sustainability

MDPI

Article

Hierarchical PPO–TD3 Energy Management for Renewable-Rich Microgrids with Safety-Filtered Action Projection

Yuanyuan Xu [1,*], Yixin Lin [1], Shuhao Li [1], Xiutao Gao [1] and Zhichun Liu [2]

[1] College of Information Science and Engineering, Huaqiao University, Xiamen 361021, China; yxlin@stu.hqu.edu.cn (Y.L.); lishuhao@stu.hqu.edu.cn (S.L.); gaoxiutao@stu.hqu.edu.cn (X.G.)
[2] School of Aerospace Engineering, Xiamen University, Xiamen 361005, China; liuzhichun@xmu.edu.cn
* Correspondence: yyxu@hqu.edu.cn

Abstract

For renewable-rich microgrids, effective energy management is essential to improve economic efficiency and renewable-energy utilization while maintaining supply adequacy and network security. However, online dispatch remains challenging because variable renewable generation and demand must be coordinated under battery operating limits, distribution-network voltage constraints, and limited forecast information. To address this challenge, this paper proposes a safety-filtered hierarchical PPO–TD3 framework for microgrid energy management. The upper PPO layer selects coarse battery operating modes for long-horizon economic scheduling, while the lower TD3 layer refines continuous battery-power commands for short-term regulation. Before execution, the fused command is projected through a model-based safety filter considering state-of-charge limits, supply-priority rules, and voltage screening under a 33-bus radial-network surrogate. Fixed-test evaluation over five random seeds and 62 chronologically held-out test days shows that the proposed controller reduces total operating cost from 1003.66 to 686.31 k c.u. (31.6%) and unserved energy from 1743.25 to 1052.14 kWh (39.6%) relative to MPC-Lite, while reducing average decision time from 180.21 to 43.55 ms per step. In a separate safety-filter ablation, the filter reduced total cost from 912.47 to 686.31 k c.u., the hard-voltage violation rate from 25.31% to 19.46%, and the executed state-of-charge violation rate from 0.40% to 0.00%. These results indicate that the proposed framework supports reliable and efficient renewable-rich microgrid operation, facilitating renewable-energy integration and enhancing operational resilience.

Keywords: microgrid energy management; renewable energy integration; hierarchical reinforcement learning; PPO–TD3; safety-filtered action projection; energy resilience

Check for updates

Academic Editor: Hugo Miguel Loureiro Algarvio

Received: 29 June 2026
Revised: 17 August 2026
Accepted: 20 August 2026
Published: 21 August 2026

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

1. Introduction

The increasing demand for low-carbon and resilient energy systems has promoted the deployment of distributed renewable energy resources, particularly photovoltaic (PV) and wind generation, in distribution networks. Compared with centralized generation, distributed renewable resources can utilize locally available clean energy and reduce the dependence of local loads on conventional dispatchable generation [1]. Microgrids provide an effective platform for integrating these distributed resources with battery energy storage systems (BESSs) and backup generation, thereby improving local energy utilization, supply reliability, and distribution-side operational flexibility [2]. Such microgrids can serve aggregated residential, commercial, or light-industrial loads. In this study, the

demand is represented by aggregated distribution-level load profiles rather than a sector-specific end-use model; therefore, the proposed framework is not restricted to a particular customer sector.

However, the intermittent and variable nature of renewable generation introduces additional challenges to distribution-level energy management. A renewable-rich microgrid requires coordinated operation of PV generation, wind generation, dispatchable diesel generation, BESSs, and time-varying loads under renewable fluctuations, load variations, battery state-of-charge (SoC) limits, device power constraints, and network voltage limits. These coupled constraints can significantly narrow the feasible operating region of the microgrid. Therefore, an online energy management strategy should not only reduce operating cost and renewable curtailment but also maintain supply adequacy and executed-action feasibility under network-related constraints [3].

1.1. Related Work

Model-based optimization methods, such as model predictive control (MPC) and mixed-integer linear programming (MILP), have been widely used for constrained microgrid dispatch. Hu et al. [4] reviewed economic MPC approaches for microgrid optimization and showed that MPC provides systematic receding-horizon decision-making with explicit constraint handling. Liu et al. [5] developed a MILP-based distributed energy management framework for networked microgrids considering network operational objectives and constraints. Nguyen et al. [6] further incorporated BESS degradation and renewable-energy uncertainty into microgrid optimal scheduling. Although these model-based approaches provide interpretable solutions and structured constraint handling, their performance depends on model fidelity and forecast accuracy, while repeated optimization may increase computational burden when network constraints, storage dynamics, and multiple operating objectives are considered. These limitations motivate the development of faster feedback-based online dispatch methods.

Recent studies have also addressed distribution-level energy management and network operation from multi-objective and multi-bus perspectives. Zeyad et al. [7] developed a multi-objective home energy management and peer-to-peer energy trading framework for a rooftop-PV-powered residential community, jointly coordinating PV generation, battery storage, and flexible demand to improve renewable-energy utilization and reduce grid dependence. Bais et al. [8] investigated data-driven faulty-line localization in the IEEE 30-bus system using a CNN-LSTM model, illustrating the application of learning-based methods to network-aware monitoring in multi-bus power systems. Mohamed et al. [9] formulated an optimal power flow problem for the IEEE 30-bus system and employed an improved particle swarm optimization method to reduce generation fuel cost while satisfying system operating constraints. These studies highlight the increasing importance of coordinated energy management, bus-level network information, and power-flow constraints in modern distribution and power-system operation. However, their primary focuses are respectively transactive residential energy management, fault localization, and optimization-based power flow, rather than execution-time safety correction of hierarchical learned dispatch actions.

Deep reinforcement learning (DRL) provides a promising alternative because it can learn dispatch policies from interactions with uncertain operating environments and generate online actions with low computational cost after training. Lee et al. [10] developed a DRL-based energy management architecture for microgrids and demonstrated the potential of learning-based control for coordinating multiple energy resources. Guo et al. [11] investigated real-time microgrid energy management under renewable and load uncertainties, highlighting the online decision-making capability of DRL. Domínguez-Barbero et al. [12]

incorporated nonlinear battery losses into a DRL-based microgrid EMS to improve the representation of battery operation. Nevertheless, these studies mainly focused on economic dispatch, uncertainty handling, or battery-related modeling, while explicit execution-time handling of network-constrained actions received less attention. More generally, standard DRL policies do not inherently enforce physical constraints during exploration or deployment, which may result in infeasible battery-power commands, SoC violations, increased load shedding, or voltage deviations in distribution-network microgrids.

To improve the operational feasibility of learning-based energy management, recent studies have introduced explicit safety mechanisms into reinforcement learning. Hao et al. [13] developed a Lyapunov-based safe reinforcement learning framework for microgrid energy management to incorporate operational constraints into the learning process. Wang et al. [14] proposed a physics-informed safe reinforcement learning approach for secure energy management of multi-energy microgrids, while Yang et al. [15] further investigated physics-shielded DRL for microgrid energy management with battery-health considerations. These studies demonstrate the value of combining physical knowledge or constraint-handling mechanisms with learned control policies. However, their safety mechanisms are mainly associated with the output of a single decision policy, whereas the execution-time feasibility of commands generated through hierarchical action coordination has received less attention.

Hierarchical reinforcement learning provides another direction for reducing the complexity of microgrid energy management by decomposing the control problem across different decision levels. Xia et al. [16] developed a hierarchical safe DRL method for decentralized coordination of networked microgrids, demonstrating the potential of hierarchical decision structures for complex energy-management tasks. Xu et al. [17] proposed a hierarchical trust-region multi-agent reinforcement learning framework for collaborative optimization of multi-energy multi-microgrid systems. Nevertheless, these hierarchical approaches mainly emphasize task decomposition and coordination among decision layers or agents, while explicit safety projection of the fused hierarchical command before physical execution remains insufficiently explored. Consequently, the coordinated integration of two-timescale hierarchical action generation and execution-time safety projection under battery and distribution-network constraints remains an open issue.

