twin module is developed to maintain a virtual representation of the residential energy system, synchronise operational states, update degradation-sensitive parameters, and generate short-term predictive information on photovoltaic (PV) generation and household load. The updated digital twin states and forecasts are incorporated into the observations of the reinforcement learning agents. In addition, a safety projection layer is incorporated to improve operational feasibility during both training and deployment. The environment considers realistic residential characteristics, including time-of-use electricity prices, battery degradation, EV mobility patterns, and grid energy trading. Simulation results show that the proposed framework reduces daily energy costs compared with rule-based baselines while maintaining EV charging reliability and operational feasibility. These results highlight the potential of combining predictive information, safety-constrained action execution, and multi-agent reinforcement learning for intelligent residential energy management.



Academic Editor: Dinko Vukadinović

Received: 21 May 2026

Revised: 24 June 2026

Accepted: 26 June 2026

Published: 30 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: residential energy management; digital twin; multi-agent reinforcement learning; soft actor-critic; electric vehicles; battery energy storage; photovoltaic generation; safety constraints

1. Introduction

The rapid deployment of distributed renewable energy resources, residential battery energy storage systems (BESS), and electric vehicles (EVs) is fundamentally transforming residential energy systems [1]. In recent years, rooftop photovoltaic (PV) generation,

household battery storage, and EV charging infrastructures have been increasingly integrated into domestic energy systems to improve renewable energy utilisation, enhance self-consumption, and reduce carbon emissions [1]. Although these technologies provide significant opportunities for more sustainable residential energy operation, they also introduce substantial operational complexity for home energy management systems (HEMS) [2]. In particular, the intermittency of PV generation, stochastic household load demand, time-varying electricity prices, and uncertain EV arrival and departure patterns create a highly dynamic and uncertain decision-making environment [3]. As a result, intelligent energy management strategies are required to coordinate multiple distributed energy resources and achieve efficient, reliable, and cost-effective operation of residential energy systems [4].

Traditional residential energy management strategies are typically based on rule-based control or optimisation-based approaches [5]. Rule-based strategies are simple to implement and easy to interpret, but they often lack flexibility under varying operating conditions and may fail to fully exploit the coordination potential among PV generation, BESS operation, and EV charging [4]. Optimisation-based methods, such as mixed-integer linear programming (MILP) and model predictive control (MPC), have been widely used to address residential energy scheduling problems by explicitly modelling system dynamics, user preferences, and operational constraints [2]. These approaches can achieve high-quality scheduling solutions when accurate system models, forecasts, and problem formulations are available [3]. However, their performance strongly depends on the accuracy of forecasts and models, while their computational burden may increase significantly as more distributed energy resources, constraints, and uncertainty sources are considered [2]. Therefore, more adaptive and scalable control strategies are needed for residential energy management in complex and uncertain environments.

In recent years, reinforcement learning (RL) has emerged as a promising approach for energy management problems due to its ability to learn control policies directly from interactions with the environment without requiring an explicit optimisation model for every operating condition [4]. Deep reinforcement learning (DRL) techniques have been investigated for residential energy management, EV charging coordination, and the joint operation of PV, BESS, and household demand under uncertainty [5]. By learning from historical data or simulated environments, RL-based controllers can adapt to stochastic user behaviour, renewable generation fluctuations, and dynamic system states [5]. Among existing RL algorithms, soft actor–critic (SAC) has attracted considerable attention because of its stability, sample efficiency, and suitability for continuous control problems [6]. Recent work has further demonstrated the potential of SAC-based residential energy management for coordinating household energy resources under stochastic EV behaviour and operational constraints [5].

Related Work and Research Gap

Recent studies have extended residential energy management research from conventional optimisation and single-agent learning methods toward more integrated digital-twin-assisted and multi-agent learning frameworks. One important research direction is the integration of digital twins with reinforcement learning and predictive energy management. Digital twins have been used to construct virtual representations of physical energy systems, support online monitoring, update system states, provide short-term forecasting, and assist intelligent decision-making. For example, Zhou et al. investigated the application of digital twins for reinforcement-learning-based optimal scheduling and reliability management enhancement of energy systems [7]. Sado et al. proposed a digital-twin-based forecasting framework for power flow management in DC microgrids, showing the value of digital twins in real-time monitoring and predictive operation [8]. In battery-related applications,

recent studies have further combined digital twins with state estimation, transfer learning, and AI-assisted modelling for lithium-ion battery management [9,10]. These studies indicate that a digital twin should not be regarded as a simple forecasting block only; rather, it can provide virtual representation, state synchronisation, parameter updating, predictive support, and controller interaction. However, the integration of digital-twin-based state updating and forecasting with safe multi-agent reinforcement learning for residential PV-BESS-EV energy management remains insufficiently studied.

A second relevant direction is safe reinforcement learning for constrained energy management. In practical residential energy systems, reinforcement learning policies may generate actions that violate physical constraints, such as battery SOC limits, EV availability constraints, charging power limits, or grid exchange limits. Therefore, recent safe RL studies have moved beyond purely penalty-based constraint handling and have introduced safety layers, action projection, physical-informed correction, and constrained policy optimisation. Wang et al. developed a physical-informed safe reinforcement learning approach for secure energy management of multi-energy microgrids [11]. More general safe RL studies have also demonstrated that action projection and runtime safety filtering can explicitly map unsafe actions into feasible action spaces [12]. In power and energy systems, projection-based deep reinforcement learning has been applied to safe distribution grid control to improve hard-constraint satisfaction during online operation [13]. These studies support the idea that operational constraints should be embedded directly into the action execution process rather than being handled only by soft penalty terms in the reward function. Nevertheless, detailed safety projection mechanisms for residential systems with coupled BESS and EV control remain underdeveloped.

A third research direction is multi-agent reinforcement learning for distributed energy storage, EV scheduling, and smart community energy management. Compared with single-agent control, multi-agent reinforcement learning can better represent the distributed nature of residential energy resources and the coupling among different controllable devices. Kaewdornhan et al. proposed a real-time multi-home energy management method with EV charging scheduling using multi-agent deep reinforcement learning [14]. Wang et al. developed a scalable multi-agent deep reinforcement learning approach for residential hybrid energy systems, highlighting the value of centralised training and decentralised execution for scalable coordination [15]. Saberi et al. further investigated a multi-agent deep constrained Q-learning method for smart building energy management under uncertainties, demonstrating the relevance of combining multi-agent learning with constraint handling in building-level energy systems [16]. These studies show that MARL is suitable for coordinating distributed energy resources under uncertainty. However, existing MARL-based energy management methods often focus on community-level coordination, EV scheduling, or economic dispatch separately, while the combined use of digital twin updating, safety-constrained action projection, and MASAC-based BESS–EV coordination in a residential PV-integrated system has not been fully addressed.

Battery health condition estimation is also closely related to long-term energy management because battery ageing affects available capacity, charging/discharging efficiency, degradation cost, and the economic value of storage scheduling. Recent battery digital twin and prognostics studies have investigated advanced SOH/RUL estimation methods, including digital-twin-supported battery state estimation using TCN-LSTM networks and transfer learning [9], physics-guided TL-LSTM networks for early-stage lithium-ion battery degradation trajectory prediction [17], and knowledge-distillation-based network compression frameworks for lithium-ion battery lifecycle management [18]. These studies indicate that battery health estimation can provide useful information for degradation-aware digital twin updating. In the present work, the digital twin framework incorporates degradation-

sensitive parameters and degradation-related cost terms, while more advanced online SOH/RUL estimation models can be integrated in future work.

Based on the above review, the research gap can be summarised as follows. Existing studies have separately investigated digital twin-assisted forecasting and decision support, safe reinforcement learning for constrained energy management, multi-agent reinforcement learning for distributed energy scheduling, and battery health estimation for digital twin applications. However, there is still a lack of an integrated residential energy management framework that combines digital-twin-based state updating and prediction, explicit safety-constrained continuous action projection, and MASAC-based coordination between stationary battery storage and EV charging flexibility. This gap motivates the digital-twin-aided safe MASAC framework proposed in this paper.

Based on the above research gaps, the main contributions of this paper are summarised as follows:

- A digital-twin-aided safe MASAC framework is developed for coordinated residential energy management with PV generation, battery storage, and electric vehicles, integrating virtual system representation, state synchronisation, degradation-aware model updating, learning-based decision-making, safety projection, and reward-based policy learning within a closed-loop architecture.
- A MASAC-based control strategy is adopted to learn continuous charging and discharging decisions for residential energy resources under stochastic PV generation, load demand, electricity prices, and EV availability.
- A safety projection layer is designed to enforce physical constraints, including battery SOC limits, EV charging availability, and grid power constraints, thereby ensuring the system's safe operation.
- A digital twin-based virtual–physical interaction mechanism is introduced to support state synchronisation, degradation-sensitive parameter updating, short-term PV/load forecasting, and augmented observation construction for the MASAC controller.
- Extensive simulation experiments demonstrate that the proposed method improves operational efficiency and economic performance compared to TD3 and conventional rule-based strategies.

The remainder of this paper is organised as follows. Section 2 introduces the system model and problem formulation of the residential energy management system. Section 3 presents the proposed digital-twin-aided safe MASAC framework. Section 4 describes the experimental setup and evaluation methodology. Section 5 discusses the simulation results and performance comparisons. Finally, Section 6 concludes the paper and outlines future research directions.

2. System Model and Problem Formulation

2.1. Residential Energy System Architecture

The residential energy management system considered in this study comprises several key components: rooftop PV generation, a household electrical load, a BESS, an EV, and a bidirectional connection to the utility grid [19]. Together, these components form a typical integrated renewable residential energy system that enables flexible energy exchange and intelligent energy scheduling within the household [19].

The photovoltaic generation unit provides renewable electricity that can be used directly to supply the household load or charge local storage devices. However, due to the intermittent, weather-dependent nature of solar power generation, PV output is inherently uncertain and varies throughout the day [19]. As a result, additional flexible resources are required to balance the mismatch between generation and consumption.

The battery energy storage system is used to improve the flexibility of the household energy system. The battery can store surplus electricity when PV generation exceeds load demand and discharge energy when household demand exceeds available renewable generation. Through appropriate charging and discharge scheduling, the battery helps reduce electricity purchase from the grid, smooth power fluctuations, and improve overall system efficiency [20].

In addition to stationary battery storage, an electric vehicle is also considered a mobile energy storage device. When the EV is connected to the residential charging point, it can participate in the household energy management by charging from the grid or local PV generation [21]. Depending on the system configuration, the EV battery may also provide energy back to the household via vehicle-to-home (V2H) [21]. However, the availability of the electric vehicle is limited by its arrival and departure patterns. Specifically, the EV is only available for charging or discharging when it is parked at home. Furthermore, the EV battery typically needs to reach a target SOC before departure to meet the user's travel requirements [19].

The residential energy system is also connected to the external power grid through a bidirectional power exchange interface. When local generation and storage are insufficient to meet the household's electricity demand, electricity can be purchased from the grid. In contrast, if local photovoltaic generation exceeds household consumption and storage capacity, surplus energy can be exported back to the grid, depending on electricity pricing policy. In this work, the time-of-use electricity price is considered, in which the electricity purchase price varies across time periods throughout the day. This pricing mechanism provides economic incentives to shift energy consumption and optimise energy storage operations [20].

Within this energy management framework, the battery charging or discharging power and the EV charging or discharging power are treated as controllable decision variables. At each decision time step, the energy management controller determines the charging and discharging actions for the battery and the EV to coordinate energy flows among PV generation, household load demand, and grid power exchange. These decisions must satisfy various physical and operational constraints, including battery power limits, SOC limits, EV availability constraints, and grid power exchange limits [19].

The objective of the residential energy management system is to minimise overall operational costs while maintaining reliable, safe system operation. The operational cost typically includes the purchase cost of the electricity from the network, potential revenue from exporting electricity, the battery degradation costs associated with the charging and discharging cycles, and penalties for operational constraints such as electric vehicle charging requirements. Due to the stochastic nature of photovoltaic generation, household load demand and EV usage patterns, determining an optimal control strategy is a challenging sequential decision-making problem that requires intelligent and adaptive energy management approaches [22].

2.2. Component Models and System Dynamics

To formulate the residential energy management problem, this subsection introduces mathematical models for the key components of the system. These models describe the energy balance of the household system, as well as the operational dynamics of the battery energy storage system and the electric vehicle [20].

2.2.1. Power Balance Model

At each time step, the power balance of the residential energy system must be satisfied. To avoid ambiguity between charging and discharging decisions, the battery and EV

are both described using signed continuous power variables. For each storage device $i \in \{b, ev\}$, $P_i(t) > 0$ denotes charging power, while $P_i(t) < 0$ denotes discharging power. The corresponding non-negative charging and discharging components are defined as

$$P_i^{ch}(t) = [P_i(t)]^+, \quad P_i^{dis}(t) = [-P_i(t)]^+, \quad (1)$$

where $[x]^+ = \max(x, 0)$. Since $P_i^{ch}(t)$ and $P_i^{dis}(t)$ are derived from the same signed power command, they are not independent decision variables. Therefore, the following mutual exclusivity relationship is naturally satisfied:

$$P_i^{ch}(t)P_i^{dis}(t) = 0, \quad \forall i \in \{b, ev\}, \forall t \in \mathcal{T}. \quad (2)$$

This formulation mathematically prevents the battery or EV from charging and discharging simultaneously within the same time interval.

