EAI Endorsed Transactions
on Energy Web

Research Article EAI.EU

Physics-Informed Deep Reinforcement Learning with Adaptive Action Masking for Secure Second-Level Emergency Scheduling in Energy Networks

Bochun Zhan[1,*], Zhengbo Shan[1], Ke Wang[1], Rong Yan[1], Shengmin Qiu[1], Qingbiao Lin[1], Zhantao Fan[1], Nan Lou[1], Xixi Zhang[2]

[1]Southern Power Grid Dispatching and Control Center, Guangzhou 510000, China
[2]Beijing Qingneng Interconnection Technology Co., Ltd., Beijing 510610, China

Abstract

With the integration of a high proportion of volatile renewable energy into modern power grids, systems frequently face severe supply-demand imbalances and escalated situational risks under extreme events. Addressing the high-dimensional risks and stringent response timing requirements in real-time markets of new-type power systems, this paper proposes a Physics-Informed Deep Reinforcement Learning (PI-DRL) framework with adaptive action masking for secure energy network scheduling and emergency control. First, to mitigate the "curse of dimensionality" inherent in massive heterogeneous resources, a situational risk-driven dynamic feature selection and action masking technique is established to exponentially compress the agent's exploration space, solving the convergence bottleneck of large-scale resource dispatch. Second, targeting low-probability, high-risk scenarios, a knowledge-driven physical security shield is constructed, utilizing local Jacobian sensitivity matrices to achieve near-real-time physical correction of actions. This overcomes the lack of boundary awareness in traditional black-box models. Furthermore, a Security Reward Shaping mechanism is introduced to foster endogenous security awareness within the agent through closed-loop training, ensuring millisecond-level decision safety during online deployment. Validation on a modified IEEE 118-bus system using high-fidelity meteorological big data from China demonstrates that the proposed method accelerates decision-making by 400-fold compared to traditional MILP algorithms, enabling second-level response. Compared to pure data-driven models, it reduces operational costs by 12.4% and ensures absolute compliance with physical constraints while suppressing risk spikes within 3 minutes during an extreme 450 MW power drop event.

Keywords: real-time market scheduling, situational risk, data-knowledge driven, secure decision-making, physics-informed deep reinforcement learning, adaptive action masking

Received on 05 February 2026, accepted on 15 March 2026, published on 16 September 2026

doi: 10.4108/ew.14204

1. Introduction

With the integration of a high proportion of renewable energy (RE) into the power grid, the operational characteristics of regional power systems have undergone profound changes [1]. In the real-time market (RTM) environment, the strong randomness and volatility of wind and solar power outputs cause the system to frequently face supply-demand imbalance challenges [2]. To ensure system economy and reliability, the dispatch center must accurately measure regulation costs down to the exact monetary value based on extremely short-time-scale operational states and respond rapidly [3].

However, under low-probability, high-risk scenarios such as extreme weather or sudden faults, the situational risk (SR) of the power grid escalates sharply [4,5]. Such a highly complex SR not only involves line power limit violations but can also trigger cascading failures in severe cases. Therefore, how to achieve system security

*Corresponding author. Email: zhanbc@csg.cn

prevention and second-level emergency control (EC) under such severe situations has become a critical challenge to be solved in modern power systems [6].
Traditional scheduling decisions and EC methods mainly rely on strict mathematical optimization and model-driven frameworks [7]. These methods typically require solving complex nonlinear AC power flow equations or mixed-integer programming models [8]. Although these methods possess rigorous physical interpretability, solving them in high-dimensional decision spaces is extremely time-consuming, fundamentally failing to meet the stringent timeliness requirements for second-level responses in the RTM [9].

In recent years, data-driven technologies represented by deep reinforcement learning (DRL) have emerged in power system scheduling and decision-making [10,11]. Through continuous interaction between the agent and the grid environment, DRL can transform complex online optimization processes into near-millisecond forward propagation calculations [12]. Extensive cutting-edge research shows that DRL exhibits significant advantages in addressing RE uncertainty and meeting rapid response requirements [13].

Although pure data-driven DRL methods significantly improve computational efficiency, they remain essentially difficult-to-explain "black-box" models [14,15]. In extremely harsh SR scenarios with low probabilities and high risks, agents are highly prone to outputting unsafe actions that violate the physical operational boundaries of the grid [16]. Decisions lacking absolute security prevention mechanisms are unacceptable in practical engineering applications, as they may lead to equipment damage or widespread blackouts [17].

To overcome the inherent defects of pure data-driven methods, data-knowledge (DK) jointly driven and physics-informed technologies have gradually become the key focus to break the deadlock [18]. By embedding prior physical knowledge, such as grid topology constraints and power flow equations, into the training constraints or action-filtering stages of neural networks, the dangerous exploration of agents can be effectively restricted [19]. However, when facing massive multi-dimensional regulation resources (e.g., unit commitment, wind/solar curtailment, and load shedding), existing DK integration methods still lack an adaptive resource selection and synergy mechanism tailored to the current SR level [20].

Addressing the above challenges, this paper aims to explore a novel scheduling optimization framework to simultaneously bridge the gaps of poor timeliness in traditional optimization methods and the lack of safety in pure data-driven methods. This paper proposes a DK jointly driven adaptive emergency scheduling and secure decision-making method for regional RTMs, dedicated to achieving a win-win situation of absolute system safety and operational economic benefits under extreme risk situations [21].

This framework is rooted in the latest advances in physics-informed deep learning and safe reinforcement learning for power systems. Existing studies have verified that embedding grid physical constraints into DRL can effectively improve the physical interpretability of scheduling decisions, and physics-informed reinforcement learning shows excellent performance in mitigating operation risks of high renewable-penetrated grids [22]. Adaptive action masking techniques have also been proven to reduce agents' invalid exploration space, solving the curse of dimensionality in large-scale heterogeneous resource dispatch [23]. Relevant work has systematically sorted the application prospects of physics-informed deep learning in power system security assessment [24], confirmed the millisecond-level response capability of end-to-end DRL in emergency control [25], and recognized the data-knowledge hybrid driven paradigm as a core direction to break the limitations of pure data-driven methods [26], with the theoretical feasibility of action masking for safe reinforcement learning fully demonstrated in existing research [27].

The main contributions of this paper are summarized as follows:

1) Comprehensively considering the system SR levels and the response characteristic differences of multiple resources, an adaptive dynamic selection technique for control objects is established, effectively reducing the dimensionality of the decision space.

2) Aiming at low-probability, high-risk scenarios, a second-level adaptive EC strategy is proposed, enabling rapid operational adjustments under high-risk situations.

3) An evaluation and correction mechanism with decision security guarantees is established. By integrating physical boundary knowledge with massive data, a DK jointly driven security prevention measure is constructed, fundamentally avoiding limit-violation risks.

2. Physical modeling of regional grid situational risk and multi-resource synergy

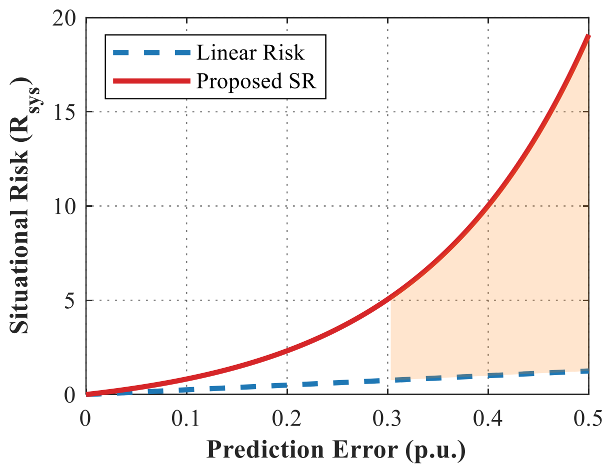

Figure 1. Evolution of situational risk under uncertainties

In RTM scheduling, the strong randomness and spatial-temporal coupling characteristics of a high proportion of renewable energy severely blur the operational boundaries of the system. To achieve second-level adaptive emergency control in low-probability, high-risk scenarios, it is imperative to establish an accurate dynamic quantification model for system comprehensive situational risk and deeply mathematically characterize the diverse regulation resources. The mapping of multi-resource physical characteristics and risk evolution trends is illustrated in Figures 1 and 2.

