Journal of Computer Science and Electrical Engineering
Print ISSN: 2663-1938
Online ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3148

INTELLIGENT SCHEDULING OF PV–STORAGE–CHARGING INTEGRATED STATIONS VIA GTRXL-PPO WITH CURRICULUM LEARNING

SongCheng Lu
School of Electrical Engineering and Automation, Hefei University of Technology, Xuancheng 242000, Hefei, China.

Abstract: To address the long-horizon sequential decision-making task, characterized by complex temporal dependencies, non-stationary dynamics, and high stochasticity in distribution-level PV–storage–charging systems, this paper develops a deep reinforcement learning framework that combines Gated Transformer-XL (GTrXL) with Proximal Policy Optimization (PPO). Cross-segment memory captures long-range temporal dependencies and time-aware encodings reinforce intraday periodicity. Training adopts a five-stage curriculum with adaptive KL control and auxiliary multi-task heads to improve sample efficiency. A group-normalized, potential-based reward unifies economic performance, grid friendliness, and storage health. In simulation, the agent learns a structured six-phase daily policy and achieves a 94.4% charging completion rate and 92.8% PV utilization, reduces average daily electricity purchase cost by 15%, and keeps grid peak power within a 60 kW soft limit. Across five seeds, returns improve by 78.4% over a feed-forward PPO baseline and by 23.6% over a vanilla GTrXL-PPO, demonstrating the benefits of long-memory RL for coordinated PV–storage–charging operation. The framework enforces feasibility via continuous action mapping with ramp-rate/jerk constraints and supports millisecond-level inference. Uncertainty-aware shaping improves robustness; gains are statistically significant across five seeds via paired tests.
Keywords: Long range temporal dependency decision making; PV–storage–charging integration; Curriculum learning; Gated Transformer XL; Proximal Policy Optimization

1 INTRODUCTION

The operational management of distribution-level PV–storage–charging systems is a complex sequential decision-making problem under profound uncertainty [1,2]. Control must contend with non-stationary inputs (renewables and EV arrivals), partially observable states, and a dynamically changing cost structure induced by time-of-use tariffs. Tradi- tional forecast-then-optimize pipelines are tightly coupled to forecast fidelity and solver latency; under model mismatch and rolling re-optimization, compounding errors and decision delays degrade performance and can jeopardize feasibility [3,4]. These challenges motivate an algorithmic approach that emphasizes long-range memory, robustness to partial observability, and feasibility-aware continuous control. By contrast, deep reinforcement learning (DRL) learns near optimal policies directly through interaction with the environment; its end to end learning capability and adaptability to uncertainty offer a new pathway to this problem [5,6]. In particular, Proximal Policy Optimization (PPO) achieves a favorable balance between training stability and sample efficiency via constrained policy updates, and has become a mainstream baseline for continuous control tasks [7].

With respect to optimization for integrated PV–storage–charging systems, relevant research has been reported in both Chinese and international literature In residential/home energy management and vehicle–grid coordination scenarios, PPObased energy management has demonstrated potential toreduce operating cost and increase renewable utilization; however, the incorporation of distribution network voltage and power constraints remains insufficient, making it difficult to directly ensure grid friendly operation [8]. In industrial microgrids, PPO augmented with forecasts can adapt to complex intraday profiles and price spreads, but it still struggles to model long time scale dependencies and cross segment credit assignment [9]. For grid friendly charging control, reinforcement learning shows promise in mitigating limit violations and improving voltage quality, yet policy robustness and action smoothness under high uncertainty remain to be strengthened [10]. In real world fleet charging contexts, coupling RL with on site renewables can deliver higher over-all benefits and stronger adaptability [11]. Nonetheless, existing methods still fall short in long horizon decision modeling, multi objective coordination, and enforcement of engineering constraints, leaving a gap to the complex requirements of practical PV–storage–charging deployments. Achieving an industry-ready balance grade balance among economic performance, grid friendliness, and real time implementability remains a central challenge [12].

Enabling all day and cross day load shifting and storage assisted renewable absorption require effective representation of long range temporal dependencies and principled credit assignment—both are key technical- bottlenecks for coordinated PV–storage–charging optimization. While recurrent neural networks (e.g., LSTM/GRU) can process sequences, they suffer from vanishing gradients and memory decay when capturing far future information [13]. In comparison, the Transformer architecture alleviates long range dependency modeling via self attention, but the vanilla Transformer incurs quadratic complexity on very long sequences [14]. Transform- er XL extends the effective context via segment level memory reuse and relative positional encoding, mitigating sequence “fragmentation” and providing a structural foundation

for capturing cross valley/peak and cross day dependencies [15]. In reinforcement learning, the Gated Transformer-XL (GTrXL) further improves numerical stability through gating mechanisms and outperforms recurrent networks on complex sequential tasks [16]. Safe/Constraint reinforcement learning provides an effective framework to embed grid interconnection constraints and feasible set mappings into policy learning [17]. However, most existing applications of GTrXL target relatively simple settings, without fully accounting for the complex constraints of PV–storage–charging systems or exploring its long memory capabilities for cross day coordinated decision making [18].

Despite progress toward engineering-realistic settings, key gaps remain: short-memory architectures struggle to capture cross-day and valley–peak long-range dependencies; robustness and action smoothness under high uncertainty are still inadequate; and real-time deployment often lacks unified designs for low-latency inference and feasibility-aware mappings (e.g., ramp-rate/jerk constraints and power soft limits). To address these issues, this paper proposes an intelligent dispatch framework for PV–storage–charging integration that couples GTrXL-based long-range memory with PPO-based policy optimization. Our solution is summarized as follows:

(1) Long range dependency capture and improved credit assignment. A GTrXL-PPO architecture with cross segment memory and temporal embeddings is used to capture long horizon dependencies and enhance credit assignment quality.

(2) Stable training and physically feasible control. A five stage curriculum learning scheme with adaptive KL regularization stabilizes training; continuous action mapping together with ramp rate (and jerk) constraints ensures smooth policy execution within the physical feasible set.

(3) Physically interpret-able multi objective reward. This paper constructs a multi objective reward that is physically interpret-able, unifying economic performance, grid friendliness, and storage health via group wise normalization and potential based shaping, and augments learning with exogenous auxiliary tasks to improve sample efficiency.

Through integrated application of the above techniques, the proposed method attains millisecond level inference latency while jointly optimizing economic efficiency, reliability, and real time performance, offering a solution with both theoretical novelty and engineering practicality for intelligent dispatch of PV–storage–charging systems. Experimental results verify important improvements over existing methods across multiple key performance indicators, demonstrating the potential of long memory reinforcement learning for optimizing complex energy systems.

