Hybrid Reinforcement Learning and Optimal Power Flow for Smart Grid Voltage Control

1st Dandi Desta
Electrical and Computer Engineering
Harvard University
Cambridge, MA, USA

2nd Sofiyat Bello
Electrical Engineering
Harvard University
Cambridge, MA, USA

***Abstract*—Modern distribution grids are increasingly challenged by high penetrations of distributed energy resources (DERs), which introduce bidirectional power flows, voltage fluctuations, and operational uncertainty. While alternating current Optimal Power Flow (AC OPF) provides a physically grounded method for safe DER dispatch, repeatedly solving a constrained nonlinear optimization problem introduces computational latency that scales poorly as grid size and real-time control requirements increase. This paper presents a hybrid control framework that combines the millisecond inference speed of deep reinforcement learning (RL) with the deterministic safety guarantees of numerical OPF. We model a 10-house low-voltage radial feeder with five DERs: three rooftop solar photovoltaic units, one wind turbine, and one natural-gas micro-generator. A Proximal Policy Optimization (PPO) agent learns to dispatch or curtail DER output in response to bus voltages, household loads, and environmental conditions. To guarantee physical reliability, each proposed RL action is rapidly verified using a pandapower AC power-flow simulation. Actions satisfying the 0.95 to 1.05 per-unit voltage band are applied directly; otherwise, an OPF fallback computes a corrected, feasible dispatch. Crucially, failed RL actions and OPF corrections are logged for online behavior cloning, allowing the policy to self-correct and continuously drive down the verification failure rate (p_{fail}). Although the 10-house network is intentionally simplified, it provides a controlled environment to validate how learned control, numerical verification, and optimization-based fallback can be combined. Ultimately, this framework demonstrates a scalable pathway to achieve the adaptive speed of agentic controllers while strictly preserving the safety of classical optimization.**

***Index Terms*—Smart grid, distributed energy resources, reinforcement learning, optimal power flow, voltage regulation, hybrid control, pandapower, PPO.**

I. Introduction

Modern electric power systems are undergoing a fundamental transition from centralized, one-directional generation toward decentralized, bidirectional operation. The rapid proliferation of distributed energy resources (DERs)—such as rooftop solar photovoltaics (PV), small wind turbines, and micro-generators—enables residential customers to actively inject power into the grid [1]. While this transition enhances local energy resilience, uneven DER penetration introduces critical operational challenges. High localized power injections during periods of low household demand frequently cause voltage rise and imbalance. Consequently, maintaining all bus voltages within the safe operating range is a primary control problem for DER-rich distribution feeders.

Traditionally, Alternating Current Optimal Power Flow (AC OPF) has served as the gold standard for power system dispatch. By formulating the grid as a constrained nonlinear optimization problem, AC OPF provides mathematically guaranteed safety and optimality, ensuring that all physical constraints—such as the standard 0.95 to 1.05 per-unit (p.u.) voltage band—are strictly respected [2]. However, repeatedly solving a full OPF problem at every control timestep introduces significant computational latency. As grids scale to include thousands of active nodes with stochastic renewable generation, the computational burden of real-time numerical optimization becomes a critical bottleneck.

Reinforcement learning (RL) offers a highly computationally efficient alternative for real-time control [3], [4]. Instead of solving an optimization problem online, an RL agent learns a policy that maps the current grid state directly to a DER control action. After training, the policy produces actions with sub-second inference latency. However, learned policies are statistical approximations. They lack inherent physical awareness and cannot deterministically guarantee that proposed actions will satisfy critical power-system constraints under rare or out-of-distribution conditions [3]. Deploying pure RL on a physical feeder carries an unacceptable risk of localized voltage violations.

To bridge the gap between the speed of learned control and the physical reliability of numerical optimization, this paper proposes a hybrid, self-correcting control architecture. We formulate the DER dispatch problem as a Markov Decision Process (MDP) with a 26-dimensional observation space (incorporating bus voltages, active loads, available DER power, and temporal features) and a 5-dimensional continuous action space representing DER curtailment fractions. A Proximal Policy Optimization (PPO) agent acts as the primary controller. Crucially, each RL-proposed action is rapidly verified using a pandapower AC power-flow simulation. If the action risks a voltage violation, it is rejected, and an AC OPF solver is invoked as an emergency fallback to compute a minimum-curtailment dispatch that satisfies constraints.

