energies

MDPI

Article

Deep Reinforcement Learning-Based Voltage-Var Control for Active Distribution Networks Using GAT-SimAM

Yunhai Zhou, Lingzhi Wang *, Zhewen Sun, Desong Wang and Hongyang Sheng

College of Electrical Engineering and New Energy, China Three Gorges University, Yichang 443002, China
* Correspondence: 13160677171@163.com

Abstract

High penetration of distributed photovoltaic (PV) generation can create spatially heterogeneous voltage excursions in active distribution networks, while conventional voltage-var control (VVC) methods struggle with uncertain operating conditions. This study proposes a soft actor-critic (SAC) VVC framework with a two-layer graph attention network (GAT) and a simple parameter-free attention module (SimAM). GAT aggregates topology-dependent information, and SimAM recalibrates the 32 feature channels within each node after graph aggregation without adding trainable parameters. The controller coordinates PV-inverter reactive power, static var compensators, and battery storage. Evaluation uses a frozen 96-day test set comprising 46,080 three-minute operating steps on a modified IEEE 123-bus feeder. Under a $\pm 5\%$ voltage band, the complete method reduces the whole-step violation rate from 15.651% without control and 0.701% with Q-V droop control to 0.106%; under the $\pm 7\%$ band specified by GB/T 12325-2008, it records no violations. A matched comparison between SAC with a GAT feature extractor and penalty-only GAT-SimAM reduces the violation rate from 0.697% to 0.109%, while average daily loss increases from 2.04 to 2.11 MWh. The results show that topology-aware feature extraction and parameter-free channel recalibration can improve voltage-risk sensitivity while maintaining millisecond-scale inference.

Keywords: active distribution network; voltage-var control; deep reinforcement learning; graph attention network; SimAM; distributed photovoltaic generation

1. Introduction

Driven by China's carbon-peaking and carbon-neutrality goals and the development of new power systems, the penetration of distributed photovoltaic (PV) generation continues to increase, and passive distribution networks are evolving into active distribution networks with bidirectional source-load interactions [1]. The stochastic and intermittent nature of PV output causes bus voltages to depart frequently from their nominal range, making voltage violations a central threat to secure distribution-network operation [2,3]. A review in [4] identifies voltage regulation under high renewable penetration as a key technical bottleneck limiting distributed-generation accommodation.

The variability of PV output that motivates the control problem has itself been the subject of recent modelling work, including hybrid approaches that combine deep learning with metaheuristic optimization for PV power forecasting [5]. Such work addresses the prediction layer that precedes control; the present study takes the PV output as a given disturbance sequence and addresses the control layer, and it is cited here as evidence that the uncertainty being handled downstream is both substantial and actively studied.

Check for updates

Received: 20 September 2026
Revised: 19 October 2026
Accepted: 20 October 2026
Published: 26 October 2026

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

Existing voltage-var control (VVC) methods can be grouped into three categories. Heuristic Q-V droop strategies depend on manually tuned coefficients and provide limited system-wide coordination. Mathematical optimization methods, including second-order cone programming and model predictive control [6–8], can obtain high-quality solutions but require accurate network parameters and may be too computationally demanding for real-time operation. Distributed and partitioned coordination methods [9–12] reduce communication burdens through network decomposition, but optimization gaps can remain at partition boundaries. These limitations motivate data-driven VVC strategies that combine rapid inference with system-level coordination.

Deep reinforcement learning (DRL) has therefore been introduced into VVC because it supports data-driven end-to-end decision making [13]. Multi-agent DQN was applied to volt-var control in unbalanced distribution systems in [14]. A two-stage method combining hourly mixed-integer second-order cone programming with deep deterministic policy gradients was proposed in [15]. Model-free multi-agent SAC with surrogate modeling and voltage-sensitivity partitioning was developed in [16], while ref. [17] coordinated storage active power and PV reactive power over two time scales. Safe DDPG with an action safety layer was studied in [18], and human-in-the-loop SAC was used to correct unsafe actions in [19]. Self-attention and large-language-model assistance have also broadened DRL-based distribution-network control [20]. Most of these methods, however, use multilayer perceptrons (MLPs) to process a one-dimensional concatenation of bus states. This representation neglects network topology and electrical coupling, making it difficult to distinguish the voltage effects of buses at different locations and potentially requiring excessive control effort to maintain voltage security.

Graph neural networks (GNNs) are naturally suited to power-network data [21] and have performed well in fault location [22,23], state estimation [24,25], and topology identification [26]. The graph attention network (GAT) in [27] uses masked self-attention to aggregate neighboring-bus information adaptively; an interpretable GAT was subsequently applied to reactive-power optimization [28]. For distribution-network VVC, GNNs have been combined with DRL for voltage and reactive-power optimization [29,30]. Other developments include physics-informed global GAT representations [31], physics-assisted multi-agent graph reinforcement learning [32], attention-based communication for emergency voltage control [33], and spatial-temporal graph representations [34]. Recent reviews emphasize topology awareness as an important factor in VVC performance [35].

Neighborhood aggregation in GNNs can nevertheless homogenize node representations. In deeper networks, repeated aggregation may make node features increasingly similar, a phenomenon known as over-smoothing [36–38]. Mathematical analysis has shown that attention weights alone do not fundamentally prevent exponential convergence of node representations [39], and dominant propagation patterns can cause GATs to neglect information associated with particular nodes as depth increases [40]. These results concern deep networks. Whether the effect also appears at the two-layer depth used here is a separate empirical question, and Section 4.2 answers it by measurement rather than by assumption [36–40].

This question is relevant to VVC because voltage risk is often concentrated at a small number of electrically sensitive buses whose operating features can differ substantially from those of adjacent buses. If aggregation blends those differences away, the policy becomes less sensitive to incipient violations and its corrective actions are delayed or too small. Rather than assert that this blending occurs at the depth used here, we measure it: the probe analysis in Section 4.2 quantifies how well the node representation predicts the voltage-deviation magnitude at each bus, before and after channel recalibration.

SimAM is a parameter-free attention module that derives feature weights from an energy function [41]. In the implementation used here, the mean and variance are calculated across feature channels within each node after GAT aggregation. SimAM therefore does not compare one bus directly with its neighbors; instead, it recalibrates the channels of each bus representation according to their within-node statistical distinctiveness. Applied after topology-aware aggregation, this operation can preserve channels carrying localized voltage-risk information without adding trainable parameters, which is attractive when power-system training data are limited.

Accordingly, this paper formulates VVC as a Markov decision process and proposes a SAC controller with a GAT-SimAM feature extractor. GAT maps feeder topology to adaptive neighborhood aggregation, and SimAM performs parameter-free channel recalibration after graph aggregation. The resulting policy issues continuous reactive-power commands for PV inverters and static var compensators (SVCs), together with storage charging and discharging commands. Training proceeds through interaction with a power-flow environment and does not require day-ahead load forecasts or discretization of the continuous action space. A unified 96-day frozen test set, engineering baselines, and matched ablation groups are used to evaluate overall effectiveness and isolate the contributions of SimAM, the training penalty, and execution-time action scaling.

Stated in three layers, the research gap addressed here is as follows. What is known: voltage control under high PV penetration is a well-characterized problem, and topology-aware graph representations combined with deep reinforcement learning are established techniques that have been reported to work on distribution feeders. What remains unresolved: voltage risk is spatially concentrated at a small number of electrically sensitive buses, while graph aggregation combines each bus with its neighbourhood, so the representation the policy acts on is not guaranteed to emphasize the bus-level detail that distinguishes an at-risk bus from its neighbours, and the learned controller must remain safe under delayed and partial observation. What is missing: a formulation that combines a parameter-free, node-wise recalibration of the aggregated representation with a voltage-constrained DRL policy and validates the two jointly on voltage security and operating economy. This paper provides such a formulation and evaluates it under a unified frozen test protocol.

2. System Modeling and Control Framework

2.1. Distribution Network and Control-Device Modeling

The active distribution network comprises distribution lines, load buses, distributed PV units, SVCs, and battery energy-storage systems. The operating model includes nodal active- and reactive-power balance, line voltage drop, PV-inverter reactive-power limits, SVC reactive-power limits, and storage state-of-charge constraints. Newton–Raphson power flow supplies the physical state transition used by the reinforcement-learning environment.

For any bus i and adjacent bus $j \in \mathcal{N}(i)$, active- and reactive-power balance and the voltage relationship are given by Equations (1)–(3).

$$\sum_{j\in\mathcal{N}(i)} P_{ij,t} = P_{i,t}^{\text{load}} - P_{i,t}^{\text{PV}} - P_{i,t}^{\text{bat}} \tag{1}$$

$$\sum_{j\in\mathcal{N}(i)} Q_{ij,t} = Q_{i,t}^{\text{load}} - Q_{i,t}^{\text{PV}} - Q_{i,t}^{\text{SVC}} \tag{2}$$

$$V_{i,t} = V_{j,t} - \frac{r_{ij}P_{ij,t} + x_{ij}Q_{ij,t}}{V_0} \tag{3}$$

where $P_{ij,t}$ and $Q_{ij,t}$ are the active and reactive power flows on branch ij at time t; $P_{i,t}^{\text{load}}$ and $Q_{i,t}^{\text{load}}$ are the active and reactive loads at bus i; $P_{i,t}^{\text{PV}}$, $Q_{i,t}^{\text{PV}}$, $Q_{j,t}^{\text{SVC}}$, and $P_{k,t}^{\text{bat}}$ denote PV active

power, PV reactive power, SVC reactive power, and storage active power, respectively; r_{ij} and x_{ij} are branch resistance and reactance; and V_0 is the reference voltage.

(1) The available reactive power of each PV inverter is constrained by its instantaneous active-power output and apparent-power rating, as expressed in Equation (4).

$$\left|Q_{i,t}^{\mathrm{PV}}\right| \leq max\left(Q_i^{\mathrm{base}}, P_{i,t}^{\mathrm{PV}}\right), \forall i \in \Omega_{\mathrm{PV}} \tag{4}$$

where Q_i^{base} is the available reactive-power magnitude of PV inverter i under its instantaneous active-power output and apparent-power rating, and Ω_{PV} is the set of buses with PV inverters.

(2) The reactive-power output of each SVC is limited by its rated capacity, as expressed in Equation (5).

$$\left|Q_{j,t}^{\mathrm{SVC}}\right| \leq Q_j^{\max} = 0.30\ \mathrm{MVAr}, \forall j \in \Omega_{\mathrm{SVC}} \tag{5}$$

where Q_j^{max} is the rated reactive-power magnitude of SVC j, and Ω_{SVC} is the set of buses equipped with SVCs.

