

Article

Attention-Enhanced Multi-Agent Deep Reinforcement Learning for Inverter-Based Volt-VAR Control in Active Distribution Networks

Wenwen Chen ¹, Hao Niu ^{1,*}, Linbo Liu ¹, Jianglong Lin ¹ and Huan Quan ²

¹ Shaoguan Power Supply Bureau, Guangdong Power Grid Co., Ltd., Shaoguan 512026, China; 13927825652@139.com (W.C.); 13927827267@139.com (L.L.); lt803002@163.com (J.L.)

² School of Electric Power Engineering, South China University of Technology, Guangzhou 510641, China; ephuan_quan@mail.scut.edu.cn

* Correspondence: louis876565244@outlook.com

Abstract

The increasing penetration of inverter-interfaced photovoltaic (PV) generation in active distribution networks (ADNs) intensifies fast voltage violations and makes real-time Volt-VAR control (VVC) challenging, especially when each inverter has only partial and noisy measurements and communication is limited. Existing local droop-type strategies lack coordination, while fully centralized optimization/learning is often impractical for online deployment. To address these gaps, an attention-enhanced multi-agent deep reinforcement learning (MADRL) framework is developed for inverter-based VVC under the centralized training and decentralized execution (CTDE) paradigm. First, the voltage regulation problem is formulated as a decentralized partially observable Markov decision process (Dec-POMDP) to explicitly account for system stochasticity and temporal variability under partial observability. To solve this complex game, an attention-enhanced MADRL architecture is employed, where an agent-level attention mechanism is integrated into the centralized critic. Unlike traditional methods that treat all neighbor information equally, the proposed mechanism enables each inverter agent to dynamically prioritize and selectively focus on the most influential states from other agents, effectively capturing complex intercorrelations while enhancing training stability and learning efficiency. Operating under the CTDE paradigm, the framework realizes coordinated reactive power support using only local measurements, ensuring high scalability and practical implementability in communication-constrained environments. Simulations on the IEEE 33-bus system with six PV inverters show that the proposed method reduces the average voltage deviation on the test set from 0.0117 p.u. (droop control) and 0.0112 p.u. (MADDPG) to 0.0074 p.u., while maintaining millisecond-level execution time comparable to other MADRL baselines. Scalability tests with up to 12 agents further demonstrate robust performance of the proposed method under higher PV penetration.

Keywords: reinforcement learning; Volt-VAR control; active distribution networks; attention mechanism; partially observable Markov decision process

MSC: 37M10; 68T05; 90C40

Academic Editor: Denis N. Sidorov

Received: 29 January 2026

Revised: 22 February 2026

Accepted: 25 February 2026

Published: 1 March 2026

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

As the penetration of distributed generations (DGs) in the active distribution network (ADN) continues to increase, it poses a huge challenge to the operation and control of the distribution system [1]. The access of a large number of DGs changes the power flow of traditional DNs, leading to the reverse of power flow, which has great influence on voltage fluctuation and power system reliability [2]. Voltage violation problems and network losses are getting increasingly serious with higher-level integration of DGs, due to the volatility, randomness, and intermittency of DGs [3]. Among numerous technical challenges, the voltage violations problem is of particular attention.

To solve this problem, various voltage control approaches have been developed, which can be broadly classified into active power-based and reactive power-based strategies. The voltage control methods based on active power, including power curtailment of photovoltaic (PV) [4] and the distributed energy storage system (DESS) [5], are suitable for low-voltage ADNs with a high resistance/reactance ratio. However, curtailing PV output inevitably reduces the utilization of DGs, and thus does not improve the PV hosting capacity of ADNs, nor can it provide voltage support during nighttime periods. Moreover, the charging and discharging scheduling of DESSs is costly and may accelerate battery degradation.

In contrast, reactive power-based strategies regulate bus voltages by coordinating voltage regulation devices such as capacitor banks, static var compensators, and inverter-interfaced DGs [6]. Among these methods, inverter-based voltage regulation has attracted increasing attention due to its fast response, flexibility, and low implementation cost [7]. By dynamically adjusting the reactive power output of inverters according to local voltage conditions, inverter-based Volt-VAR control (VVC) provides an effective means to mitigate voltage violations and fluctuations in active distribution networks without sacrificing active power generation. As most DGs are inverter-based energy resources (IB-ERs), they are reasonable to provide rapid Volt-VAR support with their free reactive power capacity [8,9]. Additionally, a large amount of PV is distributed in ADNs. Therefore, the PV inverter is selected as the voltage control equipment, and the optimal control of the system voltage is realized by cooperatively adjusting the reactive power output of the inverters to further improve the voltage quality.

Among the inverter-based VVC methods, traditional droop control methods adjust inverter reactive power based on predefined voltage–reactive power characteristic curves without explicit coordination among distributed generation units. Although these approaches are simple and require only local measurements, they inherently lack global optimization capabilities and often fail to suppress voltage violations under high DG penetration and dynamic load fluctuations [10]. Moreover, conventional optimization-based paradigms—including deterministic programming, metaheuristic optimization, and stochastic formulations—are not always well suited to highly uncertain and fast-varying ADN operations. Their reliance on accurate system models and parameters, together with considerable computation and the risk of local optimum trapping, often limits their scalability and real-time applicability [11–13]. These limitations highlight the need for a scalable and coordination-aware control framework that can operate under uncertainty and time-varying ADN conditions.

Distinct from the aforementioned control strategies, deep reinforcement learning (DRL) originates from behavioral psychology and enables agents to learn optimal decision-making policies through continuous interaction with the environment, with the objective of maximizing cumulative rewards or achieving predefined control goals [14]. In recent years, driven by advances in artificial intelligence, DRL has been increasingly applied to a wide range of domains [15–17]. Existing studies have demonstrated that DRL provides an effective data-driven framework for reactive power and voltage regulation in

ADNs. For the VVC problem of ADNs, several studies have shown that DRL can be used as a model-independent online scheduling and control means to effectively improve voltage quality and suppress over-voltage and under-voltage events under high penetration of distributed energy resources. In [18], a deep Q-network (DQN)-based Volt-VAR optimization method is proposed for ADNs. However, the Q-learning algorithm is inherently limited to discrete action spaces, restricting its applicability to continuous inverter control. In [19,20], hybrid two-timescale DRL frameworks combining DQN and deep deterministic policy gradient (DDPG) were developed, where the DQN handles discrete control variables and the DDPG manages continuous actions. Despite improved flexibility over pure DQNs, the approach treats each device independently, posing scalability challenges and coordination inefficiencies in larger networks. More recently, multi-agent deep reinforcement learning (MADRL) methods have been introduced to enhance coordinated control performance. In [21], a hybrid two-stage VVC framework integrating multi-agent deep deterministic policy gradient (MADDPG) with mixed-integer second-order cone programming (MISOCP) is proposed to suppress fast voltage violations while minimizing network power losses. As a representative state-of-the-art MADRL algorithm, MADDPG enables decentralized execution by allowing multiple agents to jointly learn continuous control policies under a centralized training paradigm. However, it still suffers from unstable training and information redundancy as network size and state complexity increase. Beyond value-based and actor–critic methods, multi-agent proximal policy optimization (MAPPO) has also been explored for power system dispatch and coordinated control tasks, owing to its improved policy stability and sample efficiency [22]. However, applying MAPPO to precise continuous VVC problems often requires careful algorithmic tuning and tailored network architectures. Furthermore, in [23], a multi-agent twin-delayed deep deterministic policy gradient (MATD3) framework for ADN decentralized VVC is proposed, demonstrating improved robustness over conventional MADDPG-based approaches. Recent studies have started to integrate attention mechanisms into MADRL-based VVC to better improve coordinated control, such as physical-assisted multi-agent graph attention RL for fast voltage regulation in PV-rich ADNs and domain knowledge-enhanced graph attention RL for VVC [24,25].

Overall, existing DRL-based VVC methods still face several practical pain points in ADNs, including partial observability and stochastic operating conditions, training instability, and information redundancy.

To address the above challenges, this paper develops an attention-enhanced multi-agent actor–critic algorithm for inverter-based VVC in ADNs. Specifically, to explicitly capture stochastic variations and imperfect local observations, we formulate inverter-based VVC as a decentralized partially observable Markov decision process (Dec-POMDP) model. To mitigate training instability and to reduce information redundancy, we embed a multi-head attention mechanism into the centralized critic. Under a centralized training and decentralized execution (CTDE) paradigm, the learned policies are executed online using only local measurements, enabling scalable and communication-efficient coordinated control for practical real-time deployment. The main contributions of this paper are highlighted as follows:

- (1) Dec-POMDP-based modeling and attention-aided centralized critic. We formulate inverter-based VVC in ADNs as a Dec-POMDP to explicitly capture stochastic variations and imperfect local observations. We embed a multi-head attention mechanism into the centralized critic to evaluate the global action-value function, enabling selective aggregation of the most relevant information from other agents during training. This attention-aided critic significantly improves training stability and learning efficiency compared with conventional MADRL approaches.

