Extended Abstract

Motivation As rooftop solar adoption accelerates in the United States, distribution grids face increasingly unpredictable voltage fluctuations that legacy controllers struggle to handle. Volt-VAR control (VVC), dispatching reactive power from smart inverters to keep bus voltages within safe limits, is one of the most critical elements of managing a constant grid environment. Traditional approaches like droop curves and optimal power flow (OPF) solvers assume full, accurate knowledge of grid parameters which are often unable to handle the real-time solar variability. Recent RL-based approaches have shown promise but have all been trained and evaluated on the same fixed simulation environment. This leaves it unclear how well these policies generalize to real solar grids that face unpredictable noise.

Method We train a Soft Actor-Critic (SAC) agent on the IEEE 13-bus distribution feeder simulated in OpenDSS, a grid physics simulator, applying domain randomization across four conditions: (1) a fixed nominal grid, (2) with solar/load perturbations only, (3) grid parameter perturbations only, and (4) both simultaneously. We additionally extend the RL techniques used in prior works, comparing a Lagrangian SAC variant which treats voltage limits as hard constraints, and a curriculum Lagrangian SAC that progressively increases the randomization distribution throughout training.

Implementation The environment simulates one full day as a 96-step episode with each step simulation a 15-minute interval. The agent observes a 17-dimensional state vector, comprising 12 bus voltage magnitudes, 4 PV active power fractions, and time of day. The actor is a 4-layer MLP outputting normalized reactive power controls ranging from -1 to 1 for each of the four PV inverters. Two target critic networks estimate Q-values. Domain randomization perturbs solar irradiance scale, load demand scale, line impedances ($\pm 20\%$), and capacitor ratings ($\times 0.5$–1.5) at episode reset. This added up to 27 models trained total: 9 configurations $\times$ 3 seeds, with 200k training steps each.

Results SAC with full domain randomization achieves 13 voltage violations per episode compared to 83 for SAC without DR and 172 for IEEE 1547 droop control — an 84% reduction. Lagrangian SAC without DR achieves zero violations on the nominal grid. Curriculum Lagrangian SAC reduces violations to 2 per episode under full DR, recovering most of the constraint enforcement lost when applying full DR to vanilla Lagrangian SAC (9 violations). Domain randomization consistently increases line losses slightly, reflecting a real safety-efficiency tradeoff.

Discussion Domain randomization acts as a regularizer introducing diversity during the training to prevent the policy from overfitting to a single grid configuration. Additional techniques like Lagrangian SAC separates constraint satisfaction from loss minimization, but requires careful scaling of the constraint signal: squared violations ($\mathcal{O}(10^{-4})$) are too small to drive meaningful dual updates without rescaling. Curriculum training bridges this gap by starting from easier conditions before exposing the agent to full randomization.

Conclusion We demonstrate that combining deep RL with domain randomization produces Volt-VAR control policies that are significantly more robust to grid model mismatch than both classical baselines and fixed-environment RL. The Lagrangian and curriculum variants further improve constraint satisfaction with minimal efficiency cost, suggesting a practical path toward deploying learned VVC policies on real distribution hardware.

RL & Domain Randomization for Volt-VAR Control in Electricity Distribution Grids

Aniket Mahajan
Department of Computer Science
Stanford University
aniketm@stanford.edu

Anish Chaudhuri
Department of Computer Science
Stanford University
anishch@stanford.edu

Abstract

As rooftop solar panel adoption increases, distribution grids face unpredictable voltage fluctuations that legacy Volt-VAR controllers struggle to handle. We train Soft Actor-Critic (SAC) agents as well as other RL models on the IEEE 13-bus feeder in OpenDSS under four domain randomization (DR) conditions, targeting the simulation to real life gap left unexplored by prior RL-based VVC work. We additionally introduce a Lagrangian SAC variant enforcing voltage limits as hard constraints and a curriculum variant that progressively widens randomization during training. SAC with full DR reduces voltage violations by 84% relative to SAC without DR, and by 92% relative to IEEE 1547 droop control, while Curriculum Lagrangian SAC achieves near-zero violations under full randomization. Our results demonstrate that domain randomization is the key lever for producing generalizable VVC policies, and that the Lagrangian formulation effectively separates safety and efficiency as distinct objectives.

