Sheffield Hallam University

Attention-Enabled Hierarchical Multi-Agent Deep Reinforcement Learning for Coordinated Voltage Control in Active Distribution Systems

RIAZ, Hafiz Mehboob, SAJJAD, Malik Intisar Ali and AKMAL, Muhammad <http://orcid.org/0000-0002-3498-4146>

Available from Sheffield Hallam University Research Archive (SHURA) at:

https://shura.shu.ac.uk/37675/

This document is the Published Version [VoR]

Citation:

RIAZ, Hafiz Mehboob, SAJJAD, Malik Intisar Ali and AKMAL, Muhammad (2026). Attention-Enabled Hierarchical Multi-Agent Deep Reinforcement Learning for Coordinated Voltage Control in Active Distribution Systems. IET Smart Grid, 9 (1): e70107. [Article]

Copyright and re-use policy

See http://shura.shu.ac.uk/information.html

Sheffield Hallam University Research Archive
http://shura.shu.ac.uk

Check for updates

IET Smart Grid

WILEY

The Institution of Engineering and Technology

ORIGINAL RESEARCH OPEN ACCESS

Attention-Enabled Hierarchical Multi-Agent Deep Reinforcement Learning for Coordinated Voltage Control in Active Distribution Systems

Hafiz Mehboob Riaz[1] | Malik Intisar Ali Sajjad[1] | Muhammad Akmal[2]

[1]Department of Electrical Engineering, University of Engineering and Technology, Taxila, Pakistan | [2]School of Engineering and Built Environment, Sheffield Hallam University, Sheffield, UK

Correspondence: Muhammad Akmal (m.akmal@shu.ac.uk)

Received: 3 April 2026 | **Revised:** 10 June 2026 | **Accepted:** 3 July 2026

Keywords: active distribution systems | attention mechanism | deep reinforcement learning | hierarchical control | hybrid action space | multi-agent systems | volt/VAR control

ABSTRACT

The increasing integration of distributed renewable energy resources (DRES) is transforming active distribution systems (ADS), introducing challenges in maintaining voltage levels and minimising power losses. Effective volt/VAR control (VVC) requires coordinated operation of conventional devices, such as switched capacitors, alongside modern inverter-based resources. Although model-based optimisation methods can achieve this coordination, their practical use is limited by high computational demands and dependence on accurate grid models. Model-free deep reinforcement learning (DRL) offers a promising alternative but often suffers from instability and limited interpretability, particularly in multiagent settings based on DDPG architectures. This paper proposes an attention-enabled hierarchical multiagent twin delayed deep deterministic policy gradient (AE-HMATD3) framework for coordinated VVC. The approach employs a two-layer structure to manage fast- and slow-responding devices at different temporal scales, improving control alignment and system reliability. Learning stability is enhanced through twin critics and delayed policy updates, whereas a Gumbel–SoftMax method enables unified handling of discrete and continuous control actions. An attention mechanism further improves coordination and scalability by capturing spatial dependencies. Validation on IEEE 33- and 118-bus systems demonstrates that the proposed approach reduces power loss by 12%–44% and achieves 84%–99% fewer voltage violations compared to hierarchical multiagent DRL baselines.

1 | Introduction

Concerns over declining fossil fuel reserves and the need to reduce carbon emissions have accelerated the evolution of traditional power distribution networks. Distributed energy resources (DERs), and renewable units in particular, are being added at a pace that has effectively transformed conventional feeders into active distribution systems (ADSs) [1]. This shift brings clear environmental benefits, it also introduces serious operational difficulties. For example, the bidirectional power flows arising from distributed generation (DG) result in significant active power losses and voltage violations. Voltage rise at DG connection points is a recurring concern, especially when peak renewable generation does not coincide with peak demand [2]. VVC addresses these issues by coordinating reactive power resources to keep voltages within statutory bounds whilst reducing losses.

VVC methods fall broadly into model-based optimisation and data-driven approaches [3]. Classical formulations, such as linear, quadratic and mixed-integer programming, together with metaheuristics, such as PSO and GA, have demonstrated effectiveness in solving VVC problems. However, they share two practical weaknesses: a heavy computational burden under real-

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *IET Smart Grid* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology.

time operation [4] and a reliance on accurate up-to-date network models that are difficult to maintain in practice [5].

Data-driven approaches circumvent the model dependency by learning control policies directly from historical data. Among these, deep reinforcement learning (DRL) has become the dominant approach for VVC, since it can adapt to changing operating conditions without an explicit network model and scales to the high-dimensional state spaces of realistic distribution networks. DRL has been applied to voltage regulation and loss minimisation across a range of settings from discrete reactive power compensation through capacitor switching [6, 7] to continuous reactive power dispatch from inverter-based resources [8, 9].

The devices these methods control differ in two ways that shape the learning problem. The first is the form of their control action. Switched shunt capacitors (SCs) and on-load tap changers (OLTCs) act through discrete on/off or tap switching, whereas inverter-based photovoltaics (IBPVs) and static VAR compensators (SVCs) deliver continuously adjustable reactive power. The second is response speed. SCs and OLTCs are electromechanical and slow, suited to gradual voltage variations and operators deliberately cap their switching frequency to limit mechanical wear [10]. IBPVs and SVCs respond on millisecond-to-second timescales through power electronics, making them effective against the fast fluctuations introduced by renewable generation. Following common usage, we refer to the slow discrete devices as slow responding devices (SRDs) and the fast continuous ones as fast responding devices (FRDs). Because the two classes differ by orders of magnitude in response time, controlling them on a single time step creates conflict: mechanical actions issued too frequently shorten device life, whereas fast devices are left compensating for slow ones operating outside their natural range. Therefore, effective VVC calls for a coordinated control approach that aligns each device with its own response timescale; without this alignment, conflicting control actions degrade overall network performance [11].

These device characteristics map directly onto the choice of DRL algorithm. Discrete switching is naturally handled by value-based methods, such as deep Q-networks (DQN) [6, 7], whereas continuous reactive power dispatch calls for policy-gradient methods, with the deep deterministic policy gradient (DDPG) algorithm being the most widely adopted [8, 9]. Since real networks contain both device classes, several works combine the two, pairing continuous control of IBPVs and SVCs with discrete switching of SCs to handle the resulting hybrid action space within one framework [11–13].

These two challenges, coordinating heterogeneous action spaces and respecting the differing response times of each device, underpin three connected categories of model-free VVC research. The first category, hierarchical control architectures, assigns devices to separate control layers according to their response speed. Early designs were model-based, using decentralised network partitioning [14], distributed gradient coordination [15] and a three-layer architecture, combining local edge intelligence with centralised optimisation [16]. A recent three-timescale formulation [17] schedules OLTCs and SCs hourly, optimises inverter setpoints every 15 min and applies droop control every minute, lending strong support to the hour/quarter-hour partitioning used here. Learning-based hierarchical schemes have since emerged, including hierarchical RL with meta-learning-based adaptation under limited communication [18] and MARL coordinated with a higher-level aggregator that activates demand flexibility only when reactive headroom is exhausted [19]. These methods exploit timescale separation effectively, but the learning-based ones so far control continuous reactive power almost exclusively, leaving discrete switching outside the learnt policy.

The second category, multiagent DRL coordination, operates under the centralised-training–decentralised-execution (CTDE) paradigm, which lowers online communication cost and improves scalability. Several methods in this category address the hybrid action space of VVC, coordinating discrete switching devices, such as OLTCs and SCs, on slower timescales alongside continuous reactive power devices, such as IBPVs and SVCs, on faster ones. However, how that hybrid space is handled varies considerably. A common strategy is to treat the two action types separately, assigning a value-based learner, such as DQN, to the slow discrete devices and a policy-gradient learner, such as DDPG, to the fast continuous ones [12, 13]. Although effective, this arrangement is hybrid only in the sense that two distinct algorithms run in parallel; the discrete and continuous decisions are optimised by separate networks rather than within a single policy. A smaller body of work pursues a genuinely unified treatment, using the Gumbel–SoftMax relaxation to make discrete switching differentiable so that both action types can be learnt end to end [11, 20]; for example, adopted this relaxation within a multiagent framework. Within both strategies, the two device classes are dispatched on separate timescales: the multiagent DDPG variants, for instance, coordinate SRDs with FRDs across separate timescales [20, 21], and a recent multiagent soft actor–critic schemes regulate voltage through PV inverters on partitioned networks [22, 23]. What ties these approaches together is their dependence on a DDPG-based backbone, which carries the well-documented Q-value overestimation that destabilises training and degrades policy quality [24]. The twin-delayed DDPG (TD3) algorithm [25] corrects this through clipped double-Q learning and delayed policy updates. Building on TD3, our earlier work [26] introduced a multiagent TD3 formulation for continuous voltage regulation devices; beyond this, TD3 in VVC has remained confined to continuous-only devices [27, 28] or to fully centralised designs covering all device types [29].

The third category targets the inter-agent coordination mechanism. Many multiagent VVC schemes assume a fixed information-sharing topology, which fits poorly with networks whose electrical coupling shifts as operating conditions change. Attention-based coordination, introduced for general multiagent RL [30], addresses this by letting each agent weight its neighbours according to current relevance. Recent VVC studies have begun to adopt the idea: attention has been combined with multiagent actor–critic learning using sensitivity-based subnetwork clustering [31], and embedded as self-attention within a multiagent soft actor–critic algorithm to improve collaboration and learning efficiency [32]. However, in both cases, attention is

25152947, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/stg2.70107 by Sheffield Hallam University, Wiley Online Library on [24/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

applied to homogeneous continuous-only agents, each controlling the reactive power of the same device class, within a single-timescale nonhierarchical setting. Its use to coordinate agents that span both discrete and continuous devices across separate timescales has not been examined.

Across these categories, progress is uneven where they overlap. Hierarchical learning-based schemes regulate continuous devices well but rarely fold discrete switching into the learnt policy [17–19]. Where hybrid action is handled, it is often handled in a decoupled fashion, a value-based learner for discrete devices running alongside a policy-gradient learner for continuous ones [12, 13], rather than within a single differentiable policy; the methods that do unify the two through Gumbel–SoftMax remain tied to a DDPG backbone [11, 20]. Where attention has recently been introduced to coordinate agents, it has been limited to continuous-only single-timescale control [31, 32]. This work targets the intersection of these gaps: a hierarchical multiagent framework that handles discrete and continuous devices within a single differentiable policy, suppresses *Q*-value overestimation through twin-critic TD3 learning, and extends attention-based coordination, for the first time, to agents operating across hybrid action spaces and separate timescales.

To clearly illustrate the research gaps addressed by this work, Table 1 presents a comprehensive comparison of existing DRL-based VVC methods, highlighting the research gap that motivates this work across four key dimensions: architecture type, hybrid action capability, *Q*-value overestimation handling and attention-based interagent coordination.

Table 1 makes the central observation of this work explicit: each existing category resolves one facet of the VVC problem whereas leaving another open. Hierarchical schemes align control with device timescales but keep discrete switching outside the learnt policy; hybrid-action methods unify discrete and continuous control but within centralised single-agent designs that scale poorly; multiagent methods scale through decentralised execution but inherit the value overestimation of their DDPG backbone and the attention-based methods that coordinate agents adaptively have only been demonstrated for homogeneous continuous-only devices. The closest prior work is our earlier multiagent TD3 formulation [26], which already coordinated continuous reactive power devices under decentralised execution with overestimation-corrected learning. The framework proposed here extends that foundation along three axes that no existing method combines: it incorporates discrete switching within a unified differentiable hybrid-action policy, organises control hierarchically so that slow and fast devices act on their natural timescales and replaces fixed interagent communication with attention-based coordination across heterogeneous device types. Therefore, the advance is not the addition of any single feature, but the resolution of a combination of failure modes that no prior method addresses together.

To bridge the identified gaps, this paper proposes a two-layer attention-enabled hierarchical multiagent twin delayed deep deterministic policy gradient (AE-HMATD3) approach with hybrid action for coordinated VVC in ADS. The proposed approach optimises the dispatch of both FRDs and SRDs on their respective time scales. The main contributions are as follows:

TABLE 1 | Categories of DRL-based VVC and their principal limitations.

Category/refs	Architecture	Action space	Overestimation handling	Attention-based coordination	Principal limitation
Hierarchical learning-based [18, 19]	Hierarchical	Continuous	✗	✗	Discrete switching kept outside the learnt policy
Decoupled (DQN + DDPG) [12, 13]	Multiagent	Hybrid (decoupled)	✗	✗	Separate networks per action type; not a unified policy
Unification of continuous and discrete actions [11, 20]	Multiagent	Hybrid (unified)	✗	✗	Unified but DDPG-based; overestimation bias
Entropy-based stabilisation using SAC [22, 23]	Multiagent	Continuous	✓	✗	Continuous-only; no discrete devices
TD3-based VVC [27, 28]	Centralised	Continuous/ Hybrid	✓	✗	Not multiagent; single point of failure
Multiagent TD3 [26]	Distributed	Continuous	✓	✗	Continuous-only; no hierarchy or attention
Attention MADRL [31, 32]	Distributed	Continuous	✗	✓	Attention over homogeneous continuous agents only
Proposed	**Hierarchical + Multiagent**	**Hybrid (unified)**	**✓ (TD3)**	✓	—

Note: The bold text in the last row shows the contributions of the proposed technique.

25152947, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/stg2.70107 by Sheffield Hallam University, Wiley Online Library on [24/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

i. We develop a temporally aligned hierarchical architecture a two-layer control structure that maps voltage regulation devices to their physical response times. The upper centralised layer (Layer 1) supervises SRDs on an hourly basis, whereas the distributed lower layer (Layer 2) regulates FRDs at 15-min intervals. Crucially, the discrete switching of SRDs is brought inside the learnt end-to-end differentiable policy rather than handled by separate rules, which is precisely what existing hierarchical learning-based schemes leave out. This separation enables coordinated operation whilst accounting for equipment operational constraints and system-wide operational security.
ii. We design a stabilised hybrid-action learning framework that alleviates the Q-value overestimation and training instability inherent in DDPG-based hybrid-action VVC. The proposed AE-HMATD3 employs twin critics with delayed policy updates to improve learning stability and policy convergence, whereas the hybrid action space is realised through Gumbel–SoftMax parameterisation, enabling unified end-to-end differentiable optimisation of discrete switching actions and continuous reactive power setpoints.
iii. We introduce attention-enabled cooperative multiagent control, where the distributed lower layer (Layer 2) uses an attention mechanism to support coordinated decision-making among geographically distributed agents. By focusing more strongly on information from electrically important neighbouring areas whilst still using local measurements, the proposed method improves interagent coordination and offers better scalability than conventional MARL approaches.
iv. We validate the proposed method on the IEEE 33-bus and IEEE 118-bus distribution systems under a wide range of load and generation conditions. Comparisons with state-of-the-art hierarchical baselines with hybrid action space, including HMA–DDPG and HMASAC, show better voltage regulation and lower active power losses.

