EAI Endorsed Transactions
on Energy Web
Research Article EAI.EU

Deep Reinforcement Learning-Based Coordinated Voltage Control of HST and DG

Hongliang Peng[1], Ruotian Yao[2], Riguang Huang[2], Shiqi Jiang[2,*], Xujun Wang[1]

[1]Huizhou Power Supply Bureau of Guangdong Power Grid Co., Ltd., Huizhou, 516000, China
[2]China Southern Power Grid Electric Power Research Institute Co., Ltd., Guangzhou, 510000, China

Abstract

With the increasing penetration of distributed generation in active distribution networks, voltage fluctuation and voltage violation problems have become more prominent, especially in weak grids. Conventional voltage control methods based on empirical parameter tuning or single-device regulation often have limited adaptability under time-varying load conditions and fluctuating distributed generation output. To address these issues, this paper proposes a deep reinforcement learning based coordinated voltage control method for HST and DG. The proposed method takes monitored node voltages, load levels, and distributed generation outputs as state inputs and uses the gain adjustments of HST and DG as control actions, thereby achieving adaptive optimization of voltage response through a continuous decision-making framework. On this basis, an experimental validation framework including training and validation datasets, multi-strategy comparison communication-constrained analysis, and repeated-run statistics is established to evaluate the control performance generalization capability and robustness of the proposed method. The results show that the proposed strategy outperforms the baseline method in voltage deviation, overshoot oscillation energy, and control effort, while maintaining good performance consistency in the validation scenario. Under mild communication constraints, the proposed strategy still exhibits acceptable adaptability and operational stability. These findings indicate that deep reinforcement learning provides an effective optimization approach for coordinated HST and DG control and offers a useful reference for voltage regulation in active distribution networks.

Keywords: distribution network voltage control, deep reinforcement learning, coordinated control, distributed generation

Received on 12 January 2026 accepted on 11 March 2026, published on 16 September 2026

doi: 10.4108/ew.14203

1. Introduction

With the increasing penetration of distributed generation in distribution networks, the traditional unidirectional power flow pattern is gradually changing, and voltage fluctuation, local overvoltage, and voltage violation have become more prominent. In weak grids, these issues are usually more sensitive, and the variations in load demand and DG output further increase the difficulty of voltage regulation. Therefore, maintaining voltage stability under complex and time-varying operating conditions has become one of the key issues in active distribution network operation and control.

Existing studies on distribution network voltage control have developed several technical routes, including local control, distributed control, and coordinated control [1, 2]. Conventional control methods based on empirical parameter tuning are simple to implement, but their adaptability is often limited when facing time-varying loads and fluctuating DG outputs. Model-based optimization methods can improve control performance, yet they usually depend on offline tuning or strong modeling assumptions, and their online responsiveness and real-time applicability remain constrained [3]. In addition, voltage regulation dominated by a single device may be

*Corresponding author. Email: 16676676716@163.com

effective in specific local scenarios, but it is often difficult to balance overall voltage quality and control effort at the network level [4].

In recent years, deep reinforcement learning has been gradually introduced into voltage control problems because of its weak dependence on accurate models, fast online decision-making capability, and adaptability to complex operating scenarios [5, 6]. Existing studies have shown that DRL has considerable potential in dealing with model uncertainty, topology variation, and multi-period operating conditions [7]. However, most available works focus on inverter-based Volt/VAR control, distributed energy storage, or general voltage regulation frameworks, while studies specifically addressing HST-DG coordinated control remain limited. At the same time, communication delay, packet loss, and bandwidth limitation can directly affect the performance of cooperative voltage control and may even lead to control degradation under high renewable penetration. Therefore, the adaptability of control strategies under communication-constrained conditions also requires further investigation [8].

Motivated by the above observations, this paper proposes a deep reinforcement learning-based coordinated control method for voltage regulation in a distribution network with HST and DG. The proposed framework takes node voltage, load, and DG output as the state input, and uses the gain adjustments of HST and DG as the control actions to construct an HST-DG coordinated control strategy. The method is evaluated under training and validation scenarios using publicly available real-world power system time-series data from Open Power System Data (OPSD). In addition, communication constraint analysis, multi-strategy comparison, and repeated-run statistical analysis are further carried out to systematically verify the effectiveness, generalization capability, and robustness of the proposed method.

The main contributions of this paper are summarized as follows.

(1) A DRL-based voltage regulation framework is established for the HST-DG coordinated control task, in which nodal operating states and controller gain adjustments are unified in a continuous decision process.

(2) A complete experimental framework, including training and validation datasets, communication-constrained scenarios, and multi-strategy comparisons, is built to assess the proposed method under different operating conditions.

(3) Repeated experiments and statistical tests are conducted to verify the stability of the proposed strategy and to improve the consistency and repeatability of the conclusions.

2. Related Work

2.1. Voltage Control Methods in Distribution Networks

Research on voltage control in distribution networks has developed along several technical routes, including local control, hierarchical control, and coordinated control. With the increasing penetration of photovoltaics and energy storage, traditional local regulation driven by fixed rules can no longer simultaneously address regional autonomy, multi-timescale operation, and multi-device coordination. As a result, recent studies have gradually shifted from single-device regulation to zonal control, multi-timescale scheduling, and hierarchical coordination frameworks [9-11].

On this basis, coordinated voltage control in active distribution networks has further evolved into distributed robust control, hierarchical energy-storage-assisted control, and data-driven voltage regulation. Such approaches enhance network-level controllability and adaptability to some extent. However, most of them still rely on sensitivity information, forecasting inputs, or manually defined partitions, and their effectiveness still depends heavily on the precision of the model and changing operating conditions. Real-time deployment and system scalability also remain difficult to balance in complex networks with rapid fluctuations [12-14].

2.2. Applications of Intelligent Optimization in Distribution Networks

To overcome the dependence of traditional voltage regulation on empirical parameter tuning, intelligent optimization methods were introduced early into voltage and reactive power control. Existing studies mainly employ robust, two-stage, and distributed optimization schemes to coordinate multiple devices under operational constraints. These approaches perform well in improving optimization quality, but most of them remain offline or quasi-online in nature, and the associated computational cost restricts their use in fast-varying scenarios [15-17].

With the development of data-driven methods, data-driven approaches such as reinforcement learning, especially deep reinforcement learning, are now widely used in Volt/VAR control, distributed energy storage regulation, and multi-agent coordinated voltage control. Existing studies show that DRL can learn control policies under weak model dependence and can handle high-dimensional states, continuous actions, and multi-device coordination. At the same time, a number of challenges still exist in the literature. Many studies still focus on generic Volt/VAR settings rather than specific coordinated device structures, and a considerable part of the literature emphasizes policy optimization itself while paying less attention to large-scale deployment, communication uncertainty, and cross-scenario stability [18-22].

2.3. Communication Constraints and Control Strategies

As voltage regulation evolves from local adjustment to multi-device coordination, communication delay, packet loss, and bandwidth limitation are gradually becoming key factors affecting control performance. Existing studies indicate that communication uncertainty may alter the timeliness of state feedback and the execution of control commands, thereby affecting system response speed, steady-state quality, and closed-loop stability of distributed voltage control [23-25].

To address these issues, previous research has explored online control, event-triggered control, and safe reinforcement learning methods aimed at reducing communication burden, mitigating delay effects, or improving control safety and system robustness. Nevertheless, communication-constrained DRL-based voltage control is still not sufficiently investigated in an integrated manner, particularly under specific coordinated control structures [26].

2.4. Research Gaps and the Position of This Work

Taken together, existing voltage control methods are mature in engineering implementation, but their parameters are typically adjusted empirically, and their capability to handle time-varying load and distributed generation remains limited. Intelligent optimization improves multi-objective regulation capability, yet many studies still emphasize offline or scenario-specific optimization. Deep reinforcement learning offers a new route for online coordinated control, but most current works concentrate on general Volt/VAR problems and still lack systematic analyses of device-specific coordination, communication-constrained adaptability, and repeated-run stability [27, 28].