This hybrid architecture is beneficial under the condition that the average control latency satisfies $T_{\text{RL}} + T_{\text{PF}} + p_{\text{fail}} T_{\text{OPF}} < T_{\text{OPF}}$, where p_{fail} is the probability that the RL policy requires OPF correction. To continuously drive down p_{fail}, the architecture implements an online behavior-cloning loop. Whenever

This work was completed for ES145 Final Project

the OPF fallback is triggered, the unsafe state and OPF-corrected action are logged and used to fine-tune the RL policy, allowing the agent to approximate OPF behavior over time.

We validate this framework on a simulated low-voltage residential feeder comprising 10 houses and five heterogeneous DERs (three 5 kW PV systems, one 5 kW wind turbine, and one 10 kW gas micro-generator). While OPF-only control remains computationally feasible at this localized scale, this demonstration serves as a controlled, high-fidelity proxy for large-scale distribution grids.

The main contributions of this paper are as follows:

- **Hybrid Control Architecture:** We propose a framework combining the low-latency inference of a PPO agent with the strict physical safety guarantees of an AC OPF fallback.
- **Deterministic Verification:** We implement a simulation-in-the-loop verification layer that physically tests proposed RL actions prior to deployment, ensuring zero real-world constraint violations.
- **Self-Correcting Policy Loop:** We design an online behavior-cloning mechanism that continuously refines the RL policy using OPF-generated expert labels, systematically reducing the verification failure rate (p_{fail}).

II. Background and Related Work

A. Traditional Voltage Regulation and Optimization

The increasing penetration of photovoltaics (PVs) and other distributed energy resources in the distribution network frequently causes voltage swings due to rapid power variations. To mitigate these issues, utilities traditionally rely on voltage regulation strategies categorized into active and reactive power control. While active power curtailment reduces the network's absorption capacity for renewable energy and battery scheduling is costly, reactive power control utilizing PV inverters is recognized as a highly effective and economical solution.

From a control architecture perspective, classical methods are generally divided into centralized optimization, decentralized implementation, and local control. Centralized strategies, such as Alternating Current Optimal Power Flow (AC OPF), provide mathematically optimal and physically safe dispatch commands. However, these centralized control strategies require fast, reliable communication links and suffer from severe computational latency as the network scales. Conversely, local control approaches (such as droop control) make immediate decisions based on local observations without communication latency, but they often fail to fully utilize system resources due to a lack of cooperation between nodes.

B. Reinforcement Learning for Grid Control

To overcome the computational burden of centralized OPF and the sub-optimality of local droop control, deep reinforcement learning (DRL) has gained significant traction. DRL moves the heavy computational lifting to an offline training phase, allowing for sub-second, inference-based decision-making in real-time.

Recent advancements have explored Multi-Agent Deep Reinforcement Learning (MADRL) to manage complex distribution networks. For instance, Cao et al. [5] proposed a MADRL-based approach for voltage regulation using coordinated PV inverters. In their framework, each PV inverter acts as an independent agent utilizing a Multi-Agent Deep Deterministic Policy Gradient (MADDPG) algorithm enhanced with an attention mechanism. By training a centralized critic network offline, the agents learn to model the policies of other inverters. Once deployed, the actors execute coordinated control strategies using strictly local information, successfully avoiding the need for costly real-time communication networks. Comparative results on standard test systems demonstrated that this MADDPG approach achieves significantly smaller voltage fluctuations than traditional droop control methods [5].

C. The Need for Safe, Hybrid Architectures

While pure reinforcement learning frameworks, such as the MADDPG approach, demonstrate exceptional speed and adaptability, they fundamentally lack deterministic physical guarantees. In these models, safety is encouraged statistically via a reward function—for example, penalizing the agent based on the absolute sum of voltage deviations across the network. However, because neural networks act as statistical approximators, they can still output unpredictable or out-of-bounds actions when presented with unseen, highly stochastic weather or load dynamics.