The remainder of this paper is organised as follows: Section 2 presents the mathematical modelling of voltage regulation devices and the problem formulation of VVC. Section 3 details the proposed AE-HMATD3 framework, including the two-timescale hierarchical control architecture, the attention-based coordination mechanism and the hybrid actor–critic learning scheme. Case studies and comprehensive result analysis are provided in Section 4, followed by conclusions in Section 5.

2 | Problem Formulation and Modelling

Mathematical modelling of voltage control devices and the associated optimisation problem for hybrid-action based VVC in ADS are presented in this section.

2.1 | Mathematical Representation of Control Devices

Modern distribution networks employ a heterogeneous mix of voltage regulation devices operating on different timescales and with different action spaces associated with continuous and discrete devices. To accommodate the different operational characteristics of fast-responding and slow-responding devices, a hierarchical temporal framework is employed. Accordingly, the control horizon is considered over a 24-h period and partitioned into hourly intervals indexed by τ, where $\tau \in \{1, 2, ..., 24\}$ as illustrated in Figure 1. At the start of each hourly interval, SRDs, in particular SCs are coordinated to compensate for voltage deviations associated with slowly varying load conditions. To account for faster network dynamics within each hour, the hourly interval τ is further divided into four shorter sub-intervals, each spanning15 min, indexed by $t \in \{1, 2, 3, 4\}$. At the onset of each sub-interval t, the FRDs, including IBPVs and SVCs, dynamically adjust their reactive power injection to counteract swift voltage variations induced by intermittent renewable generation fluctuations. Detailed mathematical models of each device are given in the following sub-section.

2.1.1 | *Inverter-Based Photovoltaic*

Smart PV inverters can modulate reactive power injection/absorption to regulate local voltage whilst maintaining active power generation. The reactive power capability of an IBPV at bus j is parameterised as follows:

$$Q_j^{\mathrm{IBPV}}(t) = \alpha_j^{\mathrm{IBPV}}(t) \cdot \overline{Q_j^{\mathrm{IBPV}}}(t), \alpha_j^{\mathrm{IBPV}}(t) \in [-1, 1], \quad (1)$$

where $\alpha_j^{\mathrm{IBPV}}(t)$ is the continuous reactive power control coefficient (negative for absorption and positive for injection) and $\overline{Q_j^{\mathrm{IBPV}}}$ is the maximum available reactive power constrained by the inverter's apparent power rating:

$$Q_j^{\mathrm{IBPV}}(t) \le \sqrt{\left(S_j^{\mathrm{IBPV}}(t)\right)^2 - \left(P_j^{\mathrm{IBPV}}(t)\right)^2}, \quad (2)$$

where $S_j^{\mathrm{IBPV}}(t)$ is the inverter's rated apparent power (typically 110% of the PV panel rating to provide reactive capability) and $P_j^{\mathrm{IBPV}}(t)$ is the instantaneous active power generation. The

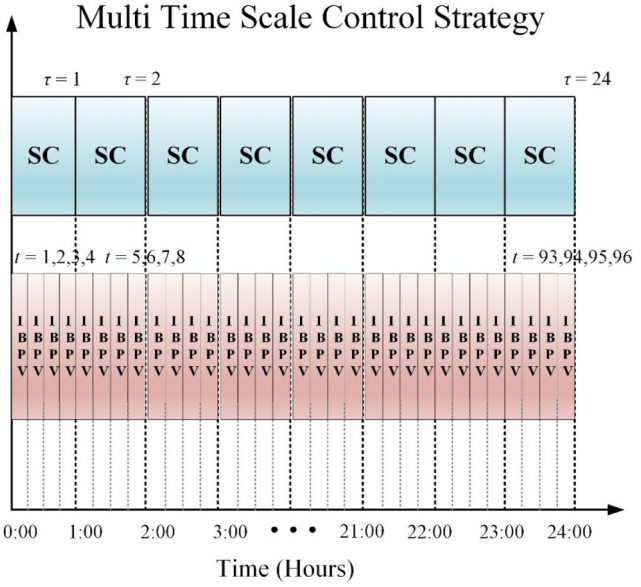

FIGURE 1 | The multitime scale control of voltage regulation devices in ADS.

25152947, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/stg2.70107 by Sheffield Hallam University, Wiley Online Library on [24/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

formulation in Equation (2) creates a time-varying action space that shrinks during peak solar generation, $P_j^{\mathrm{IBPV}}(t)$ and expands during low generation periods. IBPVs can adjust reactive power within 100–200 ms, making them suitable for fast voltage regulation responding to solar intermittency. The proposed RL framework learns to implicitly respect these bounds through reward shaping (Section 3.1).

2.1.2 | *Static VAR Compensator (SVC)*

Static VAR compensators deliver variable reactive support through thyristor-switched capacitor banks (TSCs) combined with thyristor-controlled reactor elements (TCRs). The compensator's reactive output is confined in the following range:

$$\underline{Q_j^{\mathrm{SVC}}}(\tau,t) \le Q_j^{\mathrm{SVC}}(\tau,t) \le \overline{Q_j^{\mathrm{SVC}}}(\tau,t), \tag{3}$$

where $\underline{Q_j^{\mathrm{SVC}}}$ and $\overline{Q_j^{\mathrm{SVC}}}$ represent the inductive and capacitive limits, respectively. Unlike IBPVs whose capacity varies with active power output, SVCs maintain fixed reactive power limits. We parameterise SVC control similarly to Equation (1):

$$Q_j^{\mathrm{SVC}}(t) = \alpha_j^{\mathrm{SVC}}(t) \cdot \; Q_j^{\mathrm{SVC,rated}}(t), \alpha_j^{\mathrm{SVC}}(t) \in [-1,1]. \tag{4}$$

SVCs respond within milliseconds, making them the fastest voltage control devices. However, they are expensive to install and typically deploy on critical buses only.

2.1.3 | *Switched Shunt Capacitors (SC)*

Switched shunt capacitors offer discrete reactive support via binary switching operations. The reactive power delivered by SC is defined in Equation (5),

$$Q_j^{\mathrm{SC}}(\tau) = u_j^{\mathrm{SC}}(\tau) \cdot Q_j^{\mathrm{SC,rated}}, \tag{5}$$

$$u_j^{\mathrm{SC}}(\tau) \in \{0,1\}, \forall \mathrm{j} \in \mathcal{C},$$

where $\mathcal{C}$ is the set of buses equipped with capacitor banks and $Q_j^{\mathrm{SC,rated}}$ is the rated reactive power capacity. Switching frequency constraints apply to prevent mechanical relay degradation:

$$\sum_{\tau=\tau_0}^{\tau_0+T_{\mathrm{window}}} \left| u_j^{\mathrm{SC}}(\tau) - u_j^{\mathrm{SC}}(\tau-1) \right| \le M_{\max}. \tag{6}$$

The daily switching operations of SRDs are constrained by $M_{\max}$. Capacitors operate on slow timescale (60-min intervals).

2.2 | Optimisation Framework for Voltage Control

A radial distribution network is represented as a graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ where $\mathcal{N} = \{1, 2, ..., Nb\}$ is the set of buses and $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ is the set of branches. The VVC problem aims to minimise a weighted sum of voltage deviations and active power losses by optimally dispatching all available voltage regulating devices:

Objective

$$\min_{\substack{\{Q_{\mathrm{cap}}(\tau)\} \\ \{Q_{\mathrm{IBPV}}(\tau,t)\} \\ Q_{\mathrm{SVC}}(\tau,t)}} \quad J = \lambda_v \cdot J_v + \lambda_p \cdot J_p \tag{7}$$

The term J_v penalises voltage deviations from nominal voltage (1.0 p.u.):

$$J_v = \sum_{\tau=1}^{N_\tau} \sum_{t=1}^{N_t} \sum_{i=1}^{N_b} \left[\; (v_i(\tau,t) - 1)^2 \right]. \tag{8}$$

The *active power loss term* J_p sums Ohmic losses across all branches:

$$J_p = \sum_{i=1}^{N_b} r_i \frac{P_i^2(\tau,t) + Q_i^2(\tau,t)}{v_i^2(\tau,t)} [\mathrm{MW}]. \tag{9}$$

Subject to:

i. Power Balance Constraints:

$$\sum_{i \in m(j)} \left[P_{ij}(\tau,t) - i_{ij}^2(\tau,t) r_{ij} \right] - P_j(\tau,t) = \sum_{k \in n(j)} P_{jk}(\tau,t), \tag{10a}$$

$$\sum_{i \in m(j)} \left[Q_{ij}(\tau,t) - i_{ij}^2(\tau,t) x_{ij} \right] - Q_j(\tau,t) = \sum_{k \in n(j)} Q_{jk}(\tau,t). \tag{10b}$$

ii. Voltage Drop Constraint:

$$v_j^2(\tau,t) = v_i^2(\tau,t) + \left(r_{ij}^2 + x_{ij}^2 \right) i_{ij}^2(\tau,t) - 2\big(r_{ij} P_{ij}(\tau,t) + x_{ij} Q_{ij}(\tau,t) \big). \tag{11}$$

iii. Current–Power Relationship:

$$P_{ij}^2(\tau,t) + Q_{ij}^2(\tau,t) = \left(\; i_{ij}(\tau,t) v_i(\tau,t) \right)^2. \tag{12}$$

iv. Net Power Injection:

$$P_j(\tau,t) = P_j^{\mathrm{l}}(\tau,t) - P_j^{\mathrm{IBPV}}(\tau,t), \tag{13a}$$

$$Q_j(\tau,t) = Q_j^{\mathrm{l}}(\tau,t) - Q_j^{\mathrm{IBPV}}(\tau,t) - Q_j^{\mathrm{SVC}}(\tau,t) - Q_j^{\mathrm{SC}}(\tau,t). \tag{13b}$$

v. Voltage Limits:

$$\underline{v}_i(\tau,t) \le v_i(\tau,t) \le \overline{v}_i(\tau,t), \tag{14}$$

where Equations (10a) and (10b) ensure active and reactive power balance at each node with $m(j)$ and $n(j)$ as the parent and child bus set of node j; P_{ij}, Q_{ij} are the active and reactive power flows from bus i to j; r_{ij}, x_{ij} are the resistance and reactance of the line segment (i,j); P_j^{l}, Q_j^{l} are load's active and reactive power, whereas voltage operational bounds $\underline{v}_i(\tau,t) = 0.95$ p.u. and $\overline{v}_i(\tau,t) = 1.05$ are mandated by ANSI C84.1 standard. The optimisation problem Equations (7–14) is nonconvex due to the quadratic equality constraint (12) and the product term in Equation (9). Moreover, the presence of binary variables $u_j^{\mathrm{SC}} \in \{0, 1\}$ renders the problem mixed-

25152947, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/stg2.70107 by Sheffield Hallam University, Wiley Online Library on [24/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

integer nonconvex, belonging to the NP-hard complexity class. This motivates the model-free reinforcement learning approach proposed in Section 3, which learns near-optimal policies through interaction without requiring explicit system models.

3 | Attention-Enabled Hierarchical Multi-Agent TD3 Framework (AE-HMATD3)

A conventional single-agent formulation does not reflect the physical structure of the network, so it is not directly suited to voltage control in ADS. The regulating devices are distributed across electrically distinct sub-networks, no single controller can measure the full bus-voltage profile in real time and the devices respond over different timescales. For these reasons, VVC is formulated here as a cooperative Markov game, with one agent assigned to each sub-network. The global state collects the bus voltages and power flows of Equations (10–12), whereas each agent observes only the local voltages and reactive power injections within its own area. All agents are guided by a common reward that penalises both voltage limit violations and active power losses, so that the learning signal carries the same two objectives minimised in Equation (7).

The regulating devices are of two kinds, which prevents the action space from being treated as purely continuous. The reactive injections of the IBPVs and SVCs in Equations (1–3) are continuous, whereas the switching states of the SCs in Equations (4–6) are discrete. The two are handled together through a Gumbel–SoftMax relaxation, which keeps the policy differentiable across both branches during training. The electrical coupling between sub-networks shifts with the operating point and is usually held fixed in conventional MARL; to capture it, an attention layer lets each agent weight its neighbours according to their current influence. A further difficulty is the Q-value overestimation of DDPG-based controllers, which tends to destabilise hybrid-action VVC (Table 1); this is addressed using the twin critics and delayed policy updates of TD3.

3.1 | Cooperative Markov Game Formulation

3.1.1 | Two-Layer Device-Centric Architecture Overview

We formulate the voltage control problem as a two-layer device-centric architecture designed to align with the distinct operational characteristics of grid assets as illustrated in Figure 2. The first layer manages SRDs using a centralised system-level policy,

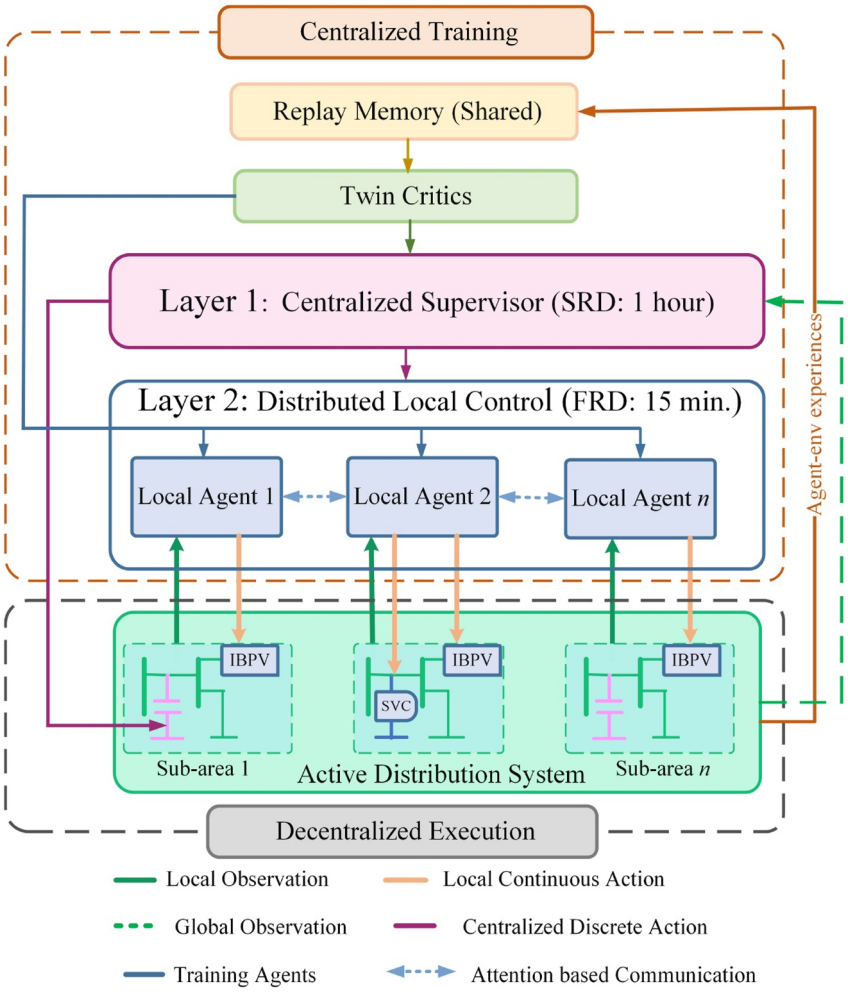

FIGURE 2 | Proposed hierarchical architecture for coordinated voltage control in ADS.

