Flow-Based Offline Reinforcement Learning for Voltage Regulation in Distribution Networks

Liyu Shan [1] **Yongli Zhu** [1]

Abstract

This paper investigates pure data-driven active voltage control in distribution networks via offline reinforcement learning (RL) to minimize the risks of online interactions. To overcome the limited policy expressivity of existing algorithms on low-quality, randomly collected datasets, we propose an improved Flow Q-Learning (improved FQL) approach featuring a Flow-Guided Base Action and Fine-Tuning Perturbation framework. This architecture utilizes a flow-matching model to accurately capture the distribution of safe behaviors, while a residual actor applies bounded perturbations to preserve physical safety while maximizing Q-values. Furthermore, Boltzmann annealing and Zero-Noise Initialization mechanisms are introduced to enhance convergence and execution stability. Simulations on IEEE 33-bus and 123-bus systems demonstrate that our approach outperforms baseline offline RL algorithms in mitigating voltage deviations and violations, despite relying entirely on randomly collected offline experience.

1. Introduction

With the large-scale integration of distributed energy resources, such as photovoltaic (PV) systems, the intermittency and uncertainty of their power generation have introduced severe voltage fluctuations and violation challenges to modern distribution networks. Since conventional active voltage control methods (such as droop control or optimization-based optimal power flow calculations) often exhibit limitations in global coordination capabilities or real-time computational efficiency, reinforcement learning (RL) has shown great potential in the voltage control domain as a model-free, data-driven approach (Wang et al., 2022). However, standard online RL requires the agent to perform extensive trial-and-error exploration in real environments, which can easily trigger irreversible equipment damage or widespread voltage violations in safety-critical power systems.

Therefore, offline reinforcement learning (Offline RL) has become a highly promising alternative, allowing agents to learn control policies entirely from pre-collected historical operational data, thus effectively decoupling the training stage from physical grid interactions. Existing studies have successfully applied algorithms like Batch-Constrained Q-learning (BCQ) and Conservative Q-learning (CQL) to voltage regulation (Yang & Zhu, 2025; Mao et al., 2025), ensuring the safe operation of the grid by constraining the action space or punishing out-of-distribution actions. Although existing offline RL methods have alleviated safety concerns, constrained by strict grid safety requirements, offline datasets can typically only be constructed through highly conservative random sampling, resulting in data is inherently noisy and lacking clear expert patterns. To overcome the limited expressivity of traditional Gaussian policies on such complex data, this paper leverages flow-based generative modeling to better capture complex action distributions for offline policy learning.

In recent years, generative models such as diffusion models and flow matching have been introduced into the RL domain (Park et al., 2025), enhancing modeling capabilities for complex action distributions. However, directly applying generative models to power grid control still poses potential risks: maximizing Q-values via backpropagation through time (BPTT) triggers training instability, and the lack of hard physical constraints can easily lead to violations during inference. To effectively resolve these inherent defects, this paper proposes a Flow-Guided Base Action and Fine-Tuning Perturbation framework. This framework first utilizes multi-step ODE integration to seek a highly reliable safe base action within the data manifold, and subsequently applies a bounded fine-tuning perturbation to it via a residual actor network, providing a robust methodological framework for safe active voltage control.

Based on this, this paper aims to introduce the improved FQL framework to the domain of active voltage regulation,

[1]School of System Science and Engineering, Sun-Yat Sen University, Guangzhou, China. Correspondence to: Yongli Zhu <yzhu16@alum.utk.edu>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

adapting it to this specific physical problem to effectively extract control knowledge from low-quality, randomly explored datasets. The main contributions of this paper are as follows:

- **Improved FQL-based active voltage control framework:** We introduce the flow-matching concept of the improved FQL algorithm into distribution network voltage regulation and design a Flow-Guided Base Action and Fine-Tuning Perturbation framework. This proposes a pure data-driven offline voltage control scheme that meets the real-time and strict safety requirements of the power grid.
- **Dual algorithm improvements for the voltage control task:** Addressing the characteristics of continuous reactive power regulation and data noise issues, we introduce Boltzmann annealing sampling to balance exploration in early stages with stable convergence in later stages, and devise a Zero-Noise Initialization strategy for optimal safe mean action extraction.
- **Comprehensive empirical evaluation:** Extensive validations on the IEEE 33-bus and 123-bus systems demonstrate that, even when relying solely on conservative randomly collected datasets, the proposed improved FQL approach outperforms existing offline RL baselines (e.g., BCQ and CQL) in mitigating voltage deviations and reducing violation rates.