1 Introduction

The rapid increase in rooftop solar systems is changing the operational conditions of electricity distribution grids. Unlike traditional unidirectional power flow from generators to consistent loads, high solar penetration introduces bidirectional power flows that cause rapid and unpredictable voltage fluctuations throughout the distribution network as a function of cloud cover or physical grid conditions like wire degradation. When solar generation exceeds local demand, excess power flows back into the grid and can push bus voltages above safe operating limits. Conversely, sudden drops in solar output due to cloud cover or time-of-day effects can cause voltage variability. Managing these fluctuations, especially in real time, is one of the main challenges of the modern grid.

Volt-VAR control (VVC) addresses this challenge by using reactive power from smart inverters to regulate bus voltages within a safe range. Traditional approaches fall into two categories. (1) Rule-based droop curves map local voltage measurements to fixed reactive power setpoints. This is purely pattern matching with no coordination across inverters and ability to anticipate solar ramps Liu et al. (2018). Optimal power flow (OPF) computes globally optimal setpoints but requires a complete and accurate network model making it too computationally expensive for real-time deployment.

Recent reinforcement learning approaches have mitigated the need for explicit grid models by learning control policies directly from simulation, demonstrating the ability to outperform both droop and OPF Liu and Wu (2020, 2021). However, all prior work trains and evaluates on the same fixed simulation environment which runs the risk of overfitting to a single grid model. In practice, real grids differ from simulators due to aging equipment, measurement error, and unmodeled dynamics. A policy that has never seen variation will likely fail when deployed on real hardware.

We address this gap by applying domain randomization (DR), a reality simulation transfer technique developed in robotics Tobin et al. (2017), to Volt-VAR control for the first time in a continuous

reactive power setting. Prior work applying DR to grid control Leurent et al. (2021) considered only binary on/off demand response scheduling, a simpler problem than the continuous, coordinated reactive power dispatch we address here. We train SAC agents across randomized distributions of solar profiles, load profiles, line impedances, and capacitor ratings, forcing the agent to learn policies robust to grid model mismatch. We then introduce a Lagrangian SAC variant that treats voltage limits as hard constraint rather than soft penalties on reward, and a curriculum variant that progressively widens the randomization distribution during training.

2 Related Work

2.1 Conventional Approaches

Traditional VVC relies on rule-based droop curves that map local voltage measurements to fixed reactive power setpoints. While simple and widely deployed, droop curves are brittle under high solar utilization where voltage can swing quickly, and cannot model inter-inverter coordination. OPF improves upon this by computing the globally optimal settings, but requires precise network data and is generally too slow for real-time responses Liu et al. (2018).

2.2 RL Approaches on a Fixed Grid

Recent work has demonstrated that deep RL can learn effective VVC policies directly from simulation without requiring explicit grid models. Liu and Wu (2020) propose a two-stage approach with adversarial offline training followed by online fine-tuning. Liu and Wu (2021) extend this to a multi-agent experiment where each inverter acts its own autonomous agent using only local voltage observations. While this work demonstrates that RL can outperform classical baselines, they all train and evaluate on the same fixed grid model, leaving ambiguity on their ability to generalize on models outside of this scope.

2.3 Domain Randomization

Domain randomization was developed for sim-to-real transfer in robotics Tobin et al. (2017), where training over randomized physical settings produced policies that transfer to real hardware without requiring additional fine-tuning. Leurent et al. (2021) applied domain randomization to grid control in a demand response setting, showing that randomizing load distributions improves generalization. However, their setting only covered discrete on/off load scheduling, a different problem from continuous reactive power control on setpoints. Here, the magnitude and coordination of actions across multiple inverters matters significantly.