- (2) CTDE paradigm for practical deployment. The proposed method adopts a CTDE paradigm, where the policy networks trained offline during the centralized training stage are executed online using only local observations. This design ensures low computational and communication overhead during real-time operation, clearly distinguishing the proposed framework from fully centralized control schemes that require extensive global information exchange. As a result, the method is well suited for ADNs with imperfect communication infrastructures and stringent real-time control requirements. Moreover, inverter-based strategies leverage existing inverter capabilities and thus avoid additional regulating devices, offering a more economical and readily implementable solution.
- (3) Effective and scalable inverter-based VVC performance. The proposed approach effectively mitigates voltage violations and suppresses voltage fluctuations without resorting to active power curtailment. By fully coordinating the reactive power capability of distributed PV inverters, cooperative control among agents is achieved, while each agent independently determines its reactive power adjustment based on shared information within the MADRL framework. Furthermore, the proposed method maintains robust control performance as the number of agents and the penetration level of DGs increase, demonstrating strong robustness and generalization capabilities. This, in turn, enhances the hosting capacity for DGs and improves the overall safety, stability, and reliability of ADNs.

2. Inverter-Based VVC Model

2.1. Principle of PV Inverter Participating in VVC

In this work, PV inverters are operated in maximum power point tracking (MPPT) mode during daytime, while they switch to STATCOM mode at night so that reactive power support can still be provided even when PV generation is unavailable [26]. Such IB-ERs can deliver both active and reactive power to the grid [27]. Specifically, the inverter adjusts its reactive power injection/absorption according to the magnitude of voltage deviation, thereby mitigating voltage deviations. During the midday hours, high PV penetration may raise local voltages, in which case the inverter tends to absorb reactive power to counteract over-voltage. In contrast, during nighttime, when the load increases and PV output is zero, the inverter can utilize its full apparent power capability to regulate reactive power and smooth voltage variations. In practice, grid-connected inverters are commonly overrated for reliability and safety considerations, which leaves residual reactive power capability even when the active power output approaches its rated value. Therefore, the feasible reactive power range is jointly constrained by the inverter's rated apparent power S_{inv} and the instantaneous active power output P_{pv} , which can be expressed as follows:

$$Q_{pv}^{\max} = \pm \sqrt{S_{inv}^2 - P_{pv}^2} \quad (1)$$

where $Q_{pv}^{\max}$ is the current maximum adjustable reactive power capacity of the inverter.

Figure 1 illustrates the relationship between reactive power capacity and active power output of the PV inverter in detail. At point a, the PV output is rated active power, and the PV inverter does not have the ability to provide reactive power to the power grid; at point b, the PV inverter can not only provide active power to the grid, but can also provide a certain reactive power output. If the reactive power demand of the grid is large, the operating point of the PV inverter moves from b to c, and the inverter runs in the rated state at point c. Moreover, if there is still a reactive power gap in the system, the active power output of PV can be appropriately reduced to improve the reactive voltage supporting capacity of PV, as shown at point d. To avoid the phenomenon of solar energy

abandonment and promote the absorption of DGs, PV active power is generated with the maximum power in MPPT control mode, without active power curtailment, in this paper.

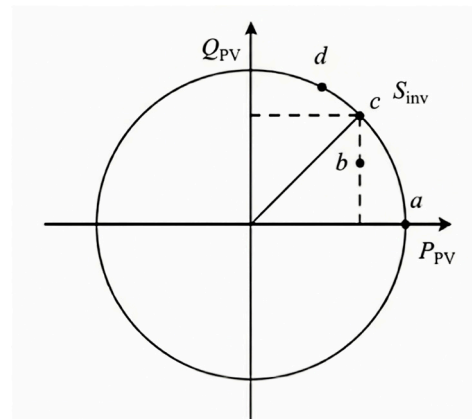


Figure 1. Diagram of photovoltaic inverter P-Q capacity.

2.2. VVC Model in Power Distribution Networks

Real-time VVC in ADNs enhances voltage quality by coordinating reactive power resources and performing continuous short-timescale adjustments, building upon the day-ahead scheduling plan. The objective function is defined as

$$\min \sum_{i=0}^m (C_{\alpha} \sum_{j=1}^n \Delta V_{i,j}) \quad (2)$$

where m is the number of instruction cycles in a day; C_{α} is additional cost factor for voltage violation; n is the node number of the distribution system; and $\Delta V_{i,j}$ is the voltage deviation of node j in the i th command cycle.

The constraints are as follows:

$$V_i^{\min} \leq V_i \leq V_i^{\max} \quad (3)$$

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

$$Q_{i,PV}^{\max 2} \leq S_{i,PV}^2 - P_{i,PV}^2 \quad (5)$$

$$P_{G,i} - P_{L,i} = V_i \sum_{j=1}^n V_j (G_{ij} \cos \delta_{ij} + B_{ij} \sin \delta_{ij}) \quad (6)$$

$$Q_{G,i} - Q_{L,i} = V_i \sum_{j=1}^n V_j (-B_{ij} \cos \delta_{ij} + G_{ij} \sin \delta_{ij}) \quad (7)$$

where Equation (3) enforces the nodal voltage limits, Equation (4) specifies the active power output bounds of PV inverters, and Equation (5) defines their reactive power capability. The power flow relationships are described by Equations (6) and (7), where $P_{G,i}$ and $Q_{G,i}$ denote the active and reactive power injections at node i ; $P_{L,i}$ and $Q_{L,i}$ represent the corresponding active and reactive demands; V_i and V_j are the voltage magnitudes at nodes i and j ; G_{ij} and B_{ij} are the line conductance and susceptance between node i and j ; and δ_{ij} is the voltage phase angle difference between nodes i and j .

3. VVC Method Based on MAAC

3.1. Multi-Agent Deep Reinforcement Learning

With the development of large-scale DGs in the power system, the power dispatch and control center is faced with a new situation of significantly increased uncertainty on both sides of the source and load, exponential increase in dispatch objects, and more complex operational characteristics, which put forward higher requirements for the automation and intelligence of power grid dispatching and operation.

DRL is a data-driven paradigm that integrates deep neural networks for representation learning with reinforcement learning for sequential decision-making [28]. By leveraging function approximation, DRL can alleviate the curse of dimensionality encountered in high-dimensional control problems. When applied to power grid scheduling, it can realize self-exploration and update of scheduling policies, and assist dispatchers to complete scheduling tasks in complex power grids. Due to the introduction of the DRL method, on the one hand, the influence of various uncertainties in the power system on the scheduling policy can be considered in advance, and the environmental adaptability of the scheduling policy can be enhanced. On the other hand, DRL can solve the rapidity of online decision-making and improve the efficiency of intra-day scheduling decisions.

DRL is mainly divided into three categories: value-based algorithms, policy-based algorithms, and AC algorithms [29]. The AC method is combined with policy gradient and value function, in which the actor uses the policy gradient function to generate actions and interact with the environment, and the critic uses the value function to evaluate the performance of the actor and guide the action of the actor in the next state. DDPG is the typical representative algorithm of the AC framework.

For ADN real-time scheduling decision problems, the AC framework has natural adaptability. At present, in the rolling calculation of intra-day scheduling plans, the scheduling plan is usually calculated based on the optimization algorithm and implemented first, and then the dispatcher determines whether the plan is risky based on the constantly updated data information, safety checks and his own scheduling experience. If it is risky, he will adjust and form the final plan. Over time, dispatchers will accumulate more and more scheduling experience, and the scheduling plan will be more stable and efficient. This closed-loop process of execution–feedback is completely consistent with the AC framework. In the training process of the AC framework, the actor constantly generates strategies based on the current experience. The critic gives a judgment on these strategies like a dispatcher, and then the actor updates and accumulates their own experience according to the judgment results, making the evaluation of the critic higher and higher in order to obtain the optimal strategy. Therefore, the AC framework is just like a virtual dispatcher continuously working and learning, and is one of the most suitable DRL methods for intra-day real-time scheduling computations.

In contrast to single-agent DRL, multi-agent settings allow agents to coordinate their behaviors, which can lead to higher global rewards in multi-agent environments. Through inter-agent coordination and information exchange, MADRL can learn cooperative policies that improve the overall system-level performance, especially for large-scale ADNs with spatially distributed devices [30]. The agent trained by offline data for online decision making can adapt to source-load uncertainty, which is helpful to solve the voltage optimization problem of ADNs considering DGs and load uncertainty.

3.2. Offline Centralized Training and Online Decentralized Execution Framework

In MADRL problems, there are usually two training frameworks: decentralized and centralized. In the decentralized structure, each agent's training is independent of other agents, and its strategy network takes actions according to local observation output. In the training process, agents regard other agents as part of the environment. However, the training of other agents will change the state transfer function in the environment, which

will destroy the Markov hypothesis followed by the reinforcement learning algorithm. Therefore, the decentralized training framework faces problems of training instability and poor convergence.