2 GTRXL PPO FRAMEWORK AND TRAINING MECHANISMS

2.1 Clipped Surrogate Objective and GAE

This paper formulates the operational optimization of the inter-grated PV–storage–charging station as a finite horizon Markov Decision Process (MDP) and adopts Proximal Policy Optimization (PPO) as the core learning framework. PPO maintains training stability by restricting the magnitude of policy updates; conceptually, it introduces a trust region inspired constraint into policy gradient optimization. The optimization objective is to maximize the discounted cumulative return:

$$J(\theta)=\mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1}\gamma^t r_t\right] \tag{1}$$

Where τ denotes a complete trajectory from start to terminal; r_t is the immediate reward at step t, sourced from multiple dimensions including plant economics, service quality, grid friendliness, and asset health; $\gamma\in(0,1)$ is the discount factor, which emphasizes near term decisions and suppresses the adverse effect of long horizon noise on training stability; T is the number of steps per episode (determined jointly by the 24 hour horizon and the discretization step)

On this basis, PPO constructs a clipped surrogate objective. The truncated (clipped) importance sampling ratio is defined as:

$$\rho_t^{\text{clip}}(\theta)=\text{clip}(\rho_t(\theta),1-\ ,1+\ ) \tag{2}$$

PPO then uses the clipped surrogate objective to constrain policy updates:

$$\mathcal{L}^{\text{clip}}(\theta)=\mathbb{E}_t\left[\min\left(\rho_t(\theta)\hat{A}_t,\text{clip}(\rho_t(\theta),1-\ ,1+\ )\hat{A}_t\right)\right] \tag{3}$$

$$\rho_t(\theta)=\frac{\pi_\theta(\mathbf{a}_t\mid\mathbf{s}_t)}{\pi_{\theta_{\text{old}}}(\mathbf{a}_t\mid\mathbf{s}_t)} \tag{4}$$

Here ϵ is the clip half width, and $\hat{A}_t$ is the advantage estimate that evaluates how good action $\mathbf{a}_t$ is in state $\mathbf{s}_t$ relative to a baseline. The advantage is estimated using Generalized Advantage Estimation (GAE), which trades off bias and variance:

$$\begin{cases}\delta_t=r_t+\gamma V_\phi(\mathbf{s}_{t+1})-V_\phi(\mathbf{s}_t)\\ \hat{A}_t=\sum_{l=0}^{L_t-1}(\gamma\lambda)^l\delta_{t+l}\end{cases} \tag{5}$$

$\lambda\in[0,1]$ controls the bias–variance trade- off.

Policy and value are jointly optimized, giving the overall training objective:

$$\begin{cases}\mathcal{J}(\theta,\phi)=-\mathcal{L}^{\text{clip}}(\theta)+c_v\mathcal{L}^{\text{VF}}(\phi)-c_e S(\pi_\theta)\\ \mathcal{L}^{\text{VF}}(\phi)=\mathbb{E}\left[(V_\phi(\mathbf{s}_t)-\hat{R}_t)^2\right]\\ S(\pi_\theta)=\mathbb{E}\left[\mathcal{H}(\pi_\theta(\cdot\mid\mathbf{s}_t))\right]\end{cases} \tag{6}$$

where c_v and c_e weight the value function regression and entropy regularization, respectively; $\mathcal{H}(\cdot)$ denotes the policy entropy, encouraging exploration in early training and reducing overfitting.

In implementation, this paper introduces adaptive KL divergence control, dynamically tuning the optimization strength according to $\mathcal{D}_{\mathrm{KL}}[\pi_{\theta_{\mathrm{old}}} \| \pi_\theta]$. The action space is continuous; a hyperbolic tangent squashing maps the network outputs to the physical feasible set $[a_{min}, a_{max}]$, ensuring that decision variables such as storage power and charging allocation satisfy engineering constraints.

2.2 Gated GTrXL Long Range Memory Mechanism

In this task, the load, electricity price, and PV sequences exhibit pronounced intraday and cross segment dependencies; a single step Markovian assumption is insufficient to capture the system dynamics. This paper therefore employs a GTrXL backbone with segment level memory and adopts gating to enhance numerical stability in the reinforcement learning setting. Given an input sequence $X \in \mathbb{R}^{L \times d_{\mathrm{model}}}$, the scaled dot- product attention is:

$$\begin{cases} Q = XW_Q \\ K = XW_K \\ V = XW_V \\ \mathrm{Att}(Q,K,V) = \mathrm{Softmax}\left(\dfrac{QK^\top}{\sqrt{d_k}}\right)V \end{cases} \tag{7}$$

Multi head attention aggregates independent subspace attentions in parallel and applies a linear projection.

$$\mathrm{MHA}(X) = \mathrm{Concat}\left(\mathrm{Att}(Q_i,K_i,V_i)\right)_{i=1}^{h} W_O \tag{8}$$

And the specific structure is shown in Figure 1:

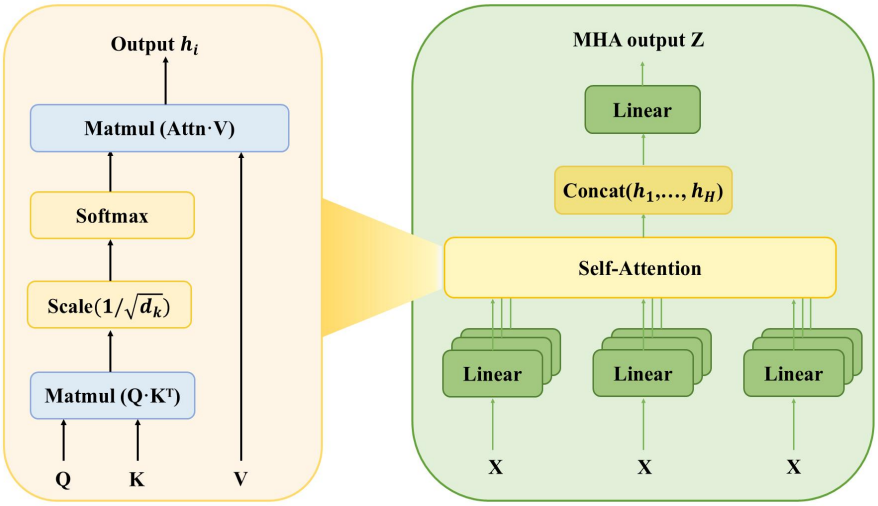

Figure 1 Scaled Dot-Product Attention with Linear Q/K/V Projections

A position wise feed forward network then performs the nonlinear transformation:

$$\begin{cases} \mathrm{FFN}(x) = W_2\sigma(W_1x + b_1) + b_2 \\ \sigma = \mathrm{ReLU} \end{cases} \tag{9}$$

To encode intraday periodicity, the raw observation is linearly projected and added to an absolute positional encoding and a time embedding vector:

$$X^{(0)} = \mathrm{Proj}([\text{raw obs}]) + \mathrm{PE}_{\mathrm{abs}} + \mathrm{TE}_{\mathrm{time}} \tag{10}$$

where TE_{time} includes embeddings for hour, weekday, and month, explicitly encoding temporal context. GTrXL caches hidden state memory across segments. Let $M_{t-1} \in \mathbb{R}^{L_m \times d_{\mathrm{model}}}$ denote the tail memory from the previous segment and X_t the current segment input; for each layer, keys and values are extended as:

$$\begin{cases} \tilde{K} = [M_{t-1}; X_t]W_K \\ \tilde{V} = [M_{t-1}; X_t]W_V \\ \mathrm{Att}(Q,\tilde{K},\tilde{V}) = \mathrm{Softmax}\left(\dfrac{Q\tilde{K}^\top}{\sqrt{d_k}}\right)\tilde{V} \end{cases} \tag{11}$$