The electricity demand of the household can be met by local PV generation, the battery energy storage system, the EV, or the electricity grid. The power balance equation is expressed as

$$P_{grid}(t) = P_{load}(t) + P_b^{ch}(t) + P_{ev}^{ch}(t) - P_{pv}(t) - P_b^{dis}(t) - P_{ev}^{dis}(t), \quad (3)$$

where $P_{grid}(t)$ denotes the power exchanged with the utility grid at time step t . A positive value of $P_{grid}(t)$ indicates electricity import from the grid, while a negative value represents electricity export to the grid. $P_{load}(t)$ represents the household load demand and $P_{pv}(t)$ denotes the power generated by the photovoltaic system.

The power exchange with the grid may also be limited by infrastructure constraints, which can be expressed as

$$|P_{grid}(t)| \leq P_{grid}^{max}, \quad (4)$$

where P_{grid}^{max} denotes the maximum allowed exchange power of the grid.

2.2.2. Battery Energy Storage Model

The battery energy storage system is modelled based on its charging and discharging dynamics. Following the signed power convention introduced above, the battery charging and discharging powers are obtained from the signed battery command $P_b(t)$ as $P_b^{ch}(t) = [P_b(t)]^+$ and $P_b^{dis}(t) = [-P_b(t)]^+$. The SOC of the battery evolves according to

$$SOC_b(t+1) = SOC_b(t) + \frac{\eta_b^{ch}P_b^{ch}(t) - P_b^{dis}(t)/\eta_b^{dis}}{E_b}\Delta t, \quad (5)$$

where $SOC_b(t)$ denotes the state of charge of the battery at time step t , E_b represents the nominal battery capacity in kWh, and Δt is the duration of one control time step. η_b^{ch} and η_b^{dis} denote the charging and discharging efficiencies of the battery, respectively.

The battery charging and discharging powers are limited by the power rating of the storage system:

$$0 \leq P_b^{ch}(t) \leq P_b^{ch,max}, \quad (6)$$

$$0 \leq P_b^{dis}(t) \leq P_b^{dis,max}. \quad (7)$$

In addition, the battery SOC must remain within its operational limits:

$$SOC_b^{min} \leq SOC_b(t) \leq SOC_b^{max}. \quad (8)$$

These constraints ensure safe operation of the battery and prevent excessive charging or deep discharging. Since $P_b^{ch}(t)$ and $P_b^{dis}(t)$ are derived from the same signed action $P_b(t)$, simultaneous battery charging and discharging are excluded by construction.

2.2.3. Electric Vehicle Model

The electric vehicle is modelled as a mobile energy storage device that can participate in household energy management when it is connected to the residential charging point. Similar to the stationary battery, the EV power command is represented by a signed continuous variable $P_{ev}(t)$, where $P_{ev}(t) > 0$ indicates charging and $P_{ev}(t) < 0$ indicates discharging or vehicle-to-home operation. The corresponding non-negative charging and discharging powers are defined as $P_{ev}^{ch}(t) = [P_{ev}(t)]^+$ and $P_{ev}^{dis}(t) = [-P_{ev}(t)]^+$.

The SOC dynamics of the EV battery are described as

$$SOC_{ev}(t+1) = SOC_{ev}(t) + \frac{\eta_{ev}^{ch} P_{ev}^{ch}(t) - P_{ev}^{dis}(t) / \eta_{ev}^{dis}}{E_{ev}} \Delta t, \quad (9)$$

where $SOC_{ev}(t)$ denotes the state of charge of the EV battery, E_{ev} is the capacity of the EV battery, and η_{ev}^{ch} and η_{ev}^{dis} are the EV charging and discharging efficiencies, respectively.

Unlike stationary storage devices, electric vehicles are not always available for charging or discharging due to their mobility. To model this characteristic, a binary availability variable $\xi_{ev}(t)$ is introduced:

$$\xi_{ev}(t) = \begin{cases} 1, & \text{if the EV is connected to the household,} \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

When the EV is not present at home, its charging and discharging powers must be zero. This is enforced through the following availability-dependent power limits:

$$0 \leq P_{ev}^{ch}(t) \leq \xi_{ev}(t) P_{ev}^{ch,max}, \quad (11)$$

$$0 \leq P_{ev}^{dis}(t) \leq \xi_{ev}(t) P_{ev}^{dis,max}. \quad (12)$$

Therefore, when $\xi_{ev}(t) = 0$, both $P_{ev}^{ch}(t)$ and $P_{ev}^{dis}(t)$ are forced to zero. Furthermore, the SOC of the EV battery must meet its operational limits:

$$0 \leq SOC_{ev}(t) \leq 1. \quad (13)$$

The EV is typically required to reach a desired SOC level before its departure time to ensure the user's mobility needs are met. Since $P_{ev}^{ch}(t)$ and $P_{ev}^{dis}(t)$ are obtained from the same signed EV action $P_{ev}(t)$, simultaneous EV charging and discharging are also mathematically prevented.

2.3. Energy Management Problem Formulation

Based on the component models and system dynamics introduced above, the residential energy management task can be formulated as a constrained sequential optimisation problem. The objective is to determine the charging and discharging actions of the battery and electric vehicle on the scheduling horizon to minimise the total operational cost while satisfying all physical and operational constraints.

Let the scheduling horizon be denoted by $\mathcal{T} = \{0, 1, \dots, T-1\}$, where T is the total number of decision steps within one operating cycle. At each time step $t \in \mathcal{T}$, the controller determines the battery power $P_b(t)$ and the electric vehicle power $P_{ev}(t)$ according to

the state of the current system and the available information. The resulting optimisation objective consists of the following terms.

2.3.1. Electricity Transaction Cost

The net cost of the electricity transaction is determined by the power exchanged with the utility grid and the corresponding time-varying electricity prices. Let $\lambda_{buy}(t)$ and $\lambda_{sell}(t)$ denote the electricity purchase and sale prices, respectively. The cost of the transaction of the grid at the time step t is defined as

$$C_{grid}(t) = \max(P_{grid}(t), 0) \lambda_{buy}(t) \Delta t - \max(-P_{grid}(t), 0) \lambda_{sell}(t) \Delta t, \quad (14)$$

where a positive $P_{grid}(t)$ indicates the import of electricity from the grid and a negative $P_{grid}(t)$ indicates the export of electricity to the grid. The numerical values of the time-of-use purchase price $\lambda_{buy}(t)$ and the export price $\lambda_{sell}(t)$ used in the simulations are provided in Table 1.

Table 1. Economic and penalty parameters used in the objective function.

Parameter	Value	Description
λ_{buy}^{off}	0.20 £/kWh	Off-peak electricity purchase price
λ_{buy}^{mid}	0.30 £/kWh	Mid-peak electricity purchase price
λ_{buy}^{peak}	0.35 £/kWh	Peak electricity purchase price
λ_{sell}	0.10 £/kWh	Electricity export price
$\lambda_{b,deg}$	0.02 £/kWh	Stationary battery degradation coefficient
$\lambda_{ev,ch}$	0.01 £/kWh	EV charging degradation coefficient
$\lambda_{ev,dis}$	0.02 £/kWh	EV discharging degradation coefficient
λ_{ramp}	0.01	Battery ramping penalty coefficient
$\Delta_{b,term}$	15.0	Battery terminal SOC penalty coefficient
Δ_{ev}	100.0	EV departure SOC penalty coefficient
Δ_{grid}	10.0	Grid violation penalty coefficient
Δt	1.0 h	Duration of one decision time step

2.3.2. Battery and EV Degradation Costs and Smoothing Penalty

Frequent charging and discharging of electrochemical storage devices can accelerate battery degradation and reduce their service life. In the proposed residential energy management problem, degradation-related costs are considered for both the stationary battery and the EV battery. For the stationary battery, the degradation cost is modelled according to the battery energy throughput:

$$C_{b,deg}(t) = \lambda_{b,deg} (P_b^{ch}(t) + P_b^{dis}(t)) \Delta t, \quad (15)$$

where $\lambda_{b,deg}$ is the degradation cost coefficient of the stationary battery.

Since the EV battery may also be used for vehicle-to-home operation, an additional EV battery degradation cost is introduced to penalise charging and discharging cycles of the EV battery, especially the discharging energy used for non-mobility purposes:

$$C_{ev,deg}(t) = \lambda_{ev,ch} P_{ev}^{ch}(t) \Delta t + \lambda_{ev,dis} P_{ev}^{dis}(t) \Delta t, \quad (16)$$

where $\lambda_{ev,ch}$ and $\lambda_{ev,dis}$ denote the EV charging and discharging degradation cost coefficients, respectively. In practical applications, $\lambda_{ev,dis}$ can be set higher than $\lambda_{ev,ch}$ to reflect users' greater sensitivity to EV battery wear caused by V2H discharging cycles. If V2H operation is disabled, $P_{ev}^{dis}(t) = 0$ and the EV discharging degradation term becomes inactive. In addition, rapid fluctuations in battery power may be undesirable in practical

operation. Therefore, a ramping penalty is incorporated to encourage smoother battery power trajectories:

$$C_{ramp}(t) = \lambda_{ramp} |P_b(t) - P_b(t-1)|, \quad (17)$$

where λ_{ramp} is the ramping penalty coefficient.

2.3.3. Battery Terminal SOC Penalty

To avoid unrealistic end-of-horizon behaviour, the battery SOC at the end of the scheduling horizon is encouraged to return to a desired reference value. Let SOC_b^{ref} denote the reference terminal SOC. The terminal penalty is defined as

$$C_{b,term} = \Lambda_{b,term} \left(SOC_b(T) - SOC_b^{ref} \right)^2, \quad (18)$$

where $\Lambda_{b,term}$ is the corresponding weighting factor.

2.3.4. EV Departure Requirement Penalty

In residential applications, the electric vehicle must satisfy the user's mobility requirement by reaching a desired SOC before departure. Let t_{dep} denote the EV departure time and SOC_{ev}^{tar} denote the target EV SOC before departure. The EV departure penalty is formulated as

$$C_{ev,dep} = \Lambda_{ev} \max \left(0, SOC_{ev}^{tar} - SOC_{ev}(t_{dep}) \right) E_{ev}, \quad (19)$$

where Λ_{ev} is the penalty coefficient and E_{ev} is the battery capacity of the EV. This term penalises insufficient charging of EV before departure and ensures user travel readiness.

2.3.5. Overall Objective Function

By combining the above cost components, the total operational cost over the scheduling horizon can be expressed as

$$J = \sum_{t=0}^{T-1} \left[C_{grid}(t) + C_{b,deg}(t) + C_{ev,deg}(t) + C_{ramp}(t) \right] + C_{b,term} + C_{ev,dep}. \quad (20)$$

The inclusion of $C_{ev,deg}(t)$ explicitly accounts for the degradation cost associated with EV battery cycling, particularly when the EV participates in V2H discharging for household energy support.

Consequently, the residential energy management problem is formulated as follows:

$$\min_{\{P_b(t), P_{ev}(t)\}_{t=0}^{T-1}} J \quad (21)$$

subject to

$$P_b^{ch}(t) = [P_b(t)]^+, \quad P_b^{dis}(t) = [-P_b(t)]^+, \quad \forall t \in \mathcal{T}, \quad (22)$$

$$P_{ev}^{ch}(t) = [P_{ev}(t)]^+, \quad P_{ev}^{dis}(t) = [-P_{ev}(t)]^+, \quad \forall t \in \mathcal{T}, \quad (23)$$

$$P_b^{ch}(t)P_b^{dis}(t) = 0, \quad P_{ev}^{ch}(t)P_{ev}^{dis}(t) = 0, \quad \forall t \in \mathcal{T}, \quad (24)$$

$$P_{grid}(t) = P_{load}(t) + P_b^{ch}(t) + P_{ev}^{ch}(t) - P_{pv}(t) - P_b^{dis}(t) - P_{ev}^{dis}(t), \quad \forall t \in \mathcal{T}, \quad (25)$$

$$SOC_b(t+1) = SOC_b(t) + \frac{\eta_b^{ch} P_b^{ch}(t) - P_b^{dis}(t) / \eta_b^{dis}}{E_b} \Delta t, \quad \forall t \in \mathcal{T}, \quad (26)$$

$$SOC_{ev}(t+1) = SOC_{ev}(t) + \frac{\eta_{ev}^{ch} P_{ev}^{ch}(t) - P_{ev}^{dis}(t) / \eta_{ev}^{dis}}{E_{ev}} \Delta t, \quad \forall t \in \mathcal{T}, \quad (27)$$

$$0 \leq P_b^{ch}(t) \leq P_b^{ch,max}, \quad 0 \leq P_b^{dis}(t) \leq P_b^{dis,max}, \quad \forall t \in \mathcal{T}, \quad (28)$$

$$SOC_b^{min} \leq SOC_b(t) \leq SOC_b^{max}, \quad \forall t \in \{0, 1, \dots, T\}, \quad (29)$$

$$0 \leq P_{ev}^{ch}(t) \leq \xi_{ev}(t)P_{ev}^{ch,max}, \quad 0 \leq P_{ev}^{dis}(t) \leq \xi_{ev}(t)P_{ev}^{dis,max}, \quad \forall t \in \mathcal{T}, \quad (30)$$

$$0 \leq SOC_{ev}(t) \leq 1, \quad \forall t \in \{0, 1, \dots, T\}, \quad (31)$$

$$|P_{grid}(t)| \leq P_{grid}^{max}, \quad \forall t \in \mathcal{T}. \quad (32)$$

The mutual exclusivity constraints in the above formulation are satisfied by construction because the charging and discharging components are derived from a single signed action for each storage device. Therefore, no additional binary decision variable is required, which is consistent with the continuous action space of the MASAC controller.