(3) Each battery is represented by a discrete-time state-of-charge model that accounts for charging and discharging efficiencies, as expressed in Equation (6).

$$\mathrm{SOC}_{k,t+1} = \begin{cases} \mathrm{SOC}_{k,t} + \frac{\eta_c P_{k,t}^{\mathrm{ch}} \Delta t}{E_k^{\mathrm{cap}}}, & a_{\mathrm{bat},k,t} \geq 0 \\ \mathrm{SOC}_{k,t} - \frac{P_{k,t}^{\mathrm{dis}} \Delta t}{\eta_d E_k^{\mathrm{cap}}}, & a_{\mathrm{bat},k,t} < 0 \end{cases} \tag{6}$$

where $\mathrm{SOC}_{k,t}$ is the state of charge of storage unit k at time t; $\eta_c = \eta_d = 0.95$ denotes charging and discharging efficiencies; $P_{k,t}^{\mathrm{ch}}$ and $P_{k,t}^{\mathrm{dis}}$ are charging and discharging power; E_k^{cap} is rated energy capacity; Δt is the control interval; and $a_{\mathrm{bat},k,t} \in [-1, 1]$ is the normalized storage action. The physical state-of-charge range is $[0, 1]$.

Together, the power-balance, equipment-capacity, and storage-state constraints define the physical transition model. Nodal balance and line voltage-drop relationships determine how voltage changes with load, PV generation, and control actions, while device constraints delimit the feasible action set. The reinforcement-learning environment uses this model to validate each action by power flow and update the system state.

Distribution-network voltage management coordinates resources with different action characteristics and time scales. These resources are divided into three categories.

The first category comprises slow discrete devices, including on-load tap changers and switched capacitor banks. Their actions are discrete, each operation has a relatively large voltage effect, and mechanical lifetime limits the number of daily operations. They are therefore suited to day-ahead or hourly voltage-reference scheduling rather than tracking minute-scale PV fluctuations. In all experiments, the substation tap operates at the same day-ahead fixed position, providing a common boundary condition for minute-scale VVC.

The second category comprises fast continuous reactive-power devices, namely SVCs and PV inverters. They can adjust reactive power continuously at second-to-minute time scales without mechanical wear and are the principal resources for tracking PV fluctuations.

The third category comprises active-power storage resources. Storage absorbs locally generated surplus power when reactive-power capability is limited, and its charging or discharging power is superimposed on the load at the installation bus in the power-flow model.

The controller therefore operates the continuous resources in the second and third categories at a three-minute interval, while the first-category devices remain fixed. This separation prevents overlap between the time scales assigned to the three resource types.

2.2. Markov Decision Process Formulation

The decision process is modeled as a Markov decision process (MDP), defined by the observation space S, action space A, transition probability P, discount factor γ, and reward R; that is, {S, A, P, γ, R}. The observation, action, and reward definitions are given below.

2.2.1. Observation Space

A dictionary observation space is used, containing node features, topology information, and global features:

$$S_t = \left\{ X_t^n, E_t, X_t^g \right\} \tag{7}$$

where $X_t^n \in R^{N \times F}$ is the node-feature matrix at time t; N is the number of buses; F is the feature dimension. In the implementation, $d = 7$ and the channels are voltage magnitude, active-power injection, reactive-power injection, PV active power, PV reactive power, SVC reactive power, and storage state of charge. The vector $\mathbf{G}_t \in R^2$ contains the system maximum voltage deviation and total active-power loss.

The edge-index matrix $\mathbf{E} \in Z^{2 \times E}$ represents feeder topology, where **E** is the number of directed edges. For each undirected branch (i, j), the matrix stores both $(i \to j)$ and $(j \to i)$; therefore, $E = 2 \times 122 = 244$. Its first row contains source-bus indices and its second row contains the corresponding target-bus indices.

2.2.2. Action Space

The continuous action-space dimension equals the number of controllable devices. It contains PV-inverter reactive-power commands, SVC reactive-power commands, and storage charging or discharging commands. Normalized actions lie in $[-1, 1]$ and are mapped linearly to physical control commands:

$$A_t = \left[a_{PV,1}, \ldots, a_{PV,15}, a_{SVC,1}, \ldots, a_{SVC,4}, a_{bat,1}, \ldots, a_{bat,4} \right]^T \tag{8}$$

where a_t is the normalized continuous action vector at time t; $a_{PV,i}$, $a_{SVC,i}$, and $a_{bat,i}$ are the normalized commands for PV inverter i, SVC i, and storage unit i, respectively. $a_{PV,i}$ maps to the feasible PV reactive-power range, $a_{SVC,i}$ maps to the SVC reactive-power range, and $a_{bat,i}$ maps to storage power, with positive values denoting charging and negative values denoting discharging.

2.2.3. Reward Function

An L1 penalty-based reward is used to represent voltage security, network losses, storage operating limits and control effort:

$$r_t = -\left(\begin{array}{l} w_1 \sum_{i=1}^{N} \left[\max(0, V_{i,t} - V_{\max}) + \max(0, V_{\min} - V_{i,t}) \right] \\ + w_2 \sum_{i=1}^{N} \left| V_{i,t} - V_{ref} \right| \cdot \mathbb{I}(V_{i,t} \in \Omega) + w_3 P_{loss,t} \\ + w_4 \frac{1}{|K|} \sum_{k=1}^{|K|} (SOC_{k,t} - 0.5)^2 + w_5 \frac{1}{|A|} \sum_{j=1}^{|A|} a_{j,t}^2 \end{array} \right) \tag{9}$$

where V_{ref} is nominal voltage; Ω denotes the voltage-security limits; $\mathbb{I}(\cdot)$ is the indicator function; $P_{loss,t}$ is total active-power loss at time t; $SOC_{k,t}$ is the state of charge of storage unit k; and w_1, w_2, w_3, w_4 and w_5 are the weights for the voltage-violation, in-band deviation,

active-power-loss, state-of-charge and control-effort penalties, respectively. Here, V_{max} and V_{min} are the upper and lower limits of the admissible voltage band, $|K|$ is the number of storage units, $|A|$ is the action-space dimension, which is 23 in this study (15 PV inverters, 4 SVCs and 4 storage units), and $a_{j,t}$ is the $_j$-th normalized action command at time $_t$.

When no bus leaves the admissible band, the reward additionally includes a zero-violation bonus and a term proportional to the remaining maximum deviation from 1.0 p.u.; the implemented training reward therefore consists of the five penalty terms listed above together with these two safety terms, and Equation (9) is stated here in the form used in the experiments.

The reward guides policy training by penalizing voltage deviation and violations, active-power loss, and sustained departure of storage state of charge from its desired range. It is used only as a training signal. Final performance is evaluated independently on the frozen test set using whole-step and bus-time violation rates, maximum voltage deviation, average daily violation steps, and average daily energy loss.

The coefficient values used in all reported experiments are as follows: $w_{voltage} = 500$ (voltage-violation penalty), $w_{minor} = 0.1$ (deviation penalty inside the admissible band), $w_{loss} = 250$ (active-power-loss penalty), $w_{soc} = 2$ (state-of-charge penalty) and $w_{action} = 8$ (control-effort penalty). A zero-violation bonus of 60 is added when no bus violates the band, together with a deviation-reduction slope of 400. These coefficients were selected by engineering reasoning about the relative cost of a voltage violation against a unit of energy loss, and were then refined on training-side validation; the frozen test set was never used to select them. The entropy coefficient of SAC is tuned automatically rather than held at a fixed small value, because a fixed small coefficient was found to drive the policy toward saturated actions and to increase energy loss without improving voltage security.

2.3. GAT-SimAM Feature Extractor and VVC Framework

Figure 1 presents the proposed VVC framework. Bus operating states and feeder topology are converted into graph observations; the GAT-SimAM extractor aggregates topology-dependent information and recalibrates node-feature channels, and the SAC policy generates continuous VVC actions. Voltage-violation penalties during training and execution-time action scaling form two distinct constraint-handling components.

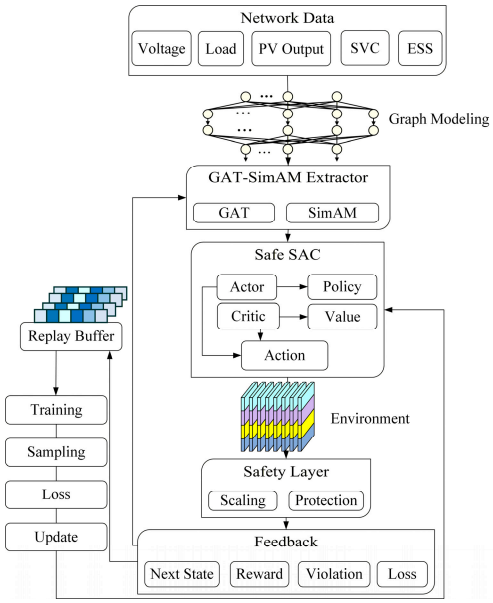


Figure 1. GAT-SimAM-based voltage-var control framework.

The GAT-SimAM extractor lies between graph-state construction and SAC decision making. The following subsections describe the parameter-free SimAM operation, GAT aggregation, and the fused network architecture.

2.3.1. Node-Wise SimAM Recalibration

SimAM calculates feature weights from a neuron-energy function without introducing trainable parameters [41]. For each node representation, channels that differ more strongly from the within-node channel statistics receive different energy values and are recalibrated accordingly. The energy function is defined as follows:

$$E_i = \frac{\hat{\sigma}^2 + \lambda}{4\left[(x_i - \hat{\mu})^2 + \lambda\right]} \tag{10}$$

where x_i is the feature vector of node i; $\hat{\mu}$ and $\hat{\sigma}^2$ are the mean and variance calculated across feature channels of that node; and λ is a numerical-stability constant.

Feature recalibration is performed as follows:

$$x_{i'} = \text{Sigmoid}(E_i) \cdot x_i \tag{11}$$

where the mean $\hat{\mu}$ and variance $\hat{\sigma}^2$ are computed independently across the feature channels of each node; Sigmoid(E_i) maps the gate to (0, 1). Channel deviations determine the energy E_i and the resulting within-node weights. This implementation is the numerical form of the original SimAM energy E_i.

SimAM uses the energy function to measure the statistical distinctiveness of each feature channel within a node and generates a parameter-free gate. It preserves the feature dimension and the parameter count while strengthening informative channels that may otherwise be weakened by graph aggregation. The operation recalibrates channels along the feature dimension of each node; it does not compare neighbouring buses, and its effect on the representation is quantified by the probe analysis reported in Section 4.2 rather than assumed.

