

Article

Sequential Restoration of Active Distribution Networks via LLM-Guided Optimistic Distributional Reinforcement Learning

Jie Zhang ¹, Yuan Yao ¹, Yan Xu ¹, Junhao Lin ¹, Jiang Zhu ¹ and Jiuxiang Chen ^{2,*}

¹ State Grid Shanghai Municipal Electric Power Company Shibe Electric Power Supply Company, Shanghai 200040, China; 13371890960@126.com (J.Z.); mzd18661393906@163.com (Y.Y.); xuyan2818@icloud.com (Y.X.); ljh7608@163.com (J.L.); 13310022406@163.com (J.Z.)

² School of Electrical Engineering and Automation, Wuhan University, Wuhan 430072, China

* Correspondence: jxchen@whu.edu.cn

Abstract

Data-driven reinforcement learning (RL) often struggles to process unstructured human semantics during distribution network restoration. To address this challenge, this paper proposes a method integrating large language model (LLM)-based semantic space guidance with optimistic distributional RL (DistRL). A fine-tuned LLM parses complex human semantics into dynamic semantic observations and action mask constraints. Concurrently, an optimistic implicit quantile network (O-IQN) employs a high-quantile optimistic preference mechanism to guide the agent toward restoration paths with higher rewards and success rates. The method is validated under random N-1, N-2 and N-3 contingencies on modified 70-bus system and 136-bus systems. Compared with baselines, it accelerates convergence and enhances cumulative rewards. Comprehensive evaluations across 50,000 scenarios on the 70-bus system and 300,000 scenarios on the 136-bus system show that O-IQN consistently ensures high recovery rates. This approach effectively integrates unstructured human semantics with data-driven control, enhancing the robustness and safety of distribution network restoration.

Keywords: load restoration; reinforcement learning; large language model; contextual Markov Decision Process; optimistic quantile regression

1. Introduction

With the widespread integration of high-penetration distributed energy resources (DERs), active distribution networks (ADNs) are prone to forming multiple islanded sections during outages. This raises the risk of concurrent faults and complicates restoration coordination [1]. Consequently, enhancing the resilience of ADNs has become a core requirement for ensuring reliable power operation [2]. In this context, sequential restoration aims to maximize orderly load recovery by coordinating DERs and switch operations [3]. Essentially, this task is a multi-step, high-dimensional, nonlinear discrete sequential decision-making problem [4].

To address the sequential restoration problem, existing research has widely adopted model-based methods to formulate restoration strategies. For instance, mixed-integer programming has been employed to maximize load recovery under topological and operational constraints [5]. Stochastic programming [6] and robust optimization [7] have also been utilized to mitigate uncertainties in load and renewable generation. Furthermore, integrated frameworks that coordinate damage assessment with maintenance scheduling have



Academic Editor: Herminio Martínez-García

Received: 3 June 2026

Revised: 3 July 2026

Accepted: 10 July 2026

Published: 13 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

attracted considerable attention [8]. However, despite their theoretical rigor, these model-based approaches often struggle with scalability and high computational overhead when facing dynamic topological changes and real-time decision-making requirements. These limitations have prompted the adoption of reinforcement learning (RL) as a promising alternative [9].

In terms of topology awareness and network reconfiguration, RL has demonstrated excellent topological adaptability [10], thereby improving the efficiency of service restoration [11,12]. With the large-scale integration of distributed renewable energy, research has further evolved toward multi-source coordination and resilient restoration. Momen et al. [13] proposed a collaborative restoration scheme based on deep reinforcement learning (DRL) for microgrid formation and mobile emergency generator scheduling. A hierarchical DRL method was introduced in [14] to enhance the distributed photovoltaic accommodation capability during fault restoration. Furthermore, Bahrami et al. [15] achieved dynamic distribution network reconfiguration with embedded AC optimal power flow, accounting for standby power sources and interruptible loads.

However, as a data-driven method, RL struggles to interpret the flexible semantic instructions provided by human dispatchers. Consequently, RL struggles to achieve seamless adaptation in real-world post-disaster distribution network restoration involving both physical grid dynamics and textual commands.

To overcome these limitations, the capability of large language models (LLMs) to extract structured information and align semantics offers a novel approach for mapping unstructured text to power grid operational features. Recent studies have attempted to utilize LLMs to extract text features for grid state estimation [16], or have integrated them as feature extractors into DRL architectures to manage load variations [17]. Moreover, LLMs have been employed to parse natural language instructions to design neural network architectures [18] or formulate reward functions [19]. These explorations indicate that LLMs can effectively serve as a bridge between semantic dispatch information and algorithmic logic, transforming complex dispatch text into structured representations suitable for downstream decision-making.

Distribution network fault restoration operates under physical safety constraints [20] and involves inherent post-fault uncertainties [21]. Integrating LLMs and RL for this task requires ensuring constraint adherence while leveraging the flexibility of restoration strategies. To address these challenges, this paper proposes a two-stage adaptive sequential restoration framework for distribution networks that integrates semantic awareness with distributional reinforcement learning (DistRL). The core contributions of this paper are summarized as follows:

1. The distribution network fault restoration problem is modeled as a contextual Markov Decision Process (Contextual MDP) enhanced by a high-dimensional semantic space. This space augments the state representation and incorporates masking constraints derived from dispatch intentions, which effectively improves both representational capacity and semantic adaptability.
2. A dispatch intent parsing and physical context fusion mechanism based on a fine-tuned LLM is proposed. By leveraging textual descriptions of grid topology and operating states alongside human semantic instructions, it generates unified vector representations from heterogeneous unstructured sources.
3. To effectively balance restoration efficiency and system robustness, an optimistic quantile regression based DistRL algorithm is introduced. It models the uncertainties inherent in both restoration resource availability and dispatcher instructions, enabling the optimization of expected return across quantiles.

2. Problem Formulation

The sequential restoration process of an ADN following faults is a highly nonlinear, spatiotemporally coupled optimization problem under operational and dispatch constraints. Upon occurrence, power supply is typically restored by closing tie switches, integrating DERs, and deploying mobile emergency power sources. Interrupted loads are then progressively reconnected to achieve full recovery. However, when resource availability fluctuates due to real-time operational conditions, the system must adaptively dispatch and dynamically coordinate multi-source assets, as illustrated in Figure 1.