The rest of the paper is structured as follows: Section 2 formulates the physical distribution network and details the offline RL control framework. Section 3 analyzes the FQL voltage control algorithm with the proposed dual improvements. Section 4 provides the experimental setup and case study results, and Section 5 concludes the paper.

2. Problem Formulation and Environment Setup

This section elaborates on the physical foundation and mathematical modeling of active voltage control in distribution networks, analyzes the Reinforcement Learning (RL) interaction environment and the design of its key components, and introduces the generation mechanism of the offline dataset used for model training.

2.1. Problem Formulation for Active Voltage Control

Active voltage control aims to maintain the voltage stability of the entire distribution network by dynamically regulating the reactive power outputs of controllable PV inverters. In each control period t, the control system needs to maintain the voltage magnitude of all buses within the safe operating limits, while the reactive power output of the inverters is limited by the remaining available capacity of their apparent power:

$$V_{\min} \leq v_i(t) \leq V_{\max}, \quad \forall i \in \mathcal{V} \tag{1}$$

$$|q_i^{\text{PV}}(t)| \leq \sqrt{(s_i^{\max})^2 - (p_i^{\text{PV}}(t))^2}, \quad \forall i \in \mathcal{V}_{\text{PV}} \tag{2}$$

where $v_i(t)$ is the voltage magnitude at bus i; $q_i^{\text{PV}}(t)$ and $p_i^{\text{PV}}(t)$ represent the reactive and active power outputs of the PV inverter at bus i, respectively; and $s_i^{\max}$ is the rated apparent power of the inverter.

Given the highly nonlinear power flow coupling characteristics, we abstract the voltage regulation task as a Constrained Markov Decision Process (CMDP). Since this paper adopts a pure data-driven offline RL paradigm, the agent's policy learning strictly depends on pre-collected historical trajectory data denoted as dataset $\mathcal{D}$:

$$\mathcal{D} = \{(s_i, a_i, r_i, s_{i+1})\}_{i=1}^{N} \tag{3}$$

2.2. Designs of State, Action, and Reward

To enable offline RL algorithms to effectively capture the operational characteristics of the distribution network, we carefully customize the key elements within the CMDP:

2.2.1. State

The system state s_t must adequately represent the current operational conditions and potential voltage violation risks. We define the state vector as a collection of critical electrical variables across the network at the current moment:

$$s_t = \{p_t^L, q_t^L, p_t^{\text{PV}}, q_t^{\text{PV}}, v_t\} \tag{4}$$

where (p_t^L, q_t^L) and $(p_t^{\text{PV}}, q_t^{\text{PV}})$ represent the load demands and PV outputs respectively, and v_t denotes the voltage magnitudes of all buses.

2.2.2. Action

This action is equivalent to the adjustment of the reactive power regulation ratio assigned to each controllable PV inverter. The model directly outputs a continuous ratio coefficient $a_{k,t} \in [-c, c]$ (where c is the maximum allowable control ratio), which is then linearly mapped to a physically executable reactive power command:

$$q_k^{\text{PV}}(t) = a_{k,t} \cdot \sqrt{(s_k^{\max})^2 - (p_k^{\text{PV}}(t))^2} \tag{5}$$

2.2.3. Reward

The reward function guides the agent in striking a balance between suppressing voltage violations and minimizing reactive power loss:

$$r_t = -\frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \alpha_1 l(v_{i,t}) - \alpha_2 \sum_{k \in \mathcal{V}_{\text{PV}}} |q_{k,t}^{\text{PV}}| \tag{6}$$

where $|\mathcal{V}|$ is the total quantity of buses, α_1 and α_2 are weighting factors, and $l(\cdot)$ represents the Voltage Barrier Function with linear L1 penalties.