In a physical power system, statistical safety is insufficient; absolute constraint satisfaction is required to prevent equipment damage. This critical gap in the existing literature motivates the architecture proposed in this paper. Rather than relying solely on the learned policy to maintain grid stability, we propose wrapping a fast RL agent (PPO) in a deterministic, physics-based safety layer. By implementing a high-speed pandapower verification step and retaining AC OPF strictly as an emergency fallback, our framework preserves the inference speed demonstrated by state-of-the-art DRL methods while guaranteeing the strict constraint satisfaction of classical centralized optimization. Furthermore, by logging OPF corrections to fine-tune the RL policy via behavior cloning, the system bridges the gap between learning-based approximation and numerical reliability.

III. System Design and Methodology

A. System Overview and Grid Model

This project implements a hybrid control framework on a simulated low-voltage (LV) radial distribution feeder using the `pandapower` network simulator. The controlled test environment consists of one 20 kV slack bus, a medium-voltage to low-voltage transformer, and ten residential buses. Five distributed energy resources (DERs) are integrated across houses 1, 3, 5, 7, and 9: three 5 kW solar PV units, one 5 kW wind turbine, and one 10 kW dispatchable natural-gas micro-generator. The objective is to dispatch or curtail DER power at every timestep to maintain bus voltages within a strict 0.95 to

1.05 per-unit (p.u.) safety band while minimizing renewable curtailment.

B. MDP Formulation

The control problem is formulated as a Markov Decision Process (MDP). At each timestep t, the agent observes the 26-dimensional grid state s_t:

$$s_t = [V_1, \ldots, V_{10}, P_1^L, \ldots, P_{10}^L, P_1^a, \ldots, P_5^a, h/24] \quad (1)$$

where V_i is the voltage at house bus i, P_i^L is the active load at house i, P_j^a is the available pre-curtailment power from DER j, and $h/24$ is the normalized time of day.

The agent outputs a 5-dimensional continuous action vector representing the curtailment fraction for each DER:

$$a_t = [a_1, a_2, a_3, a_4, a_5]^T, \quad a_j \in [0, 1] \quad (2)$$

A value of $a_j = 0$ commands full available power injection, while $a_j = 1$ commands total curtailment. The dispatched DER power is therefore:

$$P_j^d(t) = P_j^a(t)(1 - a_j(t)) \quad (3)$$

C. Reinforcement Learning Controller

The primary controller utilizes Proximal Policy Optimization (PPO), mapping the 26-dimensional observation vector to the 5-dimensional action vector. The reward function prioritizes strict voltage safety followed by curtailment minimization:

$$r_t = \begin{cases} -100 w_v n_{\text{viol}}(V) - w_c C, & n_{\text{viol}}(V) > 0 \\ b_s - w_c C, & n_{\text{viol}}(V) = 0 \end{cases} \quad (4)$$

where $n_{\text{viol}}(V)$ is the number of bus voltage violations, C is the total curtailed power in kW, w_v is the voltage penalty weight, w_c is the curtailment penalty weight, and b_s is the positive stability bonus.

D. Power-Flow Verification and OPF Fallback

Because RL policies lack deterministic safety guarantees, every proposed action is verified on a digital twin of the network using an AC power-flow simulation. If the resulting voltages satisfy the safety band:

$$0.95 \leq V_i \leq 1.05 \quad \forall i \quad (5)$$

the RL action is accepted and applied.

If the action causes a violation or the solver fails to converge, the action is rejected and an AC Optimal Power Flow (OPF) fallback is triggered. The OPF layer computes a corrected dispatch that maximizes usable DER injection subject to system constraints:

$$\max_{P_j^d} \sum_j P_j^d \quad (6)$$

subject to the bounds:

$$0 \leq P_j^d \leq P_j^a \quad (7)$$

$$0.95 \leq V_i \leq 1.05 \quad (8)$$

and the nonlinear AC power-flow mismatch equations.