The centralized training framework solves this problem by jointly modeling all agents to learn a strategy, the input of which is the joint observations of all agents, and the output is the joint actions. Nevertheless, as the number of agents grows, the input and output space will grow exponentially, making it impossible to explore and difficult to adapt to large scale multi-agent systems. In addition, limited by communication conditions, it is difficult for agents to obtain the global state in the real ADN environment.

Therefore, the framework of CTDE is adopted in this work. During offline training, communication among agents is not constrained by practical channel limitations, allowing unrestricted information sharing. Consequently, each agent can access global system information. A centralized critic network is learned through the joint observation and action of multiple agents. In the online execution phase, the agent makes decisions based on its own local observations and uses the trained policy network to realize distributed real-time execution. The CTDE framework not only meets the requirement of distributed execution of agents in the Dec-POMDP model, but also models the benefits of agents' joint actions in the training period to make agents learn to cooperate. To some extent, it overcomes the problems of environmental instability and dimensional disaster. The schematic diagram of CTDE based on the AC framework is shown in Figure 2.

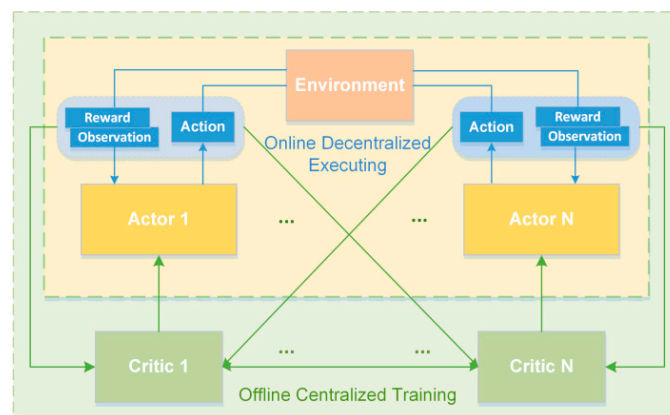


Figure 2. Schematic diagram of CTDE based on AC framework.

Figure 3 shows the offline centralized training and online decentralized execution framework of the ADN based on DRL. The offline training module builds a simulation environment based on the information of the ADN's topology, branch parameters, and equipment model to simulate the real system operation of ADNs. Combined with the simulation environment, historical scene data samples such as load data and DG data, objective function and other information are inputs of the offline agent. In the simulation environment, the actions produced by agents are simulated to generate a new power grid operation mode, and the rewards are calculated and the offline agent network parameters are updated. The offline agent iteratively interacts with the simulation environment until the parameters of the neural network gradually become stable.

The online executing module inputs the online real-time operation status, real-time scene data, objective function, and other information into the online agent. The actions generated by agents are checked and confirmed by the dispatcher, they are delivered to each scheduling object (e.g., DG inverter) for execution, and then the next moment is calculated.

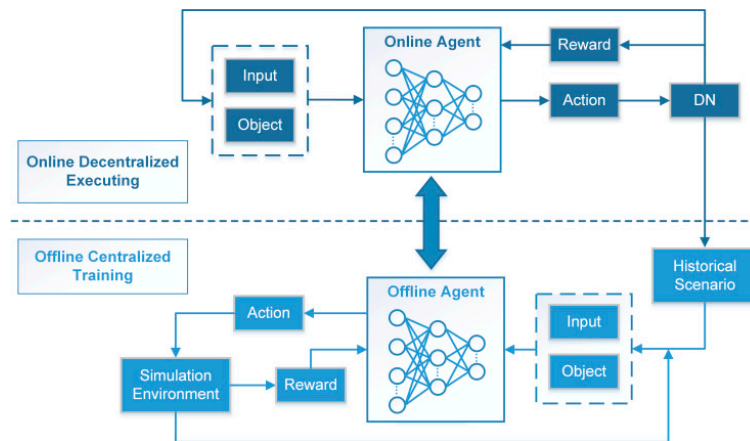


Figure 3. Architecture of offline centralized training and online decentralized execution in ADN.

3.3. Multi-Agent Actor–Attention–Critic (MAAC)

In this work, inverter-based VVC is formulated as a fully cooperative Dec-POMDP, where all agents aim to optimize a shared system-level objective. Let $o_t = (o_{1,t}, \dots, o_{N,t})$ and $a_t = (a_{1,t}, \dots, a_{N,t})$ denote the joint observation and joint action at time t . Under the CTDE paradigm, the joint policy is factorized for decentralized execution as

$$\pi(a_t|o_t) = \prod_{i=1}^N \pi_i(a_{i,t}|o_{i,t}) \quad (8)$$

where each actor π_i uses only its local observation $o_{i,t}$ online. The cooperative objective is to maximize the expected discounted return of the shared reward:

$$J(\pi) = \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t r_t \right] \quad (9)$$

where γ is the discount factor. During centralized training, each agent is equipped with a centralized critic that conditions on (o_t, a_t) to evaluate the joint interaction and provide low-variance gradients for coordinated learning, while the learned actors remain fully decentralized at execution time.

Different from the MADRL algorithms such as MADDPG [31], MAAC can maintain the control performance as the number of agents changes with the attention model. Acquiring the critic for each agent by selectively focusing on the information from other agents is the main innovation of MAAC [32]. Intuitively, each agent inquires for information regarding the observations and actions of the other agents, which is combined into the evaluation of its value function. The calculation of the value function is shown in Figure 4. The critic receives the observations and actions for calculating the value function, which is formulated as

$$Q_i^{\theta}(o, a) = f_i(g_i(o_i, a_i), x_i) \quad (10)$$

where f_i is a two-layer multi-layer perceptron (MLP), g_i is an one-layer MLP embedding function, o_i and a_i are the observations and actions of agent i , respectively, and x_i is a weighted sum of each agent's value:

$$x_i = \sum_{j \neq i} \alpha_j v_j = \sum_{j \neq i} \alpha_j h(V g_j(o_j, a_j)) \quad (11)$$

where v_j is a function of agent j 's embedding, encoded with $e_j = g_j(o_j, a_j)$ and transformed by a shared matrix V , then processed by a activation function h .

The attention weight α_j is obtained by using bilinear mapping (i.e., the query-key system), comparing the similarity values between $\mathbf{g}_j$ and $\mathbf{g}_i$, and passing these two embeddings into a softmax:

$$\alpha_j \propto \exp(e_j^T W_k^T W_q e_i) \quad (12)$$

where W_k and W_q represent the transformation matrices, which are used to prevent gradient loss.

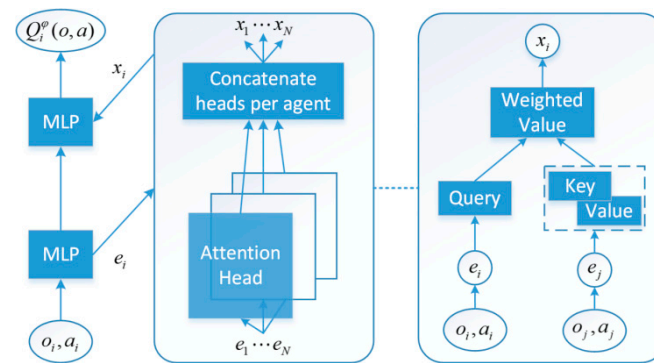


Figure 4. Calculating the value function with attention mechanism.

In the attention-based critic, each agent i forms a query from its own embedding, while each other agent j provides a key/value embedding. The attention weights quantify the relevance of agent j to agent i , and the weighted aggregation is used to compute the centralized action-value. In implementation, the reactive power commands are clipped to satisfy the inverter capability limits in Equation (5), ensuring physically feasible actions throughout training and execution.

The weighted aggregation of information contributed by other agents to agent i is determined by the parameters of each attention head, denoted as

$$Q^a = \{W_k, W_q, V\} \quad (13)$$

which are shared across all critic networks and jointly updated by minimizing the following loss function:

$$L_Q(\phi) = \sum_{i=1}^N E_{(o,a,r,o') \sim D} \left[\left(Q_i^p(o,a) - y_i \right)^2 \right] \quad (14)$$

The target value y_i is computed as

$$y_i = r_i + \gamma E_{a' \sim \pi_{\bar{\theta}}(o')} \left[Q_i^{\bar{p}}(o', a') - \alpha \log(\pi_{\bar{\theta}}(a' | o')) \right] \quad (15)$$

where D denotes the replay buffer storing transition samples, y_i represents the temporal-difference target, and $\bar{\theta}$ and $\bar{\phi}$ are the parameters of the target actors and target critics, respectively. The entropy temperature parameter α balances the trade-off between policy exploration and reward maximization, while γ is the discount factor.

To enhance exploration and learning robustness, an entropy-regularized optimization principle is incorporated into both the actor and critic updates [33]. Unlike conventional deterministic policy gradient methods such as DDPG, which may converge prematurely to suboptimal deterministic policies, the entropy regularization mechanism encourages stochastic policy exploration during training. This allows the learning process to better capture multiple high-quality action modes and improves robustness under non-stationary multi-agent environments.