2.3.2. Network Architecture

GAT aggregates node features using learned attention coefficients. For node i and neighboring node j at layer l, the coefficient is defined in Equation (12).

$$e_{ij}^{(l)} = \text{LeakyReLU}\left(a^{(l)\top}\left[W^{(l)}h_i^{(l)} \| W^{(l)}h_j^{(l)}\right]\right) \tag{12}$$

where W is the shared linear-transformation matrix; a is the learnable attention vector; $h_i^{(l)}$ and $h_j^{(l)}$ are the layer-l input features of nodes i and j, respectively.

Attention scores are normalized by softmax over the neighborhood of node i, as expressed in Equation (13).

$$\alpha_{ij}^{(l)} = \frac{\exp\left(e_{ij}^{(l)}\right)}{\sum_{k\in\mathcal{N}(i)\cup\{i\}}\exp(e_{ik}^{(l)})} \tag{13}$$

Each GAT layer uses four attention heads. Preliminary comparisons indicated that two heads provided insufficient diversity in neighborhood patterns, whereas eight heads reduced each first-layer head to eight channels and produced less stable training. Four heads provided a practical balance between per-head representation capacity and attention diversity.

As shown in Figure 2, data flow passes through input normalization, two GAT aggregation layers, SimAM recalibration, and global-feature fusion. The layers are configured as follows.

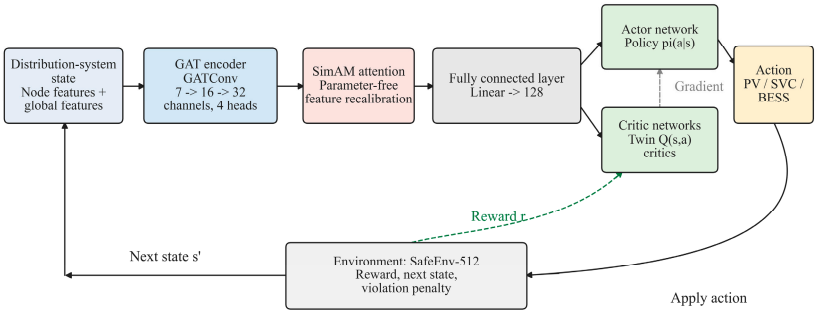


Figure 2. Architecture of the GAT-SimAM feature extractor.

(1) Input layer: The node-feature matrix $\mathbf{X}_t \in R^{N\times 7}$ contains seven features for each of N = 123 buses: voltage magnitude, active- and reactive-power injections, PV active power, PV reactive power, SVC reactive power and storage state of charge. The maximum voltage deviation and the total active-power loss are not node channels: they form the two-dimensional global feature vector, which is concatenated with the flattened graph representation before the output layer. Layer normalization removes scale differences before graph aggregation.

(2) First GAT layer: Four attention heads process the normalized input and each producing 16-channel outputs. Concatenation produces an $N \times 7$ representation from the $N \times 64$ input. Layer normalization and LeakyReLU follow the aggregation.

(3) Second GAT layer: Four heads further process the $N \times 64$ representation and their outputs are averaged to obtain an $N \times 32$ node representation. Averaging suppresses atypical single-head weights and reduces downstream computation.

(4) SimAM layer: Parameter-free channel recalibration is applied independently to every node in the $N \times 32$ representation. The output remains $N \times 32$. No trainable parameters are introduced, and statistics are computed within each node rather than across nodes.

(5) Global fusion: The $N \times 32$ SimAM output is flattened to $123 \times 32 = 3936$ features and concatenated with two global features, maximum voltage deviation and total active-power loss. A fully connected layer projects the resulting 3938-dimensional vector to 128 dimensions for SAC.

The two GAT layers follow an expand-then-compress encoder design. Preliminary experiments found that the 64-to-32-channel configuration achieved a lower violation rate than a 32-to-16 configuration and was more stable than a 128-to-64 configuration, while using approximately 60% fewer parameters than the latter.

2.4. SAC Decision Framework and Voltage-Constraint Handling

SAC is selected because the VVC task combines a high-dimensional continuous action space, nonstationary disturbances, and relatively sparse voltage-violation events. Its maximum-entropy objective balances exploration and exploitation and supports stable off-policy learning [42].

The complete controller, denoted GAT-SimAM in the results, combines the GAT-SimAM extractor with SAC. The critic minimizes a Bellman-residual objective, the policy maximizes expected return with entropy regularization, and target networks are updated softly. Voltage constraints are handled by two separable mechanisms: a violation-related penalty used during policy training and execution-time action scaling used to reduce

the magnitude of actions that would otherwise create unsafe operating points. Matched ablations distinguish the contribution of each mechanism.

Training and evaluation are strictly separated. Training uses feeder topology and full power-flow feedback in the simulation environment. Evaluation uses fixed policy parameters and the same frozen test sequence for every group; the observation supplied to the learned controller retains the specified one-control-step measurement delay.

3. Experimental Setup and Evaluation Protocol

3.1. Test Feeder and Training Configuration

The test platform is a modified IEEE 123-bus radial feeder with a nominal voltage of 4.16 kV, 123 buses, and 122 branches. Typical overhead-line parameters are 0.50 ohm/km resistance and 0.45 ohm/km reactance, with a total feeder length of 12.2 km. All 122 non-slack buses carry loads, and the simultaneous annual peak is 3.5 MW and 1.40 MVAr. Fifteen distributed PV units, each rated at 0.60 MW, yield an installed-PV-to-peak-load ratio of approximately 2.6. Four SVCs and four battery systems are also installed; their locations and capacities are given in Table 1. The substation tap is fixed at its day-ahead setting and maintains the source-bus voltage at 1.04 p.u. The feeder is implemented in pandapower 3.3.2, and Newton–Raphson power flow uses a convergence tolerance of 1×10^{-6} p.u. The control interval is 3 min, giving 480 steps per day. A frozen test set of 96 days is sampled at equal intervals from the annual sequence, resulting in 46,080 evaluation steps. These days are not used during training.

Table 1. Locations and capacities of controllable voltage-regulation devices.

Device	Number	IEEE Bus Locations	Unit Rating	Adjustable Range
Distributed PV	15	5, 11, 12, 25, 32, 35, 45, 51, 56, 66, 75, 82, 85, 101, 110	0.60 MW active; 0.667 MVA apparent	Reactive limit from inverter capability circle; ±0.29 MVAr at rated active power
SVC	4	31, 50, 74, 110	0.30 MVAr	−0.30 to +0.30 MVAr
Battery storage	4	11, 33, 55, 80	1.0 MWh	−0.20 to +0.20 MW

The distributed PV units are grid-connected through grid-following inverters that supply reactive power for voltage support within their apparent-power capability; the reactive capability corresponds to a power-factor limit of 0.90, which is how the PV units participate in voltage control through Equation (4). The units are not operated at unity power factor and no islanded or grid-forming operation is considered. The point of common coupling with the upstream grid is shown in Figure 1, and Figure 3 places the PV output, the load demand and the commands issued to the three controllable resource types on the same grid-connected feeder.

To characterize how the difficulty of the problem depends on the PV-to-load ratio, a penetration sweep was carried out at ratios of 1.0, 1.5, 2.0 and 2.57 under otherwise identical settings. The zero-action whole-step violation rate rises monotonically across the sweep, at approximately 1.0%, 6.5%, 12.5% and 16.7% respectively, and the cost of corrective action rises with the ratio: remedial control remains effective at the lower levels, whereas at the highest level the admissible action window narrows substantially. The ratio of approximately 2.6 used in the main experiments is therefore a deliberate stress case rather than the most convenient operating point, and it is the regime in which the difference between control strategies is largest.

The sweep was then extended to include the learned controllers and the engineering baseline, so that the comparison covers what each strategy achieves at each ratio rather than the uncontrolled stress alone. Five configurations were evaluated at each of the four

ratios: zero-action output, Q-V droop control, SAC with an MLP extractor, SAC with a GAT extractor and the complete GAT-SimAM controller. The measured whole-step violation rates are 0.89%, 0.00%, 0.00%, 0.52% and 0.00% at a ratio of 1.0; 6.51%, 2.40%, 1.41%, 3.07% and 0.83% at 1.5; 12.45%, 5.89%, 15.62%, 21.51% and 21.41% at 2.0; and 16.77%, 9.90%, 28.65%, 63.12% and 57.71% at 2.57, in the order of the configurations just listed. Two features of these results are stated explicitly rather than left to be inferred. First, at the ratio at which the controllers were trained, the complete controller attains the lowest violation rate of the five configurations, and the SimAM effect is present there as well, at 0.83% against 3.07% for the matched GAT extractor. Second, at ratios above the training point, the learned controllers degrade sharply, and the fixed-gain droop baseline becomes the most robust strategy across the whole sweep: at ratios of 2.0 and 2.57 it reaches 5.89% and 9.90%, where the complete controller reaches 21.41% and 57.71%, and at both of those ratios the learned controllers perform worse than zero-action output. This behaviour is not a statement about the controller's training quality so much as about the transfer of a fixed policy across operating regimes. The learned controllers were trained at a single ratio, and in this model the per-unit rating against which their continuous commands are scaled is itself a function of the PV-to-load ratio, so the meaning of an action changes with the ratio even though the command range does not. The sweep is therefore an out-of-distribution test of a fixed policy rather than a report of controllers retrained at each ratio, and it indicates that the advantage measured in the main experiments is confined to the operating regime in which the policy was trained. Retraining each configuration at each ratio would require one complete training run per cell and is not reported here.