E. Self-Correcting Learning Loop

Whenever an RL action is overridden by the OPF fallback, the system stores the unsafe state s_t and the OPF-corrected expert action a_{safe}. Periodically, the PPO actor network is fine-tuned against these records via behavior cloning, minimizing the following loss function:

$$L_{\text{BC}}(\theta) = \|\mu_\theta(s_t) - a_{\text{safe}}\|_2^2 \quad (9)$$

where $\mu_\theta(s_t)$ is the mean action predicted by the policy. This forces the RL controller to imitate safe OPF behavior in challenging states, progressively reducing the frequency of OPF overrides.

F. LLM-Based Operator Interface

To manage stochastic environmental parameters, the architecture includes an operator interface driven by a Large Language Model (LLM). Currently, human operators can input natural language commands (e.g., "approaching storm reduces solar output by 70% and prioritizes grid stability"). The LLM parses this intent and autonomously adjusts structured mathematical scaling factors for solar, wind, gas, load, and reward weights. In future real-world deployments, this LLM interface is designed to ingest data directly from external API streams—such as live weather forecasts, news feeds, and utility alerts—allowing the system to autonomously adapt DER dispatch strategies to breaking external events without human intervention.

G. Algorithm Implementation

The execution of this hybrid framework during real-time operation is summarized in Algorithm 1.

IV. Experimental Setup

A. Simulation Environment

The proposed control architecture is evaluated in a Python-based simulation environment utilizing `pandapower` for AC power-flow and OPF computations, `gymnasium` for the MDP framework, and `Stable-Baselines3` for RL training. The test system is a low-voltage radial feeder comprising one 20 kV slack bus, a 0.4 kV transformer, ten residential loads, and five DERs (three 5 kW PV units, one 5 kW wind turbine, and one 10 kW gas micro-generator).

Each simulation episode represents a 24-hour operational period evaluated at hourly timesteps. To ensure deterministic and reproducible comparative evaluation, all test episodes are executed using a fixed set of environment seeds (100, 101, and 102). This guarantees that the exogenous weather profiles, load variations, and stochastic noise trajectories remain perfectly identical across all tested controllers.

B. Training Configuration

The primary RL controller is trained utilizing Proximal Policy Optimization (PPO) with a Multi-Layer Perceptron (MLP) policy network. The agent is trained for 1,000,000 timesteps using an Adam optimizer with a learning rate of 3×10^{-4}. The training procedure utilizes a rollout length of

Algorithm 1 Self-Correcting Hybrid RL-OPF Control Loop

Inputs: PPO policy π_θ, pandapower grid environment E, correction memory buffer $\mathcal{D}$.

Initialize: Time step $t = 0$.

for each control step t **do**
  Obtain current grid state observation s_t from E.
  Step 1: RL Action Proposal
  Sample raw action $a_{\text{RL}} \sim \pi_\theta(s_t)$.
  Step 2: Deterministic Verification
  Simulate a_{RL} on digital twin E_{twin} using AC power flow.
  if all $V_i \in [0.95, 1.05]$ p.u. **then**
    Action is safe: $a_{\text{safe}} \leftarrow a_{\text{RL}}$
    Source $\leftarrow$ "RL"
  else
    Step 3: OPF Fallback
    Run constrained AC OPF on E_{twin} to maximize injection.
    if OPF converges **then**
      $a_{\text{safe}} \leftarrow a_{\text{OPF}}$
      Source $\leftarrow$ "OPF"
    else
      $a_{\text{safe}} \leftarrow$ Maximum Curtailment
      Source $\leftarrow$ "Fallback"
    end if
  end if
  Step 4: Execution and Logging
  Apply a_{safe} to live environment E.
  Store tuple $(s_t, a_{\text{RL}}, a_{\text{safe}}, \text{Source})$ in $\mathcal{D}$.
  Step 5: Self-Correction Trigger
  if retraining conditions met **then**
    Update policy π_θ via behavior cloning against $\mathcal{D}$.
  end if
  $t \leftarrow t + 1$
end for

256 steps per environment update, a batch size of 64, and 10 optimization epochs. To balance exploitation with sufficient state-space exploration, the entropy coefficient is set to 0.01. The generalized advantage estimation (GAE) parameter (λ) is set to 0.95, the discount factor (γ) to 0.99, and the trust-region clipping parameter (ϵ) to 0.2.