Policy gradient methods are commonly employed in policy-based DRL algorithms and can be expressed as

$$\nabla_{\theta_i} J(\pi_{\theta}) = \mathbb{E}_{o \sim D, a \sim \pi} [\nabla_{\theta_i} \log(\pi_{\theta_i}(a_i | o_i)) \sum_{k=1}^T \gamma^{k-1} r(o_i, a_i)] \quad (16)$$

For completeness, we outline the entropy-regularized objective used to optimize each actor. For agent i , the objective is

$$J(\pi_{\theta}) = \mathbb{E}_{o \sim D, a \sim \pi} [\sum_i \gamma' (r_i + \alpha \mathcal{H}(\pi_{\theta_i}(\cdot | o_i')))] \quad (17)$$

where $\mathcal{H}(\cdot)$ denotes the policy entropy. Using the log-derivative trick, the policy gradient can be written as

$$\nabla_{\theta_i} J(\pi_{\theta}) = \mathbb{E}_{o \sim D, a \sim \pi} [\nabla_{\theta_i} \log(\pi_{\theta_i}(a_i | o_i)) (-\alpha \log(\pi_{\theta_i}(a_i | o_i)) + Q_i^{\theta}(o, a) - b_i(o))] \quad (18)$$

where $b_i(o)$ is an action-independent baseline. In our implementation, $b_i(o)$ is chosen as the state-value estimate induced by the centralized critic.

$$b_i(o) = E_{a \sim \pi(o)} [Q_i^{\theta}(o, (a_i, a_{-i}))] \quad (19)$$

where a_{-i} represents actions except that of agent i . The baseline term is introduced to compute the advantage function:

$$A_i(o, a) = Q_i^{\theta}(o, a) - b_i(o) \quad (20)$$

which facilitates more effective credit assignment among agents by emphasizing actions that contribute more significantly to the global cooperative objective. Consequently, this design improves coordination efficiency and accelerates policy convergence. It is worth noting that actions used in the actor update are sampled directly from the current policy rather than from the replay buffer, which helps mitigate overgeneralization and further enhances cooperative learning among agents.

The introduction of the attention mechanism enables the centralized critic to adaptively assign higher weights to agents whose observations and actions are more relevant to the voltage regulation task of a specific inverter. By learning context-dependent importance weights, the attention module effectively performs dynamic information filtering, allowing the critic to focus on electrically and operationally significant interactions. This mechanism aligns well with the physical characteristics of voltage regulation in distribution networks, where voltage deviations are predominantly influenced by a limited number of strongly coupled nodes. Consequently, attention-enhanced critics reduce information redundancy, improve credit assignment among agents, and enhance training stability in high-dimensional multi-agent environments. Moreover, attention facilitates improved coordination under decentralized execution by enabling each agent to implicitly capture the influence of other agents through the learned value function, without requiring explicit communication during online operation. This property is particularly advantageous for real-time VVC, where communication constraints and latency may limit the availability of global system information. By selectively emphasizing the most informative agent interactions during training, the attention-enhanced MADRL framework achieves more robust and scalable voltage control policies, especially under varying network conditions and increasing levels of DG penetration.

The training process of the proposed method is shown in Algorithm 1. The executing process is shown in Algorithm 2. For clarity, the overall workflow of the proposed MAAC-based VVC is summarized in Figure 5.

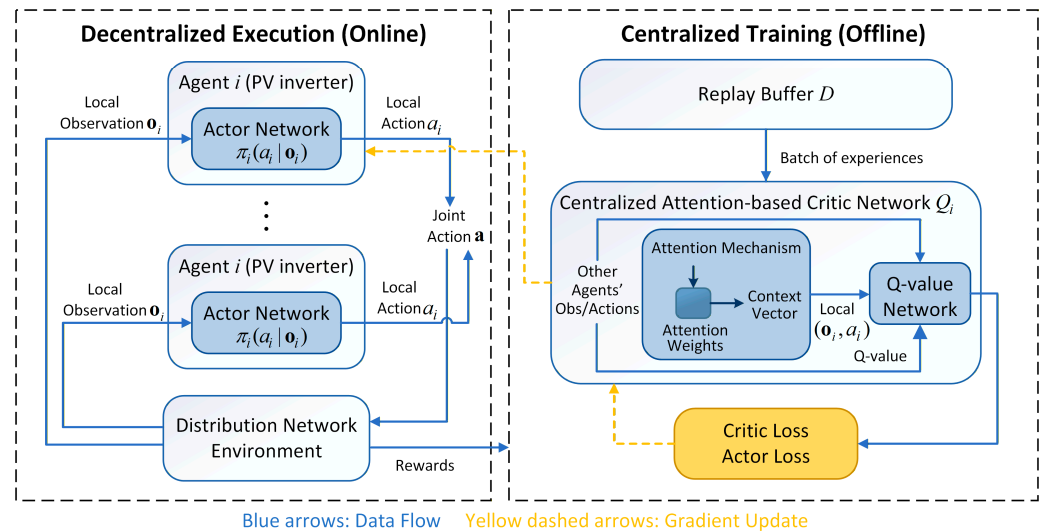


Figure 5. Overall workflow of the proposed attention-enhanced MAAC-based VVC under CTDE.

Algorithm 1 Offline Centralized Training Process

Randomly initialize parameters of actor network θ and centralized critic φ network of agent i
 Initialize target networks $\bar{\theta}$, $\bar{\varphi}$, and replay buffer D .
for episode = 1, 2, ..., H **do**
 Reset environment and obtain the initial global state s
 Obtain initial local observations o_i for each agent i
 for time step = 1, 2, ..., T per episode **do**
 Select action a_i for each agent i
 Execute joint action $a = \{a_1, \dots, a_N\}$, and receive rewards r_i and next state s' .
 Obtain next observations o'
 Store transitions (s, a, r, s', o, o') in replay buffer D
 Set $s \leftarrow s'$, $o_i \leftarrow o'_i$
 Randomly sample a minibatch from D **for** agent $i = 1, 2, \dots, N$ **do**
 Update the centralized critic φ using the attention-based Q-function according to (12) and (13)
 Update the actor network θ using the policy gradient according to (15) and (16)
 end for
 Soft update target networks:
 $\bar{\varphi} \leftarrow \tau\varphi + (1-\tau)\bar{\varphi}$ $\bar{\theta} \leftarrow \tau\theta + (1-\tau)\bar{\theta}$ **end for**
end for

Algorithm 2 Online Decentralized Executing Process

Load the trained actor parameters θ_i for each agent i
for time step $t = 1, 2, \dots, T$ **do**
 for agent $i = 1, 2, \dots, N$ **do**
 Obtain local observation o_i
 Calculate action a_i
 Output the control action to the corresponding inverter
 end for
end for

3.4. Formulation of Dec-POMDP

The VVC problem can be formulated as a Dec-POMDP, where each PV is modeled as an agent and the ADN is the environment. Let $\mathcal{N} = \{1, \dots, |\mathcal{N}|\}$ denote the set of buses and $\mathcal{I} = \{1, \dots, N\}$ denote the set of inverter agents. Define the set of zones as $\mathcal{Z} = \{1, \dots, Z\}$

and the zonal partition of the AND as $\{\mathcal{N}_z\}_{z=1}^Z, \mathcal{N} = \bigcup_z \mathcal{N}_z, \mathcal{N}_z \cap \mathcal{N}_{z'} = \emptyset$. Each agent $i \in \mathcal{I}$ belongs to a zone $\mathcal{N}_{z(i)} \in Z$ and is connected to a bus $b(i) \in \mathcal{N}_{z(i)}$.

In this paper, the Dec-POMDP is defined as a tuple $M = \langle S, \{O_i\}_{i \in \mathcal{I}}, \{A_i\}_{i \in \mathcal{I}}, T, R \rangle$, which is described as follows:

(1) State space. The global state at time step t is defined as

$$\mathbf{s}_t = [\mathbf{P}_t^L, \mathbf{Q}_t^L, \mathbf{P}_t^{\text{PV}}, \mathbf{V}_t, \mathbf{Q}_{t-1}^{\text{PV}}]^\top \in S \quad (21)$$

where $\mathbf{P}_t^L, \mathbf{Q}_t^L$ are the active and reactive load, $\mathbf{P}_t^{\text{PV}}$ is the PV active power, $\mathbf{V}_t$ is the bus voltage magnitudes, and $\mathbf{Q}_{t-1}^{\text{PV}}$ is the inverter reactive power outputs at the previous step.

(2) Observation space. Each agent i only has access to local measurements. The local observation is defined as

$$\mathbf{o}_{i,t} = [\mathbf{P}_{\mathcal{N}_{z(i)},t}^L, \mathbf{Q}_{\mathcal{N}_{z(i)},t}^L, \mathbf{P}_{\mathcal{N}_{z(i)},t}^{\text{PV}}, \mathbf{V}_{\mathcal{N}_{z(i)},t}, \mathbf{Q}_{\mathcal{N}_{z(i)},t-1}^{\text{PV}}]^\top \in O_i \quad (22)$$

where $\mathbf{o}_{i,t}$ is the local state information within the region $\mathcal{N}_{z(i)}$ where the agent i is located.