25152947, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/stg2.70107 by Sheffield Hallam University, Wiley Online Library on [24/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

operating them on an hourly basis. This layer monitors system-wide voltage violations and applies mechanical switching constraints to protect equipment life. The second layer, by contrast, comprises a cooperative multiagent framework augmented with an interagent attention mechanism. This layer enables coordinated voltage regulation that adaptively responds to the stochastic nature of renewable generation. Within this decentralised framework, the continuous reactive power outputs of FRDs, such as IBPVs and SVCs, are managed based on local measurements within each agent's designated sub-network. The set points of FRDs are updated every 15 min to balance two competing goals: maintaining voltage within allowable limits and reducing active power losses.

The two-layer structure allows discrete switching actions to complement the high-frequency continuous adjustments made by the agents. We essentially align the control strategy with the physical speed limit. This way, the framework not only keeps the voltage within acceptable limits but also protects legacy devices from excessive mechanical stress caused by over-operation.

3.1.2 | *Markov Game Definition*

This device-centric decomposition is mathematically framed as a cooperative Markov game, defined by the following tuple $\mathcal{MG}$:

$$\mathcal{MG} = \langle \mathcal{N}, \mathcal{S}, \{\mathcal{O}_b\}_{b=1}^{n}, \{\mathcal{A}_b^c\}_{b=1}^{n}, \mathcal{A}^d, \mathcal{R}, \mathcal{P}, \gamma \rangle, \tag{15}$$

where the set $\mathcal{N} = \{1, 2, ..., n\}$ comprise the number of agents assigned to different sub-networks of ADS, $\mathcal{S}$ represents the global state space that encodes all bus voltages and power flows, $\mathcal{O}_b = \{\mathcal{O}_1, \mathcal{O}_2, ..., \mathcal{O}_n\}$, $\mathcal{O}_b \subseteq \mathcal{S}$ is the set of local observation space of each agent collected from the corresponding network, $\boldsymbol{\mathcal{A}_c^b} \subseteq [-1, 1]^{m_b}$ designates continuous action set of m_b number of fast devices and $\boldsymbol{\mathcal{A}^d} \subseteq \{0, 1\}^k$ represents discrete action space for k number of slow devices. $\mathcal{R}$ is the shared reward function all agents, $\mathcal{P}$ is the transition probability of state and γ is the discount factor.

3.1.3 | *Local Observation Space*

To ensure scalability and minimise the communication burden, each agent b constructs a localised observation vector $\boldsymbol{o_b(t)} \in R^{d_b}$ from measurements within its respective controlled network. The local observation reduces in the dimensionality curse associated with global state space. The local observation vector is defined in Equation (16):

$$\boldsymbol{o_b(t)} = \left[v_b(t), Q_b^{\text{own}}(t), \alpha_b^{\text{prev}}(t), v^{\min}(t), v^{\max}(t)\right]^T, \tag{16}$$

where $v_b(t)$ characterises voltage magnitudes at all buses in sub-network b, $Q_b^{\text{own}}(t)$ represents total reactive power from voltage regulation devices controlled by agent b, $\alpha_b^{\text{prev}}(t)$ captures temporal information through the previous control action, $v^{\min}(t)$ and $v^{\max}(t)$ provide global voltage's upper and lower limits for system-wide awareness. All voltage observations are normalised to zero mean and unit variance to improve neural network training stability.

3.1.4 | *Two-Layer Action Space*

The control framework develops a hybrid action space $\boldsymbol{\mathcal{A}_b^c} \in [-1, 1]^{m_b}$ to accommodate the different timescales of the grid assets. For the distributed layer (Layer 2), the continuous actions for FRDs managed by agent b are defined in Equation (17a) as follows:

$$\boldsymbol{\mathcal{A}_b^c} = \left[\alpha_1^{\text{IBPV}}(t), ..., \alpha_{m_b}^{\text{IBPV}}(t), \alpha_1^{\text{SVC}}(t), ..., \alpha_{m_b}^{\text{SVC}}(t)\right]^T. \tag{17a}$$

These actions parameterise the exchange of reactive power from FRDs and are updated every 15 min to respond rapidly to fluctuations caused by solar intermittency. These distributed decisions are then aggregated into a global FRD action vector for system interaction in Equation (17b),

$$\boldsymbol{\mathcal{A}^c}(t) = \left[\boldsymbol{a}_1^c(t)^T + \boldsymbol{a}_2^c(t)^T ... \boldsymbol{a}_n^c(t)^T\right]^T \in R^{m_{\text{tot}}}, \tag{17b}$$

where $m_{\text{tot}} = \sum_{b=1}^{n} m_b$ denotes the total number of FRDs across all controlled sub-networks. This aggregation step combines individual agent decisions into a global continuous action vector used for environmental interaction, where each sub-vector is determined through the attention-enhanced coordination mechanism detailed in Section 3.2, ensuring that localised FRD's set-points are aware of neighbouring bus sensitivities.

Simultaneously, the slow-responding discrete actions are managed by the centralised layer in Equation (17c):

$$\boldsymbol{\mathcal{A}^d}(t) = \left[u_1^{\text{SC}}(t), u_2^{\text{SC}}(t), ..., u_{\text{k}}^{\text{SC}}(t)\right]^T \in \{0, 1\}^k, \tag{17c}$$

where $u_1^{\text{SC}}(t) \in 0, 1$ represents the binary switching state of capacitor bank (0 = OFF and 1 = ON) and k is the total number of slow switching devices in ADS. This centralised supervision of SRDs is justified by the need to enforce global mechanical switching constraints, manage system-wide transients caused by capacitor switching and preserve communication efficiency by avoiding complex distributed consensus protocols.

3.1.5 | *Multi-Timescale SRD Control Logic*

Although many recent VVC frameworks advocate for fully distributed schemes, a centralised supervisory approach is adopted for the SRD's layer for two critical reasons. First, physical coupling dictates that capacitor switching induces transients, propagating across multiple network zones. The communication requirements can also be reduced by leveraging the aforementioned centralised supervision. Rather than switching the discrete actions of SRDs based on local observations, a centralised supervisor directs system-level voltage fluctuations to control the switching of SRDs. In this way, the 4× to 8× larger communication burden of iterative distributed consensus methods can be eliminated as the centralised controller only needs to exchange scalar timer states and binary voltage-violation indicators.

To align the control actions with the inherent physical response times of the assets, we implement the multitimescale logic defined in Equation (18) as follows:

25152947, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/stg2.70107 by Sheffield Hallam University, Wiley Online Library on [24/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

$$\mathcal{A}^d(t) = \begin{cases} \pi_d(s(t)) & \text{if } V_{\text{vio}}(t) \text{ and } \tau_{\text{sw}} \geq \Delta T \\ \mathcal{A}^d(t-1) & \text{otherwise} \end{cases}. \quad (18)$$

Within this multitimescale control, the centralised SRD control policy $\pi_d(s(t))$ is executed using Gumbel–SoftMax reparameterisation in such a way that entire training procedure remains differentiable. The Boolean variable $V_{\text{vio}}(t)$ monitors the system-wide voltage violations by indicating whether any bus violates the voltage limits, whereas the switching timer τ_{sw} ensures temporal consistency by incrementing at each time step. Coordination between the two layers is defined by the timescale ratio $\Delta T = \Delta\tau_{\text{slow}}/\Delta t_{\text{fast}} = 4$, which corresponds to the 60-min switching interval for SRDs relative to the 15-min dispatch interval for FRDs. The selection of this interval is examined in Section 4.6.

This design enables event-driven switching, so SRDs only act when a voltage violation is present, and time-gated switching, which prevents chattering by enforcing a minimum four-step delay between actions.

The complete two-layer heterogeneous action space is summarised in Equation (19)

$$\boldsymbol{\mathcal{A}}(t) = \left[\boldsymbol{\mathcal{A}}^c(t); \boldsymbol{\mathcal{A}}^d(t)\right] \in R^{m_{\text{tot}}} \times \{0,1\}^k. \quad (19)$$

This formulation bridges the gap between distributed, high-frequency control of modern FRDs and the centralised low-frequency supervision of legacy SRDs, providing a unified command signal for the ADS.

3.1.6 | *Cooperative Reward Function*

The objective of all agents is to maximise a shared reward signal designed to minimise both voltage violations and active power losses:

$$r(t) = -\lambda_V \cdot r_V(t) - \lambda_P \cdot r_P(t). \quad (20)$$

The negative signs associated with two weighted components of reward function translate the minimisation problem into maximisation, which is a standard RL convention.

The voltage penalty term $r_V(t)$ comprises a quadratic barrier function in Equation (21) to penalise deviations beyond the standard limits of $\overline{v} = 1.05$, and $\underline{v} = 0.95$.

$$r_V(t) = \sum_{i=1}^{N_b} \left[\max(v_i(t) - \overline{v}, 0)^2 + \max(\underline{v} - v_i(t), 0)^2\right]. \quad (21)$$

The active power loss term comprises the sum of Ohmic losses across all network branches whilst accounting for voltage-dependent variations defined in Equation (22) as follows:

$$r_P(t) = \sum_{(i,j)\in\mathcal{E}} r_{ij} \cdot \frac{P_{ij}^2(t) + Q_{ij}^2(t)}{v_i^2(t)}. \quad (22)$$

In order to achieve the balance between the two competing objectives, we apply weighting coefficients of $\lambda_V = 60$ and $\lambda_P = 7.5$. The 8:1 ratio gives priority to the voltage security during early training, helping agents first learn stable regulation before focusing more carefully on efficiency

3.1.7 | *Joint Policy Optimisation Objective*

The main objective of the cooperative framework is to learn a set of policies that maximise the expected cumulative discounted return as expressed in Equation (23):

$$J(\theta_1, ..., \theta_n, \theta_c) = E_{\tau\sim\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]. \quad (23)$$

In this formulation, $\theta_1, ..., \theta_n$ represent the parameters of the actor networks for all agents, where each θ_b maps local to continuous FRD actions in its own sub-network. The parameter θ_c represents the centralised network that performs hourly SRD coordination. The trajectory τ consists of the sequence of states, actions, and rewards generated under the joint policy, whereas the discount factor γ balances immediate voltage stabilisation against long-term power loss reduction.

To achieve this objective, the framework follows the centralised training and decentralised execution (CTDE) paradigm. During training, all policies are updated jointly using twin centralised critics that rely on global state information to address the credit assignment problem. However, during execution, FRD agents operate independently using only local observations enhanced by the attention mechanism, whereas the SRD policy provides higher-level supervisory guidance based on system-wide voltage conditions. This design improves sample efficiency during training whilst preserving the scalability needed for real-time operation, since agents do not need continuous access to global system information once deployed.

3.2 | Attention-Based Inter-Agent Coordination

Each agent sees only the voltages and reactive injections of its own sub-network, with no direct view of how neighbouring areas are pushing its bus voltages. Therefore, a multihead attention mechanism is used so that each agent can weigh the observations of other agents according to how strongly they are electrically coupled to its own area. This process begins by projecting local and neighbouring observations into a shared latent space via learnable query, key and value matrices as defined in Equations (24a–24c):

$$\boldsymbol{Q_b} = \boldsymbol{W_Q o_b}, \quad (24a)$$

$$\boldsymbol{K_j} = \boldsymbol{W_K o_j}, \forall j \neq b, \quad (24b)$$

$$\boldsymbol{V_j} = \boldsymbol{W_V o_j} \quad \forall j \neq b, \quad (24c)$$

where $\boldsymbol{W_Q} \in R^{d_{\text{model}} \times d_{\text{obs}}}$ is the learnable query projection matrix and $d_{\text{model}} = H \times d_k = 256$ with $H = 4$ attention heads and $d_k = 64$ dimensions per head. $\boldsymbol{W_K}$ and $\boldsymbol{W_V} \in R^{d_{\text{model}} \times d_{\text{obs}}}$ are projection weights from neighbouring agents' observations.

The coupling between neighbour j and agent b is quantified by the scaled dot-product attention coefficient $\alpha_{bj}^{(h)}$ in Equation (25). To ensure training stability and prevent gradient saturation, the dot products are scaled by d_k and a learnable

25152947, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/stg2.70107 by Sheffield Hallam University, Wiley Online Library on [24/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

temperature parameter $\tau^{(h)}$. The superscript (h) denotes the hth attention head.

$$\alpha_{bj}^{(h)} = \frac{\exp\left(\frac{\boldsymbol{Q}_b^{(h)} \cdot \boldsymbol{K}_j^{(h)}}{\sqrt{d_k} \cdot \tau^{(h)}}\right)}{\sum_{j' \neq b} \exp\left(\frac{\boldsymbol{Q}_b^{(h)} \cdot \boldsymbol{K}_{j'}^{(h)}}{\sqrt{d_k} \cdot \tau^{(h)}}\right)}. \tag{25}$$

The learnable temperature parameter $\tau^{(h)}$ is updated via gradient descent defined in Equation (26):

$$\tau^{(h)} \leftarrow \tau^{(h)} + \eta_\tau \cdot \nabla_{\tau^{(h)}} \mathcal{L}, \tag{26}$$

where η_τ represents the temperature-specific learning rate, constrained within the range [0.5, 2.0]. Low temperature values promote sharp attention, focusing on neighbours with the highest electrical coupling, whereas higher temperatures encourage a more distributed awareness across the local sub-network.

Each of the four attention heads independently aggregates information from neighbouring agents using unique learnt parameters. The output of the hth head is computed as the weighted sum of the value projections in Equation (27a):

$$\text{head}_h = \sum_{j \neq b} \alpha_{bj}^{(h)} \boldsymbol{V}_j^{(h)}. \tag{27a}$$

Each of these independent heads are used to synthesise the diverse spatial dependencies, the resulting vectors are unified through the multihead concatenation defined in Equation (27b):

$$\boldsymbol{c_b} = \text{Concat}(\text{head}_1, \ldots, \text{head}_H)\boldsymbol{W_O}, \tag{27b}$$

where $\boldsymbol{W_O} \in R^{d_{\text{model}} \times d_{\text{obs}}}$ is a linear projection matrix that maps the concatenated context back to the observation space. Finally, the original local observation $\boldsymbol{o_b}(t)$ is added to the attention-derived context $\boldsymbol{c_b}(t)$ through a residual connection and layer normalisation as shown in Equation (27c):

$$\widetilde{\boldsymbol{o_b}}(t) = \text{LayerNorm}(\boldsymbol{o_b}(t) + \boldsymbol{c_b}(t)). \tag{27c}$$

The resulting augmented observation vector $\widetilde{\boldsymbol{o}}_b(t)$ serves as the comprehensive input for the actor–critic networks, effectively bridging the gap between decentralised perception and global system awareness.

