



Latent Skill Mining for Distribution Grid Switching Policy Recovery from Historical Operational Switching Logs

Elliot Hamilton, Maya Turner

School of Informatics, University of Edinburgh, Edinburgh, United Kingdom

Abstract

The increasing integration of distributed energy resources and the electrification of various sectors have introduced unprecedented complexities into modern distribution grids. Managing these complexities often requires dynamic topology reconfiguration through switching operations, which are traditionally handled by human operators or rule-based systems. As grid states become more volatile, there is a critical need to automate these switching policies using advanced machine learning techniques. This paper introduces a comprehensive framework based on latent skill mining for the recovery and optimization of distribution grid switching policies from historical operational data. By treating grid reconfiguration as a sequential decision-making process, we propose a methodology that extracts reusable, high-level behavioral primitives, referred to as latent skills, from offline datasets. These skills represent fundamental topological transition strategies that human operators intuitively deploy to resolve congestion, isolate faults, and minimize power losses. The proposed approach leverages a hierarchical architecture where a high-level policy selects continuous latent skills, and a low-level policy translates these skills into specific sequences of switch actions. Through extensive analysis and textual description of the algorithmic structures, this paper demonstrates how latent skill mining mitigates the challenges of large discrete action spaces and distribution shifts commonly found in standard imitation learning and offline reinforcement learning. The findings suggest that the recovered policies exhibit superior adaptability and robustness in unobserved grid conditions compared to traditional monolithic models.

Keywords: Distribution Grid Reconfiguration, Latent Skill Mining, Policy Recovery, Offline Reinforcement Learning, Topology Optimization.

1. Introduction

1.1 Background of Distribution Grid Operations

The transition toward a sustainable energy future has fundamentally altered the operational paradigm of electrical distribution grids. Historically, distribution networks were designed as passive, unidirectional systems that delivered power from bulk transmission substations to end-use consumers. In this traditional framework, operational planning and real-time management were relatively straightforward, governed by predictable load patterns and static topological configurations. However, the proliferation of distributed energy resources, including rooftop solar photovoltaics, residential energy storage systems, and electric vehicles, has transformed the distribution grid into an active, bi-directional energy exchange platform. This transformation introduces significant volatility and uncertainty into the system state, characterized by rapid fluctuations in voltage profiles and bi-directional power flows that can cause localized congestion and equipment overloads. Consequently, grid operators are increasingly reliant on active network management strategies, the most prominent of which is distribution grid reconfiguration. Reconfiguration involves the opening and closing of sectionalizing and tie switches to alter the network topology dynamically. This process serves multiple

operational objectives, such as minimizing active power losses, balancing phase loads, isolating faulted segments, and alleviating thermal constraints on transformers and distribution lines. Despite its critical importance, the decision-making process underlying these switching operations remains heavily dependent on the heuristic knowledge of experienced human dispatchers and deterministic optimization models that struggle to scale with the increasing complexity of the grid. As the frequency and magnitude of grid disturbances escalate, the reliance on manual intervention or computationally expensive optimization techniques becomes a primary bottleneck in achieving a fully autonomous and resilient power distribution network [1]. Modernization efforts must therefore focus on developing intelligent systems capable of learning from historical data to replicate and improve upon the switching policies employed by human experts [2].

1.2 Challenges in Switching Policy Recovery

The endeavor to automate distribution grid reconfiguration through machine learning, specifically via policy recovery from historical operational data, presents a multitude of complex challenges. The fundamental objective of policy recovery is to derive a robust decision-making agent that maps grid states to appropriate switching actions based on a dataset of previously recorded operations [3]. This dataset typically comprises Supervisory Control and Data Acquisition measurements, including bus voltages, line currents, active and reactive power injections, and the corresponding states of all controllable switches over an extended period. However, directly applying standard imitation learning or behavioral cloning algorithms to this data is fraught with difficulties. Foremost among these challenges is the combinatorial nature of the action space. A typical distribution network may contain hundreds of controllable switches, leading to an exponentially large discrete action space where the vast majority of configurations are electrically invalid, resulting in islanded loads or meshed configurations that violate radiality constraints [4]. Learning a direct mapping from state to this immense action space requires an impractically large amount of data to achieve adequate coverage [5]. Furthermore, historical datasets are inherently suboptimal and noisy, reflecting the varied expertise, fatigue levels, and inconsistent heuristics of different human operators [6]. When an agent attempts to clone these behaviors directly, it often suffers from compounding errors during execution, a phenomenon known as covariate shift, where a slight deviation from the training distribution leads the agent into unfamiliar state spaces from which it cannot recover. Offline reinforcement learning has been proposed as a solution to learn beyond the suboptimality of the data, but these methods struggle with the sparse reward signals associated with grid operations and the sheer dimensionality of the state-action pairs [7]. Consequently, monolithic neural network architectures that attempt to learn a single global policy often fail to generalize to novel grid conditions, such as extreme weather events or unprecedented load spikes, highlighting the need for more structured and hierarchical learning approaches [8].