The above problem is difficult to solve using conventional optimisation methods because the residential energy management environment is characterised by uncertainty in PV generation, load demand, electricity prices, and EV usage patterns. In addition, the problem involves coupled device dynamics, time-varying constraints, and sequential decision-making over the entire scheduling horizon. Therefore, instead of relying on an exact model-based optimisation solver, this work addresses the problem using a reinforcement learning framework that learns a control policy directly from interactions with the environment while implicitly optimising the above objective under operational constraints.

3. Methodology

3.1. Overview of the Proposed Energy Management Framework

To address the residential energy management problem formulated in Section 2.3, this paper proposes a digital-twin-aided safe multi-agent reinforcement learning framework for coordinated control of distributed energy resources in residential energy systems. The proposed framework integrates three main components: a reinforcement-learning-based control module, a digital twin module, and a safety-constrained action projection mechanism. The digital twin module is not treated as a standalone forecasting model. Instead, it is designed as a virtual representation of the residential energy system that supports state synchronisation, degradation-sensitive parameter updating, short-term prediction, and interaction with the MASAC controller. Based on the updated digital twin states and predictive information, the MASAC controller learns coordinated charging and discharging strategies for the battery and EV, while the safety projection layer ensures that the executed actions satisfy physical and operational constraints. These components jointly enable adaptive, predictive, and safe residential energy management under renewable generation and load uncertainty.

The overall architecture of the proposed framework is illustrated in Figure 1. As shown in Figure 1, the residential energy environment provides the current operating states, including the battery SOC, EV SOC, EV availability, time-of-use electricity price, PV generation, and household load demand. These state variables are first used to update the virtual states and degradation-sensitive parameters of the digital twin. The updated digital twin states are then combined with short-term PV and load forecasts to form augmented observations for the MASAC controller. The battery agent and EV agent then generate continuous control actions, which are evaluated during training by a centralised critic. Before implementation, the raw actions are passed through the safety projection layer to ensure compliance with battery SOC limits, EV availability constraints, charging power limits, and grid exchange limits. The resulting feasible actions are finally executed in the residential energy environment, while the reward/cost evaluation module provides feedback for policy learning. The framework is organised as a closed-loop energy management system that integrates the residential energy environment, the digital twin-based state updating and forecasting module, the MASAC-based decision-making module, the safety

projection layer, and the reward/cost evaluation mechanism. At each decision time step, the current state of the system is first observed from the environment. State information typically includes battery SOC, EV SOC, EV availability status, electricity prices, current photovoltaic generation, and household load demand. In addition to the instantaneous state of the system, the updated virtual states, calibrated degradation-sensitive parameters, and short-term predictive information generated by the digital twin are incorporated into the observation space. The forecast module provides a summary of future PV generation and load demand over a short look-ahead horizon, enabling the controller to anticipate future energy conditions.

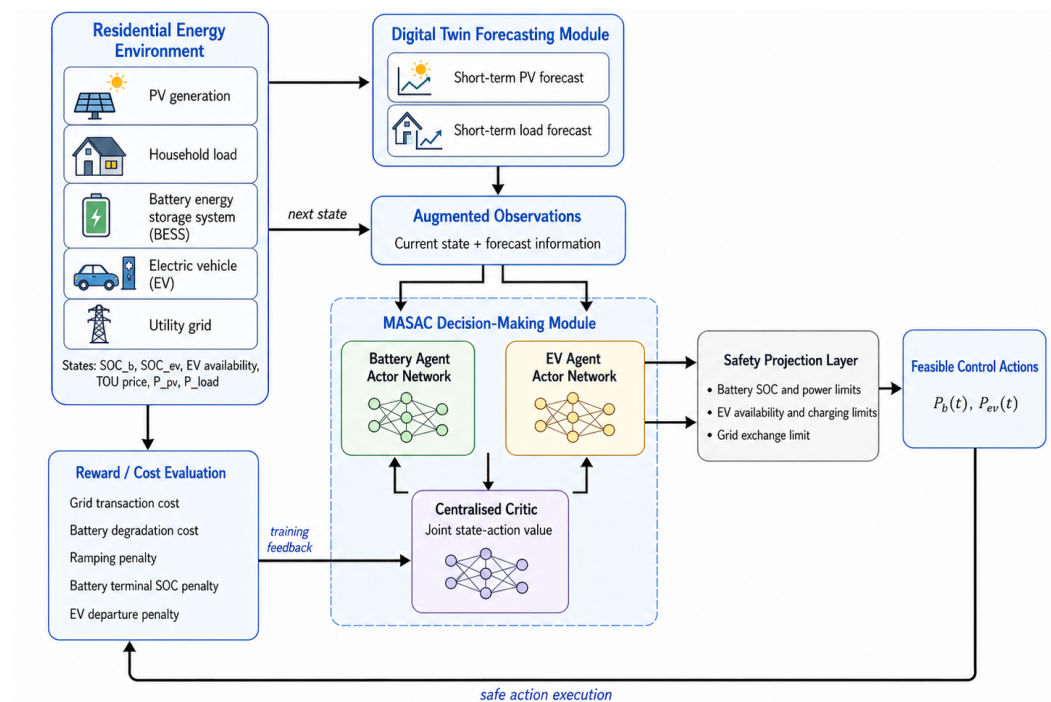


Figure 1. Architecture of the proposed digital-twin-aided safe MASAC framework for residential energy management.

Based on the observed system state and predictive information, the reinforcement learning controller determines control actions for distributed energy resources. In this work, the battery energy storage system and the electric vehicle are treated as two controllable agents that jointly participate in the energy management task. Each agent determines its charging or discharging power on the basis of local observations while considering the system's overall objective. The reinforcement learning controller, therefore, outputs continuous control actions representing the charging or discharging power commands for the battery and the EV.

Since reinforcement learning policies may occasionally generate control actions that violate physical or operational constraints, a safety projection mechanism is introduced to ensure the feasibility of the control decisions. The safety layer receives the raw actions produced by the reinforcement learning agents and projects them onto a feasible action space that satisfies all operational constraints. These constraints include battery charging and discharge power limits, battery SOC limits, EV availability constraints, EV charging and discharge limits, and grid power exchange limits. By enforcing these constraints before executing control actions, the proposed framework ensures that the energy management system operates safely at all times.

After the safety-constrained actions are applied to the residential energy environment, the system transitions to a new state according to the physical system dynamics described

in Section 2. The environment then provides a reward signal that reflects the system's operational costs, including electricity purchase costs, battery degradation costs, and penalty terms for constraint violations or unsatisfied EV charging requirements. This reward signal guides the learning process of reinforcement learning agents.

Through continuous interactions with the environment, the reinforcement learning controller gradually learns an energy management policy that minimises the long-term operational cost while respecting the constraints of the system. The use of predictive information from the digital twin module improves the controller's decision-making capability under uncertainty, while the safety projection layer ensures that all executed control actions remain physically feasible. As a result, the proposed framework enables reliable and economically efficient coordination of distributed energy resources in residential energy systems with renewable generation and electric vehicles.

3.2. Digital Twin Architecture and Synchronisation Mechanism

In the proposed framework, the digital twin is not treated merely as a forecasting model, but as a virtual representation of the residential energy system that is continuously updated using operational data from the physical or simulated household environment. The digital twin consists of five functional layers: the data acquisition layer, the virtual system model layer, the adaptive calibration layer, the forecasting layer, and the decision interaction layer.

The data acquisition layer collects time-indexed operating information from the residential energy system, including PV generation, household load demand, battery SOC, EV SOC, EV availability, electricity price, and grid power exchange. At each decision time step, these measurements are transferred from the physical system to the virtual model to update the digital representation of the household energy system. In the simulation-based implementation adopted in this study, this physical-to-virtual synchronisation is realised through the simulated environment, which provides the same categories of measurements that would be available from smart meters, battery management systems, EV chargers, and PV inverters in practical deployment.

The virtual system model layer maintains mathematical models of the PV generation unit, household load, battery energy storage system, EV battery, and grid interface. These models describe the system states, operational constraints, and energy balance relationships used by the MASAC controller. To improve consistency between the virtual model and the operating system, an adaptive calibration layer is included. This layer updates the virtual state and degradation-sensitive model parameters according to the mismatch between measured and model-predicted states.

At each decision time step, the digital twin receives a measurement vector from the residential energy environment:

$$z_t = \left[P_{pv}^{mea}(t), P_{load}^{mea}(t), SOC_b^{mea}(t), SOC_{ev}^{mea}(t), \zeta_{ev}^{mea}(t), P_{grid}^{mea}(t) \right]^T, \quad (33)$$

where the superscript "mea" denotes measured or simulated measurements. In the simulation-based implementation of this study, these measurements are provided by the residential energy environment. In practical deployment, the same information could be obtained from smart meters, PV inverters, battery management systems, EV chargers, and grid interface meters.

Before measurement correction, the digital twin predicts the virtual state using the previous virtual state, the previously executed feasible action, and the current parameter vector:

$$\hat{x}_{t|t-1} = f_{dt}(\hat{x}_{t-1}, \bar{a}_{t-1}, u_t; \theta_t), \quad (34)$$

where $\hat{x}_{t|t-1}$ is the predicted virtual state, $f_{dt}(\cdot)$ denotes the digital twin state transition model, $\bar{a}_{t-1}$ is the projected feasible action executed at the previous time step, u_t represents exogenous inputs such as PV generation, load demand, EV availability, and electricity price, and θ_t is the internal parameter vector of the digital twin.

The mismatch between the measured state and the predicted virtual state is calculated as

$$e_t = x_t^{mea} - h(\hat{x}_{t|t-1}), \quad (35)$$

where x_t^{mea} denotes the measurable state components contained in z_t , and $h(\cdot)$ maps the virtual state to the corresponding measurable outputs. The virtual state is then corrected by

$$\hat{x}_t = \hat{x}_{t|t-1} + K_x e_t, \quad (36)$$

where K_x is the state correction gain.

The internal digital twin parameter vector includes degradation-sensitive parameters, such as the available battery capacity, available EV battery capacity, battery charging/discharging efficiencies, and EV charging/discharging efficiencies. These parameters are updated according to

$$\theta_{t+1} = \text{Proj}_{\Theta}(\theta_t + K_{\theta} e_t), \quad (37)$$

where K_{θ} is the parameter calibration gain and $\text{Proj}_{\Theta}(\cdot)$ ensures that the updated parameters remain within physically reasonable bounds.

To represent degradation effects, the available capacities of the stationary battery and EV battery are linked to their state-of-health values:

$$E_b^{ava}(t) = \text{SOH}_b(t) E_b^{nom}, \quad (38)$$

$$E_{ev}^{ava}(t) = \text{SOH}_{ev}(t) E_{ev}^{nom}, \quad (39)$$

where E_b^{nom} and E_{ev}^{nom} are the nominal capacities of the stationary battery and EV battery, respectively. A simplified throughput-based degradation update can be written as

$$\text{SOH}_b(t+1) = \text{SOH}_b(t) - \rho_b \frac{(P_b^{ch}(t) + P_b^{dis}(t)) \Delta t}{E_b^{nom}}, \quad (40)$$

$$\text{SOH}_{ev}(t+1) = \text{SOH}_{ev}(t) - \rho_{ev} \frac{(P_{ev}^{ch}(t) + P_{ev}^{dis}(t)) \Delta t}{E_{ev}^{nom}}, \quad (41)$$

where ρ_b and ρ_{ev} are degradation coefficients. The updated available capacities and efficiency-related parameters are then used by the digital twin in subsequent state prediction and observation construction. This mechanism allows the digital twin to represent the influence of battery and EV ageing on the energy management process.

The decision interaction layer forms the interface between the digital twin and the MASAC controller. The digital twin sends the updated system state, calibrated model parameters, and short-term forecast information to the MASAC agents. The agents then generate charging or discharging actions for the battery and EV. After safety projection, the feasible actions are applied to the residential energy environment, and the resulting measurements are fed back to the digital twin for the next update. Therefore, a closed-loop virtual–physical interaction is established, where the digital twin supports state estimation, predictive information generation, parameter calibration, and scenario simulation for safe and adaptive residential energy management.

Although the battery and EV have different physical characteristics and availability patterns, their decisions are coupled through the household power balance, grid exchange

power, PV self-consumption, and the shared operational cost. For example, excessive EV charging during high-price periods may increase grid import unless the stationary battery discharges or sufficient PV generation is available. Similarly, aggressive battery discharging may reduce short-term grid import but may also limit the ability of the system to support future EV charging requirements. Therefore, the optimal decision for one device depends not only on its own local state, but also on the state and action of the other device.

3.3. Multi-Agent Soft Actor–Critic Energy Management Algorithm

The residential energy management problem defined in Section 2.3 is inherently a sequential decision-making problem under uncertainty. The stochastic nature of photovoltaic generation, household load demand, and EV arrival and departure patterns makes it difficult to derive an optimal control policy using conventional deterministic optimisation methods. To address these challenges, this work adopts a reinforcement learning approach to learn optimal energy management strategies through interaction with the environment.