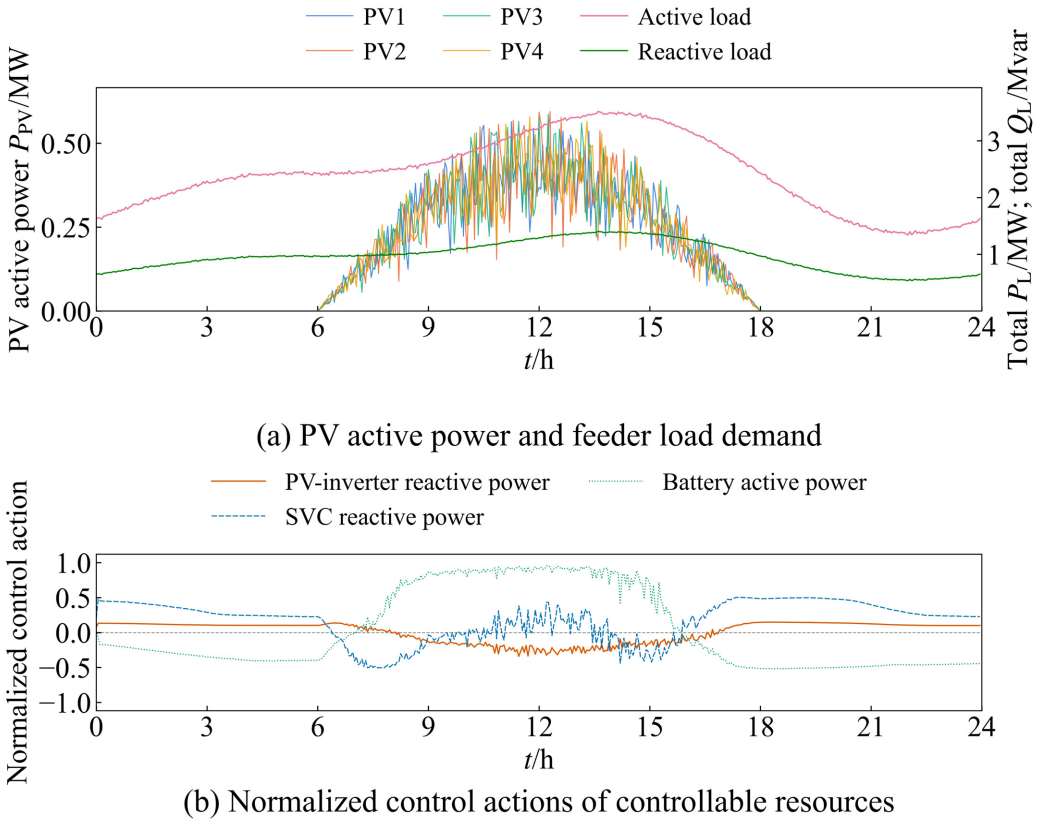


Figure 3. PV output, load demand, and controllable-resource trajectories on a representative day.

Experiments were run under Windows 11 (Microsoft Corporation, Redmond, WA, USA) with Python 3.10.19. The power flow was computed with pandapower 3.3.2, the SAC agent was built with Stable-Baselines3 2.7.1 and Gymnasium 1.2.3, and the graph

attention layers used PyTorch Geometric 2.7.0 on PyTorch 2.5.1 with CUDA 12.4. Training was carried out on an NVIDIA GeForce RTX 4070 SUPER graphics card (NVIDIA Corporation, Santa Clara, CA, USA). Four parallel environments collect training transitions synchronously from annual load and PV profiles, and model parameters are optimized with Adam. Figure 3 shows the PV and load profiles for a representative day together with the corresponding commands for the three controllable resource types.

3.2. Comparison Groups

Ten methods are evaluated under identical device-capacity constraints and the same frozen test sequence. Table 2 includes zero-action output, Q-V droop control, SAC with an MLP feature extractor, SAC with a GAT feature extractor, an attention-based variant that replaces SimAM with a squeeze-and-excitation module, three GAT-SimAM variants for voltage-constraint ablation, and centralized model-based optimization. For the matched SimAM comparison, SAC with a GAT feature extractor, the squeeze-and-excitation variant and GAT-SimAM with the training-side voltage penalty differ only in the attention module, and none of the three uses execution-time action scaling. For the voltage-constraint ablation, the four GAT-based variants use no constraint-handling mechanism with and without SimAM, the training-side voltage penalty only, and the combination of the penalty and execution-time action scaling, respectively.

Table 2. Configuration of comparison and ablation groups.

Method	Core Configuration	Purpose
Zero-action output	All control commands set to zero	Quantify the need for active control
Q-V droop control	Droop coefficients 8.0/10.0	Engineering baseline
SAC with an MLP feature extractor	MLP feature extractor; training penalty	Feature-extractor comparison
SAC with a GAT feature extractor	GAT extractor; training penalty; no SimAM or execution scaling	Matched baseline for SimAM
SAC with a squeeze-and-excitation extractor	GAT extractor with a trainable squeeze-and-excitation module; training penalty; no SimAM or execution scaling	Attention-module baseline at the same insertion point
GAT-SimAM without voltage-constraint handling	No voltage penalty; no execution scaling	Ablate both safety components
SAC with a GAT feature extractor and no voltage-constraint handling	GAT extractor; no voltage penalty; no execution scaling	Missing factorial cell: GAT without SimAM and without the penalty
GAT-SimAM with training-time voltage penalty only	Voltage penalty; no execution scaling	Isolate training-side penalty and SimAM
Full GAT-SimAM controller	Voltage penalty and execution-time action scaling	Complete method
Centralized model-based optimization	Exact model and current full-network measurement at each step	Strong same-capacity comparator

3.3. Evaluation Metrics

Performance is assessed by voltage-security, operating-economy, and computational-efficiency metrics.

Voltage security is quantified by whole-step violation rate, bus-time violation rate, average daily violation steps, and maximum voltage deviation. A whole step is counted as violating when any bus lies outside the prescribed band. The bus-time rate is the number of violating bus-time pairs divided by the product of bus count and time-step count. Maximum deviation is the largest absolute departure from 1.0 p.u. over the test period. Results are reported for the $\pm 5\%$ study band and, where stated, for the $\pm 7\%$ limit

specified for three-phase supply at 20 kV and below by GB/T 12325-2008 [43]. Operating economy is measured by average daily active-energy loss. Single-step inference time is used to assess online applicability. All aggregate metrics use the same 96-day, 46,080-step test set.

Because the reported quantities are aggregates over a distribution of operating conditions rather than over a single day, the distribution of the daily violation count is reported alongside its mean. For the complete method, 57 of the 96 test days contain no violation step at all, and the 95th percentile of the daily violation count is two steps; the remaining violations are therefore concentrated in a minority of days rather than spread evenly across the test period. Comparable figures are reported for each configuration so that the mean violation rate can be read together with how the residual violations are distributed.

4. Results

4.1. Overall Method Effectiveness

Table 3 compares the two engineering baselines, the centralized comparator, and the complete proposed method over the frozen test set.

Table 3. Overall effectiveness on the 96-day frozen test set.

Method	Whole-Step Violation Rate (%)	Daily Violation Steps	Maximum Deviation (p.u.)	Daily Loss (MWh)
Zero-action output	15.651	75.13	0.0812	1.42
Q-V droop control	0.701	3.36	0.0701	1.95
Centralized model-based optimization	0.002	0.01	0.0527	1.49
Full GAT-SimAM controller	0.106	0.51	0.0599	2.11

Relative to zero-action output, the proposed GAT-SimAM method reduces the whole-step voltage-violation rate from 15.651% to 0.106% and average daily violation steps from 75.13 to 0.51. The maximum deviation decreases from 0.0812 to 0.0599 p.u. Relative to Q-V droop control, the violation rate decreases by 84.8% and average daily violation steps decrease from 3.36 to 0.51. Under the $\pm 5\%$ study band, the maximum deviation of 0.0599 p.u. indicates that isolated instantaneous violations remain. Under the $\pm 7\%$ limit in GB/T 12325-2008, the proposed method records no violations over the 96-day test period. Centralized model-based optimization has access to current full-network measurements and an accurate network model at each control step, whereas the learned controller operates with a one-step-delayed observation. It therefore serves as a strong same-capacity comparator rather than a directly deployable policy. The proposed method requires 2.57 ms per inference step, far below the three-minute control interval.

Recent work on voltage control has moved toward graph-based and safety-constrained learning. Physics-informed graphical representations have been combined with deep reinforcement learning for robust distribution-system voltage control [31], and graph representations have also been used with multi-agent formulations, either physically assisted [32] or based on graph attention for post-contingency emergency control [33]; spatial-temporal graph information has been used for voltage control and renewable accommodation [34]. On the constraint-handling side, safe reinforcement learning with an explicit constrained formulation [18] and human-in-the-loop correction of unsafe actions [19] have both been studied. These methods are evaluated on different feeders with different observation spaces, action sets and device-capacity limits, and their implementations are not available in a form that can be executed under the protocol used here. We therefore report them as a qualitative reference for the direction of the field rather than as numerically comparable entries in Tables 3–5, and we state this explicitly. Within the identical protocol the comparison is

made against the zero-action output, a rule-based Q-V droop controller, and a centralized model-based optimization with full measurement and model information.

Figure 4 compares daily violation steps for SAC with a GAT feature extractor and the proposed GAT-SimAM method. Figure 5 presents the maximum voltage-deviation trajectory on a representative day and the daily maxima over all 96 test days. Figure 6 compares the active-energy loss of SAC with an MLP feature extractor and the proposed GAT-SimAM controller.

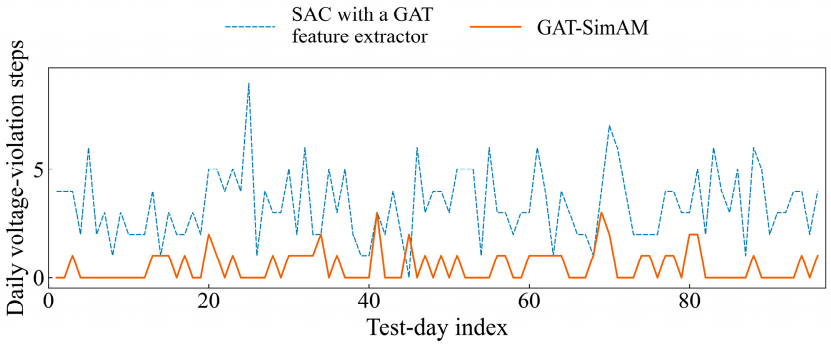


Figure 4. Daily voltage-violation steps over the 96-day frozen test set: SAC with a GAT feature extractor versus the complete GAT-SimAM method.