3 Method

3.1 Environment

We simulate the IEEE 13-bus test feeder in OpenDSS. The feeder includes four rooftop solar PV inverters (PV675, PV680, PV611, PV652) on which the agent selects the reactive power setpoints. Each episode simulates one full day as $T = 96$ timesteps of 15 minutes each.

The agent observes a 17-dimensional state vector $o \in \mathbb{R}^{17}$ consisting of

- 12 bus voltage magnitudes (the 13th slack bus is a maintained constant value),
- 4 PV active power fractions (current output / rated capacity), and
- 1 normalized time of day from 0 to 1.

The agent outputs a 4-dimensional continuous action a in the range from -1 to 1 inclusive, representing normalized reactive power setpoints for each inverter. These are then mapped to the actual reactive energy dispatched, Q_i, using $Q_i = a_i \cdot \text{kVA}_i$ where kVA is the capacity of the inverter in kilovolts.

The per-step reward is:

$$r_t = -\alpha \sum_{b \in \mathcal{B}} \left[\max(0, V_b - V_{\max})^2 + \max(0, V_{\min} - V_b)^2 \right] - \beta \frac{P_{\text{loss},t}}{P_{\text{load},t}} \tag{1}$$

whewre $V_{\min} = 0.95$ per unit, $V_{\max} = 1.05$ per unit, $\alpha = 10$, $\beta = 1$.

3.2 Domain Randomization

At the start of each training episode, grid parameters are sampled from the initialized distributions. We apply randomization under four settings:

- None: Fixed grid parameters
- Solar/Load: Solar irradiance scale initialized with a Uniform(0.5, 1.5) and load demand scale Uniform(0.7, 1.3)
- Grid: Line impedances perturbed $\pm 20\%$ per line independently and capacitor ratings scaled from 0.5 – 1.5x
- Both: Both solar/load and the grid randomization applied in conjunction

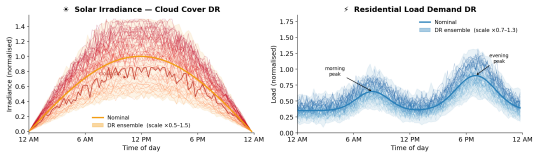

Figure 1: Domain randomization stochastic perturbations

Since the agent never directly observes the sampled parameters, it is forced to learn the policies only on observations of the perturbed voltages from the state vector.

3.3 Soft Actor-Critic

We use Soft Actor-Critic (SAC), an off-policy deep RL algorithm that maximizes a combination of expected reward and policy entropy:

$$J(\pi) = \mathbb{E}\left[\sum_t r_t + \alpha_{\text{ent}} \mathcal{H}(\pi(\cdot|s_t))\right] \tag{2}$$

The entropy bonus encourages action diversity, which is important under domain randomization where the agent must handle a wide range of grid conditions. The actor is a 4-layer MLP (17-256-256-4) with ReLU activations outputting a Gaussian distribution over actions, squashed through $\tanh$ to enforce the $[-1, 1]$ bound. Two critic networks (input dim 21, layers 256-256-1) estimate $Q(s, a)$ independently with the minimum is used to prevent overestimation with double q-learning. Critics update by minimizing the Bellman error against targets computed with a target network updated with $\tau = 0.005$.

3.4 Lagrangian SAC

Standard SAC treats voltage violations as a weighted penalty in the reward, making constraint satisfaction dependent on the choice of α. We introduce a Lagrangian formulation that treats voltage limits as hard constraints:

$$r_L = r_{\text{task}} - \lambda \cdot c_{\text{safety}} \tag{3}$$

where c_{safety} measures constraint violations and λ is a dual variable that self-adjusts every $N = 500$ steps:

$$\lambda \leftarrow \max(0, \lambda + c_safety) \tag{4}$$

where

$$c_{safety} = -\alpha \Sigma_{b \in \mathcal{B}} [\max(0, V_b - V_{max}) + \max(0, V_{min} - V_b)] \tag{5}$$

which uses *L1 norm* instead of a *sum-of-squares (L2-like)* loss. λ appears to be the maximum of 0 and $\lambda + c$ since the baseline b_0 that we subtract from $\lambda + c$ is 0, since we want 0 violations to occur. One