The memory is updated using a stop gradient tail buffer $M_t = \mathrm{SG}(\mathrm{Tail}(H_t, L_m))$. In a vectorized, parallel environment, memories are maintained independently per environment index and cleared via masking at the end of each episode to prevent cross trajectory and cross environment leakage. To further improve training stability, this paper adopts GRU style gating in GTrXL. Given the layer input $\mathbf{x}$ and a branch output $\mathbf{y}$:

$$\begin{cases} \mathbf{r} = \sigma(W_r \mathbf{y} + U_r \mathbf{x}) \\ \mathbf{z} = \sigma(W_z \mathbf{y} + U_z \mathbf{x} + \mathbf{b}_z) \\ \mathbf{h} = \tanh(W_g \mathbf{y} + U_g (\mathbf{r} \circ \mathbf{x})) \\ \mathrm{GRUGate}(\mathbf{x}, \mathbf{y}) = (1 - \mathbf{z}) \odot \mathbf{x} + \mathbf{z} \odot \mathbf{h} \end{cases} \tag{12}$$

The layers are stacked in the order "LN→MHA→ GRUgate→LN→FFN→GRUgate", which mitigates diver- gence risk along residual paths in deep networks.

2.3 Multi Task Learning and Exogenous Information Fusion

Given that exogenous variables—such as electricity price trend, PV intensity, and evening peak risk—substantially influence policy decisions yet are not in one to one correspondence with the main task return, This paper attaches three auxiliary heads to the shared representation $\mathbf{f}_t \in \mathbb{R}^{D_f}$: a price trend three class classifier, a future PV regressor, and a peak risk binary classifier, denoted $L_1 = \mathrm{CE}(\hat{\mathbf{y}}^{\mathrm{price}}, \mathbf{y}^{\mathrm{price}})$, $L_2 = \mathrm{MSE}(\hat{\mathbf{y}}^{\mathrm{pv}}, \mathbf{y}^{\mathrm{pv}})$, and $L_3 = \mathrm{CE}(\hat{\mathbf{y}}^{\mathrm{risk}}, \mathbf{y}^{\mathrm{risk}})$, respectively, and defines:

$$\mathcal{L}_{\mathrm{aux}} = L_1 + \lambda_{\mathrm{pv}} L_2 + L_3 \tag{13}$$

During training, this paper applies $\mathbf{f_t}$ to detach and mini- mizes $\mathcal{L}_{aux}$ with an independent optimizer, preventing gradients from back propagating into the policy/value backbone and thereby avoiding inappropriate leakage of supervised signals. This stabilizes the learning of long term exogenous structure without distorting the optimization direction of the main task.

2.4 Progressive Curriculum Learning Strategy

To reduce cold-start difficulty on complex tasks, this paper adopts a staged curriculum that tightens controllable environmental factors from easy to hard. Let the progress scalar $p \in [0,1]$ (e.g. $p = t / T_{\mathrm{train}}$); a monotone schedule $\mathcal{D}(p)$ contracts $\phi(p) = \{\text{fleet size}, P_{\max}^{\mathrm{soft}}, \lambda_{\mathrm{peak}}, \sigma_{\mathrm{price}}\}$ across K stages. At stage *i*, this paper runs L_i episodes, then rebuilds the environment with updated ϕ_{i+1} while carrying over policy/critic parameters and retaining VecNormalize statistics (e.g. obs_{rms}, ret_{rms}), which align train–eval distributions. Each method fits its own normalizers on its training rollouts and freezes them during evaluation to ensure fair, stable comparison. Stage transitions trigger when *p* crosses thresholds (or a moving-average return), yielding steadily increasing difficulty without destabilizing training.

3 PV–STORAGE–CHARGING COORDINATED SCHEDULING MODELING AND ALGORITHM DESIGN

3.1 State Space Modeling and Dynamic Constraints of the PV–Storage–Charging System

Discretize the 24- hour horizon with step size Δt into $T = 24 / \Delta t$ steps. At each step, the observation is $\mathbf{s}_t \in \mathbb{R}^{D_s}$, and the policy outputs the action $\mathbf{a}_t = [\alpha_t, \beta_t, \gamma_t]$. For interpretability, the three components are understood as: a storage command ratio (normalized by P_{max}^{ess}), a grid- purchase willingness coefficient, and the allocation preference between PV→EV and PV→ESS. The station- level power balance is:

$$P_t^{\mathrm{grid}} = P^{\mathrm{base}} + P_t^{\mathrm{ev,grid}} + P_t^{\mathrm{ess,grid}} - P_t^{\mathrm{pv}} \tag{14}$$

Here, P^{base} denotes the base load. The aggregate EV charging power is constrained jointly by per pile and station level limits:

$$P_t^{\mathrm{ev}} = \min\left(\sum_{i \in A_t} \min\left(\frac{R_{i,t}}{\tau_{i,t}}, P_{\max}^{\mathrm{single}} \right), P_{\max}^{\mathrm{ev}} \right) \tag{15}$$

A_t denotes the set of vehicles currently in the charging station; $R_{i,t}$ and $\tau_{i,t}$ denote vehicle *i* remaining energy demand and remaining dwell time, respectively. The storage energy evolves with asymmetric efficiency:

$$E_{t+1} = \Pi_{[0, C^{\mathrm{ess}}]} \left(E_t + \eta_c P_t^{+} \Delta t - \frac{1}{\eta_d} P_t^{-} \Delta t \right), \tag{16}$$

Here, $P_t^{+} = \max(P_t^{\mathrm{ess}}, 0)$; $P_t^{-} = \max(-P_t^{\mathrm{ess}}, 0)$; Π is the inter- val projection ensuring $E_t \in [0, C^{\mathrm{ess}}]$. In addition, this paper requires $E_t \in [\underline{\alpha}, \overline{\alpha}] \cdot C^{\mathrm{ess}}$ to protect battery lifetime. To encode prior "valley replenishment-daytime locking-evening- peak discharge", this paper introduces:

$$E_t^{\star} \in \{\alpha_{\mathrm{prep}}, \alpha_{\mathrm{morn}}, \alpha_{\mathrm{solar}}, \alpha_{\mathrm{tran}}, \alpha_{\mathrm{eve}}, \alpha_{\mathrm{rech}}\} \cdot C^{\mathrm{ess}} \tag{17}$$

The look ahead control for grid friendliness is quantified by a peak risk indicator over a window H , normalized to $[0,1]$:

$$\begin{cases} \rho_t^{\text{peak}} = \dfrac{1}{C_\rho} \max_{1\le k\le H} \{S_{t+k}\} \\ S_{t+k} = w(h_{t+k}) \cdot g(p_{t+k}, v_{t+k}, \ell_{t+k}) \end{cases} \tag{18}$$

In this expression, h is the time segment index; p, v, ℓ denote the future price, PV, and net load, respectively; $w(\cdot)$ emphasizes evening peaks; $g(\cdot)$ increases under "high price / low PV / high net load"; and C_ρ is a normalization constant.