3.3 | Hybrid Actor–Critic Architecture

3.3.1 | *Actor Network With Gumbel–SoftMax Parameterisation*

Reactive power dispatch by IBPVs and SVCs is a continuous control task that calls for a deterministic policy. TD3 is adopted as the basis of the learning algorithm because it provides this whilst curbing the value overestimation that would otherwise drive overly aggressive reactive power commands. It is modified here for the hybrid action space and the two-layer structure. Each decentralised agent b is equipped with a dedicated actor network π_{θ_b}, defined in Equation (28), which maps the attention-augmented observation $\widetilde{\boldsymbol{o}}_b$ to the continuous reactive power set-points $\boldsymbol{\mathcal{A}_b^c}$ of the IBPVs and SVCs in sub-network b.

$$\pi_{\theta_b} : \widetilde{\boldsymbol{o_b}} \rightarrow \boldsymbol{\mathcal{A}_b^c}. \tag{28}$$

The actor architecture comprises a multilayer perceptron with three hidden layers, utilising an activation function with a final tan$h$ layer to confine the reactive power set-points within the $[-1, 1]$ range.

The capacitor banks switch only in discrete on/off steps, which would normally block gradient-based training of the supervisory policy. To keep the framework end-to-end differentiable, these switching decisions are relaxed through the Gumbel–SoftMax reparameterisation shown in Equation (29):

$$\widetilde{a}_1^{\text{SRD}}[k] = \frac{\exp\left(\left(\log \pi_k + g_k\right)/\tau'\right)}{\sum_{k'=1}^{K} \exp\left(\left(\log \pi_{k'} + g_{k'}\right)/\tau'\right)}. \tag{29}$$

In this formulation, π_k denotes the probability of selecting discrete state k and g_k is a noise sample drawn from a Gumbel distribution within range (0, 1). The temperature parameter τ' is strategically annealed from 5.0 to 0.5 in the training phase to enable a smooth transition from stochastic exploration to deterministic categorical selection.

3.3.2 | *Twin-Critic Architecture With Delayed Updates*

During the centralised training phase, twin critic functions Q_{ψ_1} and Q_{ψ_2} are employed to evaluate combined reactive dispatch and capacitor switching $(\boldsymbol{\mathcal{A}^c}, \boldsymbol{\mathcal{A}^d})$ against the system-wide voltage and power-flow state as expressed in Equation (30):

$$Q_{\psi_n}(s, a) = Q_{\psi_n}\left(s, \boldsymbol{\mathcal{A}^c}, \boldsymbol{\mathcal{A}^d}\right), n \in \{1, 2\}. \tag{30}$$

The critics are optimised by using the minimisation of the mean squared error (MSE) loss defined in Equation (31):

$$\mathcal{L}(\psi_n) = E_{(s,a,r,s') \sim \mathbb{D}}\left[\left(Q_{\psi_n}(s, a) - y\right)^2\right], n \in \{1, 2\}. \tag{31}$$

A prioritised replay memory $\mathbb{D}$ with capacity depending upon the system size (30k samples for the 33-bus system and 100k for the 118-bus system) is utilised to sample the agent-environment experiences. An overestimated action value tends to make the agents over-commit reactive power and disturb the very voltages they are meant to regulate. To curb this, the TD target y is computed with the clipped double-Q technique in Equation (32):

$$y = r + \gamma \min_{n \in \{1,2\}} Q'_{\psi_n}\left(s', \widetilde{\boldsymbol{a}}'\right). \tag{32}$$

Finally, we implement target policy smoothing to prevent the policy from exploiting inaccuracies in the Q-function as shown in Equation (33):

$$\widetilde{\boldsymbol{a}_b'} = \pi'_{\theta_b}\left(\widetilde{\boldsymbol{o}_b'}\right) + \text{clip}(\epsilon, -c, c), \epsilon \sim \mathcal{N}(0, \sigma^2). \tag{33}$$

The addition of clipped Gaussian noise ($\sigma = 0.2$, $c = 0.5$) to the target actions regularises the value surface, ensuring that the learnt control policies are robust to the slight variations in voltage measurements typically encountered in ADS.

25152947, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/stg2.70107 by Sheffield Hallam University, Wiley Online Library on [24/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Finally, the framework utilises delayed policy updates to further stabilise convergence. By updating the actor parameters θ and all target networks only every $d = 2$ critic iterations, we decouple high-frequency value estimation from policy improvement. The target networks are updated via Polyak averaging defined in Equations (34a–34d):

$$\theta_b' \leftarrow \rho \cdot \theta_b + (1-\rho) \cdot \theta_b', \forall b \in \{1, ..., n\}, \tag{34a}$$

$$\theta_c' \leftarrow \rho \cdot \theta_c + (1-\rho) \cdot \theta_c', \tag{34b}$$

$$\psi_1' \leftarrow \rho \cdot \psi_1 + (1-\rho) \cdot \psi_1', \tag{34c}$$

$$\psi_2' \leftarrow \rho \cdot \psi_2 + (1-\rho) \cdot \psi_2'. \tag{34d}$$

The soft update coefficient is chosen as $\rho = 0.005$ to ensure slow and stable target network tracking. This temporal delay ensures that the centralised critics have sufficiently converged towards an accurate Q-value surface before computing the policy gradients, effectively eliminating the divergent oscillations during the training process. The overall two-timescale training procedure, incorporating the hierarchical control framework shown in Figure 1, the attention-based coordination mechanism described in Equations (24–27) and the hybrid actor–critic update rules in Equations (28–34), is presented in Table 2

TABLE 2 | Two-timescale AE-HMATD3 algorithm for coordinated voltage control.

Initialisation:

Initialise n FRD actors π_{Θ_b} (with attention module) and SRD supervisor π_{Θ_c} with Gumbel–Softmax
Initialise twin critic (Q_{ψ_1}, Q_{ψ_2}), all target networks and replay memory $\mathbb{D}$
Initialise Gumbel temperature $\tau' \leftarrow 5$ and reset counter $d \leftarrow 0$

Training loop:

For episode 1 to M **do**
  Reset environment, observe initial state
  For $\tau = 1$ to N_τ **do** // *Slow Time-Scale SRD (1-h)*
    Observe global state and voltage violation flag $V_{\text{vio}}(t)$ via Equation (18)
    If V_{vio} = True and $\tau_{\text{sw}} \geq \Delta T$ then
      Generate discrete action via Equation (29); Reset timer: $\tau_{\text{sw}} \leftarrow 0$
    Else
      $\mathcal{A}^d(\tau) = \mathcal{A}^d(\tau - 1)$ // Hold previous action
    End if
    For $t = 1$ to N_t **do** // *Fast Time-Scale FRD (15-min)*
      For each FRD agent $b = 1$ to n **do**
        Extract attention-fused observation via Equation (16)
        Compute projections, attention weights Equations (24–27)
        Get continuous action using Equation (28)
      End for
      Execute joint hybrid action $\boldsymbol{\mathcal{A}}(t)$ via Equations (17b, 17c) and (19)
      Observe ($\mathcal{A}^d(\tau)$, $\mathcal{A}^c(t)$, $r(t)$, $\boldsymbol{s}(t+1)$); store in $\mathbb{D}$
      If $|\mathbb{D}| \geq$ batch_size, then
        Sample mini-batch from $\mathbb{D}$; update Q_{ψ_1}, Q_{ψ_2} Equations (32) and (33)
        $d \leftarrow d + 1$ // Delayed policy update
        If d mod 2 = 0 then
          Update all policies $\{\{\pi_{\theta_b}, \pi_{\theta_c}\}$ via Equation (23)
          Soft update target actor and critic using Equation (34)
        End if
      End if
      $\tau_{\text{sw}} \leftarrow \tau_{\text{sw}} + 1$
    End for // FRD loop
    $\tau \leftarrow \max(0.5, 5.0 \cdot \exp(-\lambda\_\tau' \cdot t))$ // Anneal Gumbel temp.
  End for // SRD loop
End for // Episodes
Return trained policies $\{\pi_{\theta_1}, ..., \pi_{\theta_n}, \pi_{\theta_c}\}$

25152947, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/stg2.70107 by Sheffield Hallam University, Wiley Online Library on [24/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

4 | Experimental Setup and Performance Evaluation

The performance of the proposed AE-HMATD3 framework is evaluated using two IEEE benchmark distribution systems in order to demonstrate effectiveness and scalability. The IEEE 33-bus system is used for baseline validation, whereas the IEEE 118-bus system serves as a large-scale test case. In both case studies, the original network models are expanded by adding distributed energy resources and voltage regulation devices so that they better reflect the complexity of active distribution systems.

4.1 | IEEE 33-Bus System

We selected the IEEE 33 bus as our initial test network. This is a radial network operating at 12.66 kV with a total connected load of 3.715 MW and 2.3 MVAR. In order to perform hierarchical control, the network is divided into four sub-networks based on electrical proximity and coupling strength as illustrated in Figure 3. In the context of the control strategy, the network is equipped with three IBPVs at buses 17, 21 and 24, each having a capacity of 1.5 MW active power and 2.0 MVAR reactive power. In addition, a 1.5 MVAR SVC is connected at bus 32. Four cooperative agents in distributed layer control FRDs, whereas two 500 kVAR SC (SRDs) at buses 13 and 24 are managed by the centralised supervisory layer.

4.2 | IEEE 118-Bus System

The IEEE 118-bus distribution network shown in Figure 4 operates at 11 kV and features significantly higher loading and distributed generation levels, thereby serving as a stress test for the proposed hierarchical multiagent control architecture. The system is partitioned into three sub-networks. To support voltage regulation across this expanded network, eight IBPV units each rated at 1.5 MW active and 2.0 MVAR reactive power are cited at buses 33, 50, 53, 68, 74, 97, 107 and 111, whereas two SVCs, each providing up to 1.5 MVAR of dynamic reactive compensation, are placed at buses 44 and 104. Together, these IBPVs and SVCs constitute the FRD layer managed by the distributed cooperative agents through attention-based inter-agent coordination. Four SCs, each rated at 500 kVAR installed at buses 35, 75, 95 and 105 serve as SRDs, with their switching actions controlled by the centralised control layer.

4.3 | Operational Data

In order to capture realistic variations during a full day, the simulation environment uses 24-h load and generation profiles split into 96 intervals of 15 min each. These base profiles are obtained from historical data for a pilot-project data in Eastern China [33]. For a more realistic scenario, a uniform noise of $\pm$ 20% is also added. The model is trained for a period of 300 daily episodes, with a total of 28,800 time steps. Overall, the test cases cover wide range of operating conditions, with system's load fluctuating between 0.4 and 1.6 p.u. The instantaneous renewable penetration is considered within a range of 20%–80%.

4.4 | Computational Framework and Training Procedure

The AE-HMATD3 algorithm is developed using the PyTorch 1.13 library, with computational acceleration provided by an NVIDIA RTX 3080 GPU. Power flow analysis is performed through the Pandapower library [34] to ensure high-fidelity balanced three-phase results. Network topologies are adapted from standard Matpower case files, specifically modified to include the hierarchical control interfaces for both legacy and modern power electronics-based devices. To ensure that the reported results are statistically reliable and not caused by random neural network initialisation, each system is trained independently using three random seeds: 555, 777 and 999. Performance metrics are reported as mean values together with standard deviations where appropriate. During training, continuous FRD actions use epsilon-greedy exploration with an annealed noise variance σ that decreases from 0.1 to 0.05. At the same time, discrete SRD switching is handled with temperature-annealed Gumbel–SoftMax, where the temperature τ decreases from 5.0 to 0.5, allowing the learning process to move gradually from broad exploration to stable deterministic control.

4.5 | Hyperparameter Configuration

Table 3 provides brief overview of hyperparameter utilised in the implementation of AE-HMATD3. To account for the hierarchical structure, attention-enhanced coordination among agents and hybrid action space, some of the parameters are different than those reported in ref. [26]. To ensure the stabilised training, the core TD3 settings, such as delayed policy update period ($d = 2$) and soft update rate ($\rho = 0.005$), are

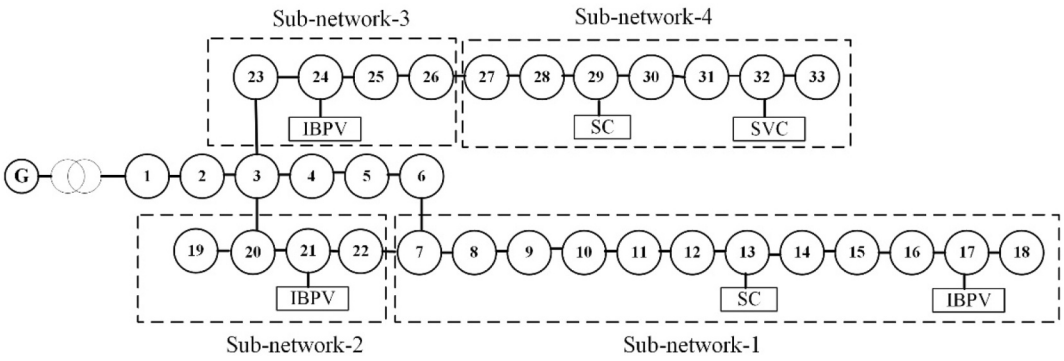

FIGURE 3 | IEEE 33-bus partitioned network.

25152947, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/stg2.70107 by Sheffield Hallam University, Wiley Online Library on [24/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

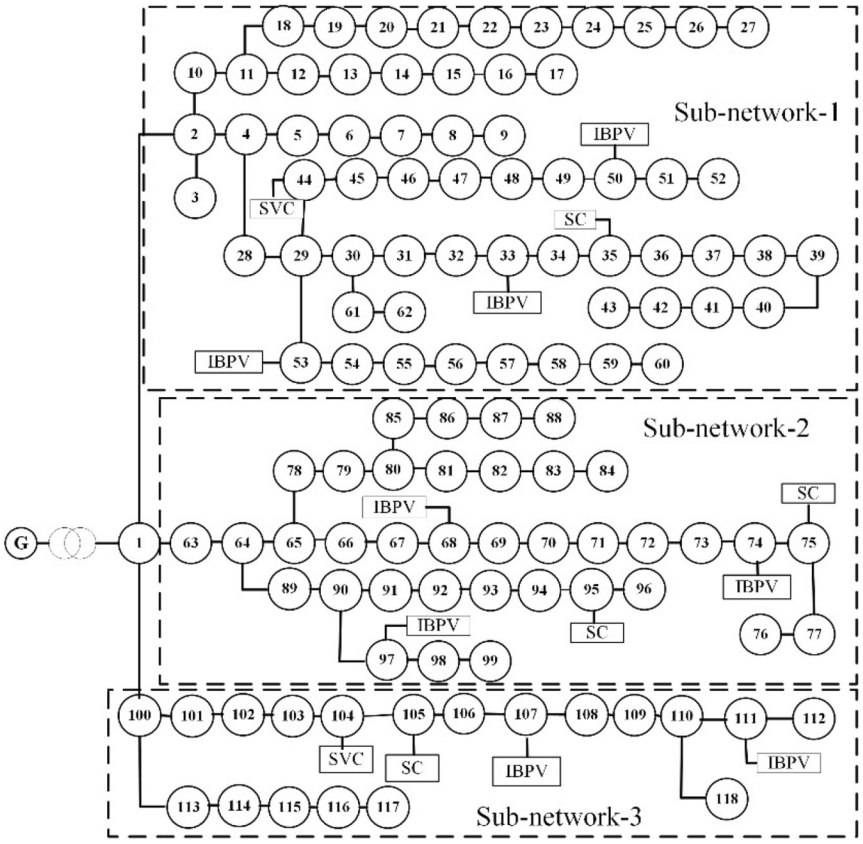

FIGURE 4 | IEEE 118-bus distribution network.

adopted exactly from ref. [25]. One notable shift from our prior work concerns the reward-shaping coefficients for voltage violation and power loss. Although prior work in refs. [11, 26] both used a 50:1, which proved too aggressive for the joint voltage–loss objective here, stalling power-loss minimisation. Rather than fix it by inspection, we determined the ratio through a sensitivity study (Section 4.7), sweeping $\lambda_V : \lambda_p$ from 1:1 to 50:1 on the 33-bus system. The 8:1 setting ($\lambda_V = 60$, $\lambda_p = 7.5$) gave the best balance, with the highest reward and lowest power loss whilst keeping voltage violations near their minimum, whereas larger ratios starved the loss term.