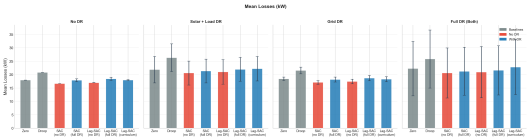

Figure 2: Aggregate Violations & Losses With Domain Randomization

Table 1: Domain-randomisation ablation for LAG-SAC. Violations and reward are averaged across all four test conditions and three seeds (mean $\pm$ std).

DR Training	**Violations** ($\downarrow$)	**Reward** ($\uparrow$)
None	76.9 ± 3.8	-0.88 ± 0.01
Solar & Load	20.0 ± 11.4	-0.92 ± 0.03
Grid	33.8 ± 9.6	-0.88 ± 0.02
Both	29.6 ± 34.5	-0.92 ± 0.03
Curriculum	$\mathbf{15.4 \pm 1.2}$	-0.93 ± 0.02

can think of this Lagrangian as formalizing a *lexicographic* scoring where constraint violations are considered more severe than additional voltage loss. λ increases when violations occur and decreases when the agent operates safely within bounds, automatically balancing the safety-efficiency tradeoff without manual tuning of penalty weights.

3.5 Curriculum Lagrangian SAC

Applying full domain randomization from the start of training presents a challenge: the environment may be too difficult early in training for the agent to learn useful signal, particularly for the Lagrangian constraint. Curriculum Lagrangian SAC addresses this by progressively widening the randomization distribution as training progresses — starting near nominal conditions and expanding to full DR as the agent's competence increases. This provides training stability while ultimately achieving the same breadth of randomization.

4 Experimental Setup

We trained 27 models total: 9 configurations (SAC, Lagrangian SAC, and Curriculum Lagrangian SAC each under None, Solar/Load, Grid, and Both DR conditions, plus Zero VAR and Droop baselines) $\times$ 3 random seeds, for 200k environment steps each. All RL models use the Stable Baselines 3 SAC implementation with default hyperparameters except where noted. Each trained policy was evaluated over 10 episodes.

Baselines include:

- **Zero VAR**: No reactive power control ($Q = 0$)
- **Droop (IEEE 1547)**: Industry-standard rule-based control mapping local voltage to reactive power via a fixed droop curve

We report three metrics: mean violations per episode (bus-timestep pairs outside $[0.95, 1.05]$ p.u.), mean line losses in kW, and mean episode reward.

5 Results

5.1 Domain Randomization Effect on Model Performance

Table 1 presents the curriculum Lagrangian SAC ablations (the best performing model) across all four DR training conditions, averaged over all test conditions and three seeds. The results show the inherent hierarchy: training without any domain randomization produces the worst generalization, with 76.9 ± 3.8 violations per episode. Solar and load randomization provides the biggest gain,