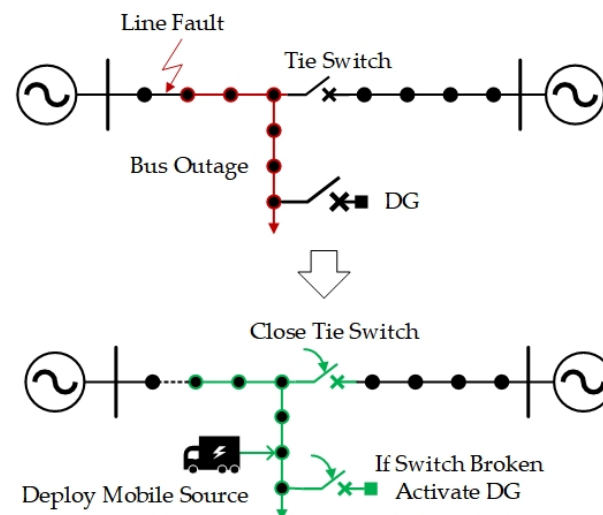


Figure 1. Load restoration with contextual resource availability.

2.1. Mathematical Modeling of Load Restoration

In a distribution network that incorporates load importance indicators, the optimization objective of fault restoration is formulated to maximize the weighted sum of restored loads over a time horizon T :

$$\max \sum_{t=1}^T w^t \cdot (P_{\text{load}}^t - P_{\text{load}}^{t-1}) \quad (1)$$

where P_{load}^t denotes the grid load at time t , and w^t represents the weight coefficient associated with the load change at time t .

During the restoration process, the distribution network must satisfy a set of operational constraints, including power flow equations:

$$\begin{cases} P_i = U_i \sum_j U_j (G_{ij} \cos \theta_{ij} + B_{ij} \sin \theta_{ij}) \\ Q_i = U_i \sum_j U_j (G_{ij} \sin \theta_{ij} - B_{ij} \cos \theta_{ij}) \end{cases} \quad (2)$$

where P_i and Q_i denote the active and reactive power injections at bus i , respectively; U_i represents the voltage magnitude of bus i ; G_{ij} and B_{ij} are the conductance and susceptance of the branch connecting bus i and bus j , respectively; and θ_{ij} represents the voltage phase angle difference between bus i and bus j .

The branches in the network must also satisfy thermal capacity limits:

$$I_{ij} \leq I_{ij}^{\max} \quad (3)$$

where I_{ij} and $I_{ij}^{\max}$ denote the actual current and the maximum current limit of the branch connecting bus i and bus j , respectively.

Furthermore, the bus voltages must remain within acceptable bounds:

$$U_i^{\min} \leq U_i \leq U_i^{\max} \quad (4)$$

where $U_i^{\min}$ and $U_i^{\max}$ represent the minimum and maximum allowable voltage magnitudes of bus i , respectively.

In practical scenarios, natural language dispatch commands d , derived from operational experience, dynamically influence both optimization objectives and system constraints. The original feasible region Ω_{phys} , defined by physical equations and static regulations, is pruned by command d into a refined subspace $\Omega_{\text{legal}} \subseteq \Omega_{\text{phys}}$. This refinement ensures that the decision space aligns with the dispatcher's real-time intent. Furthermore, when intents imply a shift in load restoration priorities, the target weights w^t are dynamically updated to reflect the preferences conveyed by d .

2.2. Contextual MDP Formulation

To capture the semantic intent of dispatchers, the sequential restoration decision-making problem is modeled as a Contextual MDP. It is defined by the tuple (S, C, A, T, R, γ) , where $\gamma \in [0, 1]$ is the discount factor. The remaining elements are defined as follows:

1. **State Space S .** The state space encompasses the real-time operational states and configurations of the distribution network:

$$S = \{S_U, S_I, S_{\text{breaker}}, S_{\text{load}}, S_{\text{fixed}}, S_{\text{movable}}, t\} \quad (5)$$

where S_U represents bus voltages; S_I denotes line currents; S_{breaker} indicates the statuses of tie switches; S_{load} signifies load demands, power supply statuses, and static load weight coefficients; S_{fixed} represents the commitment status of fixed distributed generations (DGs); S_{movable} indicates the access locations of mobile DGs; and t denotes the current time step, with $t = 0$ denoting the fault moment.

2. **Semantic Space C .** For a specific fault restoration task, the semantic vector $c \in C$ characterizes human-specified constraints and operational preferences for current episode. At the beginning of each restoration task, a semantic representation function f maps the initial observation s^0 and the natural language dispatch command d into the semantic vector c :

$$c = f(s^0, d) \quad (6)$$

Specifically, c comprises temporarily disabled tie switches c_{breaker} , temporarily disabled fixed DGs c_{fixed} , and temporarily designated critical loads c_{load} . Accordingly, the semantic space C is defined as:

$$C = \{C_{\text{breaker}}, C_{\text{fixed}}, C_{\text{load}}\} \quad (7)$$

3. **Action Space A .** To achieve load restoration in the distribution network, the action space is formulated as:

$$A = \{A_{\text{breaker}}, A_{\text{load}}, A_{\text{fixed}}, A_{\text{movable}}, A_{\text{none}}\} \quad (8)$$

where A_{breaker} is the tie switch action space corresponding to closing switches; A_{load} is the load pickup action; A_{fixed} is the fixed DG action space corresponding to activating resources; A_{movable} is the mobile DG action space corresponding to dispatching locations; and A_{none} denotes the no-op (null) action.

4. State Transition Probability Distribution T . In ADN fault restoration, the next physical state is uniquely determined by the power flow equations. Therefore, the transition probability satisfies $T(s^{t+1}|s^t, a^t, c) = 1$.
5. Reward Function R . The reward r^t at time step t depends on the previous state s^{t-1} , the previous action a^{t-1} , the current state s^t , and the current task-specific semantic vector c . Aligning with the fault restoration objectives outlined in Section 2.1, the reward function is defined as:

$$r^t = w^t(P_{\text{load}}^t - P_{\text{load}}^{t-1}) + \frac{N_{\text{bus}}^t}{N_{\text{bus}}^{\text{total}}} + r_{\text{base}} \quad (9)$$

where N_{bus}^t represents the number of energized buses in the system at time t , and $N_{\text{bus}}^{\text{total}}$ denotes the total number of system buses. Moreover, if the load restored at time t is a fixed critical load in the power grid, w^t is 5; if the load restored at time t is not a fixed critical load but belongs to C_{load} , w^t is also 5; otherwise, w^t is 1. The second term is introduced to encourage the agent to prioritize energizing unpowered buses. The final term, r_{base} , is a small negative reward applied at each time step, acting as a penalty to incentivize the agent to complete the restoration task efficiently. To guarantee operational safety, this reward r^t is granted only when the system complies with all physical and operational constraints. Conversely, any violation of these constraints incurs a penalty and triggers immediate episode termination.