3.2 Policy Network Architecture Based on GTrXL

A gated GTrXL frontend serves as the core feature extractor of the policy network. The GTrXL frontend projects the observation to d_{model} ; the projected features are added with PE_{abs}, TE_{time} , and then L layers of "LN→MHA→GRUgate→LN→FFN→GRUgate are stacked, the specific structure is shown in Figure 2. In addition, the cross step memory $\mathbf{m}_{t-1} \in \mathbb{R}^{L_m \times d\text{model}}$ is maintained independently in parallel environments; when any environment's episode terminates, the system precisely zeros the corresponding memory via a Boolean mask, strictly blocking cross trajectory information leakage. This mechanism is implemented via a custom environment wrapper that updates the memory mask synchronously whenever a done signal is returned.

Training follows the PPO configuration in Section II-A. The policy output is passed through a hyperbolic tangent activation and mapped to the continuous action space $[-1,1]^3$, corresponding to the storage power ratio, grid exchange willingness, and PV allocation preference, respectively. The value network shares the GTrXL feature extractor with the policy; the final two layers use separate MLP heads to avoid interference between policy and value gradients. Advantages are estimated via GAE to balance bias and variance. On the shared features, an exogenous "self supervised" loss is minimized, with detach($\mathbf{f}_t$) preventing back propagation into the backbone to avoid unintended leakage of supervised signals.

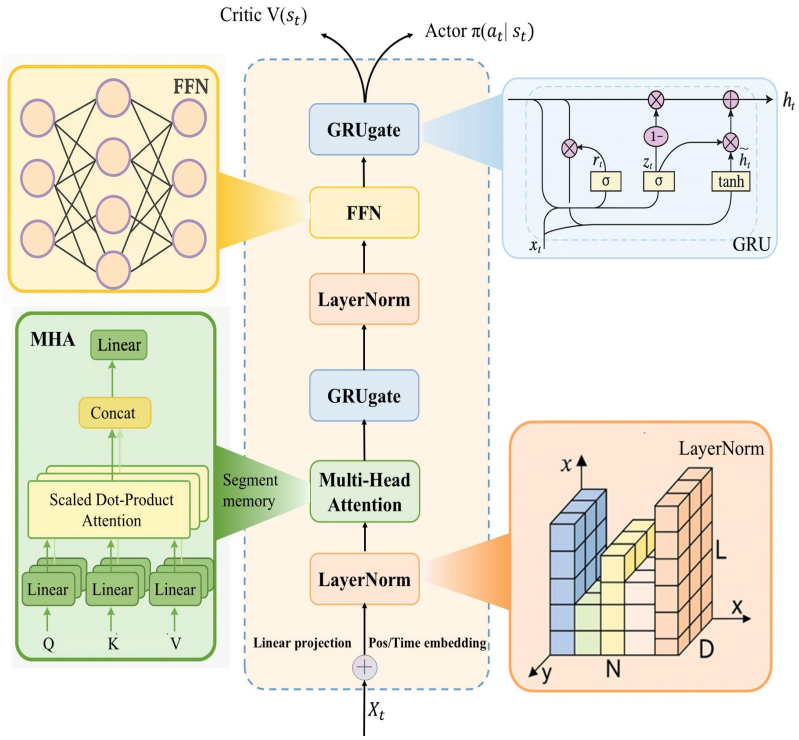

Figure 2 GTrXL Policy Network with Segment-Level Memory

3.3 Continuous Action Space Mapping and Enforcement of Engineering Constraints

Neural network actions must be mapped to executable power commands that satisfy continuity and ramping constraints. Define the storage command as $P_t^{css,cmd}=\alpha_t P_{max}^{css}$, with $\omega_{ev}=\gamma_t$ and $\omega_{ess}=1-\gamma_t$. To suppress high frequency jitter and micro cycling, apply a moving average followed by exponential smoothing:

$$\begin{cases} \tilde{\mathbf{a}}_t = \dfrac{1}{W} \sum_{i=0}^{W-1} \mathbf{a}_{t-i}, \\ \bar{\mathbf{a}}_t = (1-\omega)\tilde{\mathbf{a}}_t + \omega \bar{\mathbf{a}}_{t-1}. \end{cases} \tag{19}$$

Then impose inter-period ramp-rate constraints and a second-order rate-of-change (jerk) penalty. Here, "jerk" is the second derivative of the power set-point; penalizing it suppresses abrupt ramp acceleration, yielding smoother control and reduced converter/battery stress.

$$\begin{cases} |\ P_t^{\text{ess,cmd}} - P_{t-1}^{\text{ess}}\ | \le r(h_t) P_{\max}^{\text{ess}}, \\ \text{jerk}_t = \dfrac{|\ P_t^{\text{ess}} - 2P_{t-1}^{\text{ess}} + P_{t-2}^{\text{ess}}\ |}{P_{\max}^{\text{ess}}} \end{cases} \tag{20}$$

Simultaneously record the PV surplus power:

$$P_t^{\text{pv,ex}} = \max(0, P_t^{\text{pv}} - P^{\text{base}}) \tag{21}$$

This mapping guarantees closure of the feasible set: for any t, $P_t^{css,cmd} \in [-P_{max}^{ess}, P_{max}^{ess}]$, and the step to step variation is bounded by $r(h_t)$, facilitating subsequent feasible dispatch.

3.4 Multi Objective Reward Function Design and Normalization Strategy

To integrate multi- objective engineering constraints into a scalar reward suitable for RL—while ensuring objectivity and reproducibility—this paper decomposes the total return into physically interpretable components, each normalized to [-1,1] or [0,1], and then combine them linearly with non- negative weights. This normalization makes heterogeneous terms comparable and caps outliers, while preserving the direction of improvement:

$$\begin{aligned} r_t &= \lambda_{\text{econ}} \tilde{r}_t^{\text{econ}} + \lambda_{\text{serv}} \tilde{r}_t^{\text{serv}} + \lambda_{\text{grid}} \tilde{r}_t^{\text{grid}} \\ &+ \lambda_{\text{ess}} \tilde{r}_t^{\text{ess}} + \lambda_{\text{pv}} \tilde{r}_t^{\text{pv}} + \lambda_{\text{act}} \tilde{r}_t^{\text{act}} + \lambda_{\text{uncert}} \tilde{r}_t^{\text{uncert}}. \end{aligned} \tag{22}$$

With $\text{clip}_{[-1,1]}(x) = \max(-1, \min(1, x))$. For numerical stability, a small constant $\varepsilon > 0$ is added to all denominators.

(1) Economics (econ): electricity purchase cost and feed in at a discounted tariff:

$$\begin{cases} r_t^{\text{econ}} = -\xi_t p_t (P_t^{\text{grid}})^+ \Delta t + \eta_s p_t (-P_t^{\text{grid}})^+ \Delta t, \\ \tilde{r}_t^{\text{econ}} = \text{clip}_{[-1,1]}\left(\dfrac{r_t^{\text{econ}}}{B_{\text{econ}}} \right). \end{cases} \tag{23}$$

where ξ_t is the time slot weight, $\eta_s \in (0,1)$ is the export discount factor, and B_{econ} is a magnitude upper bound.