Overall, the existing literature has established several important directions for microgrid energy management. Model-based approaches provide explicit and interpretable constraint handling, but their online implementation may require repeated optimization and reliable model or forecast information. DRL-based approaches improve online decision efficiency, but standard learned policies do not inherently guarantee the feasibility of executed actions. Safe DRL introduces explicit constraint-handling mechanisms, whereas most existing studies apply such mechanisms to the output of a single policy. Hierarchical reinforcement learning improves task decomposition and coordination, but the feasibility of the fused command generated by different decision layers has received comparatively limited attention. Therefore, a remaining research gap lies in jointly coordinating two-timescale hierarchical action generation with execution-time safety projection under both battery and distribution-network constraints. The present study addresses this gap by combining a discrete upper-level scheduling policy, a continuous lower-level refinement policy, and a model-based safety filter applied to the fused command before execution.

1.2. Research Contributions

To address these issues, this paper develops a safety-filtered hierarchical reinforcement learning framework that coordinates two-timescale action generation, state-dependent execution-time projection, and projection-aware learning feedback for online energy man-

agement of a network-constrained microgrid. The upper layer selects coarse battery operating modes for economic scheduling, while the lower layer generates bounded continuous battery-power refinements for short-term regulation. Before execution, the fused battery command is corrected by a lightweight model-based action-projection layer considering device power bounds, tightened SoC limits, supply-priority rules, and voltage screening in a 33-bus radial network surrogate. The correction magnitude is further used as a penalty signal to discourage persistent reliance on projection-based fallback actions.

The main methodological contributions of this paper are summarized as follows.

1. A two-timescale discrete–continuous coordination mechanism is developed for battery dispatch. The upper policy provides a slowly updated operating reference, while the lower policy generates bounded step-wise power refinements without fully overriding the upper-layer schedule.
2. A state-dependent execution-time projection mechanism is introduced for the fused hierarchical command. Unlike fixed action clipping, the proposed mechanism jointly screens battery power limits, tightened SoC bounds, supply-priority requirements, and network-voltage constraints, and selects a minimum-violation fallback action when the fully feasible set is empty.
3. A projection-aware learning mechanism is developed by incorporating the safety-filter correction magnitude into the lower-layer feedback. Raw-policy actions and executed actions are also separately monitored, allowing persistent dependence on projection-based correction to be identified and discouraged.

The proposed framework is evaluated through fixed-test comparisons with MPC-based, single-layer DRL, and safety-filter ablation baselines, with emphasis on operating cost, supply adequacy, executed-action feasibility, and online decision time.

2. System Model and Problem Formulation

2.1. Network-Constrained Microgrid Model

Figure 1 shows the 33-bus radial microgrid adopted in this study. Bus 0 is the slack bus, and buses 1–32 are active buses with distributed load allocation. PV generation is assigned to buses 5, 11, 17, 23, 29, and 31, while wind generation is assigned to buses 8, 14, 24, and 30. The BESS and DG are connected to buses 18 and 25, respectively. This structure moves beyond the common-bus approximation and provides a network-aware basis for evaluating battery operation, diesel dispatch, renewable utilization, and voltage-related operating constraints.

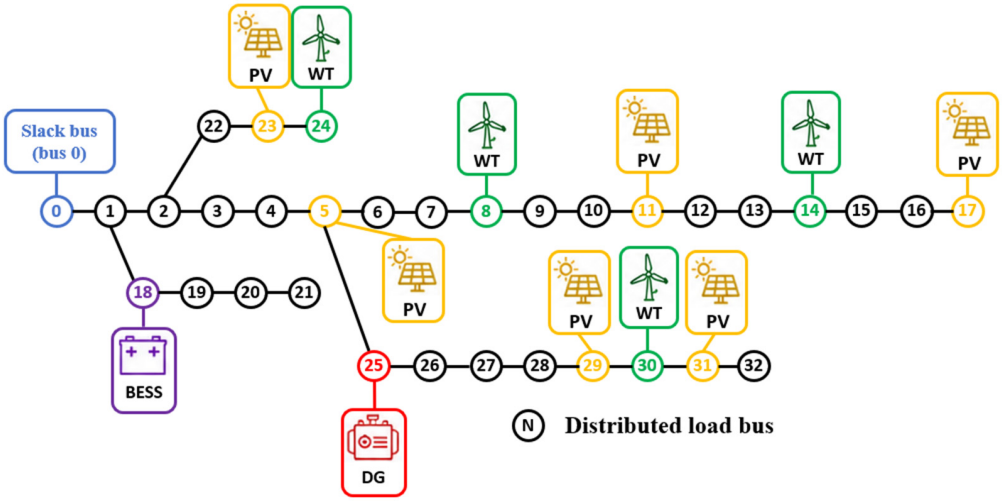


Figure 1. 33-bus radial microgrid used for network-constrained energy management.

The available PV and wind power profiles are obtained from time-series data rather than recalculated from detailed irradiance, temperature, or wind-speed conversion models. At each time step t, the active-power balance is expressed as

$$P_{pv,t} + P_{wt,t} + P_{dg,t} - P_{b,t} + P_{shed,t} = P_{L,t} + P_{loss,t} \tag{1}$$

where $P_{pv,t}$, $P_{wt,t}$, and $P_{dg,t}$ denote the utilized PV, wind, and diesel-generator powers, respectively. $P_{b,t}$ is the BESS power, where positive values indicate charging and negative values indicate discharging. Therefore, charging reduces the power available to the load, while discharging increases it. $P_{L,t}$ is the load demand, $P_{loss,t}$ is the network loss proxy, and $P_{shed,t}$ is a non-negative slack variable representing unserved load.

The total available renewable generation is defined as the sum of the available PV and wind power:

$$P_{re,t}^{ava} = P_{pv,t}^{ava} + P_{wt,t}^{ava} \tag{2}$$

The utilized PV and wind power are bounded by their available outputs:

$$0 \le P_{pv,t} \le P_{pv,t}^{ava},\ 0 \le P_{wt,t} \le P_{wt,t}^{ava} \tag{3}$$

Renewable curtailment is calculated as the difference between available renewable generation and utilized renewable generation:

$$P_{cur,t} = P_{re,t}^{ava} - P_{pv,t} - P_{wt,t} \tag{4}$$

where $P_{cur,t}$ denotes the curtailed renewable power at time step t. This formulation allows the dispatch model to evaluate supply adequacy, renewable utilization, and network-aware operating performance under the same dispatch framework.

2.2. Operating Constraints and Network Approximation

The diesel generator output is determined according to the residual demand after renewable generation and battery action, subject to capacity and ramping limits. Its output is constrained by

$$P_{dg}^{min} \le P_{dg,t} \le P_{dg}^{max} \tag{5}$$

where P_{dg}^{min} and P_{dg}^{max} denote the minimum and maximum diesel-generator outputs, respectively. The ramping behavior of the diesel generator is limited by

$$-R_{dg}^{down} \le P_{dg,t} - P_{dg,t-1} \le R_{dg}^{up} \tag{6}$$

where R_{dg}^{down} and R_{dg}^{up} denote the downward and upward ramping limits per time step. The diesel-generation cost $C_{dg,t}$ is calculated using a lightweight affine fuel-cost model.

The BESS charging and discharging power is constrained by

$$-P_{b}^{dis,max} \le P_{b,t} \le P_{b}^{ch,max} \tag{7}$$

where $P_{b}^{dis,max}$ and $P_{b}^{ch,max}$ are the maximum discharging and charging powers, respectively. According to the sign convention in this paper, $P_{b,t} > 0$ represents charging, while $P_{b,t} < 0$ represents discharging. The SoC dynamics of the BESS are described as

$$SoC_{t+1} = SoC_t + \frac{\eta_{ch}\max(P_{b,t}, 0)\Delta t}{E_b} - \frac{\max(-P_{b,t}, 0)\Delta t}{\eta_{dis}E_b} \tag{8}$$

where E_b is the battery energy capacity, η_{ch} and η_{dis} are the charging and discharging efficiencies, respectively, and $\Delta t = 1$ h is the sampling interval adopted in this study. The SoC is constrained within its lower and upper bounds:

$$SoC^{min} \leq SoC_t \leq SoC^{max} \tag{9}$$

A lightweight nonlinear battery-degradation surrogate is adopted to account for the combined effects of battery throughput, SoC swing, C-rate, and operation away from the mid-SoC region. The normalized throughput, SoC swing, C-rate proxy, and mid-SoC stress are respectively defined as:

$$\phi_t = \frac{|P_b(t)|\Delta t}{E_b} \tag{10}$$

$$d_t = |Soc(t+1) - SoC(t)| \tag{11}$$

$$c_t = \frac{|P_b(t)|}{E_b} \tag{12}$$

$$s_t = \left|\frac{Soc(t) + SoC(t+1)}{2} - 0.5\right| \tag{13}$$

The battery degradation cost is then calculated as:

$$C_{deg}(\mathrm{t}) = c_b \phi_t [1 + \alpha_d d_t^{1.3} + \alpha_c c_t^{1.2} + \alpha_s s_t^{1.1}] \tag{14}$$

where $c_b = 5.0$, $\alpha_d = 2.0$, $\alpha_c = 1.5$, and $\alpha_s = 0.8$. The sampling interval is $\Delta t = 1$ h in this study. The resulting degradation cost $C_{deg}(\mathrm{t})$ is directly subtracted from the upper-layer reward with a unit weighting coefficient. Consequently, although the normalized throughput ϕ_t and the C-rate proxy are numerically equal under the adopted hourly resolution, they represent different physical aspects of battery operation and enter the degradation surrogate with different nonlinear exponents and weighting coefficients. This degradation-aware formulation is consistent with battery-aging studies showing that battery life loss is closely related to operating profiles and cycling patterns [18]. The surrogate is intended to discourage aggressive battery operation while remaining computationally lightweight rather than reproducing a detailed electrochemical aging process.