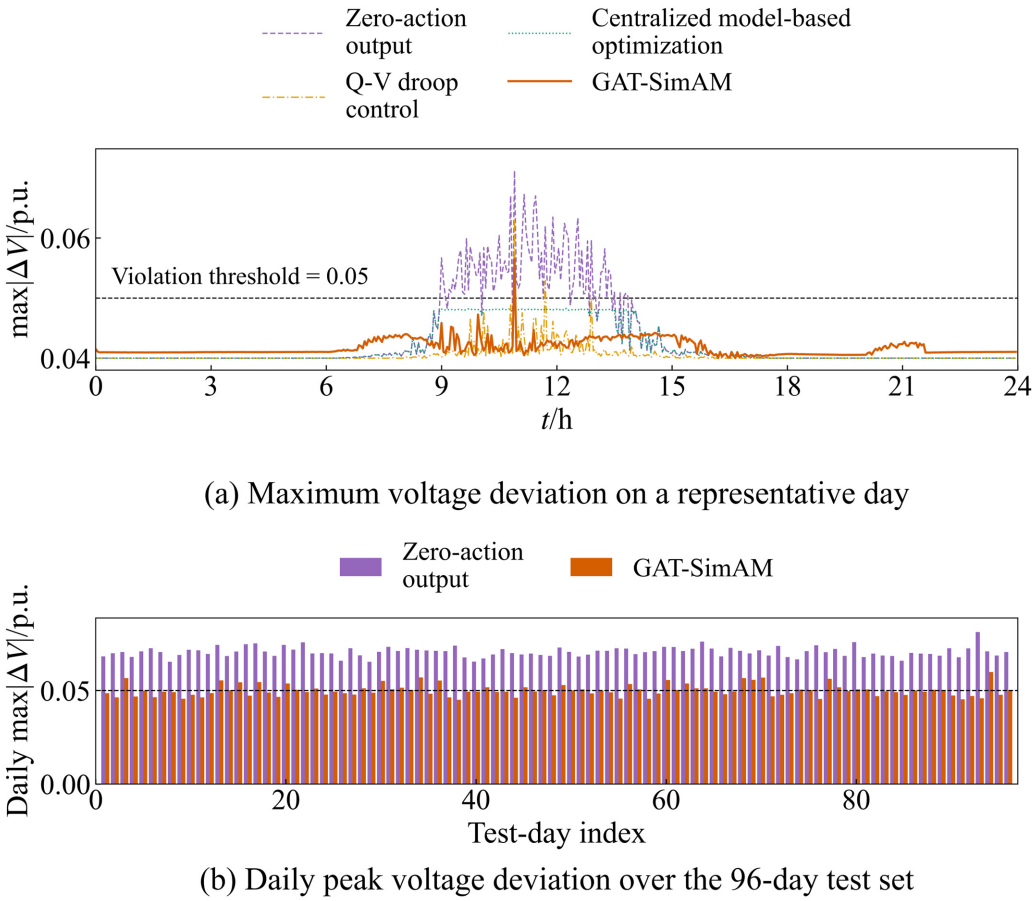


Figure 5. Maximum voltage-deviation time series on a representative day and daily peak deviations over the 96-day test set.

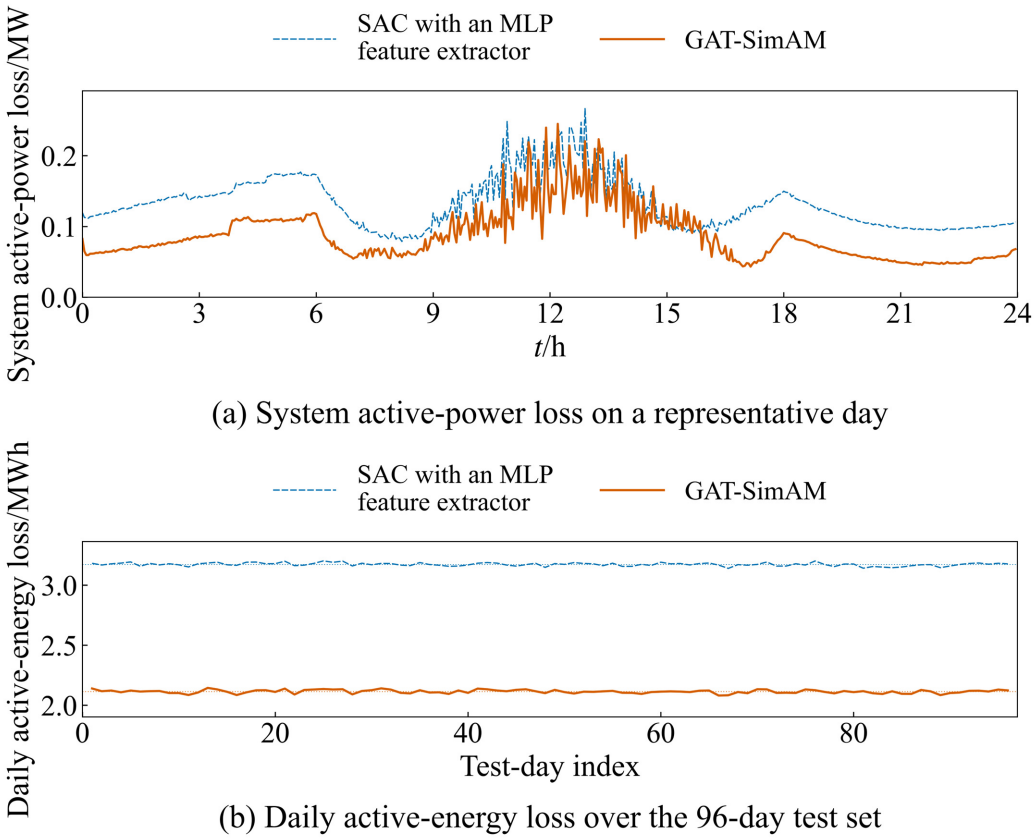


Figure 6. Active-energy-loss comparison between SAC with an MLP feature extractor and the complete GAT-SimAM method.

Table 4. Matched feature-extractor attribution on the 96-day test set.

Method	Whole-Step Violation Rate (%)	Daily Violation Steps	Daily Loss (MWh)
SAC with an MLP feature extractor	0.002	0.01	3.17
SAC with a GAT feature extractor	0.697	3.34	2.04
SAC with a squeeze-and-excitation extractor	0.534	2.56	1.94
GAT-SimAM with training-time voltage penalty only	0.109	0.52	2.11

Table 5. Ablation of voltage-constraint handling on the 96-day test set.

Method	Whole-Step Violation Rate (%)	Daily Violation Steps
GAT-SimAM without voltage-constraint handling	30.046	144.22
SAC with a GAT feature extractor, without voltage-constraint handling	26.319	126.33
GAT-SimAM with training-time voltage penalty only	0.109	0.52
Full GAT-SimAM controller	0.106	0.51

4.2. Feature-Extractor Attribution

Table 4 reports the matched feature-extractor and attention-module comparison. In Table 4, the first configuration uses a multilayer perceptron extractor, the second replaces it with GAT, the third adds a trainable squeeze-and-excitation module at the position where SimAM is applied, and the fourth adds SimAM to the second. All four retain the same

training-side voltage penalty and disable execution-time action scaling, so the comparison isolates the representation and the attention module rather than the constraint mechanism.

Two readings follow from Table 4. First, the parameter-free recalibration achieves the lowest violation rate of the four configurations, at 0.109% against 0.534% for the trainable attention module and 0.697% for the GAT extractor without attention. The squeeze-and-excitation module does improve on the bare GAT extractor, so the ordering is not an artefact of attention being unhelpful in general; at the same insertion point, however, the parameter-free variant reduces the violation rate by roughly a factor of five. The comparison is not uniformly in favour of SimAM on every metric: the squeeze-and-excitation variant records a lower maximum voltage deviation, 0.0554 p.u. against 0.0599 p.u. for the complete controller, while its violation rate is about five times higher. The two modules therefore trade a slightly better worst-case excursion against a substantially worse aggregate violation rate. Second, the advantage of the parameter-free variant is obtained without adding trainable parameters or inference cost. The two modules occupy the same position in the architecture and act on the same tensor, so the difference between them is attributable to how the recalibration is derived rather than to the depth or width of the network.

This trade-off indicates that graph-based allocation reduces excessive control effort, while the representation the policy acts on can lose some of the bus-local detail that distinguishes an at-risk bus from its neighbours. The probe analysis below quantifies how much of that detail is recovered by the recalibration step.

To connect the architecture to the representation it produces and, through it, to the control outcome, the node representations of the trained extractor were probed directly. Because the controller suppresses violations to the order of 10^{-2}% on the frozen test set, a binary violating/non-violating classification probe has no usable positive class, so a closed-form ridge-regression probe predicting the bus voltage-deviation magnitude from the node representation was used instead, over 236,160 bus-step samples with five-fold cross-validation. Predictive R^2 rises monotonically with depth: 0.0726 at the input, 0.1892 after the first GAT layer and 0.2271 after the second for SAC + GAT, and the SimAM recalibration adds a further step from 0.2538 to 0.2707. The layer-wise statistics of the representations were measured as well: the mean cosine similarity between adjacent nodes decreases with depth (0.9874 to 0.9708 to 0.9088) while the cross-node variance increases (0.0424 to 0.0896 to 0.2383). At the depth used here, therefore, the aggregation layers separate node representations rather than merging them, and the representation is not homogenized. The same representations were then separated by where the risk actually sits. At each time step the buses were divided into the quartile with the largest voltage deviation and the quartile with the smallest, and the distance between the centroids of the two groups was measured at each depth of the trained extractor, over 6240 steps. Within the complete controller the normalized centroid distance between the two groups is 0.3573 before the recalibration step and 0.3969 after it, an increase of 11.1%; the corresponding cosine distance rises from 0.1297 to 0.1673, an increase of 29.0%. The ratio of between-group to within-group spread, a Fisher-type measure of how far the two populations are separated relative to their own scatter, rises from 8.184 to 8.596 over the same step. For comparison, the GAT extractor without SimAM reaches a centroid distance of 0.3603 and a cosine distance of 0.1404 at the same depth. The recalibration step therefore increases the separation between buses that carry voltage risk and buses that do not, in the representation the policy acts on, and it does so at the layer where the change is measured. The comparison that supports this reading is the one made within a single trained model, before and after the recalibration: the two representations compared there differ only by that step. At the depth before the recalibration, the two extractors are close to indistinguishable on this measure, with the GAT extractor at a centroid distance of 0.3603 and the complete controller at 0.3573, so the

separation between the two bus populations is established by the recalibration step itself rather than by the aggregation that precedes it.

4.3. Voltage-Constraint Ablation

Table 5 separates the effects of the training-side voltage penalty and execution-time action scaling, and also reports the GAT extractor without SimAM under the two constraint settings, so that the factor design is crossed rather than incomplete.

With GAT-SimAM but without the voltage-constraint mechanism, the whole-step violation rate is 30.046% and average daily violation steps reach 144.22. Adding the training-side voltage penalty while retaining unscaled execution reduces these values to 0.109% and 0.52, respectively. Adding execution-time action scaling to form the proposed GAT-SimAM method further changes them to 0.106% and 0.51. The reduction from 30.046% to 0.109% identifies the training-side penalty as the dominant source of aggregate safety improvement. The smaller reduction from 0.109% to 0.106% indicates that execution scaling provides an incremental safeguard rather than explaining the main reduction. This three-way comparison avoids attributing the combined change from 30.046% to 0.106% to a single voltage-constraint mechanism. The same ablation is also reported for the GAT extractor without SimAM: without the training-side penalty it reaches 26.319% and 126.33 daily violation steps, and with the penalty, it reaches 0.697% and 3.34.