(2) Service completion (serv): instantaneous EV supply and "potential based" progress, with urgency increasing as tasks approach deadlines:

$$\begin{cases} E_t^{ev} = P_t^{ev} \Delta t, \\ \Phi_{prog}(q) = \kappa_{prog} \log(\varepsilon + q), \quad \text{where } q_t = 1 - \dfrac{\sum_i R_i}{\sum_i D_i}, \\ r_t^{serv} = \kappa_{ev} \dfrac{E_t^{ev}}{P_{max}^{ev} \Delta t} + \gamma \Phi_{prog}(q_{t+1}) - \Phi_{prog}(q_t), \\ \tilde{r}_t^{serv} = \text{clip}_{[-1,1]}\left(\dfrac{r_t^{serv}}{B_{serv}} \right). \end{cases} \tag{24}$$

The second term($\gamma \Phi_{prog}(q_{t+1}) - \Phi_{prog}(q_t)$) is an approximate potential based shaping function; it preserves the optimal policy partial order while mitigating training instability due to sparse rewards.

(3) Grid friendliness (grid): peak limit violations and grid side ramping:

$$\tilde{r}_t^{\text{grid}} = -\kappa_{\text{peak}} \frac{\left(P_t^{\text{grid}} - L_s^{\text{eff}}\right)^+}{P_{\text{scale}}} - \kappa_{\text{rate}} \frac{\left(\left|P_t^{\text{grid}} - P_{t-1}^{\text{grid}}\right| - P_{\text{step}}\right)^+}{P_{\text{scale}}} \tag{25}$$

(4) Storage health (ess): power magnitude, SOC deviation, smoothness, and second order jerk:

$$\tilde{r}_t^{\text{ess}} = -\kappa_e \frac{\left|P_t^{\text{ess}}\right|}{P_{\max}^{\text{ess}}} - \kappa_{\text{soc}} \frac{\left|E_t - E_t^*\right|}{C^{\text{ess}}} - \kappa_{\text{sm}} \frac{\left|P_t^{\text{ess}} - P_{t-1}^{\text{ess}}\right|}{P_{\max}^{\text{ess}}} - \kappa_{\text{jerk}} \text{jerk}_t. \tag{26}$$

Additionally, include a SOC potential based safeguard to red- uce bias:

$$\begin{cases} \Phi_{\text{soc}}(E,h) = -\kappa_{\text{soc}} \dfrac{\left|E - \alpha^{*}(h) C^{\text{ess}}\right|}{C^{\text{ess}}}, \\ r_{t,\text{shap}}^{\text{soc}} = \gamma \Phi_{\text{soc}}(E_{t+1}, h_{t+1}) - \Phi_{\text{soc}}(E_t, h_t). \end{cases} \tag{27}$$

Incorporating $r_{t,\text{shap}}^{\text{soc}}$ into the overall storage reward $\tilde{r}_t^{\text{ess}}$ provides a theoretical guarantee.

(5) 5. PV and time slot efficiency (pv): encourage local absorption and "charge more at low price / buy less at high price":

$$u_t^{\text{pv}} = \frac{E_t^{\text{pv, used}}}{\max(\varepsilon, P_t^{\text{pv}} \Delta t)} \tag{28}$$

$$\tilde{r}_t^{\text{pv}} = \kappa_{\text{pv}} u_t^{\text{pv}} \frac{P_t^{\text{pv}}}{P_{\text{ref}}} + \kappa_{\text{eff}} \left(\mathbf{1}_{\text{day-high-pv}} \frac{P_t^{\text{ev}}}{P_{\max^{\text{ev}}}} + \mathbf{1}_{\text{low-price}} \frac{P_t^{\text{ev}}}{2P_{\max}^{\text{ev}}} \right) \tag{29}$$

(6) Action quality (act): encourage smoothness and penalize large steps:

$$\tilde{r}_t^{\text{act}} = \kappa_{\text{sm-act}} \cdot \text{smoothness}(\mathbf{a}_t) - \kappa_{\text{rep}} \frac{|\mathbf{a}_t - \mathbf{a}_{t-1}|_1}{3} \tag{30}$$

(7) Uncertainty (uncert): encourage conservative decisions as forecast error rises:

$$\begin{cases} e_t^{\text{price}} = \dfrac{|\hat{p}_{t|t-H_u} - p_t|}{\sigma_p + \epsilon}, \\ e_t^{\text{pv}} = \dfrac{|\hat{v}_{t|t-H_u} - v_t|}{v_{\max} + \epsilon}, \\ \tilde{r}_t^{\text{uncert}} = \psi_p(e_t^{\text{price}}) + \psi_v(e_t^{\text{pv}}) + \psi_u(|\hat{u}_{t|t-H_u} - e_t^{\text{price}}|). \end{cases} \tag{31}$$

Where ψ is a monotone bounded function (e.g., piecewise linear/logarithmic) to avoid gradient explosion. At the end of an episode ($t = T-1$), KPI style step rewards/penalties are added (final SOC band, EV completion rate, maximum peak, whole day PV utilization, ESS cycle count, etc.), and normalized to [-1,1] by B_{final}. Through the combination of groupwise normalization and potential based terms, the overall reward remains physically interpretable and optimization stable.

3.5 Stage Wise Training Framework and Hyperparameter Configuration

Curriculum learning proceeds from easy to hard by gradually increasing fleet size, tightening soft grid limits, and strengthening evening-peak penalties. At each stage, we collect cross-day trajectories of length $n_{step}=k_{day}T$ and perform n_{epochs} minibatch updates; the entropy/value weigh- ts and the learning rate decay linearly with the progress scalar p. Stage transitions are triggered when p crosses preset thresholds (or a moving-average return criterion), yielding steadily higher difficulty without destabilizing training.

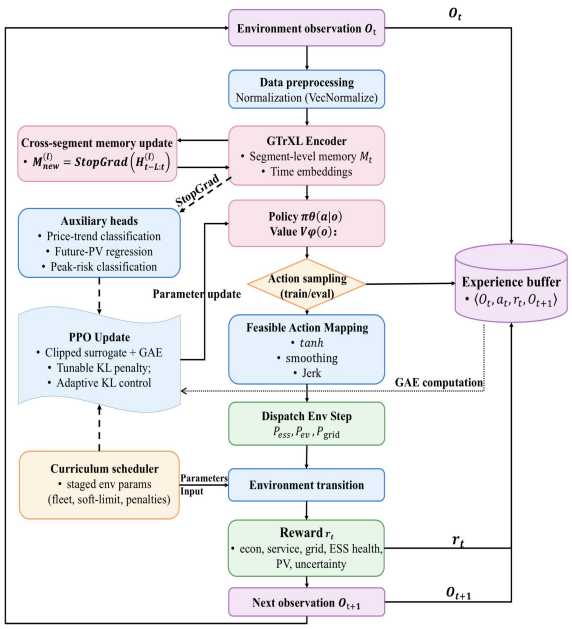

Figure 3 Overall Workflow of Cross-Segment Memory and Curriculum Learning

The coordinated PV–storage–charging scheduler follows the closed loop in Figure 3: observe O_t; a GTrXL frontend extracts long-range temporal features; the policy π_θ proposes an action, which is mapped through engineering constraints (smoothing and ramp-rate) to an executable command; the environment then transitions to O_{t+1} with reward r_t, and (O_t,a_t,r_t,O_{t+1}) updates the policy via GAE and PPO. Segment-level memory is maintained independently across parallel environments; when any episode terminates, a Boolean mask zeros the corresponding memory to strictly block cross-trajectory leakage.