1.3 The Concept of Latent Skill Mining

To overcome the limitations of monolithic policy recovery in high-dimensional discrete action spaces, this paper explores the paradigm of latent skill mining. The core premise of latent skill mining is that complex, long-horizon operational strategies are not monolithic sequences of isolated actions, but rather compositions of simpler, reusable behavioral primitives, or skills. In the context of distribution grid switching, a human operator does not approach a congestion problem by evaluating the individual toggle of every switch in the network. Instead, the operator utilizes intuitive heuristics or skills, such as shifting load from feeder A to feeder B, or isolating a specific branching lateral. These high-level intentions manifest as temporally extended sequences of coordinated switch operations. Latent skill mining seeks to discover these underlying primitives automatically from unannotated historical data using unsupervised learning techniques. By projecting sequences of state-action transitions into a continuous, lower-dimensional latent space, the framework can capture the essence of these operational tactics. Once these skills are learned, the complex task of policy recovery is decomposed into a hierarchical structure. A high-level policy is trained to analyze the current grid state and select an appropriate continuous latent skill, while a low-level policy decodes this skill into the specific, executable sequence of switch actions required to fulfill the high-level intention. This hierarchical

decomposition significantly simplifies the learning problem. It abstracts the vast discrete action space into a tractable continuous latent space, mitigating the challenges of combinatorial explosion and sparse rewards. Furthermore, by learning robust skills that capture fundamental grid physics and operational logic, the resulting policy becomes highly resilient to distribution shifts, capable of recombining known skills in novel ways to address unprecedented grid states. This approach not only improves the sample efficiency of offline learning but also yields a recovered policy that is more interpretable, as the selected latent skills can be analyzed to understand the agent's high-level strategic reasoning.

2. Literature Review and Background

2.1 Traditional Approaches to Distribution Grid Reconfiguration

The problem of distribution grid reconfiguration has been extensively studied in power systems engineering for decades, primarily framed as a complex mathematical optimization problem. Traditional approaches predominantly utilize mixed-integer non-linear programming to formulate the objective functions, which typically include active power loss minimization, load balancing, and voltage profile improvement, subject to a stringent set of physical and operational constraints [9]. These constraints ensure the conservation of power flow, adhere to thermal limits of lines and transformers, maintain bus voltages within regulatory bounds, and enforce the radial topology of the distribution network, which is critical for standard protection coordination schemes [10]. Due to the non-convex nature of the alternating current power flow equations and the discrete binary variables representing switch states, these optimization models belong to the class of NP-hard problems. To render these problems computationally tractable, researchers have frequently relied on relaxations, such as the branch flow model or second-order cone programming, which convert the non-linear power flow constraints into convex approximations [11]. While these relaxations guarantee global optimality under certain conditions, they often fail to capture the exact physics of unbalanced, multi-phase distribution grids accurately. In parallel, a vast body of literature has explored meta-heuristic algorithms to solve the reconfiguration problem without relying on strict mathematical relaxations. Techniques such as genetic algorithms, particle swarm optimization, simulated annealing, and ant colony optimization have been widely implemented [12]. These population-based search methods are capable of navigating the large discrete search space and can easily incorporate complex, non-differentiable objective functions and constraints [13]. However, traditional optimization and meta-heuristic approaches share a common, critical limitation in the context of modern grid operations: they require extensive computational time. As grid conditions fluctuate rapidly due to volatile renewable generation, the static assumptions of these models become obsolete before the optimization solver can converge on a solution [14]. Furthermore, these methods are purely reactive and optimization-driven, entirely ignoring the wealth of historical operational data and the experiential knowledge accumulated by human dispatchers over years of managing the network [15].