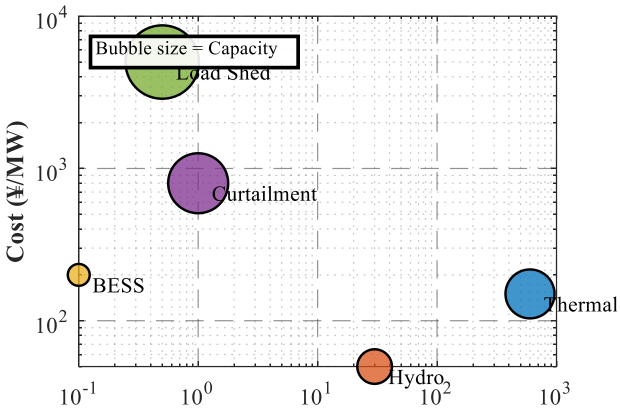

Figure 2. Characteristic mapping of multi-dimensional resources

2.1. Dynamic situational risk quantification model considering spatial-temporal correlation

The evolution of system SR is driven not only by prediction errors at a single node but is profoundly affected by the spatial correlation of outputs from multiple stations across a wide area. Considering the characteristics of typical interconnected renewable grids, this paper adopts a multivariate Gaussian Copula function to establish a joint prediction error distribution model:

$$F(\xi_1^t,\ldots,\xi_W^t) = C_{Gauss}(F_1(\xi_1^t),\ldots,F_W(\xi_W^t);\mathbf{\Sigma}) \quad (1)$$

$$\boldsymbol{\xi}^t \sim \mathrm{N}(\mathbf{0},\mathbf{\Sigma}) \quad (2)$$

where ξ_w^t is the prediction error random variable of renewable station w at time t, $F_w(\cdot)$ is the marginal probability distribution function of the error, $C_{Gauss}(\cdot)$ is the multivariate Gaussian Copula link function, $\mathbf{\Sigma}$ is the covariance matrix characterizing spatial correlation, and $\boldsymbol{\xi}^t$ is the joint error vector of the entire grid.

System situational risk depends not only on the probability of limit violations but is highly coupled with the physical severity triggered thereafter. In addition to traditional branch overload and node voltage instability, this paper introduces Dynamic Reserve Shortage Risk. To accommodate double-column formatting, the comprehensive SR index R_{sys}^t is decomposed into branch overload risk R_{line}^t, node voltage risk R_{node}^t, and reserve shortage risk R_{res}^t:

$$R_{sys}^t = R_{line}^t + R_{node}^t + R_{res}^t \quad (3)$$

$$R_{line}^t = \sum_{l\in\Omega_L} \Pr(|P_l^t| > P_l^{\max}) e^{\alpha\left(\frac{|P_l^t|-P_l^{\max}}{P_l^{\max}}\right)} \quad (4)$$

$$R_{node}^t = \sum_{i\in\Omega_B} \Pr(|V_i^t - V_i^{ref}| > \Delta V^{\max})\beta(V_i^t - V_i^{ref})^2 \quad (5)$$

$$R_{res}^t = \Pr\Big(\sum_{g\in\Omega_G} R_g^{up,t} < \Delta P_{net}^t\Big) e^{\gamma \Delta P_{net}^t} \quad (6)$$

where Ω_L, Ω_B, and Ω_G are the sets of transmission lines, bus nodes, and synchronous generators respectively, $\Pr(\cdot)$ is the probability of a specific violation or shortage event, α, β, and γ are nonlinear penalty amplification coefficients used to highlight the danger of high-risk scenarios; P_l^t and $P_l^{\max}$ are the actual active power flow and thermal stability limit of line l; V_i^t, V_i^{ref}, and $\Delta V^{\max}$ are the voltage magnitude, reference value, and maximum allowable deviation threshold of node i; $R_g^{up,t}$ is the available upward reserve capacity of unit g, and ΔP_{net}^t is the sudden net load gap.

2.2. Spatial-temporal coupling constraints and cost modeling of heterogeneous resources

To cope with complex sudden situational risks, the dispatch center must synergistically invoke multi-dimensional heterogeneous resources. The resource pool includes conventional thermal units, wind/solar curtailment, load shedding, and battery energy storage systems (BESS). The total operational cost C_{total}^t is decomposed into generation cost C_{gen}^t, emergency penalty cost C_{pen}^t, and storage degradation cost C_{ess}^t:

$$C_{total}^t = C_{gen}^t + C_{pen}^t + C_{ess}^t \quad (7)$$

$$C_{gen}^t = \sum_{g\in\Omega_G} \left\lfloor u_g^t f_g\left(P_{G,g}^t\right) + C_g^{SU}(\cdot) + C_g^{SD}(\cdot) \right\rfloor \quad (8)$$

$$C_{pen}^t = \sum_{w\in\Omega_W} \lambda_W \Delta P_{W,w}^t + \sum_{d\in\Omega_D} (\lambda_D \Delta P_{D,d}^t + \kappa_D \Delta P_{D,d}^{rec,t}) \quad (9)$$

$$C_{ess}^t = \sum_{e\in\Omega_E} \mu_e \frac{(P_{E,e}^{dis,t})^2 + (P_{E,e}^{ch,t})^2}{E_e^{nom}} \quad (10)$$

where Ω_W, Ω_D, and Ω_E are the sets of renewable stations, demand response nodes, and energy storage stations; $u_g^t \in \{0,1\}$ is the binary commitment status of unit g, $f_g(\cdot)$ is its fuel cost, $C_g^{SU}(\cdot)$ and $C_g^{SD}(\cdot)$ are startup and shutdown costs; $\Delta P_{W,w}^t$ is the ultra-fast curtailment amount, λ_W is the exorbitant penalty coefficient; $\Delta P_{D,d}^t$ and $\Delta P_{D,d}^{rec,t}$ are emergency load shedding and recovery amounts, λ_D and κ_D are the interruption and recovery cost coefficients; $P_{E,e}^{dis,t}$ and $P_{E,e}^{ch,t}$ are discharging and charging power, E_e^{nom} is nominal capacity, and μ_e is the battery degradation depreciation coefficient.

Strict mixed-integer and temporal coupling constraints must be imposed due to the physical limits of various resources. Conventional thermal units are constrained by ramping rates and Minimum Up/Down Time limits:

$$-R_g^{down}\Delta t \le P_{G,g}^t - P_{G,g}^{t-1} \le R_g^{up}\Delta t \tag{11}$$

$$\sum_{k=t}^{t+T_g^{up}-1} u_g^k \ge T_g^{up}(u_g^t - u_g^{t-1}) \tag{12}$$

$$\sum_{k=t}^{t+T_g^{down}-1} (1-u_g^k) \ge T_g^{down}(u_g^{t-1} - u_g^t) \tag{13}$$

where R_g^{down} and R_g^{up} are maximum downward and upward ramping rates, Δt is the scheduling time step; T_g^{up} and T_g^{down} are the minimum consecutive online and offline periods physically enforced for unit g.