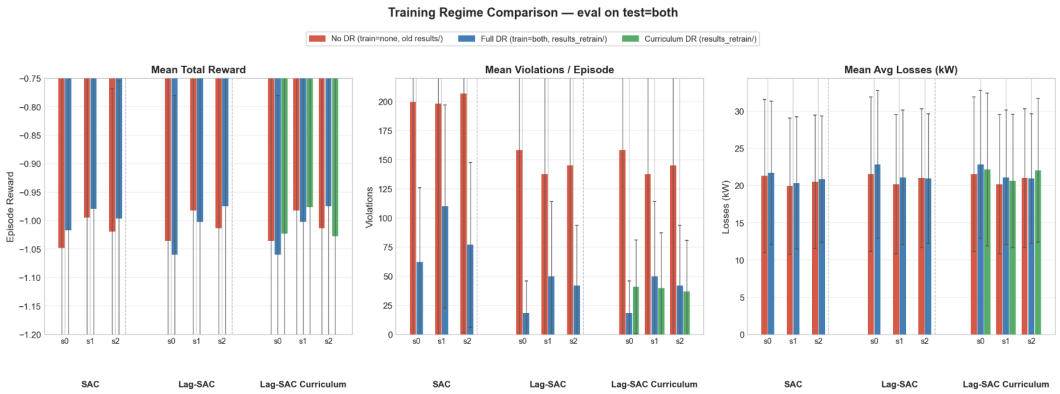

Figure 3: Training comparison on different model architectures

Table 2: Mean ± std over three seeds and 20 episodes per seed on `test=both`

Method	Reward	Violations / ep.	Losses (kW)
SAC			
No DR (`train=none`)	-1.021 ± 0.027	202 ± 5	20.6 ± 0.7
Full DR (`train=both`)	-0.998 ± 0.019	83 ± 24	21.0 ± 0.7
Lagrangian SAC			
No DR	-1.010 ± 0.027	147 ± 10	20.9 ± 0.7
Full DR (retrained)	-1.012 ± 0.044	37 ± 16	21.6 ± 1.1
Curriculum DR (retrained)	-1.009 ± 0.028	39 ± 2	21.6 ± 0.9
Baselines (retrained eval batch)			
MPC	-1.102	235	22.1
Droop (IEEE 1547)	-1.346	198	25.8
Zero VAR	-1.285	359	22.2

reducing violations to 20.0 ± 11.4, confirming that stochastic solar generation is the primary driver of out-of-distribution failure. Grid parameter perturbations provide a smaller improvement (33.8 ± 9.6), suggesting that line impedance and capacitor variation is a secondary to the distributional shift. Training on both axes simultaneously yields 29.6 ± 34.5 violations. This is performant with the mean but with higher variance coming from the compounding variance largely coming from the solar+load domain randomization.

5.2 Model Architectures

Table 2 presents the full comparison on `test=both`. Notably, all three RL methods with domain randomization outperform all three baselines on violations: full-DR SAC (83 ± 24) and full-DR Lag-SAC (37 ± 16) both beat MPC (235), droop (198), and zero VAR (359), despite baselines having access to explicit physics or deterministic rules .

Within RL methods, the architectural progression is clear. SAC without DR (202 ± 5 violations) performs comparably to droop, demonstrating that a fixed-environment policy overfits to nominal conditions. Full-DR SAC cuts this to 83 ± 24, a 59% reduction, at the cost of only 0.4 kW in additional losses. Lagrangian SAC with full DR further reduces violations to 37 ± 16, a 56% improvement over full-DR SAC, by treating voltage limits as hard constraints rather than soft penalties. Curriculum DR matches Lag-SAC on mean violations (39 ± 2) while reducing cross-seed standard deviation by an order of magnitude (± 2 vs. ± 16), making it the most reliable training recipe for deployment.

5.3 Qualitative Analysis & Training Dynamics

Figure 4 shows bus voltage profiles over a 24-hour episode for five policies.