Given that measurement devices in ADNs may be subject to noise and errors, we introduce observation disturbances to better capture the stochastic factors present in real-world operating environments. Isotropic multi-variable Gaussian distribution is added to the observation set:

$$\mathbf{o}_{i,t} = \mathbf{o}_{i,t} + \epsilon_{i,t}, \epsilon_{i,t} \sim \mathbb{N}(0, \sigma_o^2 \mathbf{I}) \quad (23)$$

where σ_o is the measurement noise's standard deviation.

(3) Action space. The action of agent i is the reactive power command of the corresponding PV inverter:

$$a_{i,t} = Q_{i,t}^{\text{PV}} \in A_i \quad (24)$$

(4) Transition function. The transition model $T: S \times A \times S \rightarrow [0, 1]$ characterizes how the environment evolves under the agents' actions. In particular, $T(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t)$ denotes the probability of reaching the next state $\mathbf{s}_{t+1}$ after taking action $\mathbf{a}_t$ in the current state $\mathbf{s}_t$, which is determined by the environment. The transition satisfies the Markov property, meaning that $\mathbf{s}_{t+1}$ depends only on $\mathbf{s}_t$ and $\mathbf{a}_t$. In ADNs, uncertainties in PV generation and load demand can be naturally represented by stochastic state transitions. In this paper, the environment dynamics are simulated via power flow calculations; therefore, the resulting state transitions inherently comply with the power flow equations and their constraints.

(5) Reward function. The instantaneous reward $r \in R$ provides immediate evaluative feedback. It is designed to be consistent with the optimization objective in (2):

$$r = -1/n \times \sum_{i=1}^n C_a |V_i - V_{\text{ref}}| \quad (25)$$

where V_{ref} represents the reference value of node voltage in the ADN.

At each time step, every agent selects an action based on its local observation and receives an immediate reward. The objective of each agent is to learn a policy that maximizes the expected discounted return. Once all agents have taken their actions, the environment transitions to the next state. The above description is a complete Dec-POMDP, which is shown in Figure 6.

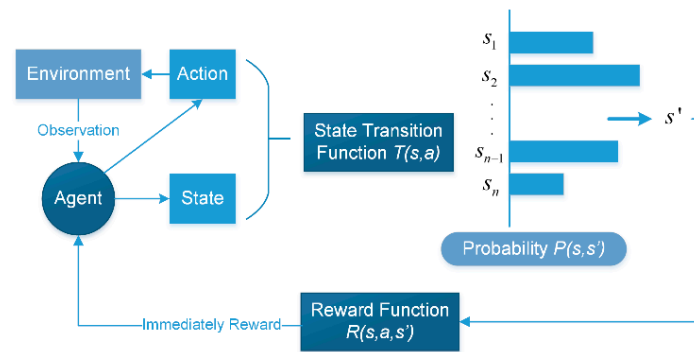


Figure 6. Schematic diagram of Dec-POMDP.

3.5. Stability and Feasibility Analysis

Since the proposed DRL-based policy directly issues reactive power setpoints to inverter controllers, theoretical stability guarantees are essential. We provide a Bounded Input–Bounded Output (BIBO) stability argument based on the physical constraints of the ADN and the formulated action space.

Assumption 1 (stable inner-loop tracking). *Each PV inverter is equipped with a fast-response inner-loop controller that is asymptotically stable. Let $q_i^{\text{cmd}}(t)$ denote the reactive power setpoint issued by the DRL agent and $q_i^{\text{inj}}(t)$ the actual injection. We assume the inner-loop dynamics are sufficiently fast such that $q_i^{\text{inj}}(t) \approx q_i^{\text{cmd}}(t)$ within the dispatch interval.*

Lemma 1 (boundedness of voltage deviation). *Consider the linearized power flow model around a nominal operating point. The relationship between the nodal voltage magnitude vector $\mathbf{v}$ and the reactive power injection vector $\mathbf{q}$ can be expressed via the sensitivity matrix $\mathbf{S}_{\text{vq}}$:*

$$\mathbf{v}(t) \approx \mathbf{v}_0 + \mathbf{S}_{\text{vp}}\mathbf{p}(t) + \mathbf{S}_{\text{vq}}\mathbf{q}(t) \quad (26)$$

where $\mathbf{v}_0$ is the base voltage, and $\mathbf{p}(t)$ is the active power injection (treated as a disturbance) at time t .

Since the MAAC agent's action space is strictly constrained by the activation function and scaled to the physical capacity Q_i^{max} , the reactive power injection is component-wise bounded:

$$|q_i(t)| \leq Q_i^{\text{max}}, \forall i, t \quad (27)$$

Let $\|\cdot\|_{\infty}$ denote the infinity norm. Leveraging the compatibility property of the vector and matrix norm, the deviation of system voltage induced by reactive power control is bounded by the following:

$$\|\Delta\mathbf{v}(t)\|_{\infty} \leq \|\mathbf{S}_{\text{vq}}\|_{\infty} \|\mathbf{q}(t)\|_{\infty} \leq \|\mathbf{S}_{\text{vq}}\|_{\infty} Q^{\text{max}} \quad (28)$$

Proof. Since the sensitivity matrix $\mathbf{S}_{\text{vq}}$ derived from the radial network topology is finite and constant (under the linearization assumption), and the input vector $\mathbf{q}(t)$ is strictly bounded by the hardware limits enforced by the DRL action formulation, the resulting voltage state $\mathbf{v}(t)$ must remain within a bounded region. This confirms that the closed-loop system satisfies BIBO stability. $\square$

Remark 1. The above proposition provides a stability-related boundedness guarantee for quasi-static grid states under enforced action constraints. The proposed control strategy will never generate unbounded control signals that could destabilize the grid physically. The worst-case voltage deviation is mathematically limited by the network topology and the inverter capacities.

4. Case Study

4.1. Simulation Example and Experiment Settings

All simulations are conducted on the modified IEEE 33-bus test system, as shown in Figure 7. PV units are connected at buses 13, 18, 22, 25, 29, and 33, and each PV inverter is rated at 1.25 MW. The nominal voltage level is 12.66 kV, and the allowable voltage variation is limited to $\pm 5\%$. The PV data are obtained from the Belgian grid [34]. All demands are modeled as time-varying constant-PQ loads in the power flow. The active power profiles are taken from the UK public dataset [35] after scaling, and the reactive power demand is constructed assuming a fixed power factor for all buses. In the baseline six-inverter case, the PV penetration is 131%. Accordingly, the voltage limits are set to $V_{\min} = 0.95$ p.u. and $V_{\max} = 1.05$ p.u. The voltage violation cost factor C_{α} is set to 10 in all experiments. The Belgian PV time series is first normalized to per-unit of its annual maximum and then scaled to each inverter by multiplying the inverter's rated active power. Similarly, the UK load profile is normalized and scaled to match the IEEE 33-bus nominal load levels [36]. The processed dataset is available in the Supplementary Materials.

The dataset spans one year. We randomly select 30 days for testing, while the remaining days are used for training. To enhance model generalization and ensure robust control under time-varying and stochastic operating conditions, the initial system states are randomly initialized during training. Gaussian perturbation with standard deviation of 0.1 is added to the dataset after properly scaling to further simulate the fluctuation of the actual PV output and load. The action interval of the agent is 3 min, which means that the real-time voltage control dispatch instruction is delivered every 3 min. The algorithm parameters settings are shown in Table 1. The simulation test is based on the hardware platform with an NVIDIA GeForce 3050 Ti GPU and an AMD Ryzen 7-5800H CPU. The proposed MAAC and all baselines were implemented in Python 3.8 with PyTorch 1.10.2. The power flow-based environment transition was implemented by the power flow solver Pandapower [36]. For each epoch, one day is randomly sampled from the training set; each day contains 480 steps given a 3 min control interval.

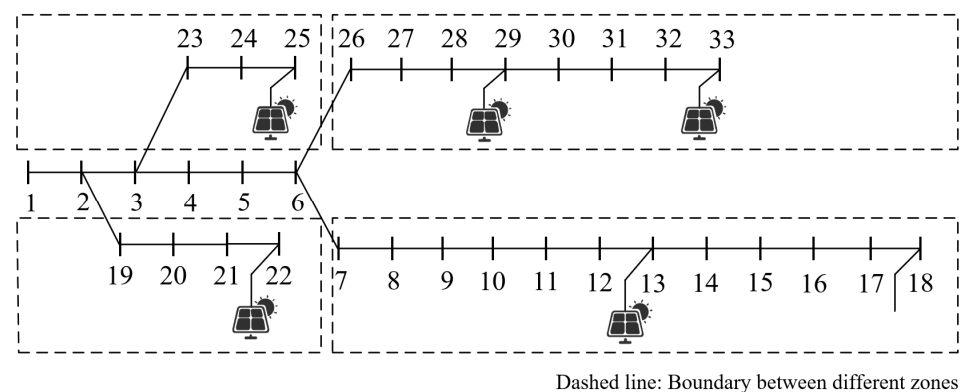


Figure 7. Topology of modified IEEE 33-bus system.