2.2.4. Objective Function

The ultimate optimization objective is to find an optimal control policy π that maximizes the cumulative discounted reward:

$$J(\pi) = \max_{\pi} \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right] \tag{7}$$

where $\gamma \in (0, 1)$ is the discount factor.

2.3. Offline Data Collection

To strictly avoid the risk of irreversible equipment damage, we employ a Conservative Random Sampling mechanism to construct the offline training dataset $\mathcal{D}$. Within safe thresholds, bounded random control perturbations are applied to the PV inverters, following a strictly truncated uniform distribution:

$$a_i \sim \mathcal{U}[-\gamma, \gamma] \tag{8}$$

where γ is the safety boundary factor for control perturbations.

3. Improved Flow-Based Q-Learning for Active Voltage Control

This section details the improved Flow-Based Q-Learning algorithm for solving the offline voltage control problem in distribution networks. First, we explain how this method achieves safe action inference through a "base action + fine-tuning" architecture while avoiding the uncontrollable risks associated with one-step models. Subsequently, addressing the specific challenges of the voltage control task, we propose two algorithmic improvements aimed at enhancing the execution stability and experience utilization efficiency of the policy.

3.1. The Flow-Based Fine-Tuning Architecture

Considering the sensitivity of distribution networks to action violations, this paper proposes a conservative architecture combining probability flow matching with residual fine-tuning:

3.1.1. Base Action Generation via Flow Matching

A continuous-time parameterized flow network $v_\theta(t, s, x)$ is established to capture the mapping vector field (Stoica et al., 2025) from standard normal distribution noise $x_0 \sim \mathcal{N}(0, I)$ to actions $x_1 = a \sim \mathcal{D}$ in the real offline dataset. The optimization objective is to minimize the flow-matching loss:

$$\mathcal{L}_{\text{Flow}}(\theta) = \mathbb{E}_{s,a\sim\mathcal{D},x_0\sim\mathcal{N}(0,I),t\sim\mathcal{U}} \left\|v_\theta(t, s, x_t) - (a - x_0)\right\|_2^2 \tag{9}$$

where t is the time step sampled from a uniform distribution $\mathcal{U}$, and $x_t = (1-t)x_0 + ta$ is the linearly interpolated state. After training, based on an ordinary differential equation (ODE) solver, this flow model can generate a base action a_{base}^{flow} representing the high-density safe action patterns in the dataset, serving as the physical boundary constraint for subsequent optimization.

3.1.2. Critic Network with CQL Penalty

To further suppress the overestimation of generated actions in out-of-distribution (OOD) regions, the optimization objective of the critic not only includes the standard temporal difference (TD) error but also introduces a Conservative Q-Learning (CQL) penalty term. Given a mini-batch of transition tuples (s, a, r, s'), the optimization objective of the Critic is:

$$\mathcal{L}_Q(\phi_i) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}[(Q_{\phi_i}(s, a) - y)^2] + \alpha_{\text{cql}}\mathcal{L}_{\text{cql}}(\phi_i) \tag{10}$$

where $y = r + \gamma \max_{a'} \min_{j=1,2} Q_{\phi'_j}(s', a')$ is the target Q-value.

3.1.3. Bounded Residual Fine-Tuning Actor

To achieve Q-value maximization while ensuring physical safety, this paper designs a residual actor network $\pi_\omega(s, a_{base}^{flow})$ parameterized by ω. Instead of generating actions from scratch, this network uses the base action a_{base}^{flow} output by the flow model as an anchor and outputs a fine-tuning perturbation Δa strictly truncated by a maximum threshold a_{max}. The final executed action is defined as:

$$a^\pi = \text{Clip}(a_{\text{base}}^{\text{flow}} + \pi_\omega(s, a_{\text{base}}^{\text{flow}}), -a_{\text{max}}, a_{\text{max}}) \tag{11}$$

The function of the actor aims to balance the Q-value maximization of the Deterministic Policy Gradient (DPG) and the Behavioral Cloning (BC) penalty:

$$\mathcal{L}_{\text{actor}}(\omega) = \mathbb{E}_{s\sim\mathcal{D}}\left[-Q(s, a^\pi) + \alpha_{\text{bc}}\left\|a_{\text{base}}^{\text{flow}} - a^\pi\right\|_2^2\right] \tag{12}$$

The latter BC regularization term α_{bc} forces the final output of the actor to not deviate from the safe action manifold defined by the flow model.