In the proposed framework, the battery energy storage system and the electric vehicle are modelled as two independent agents that learn their control policies simultaneously. Each agent determines its own charging or discharging action based on its local observations, while a centralised critic is used during training to capture the interactions among the agents. This structure enables coordinated control of multiple distributed energy resources while maintaining decentralised decision-making.

3.3.1. Multi-Agent Reinforcement Learning Formulation

The energy management process can be formulated as a Markov decision process (MDP) defined by the tuple $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}$ denotes the state space of the system, $\mathcal{A}$ represents the action space, P is the probability of transition from state, r denotes the reward function, and $\gamma \in (0, 1)$ is the discount factor.

At each time step t , the system is in state $s_t \in \mathcal{S}$, which includes battery SOC, EV SOC, electricity prices, photovoltaic generation, load demand and other relevant information. The two agents select their actions a_t^b and a_t^{ev} , corresponding to the charging and discharging powers of the battery and the EV, respectively. The joint action can therefore be expressed as

$$a_t = (a_t^b, a_t^{ev}). \quad (42)$$

After executing the joint action, the environment transitions to the next state s_{t+1} according to the system dynamics described in Section 2.2, and a reward signal r_t is obtained.

The objective of the reinforcement learning controller is to learn a policy that maximises the expected cumulative discounted reward

$$J(\pi) = \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t r_t \right]. \quad (43)$$

3.3.2. Soft Actor–Critic Framework

To address the above learning problem, this work adopts the MASAC algorithm, an off-policy actor–critic method designed for continuous control tasks. MASAC introduces an entropy regularisation term into the reinforcement learning objective to encourage exploration and improve policy robustness. The objective function of the maximum entropy reinforcement learning framework is given by

$$J(\pi) = \mathbb{E}_{(s_t, a_t) \sim \rho_{\pi}} \left[\sum_{t=0}^{\infty} \gamma^t (r_t + \alpha \mathcal{H}(\pi(\cdot | s_t))) \right], \quad (44)$$

where $\mathcal{H}(\pi(\cdot|s_t))$ denotes the entropy of the policy distribution and α is the temperature parameter that determines the relative importance of the entropy term.

In the multi-agent setting considered in this work, the joint policy can be expressed as

$$\pi(a_t|s_t) = \pi_b(a_t^b|o_t^b)\pi_{ev}(a_t^{ev}|o_t^{ev}), \quad (45)$$

where π_b and π_{ev} represent the stochastic policies of the battery and EV agents, respectively, and o_t^b and o_t^{ev} denote the local observations of the corresponding agents.

3.3.3. Centralised Critic

To capture the coupling among multiple energy devices, a centralised critic is adopted during training. The critic estimates the joint action-value function based on the feasible actions actually executed by the environment:

$$Q(s_t, \bar{a}_t^b, \bar{a}_t^{ev}), \quad (46)$$

where $\bar{a}_t^b$ and $\bar{a}_t^{ev}$ denote the projected feasible actions of the battery and EV agents, respectively. The raw actions generated by the actor networks are denoted as $\tilde{a}_t^b$ and $\tilde{a}_t^{ev}$, while the safety projection layer maps them into feasible actions:

$$\bar{a}_t = \mathcal{P}(\tilde{a}_t). \quad (47)$$

Therefore, the critic evaluates the action-value of the actions that are physically implemented in the residential energy environment, rather than the uncorrected raw actions.

In practice, two Q-networks are used to mitigate the overestimation bias commonly observed in the value function approximation. Let Q_{θ_1} and Q_{θ_2} denote the two critic networks parameterized by θ_1 and θ_2 . The target Q-value is computed as

$$y_t = r_t + \gamma \left(\min_{i=1,2} Q_{\theta'_i}(s_{t+1}, a_{t+1}) - \alpha \log \pi(a_{t+1}|s_{t+1}) \right), \quad (48)$$

where θ'_i represents the parameters of the target networks and a_{t+1} is sampled from the current policy.

The critic networks are optimised by minimising the mean-squared Bellman error

$$\mathcal{L}_Q = \mathbb{E}_{(s_t, \bar{a}_t, r_t, s_{t+1})} \left[(Q_{\theta_i}(s_t, \bar{a}_t) - y_t)^2 \right]. \quad (49)$$

3.3.4. Policy Optimisation

The actor networks generate stochastic policies for the battery and EV agents. The policy is updated by maximising the expected Q-value while considering the entropy regularisation term. The policy optimisation objective can be expressed as

$$\mathcal{L}_\pi = \mathbb{E}_{s_t \sim \mathcal{D}} [\alpha \log \pi(a_t|s_t) - Q(s_t, a_t)], \quad (50)$$

where $\mathcal{D}$ denotes the replay buffer that contains previously collected transitions. In the proposed safe MASAC framework, the replay buffer stores the projected feasible action rather than the original raw action. Specifically, after the actor networks generate raw actions $\tilde{a}_t$, the safety projection layer maps them into feasible actions $\bar{a}_t = \mathcal{P}(\tilde{a}_t)$. The projected action $\bar{a}_t$ is then executed in the residential energy environment, producing the reward r_t and the next state s_{t+1} . Therefore, each transition stored in the replay buffer is defined as

$$(s_t, o_t, \bar{a}_t, r_t, s_{t+1}, o_{t+1}), \quad (51)$$

where o_t and o_{t+1} denote the agent observations before and after the transition.

This storage strategy is adopted because the reward and the next state are generated by the projected feasible action, not by the raw action. If the raw unsafe action were stored while the environment actually executed the projected action, the replay buffer would contain an inconsistent state–action–reward–next-state relationship. Such inconsistency would make the critic approximate the value of actions that were not physically executed, which could destabilise MASAC training. By storing the projected feasible action, the critic update remains consistent with the actual environment transition.

Accordingly, the critic loss is computed using the projected action:

$$\mathcal{L}_Q = \mathbb{E}_{(s_t, \tilde{a}_t, r_t, s_{t+1}) \sim \mathcal{D}} \left[(Q_{\theta_t}(s_t, \tilde{a}_t) - y_t)^2 \right]. \quad (52)$$

During policy optimisation, the raw action sampled from the current policy is also passed through the same safety projection before being evaluated by the critic:

$$\tilde{a}_t^\pi = \mathcal{P}(\tilde{a}_t^\pi), \quad \tilde{a}_t^\pi \sim \pi(\cdot | s_t). \quad (53)$$

The actor objective is therefore written as

$$\mathcal{L}_\pi = \mathbb{E}_{s_t \sim \mathcal{D}} [\alpha \log \pi(\tilde{a}_t^\pi | s_t) - Q(s_t, \tilde{a}_t^\pi)]. \quad (54)$$

This design improves training stability by ensuring that both critic learning and actor optimisation are based on the feasible action space induced by the safety projection layer.

3.4. Safety-Constrained Action Projection

Although reinforcement learning provides a flexible framework for solving complex sequential decision-making problems, the actions generated by the learned policies may violate physical or operational constraints if they are directly applied to the residential energy system. To prevent unsafe operation, a safety-constrained action projection layer is introduced between the MASAC agents and the residential energy environment. This layer converts the raw actions generated by the actor networks into feasible actions before environment execution.

Let $\tilde{a}_t = (\tilde{P}_b(t), \tilde{P}_{ev}(t))$ denote the raw joint action generated by the battery and EV agents at time step t . Consistent with the signed power convention defined in Section 2, a positive power command represents charging, while a negative power command represents discharging. The safety projection operator $\mathcal{P}(\cdot)$ maps the raw joint action into a feasible action:

$$a_t = \mathcal{P}(\tilde{a}_t), \quad (55)$$

where $a_t = (P_b(t), P_{ev}(t))$ denotes the projected feasible action executed by the residential energy environment.

3.4.1. Battery Power and SOC Projection

The battery projection first clips the raw signed battery command into the allowable signed power range:

$$P_b^0(t) = \text{clip}(\tilde{P}_b(t), -P_b^{dis,max}, P_b^{ch,max}), \quad (56)$$

where $P_b^{ch,max}$ and $P_b^{dis,max}$ denote the maximum charging and discharging powers of the battery, respectively. Since positive power represents charging and negative power represents discharging, the corresponding non-negative charging and discharging components are obtained as

$$P_b^{ch,0}(t) = [P_b^0(t)]^+, \quad P_b^{dis,0}(t) = [-P_b^0(t)]^+, \quad (57)$$

where $[x]^+ = \max(x, 0)$. Since both components are derived from the same signed command, they satisfy

$$P_b^{ch,0}(t)P_b^{dis,0}(t) = 0. \quad (58)$$

Therefore, simultaneous battery charging and discharging are eliminated by construction.

The battery SOC at the next time step is calculated as

$$SOC_b(t+1) = SOC_b(t) + \frac{\eta_b^{ch} P_b^{ch}(t) - P_b^{dis}(t)/\eta_b^{dis}}{E_b} \Delta t. \quad (59)$$

To ensure that $SOC_b(t+1)$ remains within the admissible range $[SOC_b^{min}, SOC_b^{max}]$, the maximum feasible charging and discharging powers imposed by the current SOC are derived as

$$P_b^{ch,SOC}(t) = \frac{(SOC_b^{max} - SOC_b(t))E_b}{\eta_b^{ch}\Delta t}, \quad (60)$$

$$P_b^{dis,SOC}(t) = \frac{(SOC_b(t) - SOC_b^{min})E_b\eta_b^{dis}}{\Delta t}. \quad (61)$$

The SOC-feasible charging and discharging powers are then obtained by

$$P_b^{ch}(t) = \min(P_b^{ch,0}(t), P_b^{ch,max}, P_b^{ch,SOC}(t)), \quad (62)$$

$$P_b^{dis}(t) = \min(P_b^{dis,0}(t), P_b^{dis,max}, P_b^{dis,SOC}(t)). \quad (63)$$

Finally, the projected signed battery action is reconstructed as

$$P_b(t) = P_b^{ch}(t) - P_b^{dis}(t). \quad (64)$$

This projected battery action satisfies the battery power limits, SOC limits, and charging/discharging mutual exclusivity constraint.

3.4.2. EV Availability and SOC Projection

The EV projection follows the same signed power convention as the battery. A positive EV power command represents charging, while a negative EV power command represents discharging or vehicle-to-home operation. The raw signed EV action is first clipped into the allowable signed power range:

$$P_{ev}^0(t) = \text{clip}(\tilde{P}_{ev}(t), -P_{ev}^{dis,max}, P_{ev}^{ch,max}). \quad (65)$$

The corresponding non-negative charging and discharging components are obtained as

$$P_{ev}^{ch,0}(t) = [P_{ev}^0(t)]^+, \quad P_{ev}^{dis,0}(t) = [-P_{ev}^0(t)]^+. \quad (66)$$

Therefore, the EV charging and discharging components also satisfy

$$P_{ev}^{ch,0}(t)P_{ev}^{dis,0}(t) = 0. \quad (67)$$

The EV is available for residential energy management only when it is connected to the household charging point. Its availability is represented by the binary variable $\xi_{ev}(t)$, where $\xi_{ev}(t) = 1$ indicates that the EV is connected and $\xi_{ev}(t) = 0$ indicates that it is absent. If the EV is not connected, both charging and discharging actions are forced to zero through the availability multiplier.

The EV SOC dynamics are expressed as

$$SOC_{ev}(t+1) = SOC_{ev}(t) + \frac{\eta_{ev}^{ch} P_{ev}^{ch}(t) - P_{ev}^{dis}(t) / \eta_{ev}^{dis}}{E_{ev}} \Delta t. \quad (68)$$

To ensure that the EV SOC remains within the allowable range, the SOC-feasible EV charging and discharging powers are derived as

$$P_{ev}^{ch,SOC}(t) = \frac{(SOC_{ev}^{max} - SOC_{ev}(t)) E_{ev}}{\eta_{ev}^{ch} \Delta t}, \quad (69)$$

$$P_{ev}^{dis,SOC}(t) = \frac{(SOC_{ev}(t) - SOC_{ev}^{min}) E_{ev} \eta_{ev}^{dis}}{\Delta t}. \quad (70)$$

The final projected EV charging and discharging powers are calculated as

$$P_{ev}^{ch}(t) = \xi_{ev}(t) \min(P_{ev}^{ch,0}(t), P_{ev}^{ch,max}, P_{ev}^{ch,SOC}(t)), \quad (71)$$

$$P_{ev}^{dis}(t) = \xi_{ev}(t) \min(P_{ev}^{dis,0}(t), P_{ev}^{dis,max}, P_{ev}^{dis,SOC}(t)). \quad (72)$$

Finally, the projected signed EV action is reconstructed as

$$P_{ev}(t) = P_{ev}^{ch}(t) - P_{ev}^{dis}(t). \quad (73)$$

This projection ensures that the EV can only participate in energy management when it is available, that its SOC remains within the allowable range, and that simultaneous EV charging and discharging are prevented. Therefore, the projection layer is compatible with V2H-enabled residential energy management.