Table 1. Algorithm parameter settings.

Parameters	Values
Batch size B	64

Replay buffer size	800
Discount factor γ	0.99
Step size	480
Actor network's learning rate	0.0001
Critic network's learning rate	0.0001
Measurement noise standard deviation σ_o	0.1
Attention head H	1
Actor/Critic hidden layers	64
Target network update frequency	120
Behavior network update frequency	30
Entropy temperature parameter α	0.001
Soft update factor τ	0.1

4.2. Results and Analysis

To evaluate the effectiveness of MAAC, we compare it with several state-of-the-art multi-agent DRL baselines, including MADDPG, MATD3 [30], and MAPPO [37]. The training convergence curves of all methods are presented in Figure 8. To assess consensus behavior and robustness to random initialization and stochastic sampling, each algorithm is trained in five independent runs with different random seeds. The solid line reports the average performance across the five runs, while the shaded region indicates the minima–maxima error bounds of the five experiments.

We train the agents for 200 epochs, where each epoch corresponds to one day. At the start of each epoch, a day is randomly sampled from the training set to generate the training trajectory. In the early training phase, agents select actions at random to explore the environment and collect experience. As the neural networks are updated iteratively, the learned policy gradually improves, leading to higher cumulative returns. Eventually, the training performance stabilizes and converges within a narrow range.

As shown in Figure 8, the MAAC-based approach exhibits significantly smaller performance oscillations compared with the other benchmark algorithms, indicating more stable learning behavior during training. Benefiting from its enhanced sampling efficiency, MAAC converges more rapidly and reaches a steady policy earlier than the competing methods. Moreover, MAAC achieves the highest cumulative reward at convergence, demonstrating its superior learning effectiveness. These results confirm that the proposed attention-enhanced MADRL framework achieves faster convergence and more effective training performance for the VVC task.

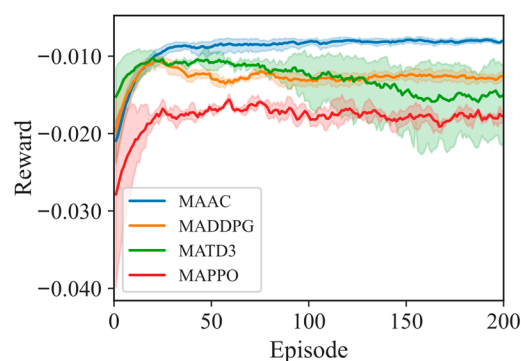


Figure 8. Comparison of training convergence process.

To evaluate the effectiveness of the trained model, the PV generation and load profiles on a representative day with the highest PV output in the test set are selected, as illustrated in Figure 9. The PV output exhibits pronounced temporal variability, with a

peak generation period occurring between 09:00 and 15:00, during which reverse power flow and over-voltage issues are more likely to arise. As shown in the figure, the PV output reaches its maximum at 13:00; therefore, this time instant is chosen for detailed analysis.

In contrast, the system load attains its maximum during nighttime hours when PV generation is unavailable, which may lead to under-voltage conditions. Accordingly, 21:00 is also selected as a representative operating point for comparative analysis.

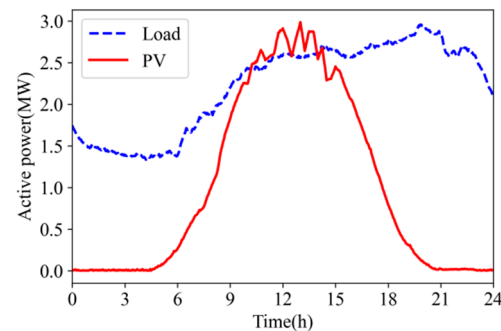


Figure 9. PV output and load curve on typical day.

The boxplot of voltage distribution of each node on a typical day after the optimization of the MAAC voltage control strategy is shown in Figure 10. It can be seen that the voltage profile in the whole day under this strategy is within the safe operating range. The MAAC method achieves a good voltage control effect.

To further compare and analyze the various voltage regulation methods, we choose four representative moments on the typical day. The voltage profiles under different voltage control strategies at different moments are shown in Figures 11–14. The original method means that the PV inverters are not controlled. The droop control adopts the QV control strategy.

It can be seen that the node's voltage fluctuates greatly without any voltage control strategy. It can be obviously observed that the voltage profile of the proposed approach is very close to the reference bus voltage value in four typical moments, which proves that MAAC outperforms other voltage control methods in suppressing voltage fluctuations. At 13:00, due to the large PV output, the load cannot be fully consumed, resulting in reverse power flow and an over-voltage risk problem. At 21:00, the power demand is large; however, the PV inverter cannot provide reactive support at night, resulting in a low-voltage problem. Both droop control and the MADRL strategy can mitigate node voltage fluctuations. However, the MADRL methods achieve better performance in suppressing the voltage violation problem compared with the droop control method, due to the coordinated control of inverters via the framework of CTDE. Moreover, the proposed MAAC further improves the voltage profile thanks to the attention model, which can help the agents learn more useful information in the training process.

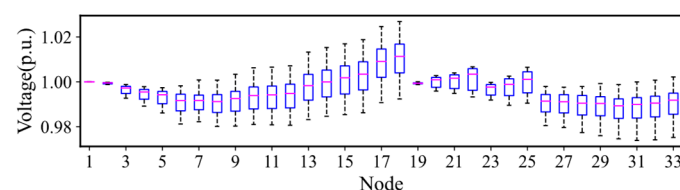


Figure 10. The boxplot of voltage profile under MAAC strategy on typical day.

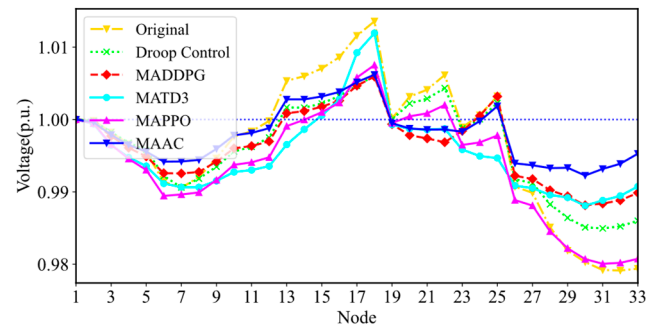


Figure 11. Voltage profile under different control strategies at 09:00.

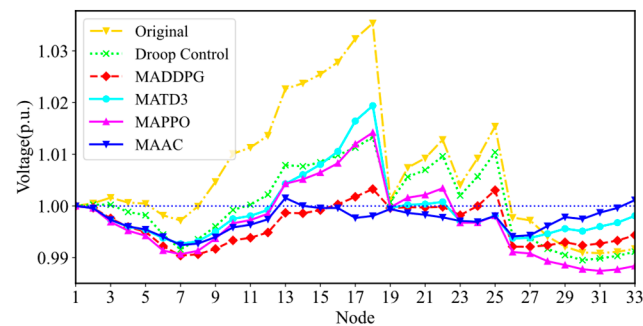


Figure 12. Voltage profile under different control strategies at 13:00.

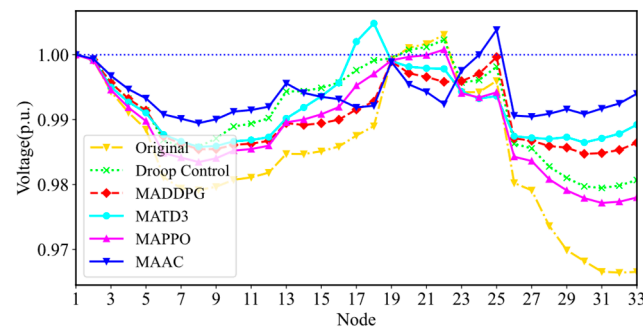


Figure 13. Voltage profile under different control strategies at 17:00.

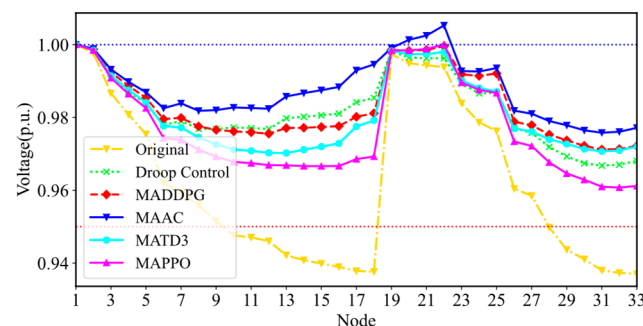


Figure 14. Voltage profile under different control strategies at 21:00.