3.2. Proposed Improvements for FQL

Addressing the issue of massive low-quality noise in the conservative random datasets, this paper proposes two key algorithmic improvements:

3.2.1. Global Energy-based Annealing Boltzmann Sampling

To prioritize the utilization of high-quality transitions, a global Boltzmann annealing mechanism is introduced to the experience replay buffer. The energy mapping is calculated as $E_i = \min(R^{(i)}_{\text{norm}}, \rho) - \rho$, where $R^{(i)}_{\text{norm}}$ is the normalized global rank of the immediate reward of transition i among all transitions in the replay buffer. The sampling temperature τ decays according to a sine curve:

$$\tau(t) = \tau_{\text{end}} + (\tau_{\text{start}} - \tau_{\text{end}}) \times \left[1 - \sin\left(\frac{\pi}{2} \min\left(1, \frac{t}{T_{\text{anneal}}}\right)\right)\right] \tag{13}$$

The sampling probability follows the Boltzmann distribution $P(i) \propto \exp(E_i/\tau(t))$. As τ decays, the mini-batch distribution gradually shifts from broader coverage toward stronger emphasis on higher-reward transitions, thereby improving training stability and data efficiency.

3.2.2. Deterministic Zero-Noise Initialization

Traditional offline RL generators typically sample randomly from Gaussian noise, which has the potential to trigger voltage violation risks in power grid control. Based on the deterministic bijective property of the Probability Flow ODE (Song et al., 2021), we eliminate the initial noise (i.e., setting $z(0) = 0$) during the base action generation phase (Liu et al., 2022). The integration path is expressed as:

$$a^{\text{flow}}_{\text{base}} = \int_0^1 v_\theta(s, z, t)dt, \quad \text{with } z(0) = 0 \tag{14}$$

This mechanism allows the ODE integration to slide smoothly along the mean trajectory, effectively extracting the safest mean action in the dataset and improving execution stability.

4. Experiments and Results

This section evaluates the improved Flow Q-Learning (improved FQL) algorithm on the standard IEEE 33-bus (Baran & Wu, 1989) and 123-bus (IEEE PES Distribution System Analysis Subcommittee) distribution systems. By comparing the proposed approach with the physical baseline without intervention (No Control) and mainstream offline reinforcement learning (RL) algorithms (BCQ, CQL), we verify the voltage regulation superiority and execution stability of the Flow-Guided Base Action and Fine-Tuning Perturbation framework and its dual improvements under low-quality random data.

4.1. Experimental Setup

The simulation experiments are conducted on the IEEE 33-bus system (integrated with 6 PV units) and the 123-bus system (integrated with 15 PV units). Both the training and testing datasets are generated according to the conservative random sampling strategy detailed in Section 2.3. The evaluation baselines include the No Control physical benchmark, as well as the BCQ and CQL algorithms.

Due to space constraints, Day 1 and Day 300 are specifically selected to concisely demonstrate the algorithm's long-term consistent generalization at both boundary extremes (initial and terminal phases). All algorithm simulations and model training are executed on a computing platform equipped with an Intel Core i5-14600KF CPU, an NVIDIA GeForce RTX 5070 (12GB) GPU, and PyTorch 2.12+cu128.

4.2. Case Study on IEEE 33-bus Distribution System

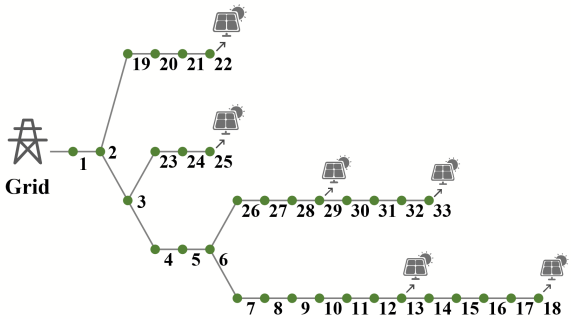

Figure 1. IEEE 33-bus radial distribution system topology.