3.4.3. Grid Power Projection

After the battery and EV projections, the resulting grid power is calculated from the household power balance:

$$P_{grid}(t) = P_{load}(t) + P_b^{ch}(t) + P_{ev}^{ch}(t) - P_{pv}(t) - P_b^{dis}(t) - P_{ev}^{dis}(t). \quad (74)$$

The grid exchange must satisfy

$$|P_{grid}(t)| \leq P_{grid}^{max}. \quad (75)$$

If the projected battery and EV actions lead to a grid power violation, a secondary correction is applied. Let $P_{grid}^0(t)$ denote the grid power before grid correction. If $P_{grid}^0(t) > P_{grid}^{max}$, the household is importing too much electricity from the grid. The excess import is calculated as

$$\Delta P_{imp}(t) = P_{grid}^0(t) - P_{grid}^{max}. \quad (76)$$

In this case, the projection first reduces flexible charging demand, starting from EV charging and then battery charging:

$$P_{ev}^{ch}(t) \leftarrow \max(0, P_{ev}^{ch}(t) - \Delta P_{imp}(t)), \quad (77)$$

$$P_b^{ch}(t) \leftarrow \max(0, P_b^{ch}(t) - \Delta P_{imp}^{rem}(t)), \quad (78)$$

where $\Delta P_{imp}^{rem}(t)$ denotes the remaining excess import after EV charging reduction.

If $P_{grid}^0(t) < -P_{grid}^{max}$, the household is exporting too much electricity to the grid. The excess export is calculated as

$$\Delta P_{exp}(t) = -P_{grid}^{max} - P_{grid}^0(t). \quad (79)$$

In this case, the projection reduces discharging power to decrease the exported power:

$$P_{ev}^{dis}(t) \leftarrow \max\left(0, P_{ev}^{dis}(t) - \Delta P_{exp}(t)\right), \quad (80)$$

$$P_b^{dis}(t) \leftarrow \max\left(0, P_b^{dis}(t) - \Delta P_{exp}^{rem}(t)\right), \quad (81)$$

where $\Delta P_{exp}^{rem}(t)$ denotes the remaining excess export after EV discharging reduction.

The signed actions are then reconstructed as

$$P_b(t) = P_b^{ch}(t) - P_b^{dis}(t), \quad (82)$$

$$P_{ev}(t) = P_{ev}^{ch}(t) - P_{ev}^{dis}(t), \quad (83)$$

and the final grid power is recalculated using the power balance equation.

3.4.4. Projection Procedure

The complete safety projection procedure is summarised as follows. First, the raw battery and EV actions generated by the MASAC actors are clipped to their device-level signed power ranges. Second, the signed actions are decomposed into charging and discharging components, which guarantees mutual exclusivity. Third, SOC-based feasible charging and discharging limits are computed for the battery and EV. Fourth, the EV availability constraint is enforced so that the EV cannot charge or discharge when it is not connected. Fifth, the grid power exchange is checked, and flexible charging or discharging actions are corrected if the grid import/export limit is violated. Finally, the corrected feasible actions are executed in the residential energy environment.

This explicit projection procedure ensures that the reinforcement learning policy does not rely solely on soft reward penalties to learn feasibility. Instead, infeasible actions are corrected before execution, which improves training stability and deployment safety. The reward function still encourages economically efficient behaviour, but physical constraints are enforced directly by the projection layer.

It should be noted that the proposed framework does not rely on the cost function to avoid simultaneous charging and discharging. Since each storage device is controlled by a single signed power command and the charging/discharging components are obtained from its positive and negative parts, the mutual exclusivity condition $P_i^{ch}(t)P_i^{dis}(t) = 0$, $i \in \{b, ev\}$, is guaranteed by construction at every decision time step. Therefore, no additional binary exclusivity variable is required.

4. Simulation Setup and Evaluation

4.1. Residential Energy System Configuration

To evaluate the performance of the proposed energy management framework, a residential energy system consisting of photovoltaic generation, household load demand, a battery energy storage system, and an electric vehicle is considered. The system is connected to the power grid and operates at an electricity rate that varies by time of use. The simulation environment is designed to capture the key characteristics of residential energy systems with distributed renewable energy and flexible energy storage devices.

Figure 2 illustrates the simulation environment used for the residential energy management case study. The environment consists of rooftop PV generation, household load demand, a stationary battery energy storage system, an EV, and a bidirectional grid connection. At each decision time step, PV generation, load demand, electricity price, battery

SOC, EV SOC, and EV availability are provided to the energy management controller. The controller then determines the battery and EV charging/discharging actions, while the grid power exchange is calculated according to the power balance relationship.

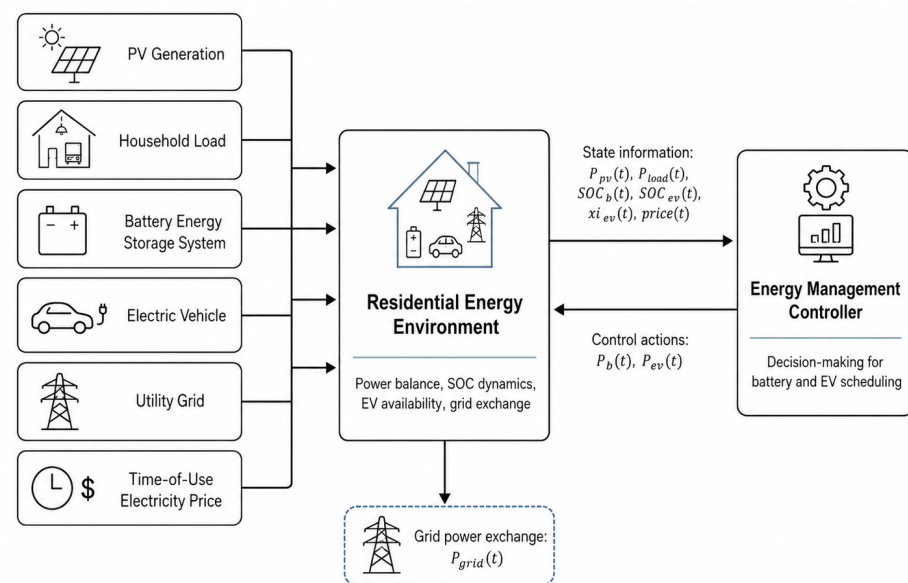


Figure 2. Simulation environment configuration for the residential energy management case study.

The photovoltaic generation system is a rooftop solar installation that is typically found in residential buildings. The PV generation profile follows a normalised daily solar generation pattern that reflects variations in solar irradiance throughout the day. The generated photovoltaic power is scaled according to the installed capacity of the photovoltaic system. To capture the stochastic nature of renewable generation, small random fluctuations are introduced into the PV profile during simulation.

The domestic electricity demand is modelled using a typical residential load profile with time-varying consumption patterns over a 24 h period. The load profile reflects common daily activities, with increased electricity consumption in the morning and evening hours. To account for variability in household energy consumption, random noise is added to the base load profile, yielding realistic variations in electricity demand across simulation episodes.

The residential energy storage system is modelled as a lithium-ion battery with specified energy capacity, charging and discharging power limits, and charging and discharging efficiencies. The battery state of charge is constrained within a predefined safe operating range to prevent excessive charging or deep discharging. The battery can store surplus photovoltaic energy for later use or discharge to support household loads when electricity prices are high.

In addition to the stationary battery storage system, an electric vehicle is incorporated into the residential energy system as a mobile energy storage device. The EV battery has its own energy capacity and charging power limits. Unlike a stationary battery, the EV is available for charging or discharging only when it is parked at home and connected to the residential charging infrastructure. The arrival and departure times of the EV are modelled to reflect typical daily commuting behaviour. When the EV returns home, its battery SOC is initialised according to a probabilistic distribution that represents the remaining energy after daily travel. Before leaving the home, the electric vehicle battery must reach a predefined target SOC to meet the user's mobility requirements.

The residential energy system interacts with the power grid through the purchase and sale of electricity. A time-of-use electricity tariff is adopted to represent realistic residential

electricity pricing schemes. Electricity purchase prices vary throughout the day according to off-peak, mid-peak and peak periods, while electricity export to the grid is compensated at a fixed selling price. These price signals provide economic incentives for energy storage devices to shift energy consumption and generation over time.

The entire energy management process is simulated on a daily operational horizon. The day is discretised into multiple decision time steps with a fixed duration Δt . At each time step, the energy management controller determines the charging or discharging actions of the battery and the electric vehicle based on the current state of the system and predicted future information. The environment then evolves according to the system dynamics described in Section 2.2, and the resulting operational cost is recorded for performance evaluation.

In this simulation environment, the proposed reinforcement learning-based energy management method can be evaluated under realistic residential operating conditions, including stochastic renewable generation, uncertain load demand, dynamic electricity pricing, and EV mobility behaviour.

To improve the reproducibility of the case study, the main system, mobility, and operational constraint parameters used in the residential energy management simulation are summarised in Table 2. These parameters define the PV-BESS-EV residential system, the EV mobility pattern, and the feasible operating region used by the proposed safe MASAC controller.

Table 2. Unified simulation parameters of the residential energy management case study.

Category	Parameter	Value
Time setting	Decision interval, Δt	0.5 h
Time setting	Daily scheduling horizon	24 steps
Time setting	Testing horizon	35 days
Training setting	Minimum episode length	48 steps
Training setting	Maximum episode length	120 steps
BESS	Battery capacity, E_b	13.5 kWh
BESS	Maximum charging power, $P_b^{ch,max}$	7.0 kW
BESS	Maximum discharging power, $P_b^{dis,max}$	7.0 kW
BESS	Charging/discharging efficiency, η_b^{ch}/η_b^{dis}	0.95/0.95
BESS	Physical SOC range, SOC_b^{min}/SOC_b^{max}	0.10/0.90
BESS	Operational SOC reserve, $SOC_b^{op,min}$	0.11
BESS	Initial SOC, $SOC_b(0)$	0.60
BESS	Terminal SOC target, SOC_b^{term}	0.30
EV	EV battery capacity, E_{ev}	40.0 kWh
EV	Maximum charging power, $P_{ev}^{ch,max}$	7.0 kW
EV	Maximum V2H discharging power, $P_{ev}^{dis,max}$	7.0 kW
EV	Charging/discharging efficiency, $\eta_{ev}^{ch}/\eta_{ev}^{dis}$	0.95/0.95
EV	SOC operating range, $SOC_{ev}^{min}/SOC_{ev}^{max}$	0.20/0.90
EV	Initial SOC, $SOC_{ev}(0)$	0.50
EV	Target SOC before departure, SOC_{ev}^{tar}	0.90
EV mobility	Daily trip energy	12.0 kWh
Grid	Maximum import power, $P_{grid}^{imp,max}$	10.0 kW
Grid	Maximum export power, $P_{grid}^{exp,max}$	10.0 kW
Electricity tariff	Off-peak purchase price, λ_{buy}^{off}	0.20 £/kWh
Electricity tariff	Mid-peak purchase price, λ_{buy}^{mid}	0.30 £/kWh
Electricity tariff	Peak purchase price, λ_{buy}^{peak}	0.35 £/kWh
Electricity tariff	Electricity export price, λ_{sell}	0.10 £/kWh

In addition to the system-level simulation parameters, the economic and penalty coefficients used in the objective function are summarised separately in Table 1. This

separate table is provided to clarify the cost calculation and avoid overloading the unified simulation parameter table. These parameters correspond to the electricity transaction cost, battery degradation cost, EV battery degradation cost, ramping penalty, battery terminal SOC penalty, and EV departure requirement penalty defined in Section 2.3. The time-of-use purchase price is divided into off-peak, mid-peak, and peak periods, while the export price is assumed to be fixed during the simulation horizon.

4.2. Training Configuration and Implementation Details

The proposed energy management framework is trained using the multi-agent soft actor–critic (MASAC) algorithm described in Section 3. The training process is conducted in the simulated residential energy environment introduced in Section 4.1, where each episode corresponds to one full-day energy management cycle.

The overall training procedure of the proposed MASAC controller is shown in Figure 3. At the beginning of each episode, the residential environment is initialised with the daily PV profile, load profile, EV availability pattern, battery SOC, and EV SOC. The battery and EV actor networks generate raw continuous actions according to their observations. These actions are processed by the safety projection layer before being executed in the environment. The environment then returns the reward, next state, and updated observations. The transition is stored in the replay buffer, and mini-batches are sampled to update the centralised critic and decentralised actor networks.

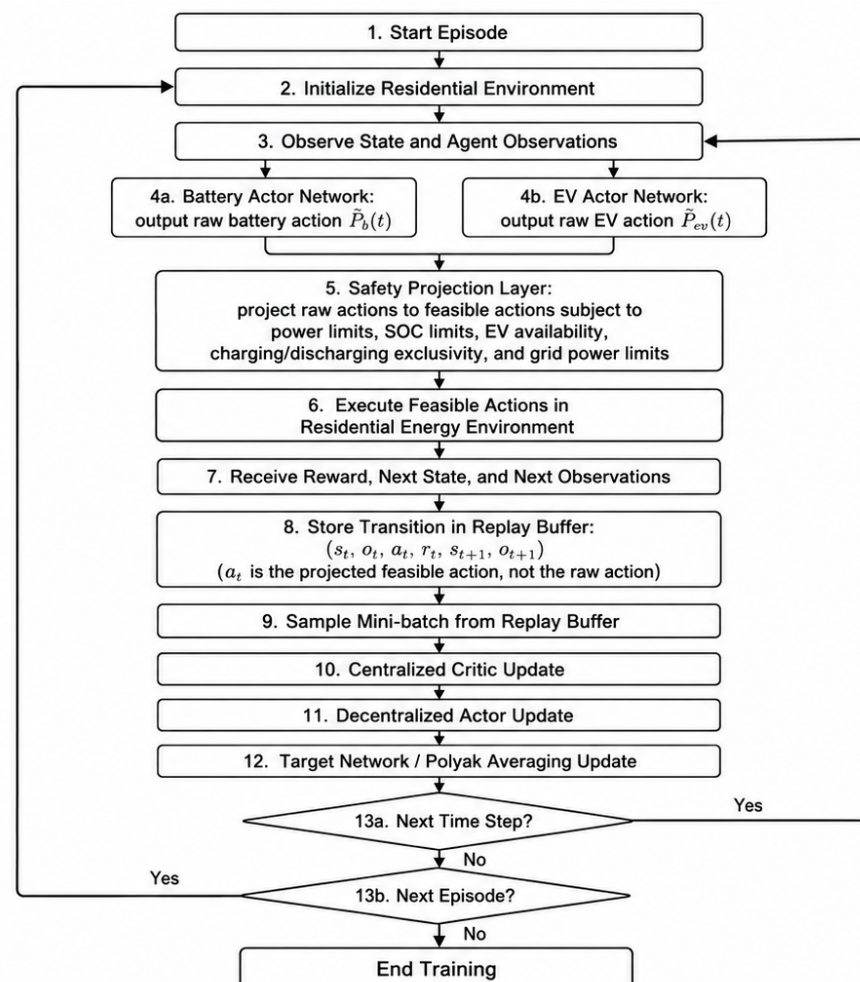


Figure 3. Training procedure of the proposed safe MASAC controller.