Motivated by the above-mentioned limitations, this study focuses on DRL-based optimization for the HST-DG coordinated control problem and performs a comprehensive investigation through training and validation, communication-constrained scenarios, multi-strategy comparison, and repeated-run statistical analysis. Compared with the existing literature, this work gives greater attention to the control specificity of a particular coordinated structure and the completeness in validation experiments.

3. Methods

3.1. Overall Experimental Design

In order to assess the proposed deep reinforcement learning-based HST-DG coordinated control method in an organized way, this study develops a unified testing framework for voltage regulation in distribution systems. In this part, the data arrangement, the compared control strategies, the general experimental procedure, and the additional validation settings. These contents serve as the foundation for the following sections on system modeling, control design, and result analysis.

The experiment uses 48 hours of sequential temporal data derived from the Open Power System Data (OPSD) time series package as the input for the distribution network simulation setup. OPSD provides publicly available real-world power system time-series data, including electricity consumption (load) and renewable generation profiles. In this study, the selected load and renewable generation series are used to construct the time-varying operating conditions of the distribution network environment. The dataset is divided in chronological order into separate training and validation subsets. The first 24 hours are allocated for model training, whereas the remaining 24 hours are reserved for independent validation. This arrangement decouples the learning phase from the performance assessment stage in the time dimension and makes it possible to examine whether the learned control policy can adapt to operating conditions that were not present during training. The training subset supports agent–environment interaction, whereas the validation subset serves to evaluate the response characteristics and control stability of the obtained policy under new operating scenarios.

Four representative control strategies are considered for comparison in this study.

(1) Baseline strategy. This strategy adopts empirical or manually tuned parameters and serves as a reference for the basic performance of the system under conventional control settings.

(2) HST-dominant strategy. This strategy assigns a dominant role to HST in the voltage regulation process and is used to analyze the control behavior when HST undertakes the main adjustment task.

(3) DG-dominant strategy. Here, the approach highlights the regulating role of distributed generation and is used to evaluate the control effect when DG provides the primary support.

(4) DRL strategy. This approach adopts deep reinforcement learning to tune the coordinated control parameters of HST and DG, with the aim of exploiting the potential of joint regulation across different operating scenarios.

The overall experimental procedure includes data input, control execution, state observation, index calculation, and result comparison. The load profile and distributed-generation-related profile obtained from the OPSD-based time-series dataset are first injected into the distribution network simulation environment to form the operating condition at each time step. The selected control strategy then adjusts HST and DG according to the current system state. During the control process, voltage responses at the monitored nodes and other relevant operating variables are recorded. Based on these observations, performance indices such as steady state error, overshoot, oscillation energy, and control power consumption are calculated. The

performance of different control strategies is then compared under a consistent evaluation framework. This procedure forms the basic route of the experimental study.

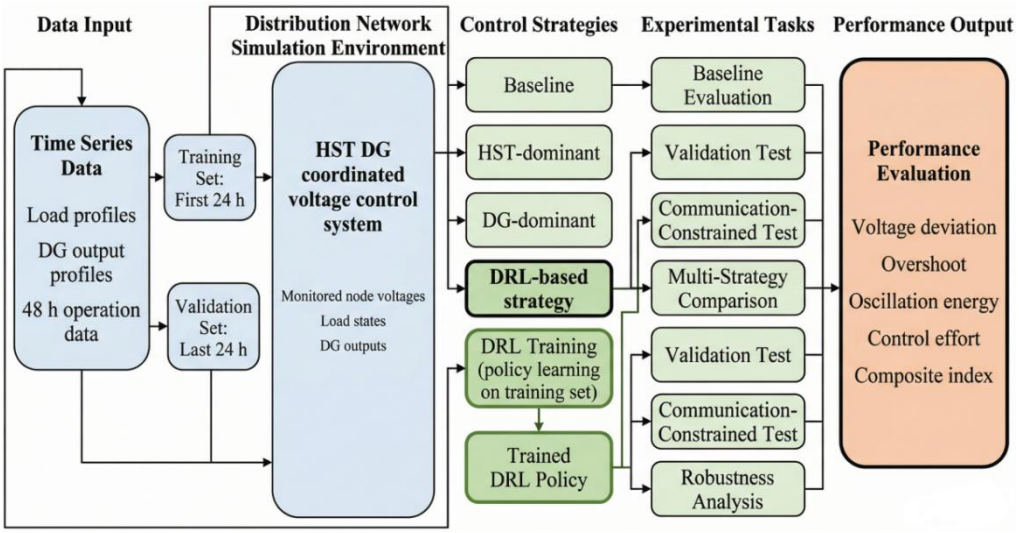

Figure 1. Overall experimental flowchart

As illustrated in Figure 1, the overall experimental design consists of policy training, policy validation, and extended testing. The training stage is based on the training set and learns the control policy through repeated agent–environment interaction. The validation stage uses the trained policy on an independent validation set in order to examine its control behavior under unseen operating conditions. On this basis, communication-constrained cases and repeated trials scenarios are further introduced to assess the applicability and stability of the proposed method in more complex scenarios.

Figure 1 is intended to present the overall structure of the experimental design, including the dataset partition, the compared control strategies, the performance evaluation process, and the extended validation settings. This figure is not used to present experimental findings. Its role is to clarify the logical relation among the main experimental components and to provide a unified framework for the subsequent sections. The later analyses of baseline performance, DRL training and validation, communication constraints, and robustness are all developed within this framework.

In addition to routine training and validation, two extended validation settings are considered.

(1) Communication constraint analysis. Delay, packet loss, and bandwidth limitation are introduced into the control loop to investigate the ability of the control strategy to sustain system performance under imperfect communication conditions.

(2) Repeated run robustness analysis. Multiple independent runs with different initial training conditions are carried out, and the key performance indices are statistically analyzed to examine the robustness and consistency of the obtained results under different random initializations.

Overall, the experimental design combines training, validation, and extended analysis under a unified model, a consistent data partition scheme, and a common set of performance indices. This design provides a clear structural basis for the subsequent methodological description and result analysis, and it also offers a complete experimental framework for assessing the effectiveness of deep reinforcement learning in HST-DG coordinated control.

3.2. System Model

3.2.1. Distribution Network Node and Device Model

The system considered in this study is a distribution network with HST units and distributed generation. According to their functions, the network buses can be classified into source buses, load buses, DG connection buses, and monitored buses. The monitored buses are used to extract key voltage responses during control, and the buses that are more sensitive to power disturbances are treated as weak buses. Let $\mathcal{N}$ denote the set of all buses and $\mathcal{N}_m$ denote the set of monitored buses. For any bus $i \in \mathcal{N}$, the voltage magnitude is denoted by V_i. The active and reactive load powers are denoted by P_i^L and Q_i^L. The DG injected powers are denoted by P_i^{DG} and Q_i^{DG}. The equivalent regulating powers of HST are denoted by P_i^H and Q_i^H.

The nodal power balance is written as

$$P_i^{inj} = P_i^{DG} + P_i^H - P_i^L \quad (1)$$

$$Q_i^{inj} = Q_i^{DG} + Q_i^H - Q_i^L \quad (2)$$

where P_i^{inj} and Q_i^{inj} are the net active and reactive power injections at bus i. These expressions indicate that the voltage state of each bus is jointly determined by the load demand, DG output, and HST regulation.

Around the nominal operating point, the voltage deviation can be approximated by a linearized sensitivity relation

$$\Delta V_i \approx \sum_{j \in \mathcal{N}} \left(R_{ij} \Delta P_j + X_{ij} \Delta Q_j \right) \quad (3)$$

where R_{ij} and X_{ij} denote the equivalent sensitivities of nodal voltage to active and reactive power variations, respectively. This relation shows that the voltage of one bus is affected not only by local power changes but also by network coupling. This provides the basis for the coordinated modeling of HST and DG.

As shown in Figure 2, the network diagram should present the main buses, the DG connection locations, the HST installation positions, and the distribution of monitored buses. This figure is used to illustrate the network structure and device layout of the control object.