This section evaluates the performance of the proposed improved FQL algorithm on the IEEE 33-bus radial distribution system, as shown in Figure 1. The test system contains 33 buses with distributed photovoltaic (PV) units integrated at six nodes, representing a typical low-voltage distribution network with high renewable penetration. We compare our method against three baselines: No Control, Batch-Constrained Q-learning (BCQ), and Conservative Q-learning (CQL), using real-world operational scenarios from Day 1 and Day 300.

As summarized in Table 1, our improved FQL outperforms all baselines across all metrics. On Day 1, it achieves the lowest average voltage deviation (0.0076 p.u.) with 0% over/under-voltage rates, while maintaining a much lower reactive power loss (0.1859 p.u.) than CQL (0.4040 p.u.). On the more challenging Day 300 scenario, it still limits the under-voltage rate to 2.5369%—which significantly outperforms BCQ (6.0986%) and CQL (3.8590%)—while achieving the highest total reward of -150.5790.

The spatial voltage distribution across all buses is visualized in Figure 2. Compared to the severe voltage drop at the feeder ends under No Control and BCQ, the improved FQL maintains voltages near 1.0 p.u. across all nodes, eliminating both over-voltage peaks and under-voltage troughs.

Table 1. Comparison on the performance of different control methods on the 33-bus system (Day 1 & Day 300)

DAYS	METRIC	NO CONTROL	BCQ	CQL	FQL (OURS)
DAY 1	AVG V DEV (P.U.)	0.0175	0.0101	**0.0075**	0.0076
	OVER-UPPER RATE (%)	0.0000	0.0000	0.1645	**0.0000**
	UNDER-LOWER RATE (%)	14.3544	0.2341	0.0000	**0.0000**
	MEAN Q LOSS (P.U.)	0.0000	0.2926	0.4040	**0.1859**
	TOTAL REWARD	-276.3267	-168.5246	-130.0246	**-124.7584**
DAY 300	AVG V DEV (P.U.)	0.0187	0.0132	0.0108	**0.0091**
	OVER-UPPER RATE (%)	0.0886	0.0000	0.0822	**0.0000**
	UNDER-LOWER RATE (%)	25.6658	6.0986	3.8590	**2.5369**
	MEAN Q LOSS (P.U.)	**0.0000**	0.2916	0.3606	0.2103
	TOTAL REWARD	-295.7298	-216.8556	-181.6233	**-150.5790**

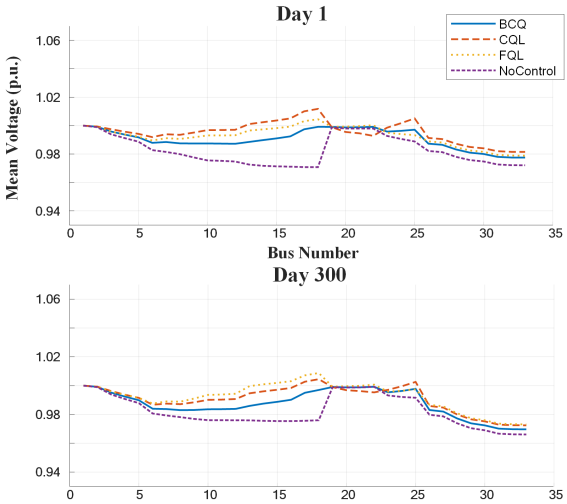

Figure 2. Spatial voltage profile of system buses under diverse control strategies on the IEEE 33-bus system.

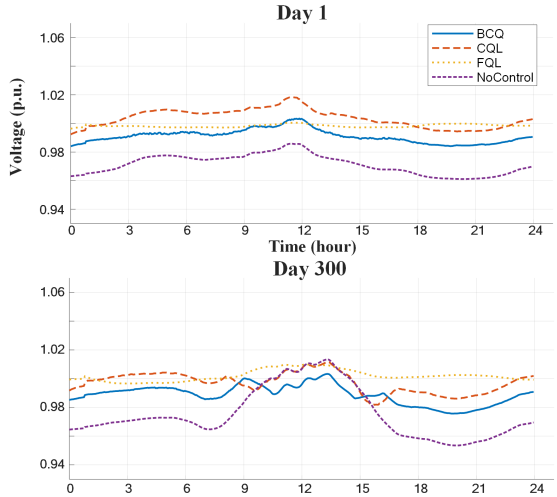

Figure 3. Temporal voltage profile of Bus 15 under diverse control strategies on the IEEE 33-bus system.