The reinforcement learning controller consists of two independent actor networks that represent the battery agent and the EV agent, respectively. Each actor network outputs continuous charging or discharging power commands for the corresponding device. A centralised critic network is used during training to evaluate the agents' joint actions based on the global system state. This centralised training strategy enables the learning process to capture the coupling relationships among different distributed energy resources.

Both the actor and critic networks are implemented using fully connected multi-layer perceptrons. The actor networks output the parameters of Gaussian action distributions, which are subsequently transformed into feasible charging and discharging power commands through a bounded activation function. The critic network estimates the joint action-value function over the system state and the joint actions of the battery and EV agents. To improve training stability and mitigate overestimation bias, two critic networks are used following the twin-Q learning mechanism.

Experience replay is used to improve the sample efficiency of the learning process. During training, the actor networks first generate raw continuous actions for the battery and EV agents. These raw actions are then processed by the safety projection layer to obtain feasible actions that satisfy power, SOC, EV availability, charging/discharging exclusivity, and grid exchange constraints. The feasible projected actions are executed in the residential energy environment and are stored in the replay buffer together with the corresponding states, observations, rewards, and next states. Therefore, each stored transition contains the actually executed projected action rather than the original raw action. Mini-batches of transitions are randomly sampled from the replay buffer to update the actor and critic networks, thereby breaking temporal correlations between consecutive samples while maintaining consistency between stored actions and actual environment transitions.

To further stabilise training, target critic networks are maintained and updated using Polyak averaging. The policy networks are optimised using stochastic gradient descent with the Adam optimiser. The entropy temperature parameter in the MASAC algorithm is automatically adjusted during training to balance exploration and exploitation. Through repeated interactions with the environment, the agents gradually learn coordinated control policies that minimise the long-term operational cost of the residential energy system.

The main hyperparameters used in the reinforcement learning training process are summarised in Table 3. These parameters were selected based on commonly used configurations in deep reinforcement learning applications for energy management and were empirically validated to ensure stable training performance.

Table 3. Key training parameters of the proposed MASAC framework.

Parameter	Value
Discount factor (γ)	0.99
Soft update coefficient (τ)	0.005
Replay buffer capacity	10^6
Batch size	256
Actor learning rate	3×10^{-4}
Critic learning rate	3×10^{-4}
Actor network hidden layers	256, 256, 256
Critic network hidden layers	256, 256, 256
Initial entropy temperature	0.2
Target entropy	adaptive
Training environment steps	10^5
Warm-up steps	5×10^3

During training, reinforcement learning agents first perform a warm-up phase in which actions are randomly sampled to populate the replay buffer. After sufficient experiences have been collected, the MASAC algorithm begins updating the actor and critic networks using sampled mini-batches. The training process continues until the specified number of environment interaction steps is reached. The trained policies are then evaluated using the testing procedures described in the following subsection.

4.3. Benchmark Strategies and Evaluation Metrics

To assess the effectiveness of the proposed multi-agent reinforcement learning framework, several benchmark control strategies are implemented for comparison. These baseline methods represent commonly used heuristic approaches for residential energy management and provide reference performance levels under identical operating conditions.

4.3.1. Benchmark Energy Management Strategies

Two rule-based energy management strategies are adopted as benchmark methods.

The first baseline is a *price-greedy strategy*, which adjusts charging and discharging actions according to the current electricity price. When the electricity price is lower than a predefined reference level, the battery is charged to store energy for future use. When electricity prices are relatively high, the battery discharges to meet household demand and reduce grid electricity purchases. Similarly, the electric vehicle is charged when electricity prices are low, provided that the vehicle is available at home. This strategy is a typical price-driven scheduling rule widely used in residential energy management.

The second baseline is a *fixed-night charging strategy*. In this strategy, the battery and electric vehicle are charged primarily at night, when electricity prices are generally lower. During the daytime hours, the battery may discharge to support the household load. This type of strategy reflects simple time-based scheduling rules commonly used in residential charging systems and EV home charging devices.

These rule-based strategies are designed without learning or prediction capabilities and therefore serve as reference methods for evaluating the advantages of the proposed reinforcement learning-based approach.

4.3.2. Evaluation Metrics

The performance of different energy management strategies is evaluated using several quantitative metrics that reflect both economic performance and operational reliability.

The primary evaluation metric is the total daily operational cost. To avoid repeating the objective function, the total cost used for evaluation is calculated according to Equation (20). It includes the electricity transaction cost, stationary battery degradation cost, EV battery degradation cost, battery ramping penalty, battery terminal SOC penalty, and EV departure requirement penalty. A lower total operational cost indicates a more economically efficient energy management strategy.

In addition to economic performance, the reliability of EV charging is evaluated using the *EV charging success rate*. This metric measures the proportion of simulation episodes in which the EV battery reaches the required target SOC before departure. A higher EV charging success rate indicates greater satisfaction with the user's mobility requirements.

To evaluate battery safety, the *battery SOC violation count* is also recorded. This metric counts the number of time steps during which the battery's SOC exceeds its allowable operating range. A well-designed energy management strategy should avoid such violations and maintain the battery SOC within its safe operating limits.

Furthermore, the *equivalent full cycle (EFC)* metric is used to estimate the battery's cycling intensity. The EFC value is calculated from the cumulative variation in the battery's SOC over the simulation horizon, providing an approximate measure of battery usage and degradation. Excessive cycling can accelerate battery wear and reduce its useful life.

By combining these evaluation metrics, the performance of different energy management strategies can be comprehensively assessed in terms of economic efficiency, operational reliability, and battery usage characteristics. The proposed reinforcement learning-based method is compared with benchmark strategies in the same simulation scenarios to demonstrate its effectiveness in coordinating distributed energy resources within residential energy systems.

5. Case Study and Results Discussion

5.1. Case Study Description

The proposed digital-twin-aided safe MASAC framework is evaluated using a residential energy management case study over a 35-day testing horizon. The testing horizon consists of five complete weekly cycles, with each week containing 168 hourly decision steps. The considered residential system includes PV generation, household load demand, a stationary battery energy storage system, an EV, and bidirectional grid power exchange. The objective is to evaluate whether the proposed controller can reduce the total operational cost while maintaining EV charging reliability and satisfying system-level operational constraints.

The proposed safe MASAC controller is compared with two benchmark strategies: a TD3-based reinforcement learning controller and a rule-based baseline. The TD3 controller represents a deterministic deep reinforcement learning benchmark, while the rule-based strategy represents a simple heuristic control method with limited adaptability. All methods are evaluated under the same PV generation, load demand, electricity price, EV availability, and operational constraint settings. Therefore, the comparison focuses on the economic performance, constraint satisfaction, and operational behaviour of the different energy management strategies.

5.2. Overall Cost Comparison

Table 4 summarises the main economic results over the 35-day testing horizon. The proposed safe MASAC controller achieves the lowest total operating cost of 365.8245, corresponding to an average daily cost of 10.4521. In comparison, the TD3 controller obtains a total cost of 391.4694, while the rule-based baseline obtains a total cost of 374.6889. Therefore, the proposed method reduces the total cost by 6.55% compared with TD3 and by 2.37% compared with the rule-based baseline.

Table 4. Overall cost comparison over the 35-day testing horizon.

Method	Total Cost	Daily Cost	Reward	Penalty
MASAC	365.8245	10.4521	−365.8245	0.0000
TD3	391.4694	11.1848	−391.4694	0.0000
Rule-based	374.6889	10.7054	−374.6889	0.0000

The cost comparison indicates that the proposed safe MASAC controller provides the best economic performance among the three methods. Since the total penalty is zero for all methods, the difference in reward is mainly caused by economic operation rather than constraint violation. This means that the proposed controller achieves cost reduction while maintaining feasible system operation. The improvement over TD3 suggests that the stochastic policy and entropy-regularised learning mechanism of SAC/MASAC can

provide more robust exploration and better energy scheduling behaviour in this residential control problem. The improvement over the rule-based strategy further shows that learning-based control can exploit storage flexibility more effectively than fixed heuristic rules.

Figure 4 illustrates the total cost comparison among the three methods. Although the rule-based controller performs better than TD3 in this case, it is still less economical than the proposed safe MASAC controller. This result shows that the rule-based baseline can provide stable operation, but its lack of predictive and adaptive decision-making limits its ability to fully utilise local PV generation and storage flexibility.

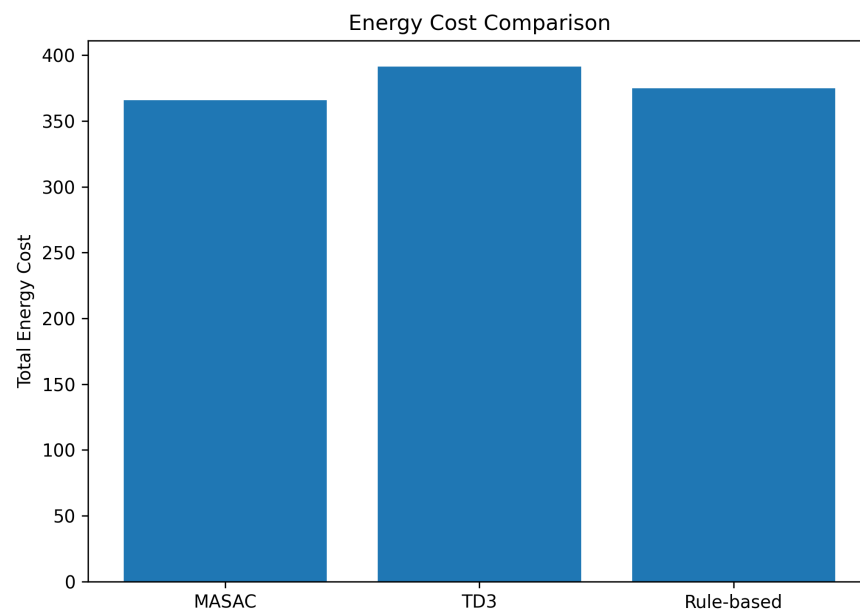


Figure 4. Total energy cost comparison over the 35-day testing horizon.

5.3. Constraint Satisfaction and Charging Reliability

Table 5 presents the constraint satisfaction results. All three methods achieve 100% EV departure success and 100% battery terminal SOC success over 35 departure events. The grid violation rate is 0% for all methods, indicating that all controllers satisfy the grid power constraint during testing.

Table 5. Constraint satisfaction and EV charging reliability.

Method	EV Success	Batt. Terminal Success	Grid Violation	Events
MASAC	100%	100%	0%	35
TD3	100%	100%	0%	35
Rule-based	100%	100%	0%	35

These results are important because they show that the cost reduction achieved by the proposed method is not obtained by relaxing operational constraints or sacrificing EV readiness. Instead, the proposed safe MASAC controller reduces the operating cost under the same feasibility conditions as the benchmark methods. The safety projection layer ensures that the executed actions remain within the admissible operating region, while the reward design encourages the EV to satisfy its departure SOC requirement.

5.4. Battery and EV Operating Behaviour

Table 6 summarises the battery and EV SOC statistics during the testing period. The proposed safe MASAC controller has a final battery SOC of 0.5785 and a mean battery SOC

of 0.5516. In contrast, TD3 maintains a much higher final battery SOC of 0.9000 and a mean battery SOC of 0.6943. This indicates that TD3 behaves more conservatively by keeping the battery at a high SOC level, which limits its ability to use the battery for economic energy shifting. The rule-based method also maintains a relatively high mean battery SOC of 0.6980, but its minimum battery SOC is 0.3661, suggesting that it does not fully exploit the lower feasible SOC region.

Table 6. Battery and EV SOC statistics during the testing horizon.

Method	Final Batt. SOC	Mean Batt. SOC	Min Batt. SOC	Final EV SOC	Mean EV SOC	Min EV SOC
MASAC	0.5785	0.5516	0.1100	0.9000	0.7512	0.6000
TD3	0.9000	0.6943	0.1100	0.9000	0.7541	0.6000
Rule-based	0.6739	0.6980	0.3661	0.6713	0.6971	0.5928

Figure 5 shows the battery SOC trajectories during the first seven days. The proposed safe MASAC controller uses the battery more actively than the rule-based method and avoids keeping the battery at a persistently high SOC level. This more flexible operation allows the battery to absorb PV surplus and discharge during economically favourable periods. TD3 also uses the battery, but it frequently drives the battery to the upper SOC boundary, indicating a stronger tendency toward over-conservative storage retention. In contrast, the rule-based strategy follows a more regular pattern and does not exploit the full operational flexibility of the battery.