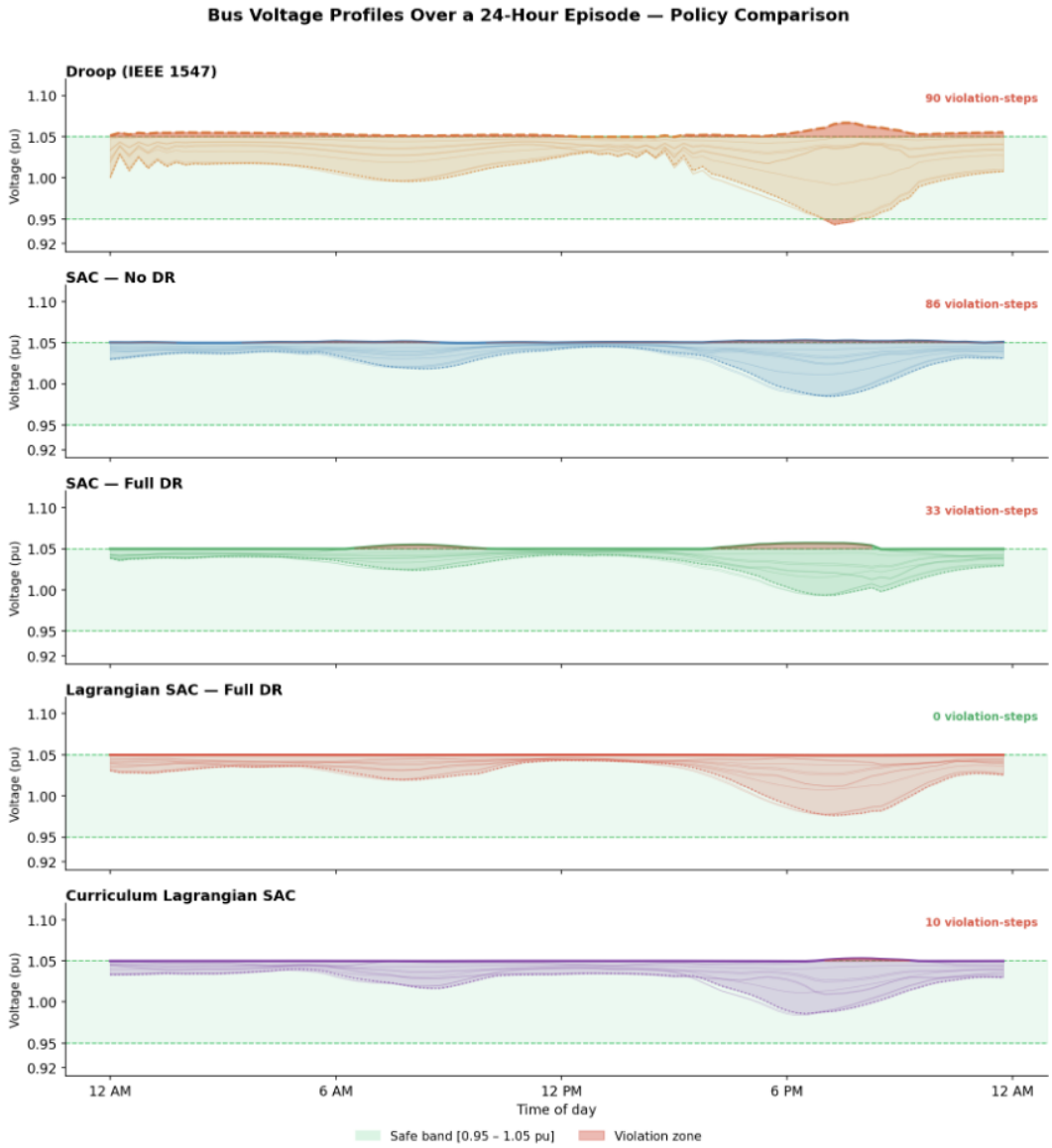

Figure 4: Bus voltage profiles over a 24-hour episode for five policies

Figure 6 shows in-training evaluation reward trajectories for the retrained agents. All three methods converge within 200k steps, with full-DR SAC and curriculum Lag-SAC showing smoother trajectories than full-DR Lag-SAC, which exhibits higher variance during mid-training as the dual variable λ adapts to the constraint signal.

The curriculum variant recovers cleanly after each phased DR transition, confirming the suspicion that the progressive exposure to harder conditions improves training stability.