Table 2 reports the average voltage deviation on the test set. The proposed method achieves the lowest average voltage deviation, indicating its superior capability to accommodate fluctuations of PV outputs and to maintain voltage security and operational stability in the ADN. Under CTDE, the centralized critic with attention is only used in offline training, while online execution only runs the actor networks. Thus, the online inference cost scales as $\mathcal{O}(N \cdot C_{\text{actor}})$, where C_{actor} denotes the forward-pass cost of one actor net-

work. In contrast, the attention-based critic involves pairwise agent interactions during training, resulting in $\mathcal{O}(B \cdot H \cdot N^2 \cdot d)$ complexity per batch, where d is the embedding dimension. The measured average online execution time is reported in Table 2 on the specified hardware platform. The proposed method can meet the real-time requirement of online VVC and has faster calculation speed than other MADRL methods.

Table 2. Voltage deviations of test set.

Method	Average (p.u.)	Execution Time (ms)
Original	0.0206	-
Droop control	0.0117	14.58
MADDPG	0.0112	34.29
MATD3	0.0162	34.77
MAPPO	0.0178	35.21
MAAC	0.0074	33.83

4.3. Proof of Scalability

In order to verify the scalability and generalization of the proposed method, we conducted tests by increasing the number of agents and the PV penetration. We define PV penetration as the ratio of the annual maximum distributed PV generation to the annual maximum regional load, which may occur at different times. The test results are shown in Table 3 below.

Table 3. Scalability results under different numbers of agents and PV penetration levels.

Number of Agents	Node Location of PV	PV Penetration	Average Voltage Deviation (p.u.)
6	13, 18, 22, 25, 29, and 33	131%	0.0074
8	5, 9, 13, 18, 22, 25, 29, and 33	172%	0.0069
10	5, 9, 13, 15, 18, 20, 22, 25, 29, and 33	211%	0.0044
12	4, 7, 10, 13, 15, 18, 20, 22, 25, 29, 31, and 33	251%	0.0038

As the number of agents grows, the MAAC method can still adapt to a large-scale multi-agent system, overcoming the problem of dimension disaster with the help of the CTDE framework. As the PV penetration increases, the total reactive power regulation capability of ADNs also increases. The average voltage deviation decreases due to the coordinated control of inverters. With its powerful learning generalization capability and robustness, the proposed method can adapt to the ADN environment where PV penetration is significantly increased.

4.4. Discussion and Insights

From a power flow perspective, voltage deviations in radial distribution feeders are primarily driven by local injections and a limited set of electrically close buses. Under high PV generation, reverse power flow increases the upstream voltage level and may trigger over-voltage, so effective VVC requires coordinated reactive power absorption at a few influential inverter locations. Under peak-load conditions, under-voltage is more likely and coordinated reactive power injection is needed to support weak-voltage buses. The proposed MAAC policy learns to allocate reactive power across inverters in a way that is consistent with voltage sensitivity. The inverters that are electrically closer to the violated buses contribute more reactive support, while others remain near-zero to avoid unnecessary circulation of reactive power. This explains why MAAC yields smaller voltage devi-

ations and fewer violations in the voltage profiles compared with droop and conventional MADRL methods.

The experimental results indicate that the proposed attention-enhanced CTDE framework achieves faster convergence and fewer voltage violations than the benchmark methods. This performance gain can be attributed to the following factors.

- (1) Better coordination and credit assignment under partial observability. In practical ADNs, each inverter has only local information. Conventional methods may produce myopic actions and struggle to coordinate across multiple inverters. By contrast, our centralized critic leverages an attention mechanism to selectively aggregate the most relevant information from other agents during training, enabling more accurate evaluation of joint actions and facilitating consistent credit assignment. As a result, agents learn cooperative behaviors that reduce system-wide voltage deviations and violations.
- (2) Improved training stability by alleviating multi-agent non-stationarity. Multi-agent environments are inherently non-stationary from the perspective of each individual agent because other agents' policies are simultaneously updated. The centralized critic helps stabilize the learning signal and reduces gradient variance, which explains the smoother training curves and better convergence behavior compared with conventional MADRL baselines.
- (3) Deployment-oriented CTDE design with physically feasible actions. Although the critic uses joint information in centralized training, the learned actors execute in a fully decentralized manner using only local observations, which matches realistic communication constraints. In addition, actions are constrained by inverter capability limits, ensuring that the learned policy remains physically feasible. This CTDE structure enables the method to retain coordination benefits without incurring heavy online computation requirements.
- (4) Scalability and robustness. The method maintains performance as the number of agents and PV penetration increases, indicating that attention-based aggregation can focus on influential interactions and avoid redundant information, thereby improving generalization to more complex operating conditions.

5. Conclusions

This paper presents an inverter-based VVC framework for ADNs using a multi-agent attention-based reinforcement learning approach. By enabling cooperative control among PV inverters, the proposed method is capable of deriving effective real-time voltage regulation strategies. Under the CTDE paradigm, the learned policies can be deployed online with low computational burden, thereby ensuring fast response and practical real-time applicability.

Another important advantage of the proposed method lies in its ability to explicitly account for environmental uncertainty through the Dec-POMDP formulation. By incorporating stochastic state transitions, the MAAC-based controller can better adapt to the inherent randomness of PV generation and load variations in real ADNs. Simulation studies conducted on the IEEE 33-bus test system demonstrate that the proposed method outperforms conventional droop-based control in mitigating voltage violations.

Furthermore, compared with representative MADRL algorithms, the proposed MAAC approach achieves superior voltage regulation performance and faster convergence, which can be attributed to the attention mechanism that selectively captures the most relevant interactions among agents. Scalability tests further confirm the robustness and generalization capability of the proposed method, as it maintains effective voltage control performance under scenarios with a substantially increased number of agents and higher DG penetration levels.

Future work will focus on spatiotemporal attention mechanisms and graph attention networks, enabling more accurate modeling of electrical coupling and information propagation among agents. These extensions are expected to further improve the scalability, robustness, and adaptability of the proposed method under highly dynamic operating conditions and complex network configurations.

Despite the superior performance demonstrated in the high-penetration IEEE 33-bus system, this study focuses primarily on the coordination mechanism and voltage control problem. Future work will extend the objective by incorporating additional operational criteria such as carbon emissions, operating costs and network losses. Such extensions naturally lead to a multi-objective trade-off that can be addressed using the Pareto-based MADRL method. In addition, further scalability tests on larger-scale distribution systems (e.g., 322-bus) and robustness tests against cyber–physical disturbances such as communication failures will be investigated in future work.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/math14050839/s1>, Table S1: The processed active load data; Table S2: The processed PV generation data.

Author Contributions: Conceptualization, W.C. and H.Q.; methodology, W.C. and H.Q.; software, W.C., H.N. and J.L.; validation, W.C., L.L. and J.L.; formal analysis, L.L.; investigation, H.N.; resources, H.N.; data curation, L.L.; writing—original draft preparation, W.C.; writing—review and editing, H.N. and H.Q.; visualization, H.Q.; supervision, H.N.; project administration, L.L.; funding acquisition, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Science and Technology Project of China Southern Power Grid Corporation, grant number GDKJXM20231229.

Data Availability Statement: The data that support the findings of this study are available in the supplementary file.

Conflicts of Interest: Authors Wenwen Chen, Hao Niu, Linbo Liu and Jianglong Lin was employed by the Shaoguan Power Supply Bureau, Guangdong Power Grid Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Yan, R.; Xu, Y. Multi-Objective and Multi-Agent Deep Reinforcement Learning for Real-Time Decentralized Volt/VAR Control of Distribution Networks Considering PV Inverter Lifetime. *IEEE Trans. Power Syst.* **2025**, *40*, 1558–1569. <https://doi.org/10.1109/TPWRS.2024.3452154>.
2. Yan, R.; Xing, Q.; Xu, Y. Multi-Agent Safe Graph Reinforcement Learning for PV Inverters-Based Real-Time Decentralized Volt/Var Control in Zoned Distribution Networks. *IEEE Trans. Smart Grid* **2024**, *15*, 299–311. <https://doi.org/10.1109/TSG.2023.3277087>.
3. Quan, H.; Peng, X.; Liu, H.; Zhou, P.; Wu, Z.; Su, H. Real time voltage optimization control method for distribution networks based on deep reinforcement learning. *Grid Technol.* **2023**, *47*, 2029–2038. <https://doi.org/10.13335/j.1000-3673.pst.2022.1472>.
4. Reinaldo, T.; Lopes, L.A.C.; El-Fouly, T.H.M. Coordinated active power curtailment of grid connected PV inverters for over-voltage prevention. *IEEE Trans. Sustain. Energy* **2011**, *2*, 139–147. <https://doi.org/10.1109/TSTE.2010.2098483>.
5. Zeraati, M.; Esmail, M.; Golshan, H.; Guerrero, J.M. Distributed control of battery energy storage systems for voltage regulation in distribution networks with high PV penetration. *IEEE Trans. Smart Grid* **2018**, *9*, 3582–3593. <https://doi.org/10.1109/TSG.2016.2636217>.
6. Yang, N.-C.; Zhong, P.-Y. Day-Ahead Scheduling of On-Load Tap Changer Transformer and Switched Capacitors by Multi-Pareto Optimality. *Mathematics* **2022**, *10*, 2969. <https://doi.org/10.3390/math10162969>.
7. Liu, H.; Zhang, C.; Chai, Q.; Meng, K.; Guo, Q.; Dong, Z.Y. Robust regional coordination of inverter-based volt/var control via multi-agent deep reinforcement learning. *IEEE Trans. Smart Grid* **2021**, *12*, 5420–5433. <https://doi.org/10.1109/TSG.2021.3104139>.