Read as a factorial design over the two factors—the feature extractor and the voltage-constraint mechanism—the four GAT-based configurations show a substantial interaction rather than two independent marginal effects. Without the training-side penalty, adding SimAM to the GAT extractor changes the violation rate from 26.319% to 30.046%; that is, it does not improve safety. With the penalty active, the same addition changes the rate from 0.697% to 0.109%, a reduction of 84.4%. The benefit of the SimAM representation therefore appears only when the voltage constraint is active, and the sign of its effect reverses when it is removed. Representation and constraint have to be assessed jointly, because an improvement in representation quality has no channel through which to reach the control outcome unless the constraint mechanism is present. This is the full two-by-two crossing of the extractor factor with the training-side penalty factor: GAT and GAT-SimAM, each with and without the penalty. Execution-time action scaling is reported separately in Table 5 and is held at its disabled setting throughout the crossing.

4.4. Multi-Metric Assessment

Across Tables 3–5, the proposed controller maintains a low voltage-risk level under a common test protocol. Under the $\pm 5\%$ study band, SAC with an MLP feature extractor attains near-zero violations but has an average daily loss of 3.17 MWh, 50.1% above the 2.11 MWh of the proposed method. Q-V droop has a lower average daily loss of 1.95 MWh but a violation rate approximately 6.6 times that of the proposed GAT-SimAM method. Under the $\pm 7\%$ national-standard band, both SAC with an MLP feature extractor and the proposed method record no violations. These results show that the proposed method does not optimize a single metric in isolation: it substantially reduces violations relative to the engineering baseline while keeping energy loss near that baseline and well below the MLP controller.

Figure 7 places the operating points of all eight evaluated configurations in the violation-rate/daily-energy-loss plane, so that the position of each configuration relative to the others is visible in one view rather than across separate tables. The trade-off is the organizing feature of the result: the configurations that achieve the lowest violation rates are not the ones with the lowest energy loss, and the proposed controller occupies an intermediate position that is close to the engineering baseline in loss while reducing the violation rate by 84.8%.

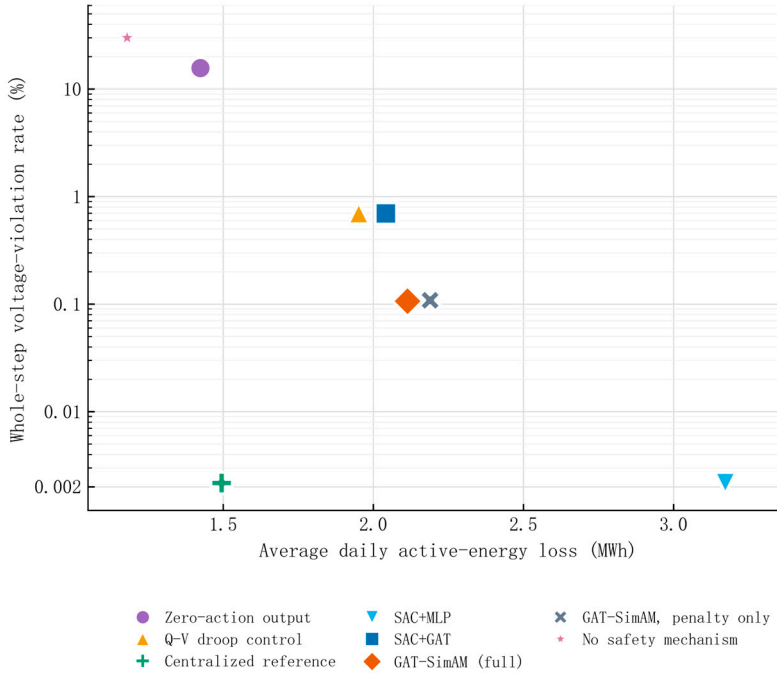


Figure 7. Security-economy operating points of the evaluated configurations on the 96-day frozen test set. The complete GAT-SimAM controller and its penalty-only variant have nearly coincident operating points, so a small horizontal offset is applied to separate them and keep both visible; the numerical values are unchanged.

Figure 8 shows how the position of the operating point in this plane is set by the loss coefficient. The coefficient w_{loss} in Equation (9), which multiplies the active-power-loss penalty, was evaluated at four settings under the identical frozen protocol: 5, 100, 250 and 500. The measured operating points are 0.109% at 4.222 MWh/d for w_{loss} = 5, 0.041% at 2.714 MWh/d for 100, 0.106% at 2.114 MWh/d for 250, and 0.668% at 2.119 MWh/d for 500. Two of the four settings are non-dominated. At w_{loss} = 5 the penalty is too small relative to the zero-violation bonus for the policy to register the cost of unnecessary action, so the point is dominated by w_{loss} = 100 on both axes: it is neither safer nor more economical. Raising the weight from 100 to 250 trades 0.065 percentage points of violation rate for a further 0.60 MWh/d of energy saving, which is the trade-off the coefficient is there to express. Beyond that the mechanism saturates and then reverses: at w_{loss} = 500 the daily loss no longer improves, while the violation rate rises to 0.668%, more than six times the rate at 250, because the loss penalty approaches the magnitude of the zero-violation bonus and the policy begins to accept violations in exchange for energy. The setting used in this paper, w_{loss} = 250, is the economical end of the measured non-dominated set, and the sweep shows that both a much smaller and a much larger coefficient degrade the operating point, in different ways. These four settings were produced during the development of the reward function rather than as a controlled one-factor sweep, and they are reported as a preliminary trade-off analysis of the operating point rather than as a definitive Pareto front; a dense front sampled over a grid of all reward coefficients would require one complete training run per grid point and is left as future work. The voltage-violation coefficient $w_{voltage}$ was held at its reported value throughout and has not been swept; only the loss coefficient was varied, so the evidence reported here bounds the operating point along one direction of the trade-off rather than over the full coefficient grid.

Figure 9 compares the same configurations simultaneously across five metrics—the number of zero-violation days and the reciprocals of the whole-step violation rate, the bus-time violation rate, the daily energy loss and the maximum voltage deviation—each

normalized to the range spanned by the evaluated configurations and oriented so that a larger area is better. No configuration dominates on every axis, which is the point: the proposed method is characterized by its balance across the five metrics rather than by a single best value.

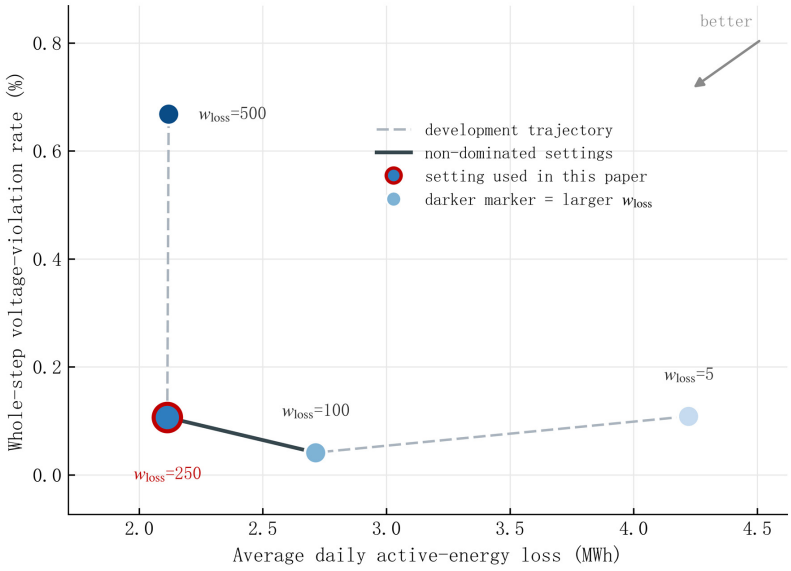


Figure 8. Preliminary trade-off analysis: operating points of the complete controller under four settings of the loss coefficient w_{loss}, evaluated on the 96-day frozen test set. The dashed line traces the development trajectory in order of increasing coefficient; the solid segment connects the two non-dominated settings. Both axes are oriented so that lower is better.

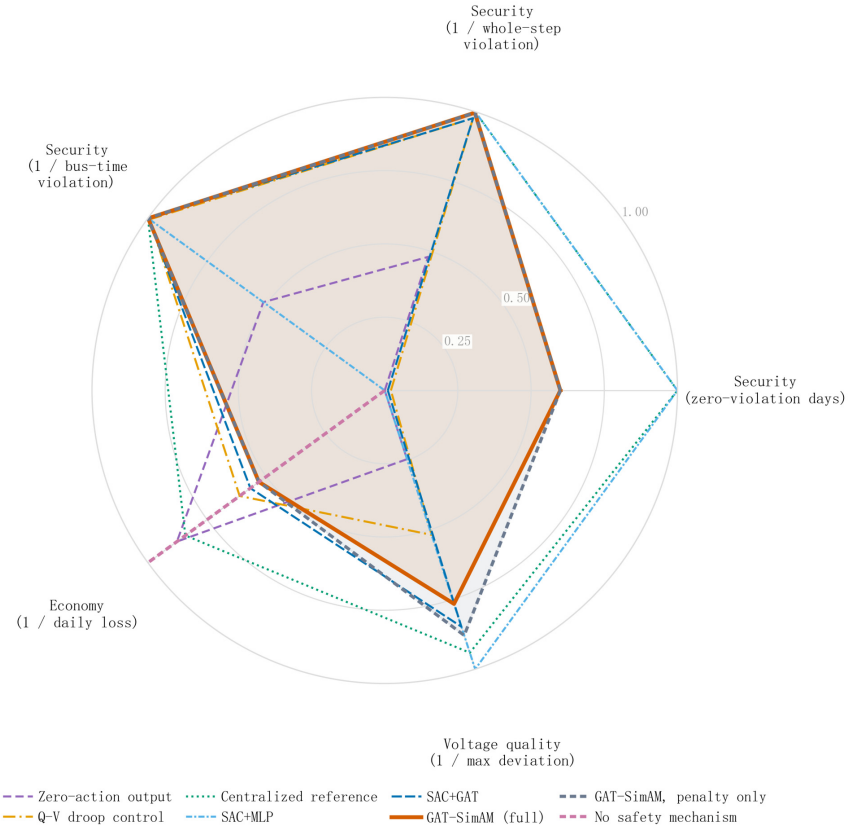


Figure 9. Multi-metric comparison of the evaluated configurations. Each axis is normalised to the range spanned by the configurations shown and oriented so that larger is better.