Meanwhile, stage-wise annealing of the KL target, entropy weight, and learning rate damps gradient oscillations and reduces sensitivity to the clipping range, improving sample efficiency without sacrificing stability. Concretely, this paper applies linear (in-progress) schedules for the target KL, for the entropy bonus, and for the optimizer step size, so exploration is gradually restrained while the effective trust region tightens as stages grow harder. To curb train–eval distribution shift, observations and rewards are normalized with shared VecNormalize statistics: running means/variances are fit on training rollouts only, then frozen for evaluation so that scaling at test time exactly matches the training distribution; each compared method fits its own normalizer to avoid cross-method leakage. A custom callback monitors both control-side KPIs (peak-grid exceedances, EV completion ratio, SoC boundary hits, ramp/jerk costs) and learning diagnostics (KL divergence, clip fraction, entropy, value loss, explained variance); it co-tunes penalty multipliers when constraint violations persist, and triggers early stopping upon plateaued moving-average returns or sustained satisfaction of soft-limit targets. Together, these mechanisms stabilize PPO updates across curriculum stages and maintain feasibility as grid limits tighten.

4 CASE STUDY

4.1 Simulation Environment and Experimental Settings

Experiments are conducted on a Python based simulation platform for integrated PV–storage–charging. The environment uses a 5 minute time step, yielding 288 periods per day. The energy storage system (ESS) is configured with a capacity of 200 kWh, a power limit of 40 kW, round trip efficiency η=0.95,and an SOC operating range of [0.2,0.95]. The EV fleet comprises 20 vehicles with per vehicle energy demand uniformly distributed in [15, 35] kWh; each charger is limited to 20 kW, and the station level power cap is 200 kW. The time of use (TOU) tariff has a base rate of 0.4 CNY/kWh, with peak/valley multipliers of 1.8/0.4; the PV installation is 120 kW. The grid soft limit is progressively tightened from 100 kW to 60 kW across five curriculum learning stages. The detailed configuration is summarized in the table 1:

Table 1 Parameter Configuration for the Five-Stage Curriculum Learning

Stage	Training Progress Threshold	Number of EVs	Per-EV Demand (kWh)	Grid Soft Limit (kW)	Evening Peak Purchase Penalty Multiplier
0	≤25%	10	20-30	100	1.0
1	≤50%	15	18-32	85	1.5
2	≤70%	20	15-35	65	2.5
3	≤85%	20	12-38	65	3.0
4	≤100%	25	10-45	60	3.5

Training progress p is defined as the fraction of total training timesteps completed, $p=t/T$. The curriculum advances to the next stage once p exceeds the specified threshold.

(t = current timestep, T = total timesteps).

Algorithm configuration:This paper adopts PPO with a GTrXL policy network, integrating time embeddings,adaptive KL scheduling, auxiliary prediction heads (price/PV/peak), and an uncertainty aware reward. Training episodes span two consecutive days to learn cross day coordination patterns. The simulation platform builds on OpenAI Gym and implements deep models in PyTorch with parallelized environments to accelerate training.

4.2 Comprehensive Performance Evaluation and Comparative Analysis

4.2.1 Comparison of multi algorithm performance metrics

Results show that, for EV charging completion rate, baseline PPO reaches 91.8%, whereas introducing LSTM reduces performance to 90.4%, indicating that basic recurrent memory struggles with cross day scheduling. GTrXL+PPO improves completion to 92.9%, confirming the efficacy of attention in capturing long range dependencies; adding curriculum learning (GTrXL+PPO+CL) attains the best performance at 94.4%. PV utilization exhibits differentiated behavior: although LSTM+PPO underperforms on completion, it achieves 91.1% PV utilization, suggesting a bias toward immediate energy balancing; by contrast, GTrXL PPO+CL sustains a high completion rate while achieving 92.8% PV absorption, validating the synergy between long memory modeling and progressive training for multi objective optimization. See the performance bar chart in Figure 4.

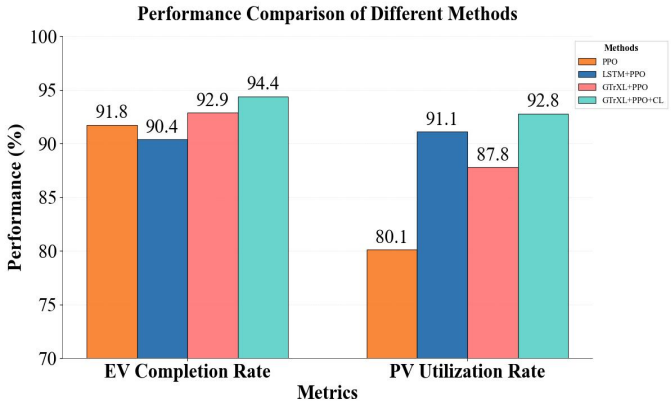

Figure 4 Comparison of Key Performance Metrics across Algorithm

4.2.2 Sample efficiency analysis

The growth rate of the cumulative return curve quantifies sample efficiency: GTrXL+PPO+CL exhibits the steepest ascent, with cumulative return at 1.5×10^6 steps approximate- ly 2.1× that of baseline PPO; GTrXL+PPO ranks second, and LSTM+PPO yields limited gains. Notably, stage wise environment reconstruction introduced by curriculum learning does not induce a cliff drop in cumulative return; instead, the progressive difficulty maintains stable learning gradients, attributable to the inheritance of VecNormalize statistics during environment resets. This observation is consistent with the staged reconstruction and statistic inheritance mechanism used during training. See Figure 5.

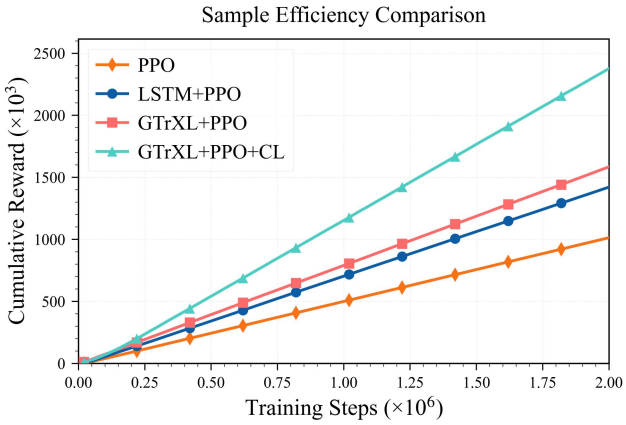

Figure 5 Convergence Efficiency Comparison

4.2.3 Training stability analysis

Figure 6 shows that convergence of average episode return reflects training stability. GTrXL+PPO+CL enters a stable region after 0.5×10^6 steps, with a plateau of [24000,26000] and the narrowest variance band (about ±1000).

Figure 6 Average Episode Return: Convergence and Variance

In contrast, baseline PPO shows pronounced periodic oscillations. LSTM+PPO rises quickly in early training (within 0.25×10^6 steps) but converges to a lower level (about 17 000) than the GTrXL architecture, likely due to vanishing gradients on long sequences, which hinder back propagation of long horizon returns. With auxiliary prediction heads and the uncertainty aware reward, policies behave more robustly at price/PV regime shifts; training callbacks also show the KL target self adjusts within [0.005,0.05] to keep policy updates stable.