For the temporal performance, the 24-hour voltage profile of Bus 15 (a representative node) is shown in Figure 3. The improved FQL effectively suppresses midday PV-induced over-voltage and evening load-induced under-voltage, keeping the voltage within the safe operating range throughout the day.

Finally, the global spatio-temporal voltage heatmaps in Figure 4 confirm the superior performance of our method. The optimal safe voltage strictly centers at the 1.0 p.u. intermediate color band (yellow-green). Unlike the uneven color distribution (indicating frequent violations) of the baselines, the improved FQL exhibits a highly uniform color tone, demonstrating consistent and reliable voltage regulation across all buses and time steps.

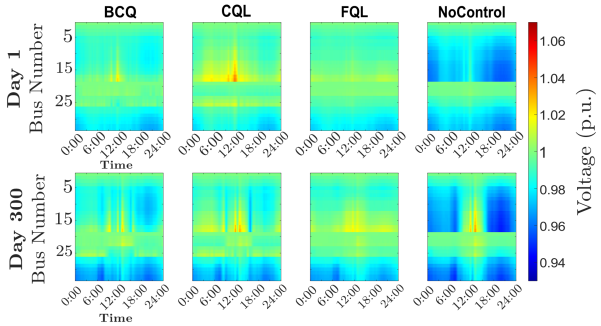

Figure 4. Voltage distribution heatmaps under diverse control strategies on IEEE 33-bus system.

4.3. Case Study on IEEE 123-bus Distribution System

To validate the scalability of the algorithm in a larger and more complex distribution network, this section conducts tests on the IEEE 123-bus system. The system incorporates 123 buses with 15 uniformly distributed PV units.

Table 2 presents the quantitative performance of each method in a large-scale system. The No Control baseline faces severe voltage violations (an 89.03% under-voltage rate on Day 1). On Day 1, the improved FQL strictly limits both over- and under-voltage violation rates to 0%, maintaining a low average voltage deviation (0.0098 p.u.). Although CQL achieves a slightly higher total reward, it incurs minor voltage violations, and its reactive power loss (0.5102 p.u.) is nearly twice that of the improved FQL (0.2742 p.u.). On Day 300, the improved FQL still maintains a 0% under-voltage rate and low reactive power consump-

Table 2. Comparison on the performance of different control methods on the 123-bus system (Day 1 & Day 300)

DAYS	METRIC	NO CONTROL	BCQ	CQL	FQL (OURS)
DAY 1	AVG V DEV (P.U.)	0.0682	0.0642	**0.0078**	0.0098
	OVER-UPPER RATE (%)	0.0000	0.0000	0.2461	**0.0000**
	UNDER-LOWER RATE (%)	89.0337	85.0077	0.1443	**0.0000**
	MEAN Q LOSS (P.U.)	0.0000	0.0550	0.5102	**0.2742**
	TOTAL REWARD	-4.0193E+03	-3.7869E+03	**-498.5075**	-595.9990
DAY 300	AVG V DEV (P.U.)	0.0598	0.0508	**0.0077**	0.0110
	OVER-UPPER RATE (%)	**0.0000**	**0.0000**	0.5397	0.3276
	UNDER-LOWER RATE (%)	76.2106	71.4446	0.0221	**0.0000**
	MEAN Q LOSS (P.U.)	**0.0000**	0.0447	0.5261	0.2483
	TOTAL REWARD	-3.5246E+03	-2.9960E+03	**-493.9968**	-666.3364

tion. This demonstrates that the improved FQL achieves a better balance between safety and economy in large-scale networks.