To include network-related operating effects, the 33-bus feeder is represented by a linearized radial-network approximation. Radial distribution-network and branch-flow models have been widely used to describe voltage-drop and branch-flow relationships in distribution-level optimization problems [19,20]. For each active bus i, the net demand is defined as

$$P_{net,i,t} = P_{L,i,t} - P_{pv,i,t} - P_{wt,i,t} - P_{dg,i,t} + P_{b,i,t} \tag{15}$$

where positive $P_{net,i,t}$ denotes net demand at bus i, while negative $P_{net,i,t}$ denotes net injection. The branch flow is accumulated from downstream buses as

$$F_{i,t} = P_{net,i,t} + \sum_{j \in c(i)} F_{j,t} \tag{16}$$

where $c(i)$ denotes the set of child buses of bus i. The predicted voltage of each bus is calculated by

$$V_{i,t}^{pred} = V_{\pi(i),t}^{pred} - \rho_i F_{i,t} \tag{17}$$

where $\pi(i)$ is the parent bus of bus i, and ρ_i is the equivalent resistance-related coefficient of the corresponding branch. To capture distribution-level voltage quality, the soft voltage operating band is imposed as

$$V^{soft,min} \leq V_{i,t}^{pred} \leq V^{soft,max} \tag{18}$$

Voltage deviations outside this soft band are penalized in the objective function, while hard-band violations are reported as evaluation metrics. The line-loss proxy is calculated from the branch-flow magnitude:

$$P_{loss,t} = \kappa_{loss} \sum_i F^2{}_{i,t} \tag{19}$$

where κ_{loss} is the loss weighting coefficient. The linearized radial-network model is adopted as a computationally efficient surrogate for online action screening. In the proposed controller, the safety filter may evaluate multiple candidate battery commands at each decision step; therefore, repeatedly solving a nonlinear AC power-flow problem would substantially increase the online computational burden. Compared with a common-bus representation, the adopted 33-bus surrogate retains the radial feeder topology, distributed power injections, branch-flow accumulation, and bus-dependent voltage variations, thereby providing a network-aware basis for action evaluation. Moreover, all controllers are evaluated using the same network model and operating conditions to ensure consistent comparisons. The approximation error may become more pronounced under severe supply-deficit or load-shedding conditions, where the operating point can deviate substantially from nominal conditions. The resulting voltage metrics are therefore interpreted as surrogate-based screening indicators rather than exact nonlinear AC power-flow solutions. Similar surrogate-based formulations have been adopted in microgrid energy management to balance network awareness and computational efficiency [21,22].

2.3. Dispatch Objective

The objective of microgrid energy management is to minimize the expected operating cost over the dispatch horizon while maintaining supply adequacy, battery feasibility, and voltage quality. The total operating cost consists of diesel-generation cost, load-shedding penalty, renewable-curtailment penalty, battery degradation cost, line-loss-related cost, and voltage-violation penalty. Therefore, the dispatch objective is formulated as

$$\min J = \mathrm{E}[\sum_{t=1}^{T} (C_{dg,t} + C_{shed,t} + C_{cur,t} + C_{deg,t} + C_{loss,t} + C_{v,t})] \tag{20}$$

where $C_{dg,t}$ is the diesel-generation cost, $C_{shed,t}$ is the load-shedding penalty, $C_{cur,t}$ is the renewable-curtailment penalty, $C_{deg,t}$ is the battery degradation cost, $C_{loss,t}$ is the line-loss-related cost, and $C_{v,t}$ is the voltage-violation penalty.

All economic terms are expressed in normalized cost units (c.u.) defined by the simulation cost function. The unit c.u. is not intended to represent a specific national currency, because this study uses synthetic operating-cost coefficients rather than location-specific market tariffs. For readability, total operating costs are reported in k c.u. (where 1 k c.u. = 1000 c.u.). All cost components, including diesel-generation, unserved-energy, curtailment, line-loss-related, voltage-violation, and battery-degradation penalties, use the same cost-unit definition.

In this objective, the load-shedding penalty reflects supply adequacy, the renewable-curtailment penalty reflects renewable-energy utilization, the battery degradation cost discourages excessive cycling, and the voltage-violation penalty encourages operation within the network voltage limits. This objective provides the system-level basis for the reward design of the proposed hierarchical reinforcement learning controller.

3. Safety-Filtered Hierarchical PPO-TD3

To address the coupling between economic scheduling and operational safety in network-constrained microgrid dispatch, a safety-filtered hierarchical PPO-TD3 framework is developed, as shown in Figure 2. The proposed controller decomposes the BESS dispatch

decision into two layers. The upper PPO layer selects a coarse battery operating mode for long-horizon economic scheduling, while the lower TD3 layer generates a bounded continuous power refinement for short-term SoC and voltage regulation. Before execution, the fused BESS command is corrected by a model-based action-projection layer considering battery power limits, tightened SoC bounds, supply-priority rules, and voltage screening under the radial-network surrogate. This design is consistent with recent safe DRL studies in energy management, where safety-aware constraint handling is introduced to improve operational feasibility before execution [23]. Unlike single-policy safety correction, the proposed filter is applied to the fused command generated by the two hierarchical decision layers, thereby serving as an execution interface between learned dispatch decisions and physical operating constraints.

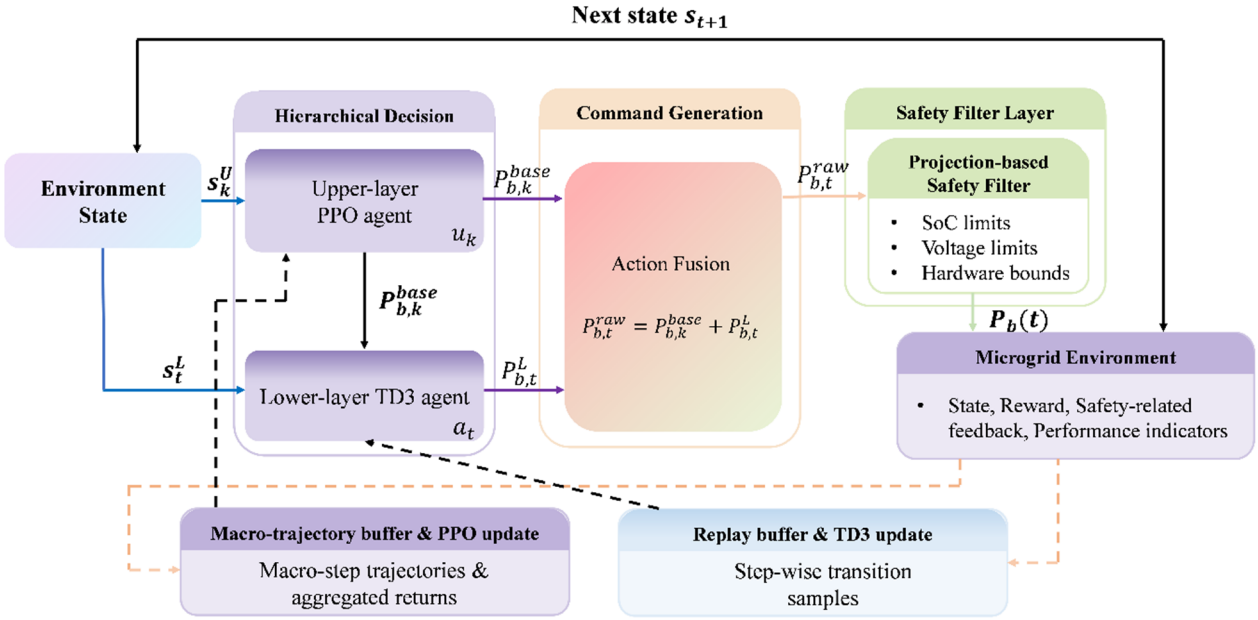


Figure 2. Structure of the safety-filtered hierarchical PPO-TD3 controller.