C. Baseline Controllers

To rigorously isolate the benefits of the proposed hybrid architecture, performance is evaluated against three distinct baseline controllers under identical environmental conditions:

1) **No-Control Baseline:** All DERs inject their maximum available real power without any curtailment ($a_t = \mathbf{0}$). This serves as the worst-case operational boundary for localized voltage violations.
2) **OPF-Only Controller:** A classical numerical controller that solves an AC OPF at every timestep to minimize curtailment subject to the 0.95 to 1.05 p.u. voltage band. If the solver identifies an infeasible state (e.g., total load vastly exceeds available supply), it conservatively defaults to full curtailment ($a_t = \mathbf{1}$).
3) **RL-Only Controller:** The trained PPO policy directly manipulates the live grid environment. The raw predicted action is applied without any verification or deterministic safety net.
4) **Proposed Hybrid Controller:** The architecture introduced in Section III, where the PPO action is filtered through a deterministic `pandapower` verification layer and backed by the AC OPF emergency fallback.

D. Evaluation Metrics

The central hypothesis of this work is that the hybrid controller safely approximates the optimality of OPF while significantly reducing computational latency. Therefore, the controllers are compared across four primary domains:

1) **Voltage Safety:** Quantified by the mean number of timestep-level violations per episode, the overall percentage of constraint-compliant steps, and the absolute minimum and maximum observed bus voltages ($v_{\min}$, $v_{\max}$).
2) **Curtailment Efficiency:** Measured by the total DER energy curtailed (in kWh) across the evaluation period, alongside the utilization percentages of the available solar and wind resources.
3) **Fallback Frequency:** For the hybrid controller, the system tracks the verification failure rate (p_{fail})—defined as the absolute count of timesteps where the RL action is rejected, triggering either the OPF solver or the conservative emergency fallback.
4) **Computational Latency:** Evaluated by the average wall-clock execution time (in milliseconds) required to compute and verify the dispatch action for a single timestep.

V. Results and Discussion

A. Overall Controller Performance

Table I summarizes the performance of the four controllers over three fixed-seed 24-hour episodes, corresponding to 72 total timesteps per controller. The Hybrid controller achieved the strongest overall performance: it recorded zero voltage-violating timesteps, maintained 100% safe operation, and obtained the highest average reward. By contrast, the RL-Only controller was fast but still produced localized voltage violations, while the OPF-Only controller was significantly slower and failed to maintain safety during structurally difficult evening-load periods.

The No-Control baseline performed poorly, with an average of 14.67 violating steps per episode and a maximum voltage of 1.080 p.u. This confirms that uncoordinated DER injection pushes the feeder outside the acceptable operational range. The OPF-Only controller slightly reduced average voltage deviation compared with No-Control, but it still produced 14.00 violating steps per episode. This result reflects a structural limitation of the simulated system: during evening peak hours, the total load exceeded the available dispatchable DER

TABLE I
AGGREGATE PERFORMANCE METRICS OVER 72 CONTROL STEPS

Controller	Viol/Ep	Safe %	$V_{\min}$	$V_{\max}$	Avg ΔV	Curt. (kWh)	Rcurt %	Solar %	Wind %	OPF %	Fbk	Step ms	Reward
No-Control	14.67	38.9	1.004	1.080	0.041	0.00	0.0	100.0	100.0	0	0	24.07	-7440.00
OPF-Only	14.00	41.7	1.004	1.079	0.040	29.57	4.5	96.1	94.7	100	0	165.05	-7267.65
RL-Only	2.00	91.7	0.997	1.051	0.030	402.86	34.2	65.3	66.6	0	0	28.30	-160.10
Hybrid	**0.00**	**100.0**	**0.997**	**1.050**	**0.028**	470.08	38.0	65.3	57.5	0	6	**80.71**	**+224.33**

capacity, resulting in an infeasible state where no numerical dispatch could fully restore all voltages to the desired band.