To handle the multiagent coordination of discrete and continuous devices, we implemented a scalable memory and attention architecture. The replay memory and mini-batch sizes were scaled proportionally with system complexity (from 33 to 118-bus) to maintain sample diversity and reduce gradient variance in higher-dimensional spaces. We adopted a learnable attention temperature (ranging from 0.5 to 2.0) to allow agents to transition between focused neighbour-specific monitoring during emergencies and broad spatial awareness during normal operation. Furthermore, the Gumbel–SoftMax temperature and exploration noise follow an exponential decay schedule, ensuring a smooth transition from stochastic exploration of discrete switching actions to deterministic exploitation as the policy converges.

4.6 | Sensitivity to the Time-Scale Partition

The controller of Figure 1 places the inverter-based resources and the SVC on a 15-min fast layer and the switched capacitors on a 1-h slow layer. The fast interval is fixed by the data: the field load and generation series are recorded at 15-min resolution, the finest available for this network, so the fast layer cannot act below this period. The slow interval is a free parameter that sets how often the discrete capacitor steps may change. Assigning fast continuous devices to a short interval and switched devices to a longer one is standard in two-time-scale Volt/VAR control [11, 20, 21], but the length of the slow interval varies between studies and is seldom supported by a sensitivity study. To confirm that the reported results do not depend on the particular 15-min/1-h pairing, the slow interval was swept over 15, 30, 45, 60 and 90 min with the fast layer held at 15 min and all other settings unchanged. Each case was trained for 100 episodes over two seeds; as Figures 5 and 6 show, the policies converge well before this horizon, so 100 episodes is sufficient.

Figure 7 shows the training curves for all five partitions, with shaded bands spanning one standard deviation across seeds. The reward and loss curves are almost indistinguishable: both settle to a common level within about 25 episodes and the bands overlap thereafter. The violation curves separate slightly, the 15-min setting sits a little higher in later episodes and the 60-min setting a little lower, but the gap stays within the seed spread throughout. The slow interval thus changes neither the convergence rate nor the converged level beyond the seed noise. Averaged over the final 20 episodes, the regulation indicators barely move across the range: mean violations stay between 13.98 and 14.78 per episode, active power loss between 17.25 and 17.61 and reward between −17.42 and −18.25, all spans smaller than the seed-to-seed scatter (the violation standard deviation

25152947, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/stg2.70107 by Sheffield Hallam University, Wiley Online Library on [24/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 3 | AH-MATD3 hyperparameter configuration.

Parameter	Value (33-bus)	Value (118 bus)	Justification
Network architecture			
Actor's hidden layer	[256, 256, 128]	[256, 256, 128]	Sufficient for local observation
Critic's hidden layer	[512, 512]	[512, 512]	Larger capacity for global state (99–400)
Attention heads (H)	4	4	Captures diverse coordination patterns
Head dimension d_k	64	64	Balances expressiveness versus computation
TD3 parameters			
Discount factor (γ)	0.95	0.95	For long-term control objectives
Soft update rate (ρ)	0.005	0.005	Standard TD3 value for stable updates [25]
Target policy noise	0.2	0.2	20% of action range to smooth targets [25]
Noise clip	0.5	0.5	Prevents excessive target deviation [25]
Policy update frequency	2	2	Delayed updates every 2 critic steps [25]
Training			
Replay memory	30,000	100,000	Scaled to ensure sample diversity as state-action space increases
Mini-batch size	128	512	Larger batches reduce gradient variance
Actor learning rate	0.0001	0.0001	To prevent unstable policy updates
Critic learning rate	0.0003	0.0003	Accelerate value learning [25]
Initial random steps	768	1536	Pretraining value: Scales with network size
Exploration noise	0.1 → 0.05	0.1 → 0.05	Exponentially decayed to favour exploitation
Gumbel–SoftMax			
Initial temperature	5	5	Promotes broad discrete action exploration
Final temperature	0.5	0.5	Reduced for near-discrete action selection
Reward shaping			
Voltage coefficient (λ_V)	60	60	Value confirmed by the reward-weight sensitivity analysis (Section 4.7)
Power coefficient (λ_P)	7.5	7.5	8:1 ratio (refined from 50:1 in [26]); identified as the best joint voltage–loss balance in the sensitivity analysis
Attention parameters			
Temperature range	[0.5, 2.0]	[0.5, 2.0]	Adapts heads b/w focused and distributed weighting
Temperature learning rate	0.001	0.001	Slow adaptation for stability
Dropout rate	0.1	0.1	Regularises attention layers and mitigate overfitting.

reaches 6.54 at 45 min). Therefore, we treat regulation quality as effectively independent of the partition, consistent with the fast layer carrying the continuous voltage-tracking duty at 15-min resolution regardless of how often the capacitors switch.

Table 4 describes converged performance and capacitor switching for each interval and includes cumulative reward, power loss per MW (Ploss/MW), volt–var ratio normalised by the square of the per-unit voltage (VVR/p.u.^{2}) and capacitor switching per day. The capacitor switching count behaves differently, falling steadily from about 26 operations per day at 15 min to 10 at 90 min, with 15 at the 60-min setting. A longer interval blocks frequent step changes and eases the duty on the capacitor switchgear, whose switching endurance is finite and is governed by dedicated device standards [35]. The benefit tapers off, however: most of the reduction is won by 45 min, and extending 60–90 min saves only about five operations a day whilst delaying the point at which the discrete devices can react to a sustained deviation. Therefore, the 60-min slow layer adopted here is a considered compromise, it secures most of the switching reduction whilst keeping the discrete layer responsive on a time scale matched to the 15-min data. Since the regulation indicators are statistically indistinguishable across the range, this choice incurs no measurable penalty in voltage quality or losses, confirming that the controller is robust to the time-scale partition and that the 15-min/1-h pairing is well justified.

25152947, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/stg2.70107 by Sheffield Hallam University, Wiley Online Library on [24/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

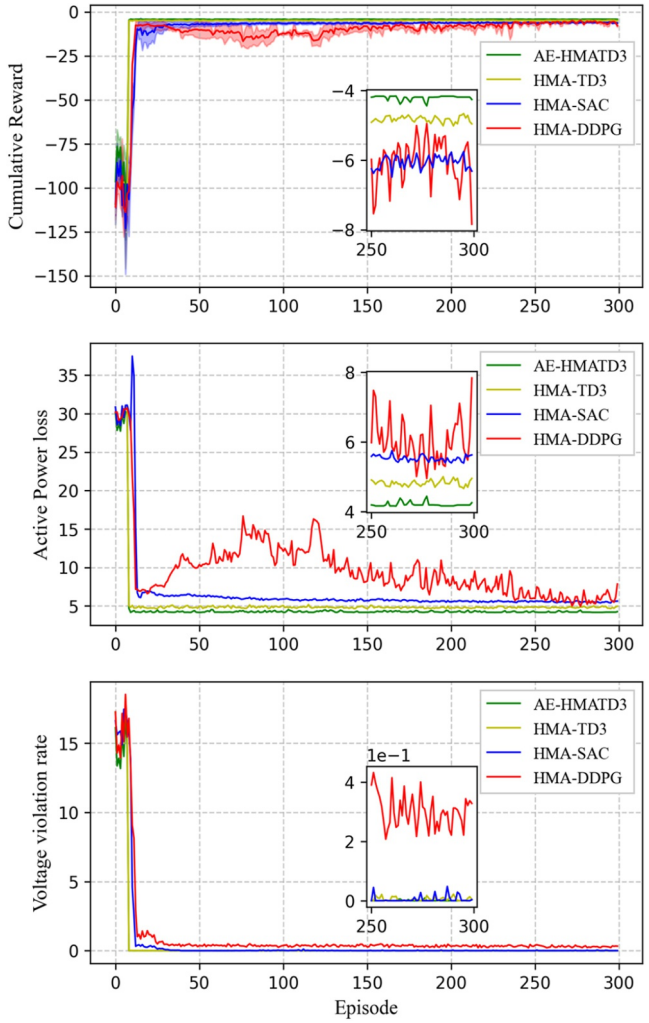

FIGURE 5 | Training convergence on IEEE 33-bus system. (Curves represent mean values across seeds; zoomed insets highlight convergence behaviour during final 50 episodes).

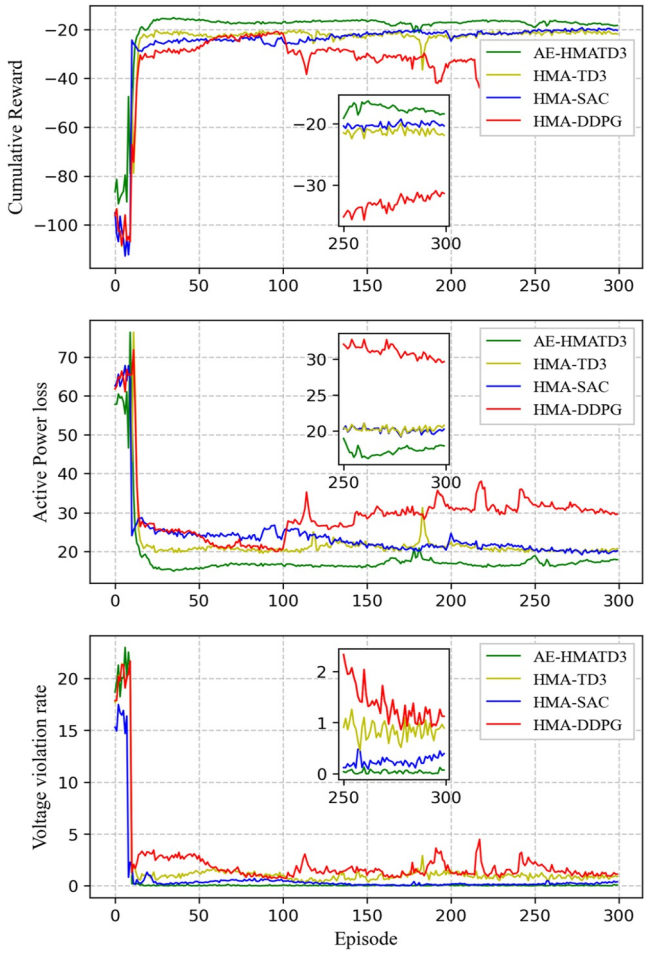

FIGURE 6 | Training convergence on IEEE 118-bus system. (Layout and conventions match Figure 5, demonstrating scalability to realistic distribution system sizes).

4.7 | Sensitivity Analysis of the Reward Weights

The reward in Equations (20–22) combines two objectives of differing nature, and the relative weighting of the voltage-violation and power-loss terms governs which objective the policy prioritises during learning. Since this weighting was set empirically, we examined its effect through a sensitivity study on the 33-bus system, sweeping the ratio λ_v: λ_p over a geometric grid of 1:1, 2:1, 4:1, 8:1, 16:1, 32:1 and 50:1. The voltage weight was held at $\lambda_v = 60$ and λ_p was varied to realise each ratio, so that the magnitude of the voltage term, and hence the scale of the critic targets, remained fixed and any change in behaviour could be attributed to the ratio alone. All configurations were trained under identical conditions, using the same two random seeds and the same 100-episode budget, so that the comparison across ratios is controlled and the differences reflect the weighting rather than the initialisation. The averaged learning curves for the episode reward, voltage-violation count and active power loss are shown in Figure 8, and the converged values (mean of the final 20 episodes) are summarised in Table 5.

Two regimes are evident. At low ratios (1:1 to 4:1) the loss term competes with the voltage objective, and the agent fails to suppress violations: the violation count stays above 16 and the reward remains poor, near −6.3 to −7.1. Increasing the ratio to 8:1 produces a sharp improvement across all three metrics, after which performance saturates and then slowly degrades as the loss term is progressively starved. This is consistent with the observation of ref. [36] that eliminating voltage violations is the more important and harder of the two objectives and therefore warrants a larger coefficient, but that an excessively large voltage weight introduces numerical-stability problems in the critic. The same trend appears here: beyond 16:1, the active power loss rises and the reward falls, and the 50:1 setting, adopted in the baseline conference work [26], is clearly suboptimal once the joint voltage–loss objective is considered, which directly motivated the re-tuning carried out in the present paper.

The 8:1 ratio gives the best overall balance. It achieves the highest episode reward (−4.35) and the lowest active power loss (4.22) of all settings, whereas keeping voltage violations low (5.18) with a tight spread across seeds ($\sigma = 0.33$). The 16:1 ratio records a lower mean violation count (3.40), but this figure is

25152947, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/stg2.70107 by Sheffield Hallam University, Wiley Online Library on [24/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

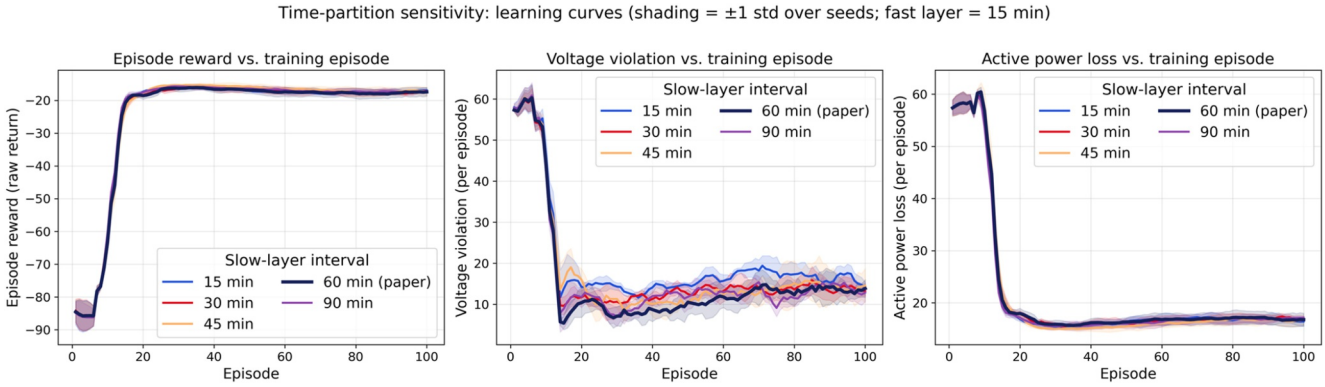

FIGURE 7 | Training curves for the time-scale partition sweep on the 118-bus system, showing episode reward, voltage-limit violations, and active power loss against training episode for slow-layer intervals of 15, 30, 45, 60 and 90 min, with the fast layer fixed at 15 min. Solid lines are the mean over two random seeds and shaded bands span one standard deviation; the 60-min interval adopted in this paper is drawn in bold.