3.1. Hierarchical Action Generation

At each macro decision step k, the upper PPO policy selects a normalized battery operating mode u_k from the predefined discrete set U = {−1, −0.5, 0, 0.5, 1}.

$$P_{b,k}^{U} = u_k P_b^{rated} \tag{21}$$

where $P_b^{\text{rated}} = 5$ kW is the rated BESS charging and discharging power. The selected upper-layer action is held for $K = 2$ environment steps, allowing the upper layer to focus on slower economic scheduling rather than step-wise corrective regulation.

At each environment step t, the lower layer is implemented using TD3, an off-policy actor–critic algorithm that employs twin critics and delayed policy updates to reduce overestimation bias [24,25]. The TD3 policy outputs a normalized continuous action $a_t^L \in [-1, 1]$, which is converted into a bounded battery-power offset according to

$$\Delta P_{b,t}^{L} = \sigma a_t^L P_b^{rated} \tag{22}$$

where $\sigma = 0.10$ is the dimensionless lower-layer offset coefficient. Therefore, the lower layer can adjust the upper-layer reference within [−0.5,0.5] kW. This coefficient is fixed across all training seeds and test days to restrict the lower layer to local regulation and prevent it from completely overriding the upper-layer operating mode.

The fused BESS command before safety filtering is then obtained as

$$P_{b,t}^{raw} = P_{b,k}^{U} + \Delta P_{b,t}^{L}, \qquad kK \le t \le (k+1)K \tag{23}$$

Following the action-fusion step, the raw command $P_{b,t}^{raw}$ is hard-clipped to the physical BESS power range $[-P_b^{dis,max},\ P_b^{ch,max}]$. The resulting clipped command is then passed to the safety filter for state-dependent projection considering the tightened SoC bounds, supply-priority rule, and voltage-screening constraints.

This two-timescale action structure allows the upper layer to determine the coarse battery operating direction, while the lower layer performs bounded step-wise refinement around the upper-layer reference.

The upper-layer observation includes load, renewable generation, SoC margin, voltage deviation, available DG capacity, deficit/surplus indicators, and time information. The lower-layer observation emphasizes short-term SoC margin, voltage deviation, and the upper-layer battery reference. The upper reward mainly reflects economic terms, including diesel generation, load shedding, renewable curtailment, and battery cycling cost. The lower reward emphasizes voltage quality, SoC safety, alignment with the upper-layer reference, and reduced reliance on safety-filter correction.

3.2. Safety-Filtered Action Projection

The safety filter is placed between the learned policy and the microgrid environment. Different from pure penalty-based constraint handling, this layer directly modifies the candidate action before execution. The fused command is first clipped to the physical BESS power range. A state-dependent admissible interval $I_t = \left[P_{b,t}^{min}, P_{b,t}^{max}\right]$ is then constructed according to the BESS power limits, the tightened SoC interval [0.12,0.88], and the supply-priority rule that prevents additional charging-induced power shortages. This execution-time correction mechanism is related to safety-layer-based action projection in safe reinforcement learning [26].

Within this interval, the filter generates 121 equally spaced candidate actions:

$$c_t = linspace(P_{b,t}^{min}, P_{b,t}^{max}, 121) \tag{24}$$

Each candidate is evaluated using the noise-free radial-network surrogate. The voltage-feasible candidate set is defined as

$$F_t = \{\mathrm{p} \in c_t : \hat{V}_t^{min}(p) \ge V_{min}^{SF}, \quad \hat{V}_t^{max}(p) \le V_{max}^{SF}\} \tag{25}$$

where the effective voltage-screening limits are $V_{min}^{SF} = 0.9745$ p.u. and $V_{max}^{SF} = 1.0255$ p.u. under the validation and test settings.

Let $P_{b,t}^{0}$ denote the raw command after physical-power, SoC, and supply-priority clipping. The executed command is selected according to

$$P_{b,t}^{exec} = \begin{cases} P_{b,t}^{0}, & P_{b,t}^{0}\ satisfies\ the\ limits, \\ \underset{p \in F_t}{arg\min} |p - P_{b,t}^{0}|, & F_t \neq \varnothing, \\ \underset{p \in c_t}{lexarg\min} (J_{v,t}(p), |p - P_{b,t}^{0}|), & F_t = \varnothing, \end{cases} \tag{26}$$

The voltage-violation score $J_{v,t}(p)$ is defined as

$$J_{v,t}(p) = \max(0, V_{min}^{SF} - \hat{V}_t^{min}(p) + \max(0, \hat{V}_t^{max}(p) - V_{max}^{SF})) \tag{27}$$

where $J_{v,t}(p)$ is the voltage-violation score. The lexicographic rule first minimizes the voltage violation and then selects the action closest to the clipped policy command when multiple candidates have the same violation score.

The safety-filter correction magnitude is recorded as

$$\Delta P_t^{SF} = |P_{b,t}^{exec} - P_{b,t}^{0}| \tag{28}$$

This correction magnitude is included in the lower-layer feedback to discourage persistent reliance on execution-time correction. Therefore, the filter not only improves executed-action feasibility but also encourages the learned policy to generate commands closer to the admissible region.

3.3. Training and Online Execution

Training follows the hierarchical decision structure. At macro decision points, the upper PPO layer selects a discrete battery operating mode and keeps the corresponding baseline power fixed within the macro interval. At each environment step, the lower TD3 layer generates a continuous offset around this baseline. The fused raw command is then passed through the safety filter before being applied to the microgrid environment.

The upper layer is updated using macro-step trajectories with aggregated economic returns, while the lower layer is trained from step-wise transition samples with safety-oriented feedback. The safety-filter correction magnitude, raw-command violation indicators, and executed-action feasibility indicators are recorded during training and evaluation. In the final setting, fixed inter-layer coordination is adopted, where the lower-layer offset scale and alignment strength are kept constant to prevent the lower layer from excessively overriding the upper-layer schedule.

To improve generalization, domain randomization is applied to exogenous load and renewable profiles during training. Specifically, during training, the load scaling factor is independently sampled from a uniform distribution over [0.95, 1.05], corresponding to a $\pm$5% variation, while the renewable-generation scaling factor is sampled from [0.90, 1.10], corresponding to a $\pm$10% variation. The same renewable scaling factor is applied to both PV and wind generation at each time step. Validation and test profiles are kept fixed without domain randomization. During online execution, the trained controller only requires the current state as input and produces the safety-filtered BESS command through the sequence: upper-layer mode selection, lower-layer offset generation, action fusion, safety projection, and environment execution.

4. Results and Discussion

4.1. Experimental Setup

The proposed method was evaluated on the radial network-constrained microgrid described in Section 2. The microgrid includes PV units, wind turbines, a diesel generator, a BESS, and distributed loads assigned to a 33-bus radial feeder. Therefore, dispatch performance was evaluated using both system-level operating indicators and network-aware indicators, including bus-voltage deviations and line-loss-related effects under the adopted linearized radial-network surrogate.

Figure 3 illustrates the dispatch components of a representative held-out test day. The black line denotes the load demand, and the stacked bars show PV generation, wind generation, diesel generation and BESS discharge. Negative bars indicate BESS charging, while red bars denote unserved energy. This profile provides an intuitive view of the dispatch task considered in the experiments, where intermittent renewable generation, diesel backup and battery operation must be coordinated under time-varying demand.