To incorporate standard grid operating regulations and the semantic constraints in C , the set of legal actions at time t , denoted as $A_{\text{legal}}^t \subseteq A$, is restricted using a mask function M :

$$A_{\text{legal}}^t = \{a \in A \mid M(a, s^t, c) = 1\} \quad (10)$$

The mask function outputs 0 if an action a violates any physical constraints dictated by s^t or semantic boundaries defined by c ; otherwise, it outputs 1.

Specifically, given the current grid state s^t , the masking mechanism prohibits the following: closing a switch when both terminal buses are already energized; connecting loads to de-energized buses; redundant reconnection of healthy loads; and redundant activation of fixed DGs. Concurrently, based on the semantic context c , the mechanism strictly blocks operations on any assets explicitly identified by c_{breaker} and c_{fixed} .

Under this Contextual MDP framework, the agent learns a mapping from the combined state and semantic context space to the action space, yielding a policy $\pi(a^t|s^t, c)$ that strictly satisfies $a^t \in A_{\text{legal}}^t$.

3. Algorithmic Framework

To incorporate semantic information into fault restoration, this paper proposes a two-stage adaptive sequential restoration framework that combines semantic awareness and DistRL. The core logic of this framework is illustrated in Figure 2.

As shown in Figure 2, the framework consists of an initialization stage and an interactive decision-making stage. In the initialization stage, the semantic instruction d and the initial state s^0 are transformed into a prompt and processed by an LLM to interpret action constraints and dynamically prioritize critical loads. This process yields a global semantic vector c generated strictly once per episode. During the interactive stage, the quantile network evaluates return distributions to estimate the maximum expected value, while the mask function M prunes the admissible action set to select the optimal action a^t . This design avoids any online LLM inference, relying solely on environmental interactions. Consequently, the semantic vector c remains strictly static throughout the entire episode, meaning the current framework cannot handle dynamic semantic updates mid-episode.

This is a deliberate architectural trade-off designed to completely eliminate LLM inference latency during the interaction phase, thereby ensuring the real-time execution of the RL.

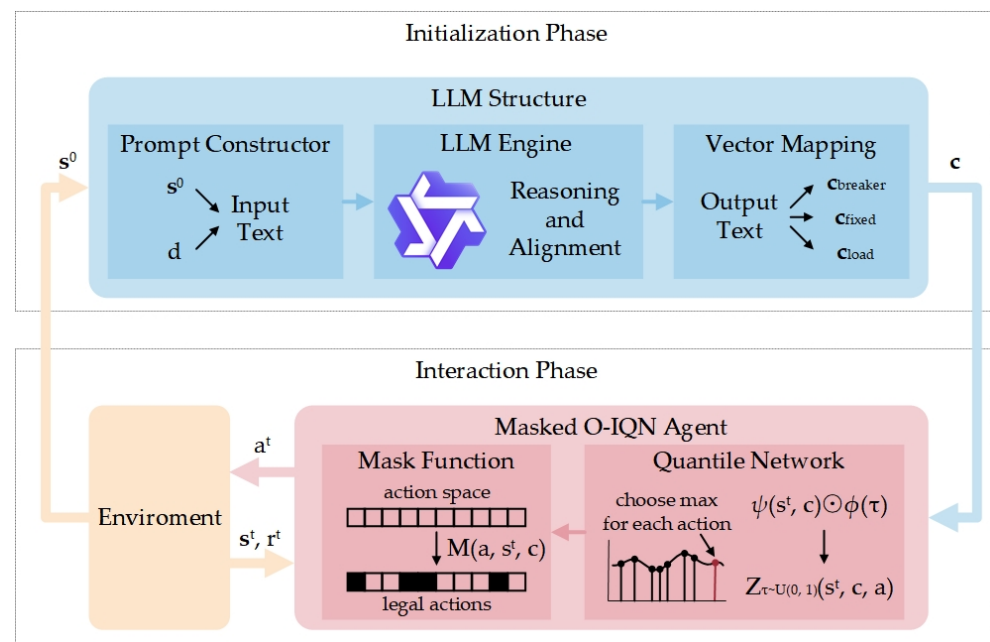


Figure 2. LLM-guided DistRL for semantic-aware restoration.

3.1. LLM-Driven Semantic Vector Extraction and Structuring

To capture the semantic nuances of power restoration, this paper performs domain-specific fine-tuning of Qwen-7B via Low-Rank Adaptation (LoRA) [22]. The fine-tuning utilizes an instruction dataset generated through knowledge distillation from a DeepSeek API, which ensures high-confidence guidance. This process enhances the model's instruction-following capability and enforces constrained output generation that strictly adheres to predefined control specifications.

To ensure the structural consistency of the model's output semantic vector c , a masked cross-entropy loss function is adopted for fine-tuning:

$$L = -\frac{1}{\sum_{i=1}^{N_{\text{token}}} \rho_i} \sum_{i=1}^{N_{\text{token}}} \rho_i \cdot \log P(y_i | x_{<i}) \quad (11)$$

where N_{token} denotes the number of tokens; ρ_i is a mask variable where $\rho_i = 1$ if the i -th token belongs to the response, and $\rho_i = 0$ otherwise; and $P(y_i | x_{<i})$ represents the probability of the model predicting the i -th token as the correct label given the preceding $i - 1$ tokens. Through this approach, Qwen-7B can be fine-tuned solely based on the response tokens. This mechanism forces the model to concentrate its attention on the logical mapping between (s^0, d) and c .

For the original parameters of Qwen-7B, denoted as $W_0 \in \mathbb{R}^{d \times k}$, two low-rank matrices $A \in \mathbb{R}^{r \times k}$ and $B \in \mathbb{R}^{d \times r}$ are introduced through the projection layers to represent the weight increment matrix ΔW , where $r \ll \min(d, k)$. Consequently, the updated weight W satisfies:

$$W = W_0 + \Delta W = W_0 + \frac{\alpha}{r} BA \quad (12)$$

where α is the scaling factor. During initialization, A is initialized using a random Gaussian distribution, and B is initialized to zero, ensuring that the weight increment ΔW is initially zero. Throughout the training process, W_0 remains frozen, and only A and B are optimized.

In the proposed framework, instead of producing high-dimensional embeddings, the fine-tuned Qwen model generates a structured string containing operational status identifiers. A deterministic parser then extracts the specific components c_{breaker} , c_{fixed} , and c_{load} from this string, aggregating them into a unified, fixed-dimensional semantic vector c .