B. RL-Only Versus Hybrid Control

The RL-Only controller substantially improved safety compared with the No-Control and OPF-Only baselines, reducing the average violation count to 2.00 steps per episode and achieving a 91.7% safe-step rate. This indicates that the PPO policy successfully learned useful DER dispatch behaviors from its training environment. However, RL-Only still violated voltage limits on 6 out of 72 total timesteps, reaching a maximum voltage of 1.051 p.u. Although this violation is marginal, it demonstrates the fundamental weakness of pure RL in grid applications: the learned policy is statistically effective but lacks deterministic safety guarantees.

The proposed Hybrid controller successfully eliminated these remaining unsafe cases. It achieved zero violations across all 72 evaluated timesteps, bounding the network tightly between $V_{\min} = 0.997$ p.u. and $V_{\max} = 1.050$ p.u. To achieve this, the Hybrid controller sacrificed additional DER energy, increasing total curtailment from 402.86 kWh (RL-Only) to 470.08 kWh. This extra curtailment represents the explicit cost of safety: the Hybrid controller curtailed more aggressively in borderline cases to ensure absolute constraint satisfaction. Consequently, its reward improved from -160.10 (RL-Only) to +224.33, reflecting a reward structure that accurately penalizes voltage violations more heavily than renewable curtailment.

C. Runtime and OPF Usage

The runtime results support the main architectural claim that the Hybrid framework drastically reduces reliance on expensive OPF solves while strictly preserving safety. The OPF-Only controller required an average of 165.05 ms per timestep. The RL-Only controller required only 28.30 ms per timestep but lacked guaranteed safety. The Hybrid controller required 80.71 ms per timestep, making it slower than pure RL but approximately twice as fast as classical OPF.

These results validate the expected timing relationship:

$$T_{\text{hybrid}} = T_{\text{RL}} + T_{\text{PF}} + p_{\text{fail}} T_{\text{OPF}} \tag{10}$$

In this evaluation, the deterministic verification layer flagged 6 borderline RL actions as unsafe. Rather than successfully solving a secondary OPF for these extreme cases, the system routed them to the mathematically guaranteed, conservative fallback path. Therefore, while the OPF invocation percentage was mathematically 0%, the verification failure rate (p_{fail}) accurately captured the 6 required interventions. The observed Hybrid runtime was dominated by RL inference, `pandapower` verification, and fallback handling, effectively bypassing the heavy computational burden of T_{OPF}.

D. Safety–Efficiency Tradeoff

The central operational tradeoff in smart grid dispatch is balancing computational speed, physical safety, and renewable DER utilization. The No-Control baseline maximized DER utilization but generated severe voltage violations. The OPF-Only controller curtailed minimal energy (29.57 kWh) but failed to maintain safety during physically constrained peak load periods. While RL-Only provided an excellent speed and utilization tradeoff, it failed to guarantee absolute safety.

The Hybrid framework achieved the safest overall operation by increasing total curtailment to 470.08 kWh, corresponding to a 38.0% renewable curtailment rate. This highlights the practical value of the Hybrid design. The objective is not to prove that RL mathematically supersedes OPF, but rather to demonstrate that a learned controller can provide fast, highly accurate approximate decisions, while a physics-based verification layer guarantees safety.

E. The Evening Peak Anomaly and OPF Limitation

The OPF-Only failure rate requires careful interpretation. Classical OPF is designed to provide physically safe dispatch when a feasible solution exists within the network parameters. However, this benchmark intentionally features structurally difficult evening hours (hours 18–21) where peak demand reaches approximately 26 kW, while total dispatchable DER capacity is limited to roughly 13 kW.

Under these conditions, the OPF solver fails not because the algorithm is conceptually unsafe, but because it acts myopically. Because OPF optimizes only for the current timestep, it returns an infeasible signal when it cannot perfectly satisfy the voltage band, subsequently defaulting to a conservative action that fails to rescue the severe undervoltage. In stark contrast, the PPO agent possesses a longer time horizon. The RL policy learned a broader strategy to anticipate this deficit, keeping the gas micro-generator operating near maximum capacity leading into and throughout the evening peak. This foresight allowed the Hybrid architecture to maintain the safety constraints exactly where the classical numerical controller mathematically failed.