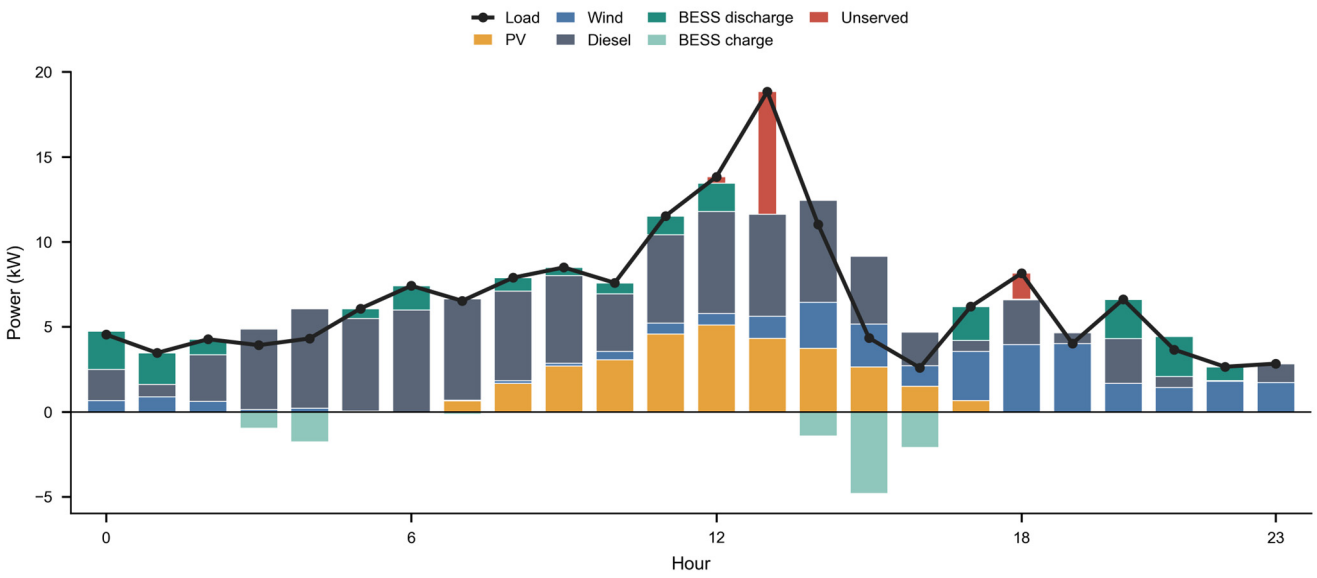


Figure 3. Representative held-out test-day dispatch components.

Each episode represents a 24-step daily dispatch cycle, with a one-hour time step ($\Delta t = 1$ h). The dataset was split chronologically into training, validation, and test subsets to avoid temporal leakage. The DRL controllers were trained for five random seeds. For each seed, the checkpoint with the lowest validation cost was selected and evaluated without further training or parameter updates on the same 62 chronologically held-out test days, corresponding to 1488 hourly decision steps. During training, load profiles were randomly scaled within 0.95–1.05 and renewable-generation profiles within 0.90–1.10, whereas the validation and test profiles were kept fixed. The main system, simulation, and controller parameters used in the experiments are summarized in Table 1.

Table 1. Key experimental and system settings.

Parameter	Value
Network configuration	33-bus radial distribution network
BESS configuration	20 kWh capacity; ±5 kW power limit; SoC range: 0.10–0.90
DG configuration	6 kW rated power; 2 kW/step ramp limit
Simulation resolution	1 h per step; 24 steps per daily episode
Safety-filter setting	Tightened SoC interval: 0.12–0.88; discrete candidate-based action projection
Training protocol	five independent random seeds; train/validation/test split
MPC-Lite baseline	Short-horizon receding optimization with persistence forecasting
Safety-filter candidates	121 equally spaced candidate actions per decision step
Voltage limits	Hard band: 0.95–1.05 p.u.; SF band: 0.9745–1.0255 p.u.
Domain randomization	Load scaling: 0.95–1.05; renewable scaling: 0.90–1.10 (training)

The proposed controller was compared with four baselines. The first baseline is MPC-Lite, a rolling-horizon MPC controller using persistence forecasts [27]. The second baseline is a hierarchical PPO–TD3 controller without the safety filter, which is included in the main comparative evaluation as a representative no-projection hierarchical baseline. The third baseline is a single-layer TD3 controller with the same safety filter, which is used to evaluate the effect of hierarchical action decomposition. The fourth baseline is a hierarchical PPO–DQN controller with the same safety filter. This variant retains the upper-level PPO scheduling layer and the safety-filtered execution mechanism but replaces the lower-level

TD3 controller with a DQN-based controller. It is included as a representative alternative hierarchical design to examine whether the observed performance depends specifically on the TD3 lower-level controller. All methods were evaluated using the same fixed 62-day test block, initial conditions, cost function, and network-aware evaluation metrics.

4.2. Main Online Dispatch Performance

Figure 4 and Table 2 summarize the fixed-test online dispatch results over five random seeds and 62 held-out test days. The proposed Hier. PPO–TD3 + SF controller achieved a mean total operating cost of 686.31 k c.u., compared with 1003.66 k c.u. for MPC-Lite, 848.27 k c.u. for the hierarchical PPO–TD3 controller without the safety filter, 717.05 k c.u. for the single-layer TD3 controller with the safety filter, and 698.47 k c.u. for the hierarchical PPO–DQN controller with the safety filter. Accordingly, the proposed controller reduced total cost by 31.6% relative to MPC-Lite, 19.1% relative to the no-safety hierarchy, 4.3% relative to the single-layer TD3 controller, and 1.7% relative to the PPO–DQN hierarchy.

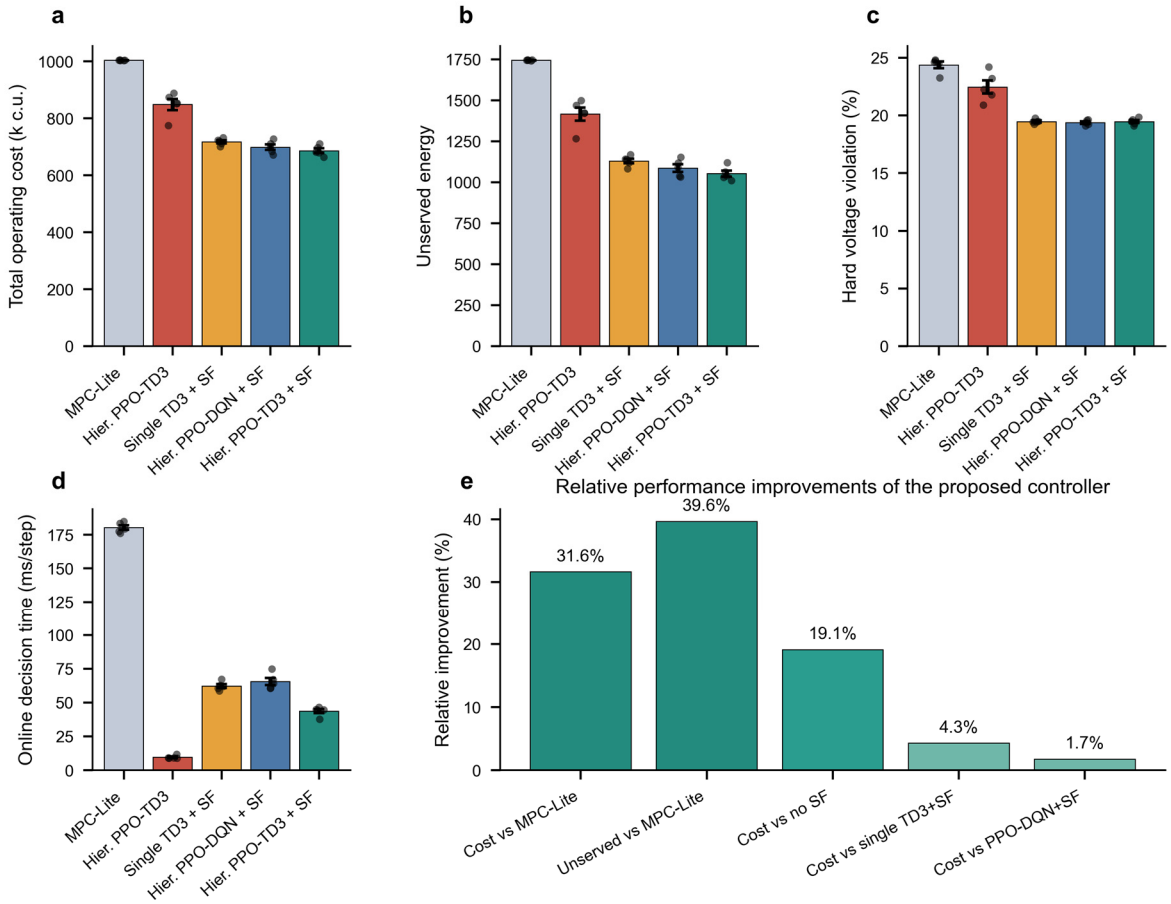


Figure 4. Fixed-test online dispatch comparison. (**a**) Total operating cost; (**b**) Unserved energy; (**c**) Hard-voltage violation rate; (**d**) Online decision time; (**e**) Relative performance improvements of the proposed controller.