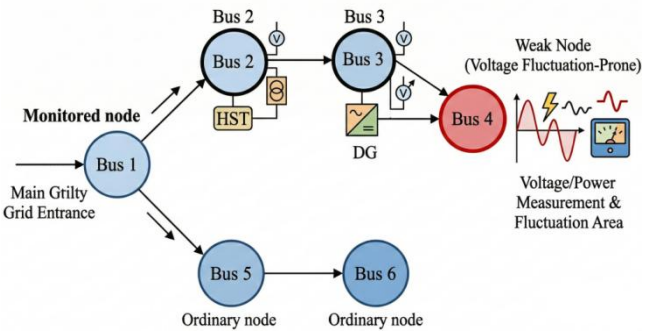

Figure 2. Distribution network structure and node layout

Figure 2 shows the correspondence among bus hierarchy, control devices, and monitoring locations. The following control model and index definitions are all established on this network structure.

3.2.2. HST-DG Coordinated Control Model

To describe the joint regulation of HST and DG on bus voltage, let V_i^{ref} be the reference voltage of bus i, and let $V_i(k)$ be the actual voltage at time step k. The voltage deviation is defined as

$$e_i(k) = V_i^{ref} - V_i(k) \quad (4)$$

The HST branch provides fast local compensation, and its control output is expressed as

$$u_i^H(k) = K_i^H e_i(k) \quad (5)$$

u_i^H where K_i^H is the HST control gain. The DG branch provides supplementary regulation, and its control output is written as

$$u_i^{DG}(k) = K_i^{DG} e_i(k) \quad (6)$$

where K_i^{DG} is the DG control gain. Considering the physical limits of the devices, the outputs of HST and DG satisfy

$$u_{i,\min}^H \le u_i^H(k) \le u_{i,\max}^H \quad (7)$$

$$u_{i,\min}^{DG} \le u_i^{DG}(k) \le u_{i,\max}^{DG} \quad (8)$$

Under the coordinated control structure, both devices participate in voltage regulation, and the total control action is given by

$$u_i(k) = u_i^H(k) + u_i^{DG}(k) \quad (9)$$

Accordingly, the voltage state update can be summarized as

$$V_i(k+1) = f\left(V_i(k), P_i^L(k), Q_i^L(k), P_i^{DG}(k), Q_i^{DG}(k), u_i(k)\right) \quad (10)$$

where $f(\cdot)$ denotes the state transition relation determined by the network dynamics. This model indicates that coordinated control aims to allocate the regulating roles of HST and DG according to the voltage deviation while satisfying the operating constraints of both devices.

As shown in Figure 3, the control block diagram should include voltage measurement, error calculation, the HST control branch, the DG control branch, and the feedback from the control output to the network state.

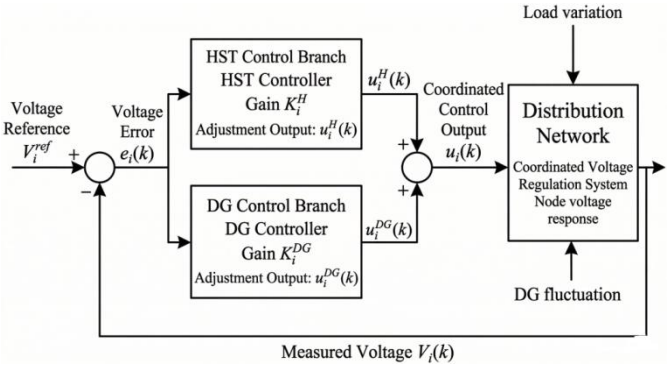

Figure 3. HST-DG coordinated control block diagram

Figure 3 shows that the error signal acts on both the HST and DG branches, and the two control outputs jointly determine the subsequent voltage evolution.

3.2.3. Voltage Control Performance Indices

To evaluate the voltage regulation effect of different control strategies, this study adopts voltage deviation, steady state error, overshoot, oscillation energy, and control power consumption as the performance indices. A composite performance index is then introduced for overall evaluation.

(1) The voltage deviation measures the average departure of monitored bus voltages from their reference values:

$$J_V = \frac{1}{T \mid \mathcal{N}_m \mid} \sum_{k=1}^{T} \sum_{i \in \mathcal{N}_m} \mid V_i(k) - V_i^{ref} \mid \tag{11}$$

(2) The steady state error describes the residual deviation after the system reaches steady operation:

$$J_{ss} = \frac{1}{\mid \mathcal{N}_m \mid} \sum_{i \in \mathcal{N}_m} \mid V_i(T_s) - V_i^{ref} \mid \tag{12}$$

where T_s is the steady state evaluation instant.

(3) The overshoot represents the maximum amount by which the voltage exceeds the reference during the response:

$$J_{os} = \frac{1}{\mid \mathcal{N}_m \mid} \sum_{i \in \mathcal{N}_m} \max_{1 \le k \le T} \left(V_i(k) - V_i^{ref}, 0 \right) \tag{13}$$

(4) The oscillation energy is used to describe the fluctuation intensity of the voltage trajectory:

$$J_{osc} = \frac{1}{\mid \mathcal{N}_m \mid} \sum_{i \in \mathcal{N}_m} \sum_{k=2}^{T} (V_i(k) - V_i(k-1))^2 \tag{14}$$

(5) The control power consumption reflects the overall magnitude of HST and DG regulation during control:

$$J_u = \frac{1}{T \mid \mathcal{N}_m \mid} \sum_{k=1}^{T} \sum_{i \in \mathcal{N}_m} (\mid u_i^H(k) \mid + \mid u_i^{DG}(k) \mid) \tag{15}$$

To balance multiple control objectives, the composite performance index is defined as

$$J_c = w_1 \tilde{J}_V + w_2 \tilde{J}_{ss} + w_3 \tilde{J}_{os} + w_4 \tilde{J}_{osc} + w_5 \tilde{J}_u \tag{16}$$

where $\tilde{J}_V, \tilde{J}_{ss}, \tilde{J}_{os}, \tilde{J}_{osc}$, and $\tilde{J}_u$ are the normalized indices, and w_1 to w_5 are the corresponding weights. These indices provide the basis for the subsequent policy training and strategy comparison.

3.3. DRL-Based Optimization Strategy

In this study, the HST-DG coordinated control process is formulated as a Markov decision process. At each control step, the agent observes the system state, outputs gain adjustment actions, and updates the policy according to the environmental feedback. Since the control variables are continuous, a continuous action deep reinforcement learning method with an actor critic structure is adopted.

3.3.1. State Space and Action Space

The state space is used to describe the current operating condition of the distribution network and the available controllable resources. For the present control task, the state vector consists of monitored node voltages, load levels, and DG outputs. Let M denote the number of monitored nodes. The state at time t is defined as

$$s_t = [V_1(t), \cdots, V_M(t), L_1(t), \cdots, L_M(t), G_1(t), \cdots, G_M(t)]^\top \tag{17}$$

where $V_i(t)$ denotes the node voltage, $L_i(t)$ denotes the load level, and $G_i(t)$ denotes the corresponding DG output. The state dimension is therefore $3M$. This state representation captures voltage deviation, load disturbance, and DG fluctuation simultaneously, which is suitable for coordinated control.

The action space represents the gain adjustment commands of HST and DG. The action is defined as a two-dimensional continuous vector

$$a_t = [\Delta K_H(t), \Delta K_{DG}(t)]^\top \tag{18}$$

where $\Delta K_H(t)$ and $\Delta K_{DG}(t)$ are the gain adjustments applied to HST and DG at time t, respectively. To ensure feasibility, the raw output of the agent is first constrained within a bounded continuous interval and then mapped to the allowable gain range of the controller. The action is updated at each control step and remains unchanged within the adjacent sampling interval.

As shown in Figure 4, the state vector is fed into the agent, the policy network generates the control action, and the action is applied to the distribution network environment. The resulting voltage response and operating feedback then form the next state.