The EV results also show that the proposed method maintains charging reliability while allowing flexible operation. The final EV SOC of the proposed method is 0.9000, and the minimum EV SOC remains at 0.6000. The rule-based method has a lower final EV SOC of 0.6713, but it still achieves 100% departure success because the EV reaches the required SOC at each departure event. Therefore, the EV-related results indicate that all methods satisfy the mobility requirement, while the proposed method achieves this with better overall economic performance.

Figure 6 compares the EV charging power profiles during the first seven days. The proposed safe MASAC controller charges the EV mainly when it is available and when charging is economically beneficial. Compared with the rule-based strategy, which follows a more predefined charging pattern, the proposed method exhibits more adaptive charging behaviour. This enables the EV to satisfy departure requirements without unnecessary charging during less favourable periods.

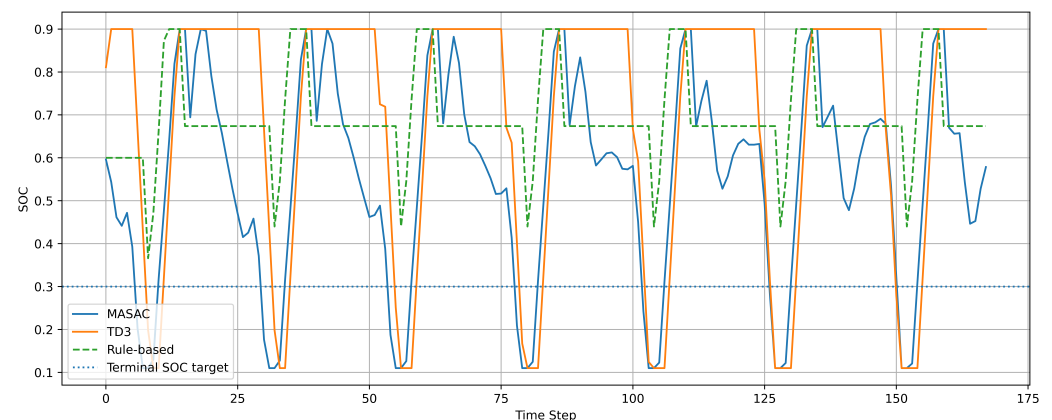


Figure 5. Battery SOC comparison during the first seven days.

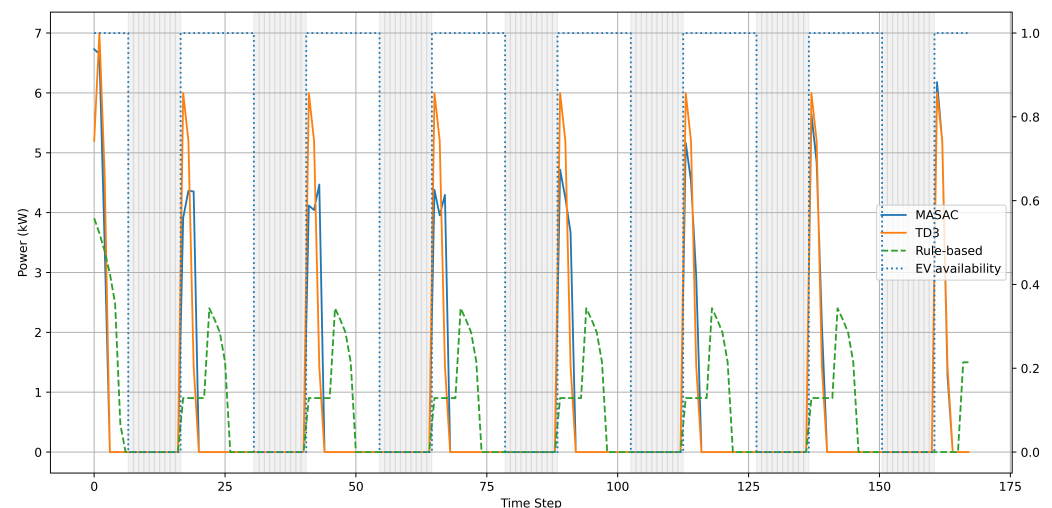


Figure 6. EV charging power comparison during the first seven days.

5.5. Grid Energy Exchange and Renewable Energy Utilisation

Table 7 compares the grid import, grid export, import cost, export revenue, and battery throughput of the three methods. The proposed safe MASAC controller achieves the lowest grid import, with 1608.48 kWh over the testing horizon. This is lower than TD3 and the rule-based method, which import 1764.28 kWh and 1863.32 kWh, respectively. The proposed method also has the lowest grid export, with 194.10 kWh, compared with 356.26 kWh for TD3 and 484.14 kWh for the rule-based method.

Table 7. Grid energy exchange and battery utilisation over the testing horizon.

Method	Grid Import (kWh)	Grid Export (kWh)	Gross Import Cost (£)	Export Revenue (£)	Battery Throughput (kWh)
Proposed safe MASAC	1608.48	194.10	385.23	19.41	960.51
TD3	1764.28	356.26	427.10	35.63	751.80
Rule-based	1863.32	484.14	423.10	48.41	436.36

These results show that the main advantage of the proposed method is not simply exporting more energy to the grid. Instead, the proposed controller imports less electricity, exports less surplus PV energy, and uses the battery more effectively to increase local energy utilisation. This is further supported by the self-consumption and self-sufficiency indicators. The proposed safe MASAC controller achieves a self-consumption ratio of 86.03% and a self-sufficiency ratio of 41.61%, both of which are higher than TD3 and the rule-based baseline. TD3 achieves a self-consumption ratio of 74.36% and a self-sufficiency ratio of 35.96%, while the rule-based method achieves 65.16% and 32.13%, respectively.

The battery throughput results also confirm the more active storage utilisation of the proposed method. The proposed safe MASAC controller has a battery throughput of 960.51 kWh, corresponding to 27.44 kWh per day. In comparison, TD3 and the rule-based method have daily battery throughputs of 21.48 kWh and 12.47 kWh, respectively. This indicates that the proposed controller uses the battery more actively to absorb local PV generation and reduce grid imports. Although higher battery throughput may increase cycling-related degradation, the degradation cost is already included in the objective function. Therefore, the obtained cost reduction reflects a trade-off between battery utilisation and economic benefit.

Figure 7 presents the grid power exchange profiles during the first seven days. Positive values represent grid import, while negative values represent grid export. The proposed

safe MASAC controller shows lower overall dependence on grid import and reduces unnecessary export compared with the benchmark methods. The rule-based strategy has a lower peak import, but it exports more energy to the grid and does not achieve the lowest total cost. This suggests that reducing peak import alone is not sufficient to minimise the overall energy cost when the objective function mainly considers energy transaction cost rather than demand charges or peak penalties.

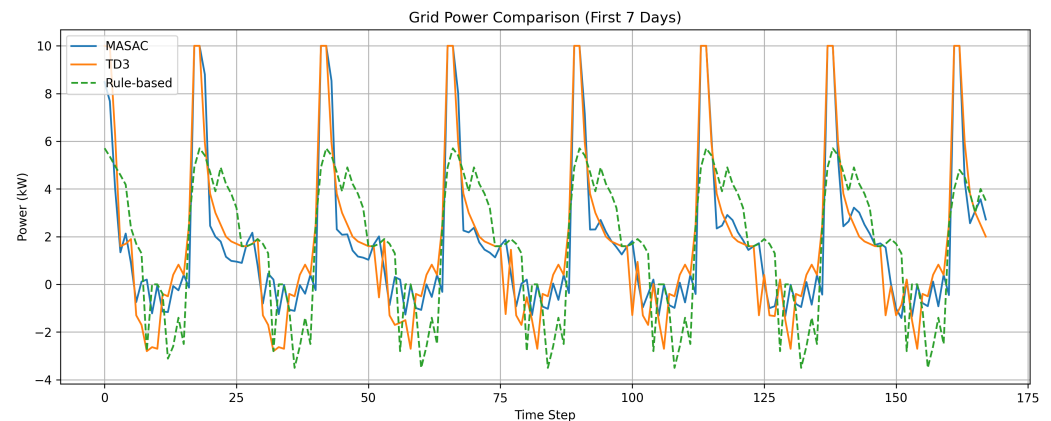


Figure 7. Grid power exchange comparison during the first seven days.

5.6. Peak Power Behaviour

Table 8 reports the peak import and peak export values. The proposed safe MASAC and TD3 controllers both reach the grid import limit of 10.00 kW during some periods, while the rule-based method has a lower peak import of 5.71 kW. However, the rule-based method also has the largest peak export magnitude of -3.50 kW, compared with -1.41 kW for the proposed method and -2.80 kW for TD3.

Table 8. Peak grid import and export comparison.

Method	Peak Import (kW)	Peak Export (kW)
Proposed safe MASAC	10.00	-1.41
TD3	10.00	-2.80
Rule-based	5.71	-3.50

This result provides an important interpretation of the optimisation objective. The proposed method is designed to minimise total operational cost, including energy transaction cost, degradation cost, ramping cost, and penalty terms. It is not explicitly designed for peak shaving. Therefore, the proposed method may occasionally use the full grid import capacity if this is economically beneficial and does not violate the grid constraint. If peak demand reduction is required as an additional objective, a demand charge term or peak power penalty could be incorporated into the reward function in future work.

5.7. Weekly Cost Stability

The weekly cost results show that all three methods experience a slightly higher cost in the first week and then converge to a stable weekly pattern. The weekly costs of the proposed safe MASAC controller are 75.7946, 72.5075, 72.5075, 72.5075, and 72.5075. The TD3 controller obtains 81.6707 in the first week and 77.4497 in the following weeks, while the rule-based method obtains 76.1755 in the first week and 74.6284 in the following weeks.

This pattern is mainly caused by the initial SOC condition at the start of the testing horizon. After the first weekly cycle, the system enters a repeated steady operating pattern. In the stable weekly cycles, the proposed safe MASAC controller consistently

achieves the lowest cost, confirming that its economic advantage is not limited to the initial transient period.

5.8. Ablation Study on Multi-Agent Coordination

To verify the benefit of the proposed multi-agent structure, an ablation study is conducted by comparing the proposed MASAC controller with a single-agent SAC controller. The single-agent SAC baseline uses one centralised actor to generate the joint battery and EV actions, while the proposed MASAC framework uses two decentralised actors and a centralised critic. Both methods are trained and tested under the same residential energy environment, reward function, safety projection layer, neural network size, replay buffer setting, learning rates, and testing horizon.

Table 9 reports the comparison between the proposed MASAC controller and the single-agent SAC baseline over the 35-day testing horizon. The proposed MASAC controller achieves a total operating cost of 365.8245, while the single-agent SAC controller obtains a total operating cost of 372.0309. Therefore, MASAC reduces the total operating cost by 1.67% compared with the single-agent SAC baseline. In addition, the MASAC controller achieves 100% EV departure success, 100% battery terminal success, and 0% grid violation. In contrast, the single-agent SAC controller achieves 100% EV departure success and 0% grid violation, but its battery terminal success rate is only 60%, with a penalty of 14.2500. This indicates that the proposed MASAC method provides better coordination between economic operation and terminal battery SOC recovery.

Table 9. Ablation comparison between MASAC and single-agent SAC over the 35-day testing horizon.

Method	Total Cost	Daily Cost	Penalty	EV Success	Batt. Terminal Success
MASAC	365.8245	10.4521	0.0000	100%	100%
Single-agent SAC	372.0309	10.6323	14.2500	100%	60%

The ablation results demonstrate the coordination advantage of the proposed MASAC framework. In the single-agent SAC baseline, the battery and EV actions are generated by one centralised policy. Although this policy can output joint actions, it does not explicitly separate the decision responsibilities of the stationary battery and EV. As a result, the single-agent SAC controller tends to operate the battery more aggressively and fails to restore the battery SOC to the required terminal range in some testing episodes. This is reflected by its lower final battery SOC of 0.3100 and lower mean battery SOC of 0.3591, compared with 0.5785 and 0.5516 under MASAC.

In contrast, the proposed MASAC controller allows the battery and EV to learn dedicated device-level policies while the centralised critic evaluates their joint effect on the system-level objective. This centralised-training and decentralised-execution structure improves the coordination between battery energy shifting and EV charging requirements. Compared with single-agent SAC, MASAC reduces grid import from 1683.7874 kWh to 1608.4795 kWh and reduces grid export from 296.4633 kWh to 194.0974 kWh. These results indicate that MASAC makes more effective use of local PV generation and battery flexibility, while maintaining better terminal SOC feasibility.

Overall, the single-agent SAC ablation confirms that the performance improvement of the proposed method is not only due to the use of SAC itself, but also due to the multi-agent coordination structure. The MASAC framework achieves lower cost, better battery terminal SOC satisfaction, and lower grid dependence under the same environment and safety projection settings.