The proposed controller also reduced unserved energy to 1052.14 kWh, compared with 1743.25 kWh for MPC-Lite. This corresponds to a 39.6% reduction relative to MPC-Lite. The reductions relative to the no-safety hierarchy, single-layer TD3 + SF, and hierarchical PPO–DQN + SF were 25.6%, 6.8%, and 3.1%, respectively. The average online decision time of the proposed controller was 43.55 ms/step, compared with 180.21 ms/step for MPC-Lite, corresponding to an approximately 4.14-fold speed advantage over MPC-Lite.

Table 2. Fixed-test online dispatch comparison over five seeds and 62 held-out days.

Method	Total Cost (k c.u.)	Unserved Energy (kWh)	Hard Voltage Violation (%)	Executed SoC Violation (%)	Decision Time (ms/Step)
MPC-Lite	1003.66 ± 0.89	1743.25 ± 1.80	24.37 ± 0.64	0.00 ± 0.00	180.21 ± 3.76
Hier. PPO-TD3	848.27 ± 43.71	1414.97 ± 89.60	22.46 ± 1.27	0.52 ± 1.06	9.50 ± 1.20
Single TD3 + SF	717.05 ± 11.94	1128.97 ± 31.97	19.46 ± 0.19	0.00 ± 0.00	62.20 ± 3.21
Hier. PPO–DQN + SF	698.47 ± 21.95	1085.55 ± 51.79	19.37 ± 0.22	0.00 ± 0.00	65.36 ± 5.86
Hier. PPO-TD3 + SF (Proposed)	686.31 ± 18.30	1052.14 ± 42.15	19.46 ± 0.29	0.00 ± 0.00	43.55 ± 3.36

The no-safety hierarchical controller remained the fastest method, with an average decision time of 9.50 ms/step, because it does not perform safety projection. Nevertheless, the proposed controller was faster than the single-layer TD3 + SF and hierarchical PPO–DQN + SF baselines. Compared with the single-layer TD3 + SF controller, the proposed method achieved lower cost and unserved energy while maintaining the same executed SoC violation rate of zero. The hard-voltage violation rates of the proposed PPO–TD3 + SF and PPO–DQN + SF controllers were comparable, although the PPO–DQN variant showed a marginally lower value. These results indicate a framework-level safety–economy–computation trade-off.

To assess whether the aggregate performance advantage was driven by a single random initialization, Table 3 reports the seed-wise total operating costs and relative improvements for the proposed controller and MPC-Lite. Across all five seeds, the proposed controller reduced total operating cost by 29.2–34.0%, with a mean improvement of 31.6 ± 1.8%. This consistent seed-wise improvement indicates that the reported advantage was not determined by a single favorable initialization.

Table 3. Seed-wise total operating cost comparison over five random seeds and 62 held-out test days.

Seed	Proposed Cost (k c.u.)	MPC-Lite Cost (k c.u.)	Relative Improvement
42	663.14	1004.79	34.0%
2024	680.78	1002.53	32.1%
2026	711.06	1004.27	29.2%
3407	679.59	1003.56	32.3%
5200	697.00	1003.17	30.5%
Mean ± SD	686.31 ± 18.30	1003.66 ± 0.89	31.6 ± 1.8%

4.3. Ablation Analysis of the Safety Filter

To further isolate the contribution of the safety-filtered execution layer, a separate paired ablation was conducted using the PPO–TD3 implementation, as shown in Figure 5. Over the same 62 held-out test days, adding the safety filter reduced the mean total operating cost from 912.47 k c.u. to 686.31 k c.u., corresponding to a 24.8% reduction. Unserved energy decreased from 1545.98 kWh to 1052.14 kWh, while the hard-voltage violation rate decreased from 25.31% to 19.46%. The paired seed-level results in Figure 5a–c show that these improvements were observed consistently across the five random seeds.

The raw-policy and executed-action metrics in Figure 5d–f further clarify the source of the safety improvement. Under the safety-filtered controller, the raw SoC violation rate was 7.92%, whereas the executed SoC violation rate was 0.00%. Without the safety filter, the executed SoC violation rate was 0.40%. These results show that the learned policy should not be described as intrinsically safe in every state. Instead, the action-projection layer transforms infeasible candidate commands into feasible executed actions

before implementation. The filter was active in 85.09% of the evaluated steps, with a mean correction magnitude of 2.19 kW.

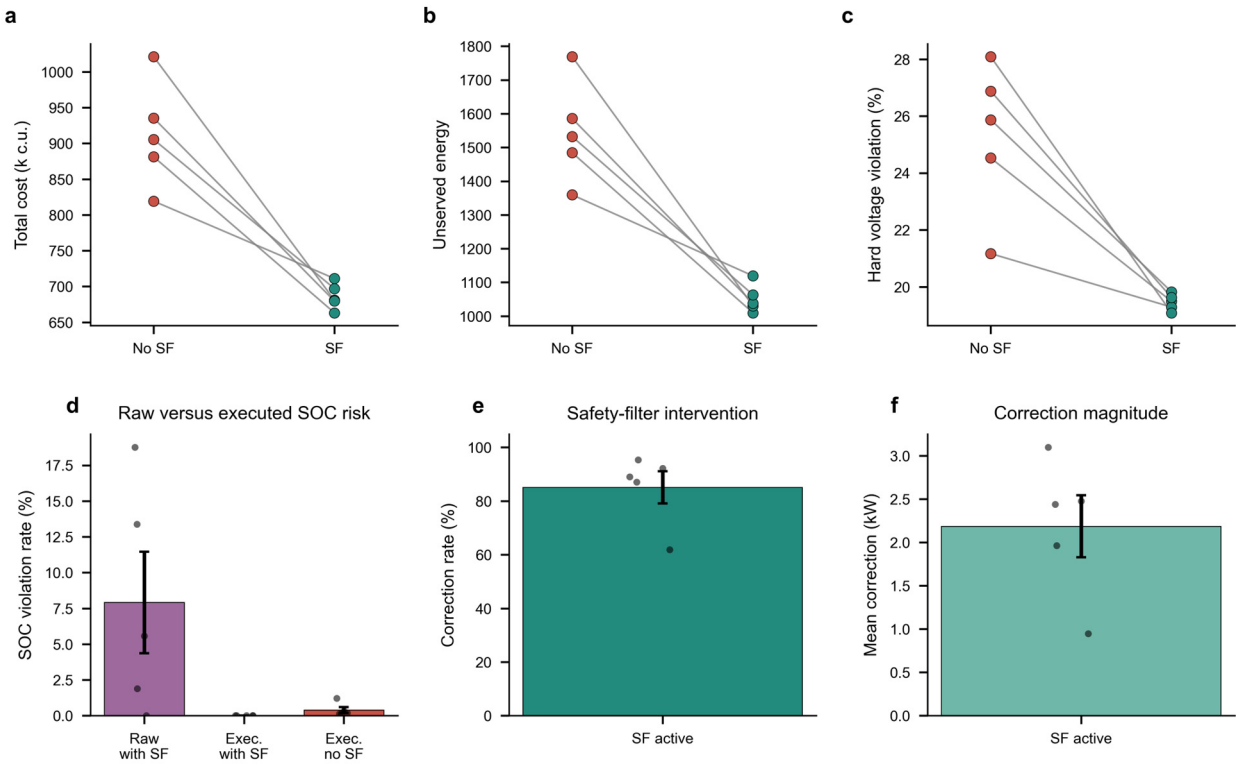


Figure 5. Safety-filter ablation. (**a**) Total operating cost; (**b**) Unserved energy; (**c**) Hard-voltage violation rate; (**d**) Raw and executed SoC-violation rates; (**e**) Intervention rate. (**f**) mean correction magnitude.