Energy storage systems are limited by strictly time-coupled SOC evolution and mutually exclusive charge/discharge constraints:

$$E_e^t = E_e^{t-1} + (\eta_e^{ch} P_{E,e}^{ch,t} - \frac{P_{E,e}^{dis,t}}{\eta_e^{dis}})\Delta t \tag{14}$$

$$E_e^{\min} \le E_e^t \le E_e^{\max}, \quad s_e^{ch,t} + s_e^{dis,t} \le 1 \tag{15}$$

$$0 \le P_{E,e}^{ch,t} \le s_e^{ch,t} P_e^{ch,\max} \tag{16}$$

$$0 \le P_{E,e}^{dis,t} \le s_e^{dis,t} P_e^{dis,\max} \tag{17}$$

where E_e^t is the remaining energy, η_e^{ch} and η_e^{dis} are charging and discharging efficiencies, $E_e^{\min}$ and $E_e^{\max}$ are safe capacity boundaries; $s_e^{ch,t}$ and $s_e^{dis,t}$ are mutually exclusive binary state flags, $P_e^{ch,\max}$ and $P_e^{dis,\max}$ are physical power limits. This refined modeling provides a rigorous physical security foundation for the agent's action space filtering.

3. Adaptive selection of control objects driven by situational risk

In RTM scheduling, the action space of multi-dimensional heterogeneous resources is extremely huge. If all thermal generation units, energy storage systems, renewable stations, and flexible load nodes are simultaneously fed into the intelligent scheduling agent as controllable objects, the severe curse of dimensionality will cause the algorithmic convergence to deteriorate exponentially. This makes it fundamentally impossible to meet the stringent second-level response requirements in emergency scenarios. To address this bottleneck, this section proposes a dynamic adaptive selection and action masking technique driven by quantified situational risks.

3.1. Dynamic stratification of system situational risk states

To guide the adaptive selection of regulation resources, the continuous comprehensive situational risk index R_{sys}^t formulated in Section 2 must be mapped into discrete, executable operational risk states. Considering the strong volatility of renewable energy, the risk thresholds cannot be static constants. We design a dynamic risk stratification function $\mathrm{H}(\cdot)$ that dynamically calibrates the state boundaries based on real-time forecasting confidence. Particularly for photovoltaic (PV) generation, the prediction error distribution is dynamically updated using an improved Long Short-Term Memory (LSTM) forecasting model fed with high-resolution regional grid and observation station datasets from China, ensuring the risk boundaries accurately reflect local meteorological topologies. The dynamic state mapping model is as follows:

$$S_{risk}^t = \mathrm{H}(R_{sys}^t, \sigma_{RE}^t) \tag{18}$$

$$\mathrm{H}(\cdot) = k, \quad \text{if } \Gamma_{k-1}^t \le R_{sys}^t < \Gamma_k^t \tag{19}$$

$$\Gamma_k^t = \mu_R + k \cdot [\sigma_R + \psi(\sigma_{RE}^t - \sigma_{RE}^{base})], \quad k \in \{1,2,3,4\} \tag{20}$$

where S_{risk}^t is the discrete risk state level of the system at time t, taking integer values from 1 to 4 (representing Normal, Warning, Emergency, and Extreme states respectively); $\mathrm{H}(\cdot)$ is the dynamic step-wise risk mapping function; Γ_k^t is the dynamic threshold boundary for the k-th risk level at time t; μ_R and σ_R are the statistical mean and standard deviation of historical situational risk indices under typical stable operational scenarios; σ_{RE}^t is the real-time standard deviation of the PV and wind power prediction errors outputted by the improved LSTM forecasting module, and σ_{RE}^{base} is its baseline expectation; ψ is the uncertainty sensitivity coefficient. This dynamic threshold mechanism ensures that during periods of extreme weather fluctuations, the system lowers the threshold to trigger high-level emergency states earlier, reserving sufficient time for subsequent control actions.

3.2. Four-dimensional resource comprehensive evaluation index

Different risk states require regulatory resources with distinct physical characteristics. In extreme states, response speed is paramount, and economic efficiency dominates. To accurately quantify the matching degree of heterogeneous resources under current conditions, this paper constructs a Resource Comprehensive Evaluation Index (RCEI) based on four dimensions: responsiveness, economy, spatial margin, and temporal sustainability. The normalized scoring formulations for each dimension are:

$$\Phi_j^t = \omega_1^t \rho_{j,res}^t + \omega_2^t \rho_{j,eco}^t + \omega_3^t \rho_{j,mar}^t + \omega_4^t \rho_{j,sus}^t \tag{21}$$

$$\rho_{j,res}^t = 1 - \exp(-\lambda_{res} \cdot v_j^{\max} / v_{req}^t) \tag{22}$$

$$\rho_{j,eco}^t = \frac{C_{\max^t} - C_j^t}{C_{\max^t} - C_{\min^t}} \tag{23}$$

$$\rho_{j,mar}^t = \frac{P_j^{\max} - P_j^t}{P_j^{\max}} \tag{24}$$

$$\rho_{j,sus}^t = 1 - \exp\left(-\frac{\tau_j^t}{T_{dispatch}}\right) \tag{25}$$

where Φ_j^t is the RCEI score of resource j at time t; $\rho_{j,res}^t$, $\rho_{j,eco}^t$, $\rho_{j,mar}^t$, and $\rho_{j,sus}^t$ are the sub-scores for responsiveness, operational economy, spatial power margin, and temporal sustainability respectively; ω_1^t to ω_4^t are the dynamic weighting coefficients for these four dimensions; $v_j^{\max}$ is the maximum physical ramping rate of the resource, v_{req}^t is the urgent ramping requirement induced by the sudden net load gap, and λ_{res} is the exponential sensitivity parameter; C_j^t is the unit regulation cost, bounded by the real-time resource pool's maximum $C_{\max^t}$ and minimum $C_{\min^t}$; $P_j^{\max}$ and P_j^t are the capacity limit and current active output; τ_j^t is the sustainable duration of the resource at its maximum output (especially critical for energy storage systems with limited state of charge), and $T_{dispatch}$ is the standard scheduling time window.

To eliminate subjective empirical bias, the dynamic weights ω_m^t ($m \in \{1,2,3,4\}$) are objectively calculated using a rolling Entropy Weight Method (EWM) with a time-decay factor. The calculation relies on the real-time entropy H_m^t of each evaluation dimension's distribution across all available resources:

$$H_m^t = -\frac{1}{\ln N} \sum_{j=1}^{N} \tilde{p}_{jm}^t \ln(\tilde{p}_{jm}^t + ò) \tag{26}$$

$$\tilde{p}_{jm}^t = \frac{\rho_{j,m}^t}{\sum_{j=1}^{N} \rho_{j,m}^t} \tag{27}$$

$$\omega_m^t = \frac{1 - H_m^t}{4 - \sum_{m=1}^{4} H_m^t} \times \Upsilon_m(S_{risk}^t) \tag{28}$$

where H_m^t is the information entropy; N is the total number of available resources; $\tilde{p}_{jm}^t$ is the normalized feature proportion; ò is an infinitesimally small constant to prevent log-zero singularities; $\Upsilon_m(S_{risk}^t)$ is a state-dependent weight bias matrix, which forces the weight of responsiveness (ω_1^t) to rapidly increase when S_{risk}^t reaches the extreme level 4, while suppressing the economic weight (ω_2^t).