4.2.4 Comparison with external engineering baselines

To benchmark against mainstream engineering approach- es, this paper includes: (1) MPC with a 15 minute prediction horizon and 5 minute control horizon under rolling optimization; and (2) a rule based peak–valley scheduler that executes pre set charge/discharge strategies based on TOU and PV forecasts.

Results (mean ± standard deviation over five random seeds) show that MPC can approach the theoretical optimum with perfect forecasts but incurs significant computational overhead (0.85 s/step), limiting real time deployment. Rule based control is mediocre in both performance and compute; although it has the highest throughput, it degrades markedly under complex constraints. In contrast, GTrXL+PPO+CL achieves the best overall performance with millisecond level inference, notably surpassing traditional methods in charging completion (94.4%). See Table 2.

Table 2 Comprehensive Performance Comparison of Different Methods

Method	EV Completion Rate (%)	PV Utilization Rate (%)	Peak Power (kW)	Comp. Time (s/step)
Rule-based	85.2±2.1	88.4±3.2	72.8±5.3	0.001
MPC	91.6±1.3	88.7±1.8	63.5±2.7	0.85
PPO	91.8±1.5	89.1±2.0	68.2±3.1	0.003
LSTM+PPO	90.4±1.9	91.1±1.6	65.4±2.8	0.004
GTrXL+PPO	92.9±1.2	91.8±1.3	61.7±2.2	0.005
GTrXL+PPO+CL	94.4±0.9	92.8±1.0	59.3±1.8	0.005

4.2.5 Evaluation protocol and statistical testing

In order to ensure statistical reliability, this paper adopts a rigorous protocol:

1. Multiple random seeds. All experiments are repeated with five independent seeds. After training each seed for 2×10^6 steps, we evaluate on 100 fixed test episodes. The statistics for each method are summarized in Table 3.

Table 3 Performance Stability Analysis under Multiple Random Seeds

Method	EV Completion Rate (%)	PV Utilization Rate (%)	CV (%)	Standard Deviation of total returns
Rule-based	85.2±2.1	88.4±3.2	4.82	523
MPC	91.6±1.3	88.7±1.8	2.21	412
PPO	91.8±1.5	89.1±2.0	3.68	523
LSTM+PPO	90.4±1.9	91.1±1.6	3.99	612
GTrXL+PPO	92.9±1.4	91.8±1.5	2.35	481
GTrXL+PPO+CL	94.4±1.2	92.8±1.3	1.80	450

Note: CV is the Coefficient of Variation of total returns.

The proposed GTrXL+PPO+CL achieves the smallest variability across seeds (coefficient of variation 1.80%, standard deviation 450), demonstrating stability.

2. Significance testing. This paper uses a paired t- test to assess statistical significance of performance improvements. For paired samples x_i and y_i with differences $d_i = x_i - y_i$, the test statistic and Cohen's effect size quantify improvement as:

$$t = \frac{\bar{d}\sqrt{n}}{s_d} \tag{32}$$

$$d_s = \frac{t}{\sqrt{n}} \tag{33}$$

Where $\bar{d}$ is the mean difference, s_d is the standard deviation of differences, and n is the number of pairs. With degrees of freedom $df = n-1$, reject the null hypothesis if $|t| > t_{critical}$. Table 4 reports the paired- test results of GTrXL+PPO+CL versus other methods, evaluated over 5 seeds ×100 fixed test episodes (pairing unit: episode, $n = 500$). Two- sided tests use $\alpha = 0.05$, and the effect size is computed as equation 33. All pairwise comparisons against the proposed method are statistically significant, with effect sizes between 0.30 and 0.68 (small to medium).

Table 4 Paired t-test Results (GTrXL+PPO+CL vs. Other Methods)

Method	t-statistic	p-value	Cohen's d	Avg. Improvement (%)
vs.Rule-based	15.21	<0.001	0.68	133.3
vs.MPC	8.53	<0.001	0.38	35.8
vs.PPO	12.14	<0.001	0.54	78.4
vs.LSTM+PPO	10.31	<0.001	0.46	65.2
vs.GTrXL+PPO	6.71	<0.001	0.30	23.6

Note: Avg. Improvement refers to the average improvement in total returns.

4.3 Intraday Six Phase Operational Strategy Analysis

The typical day operating curves elucidate the policy's decision logic across time. The 24 h horizon is partitioned into six operational phases—night preparation → morning peak → daytime charging→evening transition→evening peak → night replenishment—that reflect distinct control intents, see Figure 7.

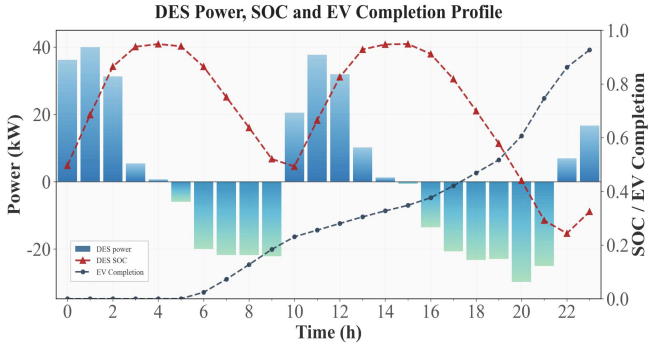

Figure 7 24-Hour Operational Strategy Decomposition on a Typical Day

• Nighttime valley period (00:00–06:00). The SOC ramps from 0.3 to 0.95 to build an energy buffer and reserve discharge capacity for the morning peak. This behavior arises from the synergy between time slot prioritization in the algorithm and relaxed nighttime soft constraints.
•Morning peak (06:00–10:00). The ESS discharges at a moderate level in coordination with EV demand to shave peaks, keeping grid active power below the soft limit.
•PV abundant midday (10:00–16:00).A"locking" strategy is executed: SOC is held nearly constant while PV is priorit- zed to directly supply EVs. If PV power exceeds 50 kW and SOC surpasses 0.7, the policy enters an energy holding mode, constraining discharge to preserve energy for the evening peak.
•Evening transition (16:00–18:00). If a peak price is fore- cast and SOC is below the stage target, the policy prefers pre-charging.
•Evening peak (18:00–22:00). Deep discharge is employ- ed for peak shaving, driving SOC toward the 0.3 lower bound. When grid power approaches 0.8× the soft limit, a forced discharge mechanism is triggered to further reduce the peak.
•Late night replenishment (22:00–24:00). The valley tariff window is used again to replenish SOC at a gentle charging rate, while a second order power change constraint suppresses oscillations. Across the day, grid exchange power is strictly kept within the 60 kW soft limit and closely tracks the constraint boundary during the evening peak, evidencing a precise balance between economic efficiency and operational security.