5.9. Sensitivity Analysis Under Forecast Errors

To evaluate the robustness of the proposed digital-twin-aided MASAC controller under degraded forecast quality, a forecast error sensitivity analysis was conducted. The trained MASAC policy was tested under different levels of forecast noise without additional retraining. Multiplicative Gaussian noise was added to the short-term PV and load forecasts generated by the digital twin, while the actual PV generation, load demand, EV availability, system dynamics, reward calculation, and safety projection layer remained unchanged. Therefore, this experiment evaluates the influence of digital twin prediction errors rather than disturbances in the physical environment.

For each forecast noise level, five random seeds were tested over the same 35-day testing horizon. Table 10 summarises the mean and standard deviation of the main performance indicators. The original testing condition without additional forecast noise corresponds to the 0% case, where the proposed MASAC controller achieves a total operating cost of 365.8245. When the forecast noise level increases to 5%, 10%, 20%, and 30%, the total operating cost increases by 0.10%, 0.58%, 2.86%, and 5.17%, respectively. This indicates that forecast accuracy influences the economic performance of the controller, because less accurate predictive information reduces the ability of the policy to schedule battery and EV operation in advance.

Table 10. Sensitivity analysis under different forecast noise levels.

Forecast Noise	Total Cost	Cost Increase	Feasibility Indicators
0%	365.8245 $\pm$ 0.0000	0.00%	100%/100%/0%
5%	366.1923 $\pm$ 0.2936	0.10%	100%/100%/0%
10%	367.9626 $\pm$ 0.8024	0.58%	100%/100%/0%
20%	376.2705 $\pm$ 1.0971	2.86%	100%/100%/0%
30%	384.7290 $\pm$ 0.5962	5.17%	100%/100%/0%

Note: The feasibility indicators are reported as EV departure success/battery terminal SOC success/grid violation.

Although higher forecast errors increase the operating cost, the operational feasibility of the proposed controller remains stable. Across all tested noise levels, the EV departure success rate remains 100%, the battery terminal SOC success rate remains 100%, and the grid violation rate remains 0%. The average penalty also remains close to zero under all noise levels, confirming that the safety-related constraints are not violated by forecast degradation. This result shows that forecast information mainly supports anticipatory economic scheduling, while the physical feasibility of the executed actions is protected by the safety projection layer.

The detailed energy exchange results further explain the cost increase under higher forecast errors. As the forecast noise level increases from 0% to 30%, the grid import rises from 1608.4795 kWh to 1730.0062 kWh, and the grid export rises from 194.0974 kWh to 312.3247 kWh. This indicates that inaccurate forecasts reduce the controller's ability to align battery operation with future PV surplus and load demand, leading to less effective local energy utilisation. Nevertheless, the proposed MASAC controller maintains reliable EV charging and feasible grid operation because all actions are corrected by the safety projection layer before execution.

The sensitivity analysis therefore confirms that the proposed framework is more sensitive to forecast errors in terms of economic performance than in terms of safety. Even when the digital twin predictions become less accurate, infeasible battery, EV, and grid actions are corrected before being applied to the residential energy environment. As a result, the proposed MASAC controller maintains EV charging reliability and operational

feasibility under forecast uncertainty. Further robustness improvement can be achieved by training the policy with forecast-noise augmentation or by incorporating uncertainty-aware forecasting models in the digital twin.

5.10. Discussion

The case study results demonstrate that the proposed digital-twin-aided safe MASAC framework achieves the best overall economic performance while maintaining full operational feasibility. Over the 35-day testing horizon, the proposed method reduces the total operating cost from 391.4694 to 365.8245 compared with TD3, corresponding to a 6.55% reduction. Compared with the rule-based baseline, the cost is reduced from 374.6889 to 365.8245, corresponding to a 2.37% reduction. At the same time, the proposed method maintains 100% EV departure success, 100% battery terminal SOC success, and 0% grid violation rate.

The superior performance of the proposed method can be explained by its more effective use of local renewable energy and battery flexibility. Compared with TD3 and the rule-based baseline, the proposed controller achieves the lowest grid import and the lowest grid export. This indicates that it is better at consuming local PV generation within the household rather than exporting surplus energy to the grid. The higher self-consumption ratio and self-sufficiency ratio further confirm this advantage.

The comparison with TD3 shows that satisfying constraints alone is not sufficient to achieve the best economic performance. Although TD3 also achieves 100% constraint satisfaction, it tends to maintain a high battery SOC and exports more energy to the grid. This conservative storage behaviour reduces its ability to shift energy across time and leads to a higher total cost. In contrast, the proposed safe MASAC controller uses the battery more actively and economically.

The rule-based baseline provides stable and feasible operation, but its fixed control logic limits its flexibility. It has a lower peak import than the learning-based methods, but it imports more total electricity and exports more PV energy. This shows that the rule-based strategy is conservative and easy to implement, but it cannot fully exploit the economic value of battery storage, EV flexibility, and PV self-consumption.

Overall, the results confirm that the proposed safe MASAC framework can achieve a better balance between cost minimisation, EV charging reliability, battery utilisation, and operational safety. The method is particularly effective because it combines learning-based decision-making, safety-constrained action execution, and adaptive use of distributed energy resources in a unified residential energy management framework.

6. Conclusions

This paper proposed a digital-twin-aided safe MASAC framework for renewable-integrated residential energy management. The proposed method coordinates the operation of PV generation, household load, battery storage, EV charging, and grid power exchange under time-varying electricity prices and uncertain operating conditions. The digital twin module was used to maintain an updated virtual representation of the residential energy system by synchronising operating states, updating degradation-sensitive parameters, and generating short-term PV/load forecasts for augmented agent observations. Meanwhile, a safety projection layer was introduced to ensure that the executed control actions satisfy key physical and operational constraints, including battery SOC limits, EV availability requirements, and grid power exchange limits.

A 35-day residential case study was conducted to compare the proposed MASAC controller with TD3 and a rule-based baseline. The results show that the proposed method achieved the lowest total operating cost while maintaining full operational feasibility. Specifically, MASAC reduced the total cost by 6.55% compared with TD3 and by 2.37% compared with the rule-based baseline. All three methods achieved 100% EV departure success, 100% battery terminal success, and 0% grid violation rate, indicating that the cost improvement of MASAC was obtained under the same feasibility conditions rather than through constraint relaxation.

The detailed operational analysis further showed that the proposed MASAC controller used the battery more actively and improved local renewable energy utilisation. Compared with the benchmark methods, MASAC achieved lower grid import, lower grid export, a higher PV self-consumption ratio, and a higher self-sufficiency ratio. These results indicate that the proposed framework can better exploit the flexibility of residential storage and EV charging to reduce grid dependence and improve economic performance.

Future work will extend the proposed framework in several directions. First, more realistic forecasting models, real-world residential datasets, and advanced battery health estimation models will be incorporated to evaluate the robustness of the controller under practical uncertainty and long-term degradation. In particular, physics-guided degradation trajectory prediction and lightweight SOH/RUL estimation models can be integrated into the digital twin to support more accurate degradation-aware energy management.

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

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors appreciate the support provided by the Research Project PINV01-272 of the National Council of Science and Technology (CONACYT), A7C200 IRCF Project from the University of Nottingham, the Programa de Redução de Assimetrias na Pós-Graduação (PRAPG)—Edital n° 14/2023-DRI-CAPES. ID Number: 046.821.818-15 and FONDECYT Iniciación Project no. 11261540 and 24EVDI-262305 CORFO Project.

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

Abbreviations

The following abbreviations are used in this manuscript:

BESS	Battery Energy Storage System
DER	Distributed Energy Resource
DRL	Deep Reinforcement Learning
EV	Electric Vehicle
HEMS	Home Energy Management System
MASAC	Multi-Agent Soft Actor–Critic
PV	Photovoltaic
RES	Renewable Energy Source
RL	Reinforcement Learning

SAC	Soft Actor-Critic
SOC	State of Charge
TOU	Time-of-Use

References

1. Aldaliee, A.A.; Mansor, N.N.; Mokhlis, H.; Ramasamy, A.K.; Awalin, L.J. Pareto-Based Optimization of PV and Battery in Home-PV-BES-EV System with Integrated Dynamic Energy Management Strategy. *Sustainability* **2025**, *17*, 7364.
2. Coccato, S.; Barhmi, K.; Lampropoulos, I.; Golroodbari, S.; van Sark, W. A Review of Battery Energy Storage Optimization in the Built Environment. *Batteries* **2025**, *11*, 179.
3. Deng, X.; Li, J.; Bao, H.; Zhao, Z.; Su, X.; Huang, Y. Efficient Energy Management for Smart Homes with Electric Vehicles Using Scenario-Based Model Predictive Control. *Sustainability* **2025**, *17*, 7678.
4. Mahmoud, M.; Ben Slama, S. Deep Learning-Based Home Energy Management Incorporating Vehicle-to-Home and Home-to-Vehicle Technologies for Renewable Integration. *Energies* **2025**, *18*, 129.
5. Mansour, S.H.; Azzam, S.M.; Hasanien, H.M.; Tostado-Véliz, M.; Alkuhayli, A.; Jurado, F. Deep Reinforcement Learning-Based Plug-in Electric Vehicle Charging/Discharging Scheduling in a Home Energy Management System. *Energy* **2025**, *316*, 134420.
6. Haarnoja, T.; Zhou, A.; Abbeel, P.; Levine, S. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *Proceedings of the 35th International Conference on Machine Learning*; Dy, J.; Krause, A., Eds.; Proceedings of Machine Learning Research; PMLR: Stockholm, Sweden, 2018; Volume 80, pp. 1861–1870.
7. Zhou, J.; Ahmad, T.; Zhang, H.; Yan, J.; Ahmad, A. Digital Twin Application for Reinforcement Learning Based Optimal Scheduling and Reliability Management Enhancement of Systems. *Sol. Energy* **2023**, *255*, 302–316.
8. Sado, K.; Peskar, J.; Downey, A.; Khan, J.; Booth, K. A digital twin based forecasting framework for power flow management in DC microgrids. *Sci. Rep.* **2025**, *15*, 6430.
9. Zhao, K.; Liu, Y.; Zhou, Y.; Ming, W.; Wu, J. Digital Twin-Supported Battery State Estimation Based on TCN-LSTM Neural Networks and Transfer Learning. *CSEE J. Power Energy Syst.* **2025**, *11*, 567–579.
10. Liu, M.; Wang, Y.; Li, J.; Yang, K.; Guo, S.; Zhao, L.; Wang, Z. AI-Assisted Digital Twin Modeling Technology for Lithium-Ion Batteries. *Energy* **2025**, *331*, 137016.
11. Wang, Y.; He, H.; Sun, C. Secure energy management of multi-energy microgrid: A physical-informed safe reinforcement learning approach. *Appl. Energy* **2023**, *334*, 120634.
12. Kochdumper, N.; Krasowski, H.; Wang, X.; Bak, S.; Althoff, M. Provably Safe Reinforcement Learning via Action Projection Using Reachability Analysis and Polynomial Zonotopes. *IEEE Open J. Control Syst.* **2023**, *2*, 79–92.
13. Zhang, M.; Guo, G.; Zhao, T.; Xu, Q. DNN Assisted Projection Based Deep Reinforcement Learning for Safe Control of Distribution Grids. *IEEE Trans. Power Syst.* **2024**, *39*, 5687–5698.
14. Kaewdornhan, N.; Intarakumnerd, P.; Chatthaworn, R.; Chaitusaney, S. Real-Time Multi-Home Energy Management with EV Charging Scheduling Using Multi-Agent Deep Reinforcement Learning Optimization. *Energies* **2023**, *16*, 2357.
15. Wang, Z.; Xiao, F.; Ran, Y.; Li, Y.; Xu, Y. Scalable Energy Management Approach of Residential Hybrid Energy System Using Multi-Agent Deep Reinforcement Learning. *Appl. Energy* **2024**, *367*, 123414.
16. Saberi, H.; Zhang, C.; Dong, Z.Y. A Multi-Agent Deep Constrained Q-Learning Method for Smart Building Energy Management Under Uncertainties. *IEEE Trans. Smart Grid* **2024**, *15*, 4649–4661.
17. Liu, Q.; Shang, Z.; Lu, S.; Liu, Y.; Liu, Y.; Yu, S. Physics-Guided TL-LSTM Network for Early-Stage Degradation Trajectory Prediction of Lithium-Ion Batteries. *J. Energy Storage* **2025**, *106*, 114736.
18. Lu, S.; Yan, Y.; Zhang, Y.; Liu, Y.; Yu, S.; Gao, Z.W. A Knowledge Distillation-Based Network Compression Framework for Lifecycle Management of Lithium-Ion Batteries. *J. Supercomput.* **2025**, *81*, 1147.
19. Wu, X.; Hu, X.; Moura, S.; Yin, X.; Pickert, V. Stochastic control of smart home energy management with plug-in electric vehicle battery energy storage and photovoltaic array. *J. Power Sources* **2016**, *333*, 203–212.
20. Zou, B.; Peng, J.; Yin, R.; Luo, Z.; Song, J.; Ma, T.; Li, S.; Yang, H. Energy management of the grid-connected residential photovoltaic-battery system using model predictive control coupled with dynamic programming. *Energy Build.* **2023**, *279*, 112712.
21. Datta, U.; Kalam, A.; Shi, J. Electric vehicle in home energy management to reduce electricity cost. *J. Sci. Ind. Res.* **2018**, *77*, 559–565.
22. Beaudin, M.; Zareipour, H. Home energy management systems: A review of modelling and complexity. *Renew. Sustain. Energy Rev.* **2015**, *45*, 318–335.

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.