3.2. Optimistic Contextual Implicit Quantile Network

In a Contextual MDP, the agent's objective is to find an optimal policy π^* that maximizes the expected discounted cumulative reward. The optimal policy can be formally expressed as:

$$\pi^* = \arg \max_{\pi} \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r^{t+1} \right] \quad (13)$$

DistRL models the cumulative reward as a random variable $Z(s, c, a)$. Since the true distributional characteristics of Z are unknown, the entire distribution is implicitly characterized by fitting its quantile function F_Z^{-1} . Specifically, a quantile level τ is sampled from a uniform distribution $\text{Unif}(0, 1)$ to compute the corresponding quantile value:

$$z = F_Z^{-1}(\tau, s, c, a) \quad (14)$$

where z denotes the maximum expected return under the τ -th quantile scenario, given that the agent takes action a under state s and semantic vector c . Since $\tau \sim \text{Unif}(0, 1)$, repeatedly sampling τ allows F_Z^{-1} to map the complete probability distribution of the cumulative reward.

The features extracted from s and c are typically high-dimensional, so directly concatenating the scalar τ as a numerical value into the input risks having its information overwhelmed during forward propagation. To address this, a cosine embedding maps τ to a high-dimensional vector $\phi(\tau)$, which is then combined with the joint features $\psi(s, c)$ via a Hadamard product to yield the fused hidden feature h :

$$h = \phi(\tau) \odot \psi(s, c) \quad (15)$$

The fused feature h is fed into the fully connected network, which outputs the estimated cumulative reward z for each action under a specific probability τ . For one (s, c, a) combination, the values of z obtained by repeatedly sampling different τ collectively constitute a discrete estimation of the action-value probability distribution.

During the decision-making phase, the agent selects the optimal action by extracting the upper envelope of the quantile distribution. Specifically, K quantile levels are randomly drawn from $\text{Unif}(0, 1)$ to capture the upper envelope of the quantile distribution, which is then combined with a mask function to select valid actions. This design aims to accelerate exploration efficiency in complex tasks by magnifying extreme positive reward signals:

$$a^* = \arg \max_a \left[M(a, s, c) \cdot \max_{k \in \{1, \dots, K\}} F_Z^{-1}(\tau_k, s, c, a) - (1 - M(a, s, c)) \cdot \infty \right] \quad (16)$$

During the parameter update phase of the neural network, the sampled Temporal Difference (TD) error is first computed as:

$$e_{\tau^t, \tau^{t+1}} = r^t + \gamma F_Z^{-1}(\tau^{t+1}, s^{t+1}, c^{t+1}, a^{t+1}) - F_Z^{-1}(\tau^t, s^t, c^t, a^t) \quad (17)$$

To evaluate the improvement in semantic parsing capability of the fine-tuned LLM and to justify its necessity over conventional approaches, a test set of 600 independent samples is constructed. Among them, 300 are low-complexity texts featuring brief, standard commands, while the other 300 are high-complexity texts characterized by intricate, dense operational narratives with jargon. We compare a conventional rule-based parsing method, the base model (Qwen2.5-7B), and our fine-tuned model under the same hardware environment. The evaluation results are summarized in Table 1.

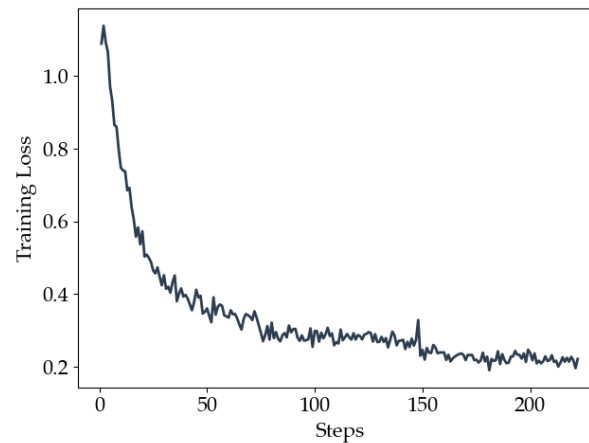


Figure 4. Qwen2.5-7B fine-tuned loss.

As shown in the table, the rule-based approach perfectly extracts standard instructions under low complexity. However, it exhibits a significant performance drop on high-complexity texts, as predefined pattern-matching rules struggle to accommodate unforeseen linguistic variations and complex combinations of power system terminology. Furthermore, the base model achieves only 72.33% valid rate under low complexity, owing to its lack of comprehension of power system terminology and output format specifications. In contrast, the fine-tuned model fails on only a single sample.

Table 1. Semantic parsing accuracy comparison.

Complexity	Approach	Valid	Invalid
Low	Rule-based model	300	0
	Base model	217	83
	Fine-tuned	299	1
High	Rule-based model	66	234
	Base model	204	96
	Fine-tuned model	270	30

Crucially, when facing high-complexity scenarios, the fine-tuned model maintains robust parsing capabilities, successfully interpreting 90% samples, substantially outperforming both the base model and the rule-based approach. These results demonstrate that while conventional parsing methods are sufficient for simple, structured commands, the fine-tuned LLM is indispensable for real-world applications. Fine-tuning not only equips the model with the ability to generate standardized power system terminology, but also significantly deepens its comprehension of complex, unstructured human semantics and scenario-specific constraints that transcend the capabilities of rigid rule-based frameworks.

4.2. Load Restoration Experiments

Random N-1/N-2/N-3 faults are applied to the modified 70-bus system. The agent's decision sequence is limited to 50 steps; otherwise, the task fails. The agent interacts with the environment 1,000,000 times. For every 10 training sessions of the neural network, 40 scenarios are sampled for evaluation. DQN [23], Double DQN [24], Dueling DQN [25], QRDQN [26], IQN [27], A2C [28] and PPO [29] serve as the baselines. The key hyperparameters of RL are shown in Table A1. Figure 5 illustrates the training reward curves of O-IQN and the baselines.

All algorithms exhibit a similar upward trend in the initial training phase. However, DQN and Dueling DQN subsequently suffer severe performance collapse, converging to a low-value interval between -0.4 and -0.5 . In contrast, O-IQN, A2C, PPO, QRDQN, and Double DQN maintain stable high rewards. Notably, the reward curve of O-IQN consistently remains at a high level, converging near 1.1.