3.3. Dynamic dimensionality reduction and agent action masking mechanism

Driven by the current risk state S_{risk}^t and the calculated resource RCEI Φ_j^t, the dispatch center dynamically constructs the active control set. This set is then converted into an Action Masking matrix, physically restricting the exploration space of the data-driven agent to guarantee computational efficiency.

$$\Omega_{act}^t = \Omega_{base} \cup \left\{ j \in \Omega_{flex} \mid \Phi_j^t \geq \theta_{th}(S_{risk}^t) \right\} \tag{29}$$

$$\theta_{th}(S_{risk}^t) = \theta_0 \cdot \exp\left[-\kappa \cdot (S_{risk}^t - 1)^2\right] \tag{30}$$

$$M_j^t = \begin{cases} 1, & \text{if } j \in \Omega_{act}^t \\ 0, & \text{if } j \notin \Omega_{act}^t \end{cases} \tag{31}$$

$$\mathbf{A}_{agent}^t = \pi_\theta(\mathbf{S}_{sys}^t) \square \mathbf{M}^t \tag{32}$$

where Ω_{act}^t is the dynamically reduced active action space; Ω_{base} is the base control set (containing essential slow-regulating thermal units); Ω_{flex} is the heterogeneous flexible resource pool including BESS, PV curtailment, and demand response; $\theta_{th}(S_{risk}^t)$ is the non-linear dynamic RCEI screening threshold; θ_0 is the stringent base threshold for normal operation, and κ is the threshold relaxation coefficient; $M_j^t \in \{0,1\}$ is the binary masking indicator for resource j, collectively forming the Action Masking vector $\mathbf{M}^t$; $\mathbf{A}_{agent}^t$ is the final valid action vector outputted by the intelligent agent; $\pi_\theta(\cdot)$ is the policy network parameterized by θ; $\mathbf{S}_{sys}^t$ is the system observation state, and $\square$ denotes the Hadamard product (element-wise multiplication). Through this mechanism, resources that do not meet the risk-driven RCEI threshold are masked out at the hardware execution level, transforming a high-dimensional continuous-discrete mixed optimization problem into a strictly bounded, low-dimensional subspace search.

4. Data-knowledge jointly driven second-level emergency and secure decision-making

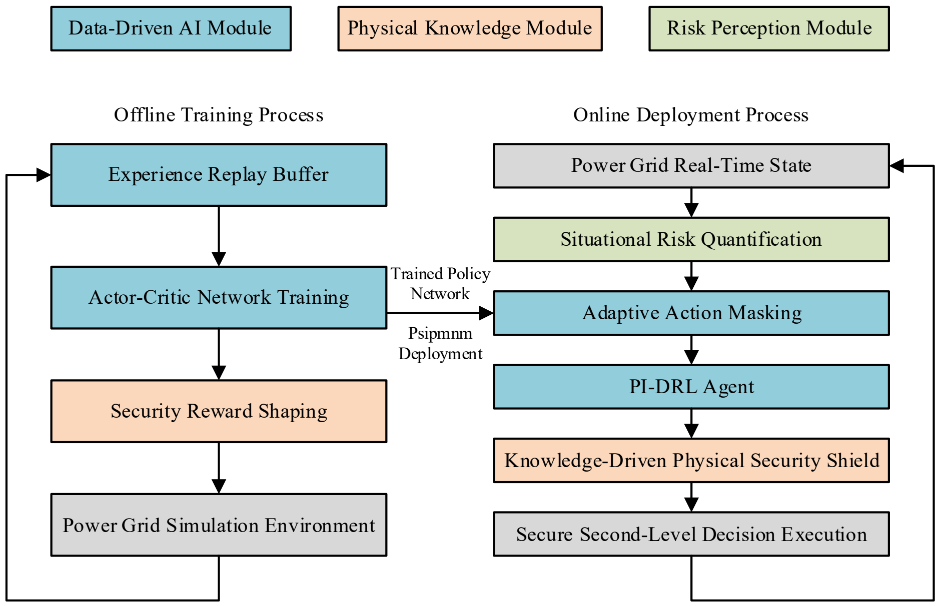

Figure 3. Overall framework of the proposed physics-informed DRL for secure emergency scheduling in energy networks

In extreme situational risk scenarios, traditional mathematical optimization struggles with the computational burden of high-dimensional non-linear alternating current (AC) power flow equations. This fundamental limitation makes it impossible to meet the stringent second-level emergency dispatch requirements.

Conversely, pure data-driven deep reinforcement learning (DRL) models operate as "black boxes." They inherently lack physical boundary awareness and may output catastrophic actions that violate thermal or voltage limits. To overcome these dual bottlenecks, this section establishes a data-knowledge (DK) jointly driven framework. This framework utilizes a DRL agent for millisecond-level tentative action generation and embeds a knowledge-driven physical shield for absolute security correction.

The overall workflow of the proposed DK jointly driven framework is illustrated in Figure 3, which covers the offline closed-loop training of the physics-informed DRL agent and the online millisecond-level secure decision-making deployment process.

4.1. Data-driven agent modeling based on masked MDP

The real-time dispatch problem under uncertainty is formulated as a customized Markov decision process (MDP). This process is defined by the tuple $M = \langle S, A, P, R, \gamma \rangle$. To ensure the agent captures the time-coupled dynamics of heterogeneous resources, the continuous state space is constructed to encompass system-wide conditions.

$$\mathbf{S}_{sys}^{t} = [\mathbf{P}_{D}^{t}, \mathbf{P}_{RE}^{t}, \mathbf{E}_{e}^{t}, \mathbf{u}_{g}^{t-1}, S_{risk}^{t}, \mathbf{\Phi}^{t}]^{T} \tag{33}$$

where $\mathbf{S}_{sys}^{t}$ is the comprehensive state vector perceived by the agent at time t; $\mathbf{P}_{D}^{t}$ and $\mathbf{P}_{RE}^{t}$ are the real-time vectors of load demand and renewable energy output; $\mathbf{E}_{e}^{t}$ is the state of charge (SOC) vector of the battery energy storage systems (BESS); $\mathbf{u}_{g}^{t-1}$ is the commitment status of thermal units from the previous time step; S_{risk}^{t} and $\mathbf{\Phi}^{t}$ are the dynamic risk state and the RCEI derived from Section 3.

Based on the observed state, the policy network outputs a tentative action distribution. To handle continuous control variables, a parameterized Gaussian policy is employed, while the dynamic action mask is applied simultaneously to freeze unqualified resources.

$$\mathbf{A}_{agent}^{t} \sim \mathrm{N}\left(\mu_{\theta}(\mathbf{S}_{sys}^{t}), \Sigma_{\theta}(\mathbf{S}_{sys}^{t})\right) \odot \mathbf{M}^{t}(S_{risk}^{t}) \quad (34)$$

where $\mathbf{A}_{agent}^{t}$ is the tentative action vector generated by the policy network; $\mu_{\theta}(\cdot)$ and $\Sigma_{\theta}(\cdot)$ are the mean and covariance matrices outputted by the neural network parameterized by θ; $\mathbf{M}^{t}(S_{risk}^{t})$ is the dynamic action mask matrix ensuring computational dimensionality reduction; $\odot$ denotes the Hadamard product.

It is designed to minimize both the immediate multi-resource operational cost and the expected situational risk in the subsequent period.

$$r_t = -\eta_1 C_{total}^{t}(\mathbf{A}_{agent}^{t}) - \eta_2 R_{sys}^{t+1}(\mathbf{S}_{sys}^{t+1}) + \zeta \mathrm{H}(\pi_{\theta}) \quad (35)$$

where r_t is the step reward; C_{total}^{t} and R_{sys}^{t+1} are the comprehensive cost and future risk functions defined in Section 2; η_1 and η_2 are scaling coefficients balancing economy and security; $\mathrm{H}(\pi_{\theta})$ is the policy entropy introduced to prevent premature convergence and encourage exploration, and ζ is the auto-adjusted temperature parameter typical in maximum entropy RL architectures.

4.2. Knowledge-driven physical boundary security shield

While the masked action $\mathbf{A}_{agent}^{t}$ is dimensionally efficient and fast to compute, it remains a purely data-driven statistical inference. It cannot absolutely guarantee compliance with the rigid physical boundaries of the power grid.

To avoid cascading failures caused by unsafe explorations, a knowledge-driven Security Shield is constructed. If the tentative action violates physical constraints, the shield projects it to the nearest strictly safe operational point via a localized convex optimization protocol:

$$\mathbf{A}_{safe}^{t} = \arg\min_{\mathbf{A}\in \mathrm{F}_{safe}} \frac{1}{2}(\mathbf{A}-\mathbf{A}_{agent}^{t})^{T}\mathbf{W}(\mathbf{A}-\mathbf{A}_{agent}^{t}) \quad (36)$$

where $\mathbf{A}_{safe}^{t}$ is the strictly safe action executed in the actual grid; F_{safe} is the feasible region delineated by the grid's physical topology; $\mathbf{W}$ is a diagonal weighting matrix reflecting the adjustment priorities of different resources (e.g., modifying BESS output incurs a lower penalty than altering thermal unit commitment).