These results support the claim that the proposed framework improves executed-action feasibility through safety-filtered execution. The safety filter should therefore be interpreted as a model-based safety-screening mechanism rather than a formal certificate of closed-loop safety. The raw policy can still generate infeasible commands, while the executed command is corrected before implementation. Although the safety filter substantially reduces voltage and SoC risk, it does not eliminate all hard-voltage violations.

4.4. Representative Dispatch Behavior, Discussion, and Limitations

Figure 6 illustrates a representative held-out test-day dispatch profile. Compared with MPC-Lite and the no-safety hierarchical policy, the safety-filtered controllers use the BESS more actively during high-demand periods, reducing unserved energy while maintaining executed SoC feasibility. The proposed controller also shows frequent but bounded safety-filter intervention, indicating that the projection layer is an active execution-time safety mechanism rather than a negligible correction. The voltage envelope is mostly maintained within the hard operating band, with occasional excursions consistent with the aggregate voltage-violation statistics reported in Table 2.

Overall, the comparative results demonstrate that the proposed Hier. PPO-TD3 + SF framework achieves a favorable balance among economic performance, supply adequacy, operational feasibility, and online computational efficiency. Compared with MPC-Lite, the proposed controller reduces total operating cost and unserved energy while substantially shortening the online decision time. Compared with the hierarchical PPO-TD3 controller without the safety filter, the proposed method further improves both economic performance and executed-action feasibility, confirming the contribution of the safety-filtered projection

mechanism. Relative to the single-layer TD3 + SF controller, the proposed hierarchical structure achieves lower operating cost and unserved energy under the same safety-filtered execution setting. In addition, compared with the hierarchical PPO-DQN + SF variant, the proposed PPO-TD3 + SF framework provides slightly better economic and supply-adequacy performance while maintaining comparable voltage-related performance. These results indicate that the performance improvement is mainly associated with the coordinated contribution of hierarchical action decomposition and execution-time safety projection.

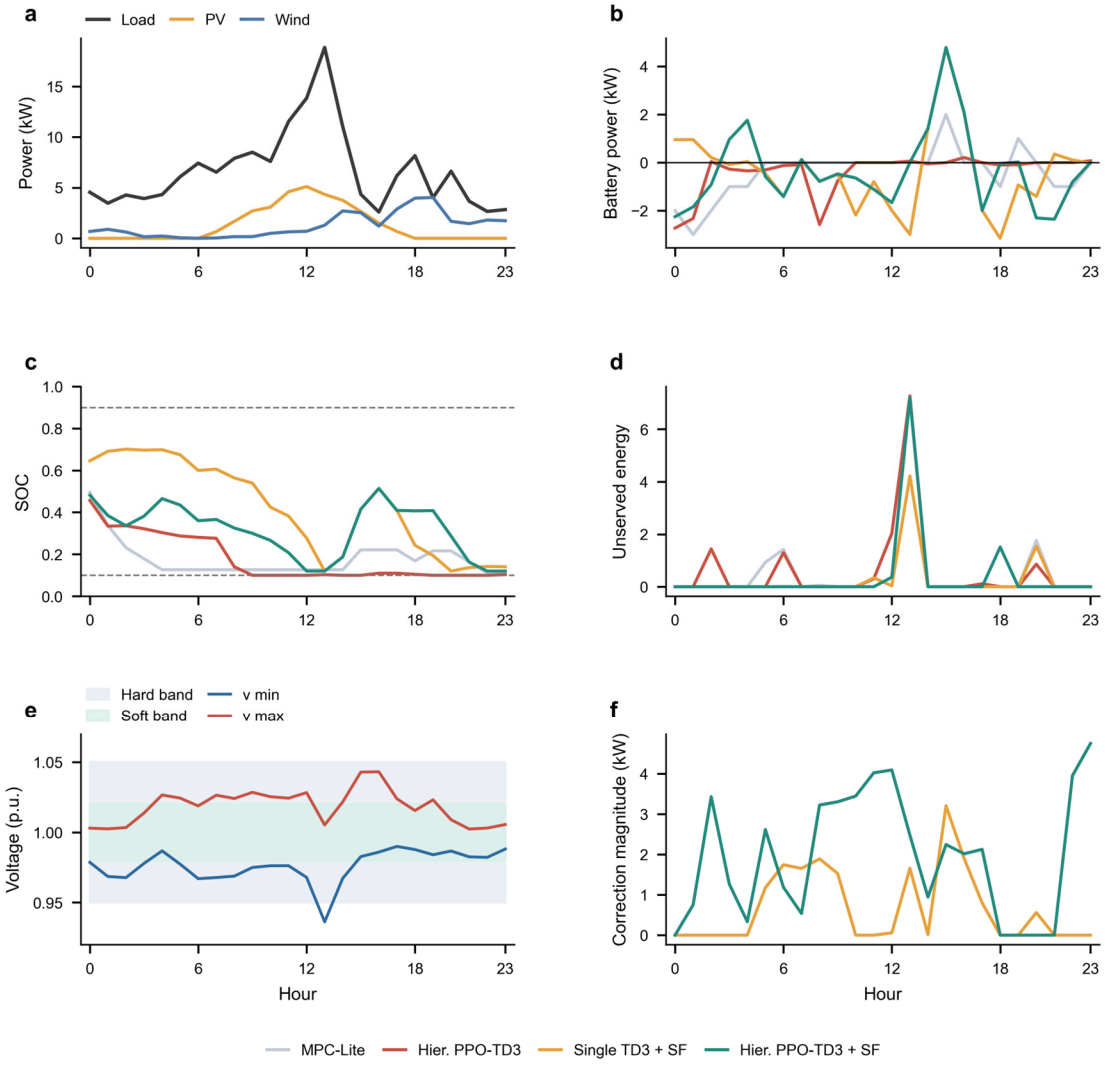


Figure 6. Representative held-out test-day controller behavior. (**a**) Load and renewable-generation profiles; (**b**) BESS power trajectories; (**c**) State-of-charge trajectories; (**d**) Unserved-energy profiles; (**e**) Voltage envelope. (**f**) Safety-filter correction magnitude.

5. Conclusions

This paper proposed a safety-filtered hierarchical PPO-TD3 framework for online energy management of renewable-rich microgrids under network constraints. The upper PPO layer selects coarse battery operating modes for long-horizon economic scheduling, while the lower TD3 layer generates bounded continuous battery-power refinements for short-term regulation. Before execution, the fused BESS command is corrected by a lightweight model-based action-projection layer considering device power limits, tightened SoC bounds, supply-priority rules, and voltage screening under a 33-bus radial-network surrogate.

Fixed-test simulations over five random seeds and 62 chronologically held-out test days show that the proposed controller improves the online safety–economy–computation trade-off under the adopted evaluation protocol. Relative to MPC-Lite, the proposed controller reduced total operating cost by 31.6%, unserved energy by 39.6%, and average decision time from 180.21 to 43.55 ms per step. The expanded comparison with hierarchical PPO–DQN + SF and single-layer TD3 + SF showed that the proposed PPO–TD3 + SF achieved the lowest mean cost and unserved energy among the evaluated controllers while maintaining zero executed SoC violations. The separate paired safety-filter ablation further demonstrates that the action-projection layer is an active contributor to reliable operation: it converts unsafe candidate commands into feasible executed actions, reduces hard-voltage violation exposure, and eliminates executed SoC violations in the evaluated cases. Taken together, the 62-day evaluation shows that the proposed PPO–TD3 hierarchy with safety-filtered execution offers an effective, computationally practical, and consistent framework for renewable-rich distribution-network dispatch.

From a sustainability perspective, the proposed framework supports the reliable and efficient operation of renewable-rich microgrids. The reductions in operating cost and unserved energy contribute to economic efficiency and supply reliability, while the safety-filtered execution mechanism improves operational feasibility under battery and network constraints. In addition, the coordinated management of renewable generation, battery storage, and dispatchable backup resources provides a practical basis for integrating variable renewable energy while maintaining operational security. Together, these improvements support operational resilience and provide a practical basis for more sustainable microgrid energy management.

Several limitations should be noted. First, the voltage model used in this study is a linearized radial-network surrogate rather than a full nonlinear AC power-flow solver. Therefore, the reported voltage metrics are interpreted as network-aware screening indicators under the adopted approximation. Second, although renewable-profile randomization is introduced during training to improve exposure to operating variability, the influence of explicitly modeled renewable forecast errors has not been separately isolated in the present evaluation. A more systematic analysis under different forecast-error levels would provide further insight into the robustness of the proposed controller. Third, the current numerical evaluation is conducted on a 33-bus radial feeder. The proposed framework itself is not restricted to this network size; however, larger networks may increase the computational effort associated with repeated voltage screening in the safety filter. Evaluating this computational trade-off on larger distribution systems is therefore an important extension. Finally, the battery degradation model is represented by a lightweight surrogate, and higher-fidelity battery and power-flow models can be incorporated in future studies.