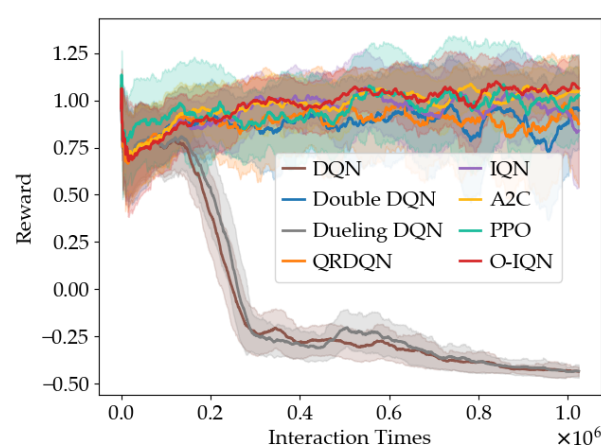


Figure 5. Training rewards in modified 70-bus system for each algorithm.

Figure 6 further presents the restoration success rates during the evaluation phase. The overall performance and convergence trends across all algorithms are highly consistent with those observed in Figure 5. O-IQN delivers performance on par with A2C and PPO, clearly outperforms Double DQN and QRDQN, and shows a slight improvement over IQN. It stabilizes at a high success rate of 0.8–0.9 after 300,000 interactions. These results confirm its competitive policy learning efficiency and decision-making robustness, matching the level of leading on-policy methods while surpassing the value-based baselines.

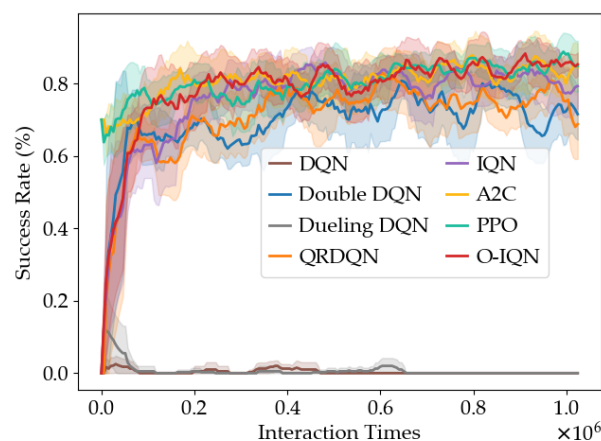


Figure 6. Evaluating success rate in modified 70-bus system for each algorithm.

To comprehensively evaluate performance, the converged models are tested across 50,000 scenarios, as summarized in Table 2. The action efficiency is newly introduced and defined as the ratio of the number of disconnected loads to the total number of interaction steps within a single episode. This metric reflects the agent's ability to complete the restoration process with minimal redundant commands. O-IQN substantially outperforms the base-lines across all core metrics. Notably, while A2C and PPO achieve competitive success rates, their significantly lower action efficiencies indicate a tendency to execute unnecessary or suboptimal step sequences. In contrast, O-IQN achieves the highest action efficiency 86.73% alongside a superior success rate 86.32% and total reward 1.0758. This demonstrates that O-IQN more accurately models and evaluates the action-value distribution, effectively minimizing redundant operations and enhancing overall fault restoration efficiency while ensuring a highly reliable system recovery.

Table 2. Testing results in modified 70-bus system for chosen algorithms.

Algorithm	Reward	Success Rate	Action Efficiency
O-IQN	1.0758	0.8632	0.8673
Double DQN	0.9615	0.7966	0.7616
QRDQN	0.9267	0.7814	0.6101
IQN	0.9887	0.8114	0.7260
A2C	1.0508	0.8605	0.6585
PPO	1.0388	0.8418	0.7954

Figure 7 compares the algorithmic robustness across varying fault scales. With fewer loads to restore, all algorithms maintain a load restoration rate above 90%. However, as the number of de-energized loads increases, the performance of QRDQN and PPO exhibit notable performance degradation while Double DQN and IQN displaying significant volatility. Conversely, O-IQN and A2C remain relatively stable. Moreover O-IQN maintains a minimum average load restoration rate of 87.80%, which is still higher than A2C's average rate of 84.40%. Its curve consistently remains at the top with a narrower shaded area, indicating smaller decision variance and higher stability.

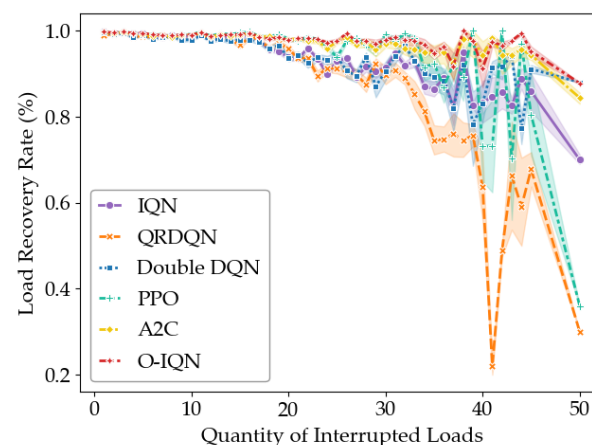


Figure 7. Load recovery rate for different interrupted load quantities in modified 70-bus system.

4.3. Ablation Experiments

To dissect the impact of the optimistic decision mechanism, we depict the kernel density estimation (KDE) distributions of recovery rates for IQN and O-IQN in Figure 8.

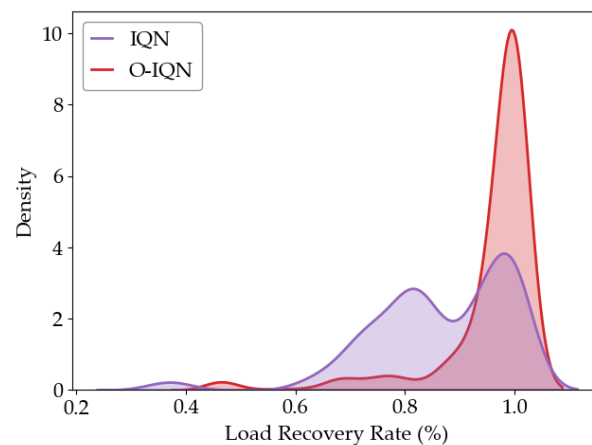


Figure 8. KDE of recovery rate in modified 70-bus system (>35 interrupted loads).