TABLE 4 | Converged performance and capacitor switching for each slow-layer interval (fast layer fixed at 15 min).

Slow layer (min)	Cumulative reward	P_{loss}/MW	VVR/p.u^2.	Capacitor switching (per day)
15	-17.92 ± 0.98	17.30 ± 1.14	14.30 ± 2.83	26 ± 3
30	-18.25 ± 1.05	17.61 ± 0.89	14.78 ± 3.01	19 ± 4
45	-17.95 ± 1.84	17.28 ± 1.44	14.78 ± 6.54	16 ± 4
60	$\mathbf{-17.42 \pm 1.05}$	$\mathbf{17.25 \pm 1.43}$	$\mathbf{14.50 \pm 1.98}$	$\mathbf{15 \pm 1}$
90	-18.09 ± 0.91	17.56 ± 0.97	13.98 ± 0.46	10 ± 0.3

Note: Values are reported as mean ± standard deviation over two random seeds, averaged across the final 20 training episodes. The bold values indicate the best results of the proposed technique.

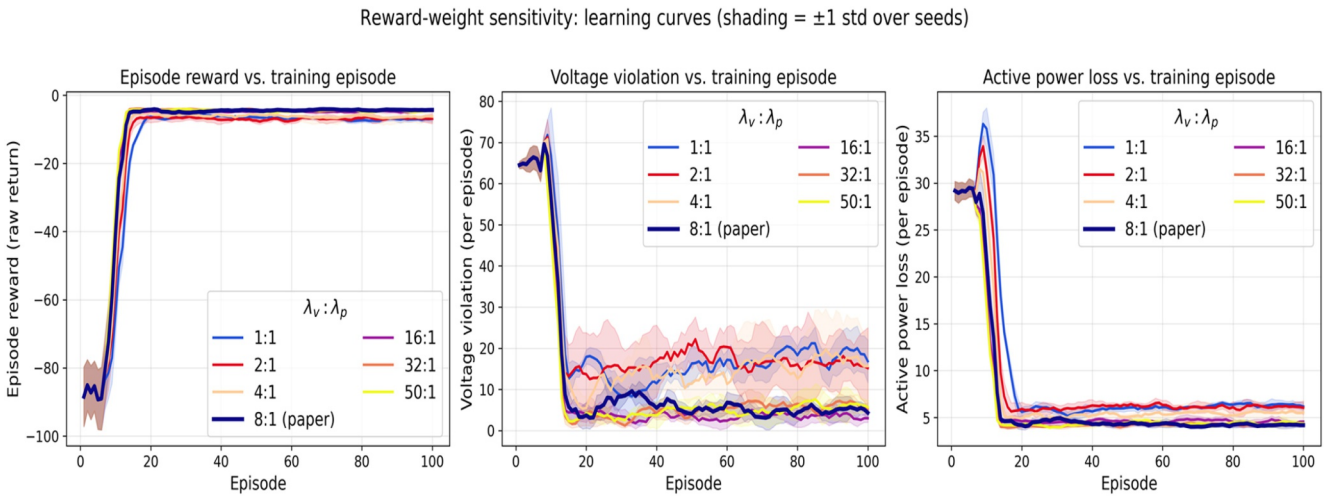

FIGURE 8 | Learning curves of the reward-weight sensitivity study on the 33-bus system, showing episode reward, voltage-violation count, and active power loss for reward-weight ratios ranging from 1:1 to 50:1. Curves are the mean over two seeds; shaded bands denote ± 1 standard deviation.

not robust: across the same two seeds its standard deviation is 1.50, more than four times that of the 8:1 case, so its apparent advantage falls well within run-to-run variability and cannot be treated as a genuine improvement. Moreover, the lower violation count at 16:1 is obtained at the cost of higher power loss (4.45 against 4.22) and lower reward (−4.60 against −4.35), both of which are stable across seeds. The reward and loss optima therefore coincide unambiguously at 8:1, and the metrics fall off on either side of it. On this basis we retain $\lambda_v = 60$ and $\lambda_p = 7.5$ (an 8:1 ratio) as the operating point, since it represents the best compromise between the two competing objectives rather than the optimum of any single one.

4.8 | Performance Analysis

To evaluate the specific contribution of the proposed attention mechanism, AE-HMATD3 is compared with three hierarchical

25152947, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/stg2.70107 by Sheffield Hallam University, Wiley Online Library on [24/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

multiagent baselines that share the same two-layer device architecture but differ in their coordination strategy. HMA–DDPG extends the original DDPG framework [37] to a multiagent hierarchical setting, whereas HMA–SAC builds upon the entropy-regularised SAC formulation [38]. Both methods are adapted to hybrid action spaces using the Gumbel–SoftMax relaxation, following the implementation strategy adopted from [11]. This ensures consistent handling of continuous FRD actions and discrete capacitor switching across all methods.

The primary ablation baseline is HMA-TD3, which employs the twin-critic structure of TD3 [25] with the same hierarchical and hybrid configuration as AE-HMATD3 but without the attention mechanism. In order to evaluate the contribution of attention mechanism, we compare AE-HMATD3 against HMATD3 with same architecture elements. To confirm a fair assessment, all the methods are evaluated using the same network structures, reward coefficients, training schedules and learning rates. Furthermore, tuning of each baseline method was independently performed using reasonable bounds so that differences in the results indicate algorithm design rather than configuration choices. The average values of all performance metrics over the final 50 episodes across three random seeds is reported in Table 6.

For 33 bus system, AE-HMATD3 has shown 23% reduction in active power loss (4.207 MW) compared to HMA-SAC (5.528 MW) and has reached up to 30% reduction compared to HMA-DDPG (6.033 MW). For voltage regulation, the proposed method has shown significantly less steady-state violation rate of 5.696×10^{-4} compared to HMA-DDPG (3.009×10^{-1}) and HMA-SAC (4.982×10^{-3}). Addition of attention mechanism to 33-bus system brings active power loss down from 4.82 MW (HMA-TD3) to 4.21 MW, a 12% improvement whereas voltage violation reduced by 87%. The 118-bus results follow the same trend, where AE-HMATD3 demonstrates 84%–97% reduction in voltage violations, and 14%–44% drop in power loss with highest cumulative reward compared to all baselines.

If we compare the HMA-TD3 against HMA–SAC, it indicates that entropy regularisation offers training stability comparable to twin critics when measured by cumulative reward alone. The difference is illustrated in loss reduction, SAC without a structured coordination layer cannot match the reactive power dispatch quality of the attention-enhanced method. This indicates that only stable training is not sufficient; efficient coordination of reactive power resources is vital for achieving both voltage improvement and loss minimisation.

4.9 | Training Dynamics and Convergence Analysis

The training progress for IEEE 33-bus and 118-bus systems is shown in Figures 5 and 6, respectively. It can be observed from Figure 5 that, all algorithms are able to reach feasible operating policies within about 50 episodes, which is mainly due to the similarity in the adopted hierarchical architecture. However, the rate of convergence differs in quality. Noticeable oscillations between episodes 50 and 200 are observed for the HMA–DDPG, where the rewards fluctuate between −5 and −8. This is a clear sign of Q-value overestimation, which results in aggressive policy updates, causing the control actions to exceed the optimal

TABLE 5 | Converged performance versus reward-weight ratio (mean ± std over two seeds and final 20 episodes).

Ratio (λ_v:λ_p)	λ_v	λ_p	Cumulative reward	VVR/p.u^2.	P_{loss}/MW
1:1	60	60	−7.09 ± 0.56	18.50 ± 2.20	6.34 ± 0.33
2:1	60	30	−6.80 ± 0.78	16.55 ± 7.80	6.13 ± 0.32
4:1	60	15	−6.34 ± 1.62	17.03 ± 8.93	5.51 ± 0.96
8:1	**60**	**7.5**	**−4.35 ± 0.22**	**5.18 ± 0.33**	**4.22 ± 0.22**
16:1	60	3.75	−4.60 ± 0.35	3.40 ± 1.50	4.45 ± 0.22
32:1	60	1.875	−4.40 ± 0.11	5.95 ± 0.10	4.23 ± 0.05
50:1	60	1.2	−4.63 ± 0.44	6.25 ± 3.05	4.40 ± 0.32

Note: The bold values indicate the best results of the proposed technique.

TABLE 6 | Performance metrics averaged over episodes 250–300 across three random seeds (mean ± std).

Test sys.	Algorithm	↓ P_{loss}/MW	↓ VVR/p.u^2.	↑ cumulative reward
33-bus	HMA–DDPG	6.033 ± 0.64	3.009×10^{-1} ± 0.055	−6.063 ± 0.18
	HMA–SAC	5.528 ± 0.08	4.982×10^{-3} ± 0.015	−6.039 ± 0.64
	HMA-TD3	4.821 ± 0.08	4.521×10^{-3} ± 0.093	−4.828 ± 0.09
	AE-HMATD3	**4.207 ± 0.06**	**5.696×10^{-4} ± 0.001**	**−4.207 ± 0.06**
118 bus	HMA–DDPG	31.045 ± 0.84	1.340 ± 0.362	−32.9151 ± 1.20
	HMA–SAC	20.126 ± 0.37	2.429×10^{-1} ± 0.092	−20.169 ± 0.38
	HMA-TD3	20.251 ± 0.36	8.494×10^{-1} ± 0.165	−21.242 ± 0.52
	AE-HMATD3	**17.232 ± 0.60**	**3.706×10^{-2} ± 0.032**	**−17.508 ± 0.64**

Note: The bold values indicate the best results of the proposed technique.

25152947, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/stg2.70107 by Sheffield Hallam University, Wiley Online Library on [24/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

reactive power setpoints. HMA-TD3 and AE-HMATD3, on the other hand, make use of a twin-critic mechanism which reduces this overestimation effect since a conservative estimate is used, that is, min{$Q1$, $Q2$}. This is the reason for the 25%–30% performance gap between HMA–DDPG and HMA-TD3 as reported in Table 6.

A comparison between HMA-TD3 and HMA–SAC gives additional insight into their stabilisation mechanisms. On the 33-bus system, HMA–SAC achieves a power loss of 5.528 ± 0.08 MW with voltage violations of $4.982 \times 10^{-3} \pm 0.015$ p.u.^{2}, which are comparable to HMA-TD3 (4.821 ± 0.08 MW and $4.521 \times 10^{-3} \pm 0.093$ p.u.^{2}). This suggests that entropy-regularised exploration in SAC offers a stabilising effect similar to twin-critic learning in TD3, even lacking conservative value estimation.

Despite these results, both methods fail to outperform AE-HMATD3. Table 6 clearly demonstrates that HMA–SAC results in an increased active power loss of 31.3% than AE-HMATD3 (5.53 vs. 4.21 MW), whereas HMA-TD3 displays a 14.5% increase (4.82 vs. 4.21 MW). A similar behaviour is observed for voltage violations: both SAC and TD3 converge to steady-state values on the order of 10^{-3} p.u.^{2}, whereas AE-HMATD3 shows reduced violations of about 5.7×10^{-4} p.u.^{2}, which is about an order-of-magnitude improvement. In a nutshell, these results suggest that although twin critics and entropy regularisation both help improve convergence trend, no approach alone can achieve the level of coordinated reactive power optimisation enabled by attention-based interaction.

4.10 | Scalability to Larger Network

The performance gap becomes even larger on the 118-bus system, as shown in Figure 6, where the reward improvement grows from 12.9% on the 33-bus network to 17.6% on the 118-bus network. AE-HMATD3 still has the best performance in voltage regulation. However, an interesting phenomenon is observed for the baseline methods. For the larger system with 118 buses, HMA-SAC surprisingly performs better than HMA-TD3, although HMA–SAC was the weaker method for the 33-bus system. For example, HMA-SAC has less power loss (20.13 MW compared to 20.25 MW for HMA-TD3), and its voltage violation is significantly less (0.243 p.u.^{2} vs. 0.849 p.u.^{2} for HMA-TD3) which is about 71% less. This is also reflected in the cumulative reward value for HMA–SAC, which is -20.17, compared to -21.24 for HMA-TD3.

This shift highlights fundamental scalability problem. For the smaller system with 33 buses and only four agents, although HMA-TD3 takes advantage of its twin critic to achieve stable learning with more conservative value estimates, SAC maintains stability through entropy regularisation and thus has comparable performance to HMA-TD3. However, for the larger system with 118 buses and 10 agents, the input size of critic abruptly increases. The critic's dimensions also expand from 140 to almost 400 dimensions for the large case, since each agent provides its own local information. This is a threefold expansion and makes learning more data-intensive, especially when keeping within the same 300 episodes. Under such a setup, TD3 is unable to perform well, and this is characterised by less accurate value function estimates, reduced coordination and increased voltage violations, rising from 5.44×10^{-3} p.u.^{2} for the 33-bus case to 0.849 p.u.^{2} for the 118-bus case. SAC is more adept at handling this expansion. Its entropy-driven reward function allows for further exploration even when the value function is no longer accurate. This is why it is able to perform better and avoid being trapped in poor deterministic policies, thus achieving lower voltage violations of 0.243 p.u.^{2}. Still, both methods are obviously outperformed by AE-HMATD3. In the case of the 118-bus network, voltage violations are limited to merely 0.037 p.u.^{2} and power loss is reduced only to 17.23 MW. In comparison to SAC, this translates into a reduction of 85% in voltage violations and 14% in power loss. This shows that even though the robustness of the system can be ensured by using entropy regularisation as the system scales up attention-based coordination is what truly enables consistent multiagent performance.

The key advantage of AE-HMATD3 over other methods lies in its approach to coordination. Unlike other methods, where all agents are treated equally, this approach learns to focus attention on only the agents that are most electrically relevant. This helps to restrict the growth of useful input information as the number of agents increases from 4 to 10. This way, the two critics are always reliable, even with limited training data, and the approach can maintain its near-optimal performance.