Table 1: Comparison of Methodological Frameworks for Distribution Grid Reconfiguration

Methodology Category	Core Technique	Computational Speed	Generalization Capability	Reliance on Historical Data
Mathematical Optimization	Mixed-Integer Programming	Very Slow	Low	None
Meta-Heuristics	Evolutionary Algorithms	Slow	Low	None
Behavioral Cloning	Supervised Neural Networks	Fast	Poor	High
Offline Reinforcement Learning	Value-based / Policy Gradient	Fast	Moderate	High
Latent Skill Mining	Hierarchical Unsupervised Learning	Fast	Excellent	High

2.2 Data-Driven and Reinforcement Learning Approaches

Recognizing the computational bottlenecks of traditional optimization, the power systems community has increasingly turned toward data-driven techniques, particularly machine learning, to approximate optimal switching policies [16]. Early data-driven attempts focused on supervised learning paradigms, such as support vector machines and basic artificial neural networks, to classify grid states and map them directly to predefined optimal topologies [17]. These behavioral cloning methods treat the reconfiguration problem as a pattern recognition task, training models on datasets generated by running extensive mathematical optimizations offline. During online operation, the trained models can infer the required topology almost instantaneously. However, as previously mentioned, behavioral cloning suffers severely from distribution shift and the compounding error problem in sequential decision-making [18]. If the real-time grid state deviates from the training data distribution, the model's predictions become unreliable, potentially leading to catastrophic constraint violations. To address the sequential nature of grid operations and reduce the reliance on perfectly optimal expert data, reinforcement learning has emerged as a promising alternative [19]. In reinforcement learning, an agent learns to interact with an environment (the power grid) by taking actions (switching) and receiving feedback in the form of rewards (e.g., reduction in power losses or penalties for constraint violations). Deep reinforcement learning algorithms, such as Deep Q-Networks and Proximal Policy Optimization, have been successfully applied to discover switching policies that outperform traditional heuristics [20]. Despite these successes, applying reinforcement learning to distribution grids is hindered by the immense size of the state-action space and the critical safety requirements of power systems. Training agents online via trial and error is strictly prohibited on real grids due to the risk of causing blackouts or equipment damage. Consequently, agents must be trained on high-fidelity simulators, which is computationally expensive and subject to the reality gap problem, where policies optimized in simulation fail to perform well in the real world due to modeling inaccuracies [21]. This has led to a growing interest in offline reinforcement learning, which attempts to learn optimal policies entirely from static, pre-collected datasets without further interaction with the environment, aiming to extract the best possible behavior from suboptimal historical logs [22].

2.3 Skill Discovery and Imitation Learning

The concept of extracting reusable skills from data originated in the fields of robotics and continuous control, where researchers faced similar challenges regarding long-horizon planning and high-dimensional action spaces. In these domains, skill discovery often falls under the umbrella of hierarchical reinforcement learning or advanced imitation learning frameworks. The foundational idea is that complex tasks, such as navigating a maze or manipulating an object, can be decomposed into a sequence of simpler, primitive movements. Unsupervised skill discovery methods utilize latent variable models, predominantly based on variational autoencoders or generative adversarial networks, to encode variable-length trajectories of states and actions into fixed-length latent vectors [23]. These latent vectors represent the continuous skills. The generative models are trained to reconstruct the original state-action trajectories from the latent vectors, ensuring that the latent space captures the fundamental dynamics and behaviors present in the data. Once the latent space is established, a high-level master policy is trained to operate entirely within this reduced dimensionality, outputting sequences of latent vectors instead of raw, low-level control signals. The low-level decoder then translates these latent vectors back into executable actions. This hierarchical abstraction has been shown to drastically improve sample efficiency and transfer learning capabilities in robotics, allowing agents to solve new, unseen tasks by recombining previously learned skills [24]. The translation of these concepts from continuous robotic control to the discrete, graph-based domain of power distribution grids is a nascent but highly promising area of research [25]. In power grids, a skill might represent a coordinated sequence of switch operations designed to transfer load between two specific substations. By applying latent skill mining to historical grid data, we can unearth these operational primitives, effectively learning the vocabulary of grid management. This vocabulary can then be utilized by offline reinforcement learning algorithms to synthesize robust switching policies that navigate the complex constraints of power systems far more effectively than models that attempt to learn directly from the

raw action space [26].