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

Funding: This research was funded by the External Cooperation Program of Science and Technology Planning of Fujian Province. (Grant No. 2022I0015).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author; the data are not publicly available due to privacy restrictions.

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

Abbreviations

The following abbreviations are used in this manuscript:

AC	Alternating current
BESS	Battery energy storage system
DG	Diesel generator
DRL	Deep reinforcement learning
MILP	Mixed-integer linear programming
MPC	Model predictive control
DQN	Deep Q-network
PPO	Proximal policy optimization
PPO-TD3	Proximal policy optimization and twin delayed deep deterministic policy gradient
PV	Photovoltaic
SF	Safety filter
SoC	State of charge
TD3	Twin delayed deep deterministic policy gradient
WT	Wind turbine

References

1. Twaisan, K.; Barışçı, N. Integrated distributed energy resources (DER) and microgrids: Modeling and optimization of DERs. *Electronics* **2022**, *11*, 2816. [CrossRef]
2. Yang, T.; Zhang, L.; Zhen, L.; Liu, Y.; Song, Q.; Tang, W. Fast microgrids formation of distribution network with high penetration of DERs considering reliability. *Energy* **2021**, *236*, 121524. [CrossRef]
3. Ye, Y.; Wang, H.; Chen, P.; Tang, Y.; Strbac, G. Safe deep reinforcement learning for microgrid energy management in distribution networks with leveraged spatial–temporal perception. *IEEE Trans. Smart Grid* **2023**, *14*, 3759–3775. [CrossRef]
4. Hu, J.; Shan, Y.; Yang, Y.; Parisio, A.; Li, Y.; Amjady, N.; Islam, S.M.; Cheng, K.W.E.; Guerrero, J.M.; Rodriguez, J. Economic model predictive control for microgrid optimization: A review. *IEEE Trans. Smart Grid* **2024**, *15*, 472–484. [CrossRef]
5. Liu, G.; Ferrari, M.F.; Chen, Y. A mixed integer linear programming-based distributed energy management for networked microgrids considering network operational objectives and constraints. *IET Energy Syst. Integr.* **2023**, *5*, 320–337. [CrossRef]
6. Nguyen, Q.M.; Nguyen, D.L.; Nguyen, T.K. A mixed-integer linear programming model for microgrid optimal scheduling considering BESS degradation and RES uncertainty. *J. Energy Storage* **2024**, *104*, 114663. [CrossRef]
7. Zeyad, M.; Celik, B.; Hansen, T.M.; Locment, F.; Sechilariu, M. Comparative analysis of peer-to-peer energy trading with multi-objective optimization in rooftop photovoltaics-powered residential community. *Energies* **2026**, *19*, 1231. [CrossRef]
8. Bais, P.K.; Londhe, N.D.; Raj, R.; Sachdev, H.S. Faulty line localization in IEEE 30 bus system using CNN-LSTM. In Proceedings of the 2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), Mandya, India, 26–27 December 2022; pp. 1–5. [CrossRef]
9. Mohamed, S.A.; Anwer, N.; Mahmoud, M.M. Solving optimal power flow problem for IEEE-30 bus system using a developed particle swarm optimization method: Towards fuel cost minimization. *Int. J. Model. Simul.* **2025**, *45*, 307–320. [CrossRef]
10. Lee, S.; Seon, J.; Sun, Y.G.; Kim, S.H.; Kyeong, C.; Kim, D.I.; Kim, J.Y. Novel architecture of energy management systems based on deep reinforcement learning in microgrid. *IEEE Trans. Smart Grid* **2024**, *15*, 1646–1658. [CrossRef]
11. Guo, C.; Wang, X.; Zheng, Y.; Zhang, F. Real-time optimal energy management of microgrid with uncertainties based on deep reinforcement learning. *Energy* **2022**, *238*, 121873. [CrossRef]
12. Domínguez-Barbero, D.; García-González, J.; Sanz-Bobi, M.Á.; García-Cerrada, A. Energy management of a microgrid considering nonlinear losses in batteries through deep reinforcement learning. *Appl. Energy* **2024**, *368*, 123435. [CrossRef]
13. Hao, G.; Li, Y.; Li, Y.; Jiang, L.; Zeng, Z. Lyapunov-based safe reinforcement learning for microgrid energy management. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 9985–9999. [CrossRef] [PubMed]
14. Wang, Y.; Qiu, D.; Sun, M.; Strbac, G.; Gao, Z. Secure energy management of multi-energy microgrid: A physical-informed safe reinforcement learning approach. *Appl. Energy* **2023**, *335*, 120759. [CrossRef]
15. Yang, J.; Du, S.; Chen, P.; Liu, S.; Chen, B. Physics-shielded deep reinforcement learning for safe energy management of microgrids with battery health consideration. *J. Frankl. Inst.* **2025**, *362*, 108112. [CrossRef]
16. Xia, Y.; Xu, Y.; Feng, X. Hierarchical coordination of networked-microgrids toward decentralized operation: A safe deep reinforcement learning method. *IEEE Trans. Sustain. Energy* **2024**, *15*, 1981–1993. [CrossRef]

17. Xu, X.; Xu, K.; Zeng, Z.; Tang, J.; He, Y.; Shi, G.; Zhang, T. Collaborative optimization of multi-energy multi-microgrid system: A hierarchical trust-region multi-agent reinforcement learning approach. *Appl. Energy* **2024**, *375*, 123923. [CrossRef]
18. Safavi, V.; Mohammadi Vaniar, A.; Bazmohammadi, N.; Vasquez, J.C.; Keysan, O.; Guerrero, J.M. A battery degradation-aware energy management system for agricultural microgrids. *J. Energy Storage* **2025**, *108*, 115059. [CrossRef]
19. Yang, T.; Guo, Y.; Deng, L.; Sun, H.; Wu, W. A linear branch flow model for radial distribution networks and its application to reactive power optimization and network reconfiguration. *IEEE Trans. Smart Grid* **2021**, *12*, 2027–2036. [CrossRef]
20. Meng, W.; Song, D.; Deng, X.; Dong, M.; Yang, J.; Rizk-Allah, R.M.; Snášel, V. Dynamic optimal power flow of active distribution network based on LSOCR and its application scenarios. *Electronics* **2023**, *12*, 1530. [CrossRef]
21. Zhao, T.; Liu, X.; Wang, P.; Blaabjerg, F. More efficient energy management for networked hybrid AC/DC microgrids with multivariable nonlinear conversion losses. *IEEE Syst. J.* **2023**, *17*, 3212–3223. [CrossRef]
22. Sun, S.; Guo, W.; Wang, Q.; Tao, P.; Li, G.; Zhao, Z. Optimal scheduling of microgrids considering real power losses of grid-connected microgrid systems. *Front. Energy Res.* **2024**, *11*, 1324232. [CrossRef]
23. Hu, B.; Gong, Y.; Liang, X.; Chung, C.Y.; Noble, B.F.; Poelzer, G. Safe deep reinforcement learning-based real-time multi-energy management in combined heat and power microgrids. *IEEE Access* **2024**, *12*, 193581–193593. [CrossRef]
24. Hutsebaut-Buysse, M.; Mets, K.; Latré, S. Hierarchical reinforcement learning: A survey and open research challenges. *Mach. Learn. Knowl. Extr.* **2022**, *4*, 172–221. [CrossRef]
25. Wang, X.; Wang, S.; Liang, X.; Zhao, D.; Huang, J.; Xu, X.; Dai, B.; Miao, Q. Deep reinforcement learning: A survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 5064–5078. [CrossRef] [PubMed]
26. Wachi, A.; Shen, X.; Sui, Y. A survey of constraint formulations in safe reinforcement learning. In Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, Jeju, Republic of Korea, 3–9 August 2024; pp. 8262–8271. [CrossRef] [PubMed]
27. Lim, S.; Lee, J.; Lee, S. Model predictive control-based energy management system for cooperative optimization of grid-connected microgrids. *Energies* **2025**, *18*, 1696. [CrossRef]

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