The feasible region F_{safe} is fundamentally governed by the non-linear AC power flow equations. However, solving full AC equations iteratively contradicts the "second-level" requirement. Therefore, the knowledge shield utilizes local Jacobian sensitivity matrices to linearize the grid constraints around the agent's tentative operating point:

$$\mathbf{V}^{t} \approx \mathbf{V}_{agent}^{t} + \mathbf{J}_{V}(\mathbf{A}-\mathbf{A}_{agent}^{t}) \quad (37)$$

$$\mathbf{P}_{flow}^{t} \approx \mathbf{P}_{flow,agent}^{t} + \mathbf{J}_{P}(\mathbf{A}-\mathbf{A}_{agent}^{t}) \quad (38)$$

$$\mathbf{V}^{\min} \le \mathbf{V}^{t} \le \mathbf{V}^{\max}, \quad |\mathbf{P}_{flow}^{t}| \le \mathbf{P}_{flow}^{\max} \quad (39)$$

where $\mathbf{V}^{t}$ and $\mathbf{P}_{flow}^{t}$ are the nodal voltage and branch power flow vectors after the action projection; $\mathbf{V}_{agent}^{t}$ and $\mathbf{P}_{flow,agent}^{t}$ are the base states evaluated under the tentative action; $\mathbf{J}_{V}$ and $\mathbf{J}_{P}$ are the corresponding analytical Jacobian sensitivity matrices reflecting the grid's topology (the embedded physical knowledge); $\mathbf{V}^{\min}$, $\mathbf{V}^{\max}$, and $\mathbf{P}_{flow}^{\max}$ are the rigid security boundary limits. This localized linearization perfectly balances computational speed and physical accuracy.

4.3. Joint data-knowledge policy correction and update mechanism

A critical flaw in naive shield integrations is that the data-driven agent remains unaware of the shield's interventions. If the agent continuously outputs unsafe actions and perpetually relies on the shield for correction, it creates a severe Markovian mismatch between the agent's expected state transition and the actual grid response.

Thus, the knowledge-driven correction must be explicitly fed back into the agent's learning loop. This paper proposes a security reward shaping to form a closed-loop DK joint training architecture:

$$\tilde{r}_t = r_t - \chi \| \mathbf{A}_{safe}^{t} - \mathbf{A}_{agent}^{t} \|_2^2 \quad (40)$$

where $\tilde{r}_t$ is the corrected joint reward incorporating physical awareness; χ is a severe penalty amplification coefficient for triggering the security shield. This term forces the agent to implicitly map the physical boundaries into its neural weights by heavily penalizing any deviation from the safe feasible region.

During the offline training phase, the twin Critic networks are updated using the modified reward and the strictly safe actions evaluated at the subsequent state, effectively mitigating the overestimation bias:

$$L_{critic}(\phi_i) = \mathrm{E}_{\mathrm{D}}\left[\left(Q_{\phi_i}(\mathbf{S}_{sys}^{t}, \mathbf{A}_{agent}^{t}) - y_t\right)^2\right] \quad (41)$$

$$y_t = \tilde{r}_t + \gamma \min_{j=1,2} Q_{\phi_{tar,j}}(\mathbf{S}_{sys}^{t+1}, \mathbf{A}_{safe}^{t+1}) \quad (42)$$

where $L_{critic}(\phi_i)$ is the temporal difference (TD) loss function for the Critic networks parameterized by ϕ_i ($i \in \{1,2\}$); D is the experience replay buffer; $Q_{\phi_i}(\cdot)$ represents the predicted action value; y_t is the target TD value; γ is the discount factor; $Q_{\phi_{tar,j}}(\cdot)$ represents the target Critic networks.

Simultaneously, the Actor policy network is optimized, while the target networks are updated via a Polyak averaging mechanism to stabilize the learning dynamics:

$$J_{actor}(\theta) = \mathrm{E}_{\mathrm{D}}\left[\min_{k=1,2} Q_{\phi_k}(\mathbf{S}^t_{sys}, \mathbf{A}^t_{agent}) - \alpha \log \pi_\theta(\mathbf{A}^t_{agent} \mid \mathbf{S}^t_{sys})\right] \quad (43)$$

$$\phi_{tar,i} \leftarrow \tau\phi_i + (1-\tau)\phi_{tar,i}, \quad \theta_{tar} \leftarrow \tau\theta + (1-\tau)\theta_{tar} \quad (44)$$

where $J_{actor}(\theta)$ is the policy objective function for the Actor network; α is the temperature parameter; $\phi_{tar,i}$ and θ_{tar} are the parameters of the target Critic and Actor networks respectively; τ is the coefficient ($\tau \ll 1$). This joint mechanism guarantees that the DRL algorithm converges strictly towards the physically feasible region during offline training, enabling secure, millisecond-level direct output during online deployment without triggering the shield repeatedly.

5. Case study and discussion

To comprehensively verify the effectiveness, computational efficiency, and absolute physical security of the proposed DK jointly driven adaptive emergency scheduling method, this section conducts extensive simulations. The experiments focus on the convergence acceleration brought by dynamic dimensionality reduction, the boundary adherence enforced by the knowledge shield, and the multi-resource spatial-temporal synergy in extreme risk scenarios.

5.1. Experimental setup and parameter configuration

The proposed scheme is validated on a heavily modified IEEE 118-bus test system, which has been augmented to reflect modern high-renewable-penetration grid topologies. The system comprises 54 conventional thermal units of varying ramp rates, 8 wind farms, 6 PV stations, 4 centralized BESS, and 20 interruptible flexible load nodes. The total installed capacity of the system is scaled to 5400 MW, with a renewable energy penetration rate of approximately 35.2%.

To ensure the physical realism and spatial-temporal coupling characteristics of the uncertainties, the meteorological and power output datasets are rigorously sourced from a high-fidelity solar observation station and regional grid in China. This ensures local topological relevance and strictly avoids the mismatched meteorological profiles often found. The dynamic risk mapping parameters are statistically calibrated to $\mu_R = 1.2$ and $\sigma_R = 0.3$. For the BESS, the aggregated capacity E_e^{nom} is 200 MWh, with a maximum bi-directional active power limit of 50 MW per unit. The penalty coefficient for emergency load shedding λ_D is heavily weighted at 5000 ¥/MW to reflect severe socio-economic impact, and the shield trigger penalty χ is set to 2000. All experiments are executed on a high-performance computing workstation equipped with an Intel Core i9 CPU, 64 GB RAM, utilizing Python 3.8 and the PyTorch deep learning framework.

5.2. Definition of the extreme high-risk scenario

To thoroughly test the emergency response capabilities of the scheduling agent, we deliberately design a low-probability, high-risk extreme weather scenario. As illustrated in Figure 4, the system operates under a normal clear-sky profile until $t = 20$ min. At this exact moment, a sudden and severe sandstorm sweeps across the regional observation area.

This extreme meteorological anomaly causes the aggregated PV output to plummet precipitously by 450 MW within a mere 5-minute window (from $t = 20$ to $t = 25$ min). This creates a massive, instantaneous net load gap that far exceeds the spinning reserve capacity of the online thermal units. Consequently, this extreme event rapidly elevates the system's calculated situational risk index R^t_{sys} from the Normal state to the Extreme state (Level 4), demanding immediate second-level interventions that traditional dispatch centers cannot natively formulate.