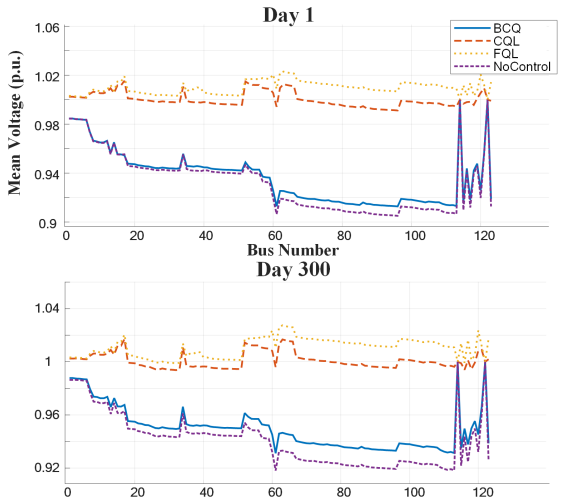

Figure 5. Spatial voltage profile of system buses under diverse control strategies on the IEEE 123-bus system

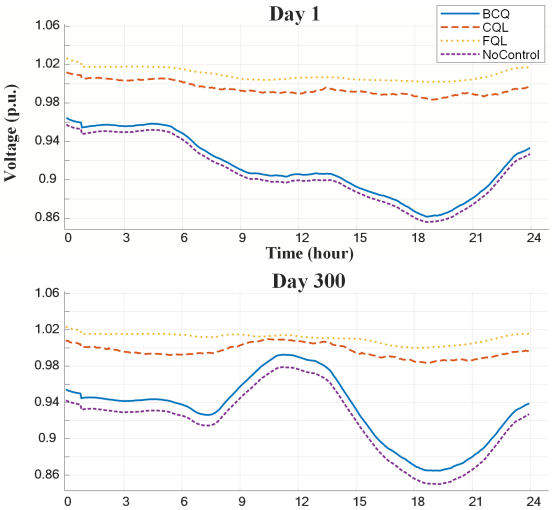

Figure 6. Temporal voltage profile of Bus 111 under diverse control strategies on the IEEE 123-bus system.

Figure 5 compares the spatial voltage distribution. The No Control and BCQ methods suffer from sharp voltage drops in the middle and end sections of the feeder. The improved FQL effectively overcomes this issue, smoothly maintaining the average voltage of all buses around 1.0 p.u., exhibiting stable spatial regulation comparable to CQL.

Figure 6 illustrates the dynamic voltage profile at Bus 111, a highly vulnerable node, over 24 hours. No Control and BCQ experience severe under-voltage (dropping to around 0.86 p.u.) during the evening load peak. Through all-day reactive power compensation, the improved FQL successfully confines the dynamic voltage fluctuations within the safe tolerance band.

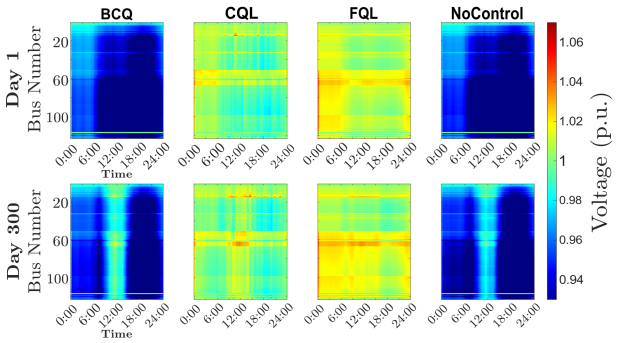

Figure 7. Voltage distribution heatmaps under diverse control strategies on the IEEE 123-bus system.

The global spatio-temporal heatmaps in Figure 7 visually reflect the overall control effects. No Control and BCQ exhibit large, deep blue areas indicating severe under-voltage. In contrast, the heatmap of the improved FQL presents a relatively uniform yellow-green tone, eliminating the under-voltage bands across spatio-temporal domains. This further verifies the scalability and global voltage control capability of the proposed method in complex distribution networks.

5. Conclusion

In conclusion, this study demonstrates the effectiveness of flow-based offline reinforcement learning for active voltage control in distribution networks. The proposed improved FQL framework combines a Flow-Guided Base Action and

Fine-Tuning Perturbation architecture with Boltzmann annealing and deterministic zero-noise initialization to achieve stable and safe policy learning from conservative random datasets. Experimental results on IEEE 33-bus and 123-bus systems show that the proposed method outperforms existing offline RL baselines such as BCQ and CQL in reducing voltage deviations and mitigating voltage violations. These findings indicate that flow-based offline RL is a promising approach for safe decision-making in power systems where online interaction is costly or infeasible. Future work will explore more complex grid scenarios and offline-to-online adaptation strategies.