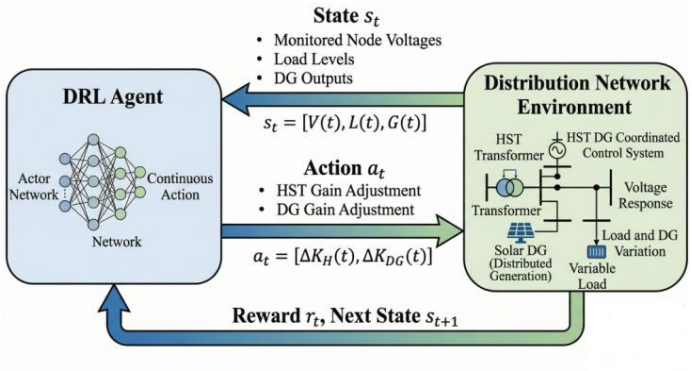

Figure 4. Interaction among state, action, and environment

Figure 4 illustrates the closed loop interaction between the DRL agent and the distribution network environment. The state variables describe the current operating condition, the action variables correspond to the gain adjustments of HST and DG, and the environmental feedback is used for subsequent policy updates.

3.3.2. Reward Function Design

The reward function describes the optimization objective of the agent. In order to maintain voltage quality, suppress dynamic fluctuation, and reduce control effort, the reward is composed of voltage deviation, overshoot, oscillation energy, and control power. The step reward is defined as

$$r_t = -(1.30\,\tilde{J}_V(t) + 1.10\,\tilde{J}_{os}(t) + 0.55\,\tilde{J}_{osc}(t) + 0.10\,\tilde{J}_u(t)) \tag{19}$$

$\tilde{J}_{osc}(t)$, and $\tilde{J}_u(t)$ denote the voltage deviation, overshoot, oscillation energy, and control effort after unified scaling.

A larger reward corresponds to a better control action in the overall sense.

The weighting factors reflect the relative importance of different control objectives during training. Voltage deviation and overshoot are directly related to voltage security and dynamic quality, and therefore receive larger weights. Oscillation energy is used to suppress excessive fluctuation during transient regulation and is assigned a moderate weight. The control effort term is introduced to avoid overly aggressive actions and to prevent the policy from relying only on large control outputs. The steady state error is reserved for offline performance evaluation and is not included as an individual step reward term, so that duplicated feedback with the voltage deviation term can be avoided in sequential learning.

3.3.3. Policy Training and Convergence Criterion

Since the action variables are continuous gain adjustments, the policy is trained using an actor critic framework. The policy network outputs the action according to the current state, while the value network evaluates the long-term return of the state action pair. During training, the action selection is written as

$$a_t = \mu(s_t \mid \theta^{\mu}) + \epsilon_t \tag{20}$$

where $\mu(\cdot)$ denotes the policy network, θ^{μ} denotes its parameters, and ϵ_t is the exploration noise. To improve sample efficiency and reduce training fluctuation, an experience replay mechanism is used to store transition samples, and the network parameters are updated using mini-batch training.

The target value of the critic network is written as

$$y_t = r_t + \gamma Q'(s_{t+1}, \mu'(s_{t+1} \mid \theta^{\mu'}) \mid \theta^{Q'}) \tag{21}$$

and the corresponding loss function is

$$L(\theta^{Q}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - Q(s_i, a_i \mid \theta^{Q}))^2 \tag{22}$$

where $Q(\cdot)$ denotes the critic network, $Q'(\cdot)$ and $\mu'(\cdot)$ denote the target critic network and target actor network, γ is the discount factor, and N is the mini batch size. The actor network is updated along the gradient direction provided by the critic so as to improve the long-term accumulated reward.

During training, the environment performs state observation, action execution, reward calculation, and sample storage at each control step, followed by network updating. The main hyperparameters include the learning rate, discount factor, replay buffer capacity, and batch size, which are determined through preliminary experiments. The stopping condition follows a dual criterion. Training is terminated when the maximum number of episodes is reached, or when the moving average reward becomes stable within a sliding window for several consecutive episodes. This criterion is used to balance convergence and training cost.

3.4. Communication Constraint Modeling

To examine the applicability of the coordinated control strategy under non ideal communication conditions, three types of communication constraints are introduced into the control loop, namely delay, packet loss, and bandwidth limitation. These communication constraints do not alter the physical model of the distribution network itself. Instead, they act on the transmission of state measurements and control commands, so that the observation received by the agent may differ from the actual system state, and the command executed by the plant may differ from the original decision. A unified framework is adopted for three communication scenarios, including no constraint, mild constraint, and severe constraint. In the present setting, the mild scenario is defined by a delay of 1 step, a packet loss probability of 0.05, and a bandwidth coefficient of 0.95, while the severe scenario is defined by a delay of 4 steps, a packet loss probability of 0.20, and a bandwidth coefficient of 0.70.

(1) Delay model

Communication delay is used to describe the fixed step lag caused by information transmission. Let the actual system state be s_t, and let the observed state received by the agent at time t be denoted by $\hat{s}_t$. The delayed observation is written as

$$\hat{s}_t = s_{t-\tau} \tag{23}$$

where τ denotes the communication delay in discrete control steps. When $\tau = 0$, no extra delay is imposed. When $\tau > 0$, the agent makes decisions based on historical states. The same treatment can also be applied to the command channel, where the actuator receives a previously generated control signal instead of the most recent one.

(2) Packet loss model

Packet loss is used to describe the case in which part of the state data or control commands cannot be successfully transmitted. Let δ_t be a Bernoulli random variable representing the transmission result at time t, then

$$\delta_t \sim \text{Bernoulli}(1-p) \tag{24}$$

where p is the packet loss probability. When $\delta_t = 1$, the current data packet is successfully delivered. When $\delta_t = 0$, the current packet is lost. For the control command, the actually executed input is written as

$$\tilde{u}_t = \delta_t u_t + (1-\delta_t)\tilde{u}_{t-1} \tag{25}$$

where u_t is the original control output and $\tilde{u}_t$ is the executed control command. This expression means that once the current command is lost, the actuator holds the previously received control input.

(3) Bandwidth limitation model

Bandwidth limitation is used to describe the reduction in effective information transmission within one control interval, which weakens the update capability of the controller. In this study, a bandwidth coefficient $\beta \in (0,1]$ is introduced to abstract the executed control input as

$$u_t^b = \beta\tilde{u}_t + (1-\beta)u_{t-1}^b \tag{26}$$

where u_t^b denotes the final control input under bandwidth limitation. When $\beta = 1$, the current control command is

fully delivered and executed. When $\beta < 1$, only part of the current control action is effectively transmitted, and the remaining part is inherited from the previously executed input. This form captures the sparse and smoothed update characteristic caused by limited communication bandwidth.

Under the unified experimental setting, the no constraint scenario is described by $\tau = 0$, $p = 0$, and $\beta = 1$. The mild constraint scenario is described by $\tau = 1$, $p = 0.05$, and $\beta = 0.95$. The severe constraint scenario is described by $\tau = 4$, $p = 0.20$, and $\beta = 0.70$. The policy network structure and the control objective remain unchanged across all communication scenarios. Only the communication parameters are adjusted to characterize different information channel conditions, so that the variation in control performance can be attributed to the communication constraints themselves rather than to changes in the control model.

4. Experiments and Results

This chapter evaluates the proposed DRL-based coordinated control strategy on the basis of the models and methods introduced in the previous chapter. The experiments include dataset description, platform and parameter settings, strategy comparison, and the analyses of training, validation, communication constraints, and robustness. The focus of this chapter is on the difference among control strategies in terms of voltage regulation quality and control effort, rather than on methodological derivation.

4.1. Experimental Overview

4.1.1. Dataset and Scenario Settings

The experimental input is taken from the Open Power System Data (OPSD) time series package, which provides publicly available real-world power system time-series data, including electricity consumption (load) and renewable generation profiles. In this study, the selected load and DG-related generation series are used as external time-varying inputs to drive the distribution network simulation. To keep the training and testing scenarios temporally independent, the first 24 hours are used as the training set, and the remaining 24 hours are used as the validation set. This partition preserves the continuity of adjacent operating periods while avoiding repeated use of the same data segment for both policy learning and performance evaluation.

As shown in Figure 5, the load profiles exhibit clear time-varying behavior within the 24-hour horizon, and different nodes present different peak and valley characteristics. The DG output shows stronger time dependence, with higher output during daytime and nearly zero output during nighttime. The training set and the validation set follow similar overall trends, while local fluctuations and peak positions are not exactly the same. This makes the dataset suitable for examining the adaptability of the DRL policy under unseen operating conditions.