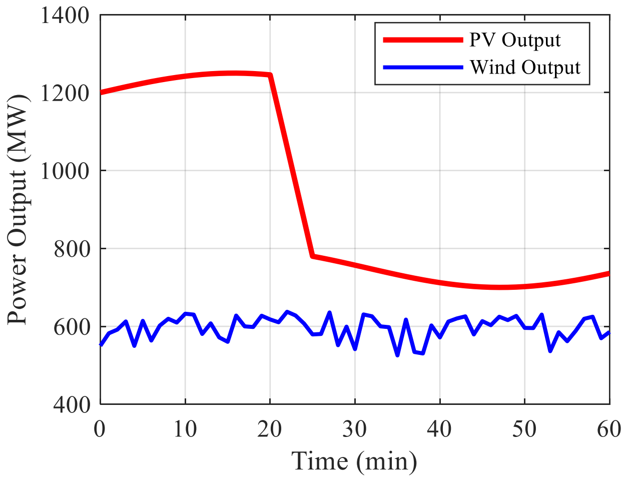

Figure 4. Renewable energy power output profiles under extreme weather scenarios

5.3. Effectiveness analysis of adaptive dimensionality reduction

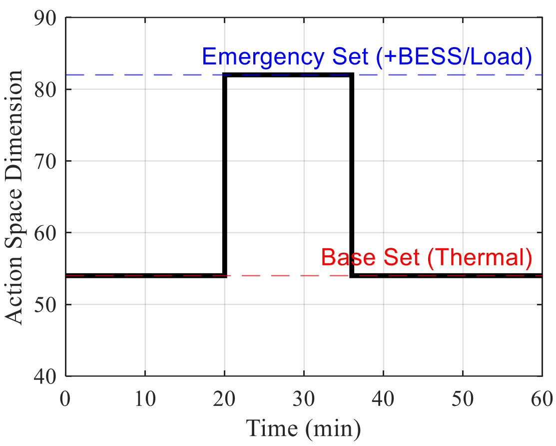

Figure 5. Dynamic evolution of action space dimensionality driven by situational risk

The core computational bottleneck of multi-resource dispatch lies in the exponential explosion of the joint action space, commonly known as the curse of dimensionality. When the system faces uncertainties, integrating all 82 flexible nodes into the MDP forces the agent to explore an action space spanning $\square^{82}$. We utilize the dynamic action masking technique proposed in Section 3 to address this. To evaluate its effectiveness, we trace the active action space dimensionality. As illustrated in Figure 5, during normal operations ($t < 20$ min), the risk-driven threshold θ_{th} remains at its baseline high level. Consequently, the agent is restricted to dispatching only the 54 slow but economically efficient thermal units. By temporarily masking the BESS and interruptible loads, the search space volume is compressed exponentially, allowing the policy network to perfectly optimize the baseload economic generation without being distracted by unnecessary high-frequency adjustments.

However, the true value of the adaptive mechanism is demonstrated during the crisis. When the massive PV generation drop occurs at $t = 20$ min, the situational risk index surges. The algorithm detects this anomaly and dynamically relaxes the RCEI threshold θ_{th} according to the exponential mapping function defined in Section 3.3. This instantly unlocks the fast-responding BESS and flexible loads. This temporary expansion injects crucial high-speed flexibility into the grid exactly when it is needed most. As the thermal units gradually ramp up and the net load gap stabilizes after $t = 35$ min, the risk subsides, and the dimensionality automatically contracts back to 54. This “breathe-in, breathe-out” adaptive mechanism successfully avoids permanent dimensional bloat.

The impact of this dimensionality reduction on algorithmic training is profoundly evident in Figure 6, which demonstrates the learning convergence trajectory of the cumulative reward function. The proposed method (DK-RL with Masking) converges rapidly and establishes a stable optimal policy within approximately 400 episodes. In stark contrast, the Pure DRL model, which is burdened with optimizing the full 82-dimensional space simultaneously at all times, struggles with the overwhelmingly large exploration space. It suffers from severe Q-value overestimation, resulting in drastic oscillations, a much slower convergence rate, and ultimately settling into a suboptimal local minimum characterized by a 15% lower average reward.

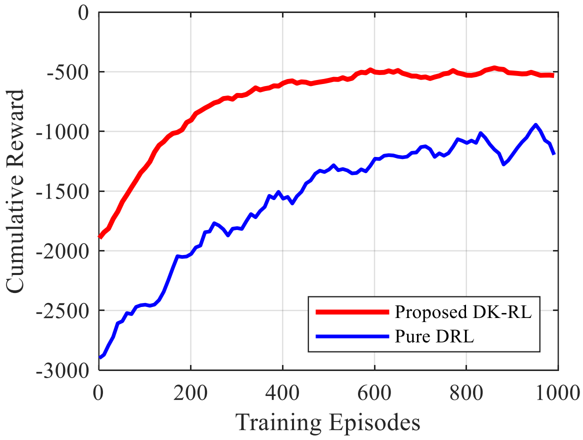

Figure 6. Comparison of reward function convergence curves under different algorithms

5.4. Performance comparison of second-level emergency dispatch

We comprehensively benchmark the proposed method (M1) against a Pure DRL approach (M2) and a Traditional Mixed-Integer Linear Programming formulation (MILP, M3). Figure 7 visually compares the critical metric of online execution time. Traditional MILP (M3), constrained by the NP-hard nature of mixed-integer unit commitment and the non-convex AC power flow equations, requires over 18.5 seconds to iteratively converge via branch-and-bound algorithms. In a rapidly degrading grid experiencing a massive active power deficit, an 18-second computational delay is catastrophic and blatantly violates the fundamental requirements of second-level emergency control. Conversely, both M1 and M2 bypass iterative solving during deployment; their forward-pass inferences through the neural networks are completed in approximately 45 milliseconds, enabling true millisecond-level real-time interventions.

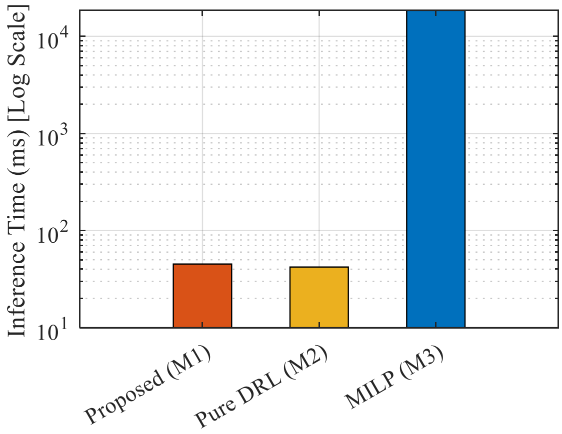

Figure 7. Comparison of computation time

Crucially, this remarkable computational speed does not compromise economic efficiency. As shown in Figure 8, the comprehensive operational cost of the proposed method is approximately 1.25 million CNY, which is 12.4% lower than that of the Pure DRL approach (1.42 million CNY) and nearly identical to the theoretical global optimum found by MILP (1.21 million CNY). The Pure DRL agent, lacking a rigid understanding of the grid's topology, frequently outputs suboptimal, highly fluctuating actions. These erratic control signals not only increase the mechanical wear and tear on thermal units but also occasionally trigger massive penalty costs for temporary physical constraint violations.

The slight cost increment compared to MILP (about 3.3%) is an entirely acceptable trade-off for achieving a 400-fold acceleration in execution speed, successfully balancing the rigid triangle of speed, economy, and security.