References

Baran, M. and Wu, F. Network reconfiguration in distribution systems for loss reduction and load balancing. *IEEE Transactions on Power Delivery*, 4(2):1401–1407, 1989. doi: 10.1109/61.25627.

IEEE PES Distribution System Analysis Subcommittee. Ieee distribution test feeders. https://cmte.ieee.org/pes-testfeeders/resources/. Accessed: 2026-05-05.

Liu, X., Gong, C., and Liu, Q. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022. URL https://arxiv.org/abs/2209.03003.

Mao, Y., Qu, Y., Wang, Q., and Ji, X. Adaptive neighborhood-constrained q learning for offline reinforcement learning, 2025. URL https://arxiv.org/abs/2511.02567.

Park, S., Li, Q., and Levine, S. Flow q-learning, 2025. URL https://arxiv.org/abs/2502.02538.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations, 2021. URL https://arxiv.org/abs/2011.13456.

Stoica, G., Ramanujan, V., Fan, X., Farhadi, A., Krishna, R., and Hoffman, J. Contrastive flow matching, 2025. URL https://arxiv.org/abs/2506.05350.

Wang, J., Xu, W., Gu, Y., Song, W., and Green, T. C. Multi-agent reinforcement learning for active voltage control on power distribution networks, 2022. URL https://arxiv.org/abs/2110.14300.

Yang, S. and Zhu, Y. Offline reinforcement learning for microgrid voltage regulation, 2025. URL https://arxiv.org/abs/2505.09920.

A. Detailed Pseudocode of the Improved FQL Algorithm

This appendix provides a complete training procedure of the improved FQL approach introduced in Section 3. As detailed in Algorithm 1, the practical implementation adopts a unified training loop, where global Boltzmann resampling, flow matching, critic optimization, and residual actor optimization are performed sequentially within each offline RL iteration.

In particular, Step 2.2 details the implementation of the *Deterministic Zero-Noise Initialization*. Using the deterministic bijective property of the Probability Flow ODE and setting the initial noise $z_0 = \mathbf{0}$, the integration slides smoothly along the mean trajectory. The comprehensive procedure is summarized below.

Algorithm 1 Training of Improved Flow Q-Learning (FQL) for Voltage Regulation

Input: Offline dataset $\mathcal{D}$, iterations N, temperature τ, weight λ_{base}
Initialize: Flow network v_θ, Critic networks $Q_{\phi_{1,2}}$, Residual Actor ξ_ψ

// Reward ranking for Boltzmann resampling
Compute normalized reward ranks and energy values for all transitions in $\mathcal{D}$

// Training loop
for $t = 1$ **to** N **do**
  Update sampling probability $P(i) \propto \exp(E_i/\tau(t))$
  Sample batch transitions (s, a, r, s') from $\mathcal{D}$ per $P(i)$
  // Step 1: Flow Matching Update
  Sample $t \sim \text{Uniform}(0, 1)$, $\epsilon \sim \mathcal{N}(0, I)$
  Construct flow path: $a_t = (1 - t)\epsilon + ta$
  Optimize θ by minimizing:
    $\mathcal{L}_{flow}(\theta) = \mathbb{E}\left[\|v_\theta(s, a_t, t) - (a - \epsilon)\|^2\right]$
  // Step 2: Base Action and Residual Perturbation
  Set initial noise $z_0 = \mathbf{0}$
  Solve ODE $\frac{dz}{d\tau} = v_\theta(s, z, \tau)$ ($\tau \in [0, 1]$) to obtain a_{base}
  Generate perturbation: $\Delta a = \xi_\psi(s, a_{base})$
  Clip action: $a = \text{clip}(a_{base} + \Delta a, -a_{max}, a_{max})$
  // Step 3: Critic & Actor Update
  Compute target y via flow-guided candidate actions at s'
  Update ϕ_1, ϕ_2 to minimize $\mathcal{L}_{critic} + \alpha_{cql}\mathcal{L}_{cql}$
  Update ψ to maximize $Q(s, a_{base} + \Delta a)$ with BC penalty
end for