Figure 5 presents the load and DG time-series characteristics of the training set and the validation set. All subsequent experiments are conducted on this dataset. The training set is mainly used for policy learning, while the validation set is used for performance testing and generalization analysis.

4.1.2. Compared Strategies and Evaluation Indices

The main strategies compared in this chapter include Baseline, HST-dominant, DG-dominant, and DRL. The Baseline strategy uses empirical parameter settings and represents the conventional control level. The HST-dominant and DG-dominant strategies emphasize single-sided regulation, respectively. The DRL strategy realizes coordinated control of HST and DG through policy learning. If a PSO curve is retained in a particular figure, it can be treated as a supplementary reference, while the main comparison in this chapter still focuses on the above four strategies.

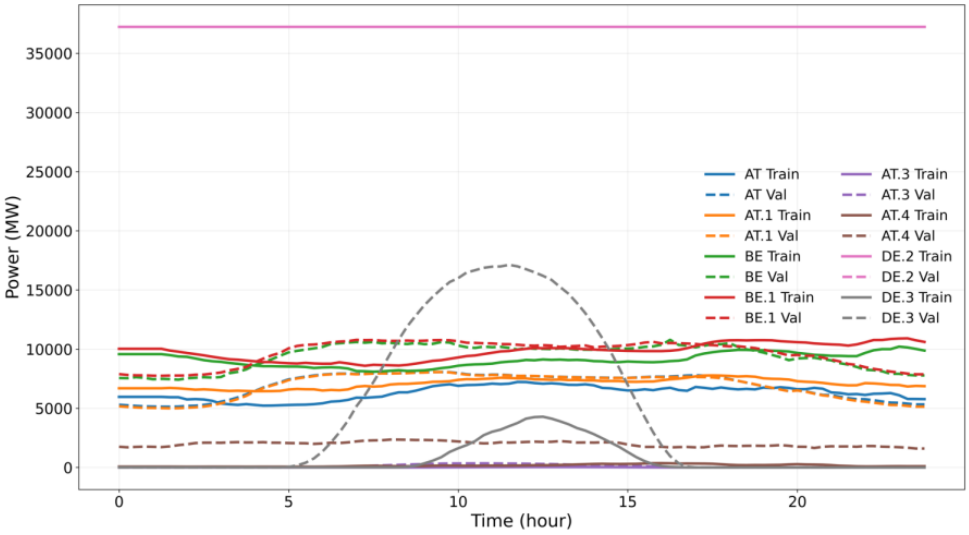

Figure 5. Load and DG time series.

Table 1 summarizes the compared strategies. It should be noted that the parameters of Baseline, HST-dominant, and DG-dominant remain fixed during the experiments, whereas the control action of DRL is provided online by the learned policy network.

Table 1. Description of compared strategies

Strategy	Setting	Description
Baseline	$K_H = 0.35, K_{DG} = 0.20$	Empirical parameter control
HST-dominant	$K_H = 0.60, K_{DG} = 0$	HST-dominant regulation
DG-dominant	$K_H = 0, K_{DG} = 0.05$	DG-dominant regulation
DRL	Online output of policy network	Coordinated regulation of HST and DG

This chapter adopts the evaluation framework defined in Section 3.2. The reported results mainly focus on voltage deviation, overshoot, oscillation energy, control effort, and the composite performance metric. For important nodes, voltage response curves are also used to analyze the dynamic characteristics of the control process. Under the unified indices and scenario settings, differences in performance across control strategies can be compared in a clear manner.

4.2. Baseline Performance Evaluation

This subsection presents the performance of the baseline strategy, which is used as the reference for all subsequent comparisons. The baseline controller adopts empirical parameters, where the HST gain is set to 0.35, and the DG gain is set to 0.20. All other experimental conditions remain the same as those used in the following comparative studies.

Under this setting, the system is able to achieve basic regulation, but the voltage fluctuation at key nodes is still evident. In the validation scenario, the baseline strategy yields a voltage deviation of 0.2245, an overshoot of 0.0824, an oscillation energy of 0.001424, and a control effort of 3.2640. The corresponding composite index is 0.7097. These results indicate that empirical parameter adjustment can provide a certain control effect, but it does not achieve a satisfactory balance between voltage quality and control cost.

As shown in Figure 6, the weak node AT.4 exhibits evident dynamic fluctuation under the baseline strategy. The voltage curve drops markedly in the early period and rises to a relatively high level in the later period, indicating that this strategy has a restricted capability in adapting to time-varying load and DG fluctuations. This observation is consistent with the large voltage deviation and overshoot reported in Table 2 and serves as the rationale for introducing an adaptive coordinated control strategy in the following subsection.

Figure 6 gives the dynamic voltage response of the baseline strategy at the key node. It serves as a direct visual reference for the later comparison with the DRL strategy.

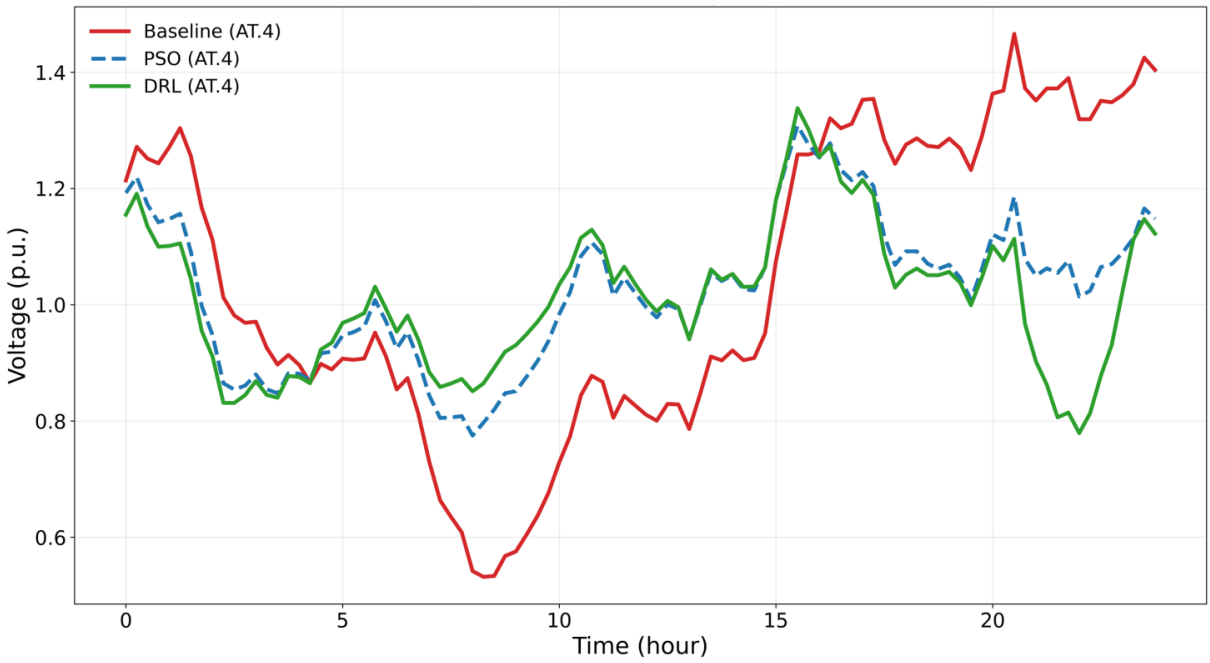

Figure 6. Voltage response of the key weak node under the baseline strategy

Table 2 shows that the baseline strategy leaves considerable potential for further improvement in all reported indices. In particular, the relatively high control effort indicates low utilization efficiency of regulating resources, while the large voltage deviation and overshoot imply that the voltage response is still unsatisfactory.

Table 2. Performance indices of the baseline strategy

Index	Value
Voltage deviation	0.2245
Overshoot	0.0824
Oscillation energy	0.001424
Control effort	3.2640
Composite index	0.7097

4.3. DRL Training and Validation

In this subsection, the DRL policy undergoes training on the training set before being tested on an independent validation set. The training phase mainly centers on reward convergence and policy robustness, whereas the validation stage focuses on the performance difference of DRL compared with the baseline strategy.