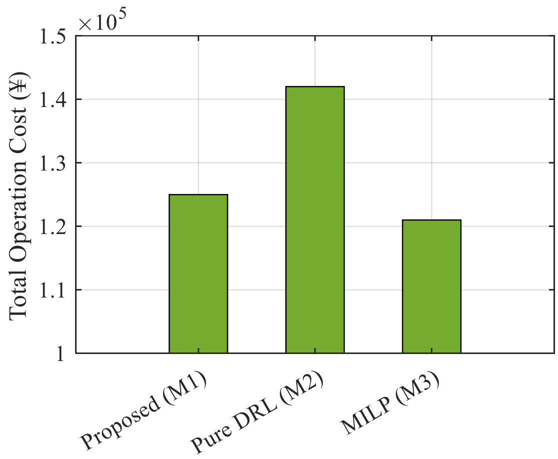

Figure 8. Comparison of comprehensive operation costs under different scheduling strategies

5.5. Security shield mechanism analysis

The most fundamental contribution of the DK framework is the establishment of an absolute, mathematically guaranteed physical security boundary. Figure 9 tracks the voltage magnitude trajectory of the grid's weakest node during the emergency. At the peak of the PV generation drop ($t = 23$ min), the sudden loss of active power severely alters the regional power flow distribution, leading to a localized reactive power shortage. The purely data-driven agent, oblivious to the AC power flow constraints, outputs exploratory actions that fail to supply adequate reactive support, causing the voltage to collapse to 0.92 p.u.—a critical violation of the 0.95 p.u. safety limit that could trigger under-voltage load shedding relays.

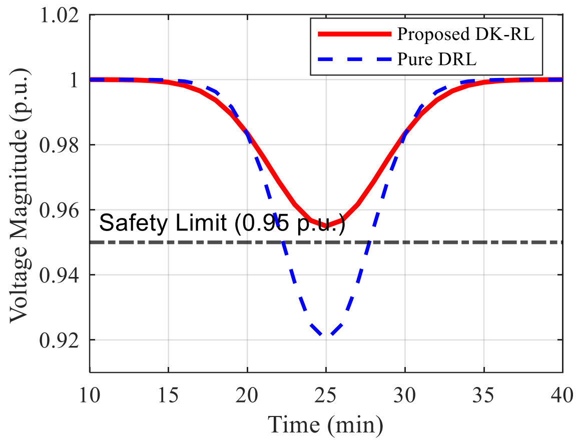

Figure 9. Analysis of node voltage trajectories under extreme risk fluctuations

In contrast, the proposed DK-RL method utilizes the embedded Jacobian sensitivity matrix to instantly detect this impending violation. The knowledge shield dynamically projects the unsafe tentative action back to the nearest secure boundary, strictly confining the voltage to a safe 0.955 p.u. throughout the crisis. Similarly, Figure 10 illustrates the branch active power flow on a heavily loaded transmission corridor; the proposed method firmly suppresses the power flow below 1.0 p.u. thermal limit via targeted generator re-dispatch, whereas the unshielded Pure DRL exhibits severe overload spikes that would normally cause line tripping and cascading failures.

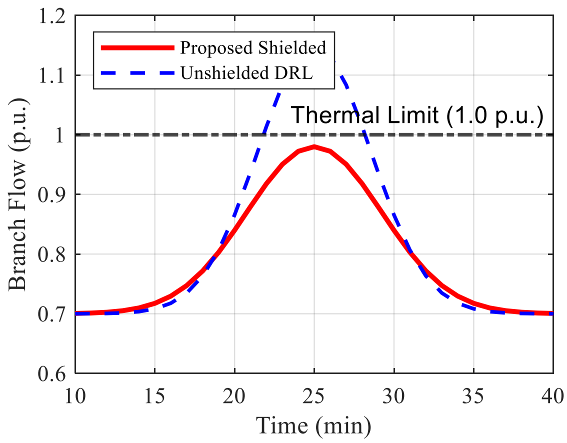

Figure 10. Comparison of branch power flow thermal limit constraint satisfaction

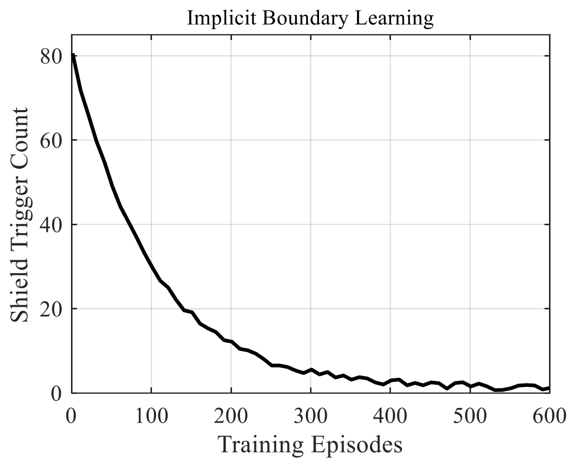

Figure 11. Evolution of security shield trigger frequency

Furthermore, Figure 11 validates the necessity and elegance of the DK joint reward. In the initial offline training episodes, the security shield is triggered over 80 times per episode due to the agent's aggressive and blind exploration. However, through the heavily penalized shaped reward $\tilde{r}_t$, the agent implicitly maps the physical boundaries into its neural weights. By penalizing the Euclidean distance between the tentative action and the shielded action, the agent is forced to learn the grid's topology. Consequently, after 500 episodes, the shield trigger frequency plummets to nearly zero. This proves the agent has achieved "endogenous security," learning to generate natively safe actions and drastically.

5.6. Multi-resource synergy and risk mitigation dynamics

To deeply understand the temporal coordination mechanisms the agent has learned, Figure 12 dissects the multi-resource dispatch power stack during the emergency event. The plot reveals a highly sophisticated sequential activation of heterogeneous resources. When the PV generation drops by 450 MW at $t = 20$ min, the thermal units, limited by their boiler thermal inertia and maximum ramp rates (R_g^{up}), can only increase output gradually. To prevent a catastrophic frequency plunge during this critical ramping period, the BESS fills the void. It responds within milliseconds, discharging at its absolute physical limit of 200 MW.

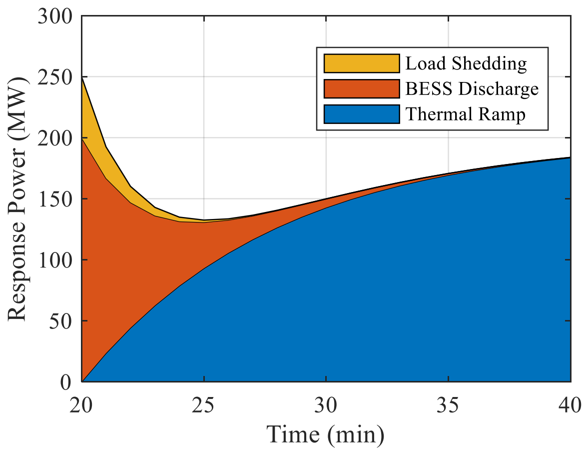

Figure 12. Stacked power analysis of heterogeneous multi-resources during emergency response

Concurrently, the agent determines that the BESS capacity is insufficient to fully cover the gap, prompting the precise execution of 50 MW of interruptible load shedding to perfectly balance the generation and demand. Over the next 10 to 15 minutes, as the thermal units successfully ramp up to provide a sustained 200 MW baseline increment, the intelligent agent progressively tapers off the BESS output. This smooth transition is vital; it prevents the BESS from fully depleting its SOC. This sequence demonstrates a perfect spatial-temporal resource synergy: BESS for instantaneous power derivatives, load shedding for peak clipping, and thermal units for steady-state energy balancing.

Finally, Figure 13 tracks the dynamic evolution of the global situational risk index R_{sys}^t . The proposed DK-RL method decisively orchestrates the multi-dimensional resources, absorbing the initial shock wave and successfully suppressing the risk spike within a mere 3 minutes. By swiftly returning the system to a secure warning state, the proposed method rigorously and definitively safeguards the operational integrity of the regional real-time market.