4.4 Economic Benefits and Parameter Sensitivity Tests

Using a per day statistical window, this paper compares methods in terms of grid purchase cost, peak suppression, and PV absorption. Costs are computed via C_{grid}; feed in revenue is credited at $0.6 \times \pi_t$. Benchmarking the final model (GTrXL+PPO+CL) yields: daily grid purchase ≈ 190 kWh, cost ≈ 88 CNY, PV utilization ≈ 92.8%, and grid peak power ≈ 56 kW—each below the 60 kW soft limit. With total EV energy demand ≈ 500 kWh, the EV completion rate reaches 94.4%. Relative to GTrXL+PPO (without curriculum learning), cost decreases by about 10%–15% (≈ 15% under a representative configuration) and the peak drops by ≈ 10 kW, indicating that the cross segment combination "evening peak release + evening pre charging + midday locking" confers a clear economic advantage under the TOU schedule and PV profile. These magnitudes are consistent with the return improvement trend in Figure 7.
Under a 10% increase in electricity price, the policy advances its pre charging preference (greater weight near 16 h), and grid purchase cost rises by 7%–9%. If a "cloudy day" PV scenario reduces the midday peak by 30% (ESS scale unchanged), PV utilization falls to 84%–86% and cost rises by ≈ 6%, yet the evening peak remains constrained near the soft limit. This stability stems from the auxiliary prediction heads and uncertainty aware reward used in training, which steer the policy toward a conservative yet safe purchase-charge/dis- charge mix when short term forecast errors increase.

4.5 Ablation Study

To elucidate the mechanism and marginal contribution of each key module, this paper designs a systematic ablation study that employs a controlled incremental approach to component manipulation. Specifically, this paper adopts a bidirectional ablation strategy that both progressively removes components from the full model and incrementally adds components to a minimal baseline, allowing us to isolate the individual impact of each module. This paper selects GTrXL+PPO as our foundational baseline due to its established effectiveness and widespread adoption in the field, providing a robust starting point for comparative analysis. Building upon this baseline, six modules are strategically layered according to a hierarchical logic that progresses from fundamental to advanced capabilities: "operating mechanism → resource coordination → robustness → training optimization." This structured progression ensures that each layer builds meaningfully upon the previous ones, forming a comprehensive ablation suite that systematically evaluates the contribution of each architectural decision, see Table 5.

Table 5 Ablation Study Results for Key Functional Modules

Method	EV Completion Rate (%)	PV Utilization Rate (%)	Average Return
GTrXL+PPO	92.9	87.8	20484
+ Evening Peak Discharge Incentive	93.4	88.1	22099
+ Dusk Pre-charging	93.7	88.6	23198
+ Daytime Energy Holding Strategy	94.0	90.6	24102
+ Dynamic Ramp Rate Constraint	94.1	91.5	24573
+ Auxiliary Prediction & Uncertainty Reward	94.3	92.1	24981
+ Curriculum Learning	94.4	92.8	25318

Results exhibit progressive gains. The baseline achieves 92.9% EV completion, 87.8% PV utilization, and an average return of 20484. Adding “evening peak discharge incentive + evening pre charging” lifts the average return to 22099 (+7.9%) by directly lowering purchase cost through TOU aware scheduling. Incorporating the “midday locking” strategy raises PV utilization to 88.6% and increases the return to 23198, effectively addressing long horizon capacity allocation. Imposing dynamic ramp constraints ensures engineering feasibility while increasing PV utilization to 90.6%. Auxiliary prediction combined with an uncertainty aware reward further enhances robustness; together they raise the return to 24981. Finally, curriculum learning with progressive constraint tightening attains the best overall performance: 94.4% EV completion, 92.8% PV utilization, and an average return of 25318, a 23.6% improvement over the baseline—validating the synergy of the multi level optimization framework.

5 CONCLUSION

This paper proposes a deep reinforcement learning method that integrates Gated Transformer-XL with PPO, leveraging cross-segment memory mechanisms and a five stage curriculum to achieve long horizon decision optimization for integrated PV–storage–charging.

Empirical results demonstrate the effectiveness of GTrXL PPO with curriculum learning: under our setup, EV charging completion reaches 94.4%, PV utilization 92.8%, daily grid purchase cost decreases by 15%, and grid peak power remains within the 60 kW soft limit. Compared with engineering baselines such as MPC and rule based scheduling, the proposed approach delivers millisecond level inference and superior overall performance. Relative to the GTrXL-PPO baseline, the average return improves by 23.6%; in paired comparisons with n=500, all tests are statistically significant (p<0.001; effect sized_s=0.30 ~ 0.68), confirming the method’s stability and reliability. Future work will target multi station coordinated scheduling and the incorporation of wind power and other renewables, further enhancing applicability and economic value to support the construction of the new type power system.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Rehman W U, Bo R, Mehdipourpich H, et al. Sizing battery energy storage and PV system in an extreme fast charging station considering uncertainties and battery degradation. Applied Energy, 2022, 313: 118745.
[2] Zhao Z, Lee C K M, Yan X, et al. Reinforcement learning for electric vehicle charging scheduling: a systematic review. Transportation Research Part E: Logistics and Transportation Review, 2024, 190: 103698.
[3] Liao X, Cao N, Li M, et al. Research on short-term load forecasting using XGBoost based on similar days. 2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS). Changsha, China, 2019: 675-678.

[4] Zhu K, Jiang Y, Wang K, et al. Day-ahead campus load interval forecast based on similar day and kernel function estimation. 2020 International Conference on Intelligent Computing, Automation and Systems (ICICAS). 2020: 145-148.
[5] Sutton R S, Barto A G. Reinforcement Learning: An Introduction. 2nd ed. Cambridge, MA: MIT Press, 2018.
[6] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. Nature, 2015, 518(7540): 529-533.
[7] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms. arXiv, 2017. DOI: 10.48550/arXiv.1707.06347.
[8] Alonso M, Amaris H, Martin D, et al. Proximal policy optimization for energy management of electric vehicles and PV storage units. Energies, 2023, 16(15): 5689.
[9] Upadhyay S, Ahmed I, Mihet-Popa L. Energy management system for an industrial microgrid using optimization-algorithms-based reinforcement learning technique. Energies, 2024, 17(16): 3898.
[10] Heendeniya C B, Nespoli L. A stochastic deep reinforcement learning agent for grid-friendly electric vehicle charging management. Energy Informatics, 2022, 5(S1): 28.
[11] Li H, Dai X, Goldrick S, et al. Reinforcement learning for EV fleet smart charging with on-site renewable energy sources. Energies, 2024, 17(21): 5442.
[12] Hermans B A L M, Walker S, Ludlage J H A, et al. Model predictive control of vehicle charging stations in grid-connected microgrids: an implementation study. Applied Energy, 2024, 368: 123210.
[13] Pascanu R, Mikolov T, Bengio Y. On the difficulty of training recurrent neural networks. Proceedings of the 30th International Conference on Machine Learning (ICML 2013). PMLR 28(3), 2013: 1310-1318.
[14] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS 2017). 2017: 5998-6008.
[15] Dai Z, Yang Z, Yang Y, et al. Transformer-XL: attentive language models beyond a fixed-length context. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019). 2019: 2978-2988.
[16] Parisotto E, Song H F, Rae J W, et al. Stabilizing transformers for reinforcement learning. Proceedings of the 37th International Conference on Machine Learning (ICML 2020). PMLR 119, 2020: 7487-7498.
[17] Li H, Wan Z, He H. Constrained EV charging scheduling based on safe deep reinforcement learning. IEEE Transactions on Smart Grid, 2020, 11(3): 2427–2439.
[18] Li W, Luo H, Lin Z, et al. A survey on transformers in reinforcement learning. arXiv, 2023. DOI: 10.48550/arXiv.2301.03044.