The training results demonstrate a clear convergence trend of the reward. The 15-step moving average reward increases from -65.74 to -19.96, corresponding to a relative improvement of about 69.63%, and the tail stability ratio is 0.00291. In addition, under the same initial random seed, the differences in reward curve, action output, and performance indicators are zero, suggesting good repeatability of the training procedure.

As shown in Figure 7, the reward fluctuates strongly at the beginning of training and then gradually rises before becoming stable. This evolution suggests that the agent progressively acquires effective control behavior via continuous interaction with the environment and reaches a relatively stable stage in the later episodes.

Figure 7 presents the reward evolution during training and is used to illustrate the convergence behavior of the learned policy.

On the training set, DRL already outperforms the baseline strategy. The voltage deviation decreases from 0.1740 to 0.0655, the overshoot decreases from 0.0518 to 0.0139, the oscillation energy decreases from 0.005540 to 0.004957, and the control effort decreases from 2.6680 to 1.0725. The corresponding reductions across the four metrics are 62.37%, 73.13%, 10.54%, and 59.80%, respectively. Meanwhile, the total reward improves from -53.09 to -20.20.

On the validation set, DRL maintains consistent improvement. The voltage deviation decreases from 0.2245 to 0.0740, the overshoot decreases from 0.0824 to 0.0180, the oscillation energy decreases from 0.001424 to 0.001001, and the control effort decreases from 3.2640 to 1.3086. Compared with the baseline strategy, the four indices are reduced by 67.05%, 78.16%, 29.67%, and 59.91%, respectively. At the same time, the validation reward improves from -68.13 to -23.75, and all metric trends are aligned with those observed on the training set.

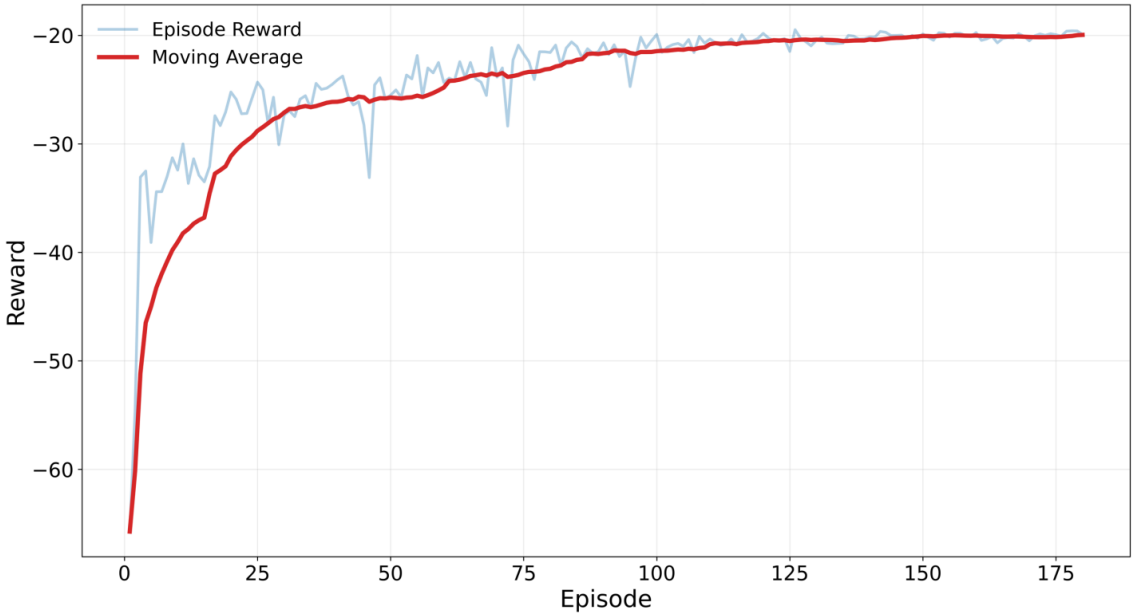

Figure 7. DRL reward convergence curve

Table 3. Comparison of performance metrics in the training and validation stages

Scenario	Strategy	Voltage deviation	Overshoot	Oscillation energy	Control effort
Training set	Baseline	0.1740	0.0518	0.005540	2.6680
Training set	DRL	0.0655	0.0139	0.004957	1.0725
Validation set	Baseline	0.2245	0.0824	0.001424	3.2640
Validation set	DRL	0.0740	0.0180	0.001001	1.3086

Table 3 indicates that DRL achieves lower voltage deviation, overshoot, and control effort across both the training and the validation set. More importantly, the direction of improvement remains unchanged from training to validation, which implies that the derived coordinated control policy is not restricted only to the training dataset but also retains good adaptability under unseen time-series conditions.

4.4. Impact of Communication Constraints

This subsection evaluates the adaptability of the DRL strategy under communication-constrained conditions. Three scenarios are considered, namely no constraint, mild constraint, and severe constraint. The mild scenario corresponds to a delay of 1 step, a packet loss probability of 0.05, and a bandwidth coefficient of 0.95. The severe scenario corresponds to a delay of 4 steps, a packet loss probability of 0.20, and a bandwidth coefficient of 0.70. All other experimental settings remain unchanged.

Table 4 reports the main indices under various communication conditions. Compared with the no-constraint case, the mild constraint case shows only limited variation. The voltage deviation changes from 0.0740 to 0.0758, and the overshoot changes from 0.0180 to 0.0188. Oscillation energy and control effort do not deteriorate noticeably. The settling step changes from 12 to 13, which indicates that mild communication disturbances have only a limited effect on the control process.

Under severe constraints, the performance degradation becomes more visible. The voltage deviation rises to 0.0938, the overshoot rises to 0.0286, the oscillation energy rises to 0.001660, and the control effort rises to 1.6362. The settling step is extended from 12 to 23. These results indicate that stronger delay, packet loss, and bandwidth limitation jointly slow down the response and weaken the dynamic smoothness, with overshoot and oscillation energy showing the most evident deterioration.

As shown in Figure 8, the voltage deviation curves under no constraint and mild constraint remain close to each other, whereas the severe constraint curve is lifted upward and exhibits more obvious fluctuation. This observation is consistent with the numerical results in Table 4, which suggests that the proposed strategy can still retain reasonable performance under mild communication constraints, while stronger communication disturbances result in slower system response.

Figure 8 is used to illustrate the variation trend of voltage deviation under various communication conditions and to show the direct impact of communication conditions on control performance.

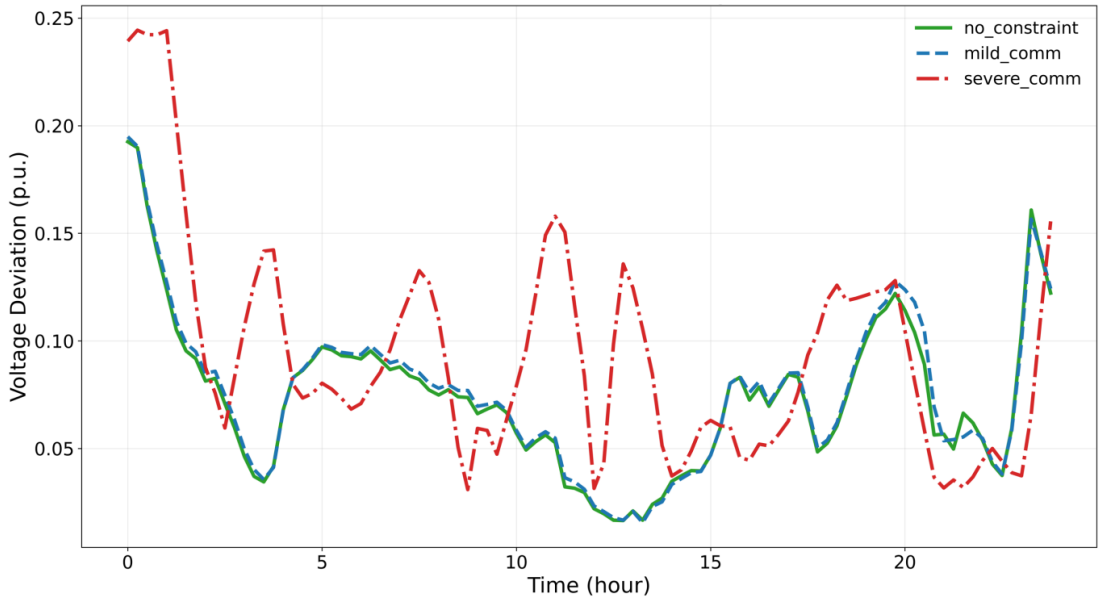

Figure 8. Voltage deviation curves under communication constraints

Table 4. Performance indices under various communication conditions

Scenario	Voltage deviation	Overshoot	Oscillation energy	Control effort	Settling step
No constraint	0.0740	0.0180	0.001001	1.3086	12
Mild constraint	0.0758	0.0188	0.001000	1.2938	13
Severe constraint	0.0938	0.0286	0.001660	1.6362	23