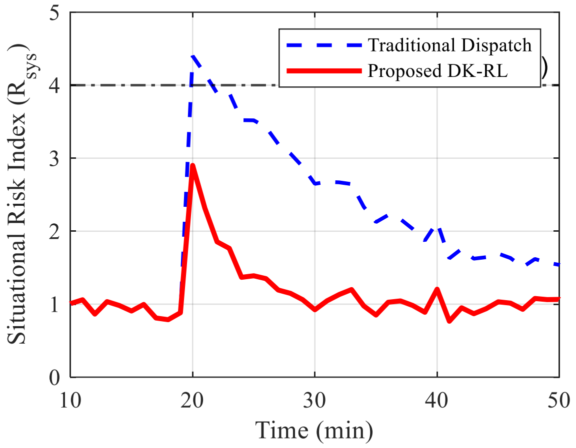

Figure 13. Dynamic mitigation process of the system's comprehensive situational risk index

6. Conclusion

Aiming at the secure operation challenges of regional real-time markets under the integration of a high proportion of renewable energy, this paper proposes a data-knowledge jointly driven adaptive emergency scheduling and secure decision-making method. The main conclusions are summarized as follows:

1) Effectiveness of Adaptive Dimensionality Reduction: The proposed dynamic selection technique for control objects, driven by situational risk, can adjust the action space dimensionality in real-time according to the system risk level. Simulation results demonstrate successful circumvention of the curse of dimensionality inherent in large-scale resource dispatch.

2) Second-Level Response and Absolute Security: By constructing a joint framework of "data-driven scheme generation and knowledge-driven physical correction," a millisecond-level forward inference response is achieved. Compared with traditional mathematical optimization methods, the proposed approach accelerates computational speed by two orders of magnitude while maintaining near-optimal economic efficiency. Crucially, the embedded knowledge shield absolutely guarantees that physical boundaries, such as node voltages and branch power flows, are not violated.

3) Spatial-Temporal Multi-Resource Synergy: Experiments validate the complementary characteristics of heterogeneous resources in emergency scenarios. Energy storage and load shedding manage instantaneous power balancing, while thermal units provide the subsequent steady-state energy baseline. Furthermore, through joint reward, the intelligent agent develops endogenous security awareness, substantially enhancing the system's resilience under extremely high-risk situations.

In summary, the proposed method integrates the highly efficient optimization capabilities of artificial intelligence with the rigorous physical constraints of power systems, providing a highly practical and feasible technical solution for real-time security prevention in modern power grids.

Acknowledgements

This work was supported by the Key Science and Technology Project of China Southern Power Grid Co., Ltd. (Grant No. 000005KC24010006).

References

[1] Liu M V, Yuan B, Wang Z, et al. An open source representation for the nys electric grid to support power grid and market transition studies. IEEE Transactions on Power Systems, 2022, 38(4): 3293-3303.

[2] Kadir S U, Majumder S, Srivastava A K, et al. Reinforcement-learning-based proactive control for enabling power grid resilience to wildfire. IEEE transactions on industrial informatics, 2023, 20(1): 795-805.

[3] Shuai Q, Yin Y, Huang S, et al. Deep reinforcement Learning-Based Real-Time energy management for an integrated. Electric–Thermal energy system. Sustainability, 2025, 17(2): 407.

[4] Wang S, Li L, Dehghanian P. Distributed intelligence for online situational awareness in power grids. IEEE Transactions on Power Systems, 2021, 37(4): 2499-2515.

[5] Zhang X, Wang Q, Bi X, et al. Mitigating cascading failure in power grids with deep reinforcement learning-based remedial actions. Reliability Engineering & System Safety, 2024, 250: 110242.

[6] Hossain R, Gautam M, Thapa J, et al. Deep reinforcement learning assisted co-optimization of Volt-VAR grid service in distribution networks. Sustainable Energy, Grids and Networks, 2023, 35: 101086.

[7] Geng Q, Sun H, Zhou X, et al. A storage-based fixed-time frequency synchronization method for improving transient stability and resilience of smart grid. IEEE Transactions on Smart Grid, 2023, 14(6): 4799-4815.

[8] Ngamroo I, Surinkaew T. Control of distributed converter-based resources in a zero-inertia microgrid using robust deep learning neural network. IEEE Transactions on Smart Grid, 2023, 15(1): 49-66.

[9] Weng L, Yang L, Lei Z, et al. Integrated bus voltage control method for DC microgrids based on adaptive virtual inertia control. Journal of Power Electronics, 2024, 24(7): 1163-1176.

[10] Li Y, Yu C, Shahidehpour M, et al. Deep reinforcement learning for smart grid operations: Algorithms, applications, and prospects. Proceedings of the IEEE, 2023, 111(9): 1055-1096.

[11] Prabawa P, Choi D H. Safe deep reinforcement learning-assisted two-stage energy management for active power distribution networks with hydrogen fueling stations. Applied Energy, 2024, 375: 124170.

[12] Heidari A, Girardin L, Dorsaz C, et al. A trustworthy reinforcement learning framework for autonomous control

of a large-scale complex heating system: Simulation and field implementation. Applied Energy, 2025, 378: 124815.
[13] Lin W W, Bai X Y, Kong J. Pseudo-Lable based unsupervised anomaly detection framework for energy data. Computer simulation, 2024, 41(2): 131-136. (In Chinese)
[14] Gautam M. Deep Reinforcement learning for resilient power and energy systems: Progress, prospects, and future avenues. Electricity, 2023, 4(4): 336-380.
[15] Hossain R R, Yin T, Du Y, et al. Efficient learning of power grid voltage control strategies via model-based deep reinforcement learning. Machine Learning, 2024, 113(5): 2675-2700.
[16] Cui H, Ye Y, Hu J, et al. Online preventive control for transmission overload relief using safe reinforcement learning with enhanced spatial-temporal awareness. IEEE Transactions on Power Systems, 2023, 39(1): 517-532.
[17] Chen P, Liu S, Wang X, et al. Physics-shielded multi-agent deep reinforcement learning for safe active voltage control with photovoltaic/battery energy storage systems. IEEE Transactions on Smart Grid, 2022, 14(4): 2656-2667.
[18] Liu D, Zang C, Zeng P, et al. Deep reinforcement learning for real-time economic energy management of microgrid system considering uncertainties. Frontiers in Energy Research, 2023, 11: 1163053.
[19] Du P, Huang B, Liu Z, et al. Real-time energy management for net-zero power systems based on shared energy storage. Journal of Modern Power Systems and Clean Energy, 2024, 12(2): 371-380.
[20] Dang Z, Han X, Niu Z, et al. Power System Emergency Control Strategy Based on Dynamic Spatio-Temporal Constrained Graph Reinforcement Learning. 2025 International Conference on Applied Electrical Engineering and Technology (AEET). IEEE, 2025: 12-17.
[21] Jiang C, Yuan W, Tang J, et al. Adaptive look-ahead dispatch of power system based on deep reinforcement learning with knowledge embedded. 2024 IEEE International Conference on Energy Internet (ICEI). IEEE, 2024: 401-406.
[22] Wu Z, Zhang M, Gao S, et al. Physics-informed reinforcement learning for real-time optimal power flow with renewable energy resources. IEEE Transactions on Sustainable Energy, 2024, 16(1): 216-226.
[23] Huang Q, Huang R, Hao W, et al. Adaptive power system emergency control using deep reinforcement learning. IEEE Transactions on Smart Grid, 2019, 11(2): 1171-1182.
[24] Huang B, Wang J. Applications of physics-informed neural networks in power systems-a review. IEEE Transactions on Power Systems, 2022, 38(1): 572-588.
[25] Cheng L, Wang H, Su Y, et al. Response-Driven Optimal Emergency Control of Power Systems via Deep Learning-Based Sensitivity Embedded Optimization. Energies, 2026, 19(5): 1284.
[26] Shi W, Wu X, Zhang L, et al. Data-Driven and Knowledge-Enhanced Approaches for Real-Time Grid Dispatch: A Comprehensive Survey. 2025 4th Asian Conference on Frontiers of Power and Energy (ACFPE). IEEE, 2025: 244-249.
[27] Gu S, Yang L, Du Y, et al. A review of safe reinforcement learning: Methods, theories, and applications. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 11216-11235.