6 Discussion

Domain randomization makes a huge difference. When training under violations that perturb both solar & load perturbations as well as grid perturbations that represent capacitance and wear & tear, full-DR cuts, even on less-constrained methods like SAC, violations by more than half relative to no-DR training with almost no change in reward (Table 2, Figure 3). Exposing the agent to solar, load, and grid variability during training forces it to learn reactive-power strategies that transfer,

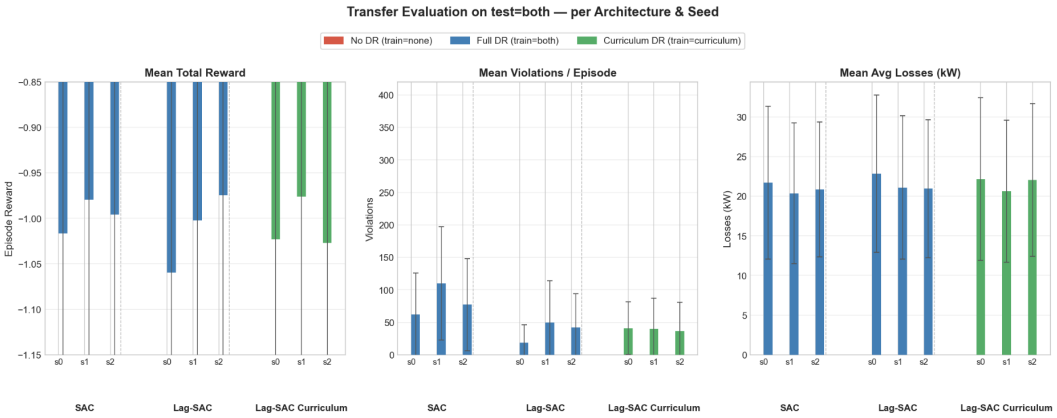

Figure 5: Trained architectures

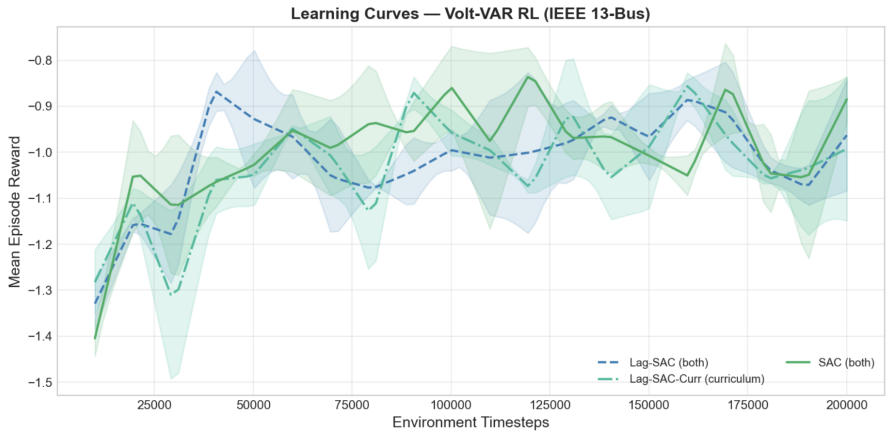

Figure 6: In-training evaluation reward trajectories for retrained agents. Shaded bands: ± 1 std across seeds.

rather than memorize responses to a single daily profile. An initial ablation on the nominal feeder (`test=none`) further illustrates why DR matters: Lag-SAC without DR achieves zero violations on the fixed grid, but this specialization does not carry over to `test=both` ($\sim$147 violations).

Lagrangian SAC reallocates optimization toward safety at negligible reward cost. Once Domain Randomization is in place, Lag-SAC provides the second performance tier: violations drop from $\sim$83 (full-DR SAC) to $\sim$37 (full-DR Lag-SAC) with a mean reward change of only 0.002 relative to no-DR Lag-SAC. The safety–efficiency tradeoff is visible but small: full-DR Lag-SAC adds $\sim$0.7 kW in average losses relative to its no-DR counterpart on `test=both`, an exchange most operators would accept for a increased reduction in voltage-limit breaches. The episode-voltage traces (Figure 4) make the mechanism tangible: Lag-SAC holds the full fleet inside the ANSI band on a representative shifted episode, while SAC and droop exhibit prolonged excursions.