Table 5. Multi-strategy comparison results

Strategy	Voltage deviation	Overshoot	Oscillation energy	Control effort	Composite index
DRL	0.0740	0.0180	0.001001	1.3086	0.2474
HST-dominant	0.1016	0.0323	0.001011	0.8030	0.2484
DG-dominant	0.1738	0.0765	0.001108	0.7775	0.3884
Baseline	0.2245	0.0824	0.001424	3.2640	0.7097

Table 4 indicates that communication constraints directly affect control performance, and the effect depends on the severity of the constraint. Under mild constraints, the strategy still maintains good control quality. Under severe constraints, the degradation is mainly reflected by a slower system response, larger overshoot, and stronger fluctuation.

4.5. Multi-Strategy Performance Comparison

This subsection compares Baseline, HST-dominant, DG-dominant, and DRL under the same validation scenario. The evaluation considers both individual indices and the ranking of the composite index.

Table 5 summarizes the results of the four strategies. In terms of the composite index, DRL achieves the lowest value of 0.2474. The HST-dominant strategy yields 0.2484, which is very close to DRL. The DG-dominant strategy yields 0.3884, and the Baseline strategy yields 0.7097, which is the worst among the four. According to the composite index, the ranking is DRL, HST-dominant, DG-dominant, and Baseline. Compared with Baseline, DRL reduces the composite index by 65.14%.

From the perspective of individual indices, DRL performs best in voltage deviation and overshoot, with values of 0.0740 and 0.0180, respectively. The HST-dominant strategy achieves 0.1016 in voltage deviation and 0.0323 in overshoot, which are clearly superior to the Baseline and DG-dominant but still higher than DRL. The DG-dominant strategy has the lowest control effort of 0.7775, but its voltage deviation and overshoot remain at 0.1738 and 0.0765. This indicates that relying solely on DG can reduce control effort, but it is insufficient to maintain voltage quality. The Baseline strategy shows the highest control effort and the worst voltage quality among all strategies.

Table 5 shows that DRL achieves the best overall performance. Compared with the single-device strategies, its advantage mainly lies in voltage deviation and overshoot control. Compared with Baseline, its advantage is reflected in the overall improvement of all reported indices. Overall, DRL behaves as a more balanced coordinated control strategy, while HST-dominant and DG-dominant each reflect only one-sided regulation characteristics.

4.6. Robustness Analysis

This subsection evaluates whether the performance improvement of the DRL strategy is accidental by repeating the experiment under different random initializations. A total of 10 independent runs are conducted with random seeds from 4200 to 4209, while all other experimental settings remain unchanged. The composite performance index is adopted as the main evaluation measure, and its mean, standard deviation, minimum, and maximum values are reported.

Table 6 summarizes the statistical results of all strategies. The DRL strategy achieves a mean composite index of 0.3256, a standard deviation of 0.0605, a minimum of 0.2564, and a maximum of 0.4337, with a coefficient of variation of 0.1860. By comparison, the mean values of Baseline, HST-dominant, and DG-dominant are 0.7191, 0.3700, and 0.8725, respectively. These results indicate that although DRL exhibits some fluctuation under different random initializations, its overall level remains clearly better than Baseline and DG-dominant and stays close to HST-dominant.

As shown in Figure 9, the composite index of DRL remains within a relatively low range across the 10 repeated runs, without obvious instability or abnormal deviation. Combined with the standard deviation and coefficient of variation reported in Table 6, this suggests that DRL maintains good consistency under repeated independent training and does not rely on a single favorable initialization.

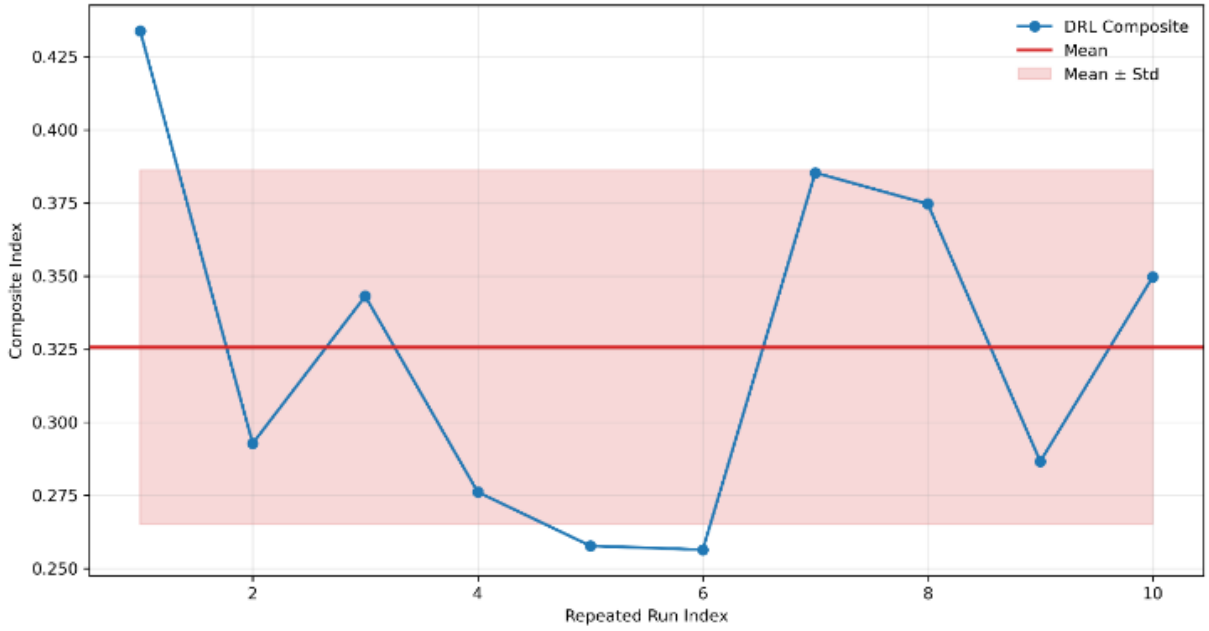

Figure 9. Robustness results under repeated runs

Figure 9 is used to illustrate the variation range of the DRL composite index under different random initializations. It mainly reflects the dispersion and overall stability of the results.

In addition, the Mann–Whitney U test and the t test are used to compare DRL with the other strategies. Compared with Baseline, the Mann–Whitney p value is $3.19 \times 10^{-5}$and the t-test p value is 3.57×10^{-9}. Compared with DG-dominant, the Mann–Whitney p value is $3.19 \times 10^{-5}$and the t-test p value is 1.92×10^{-10}. These results both indicate a significant advantage of DRL over the two strategies. Compared with HST-dominant, the Mann–Whitney p value is 0.0576, which is at a marginal significance level, while the t-test p value is 0.0228, still supporting a statistically meaningful advantage of DRL.