In high-stress conditions with more than 35 loads to restore, the density of IQN exhibits a highly dispersive and bimodal distribution with a pronounced left-skewed heavy tail. This profile reveals that standard value estimation tends to undervalue potentially optimal policies under severe uncertainty, resulting in cautious and suboptimal action selections. Conversely, O-IQN utilizes the maximization operator to actively steer the policy toward high-reward regions. Crucially, this optimistic drive is rendered both safe and effective because of the mask function. Capitalizing on this constrained action space, O-IQN converts the dual-peaked pattern into a sharp, concentrated peak centered near the optimal recovery rate, significantly reducing the probability of low-recovery outcomes in challenging test cases.

To investigate how the semantic space affects RL policy training and convergence, three configurations are evaluated: introducing only c_{load} (no LLM mask), introducing only c_{breaker} and c_{fixed} (no LLM obs), and omitting the semantic space entirely (no LLM).

Figure 9 shows the training rewards under different modes. Without the semantic space (no LLM), the agent cannot perceive external semantics, causing the reward to fluctuate severely at a low level without converging. When only c_{load} is used as a supplementary observation, the performance drops sharply, which confirms that semantic action masks are critical for eliminating invalid action spaces and accelerating exploration. Using only c_{breaker} and c_{fixed} as masks maintains baseline performance. However, the lack of perception of the spatial distribution of temporary critical loads traps the policy in local sub-optimal solutions, capping the reward upper bound. Therefore, full semantic space integration overcomes the semantic bottlenecks of purely data-driven RL in dynamic, human-intervened scenarios, significantly improving the agent's decision logic and restoration efficiency.

4.4. Scalability Analysis

A more complex 136-bus system [30] is built to evaluate the scalability of O-IQN. Double DQN, QRDQN, IQN, A2C and PPO were chosen as baselines. All algorithms were retrained on this larger-scale scenario to ensure that the model fully adapts to the increased topological complexity.

As shown in Figures 10 and 11, both results demonstrate that O-IQN exhibits exceptional robustness, matching the top-tier performance of PPO and A2C while significantly outperforming the Double DQN and QRDQN baselines. Specifically, O-IQN achieves rapid convergence to a high reward plateau of approximately 1.6 and consistently maintains this peak without the catastrophic degradation witnessed in its DQN-based counterparts during the latter half of the training. In terms of success rate, the algorithm demonstrates superior

sample efficiency by quickly exceeding 90% and sustaining this high level, whereas the competing DQN variants suffer from severe oscillation and notably lower final rates.

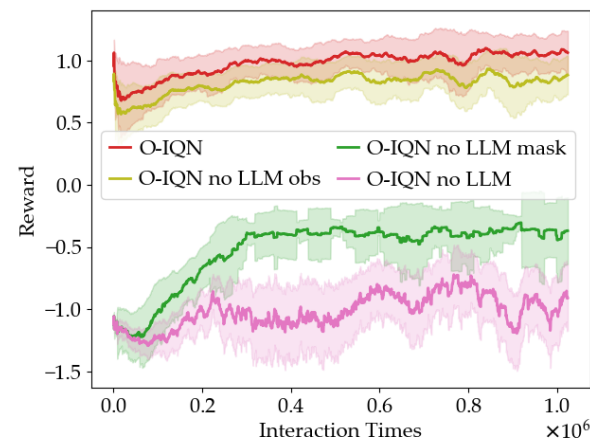


Figure 9. Training rewards with different semantic space in modified 70-bus system.

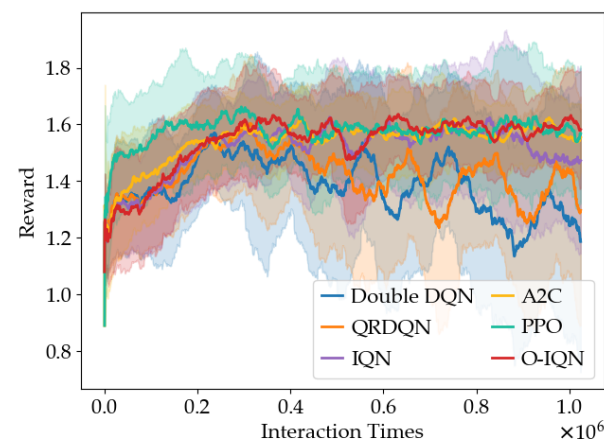


Figure 10. Training rewards in modified 136-bus system.

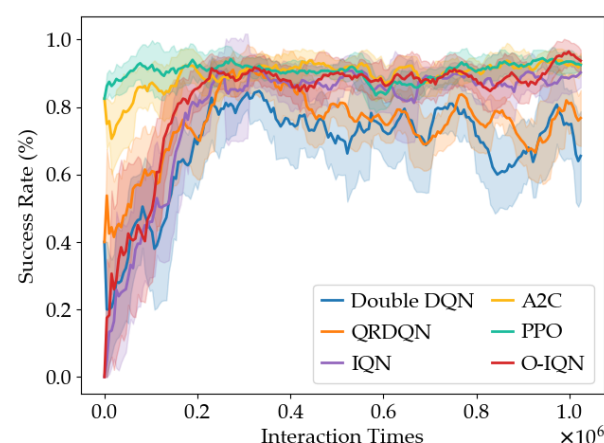


Figure 11. Evaluating success rate in modified 136-bus system.

We evaluated the trained models on 300,000 test samples divided into three subgroups based on interrupted loads (1–20, 21–35, and 36+). As shown in Figure 12, all algorithms achieve a high recovery rate centered around 1.0 across all groups. However, a distinct divergence emerges in worst-case performance. While all methods effectively handle minor interruptions (1–20 group), O-IQN exhibits a performance edge in the 21–35 group by

maintaining a higher lower bound. Most notably, in the challenging 36+ group, O-IQN significantly outperforms other methods by sustaining a much higher minimal recovery rate under severe contingencies.

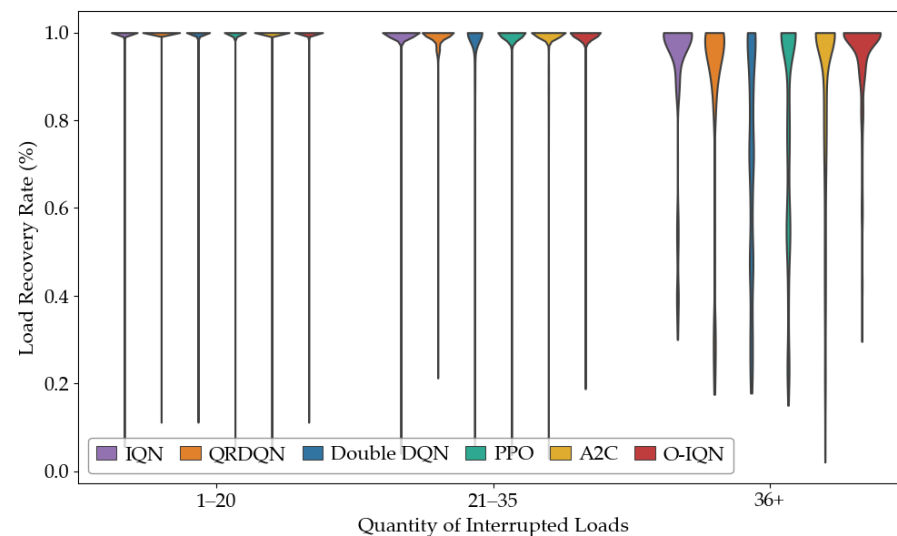


Figure 12. Violin plot of testing results in modified 136-bus system.