Curriculum training improves reproducibility, not peak safety. On `test=both`, curriculum Lag-SAC does *not* beat full-DR Lag-SAC on mean violations (39 vs. 37) — contrary to what easier `test=none` evaluations in the initial batch suggested. Its value in the retrained experiment is instead **lower cross-seed variance** (39 ± 2 vs. 37 ± 16) and a training trajectory that recovers after the phased DR transitions (Figure 6). Phased exposure to distributional shift is therefore best viewed as a *stability* mechanism for constraint-satisfying policies, not a guarantee of lower violation counts at deployment.

Limitations & Future Work First, the fixed reward defined by α and β above as the final evaluation metric for reward is contestable. There is no concrete relationship between α and β that should be defined. Future work involves adjusting α and β to be values that are more emblematic of a tradeoff which operators are likely to accept. In addiiton, our work remains synthetic; next steps could include incorporate NREL real data instead of current pure simulation evaluationd. Finally, we can test this methodology on smaller & larger bus number simulations, as well as incorporate the fact that most grids, hardware-wise optimize decentrally, with minimal view of the whole grid's statistics; thus, a multi-agent approach may be more appropriate.

7 Conclusion

We demonstrate that combining deep RL with domain randomization produces Volt-VAR control policies significantly more robust to grid model mismatch than both classical baselines and fixed-environment RL. Under the hardest evaluation condition (`test=both`), full-DR SAC reduces violations by 59% relative to no-DR SAC, and full-DR Lagrangian SAC achieves a further 56% reduction over full-DR SAC — outperforming MPC, droop, and zero VAR by a wide margin despite those baselines having access to explicit physics or deterministic rules. Curriculum Lagrangian SAC matches the safety performance of full-DR Lag-SAC while reducing cross-seed variance by an order of magnitude, offering the most reproducible training recipe for deployment. Together these results suggest a practical path toward deploying learned VVC policies on real distribution hardware — addressing the sim-to-real gap that has limited all prior RL-based VVC work.

We perturbed values of $\lambda_{init} \in \{0.0001, 0.001, ...1\}$. We found that λ of 0.001 worked best, with minimal mobility changes based on λ_{init}.

8 Team Contributions

- **Aniket Mahajan:** Lead on the simulation environment — building and validating the OpenDSS Gymnasium wrapper, integrating solar and load profiles, collecting droop baseline, constructing the evaluation set, writing Abstract, Introduction, and Results sections.
- **Anish Chaudhuri:** Lead on the RL training pipeline — implementing and tuning SAC, Lagrangian SAC, and Curriculum variants, designing the domain randomization sampling procedure, running ablation experiments, writing Methods and Conclusions sections.

Changes from Proposal The core hypothesis and objective remain unchanged. We added the Lagrangian SAC and Curriculum Lagrangian SAC variants beyond the originally proposed SAC baseline, and replaced the planned PPO/TD3 comparisons with a more thorough ablation across DR conditions for the constrained RL variants. MPC was evaluated as an additional baseline beyond the original proposal scope.

References

E. Leurent et al. 2021. Sample Efficient Reinforcement Learning with Domain Randomization for Automated Demand Response in Low-Voltage Grids. *IEEE Transactions on Smart Grid* (2021).

H. Liu and W. Wu. 2020. Two-Stage Deep Reinforcement Learning for Inverter-Based Volt-VAR Control in Active Distribution Networks. (2020). arXiv:2005.11142 [eess.SY]

H. Liu and W. Wu. 2021. Online Multi-Agent Reinforcement Learning for Decentralized Inverter-Based Volt-VAR Control. (2021). arXiv:2006.12841 [eess.SY]

M. Liu, L. Ochoa, and S. Low. 2018. Using OPF for Smart Grids: From Concept to Reality. *IEEE Smart Grid Newsletter* (2018).

J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel. 2017. Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World. In *International Conference on Intelligent Robots and Systems (IROS)*.