5. Discussion

5.1. Sources of Performance Improvement

The ablation results indicate that the two mechanisms introduced in this work act on different parts of the control problem. Topology-aware aggregation and parameter-free channel recalibration change how the controller represents the network, whereas the training-side voltage penalty and execution-time scaling change how strictly it is driven toward feasible operating points. The matched comparison in Section 4.2 isolates the representation effect by holding the constraint-handling configuration fixed, and the three-way comparison in Section 4.3 isolates the constraint-handling effect while holding the extractor fixed. Read together, they show that the dominant reduction in aggregate voltage risk originates from the training-side penalty, while the extractor determines how much control effort the policy expends to reach comparable voltage security. One further point concerns the representational mechanism. The homogenization concern discussed in Section 1 is a property reported for deep graph networks, and the layer-wise measurements in Section 4.2 show that it is not present at the depth used here: adjacent-node similarity decreases with depth while cross-node variance increases. The benefit of the recalibration step is therefore attributed to its effect on how strongly each node's representation encodes its own voltage-deviation level, not to any restoration of information that aggregation had removed.

5.2. Implications for Deployment

This separation has practical consequences for deployment. Because the voltage penalty is applied only during training, no constraint-solving routine is required at execution time, and the deployed controller performs a single feed-forward inference per control interval. The measured inference time is therefore governed by the extractor rather than by any online optimization. The parameter-free nature of the recalibration step also means that the representation change does not increase the number of trainable parameters, so improved voltage-risk sensitivity is obtained without a corresponding increase in model size or inference cost. The method remains compatible with the three-minute control interval and with the existing device set, which is relevant for feeders where communication and computation budgets are limited.

5.3. Scope of the Evaluation

The evaluation was conducted primarily on a modified IEEE 123-bus feeder with a fixed device placement, one year of disturbance data, and the simulation protocol described in Section 3. To indicate how far the reported behavior depends on the size of the feeder, the same procedure was then applied in full to two further systems at smaller scale: a modified IEEE 33-bus feeder at a PV penetration of approximately 1.5 and a modified IEEE 69-bus feeder at a PV penetration of approximately 1.0, each evaluated over its own 286-day frozen test set drawn from the same three-year annual profile. On the 33-bus feeder five configurations were evaluated under that protocol: zero-action output, Q-V droop control, SAC with an MLP feature extractor, SAC with a GAT feature extractor, and the complete GAT-SimAM controller. The whole-step violation rate was 6.356% without control and 2.524% with Q-V droop control; among the learned configurations, it was 1.530% with the MLP extractor, 3.311% with the GAT extractor and 0.994% with the complete controller. The complete controller therefore attains the lowest violation rate of the five configurations at this feeder scale, and the SimAM recalibration reduces the violation rate by 70.0% relative to the matched GAT extractor. On the 69-bus feeder the same five configurations plus the unpenalized GAT-SimAM variant were evaluated. The violation rate was 15.807% without control and 4.079% with Q-V droop control; among the learned configurations,

it was 3.056% with the MLP extractor, 7.598% with the GAT extractor and 1.985% with the complete controller. The complete controller again attains the lowest violation rate, and the recalibration reduces it by 73.9% relative to the matched GAT extractor. The three reductions—84.4% on the 123-bus feeder between the matched pair in Table 4, 70.0% on the 33-bus feeder and 73.9% on the 69-bus feeder—agree in direction and in order of magnitude across a fourfold range of feeder size. Figure 10 places the three comparisons side by side. The droop gains were calibrated separately for each feeder—4.0 on the 33-bus feeder and 32.0 on the 69-bus feeder, against 8.0 for the PV inverters and 10.0 for the SVCs on the 123-bus feeder—using the same rule in each case: the largest gain before the violation rate begins to reverse, so that the baseline operates at its own best point rather than a handicapped one.

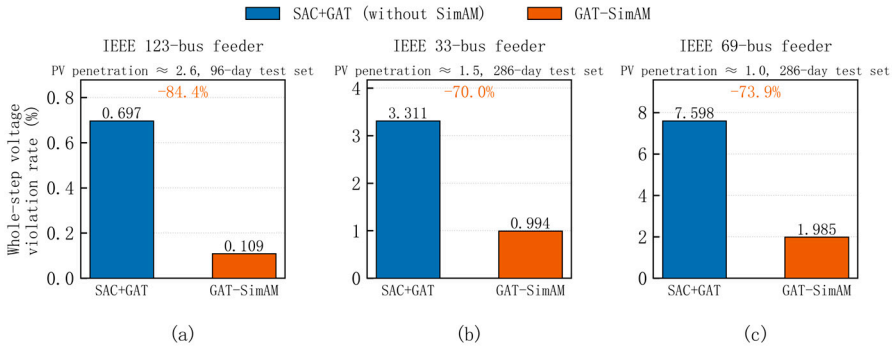


Figure 10. SimAM ablation at three feeder scales: (**a**) modified IEEE 123-bus feeder; (**b**) modified IEEE 33-bus feeder; (**c**) modified IEEE 69-bus feeder. Whole-step voltage violation rate over the respective frozen test sets. Note that the vertical axes are not shared, since the violation rates of the three feeders differ by up to an order of magnitude.

One feature of these supplementary comparisons is stated explicitly rather than left implicit. On both smaller feeders the GAT extractor on its own is weaker than the MLP extractor, at 3.311% against 1.530% on the 33-bus feeder and 7.598% against 3.056% on the 69-bus feeder; the same ordering holds on the 123-bus feeder, where the MLP extractor reaches 0.002% against 0.697% for the GAT extractor, although it does so at 3.17 MWh/d against 2.04 MWh/d of daily loss. Adding the parameter-free recalibration step changes the outcome: the GAT-based controller then reaches 0.994% on the 33-bus feeder and 1.985% on the 69-bus feeder, below the MLP extractor in both cases, and 0.106% on the 123-bus feeder at a lower daily loss than the MLP extractor. On the 69-bus feeder the complete controller is ahead of the MLP extractor on four measures at once—whole-step violation rate, bus-time violation rate (0.153% against 0.394%), maximum voltage deviation (0.0814 against 0.0966 p.u.) and daily energy loss (2.247 against 2.407 MWh/d). The contribution of the SimAM step is thus consistent across all three feeder scales, and at the smaller scales it more than compensates for the weaker stand-alone graph representation. Read together, the loss of stand-alone GAT performance at small scale and the recovery obtained from channel recalibration are the same phenomenon seen from two sides rather than two unrelated results. The unpenalized GAT-SimAM variant is reported for completeness on the 69-bus feeder and reaches 16.783%, close to the 15.807% of the zero-action baseline, which is the same dependence on the training-side penalty that the factorial ablation establishes in Section 4.3 at the primary scale.

This behavior is consistent with what has been reported for graph-based policies in voltage-var control. A comparative study of graph-based and vector-based policies [44] found that the two converge to comparable asymptotic rewards, with the graph policy converging more slowly, and that the advantage of the graph representation appears chiefly

as robustness to data acquisition errors—sensor communication failure and measurement misalignment—rather than as a uniform gain in reward. That finding matches the observation above: on a small feeder with complete observations, the benefit of an explicit topology representation is difficult to separate from that of a well-tuned dense representation, while the representation continues to matter where the network is large or the observations are incomplete. The absolute violation levels are not comparable between the three feeders, and the supplementary evaluations are reported as consistency checks on the direction of the SimAM effect rather than as a transfer of absolute performance. The reported violation rates are also conditional on the stated voltage bands, and the two bands lead to different conclusions about whether isolated instantaneous excursions remain. The conclusions of this study should therefore be read as applying to the feeders and protocol examined here rather than as general claims about arbitrary distribution networks [44].

6. Conclusions

This paper presented a SAC-based VVC framework with a GAT-SimAM feature extractor for an active distribution network with high PV penetration. A unified 96-day frozen test protocol was used to evaluate engineering baselines, feature-extractor variants, and voltage-constraint ablations. The principal findings are as follows.

(1) In the matched comparison between SAC with a GAT feature extractor and GAT-SimAM with the training-side voltage penalty and no execution-time scaling, parameter-free SimAM recalibration reduced the whole-step violation rate from 0.697% to 0.109% and average daily violation steps from 3.34 to 0.52, while average daily loss increased from 2.04 to 2.11 MWh.

(2) The three-way voltage-constraint ablation showed that the training-side voltage penalty produced the dominant safety improvement, reducing the violation rate from 30.046% without either constraint-handling mechanism to 0.109% with the training-side penalty alone, while adding execution-time scaling provided a smaller incremental reduction to 0.106%.

(3) Relative to zero-action output and Q-V droop control, the complete method reduced the $\pm 5\%$ whole-step violation rate by 99.3% and 84.8%, respectively, and reduced maximum deviation to 0.0599 p.u. Under the $\pm 7\%$ GB/T 12325-2008 band, no violations occurred over the frozen test set.

(4) Average daily loss was 2.11 MWh, close to the 1.95 MWh droop baseline and substantially below the 3.17 MWh SAC with an MLP feature extractor result. A 2.57 ms inference time was negligible relative to the three-minute control interval.

(5) The advantage of the proposed controller is conditional on the operating regime. At the PV-to-load ratio at which the controller was trained, it attains the lowest violation rate of the evaluated configurations, but at higher ratios the learned controllers degrade substantially, and a fixed-gain Q-V droop controller proves the most robust strategy across the sweep. A policy trained at one ratio should therefore not be expected to transfer to a substantially different one without retraining, and the absolute results reported here apply to the feeders and the operating regime examined.

Author Contributions: Conceptualization, L.W. and Y.Z.; methodology, L.W.; software, L.W.; validation, L.W., Z.S., D.W. and H.S.; formal analysis, L.W.; investigation, L.W.; data curation, L.W.; writing—original draft preparation, L.W.; writing—review and editing, Y.Z., Z.S., D.W. and H.S.; visualization, L.W.; supervision, Y.Z.; project administration, L.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 52107107.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

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

Abbreviations

DRL, deep reinforcement learning; GAT, graph attention network; GNN, graph neural network; MDP, Markov decision process; MLP, multilayer perceptron; PV, photovoltaic; SAC, soft actor-critic; SimAM, simple parameter-free attention module; SVC, static var compensator; VVC, voltage-var control.