F. Summary of Findings

Overall, the evaluation validates the proposed hybrid architecture. Across 72 stochastic timesteps, the Hybrid controller

achieved 100% safe operation, reduced voltage violations to zero, and improved the average system reward to +224.33. Operating at 80.71 ms per timestep, it achieved control latencies approximately 2.0 times faster than a pure OPF approach. These results confirm that while deep reinforcement learning provides highly efficient real-time control approximations, a deterministic verification and fallback layer remains mandatory for safe physical deployment.

The primary limitation of this evaluation is its reliance on a localized 10-house feeder. Consequently, these results serve as a foundational proof-of-concept for scalable architecture. Future work will extend this framework to larger distribution nodes, incorporating reactive-power inverter control, integrated battery storage systems, and Multi-Agent PPO (MAPPO) topologies to manage higher penetrations of stochastic renewable generation.

VI. Conclusion and Future Work

This paper presented a hybrid, self-correcting control architecture for safe distributed energy resource (DER) dispatch in stochastic low-voltage smart grids. By combining the sub-second inference speed of a Proximal Policy Optimization (PPO) agent with the deterministic physical guarantees of a `pandapower`-based verification layer and an AC Optimal Power Flow (OPF) fallback, the proposed framework successfully bridges the gap between learning-based adaptability and numerical reliability. Experimental evaluation on a 10-house distribution feeder demonstrated that the hybrid controller achieved strictly 100% safe operation—eliminating the localized voltage violations inherent to pure RL policies—while executing at approximately twice the speed of a traditional OPF-only approach. Furthermore, the RL agent demonstrated critical strategic foresight by proactively managing dispatchable resources during structurally constrained evening load peaks, successfully navigating scenarios where myopic numerical solvers mathematically failed.

While the 10-house simulation serves as a robust proof-of-concept, scaling this architecture to complex, real-world distribution networks presents both data and computational challenges. Future work will focus on expanding the core RL policy into a large-scale foundation model for grid control. This will be achieved by leveraging massive, open-source power system databases—such as the high-fidelity telemetry and grid topology datasets maintained by the National Renewable Energy Laboratory (NREL) and the Pacific Northwest National Laboratory (PNNL). By training very large, high-capacity deep learning models on these expansive datasets, the system can predict high-quality, generalized dispatch actions across diverse, multi-feeder grid topologies before passing them to the physical verification layer.

We acknowledge that as the number of active nodes increases and the predictive models grow in complexity, computational latency for both inference and verification will inherently scale. To maintain the millisecond-level responsiveness required for real-time grid stability, future research must explore the development of Application-Specific Integrated Circuits (ASICs). By designing custom hardware architectures explicitly tailored for highly parallelized neural network inference and sparse-matrix AC power-flow computations, the computational bottlenecks of scaling can be bypassed. Ultimately, the integration of massive predictive models, deterministic physical verification, and dedicated ASIC hardware provides a comprehensive roadmap toward a highly stable, autonomous, and ultra-low-latency smart grid system.

References

[1] J. M. Carrasco et al., "Power-Electronic Systems for the Grid Integration of Renewable Energy Sources: A Survey," *IEEE Transactions on Industrial Electronics*, vol. 53, no. 4, pp. 1002–1016, 2006.

[2] S. Frank, I. Steponavice, and S. Rebennack, "Optimal power flow: A bibliographic survey I: Formulations and deterministic methods," *Energy Systems*, vol. 3, no. 3, pp. 221–258, 2012.

[3] M. Glavic, R. Fonteneau, and D. Ernst, "Reinforcement Learning for Electric Power System Decision and Control: Past Considerations and Perspectives," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 6918–6927, 2017.

[4] Y. Wang et al., "Safe Off-Policy Deep Reinforcement Learning Algorithm for Volt-VAR Control in Power Distribution Systems," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3008–3018, 2020.

[5] D. Cao, W. Hu, J. Zhao, Q. Huang, Z. Chen, and F. Blaabjerg, "A Multi-Agent Deep Reinforcement Learning Based Voltage Regulation Using Coordinated PV Inverters," *IEEE Transactions on Power Systems*, vol. 35, no. 5, pp. 4120–4123, 2020.