8. Mahmoud, K.; Lehtonen, M. Comprehensive analytical expressions for assessing and maximizing technical benefits of photo-voltaics to distribution systems. *IEEE Trans. Smart Grid* **2021**, *12*, 4938–4949. <https://doi.org/10.1109/TSG.2021.3097508>.
9. Weckx, S.; Driesen, J. Optimal local reactive power control by PV inverters. *IEEE Trans. Sustain. Energy* **2016**, *7*, 1624–1633. <https://doi.org/10.1109/TSTE.2016.2572162>.
10. Singhal, A.; Ajarapu, V.; Fuller, J.; Hansen, J. Real-Time local volt/var control under external disturbances with high PV penetration. *IEEE Trans. Smart Grid* **2019**, *10*, 3849–3859. <https://doi.org/10.1109/TSG.2018.2840965>.
11. Du, Z.; Lin, X.; Zhong, G.; Liu, H.; Zhao, W. Data-Driven Voltage Control Method of Active Distribution Networks Based on Koopman Operator Theory. *Mathematics* **2024**, *12*, 3944. <https://doi.org/10.3390/math12243944>.
12. Dou, X.; Li, C.; Niu, P.; Sun, D.; Zhang, Q.; Dou, Z. An Optimal Scheduling Method for Power Grids in Extreme Scenarios Based on an Information-Fusion MADDPG Algorithm. *Mathematics* **2025**, *13*, 3168. <https://doi.org/10.3390/math13193168>.
13. Cao, D.; Hu, W.; Zhao, J.; Zhang, G.; Zhang, B.; Liu, Z.; Chen, Z.; Blaabjerg, F. Reinforcement Learning and Its Applications in Modern Power and Energy Systems: A Review. *J. Mod. Power Syst. Clean Energy* **2020**, *8*, 1029–1042. <https://doi.org/10.35833/MPCE.2020.000552>.
14. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; Petersen, S. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533.
15. Zhang, Z.; Zhang, D.; Qiu, R.C. Deep reinforcement learning for power system applications: An overview. *CSEE J. Power Energy Syst.* **2020**, *6*, 213–225. <https://doi.org/10.17775/CSEEJPES.2019.00920>.
16. Pan, Z.; Quan, H.; Lin, X.; Zhou, H.; Yu, M.; Kang, H.; Chen, L. Reinforcement Learning Based Reactive Power Real-Time Dispatch Optimization in Distribution Networks. In *Proceedings of the 2023 5th International Conference on Power and Energy Technology (ICPET)*, Tianjin, China, 27–30 July 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 642–647. <https://doi.org/10.1109/ICPET59380.2023.10367680>.
17. Gao, X.; Xin, H.; Liu, J.; Li, T. Event-Driven Prescribed Optimal Disturbance Rejection for Dynamic Positioning of Ships via Reinforcement Learning. *IEEE Trans. Autom. Sci. Eng.* **2026**, early access. <https://doi.org/10.1109/TASE.2026.3660669>.
18. Zhang, Y.; Wang, X.; Wang, J.; Zhang, Y. Deep reinforcement learning based volt-var optimization in smart distribution systems. *IEEE Trans. Smart Grid* **2021**, *12*, 361–371. <https://doi.org/10.1109/TSG.2020.3010130>.
19. Zhang, J.; Li, Y.; Wu, Z.; Rong, C.; Wang, T.; Zhang, Z.; Zhou, S. Deep-Reinforcement-Learning-Based two-timescale voltage control for distribution systems. *Energies* **2021**, *14*, 3540. <https://doi.org/10.3390/en14123540>.
20. Hossain, R.; Gautam, M.; Thapa, J.; Livani, H.; Benidris, M. Deep reinforcement learning assisted co-optimization of Volt-VAR grid service in distribution networks. *Sustain. Energy Grids Netw.* **2023**, *35*, 101086. <https://doi.org/10.1016/j.segan.2023.101086>.
21. Sun, X.; Qiu, J. Two-Stage volt/var control in active distribution networks with multi-agent deep reinforcement learning method. *IEEE Trans. Smart Grid* **2021**, *12*, 2903–2912. <https://doi.org/10.1109/TSG.2021.3052998>.
22. Zuo, J.; Ai, Q.; Wang, W.; Tao, W. Day-Ahead Economic Dispatch Strategy for Distribution Networks with Multi-Class Distributed Resources Based on Improved MAPPO Algorithm. *Mathematics* **2024**, *12*, 3993. <https://doi.org/10.3390/math12243993>.
23. Cao, D.; Zhao, J.; Hu, W.; Ding, F.; Huang, Q.; Chen, Z. Attention Enabled Multi-Agent DRL for Decentralized Volt-VAR Control of Active Distribution System Using PV Inverters and SVCs. *IEEE Trans. Sustain. Energy* **2021**, *12*, 1582–1592.
24. Chen, Y.; Liu, Y.; Zhao, J.; Qiu, G.; Yin, H.; Li, Z. Physical-assisted multi-agent graph reinforcement learning enabled fast voltage regulation for PV-rich active distribution network. *Appl. Energy* **2023**, *351*, 121743. <https://doi.org/10.1016/j.apenergy.2023.121743>.
25. Luo, F.; Wang, S.; Lv, Y.; Mu, R.; Fo, J.; Zhang, T.; Xu, J.; Wang, C. Domain knowledge-enhanced graph reinforcement learning method for Volt/Var control in distribution networks. *Appl. Energy* **2025**, *398*, 126409. <https://doi.org/10.1016/j.apenergy.2025.126409>.
26. Kabiri, R.; Holmes, G.; McGrath, B.P.; Meegahapola, L.G. LV grid voltage regulation using transformer electronic tap changing, with PV inverter reactive power injection. *IEEE J. Emerg. Sel. Top. Power Electron.* **2015**, *3*, 1182–1192. <https://doi.org/10.1109/JESTPE.2015.2443839>.
27. Mahmoud, K.; Yorino, N.; Ahmed, A. Optimal distributed generation allocation in distribution systems for loss minimization. *IEEE Trans. Power Syst.* **2016**, *31*, 960–969. <https://doi.org/10.1109/TPWRS.2015.2418333>.
28. Li, Y.; Hu, X.; Zhuang, Y.; Gao, Z.; Zhang, P.; El-Sheimy, N. Deep reinforcement learning (DRL): Another perspective for unsupervised wireless localization. *IEEE Internet Things J.* **2020**, *7*, 6279–6287. <https://doi.org/10.1109/JIOT.2019.2957778>.
29. Nguyen, T.T.; Nguyen, N.D.; Nahavandi, S. Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications. *IEEE Trans. Cybern.* **2020**, *50*, 3826–3839. <https://doi.org/10.1109/TCYB.2020.2977374>.

30. Fujimoto, S.; Van Hoof, H.; Meger, D. Addressing function approximation error in actor-critic methods. In *Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018*; PMLR: Cambridge, MA, USA, 2018; pp. 1587–1596.
31. Lowe, R.; Wu, Y.; Tamar, A.; Harb, J.; Abbeel, P.; Mordatch, I. Multiagent actor-critic for mixed cooperative-competitive environments. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017*; pp. 6379–6390.
32. Shariq, I.; Sha, F. Actor-attention-critic for multi-agent reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019*; PMLR: Cambridge, MA, USA, 2019; pp. 2961–2970.
33. Haarnoja, T.; Zhou, A.; Abbeel, P.; Levine, S. Soft actor-critic: Offpolicy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018*; PMLR: Cambridge, MA, USA, 2018; pp. 1861–1870.
34. Elia Group. Transparency on Grid Data: Solar Power-Generation. Available online: <https://www.elia.be/en/grid-data/generation-data/solar-power-generation> (accessed on 18 January 2025).
35. UK Power Networks. SmartMeter Energy Consumption Data in London Households. Available online: <https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households> (accessed on 18 January 2025).
36. Thurner, L.; Scheidler, A.; Schäfer, F.; Menke, J.H.; Dollichon, J.; Meier, F.; Meinecke, S.; Braun, M. Pandapower-an open-source pythontool for convenient modeling, analysis, and optimization of electric power systems. *IEEE Trans. Power Syst.* **2018**, *33*, 6510–6521. <https://doi.org/10.1109/TPWRS.2018.2829021>.
37. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal policy optimization algorithms. *arXiv* **2017**, arXiv:1707.06347. <https://doi.org/10.48550/arXiv.1707.06347>.

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