5. Conclusions

This paper proposes a distribution network restoration method combining LLM semantic space guidance with O-IQN. It maps human dispatch instructions into semantic space, which enhances the state space and mask function. Furthermore, the integrated high-quantile optimistic preference mechanism efficiently overcomes traditional policy conservatism. This mechanism guide the RL toward globally optimal restoration paths. Consequently, the proposed framework significantly mitigates the risk of catastrophic restoration failures under extreme fault scenarios. It also demonstrates exceptional robustness and scalability across different network scalability.

Future work will focus on incorporating online LLM inference and accounting for the complex spatiotemporal coupling factors inherent in sequential restoration. These efforts aim to further enhance the practicality and robustness of LLM-guided decision-making in real-world power systems.

Author Contributions: Conceptualization, J.Z. (Jie Zhang) and Y.Y.; methodology, J.Z. (Jie Zhang) and J.C.; software, Y.Y. and Y.X.; validation, J.Z. (Jie Zhang) and Y.X.; formal analysis, J.L.; investigation, J.L.; resources, J.Z. (Jiang Zhu); data curation, J.Z. (Jiang Zhu); writing—original draft preparation, J.Z. (Jie Zhang); writing—review and editing, J.C.; visualization, Y.Y.; supervision, J.L.; project administration, Y.X.; funding acquisition, J.Z. (Jiang Zhu). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by State Grid Shanghai Municipal Electric Power Company Technology Project: “Research on Human-in-the-Loop Hybrid Augmented Intelligence Technology for Load Transfer Decision Generation in Distribution Networks” (520914250001).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: During the preparation of this study, the authors used DeepSeek-V4-Flash for the purposes of translating the main text and generating the fine-tuned data for Qwen2.5-7B model. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: Authors Jie Zhang, Yuan Yao, Yan Xu, Junhao Lin and Jiang Zhu were employed by State Grid Shanghai Municipal Electric Power Company Shibe Electric Power Supply Company. The authors declare that this study received funding from the State Grid Shanghai Municipal Electric Power Company. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A

Table A1. RL hyperparameters.

Parameter	Value
total steps	1,000,000
hidden size	512
layer number	2
learning rate	1×10^{-4}
γ	0.99
start epsilon	1
minimum epsilon	0.02
off-policy batch size	128
on-policy batch size	256
K in 70-bus system	64
K in 136-bus system	128
dimension of $\phi(\tau)$	64

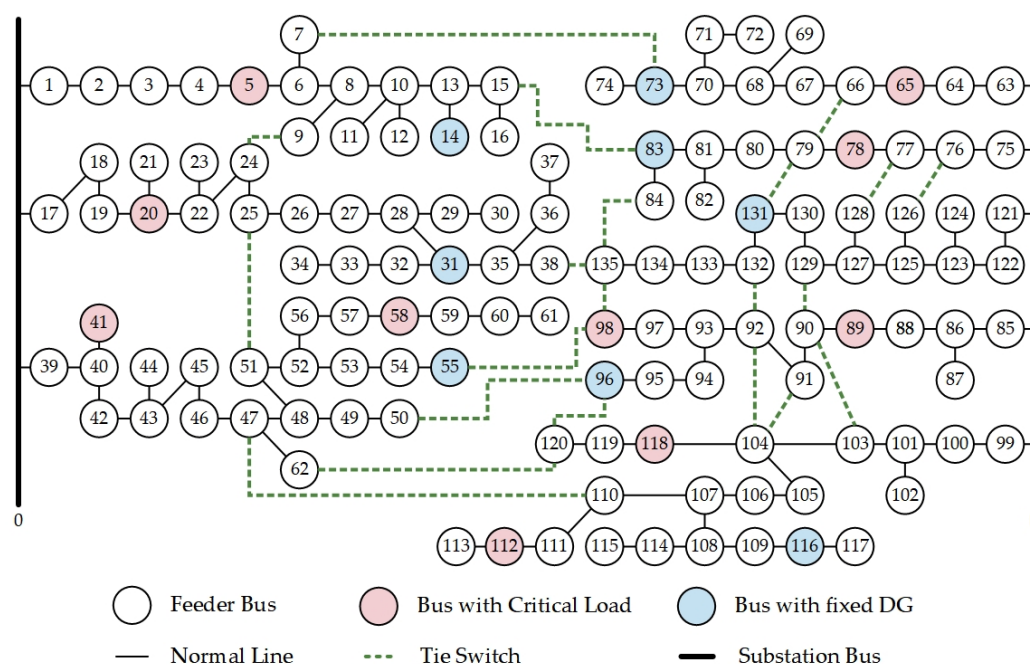


Figure A1. Modified 136-bus system.

References

1. Poudel, S.; Dubey, A.; Schneider, K.P. A Generalized Framework for Service Restoration in a Resilient Power Distribution System. *IEEE Syst. J.* **2022**, *16*, 252–263.
2. Jalilian, A.; Taheri, B.; Molzahn, D.K. Co-Optimization of Damage Assessment and Restoration: A Resilience-Driven Dynamic Crew Allocation for Power Distribution Systems. *IEEE Trans. Power Syst.* **2025**, *40*, 676–688.
3. Arif, A.; Cui, B.; Wang, Z. Switching Device-Cognizant Sequential Distribution System Restoration. *IEEE Trans. Power Syst.* **2022**, *37*, 317–329.
4. Zhang, M.; Cai, S.; Xie, Y.; Zhou, B.; Zheng, W.; Wu, Q.; Wen, J. Supply Resilience Constrained Scheduling of MEGs for Distribution System Restoration: A Stochastic Model and FW-PH Algorithm. *IEEE Trans. Smart Grid* **2025**, *16*, 194–208.