4.11 | Operational Control Performance

The convergence curves in Figures 5 and 6 show how the policies learn, but not how the trained controller behaves in operation. The converged AE-HMATD3 policy was therefore evaluated over the test horizon. The resulting operating conditions are summarised in Table 7, and the per-bus voltage distribution together with the per-device reactive usage is shown in Figure 9, in panels (a and b) for the 33-bus system and panels (c and d) for the 118-bus system. All operational metrics are reported as the mean over 100 test episodes for three random seeds (555, 777 and 999), consistent with the convergence study.

On the 33-bus feeder, the median voltage at every bus stay inside the 0.95–1.05 p.u. band, and no over-voltage occurs at any point in the horizon ($V_{\max} = 1.026$ p.u.). The lowest voltages appear at the radial extremity, buses 29 to 33, where the worst case reaches 0.945 p.u. during coincident peak load and low photovoltaic output. The deviation is small, at most 0.005 p.u.,

TABLE 7 | Operational control metrics for the proposed AE-HMATD3.

System	$V_{\min}$ (p.u.^{2})	$V_{\max}$ (p.u.^{2})	Max. voltage violation (p.u.^{2})	Mean reactive utilisation (%)	Reserve retained (%)
33-bus	0.945	1.026	0.005	10.5	89.5
118-bus	0.934	1.010	0.016	22.0	78.0

25152947, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/stg2.70107 by Sheffield Hallam University, Wiley Online Library on [24/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

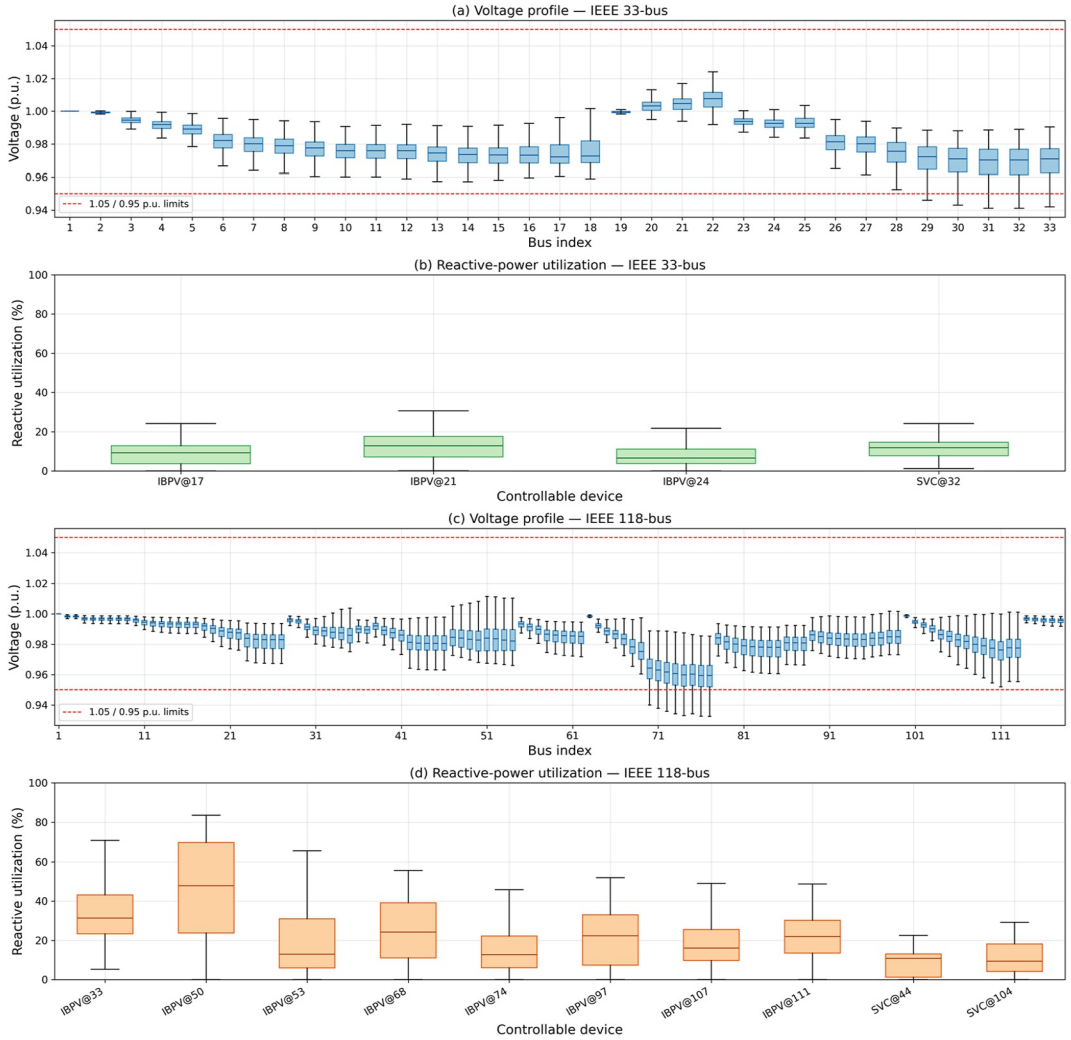

FIGURE 9 | Operational performance of the proposed AE-HMATD3 controller, showing the per-bus voltage profile and per-device reactive-power utilization for the IEEE 33-bus and IEEE 118-bus systems. Reactive-power utilization is expressed as a percentage of each device's rated capability.

and limited to short peak-load intervals. Figure 9b shows that this is achieved with modest reactive effort: median utilisation of the four controllable units stays near 10% of rated capability and never exceeds 30%, leaving close to 90% of the reactive reserve available (Table 7). The unit at bus 21 is the most active and the unit at bus 24 the least. Because the inverters and the SVC meet the reactive demand on their own, the switched capacitors are not operated during the evaluation, which removes mechanical switching and the wear it causes.

The 118-bus system follows the same pattern under heavier loading. Voltages stay within limits across most of the network, with the lowest values forming a cluster around buses 70–80 where the worst case settles at 0.934 p.u. and the largest deviation is 0.016 p.u. As expected for a larger and more heavily loaded network, the devices work harder: mean reactive utilisation rises to 22% and the retained reserve falls to about 78% (Table 7), with the unit at bus 50 carrying the most support at a median near 48% (Figure 9d). No over-voltage occurs ($V_{\max}$ = 1.010 p.u.). The controller therefore holds the voltage profile and keeps a substantial reactive margin on both systems, and the margin narrows with network size in a predictable way rather than through any loss of control quality.

4.12 | Attention Mechanism Contribution and Interpretability

The significance of role of attention mechanism becomes even clearer if we focus on the actual process of learning communication. In case of 33-bus system with four agents, there are six different pairs of agents that are capable of coordinating with each other C(4,2) = 6. However, the learnt attention weights show that the agents are not interacting with all the other agents equally. Instead, each agent is interacting with one or two agents and giving them relatively higher attention weights (> 0.30). In contrast, the agents that are farther away are not being given any attention, with attention weights < 0.05. As a result of this, the actual communication network is much less complex than the theoretical communication network. In this case, only about two of the six possible agent pairs remain actively engaged, which corresponds to roughly 67% sparsity in the communication. This benefit becomes even more apparent as the system increases in size. For example, in the 118 bus system, which has 10 agents, a fully connected system will have 45 possible pairs (C(10,2) = 45). This is a 7.5× increase in the number of communications, even though the number of agents only increases by 2.5×. In baseline approaches, which have a

25152947, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/stg2.70107 by Sheffield Hallam University, Wiley Online Library on [24/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

fixed communication structure, such as HMA-TD3 and HMA–SAC, each agent must process all inputs from the other nine agents. This results in a large increase in the critic's dimensionality, from 140 dimensions for the 33-bus system to 400 dimensions for the 118-bus system. However, with the use of learnt attention, the agents continue to only communicate with the most relevant two or three agents, which are electrically important, no matter the total number of agents in the system. This reduces effective coordination to approximately 10–12 active pairs system-wide, yielding ~73% sparsity slightly higher than on the smaller network. The preserved sparsity directly alleviates observation dimensionality growth and explains the favourable scalability behaviour observed in Section 4.10.

Table 8 isolates the impact of the attention mechanism through a controlled ablation against HMA-TD3, where the two architectures differ only by the inclusion of attention mechanism. On the 33-bus system, attention reduces power loss by 12.7% (4.82 → 4.21 MW) and decreases voltage violations by 87.4% ($4.52 \times 10^{-3} \rightarrow 5.70 \times 10^{-4}$ p.u.^{2}), resulting in a 12.9% improvement in cumulative reward. These gains come with only modest additional computation. Training time increases by about 4 s per episode, which corresponds to approximately 14.6% overhead, whereas inference time rises by 3.9 ms per control step or 46.4%.

Importantly, the extra computation introduced by the attention mechanism remains manageable as the system grows larger. For the 33-bus system, the average training time rises from 0.48 to 0.55 min per episode, corresponding to a 14.6% increase. On the larger 118-bus system, the increase is slightly lower at 11.8%, with training time growing from 1.27 to 1.42 min per episode, even though the number of coordinated agents is 2.5 times higher. A similar trend is observed during inference. The latency increase drops from 46.4% on the 33-bus system (8.4–12.3 ms) to 40.9% on the 118-bus system (27.4–38.6 ms). This behaviour indicates that the computational overhead grows more slowly than the system size, whereas the performance gains become more pronounced. In particular, the improvement in cumulative reward rises from 12.9% on the 33-bus network to 17.6% on the 118-bus network, suggesting that the benefit of attention increases as coordination requirements become more complex. By contrast, fixed-topology methods behave much less favourably. For example, HMA-TD3 experiences a 156-fold increase in voltage violations when moving from the 33-bus system to the 118-bus system.

To make the deployment cost explicit, Table 8 additionally reports the online execution time and the interagent communication burden alongside the training and inference overheads. The online execution time is the per-step latency required to map a fresh observation to a control action and is therefore identical to the inference time measured over 1000 forward passes. For AE-HMATD3, this latency is 12.3 ms on the 33-bus system and 38.6 ms on the 118-bus system. Since the fast devices are dispatched only once every 15-min sub-interval, these values occupy $< 0.005\%$ of the available control window, leaving more than four orders of magnitude of timing headroom; the negligible interagent message exchange over a local network does not alter this conclusion. The communication burden is reported as the number of active agent pairs that exchange observations per control cycle, which is governed entirely by the active-link topology established by the learnt attention weights discussed earlier in this section. Because each agent attends only to a small number of its most electrically relevant neighbours, the number of active pairs falls from 6 to about 2 on the 33-bus system and from 45 to about 12 on the 118-bus system, that is, communication reductions of roughly 67% and 73%, respectively. The key observation is that this burden scales with the number of electrically coupled neighbours rather than with the total agent count, which is the underlying reason the framework stays within the resources of conventional distribution management hardware as the network grows.

TABLE 8 | Isolated attention mechanism contribution (pure ablation comparing HMA-TD3 vs. AE-HMATD3) (mean ± std).

System	Metric	HMA-TD3	AE-HMATD3	Improvement
33-bus	Power loss (MW)	4.821 ± 0.08	4.207 ± 0.06	**−12.7%**
	Voltage violations	$4.521 \times 10^{-3} \pm 0.093$	$5.696 \times 10^{-4} \pm 0.001$	**87.4%**
	Cumulative reward	-4.828 ± 0.09	-4.207 ± 0.06	**+12.9%**
	Training time (min/epi)	0.48 ± 0.03	0.55 ± 0.03	**+14.6% overhead**
	Online execution time (ms/step)	8.4 ± 0.3	12.3 ± 0.4	**+46.4% overhead**
	Interval utilisation (15-min window)	0.0009%	0.0014%	**≈ 730,00× headroom**
	Communication burden (active pairs)	6 (full)	≈ 2	**−67%**
118 bus	Power loss (MW)	20.251 ± 0.36	17.232 ± 0.60	**−14.9%**
	Voltage violations	$8.494 \times 10^{-1} \pm 0.165$	$3.706 \times 10^{-2} \pm 0.032$	**−95.6%**
	Cumulative reward	-21.242 ± 0.52	-17.508 ± 0.64	**+17.6%**
	Training time (min/epi)	1.27 ± 0.05	1.42 ± 0.06	**+11.8% overhead**
	Online execution time (ms/step)	27.4 ± 1.2	38.6 ± 1.5	**+40.9% overhead**
	Interval utilisation (15-min window)	0.0030%	0.0043%	**≈ 230,00× headroom**
	Communication burden (active pairs)	45 (full)	≈ 12	**−73%**

Note: Bold values denote the results of the proposed AE-HMATD3 method.

25152947, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/stg2.70107 by Sheffield Hallam University, Wiley Online Library on [24/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

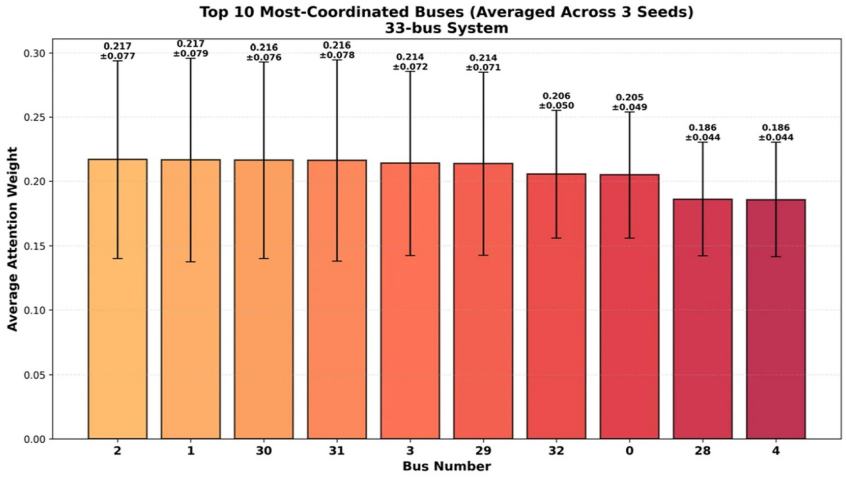

FIGURE 10 | Learned attention coordination patterns (IEEE 33-bus).

Figure 10 illustrates the coordination patterns learnt by the attention mechanism on the 33-bus system. Six buses 2, 1, 30, 31, 3 and 29 consistently receive the highest attention weights, ranging from 0.214 to 0.217 with a standard deviation of approximately 0.07. These six buses not only define a connected feeder section with several junction nodes but also represent the points where voltage is most sensitive to reactive power injection. This indicates that the model can recognise electrically critical points on its own without any topology information or Jacobian calculations.

The buses where controllable devices are present tend to have higher attention. For example, Bus 32, where Agent 3's SVC device is present, has an attention value of 0.206 ± 0.050. Similarly, Bus 0 has an attention value of 0.205 ± 0.049. This bus electrically connects to the inverter device at Bus 17. This shows that the coordination policy focuses on both controllable devices and electrically critical points. The attention value gradually decreases from around 0.217 at the central junctions towards the endpoints of the feeder. This is natural as electrical coupling decreases along a radial network. However, the performance of the policy across the network is consistent over multiple training runs. The standard deviation varies between 0.044 and 0.079, which is around 20%–36% of the mean.