Table 6. Statistical results of robustness analysis

Strategy	Mean	Standard deviation	Minimum	Maximum
DRL	0.3256	0.0605	0.2564	0.4337
Baseline	0.7191	0.0000	0.7191	0.7191
HST-dominant	0.3700	0.0000	0.3700	0.3700
DG-dominant	0.8725	0.0000	0.8725	0.8725

Table 6 shows that DRL maintains a relatively stable average performance across repeated runs and still outperforms the baseline and DG-dominant strategies overall. Taken together with Figure 9 and the significance tests, the performance gain reported in this study cannot be regarded as an accidental outcome, and the proposed coordinated control strategy can be considered stable and reliable.

5. Conclusion

This study proposed a deep reinforcement learning-based HST-DG coordinated control method for voltage regulation in distribution networks and established a complete experimental framework including baseline comparison, training and validation, communication constraint analysis, multi-strategy comparison, and repeated-run statistics. The proposed method uses node voltage, load, and DG output as state inputs, and takes the gain adjustment of HST and DG as control actions to achieve coordinated optimization of voltage response.

The results demonstrate that the proposed DRL strategy provides effective and robust voltage regulation performance under different operating conditions. Compared with the reference strategies, it achieves better overall control performance and maintains acceptable adaptability under communication-constrained scenarios, indicating its potential for coordinated voltage control in active distribution networks.

Several limitations remain in the present work. The experiments are mainly conducted on a given distribution network model and a limited-scale scenario, and the applicability to more complex topologies and larger systems still requires further verification. In addition, the current study adopts a single-agent framework, and the issue of multi-device and multi-area coordinated control

has not been fully explored. Moreover, although the proposed method was validated using publicly available real-world power system time-series data from OPSD, the closed-loop evaluation was still carried out in a simulation environment and has not yet been verified on hardware-in-the-loop platforms or practical engineering systems. Future work will therefore focus on extending the method to more complex networks, introducing multi-agent coordination mechanisms, and carrying out validation under more realistic operating conditions.

Acknowledgements

This work was financially supported by the Science and Technology Project of China Southern Power Grid, entitled "Research on Hybrid Series-Compensated Transformer for Wide-Frequency Voltage Stability in Weak Grids and Its Coordinated Control with Distributed Generation" (Project No. 031300KC23120024).

References

[1] Hua D, Peng F, Liu S, et al. Coordinated volt/var control in distribution networks considering demand response via safe deep reinforcement learning. Energies, 2025, 18(2): 333.

[2] Song G, Wu Q, Jiao W, et al. Distributed coordinated control for voltage regulation in active distribution networks based on robust model predictive control. International Journal of Electrical Power & Energy Systems, 2025, 166: 110529.

[3] Suchithra J, Rajabi A, Robinson D A. Enhancing PV hosting capacity of electricity distribution networks using deep reinforcement learning-based coordinated voltage control. Energies, 2024, 17(20): 5037.

[4] Zhang L, Yang F, Yan D, et al. Multi-agent deep reinforcement learning-based distributed voltage control of flexible distribution networks with soft open points. Energies, 2024, 17(21): 5244.

[5] Kabir F, Yu N P, Gao Y Q, et al. Deep reinforcement learning-based two-timescale Volt-VAR control with degradation-aware smart inverters in power distribution systems. Applied Energy, 2023, 335: 120629.

[6] Yang, T.; Liu, Y.; Li, H.; Chen, Y.; Pen, H. Cooperative Voltage Control in Distribution Networks Considering Multiple Uncertainties in Communication. Sustainable Energy, Grids and Networks, 2024, 39: 101459.

[7] Nematshahi, S.; Shi, D.; Wang, F.; Yan, B.; Nair, A. Deep Reinforcement Learning Based Voltage Control Revisited. IET Generation, Transmission & Distribution, 2023, 17(6).

[8] Murray W, Adonis M, Raji A. Voltage control in future electrical distribution networks. Renewable and Sustainable Energy Reviews, 2021, 146: 111100.

[9] Abdelkader S M, Kinga S, Ebinyu E, et al. Advancements in data-driven voltage control in active distribution networks: A Comprehensive review. Results in Engineering, 2024, 23: 102741.

[10] Wang J, Lan J, Wang L, et al. Voltage hierarchical control strategy for distribution networks based on regional autonomy and photovoltaic-storage coordination. Sustainability, 2024, 16(16): 6758.

[11] Yan Q, Chen X, Xing L, et al. Multi-Timescale Voltage Regulation for Distribution Network with High Photovoltaic Penetration via Coordinated Control of Multiple Devices. Energies, 2024, 17(15): 3830.

[12] Zhang T J, Yu L, Yue D, et al. Coordinated voltage regulation of high renewable-penetrated distribution networks: An evolutionary curriculum-based deep reinforcement learning approach. International Journal of Electrical Power & Energy Systems, 2023, 149: 108995.

[13] Ma, C.; Xiong, W.; Tang, Z.; Li, Z.; Xiong, Y.; Wang, Q. A Hierarchical Voltage Control Strategy for Distribution Networks Using Distributed Energy Storage. Electronics, 2025, 14(9): 1888.

[14] Xiong, W.; Tang, Z.; Cui, X. Distributed data-driven voltage control for active distribution networks with changing grid topologies. Control Engineering Practice, 2024, 147: 105933.

[15] Abbas, G.; Zhi, W.; Ali, A. A two-stage reactive power optimization method for distribution networks based on a hybrid model and data-driven approach. IET Renewable Power Generation, 2024, 18(16): 3967-3979.

[16] Li Z, Wu Q, Chen J, et al. Double-time-scale distributed voltage control for unbalanced distribution networks based on MPC and ADMM. International Journal of Electrical Power & Energy Systems, 2023, 145: 108665.

[17] Kabir F, Yu N P, Gao Y Q, Wang W Y. Deep reinforcement learning-based two-timescale Volt-VAR control with degradation-aware smart inverters in power distribution systems. Applied Energy, 2023, 335: 120629.

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

[19] Su, S.; Zhan, H.; Zhang, L.; Xie, Q.; Si, R.; Dai, Y.; Gao, T.; Wu, L.; Zhang, J.; Shang, L. Volt-VAR Control in Active Distribution Networks Using Multi-Agent Reinforcement Learning. Electronics, 2024, 13(10): 1971.

[20] Yan R D, Xing Q, Xu Y. Multi-agent safe graph reinforcement learning for PV inverters-based real-time decentralized Volt/VAR control in zoned distribution networks. IEEE Transactions on Smart Grid, 2024, 15(1): 299-311.

[21] Liu Q, Guo Y, Xu T. Robust deep reinforcement learning for inverter-based Volt-VAR control in partially observable distribution networks. Applied Energy, 2025, 399: 126445.

[22] Hossain, R.; Gautam, M.; Olowolaju, J.; Livani, H.; Benidris, M. Multi-agent voltage control in distribution systems using GAN-DRL-based approach. Electric Power Systems Research, 2024, 234: 110528.

[23] Wang L, Xie L, Yang Y, et al. Distributed online voltage control with fast PV power fluctuations and imperfect communication. IEEE Transactions on Smart Grid, 2023, 14(5): 3681-3695.

[24] Guo G D, Zhang M F, Gong Y F, Xu Q W. Safe multi-agent deep reinforcement learning for real-time decentralized control of inverter based renewable energy resources considering communication delay. Applied Energy, 2023, 349: 121648.

[25] Kang, W.; Guan, Y.; Wang, H.; Vasquez, J. C.; Agundis-Tinajero, G. D.; Guerrero, J. M. Distributed control of virtual energy storage systems for voltage regulation in low voltage distribution networks subjects to varying time delays. Applied Energy, 2024, 376: 124295.

[26] Zhang, F. Adaptive Event-Triggered Voltage Control of Distribution Network Subject to Actuator Attacks Using Neural Network-Based Sliding Mode Control Approach. Electronics, 2024, 13(15): 2960.

[27] Xiong, K.; et al. Multi-agent deep reinforcement learning-enabled voltage regulation approach for partitioned active distribution network using heterogeneous PV inverters. Energy Conversion and Economics, 2025, 6(6).
[28] Yu, P.; Zhang, H.; Song, Y.; Wang, Z.; Dong, H.; Ji, L. Safe reinforcement learning for power system control: A review. Renewable and Sustainable Energy Reviews, 2025, 223: 116022.