5. Chen, B.; Ye, Z.; Chen, C.; Wang, J. Toward a MILP Modeling Framework for Distribution System Restoration. *IEEE Trans. Power Syst.* **2019**, *34*, 1749–1760.
6. Cai, S.; Xie, Y.; Zhang, M.; Jin, X.; Wu, Q.; Guo, J. A Stochastic Sequential Service Restoration Model for Distribution Systems Considering Microgrid Interconnection. *IEEE Trans. Smart Grid* **2024**, *15*, 2396–2409.
7. Cai, S.; Xie, Y.; Wu, Q.; Xiang, Z.; Jin, X.; Zhang, M. Robust Coordination of Multiple Power Sources for Sequential Service Restoration of Distribution Systems. *Int. J. Electr. Power Energy Syst.* **2021**, *131*, 107068.
8. Chen, B.; Ye, Z.; Chen, C.; Wang, J.; Ding, T.; Bie, Z. Toward a Synthetic Model for Distribution System Restoration and Crew Dispatch. *IEEE Trans. Power Syst.* **2019**, *34*, 2228–2239.
9. Zhao, J.; Li, F.; Mukherjee, S.; Sticht, C. Deep Reinforcement Learning-Based Model-Free On-Line Dynamic Multi-Microgrid Formation to Enhance Resilience. *IEEE Trans. Smart Grid* **2022**, *13*, 2557–2567.
10. Guo, Z.; Lai, J.; Kou, X. Intelligent Distribution Network Active Protection and Fault Recovery Mechanism Based on DRL and Graph Optimization. *Electr. Power Syst. Res.* **2026**, *254*, 112623.
11. Yin, Z.; Fang, C.; Fang, Y.; Xie, M. A Variational Bayesian Deep Reinforcement Learning Approach for Resilient Post-Disruption Recovery of Distribution Grids. *Reliab. Eng. Syst. Saf.* **2025**, *257*, 110840.
12. Jo, S.; Oh, J.Y.; Yoon, Y.T.; Jin, Y.G. Self-Healing Radial Distribution Network Reconfiguration Based on Deep Reinforcement Learning. *Results Eng.* **2024**, *22*, 102026.
13. Momen, H.; Jadid, S. Resilience Enhancement of Power Distribution System Using Fixed and Mobile Emergency Generators Based on Deep Reinforcement Learning. *Eng. Appl. Artif. Intell.* **2024**, *137*, 109118.
14. Wu, Y.; Li, Q.; Qiu, J. Hierarchical Deep Reinforcement Learning for Dynamic Reconfiguration of Photovoltaic Integration in Distribution Network. *Electr. Power Syst. Res.* **2025**, *246*, 111587.
15. Bahrami, S.; Chen, Y.C.; Wong, V.W.S. Dynamic Distribution Network Reconfiguration with Generation and Load Uncertainty. *IEEE Trans. Smart Grid* **2024**, *15*, 5472–5484.
16. Li, Y.; Cheng, G.; Zhao, J.; Liu, Y. In-Context Learning Enhanced Large Language Model for Robust Distribution System State Estimation. *Appl. Energy* **2026**, *407*, 127344.
17. Yin, L.; Wei, K. Large Language Models Driven Adaptive Proximal Policy Optimization for Clean Energy System Frequency Stability Control. *Sustain. Energy Technol. Assess.* **2025**, *84*, 104739.
18. Deng, K.; Zhou, Y.; Zeng, H.; Wang, Z.; Guo, Q. Power Grid Model Generation Based on the Tool-Augmented Large Language Model. *IEEE Trans. Power Syst.* **2025**, *40*, 5487–5490.
19. Yan, Z.; Xu, Y. Real-Time Optimal Power Flow with Linguistic Stipulations: Integrating GPT-Agent and Deep Reinforcement Learning. *IEEE Trans. Power Syst.* **2024**, *39*, 4747–4750.
20. Gholizadeh, N.; Musilek, P. Explainable Reinforcement Learning for Distribution Network Reconfiguration. *Energy Rep.* **2024**, *11*, 5703–5715.
21. Xie, H.; Tang, L.; Zhu, H.; Cheng, X.; Bie, Z. Robustness Assessment and Enhancement of Deep Reinforcement Learning-Enabled Load Restoration for Distribution Systems. *Reliab. Eng. Syst. Saf.* **2023**, *237*, 109340.
22. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. In Proceedings of the International Conference on Learning Representations, Virtual, 25–29 April 2022.
23. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; Riedmiller, M. Playing Atari with Deep Reinforcement Learning. In Proceedings of the NIPS Deep Learning Workshop, Lake Tahoe, NV, USA, 9–10 December 2013.
24. van Hasselt, H.; Guez, A.; Silver, D. Deep reinforcement learning with double Q-Learning. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*; AAAI Press: Menlo Park, CA, USA, 2016; AAAI'16; pp. 2094–2100.
25. Wang, Z.; Schaul, T.; Hessel, M.; Van Hasselt, H.; Lanctot, M.; De Freitas, N. Dueling network architectures for deep reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning—Volume 48*; JMLR.org: Norfolk, MA, USA, 2016; ICML'16; pp. 1995–2003.
26. Dabney, W.; Rowland, M.; Bellemare, M.G.; Munos, R. Distributional reinforcement learning with quantile regression. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*; AAAI Press: Menlo Park, CA, USA, 2018.
27. Dabney, W.; Ostrovski, G.; Silver, D.; Munos, R. Implicit Quantile Networks for Distributional Reinforcement Learning. In *Proceedings of the 35th International Conference on Machine Learning*; Proceedings of Machine Learning Research; Dy, J., Krause, A., Eds.; PMLR: Cambridge, MA, USA, 2018; Volume 80, pp. 1096–1105.
28. Mnih, V.; Badia, A.P.; Mirza, M.; Graves, A.; Lillicrap, T.P.; Harley, T.; Silver, D.; Kavukcuoglu, K. Asynchronous Methods for Deep Reinforcement Learning. In *Proceedings of the Deep Reinforcement Learning Workshop*; NIPS: Grenada, Spain, 2016.

29. OpenAI. *Proximal Policy Optimization (PPO) in OpenAI Spinning Up*; OpenAI: San Francisco, CA, USA, 2018. Available online: <https://spinningup.openai.com/en/latest/algorithms/ppo.html> (accessed on 25 June 2026).
30. Ma, L.; Wang, L.; Liu, Z. Soft Open Points-Assisted Resilience Enhancement of Power Distribution Networks Against Cyber Risks. *IEEE Trans. Power Syst.* **2023**, *38*, 31–41.

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