Similarly, the same trend can be seen in the 118-bus system, where the agents at critical junction points and major load centres continue to have higher attention weights. Thus, the mechanism learns a data-driven representation of voltage sensitivity, which improves the coordination of the system while also offering a better understanding of the interactions between the agents.

5 | Conclusion

This work presented an attention-enabled hierarchical multi-agent DRL framework for coordinated voltage control in active distribution system. The framework decouples the control of slow devices from fast devices to activate each device type at its optimal timescale without compromising overall coordination. Specifically, Layer 1 centrally schedules the slow-responding devices (SRDs) on an hourly basis, whereas Layer 2 is the attention-enabled multiagent layer that regulates the fast-responding devices (FRDs) every 15 min. A key innovation of the present work is the incorporation of an attention mechanism within the actor networks, which allows agents to autonomously identify and prioritise electrically significant neighbours. This eliminates the need of computational burden associated with centralised sensitivity calculations. The effectiveness of the proposed control scheme was validated using IEEE 33-bus and IEEE 118-bus distribution networks. Simulation results demonstrated that proposed approach reduced voltage violations to 87.4% (4.521×10^{-3} p.u.$^2 \rightarrow 5.696 \times 10^{-4}$ p.u.^{2}) whereas exhibiting 12.7% improvement in active power loss reduction over HMA-TD3 for IEEE 33 bus system. These improvements were more convincing on the larger 118-bus system, where voltage violations declined to a remarkable 95.6% and power loss reduced by 14.9%. A further operational evaluation over 100 test episodes confirmed that the converged policy keeps bus voltages within the 0.95–1.05 p.u. band with no overvoltage, drawing only 10.5% and 22.0% of the available reactive capability on the 33- and 118-bus systems, respectively, and thus retaining a large reactive reserve. However, these improvements were attained with a training time overhead of about 12%–14% per episode and a per-step online execution time of 12.3 and 38.6 ms on the 33- and 118-bus systems, well within the 15-min control window. This overhead in the training time is acceptable since it is done offline, and the improvements in voltage regulation and loss reduction are quite significant.

One particularly noteworthy finding that may be observed from the scalability experiments is how the attention mechanism adapts its communication pattern as the network grows Instead of being active with all the neighbouring agents, each agent is observed to interact with two or three agents that are electrically close to it whereas ignoring the rest. In case of the 33-bus system, almost 67% of the possible pairs had their attention weights below 0.05, making them inactive. Even in the case of the 118-bus system, where the number of interactions increased substantially, only 10–12 pairs were active among approximately 45, making up almost 73%.

Apart from the control performance, another interesting feature is observed. As all training runs are considered, a clear tendency

25152947, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/stg2.70107 by Sheffield Hallam University, Wiley Online Library on [24/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

is that the attention weights are consistently focusing on the electric sensitive buses, even without any specific information given regarding the network topology. This could be an interesting feature as this could be a useful tool for fault detection. It points towards a promising direction for future work as well developing physics-aware data-driven control approaches.

Author Contributions

Hafiz Mehboob Riaz: conceptualisation, data curation, investigation, software, writing – original draft. **Malik Intisar Ali Sajjad:** conceptualisation, data curation, formal analysis, investigation, methodology, project administration, resources, software, supervision, validation, visualisation, writing – review and editing. **Muhammad Akmal:** supervision, validation, visualisation, writing – review and editing.

Funding

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from any author upon reasonable request.

References

1. D. Cao, J. Zhao, W. Hu, et al., "Model-Free Voltage Control of Active Distribution System With PVs Using Surrogate Model-Based Deep Reinforcement Learning," *Applied Energy* 306 (January 2022): 117982, https://doi.org/10.1016/j.apenergy.2021.117982.

2. A. A. Stephen, K. Musasa, and I. E. Davidson, "Voltage Rise Regulation With a Grid Connected Solar Photovoltaic System," *Energies (Basel)* 14, no. 22 (November 2021): 7510, https://doi.org/10.3390/en14227510.

3. S. M. Abdelkader, S. Kinga, E. Ebinyu, et al., *Advancements in Data-Driven Voltage Control in Active Distribution Networks: A Comprehensive Review* (Elsevier B.V., 2024), https://doi.org/10.1016/j.rineng.2024.102741.

4. H. Mataifa, S. Krishnamurthy, and C. Kriger, "Volt/VAR Optimization: A Survey of Classical and Heuristic Optimization Methods," *IEEE Access* 10 (2022): 13379–13399, https://doi.org/10.1109/ACCESS.2022.3146366.

5. T. Xu, W. Wu, Y. Hong, J. Yu, and F. Zhang, "Data-Driven Inverter-Based Volt/VAR Control for Partially Observable Distribution Networks," *CSEE Journal of Power and Energy Systems* 9, no. 2 (March 2023): 548–560, https://doi.org/10.17775/CSEEJPES.2020.05920.

6. R. Diao, Z. Wang, D. Shi, Q. Chang, J. Duan, and X. Zhang, "Autonomous Voltage Control for Grid Operation Using Deep Reinforcement Learning," in *2019 IEEE Power & Energy Society General Meeting (PESGM)* (IEEE, 2019), 1–5.

7. R. Dutta and K. S. Swarup, "Voltage Control in Radial Distribution Systems Using Deep Reinforcement Learning Technique," in *2024 IEEE PES Innovative Smart Grid Technologies-Asia (ISGT Asia)* (IEEE, 2024), 1–5.

8. C. Li, Y.-A. Chen, C. Jin, R. Sharma, and J. Kleissl, "Online PV Smart Inverter Coordination Using Deep Deterministic Policy Gradient," *Electric Power Systems Research* 209 (2022): 107988, https://doi.org/10.1016/j.epsr.2022.107988.

9. Q. Liu, Y. Guo, L. Deng, H. Liu, D. Li, and H. Sun, "Residual Deep Reinforcement Learning for Inverter-Based Volt-VAR Control," preprint, arXiv:2408.06790 (2024).

10. H. Liu, W. Wu, and Y. Wang, "Bi-Level Off-Policy Reinforcement Learning for Two-Timescale Volt/VAR Control in Active Distribution Networks," *IEEE Transactions on Power Systems* 38, no. 1 (January 2023): 385–395, https://doi.org/10.1109/TPWRS.2022.3168700.

11. H. M. Riaz and M. I. A. Sajjad, "Multi-Agent Double Time Scale Two Critic Deep Reinforcement Learning for Voltage Control in Active Distribution Systems," *Sustainable Energy, Grids and Networks* 45 (2026): 102077, https://doi.org/10.1016/j.segan.2025.102077.

12. R. Hossain, M. Gautam, J. Thapa, H. Livani, and M. Benidris, "Deep Reinforcement Learning Assisted Co-Optimization of Volt-VAR Grid Service in Distribution Networks," *Sustainable Energy, Grids and Networks* 35 (September 2023): 101086, https://doi.org/10.1016/j.segan.2023.101086.

13. J. Zhang, Y. Li, Z. Wu, et al., "Deep-Reinforcement-Learning-Based Two-Timescale Voltage Control for Distribution Systems," *Energies (Basel)* 14, no. 12 (June 2021): 3540, https://doi.org/10.3390/en14123540.

14. C. Luo, H. Wu, Y. Zhou, Y. Qiao, and M. Cai, "Network Partition-Based Hierarchical Decentralised Voltage Control for Distribution Networks With Distributed PV Systems," *International Journal of Electrical Power & Energy Systems* 130 (2021): 106929, https://doi.org/10.1016/j.ijepes.2021.106929.

15. H. Fallahzadeh-Abarghouei, M. Nayeripour, S. Hasanvand, and E. Waffenschmidt, "Online Hierarchical and Distributed Method for Voltage Control in Distribution Smart Grids," *IET Generation, Transmission and Distribution* 11, no. 5 (2017): 1223–1232, https://doi.org/10.1049/iet-gtd.2016.1096.

16. A. R. Malekpour, A. M. Annaswamy, and J. Shah, "Hierarchical Hybrid Architecture for Volt/VAR Control of Power Distribution Grids," *IEEE Transactions on Power Systems* 35, no. 2 (2019): 854–863, https://doi.org/10.1109/tpwrs.2019.2941969.

17. M. Liu, H.-D. Chiang, and T. Li, "Three-Timescale Hierarchical Multi-Objective Volt/VAR Control in Active Distribution Network With Inverters Droop Control," *IEEE Access* 12 (2024): 197602–197615, https://doi.org/10.1109/access.2024.3519358.

18. H. Cheng, H. Luo, Z. Liu, C. Wu, W. Sun, and W. Li, "Hierarchical Reinforcement Learning for Volt/VAR and Wireless Communication Co-Scheduling in Active Distribution Network," *IEEE Internet of Things Journal* 12, no. 20 (2025): 41496–41509, https://doi.org/10.1109/JIOT.2025.3601580.

19. M. Hashemnezhad and P. Aristidou, "Learning-Based Hierarchical Volt/VAR and Demand Flexibility Control in Active Distribution Networks," paper presented at the IFAC World Congress, (IFAC 2026), https://sps-lab.org/publication/2026chashemnezhad/.

20. Z. Wu, Y. Li, W. Gu, et al., "Multi-Timescale Voltage Control for Distribution System Based on Multi-Agent Deep Reinforcement Learning," *International Journal of Electrical Power & Energy Systems* 147 (May 2023): 108830, https://doi.org/10.1016/j.ijepes.2022.108830.

21. X. Sun and J. Qiu, "Two-Stage Volt/VAR Control in Active Distribution Networks With Multi-Agent Deep Reinforcement Learning Method," *IEEE Transactions on Smart Grid* 12, no. 4 (July 2021): 2903–2912, https://doi.org/10.1109/TSG.2021.3052998.

22. K. Xiong, C. Ye, B. Liu, and X. Suo, "Multi-Agent Deep Reinforcement Learning-Enabled Voltage Regulation Approach for Partitioned Active Distribution Network Using Heterogeneous PV Inverters," *Energy Conversion and Economics* 6, no. 6 (2025): 359–370, https://doi.org/10.1049/enc2.70026.

25152947, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/stg2.70107 by Sheffield Hallam University, Wiley Online Library on [24/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

23. X. Liu, C. Zhang, S. Li, and J. Zhu, "Data-Driven VVC for Unbalanced Networks With Tuning-Friendly Deep Reinforcement Learning," in *2024 IEEE Power & Energy Society General Meeting (PESGM)* (IEEE, 2024), 1–5.

24. S. Fujimoto, D. Meger, and D. Precup, "Off-Policy Deep Reinforcement Learning Without Exploration," [Online], https://github.com/sfujim/BCQ.

25. S. Fujimoto, H. Van Hoof, and D. Meger, "Addressing Function Approximation Error in Actor-Critic Methods," (2018) [Online], https://github.com/.

26. H. M. Riaz, M. I. A. Sajjad, and M. Akmal, "Multiagent Twin Delayed Deep Deterministic Policy Gradient Approach for Voltage Control of Distribution System," in *2025 60th International Universities Power Engineering Conference (UPEC)* (2025), 1–6, https://doi.org/10.1109/UPEC65436.2025.11279918.

27. Y. Qi, Z. Dang, P. Yang, L. Dai, X. Xu, and Y. Liu, "TD3-Based Voltage Regulation for Distribution Networks With PV and Energy Storage System," in *Proceedings – 2023 Panda Forum on Power and Energy, PandaFPE 2023* (Institute of Electrical and Electronics Engineers Inc., 2023), 505–509, https://doi.org/10.1109/PandaFPE57779.2023.10140334.

28. F. Liu and Q. Xu, "Safe Deep Reinforcement Learning Based Volt-VAR Control for Three-Phase Unbalanced Distribution System With PV Integration," in *2024 IEEE Energy Conversion Congress and Exposition (ECCE)* (IEEE, 2024), 1823–1828.

29. Y. Zhou, Y. Peng, L. Ge, L. Hou, Y. Wang, and H. Niu, "Two-Timescale Volt/VAR Control Based on Reinforcement Learning With Hybrid Action Space for Distribution Networks," *Journal of Modern Power Systems and Clean Energy* 13, no. 4 (2025): 1261–1273, https://doi.org/10.35833/MPCE.2024.000643.

30. S. Iqbal and F. Sha, "Actor-Attention-Critic for Multi-Agent Reinforcement Learning," (2019) [Online], http://arxiv.org/abs/1810.02912.

31. D. Cao, J. Zhao, W. Hu, F. Ding, Q. Huang, and Z. Chen, "Attention Enabled Multi-Agent DRL for Decentralized Volt-VAR Control of Active Distribution System Using PV Inverters and SVCs," *IEEE Transactions on Sustainable Energy* 12, no. 3 (2021): 1582–1592, https://doi.org/10.1109/tste.2021.3057090.

32. J. Huang, H. Zhang, D. Tian, Z. Zhang, C. Yu, and G. P. Hancke, "Multi-Agent Deep Reinforcement Learning With Enhanced Collaboration for Distribution Network Voltage Control," *Engineering Applications of Artificial Intelligence* 134 (2024): 108677, https://doi.org/10.1016/j.engappai.2024.108677.

33. H. Liu and W. Wu, "Two-Stage Deep Reinforcement Learning for Inverter-Based Volt-VAR Control in Active Distribution Networks," *IEEE Transactions on Smart Grid* 12, no. 3 (May 2021): 2037–2047, https://doi.org/10.1109/TSG.2020.3041620.

34. L. Thurner, A. Scheidler, F. Schafer, et al., "Pandapower – An Open-Source Python Tool for Convenient Modeling, Analysis, and Optimization of Electric Power Systems," *IEEE Transactions on Power Systems* 33, no. 6 (November 2018): 6510–6521, https://doi.org/10.1109/TPWRS.2018.2829021.

35. W. Zheng, W. Wu, B. Zhang, and Y. Wang, "Robust Reactive Power Optimisation and Voltage Control Method for Active Distribution Networks via Dual Time-Scale Coordination," *IET Generation, Transmission & Distribution* 11, no. 6 (2017): 1461–1471, https://doi.org/10.1049/iet-gtd.2016.0950.

36. Q. Liu, Y. Guo, L. Deng, et al., "Two-Critic Deep Reinforcement Learning for Inverter-Based Volt-VAR Control in Active Distribution Networks," *IEEE Transactions on Sustainable Energy* 15, no. 3 (July 2024): 1768–1781, https://doi.org/10.1109/TSTE.2024.3376369.

37. T. P. Lillicrap, J. J. Hunt, A. Pritzel, et al., "Continuous Control With Deep Reinforcement Learning," in *Proceedings of the 4th International Conference on Learning Representations (ICLR)* (ICLR, 2016), https://arxiv.org/abs/1509.02971.

38. T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning With a Stochastic Actor," in *International Conference on Machine Learning* (PMLR, 2018), 1861–1870.

25152947, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/stg2.70107 by Sheffield Hallam University, Wiley Online Library on [24/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License