References

1. Chen, X.; Zhang, Y.J.; Huang, X.M. Review of reactive-power and voltage-control methods for active distribution networks. *Autom. Electr. Power Syst.* **2016**, *40*, 143–151. (In Chinese)
2. Qi, N.; Cheng, L.; Tian, L.T.; Guo, J.B.; Huang, R.L.; Wang, C.P. Review and prospect of distribution-network planning with flexible loads. *Autom. Electr. Power Syst.* **2020**, *44*, 193–207. (In Chinese)
3. Yang, L.Y.; Ren, X.C.; Cai, Z.P.; Qi, Q.W.; Li, Z.J.; Xia, X.J.; Nian, H. Review of voltage-control strategies for distribution networks with high photovoltaic penetration. *Power Syst. Technol.* **2024**, *48*, 5056–5070. (In Chinese)
4. Mahmud, N.; Zahedi, A. Review of control strategies for voltage regulation of the smart distribution network with high penetration of renewable distributed generation. *Renew. Sustain. Energy Rev.* **2016**, *64*, 582–595. [CrossRef]
5. Zayed, M.E.; Rehman, S.; Elgendy, I.A.; Al-Shaikhi, A.; Mohandes, M.A.; Irshad, K.; Abdelrazik, A.S.; Alam, M.A. Benchmarking reinforcement learning and prototyping development of floating solar power system: Experimental study and LSTM modeling combined with brown-bear optimization algorithm. *Energy Convers. Manag.* **2025**, *332*, 119696. [CrossRef]
6. Zhang, Z.H.; Lei, D.Y.; Jiang, C.H.; Luo, J.W.; Xu, Y.Y.; Li, J. Bi-level planning model and solution method for AC/DC hybrid distribution networks based on second-order cone programming and NNC. *Proc. CSEE* **2023**, *43*, 70–85. (In Chinese)
7. Zhang, Z.; da Silva, F.F.; Guo, Y.; Bak, C.L.; Chen, Z. Double-layer stochastic model predictive voltage control in active distribution networks with high penetration of renewables. *Appl. Energy* **2021**, *302*, 117530. [CrossRef]
8. Jiang, Y.; Wan, C.; Wang, J.; Song, Y.; Dong, Z.Y. Stochastic receding-horizon control of active distribution networks with distributed renewables. *IEEE Trans. Power Syst.* **2019**, *34*, 1325–1341. [CrossRef]
9. Ruan, H.; Gao, H.; Liu, Y.; Wang, L.; Liu, J. Distributed voltage control in active distribution networks considering renewable energy: A novel network-partitioning method. *IEEE Trans. Power Syst.* **2020**, *35*, 4220–4231. [CrossRef]
10. Yu, P.; Wan, C.; Sun, M.; Zhou, Y.; Song, Y. Distributed voltage control of active distribution networks with global sensitivities. *IEEE Trans. Power Syst.* **2022**, *37*, 4214–4228. [CrossRef]
11. Yu, P.; Wan, C.; Qin, H.; Lao, K.W.; Song, Y.; Ju, P. Centralized-distributed coordinated voltage control of active distribution networks with renewables. *IEEE Trans. Sustain. Energy* **2025**, *16*, 1504–1517. [CrossRef]
12. Hou, Q.; Dai, N.; Huang, Y. Voltage-regulation-enhanced hierarchical coordinated volt/var and volt/watt control for active distribution networks with soft open points. *IEEE Trans. Sustain. Energy* **2024**, *15*, 2021–2037. [CrossRef]
13. Li, Q.; Qiao, Y.; Zhang, Y.J. Deep-reinforcement-learning method for continuous reactive-power optimization in distribution networks. *Power Syst. Technol.* **2020**, *44*, 1473–1480. (In Chinese)
14. Zhang, Y.; Wang, X.; Wang, J.; Zhang, Y. Deep-reinforcement-learning-based volt-var optimization in smart distribution systems. *IEEE Trans. Smart Grid* **2021**, *12*, 361–371. [CrossRef]
15. Sun, X.; Qiu, J. Two-stage volt/var control in active distribution networks with a multi-agent deep-reinforcement-learning method. *IEEE Trans. Smart Grid* **2021**, *12*, 2903–2912. [CrossRef]
16. Cao, D.; Zhao, J.; Hu, W.; Ding, F.; Huang, Q.; Chen, Z.; Blaabjerg, F. Data-driven multi-agent deep reinforcement learning for distribution-system decentralized voltage control with high penetration of PVs. *IEEE Trans. Smart Grid* **2021**, *12*, 4137–4150. [CrossRef]
17. Hu, D.; Ye, Z.; Gao, Y.; Ye, Z.; Peng, Y.; Yu, N. Multi-agent deep reinforcement learning for voltage control with coordinated active- and reactive-power optimization. *IEEE Trans. Smart Grid* **2022**, *13*, 4873–4886. [CrossRef]
18. Kou, P.; Liang, D.; Wang, C.; Wu, Z.; Gao, L. Safe-deep-reinforcement-learning-based constrained optimal control scheme for active distribution networks. *Appl. Energy* **2020**, *264*, 114772. [CrossRef]
19. Sun, X.; Xu, Z.; Qiu, J.; Liu, H.; Wu, H.; Tao, Y. Optimal volt/var control for unbalanced distribution networks with human-in-the-loop deep reinforcement learning. *IEEE Trans. Smart Grid* **2024**, *15*, 2639–2651. [CrossRef]
20. Liu, S.; Feng, B.; Guo, C.X.; Ji, W.X.; Wang, W.; Zhang, Y. Deep-reinforcement-learning strategy for reactive-power and voltage control in distribution networks considering self-attention and temporal memory. *Proc. CSEE* **2025**, *45*, 565–577. (In Chinese)

21. Liao, W.; Bak-Jensen, B.; Radhakrishna Pillai, J.; Wang, Y.; Wang, Y. A review of graph neural networks and their applications in power systems. *J. Mod. Power Syst. Clean Energy* **2022**, *10*, 345–360. [CrossRef]
22. Chen, K.; Hu, J.; Zhang, Y.; Yu, Z.; He, J. Fault location in power distribution systems via deep graph convolutional networks. *IEEE J. Sel. Areas Commun.* **2020**, *38*, 119–131. [CrossRef]
23. Huang, N.T.; Cheng, D.; Cai, G.W. Fault location in high-penetration active distribution networks based on an improved spatial-temporal graph neural network. *Autom. Electr. Power Syst.* **2025**, *49*, 112–122. (In Chinese)
24. Ngo, Q.H.; Nguyen, B.L.H.; Vu, T.V.; Zhang, J.; Ngo, T. Physics-informed graphical neural network for power-system state estimation. *Appl. Energy* **2024**, *358*, 122602. [CrossRef]
25. Moshtagh, S.; Azimian, B.; Golgol, M.; Pal, A. Topology-aware graph-neural-network-based state estimation for PMU-unobservable power systems. *IEEE Trans. Power Syst.* **2025**, *40*, 4547–4560. [CrossRef]
26. Han, Z.A.; Chen, C.; Cao, Y.J.; Wang, L.L.; Li, Y.; Sun, C.H. Distribution-network topology detection and state estimation based on a physics-consistent dynamic spatial-temporal graph neural network. *Proc. CSEE* **2025**. (In Chinese)
27. Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph attention networks. *arXiv* **2017**, arXiv:1710.10903.
28. Liu, X.F.; Yin, S.; Zhang, Y.H.; Deng, F.Z. Reactive-power optimization for new power systems based on an interpretable graph attention network. *Power Syst. Technol.* **2026**, *50*, 1008–1018. (In Chinese)
29. Zhu, T.; Hai, D.; Li, W.Y.; Huang, W.; Zhou, S.C.; Wu, M.H.; Wang, Y.F. Voltage and reactive-power optimization of distribution networks based on graph neural networks and reinforcement learning. *South. Power Syst. Technol.* **2024**, *18*, 67–78. (In Chinese)
30. Zhu, Z.S.; Chen, W.L.; Zhang, X.B. Distribution-network voltage control based on graph deep reinforcement learning. *Electr. Power Autom. Equip.* **2026**, *46*, 137–145. (In Chinese)
31. Cao, D.; Zhao, J.; Hu, J.; Pei, Y.; Huang, Q.; Chen, Z.; Hu, W. Physics-informed graphical-representation-enabled deep reinforcement learning for robust distribution-system voltage control. *IEEE Trans. Smart Grid* **2024**, *15*, 233–246. [CrossRef]
32. Chen, Y.; Liu, Y.; Zhao, J.; Qiu, G.; Yin, H.; Li, Z. Physical-assisted multi-agent graph reinforcement learning enabled fast voltage regulation for PV-rich active distribution networks. *Appl. Energy* **2023**, *351*, 121743. [CrossRef]
33. Zhang, Y.; Yue, M.; Wang, J.; Yoo, S. Multi-agent graph-attention deep reinforcement learning for post-contingency grid emergency voltage control. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 3340–3350. [CrossRef]
34. Li, J.; Zhang, R.; Wang, H.; Liu, Z.; Lai, H.; Zhang, Y. Deep reinforcement learning for voltage control and renewable accommodation using spatial-temporal graph information. *IEEE Trans. Sustain. Energy* **2024**, *15*, 249–262. [CrossRef]
35. Jiang, C.X.; Lin, J.C.; Guo, C.; Yuan, Y.J.; Lin, J.J. Review of graph reinforcement learning for distribution-network optimization. *Power Syst. Technol.* **2026**, *50*, 922–947. (In Chinese)
36. Liu, M.; Gao, H.; Ji, S. Towards deeper graph neural networks. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Virtual Event, 23–27 August 2020; pp. 338–348.
37. Chen, D.; Lin, Y.; Li, W.; Li, P.; Zhou, J.; Sun, X. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 3438–3445.
38. Rusch, T.K.; Bronstein, M.M.; Mishra, S. A survey on oversmoothing in graph neural networks. *arXiv* **2023**, arXiv:2303.10991.
39. Wu, X.; Ajorlou, A.; Wu, Z.; Jadbabaie, A. Demystifying oversmoothing in attention-based graph neural networks. In Proceedings of the 37th Conference on Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023; pp. 35084–35106.
40. Su, G.; Wang, M.; Zhang, Y.; Wang, M. Simple and deep graph attention networks. *Knowl.-Based Syst.* **2024**, *293*, 111649. [CrossRef]
41. Yang, L.; Zhang, R.Y.; Li, L.; Xie, X. SimAM: A simple, parameter-free attention module for convolutional neural networks. In *Proceedings of the 38th International Conference on Machine Learning, Virtual Event, 18–24 July 2021*; PMLR: New York, NY, USA, 2021; pp. 11863–11874.
42. Haarnoja, T.; Zhou, A.; Abbeel, P.; Levine, S. Soft actor-critic: Off-policy maximum-entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018*; PMLR: New York, NY, USA, 2018; Volume 80, pp. 1861–1870.
43. *Standard's GB/T 12325-2008*; Power Quality—Deviation of Supply Voltage. Standards Press of China: Beijing, China, 2008.
44. Lee, X.Y.; Sarkar, S.; Wang, Y. A graph policy network approach for volt-var control in power distribution systems. *Appl. Energy* **2022**, *323*, 119530. [CrossRef]

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