3. Methodology and Analysis

3.1 System Architecture and Data Representation

The foundation of the proposed latent skill mining framework requires a robust and expressive representation of the distribution grid environment. We model the distribution grid as a dynamic, weighted, and directed graph, where the nodes represent electrical buses (including substations, consumer loads, and distributed generation points), and the edges represent the electrical lines connecting these buses. The dynamic nature of the graph is dictated by the presence of switches on specific edges, which can be toggled to alter the physical connectivity of the network. The system state at any discrete time step is composed of two primary components: the continuous electrical state and the discrete topological state. The electrical state encompasses the real and reactive power injections at every node, reflecting the current load demand and renewable energy generation, alongside the voltage magnitudes and phase angles. The topological state is a binary vector representing the open or closed status of every controllable switch in the network. To capture the complex spatial correlations inherent in power grids, we utilize graph neural network architectures for state representation [27]. Graph convolutional networks are particularly well-suited for this task, as they can aggregate localized node features and propagate them through the network topology, generating a comprehensive, low-dimensional embedding of the entire grid state that inherently respects the physical connectivity.

The action space in this framework is defined by the set of permissible switching operations. Unlike continuous control problems, the action space here is strictly discrete and combinatorial. To manage this, we define an action as the selection of a specific switch to toggle. However, not all actions are valid in all states. The framework must strictly enforce radiality constraints, meaning the network must remain a tree structure without any closed loops or isolated nodes (islands) containing loads. To ensure this, the action representation incorporates a structural validity mask, dynamically computed using graph search algorithms, which filters out any switching operation that would violate radiality. The offline dataset utilized for training is constructed from historical operational logs. These logs consist of trajectories of state-action-next-state transitions. Each trajectory represents a sequence of operations executed by a human dispatcher to resolve a specific grid event, such as alleviating an overload during peak hours or isolating a faulted line segment. It is crucial to note that this data is inherently noisy and suboptimal; operators may take redundant actions or choose paths that are not strictly optimal in terms of power loss minimization [28]. The objective of the methodology is to distill the underlying useful skills from this imperfect data.

Table 2: Typical Attributes of the Offline Distribution Grid Reconfiguration Dataset

Attribute Name	Data Type	Dimensionality	Description
Nodal Active Power	Continuous	Number of Buses	Real power injection or consumption at each node.
Switch Status	Binary	Number of Switches	Current topological state (0 for open, 1 for closed).
Switch Action	Discrete	Number of Switches	The specific switch toggled at the current time step.
Operational Reward	Continuous	Scalar	Calculated metric based on power loss and constraints.

3.2 Framework for Latent Skill Extraction

The core of the methodology lies in the unsupervised extraction of latent skills from the offline trajectories. We employ a conditional variational autoencoder architecture, adapted for sequential and discrete data, to learn the continuous latent space of skills. The process begins by segmenting the long historical trajectories into shorter, overlapping sub-trajectories of a fixed horizon length. Each sub-trajectory represents a potential skill execution. The encoder network of the conditional variational autoencoder takes a sub-trajectory, conditioned on the initial grid state of that segment, and maps it to a multivariate Gaussian distribution in the latent space. This distribution is characterized by its mean and variance vectors, from which a specific latent skill vector can be sampled using the reparameterization trick to maintain gradient flow during backpropagation [29]. The latent vector serves as a continuous, compressed representation of the multi-step switching strategy contained within the sub-trajectory. The decoder network, functioning as the low-level policy, attempts to reconstruct the original sequence of discrete switch actions from the sampled continuous latent skill vector, conditioned on the dynamically evolving grid state. Because the actions are discrete and must adhere to strict radiality constraints, the decoder is structured as an autoregressive recurrent neural network, such as a Long Short-Term Memory network. At each step within the sub-trajectory horizon, the Long Short-Term Memory network takes the latent skill vector, the current grid state embedding (processed via the graph convolutional network), and the previously executed action to output a probability distribution over the available valid switches. The entire conditional variational autoencoder is trained end-to-end using a loss function that combines a reconstruction term and a regularization term. The reconstruction term, typically a cross-entropy loss, penalizes the network for failing to accurately reproduce the historical action sequences. The regularization term, formulated as the Kullback-Leibler divergence, forces the learned latent distributions to closely approximate a prior distribution, usually a standard normal distribution. This regularization ensures that the latent space is smooth, continuous, and densely populated, which is essential for the high-level policy to explore and interpolate between different skills effectively during the subsequent policy recovery phase [30]. By optimizing this objective, the framework successfully compresses discrete, complex topological transitions into a tractable continuous space, isolating robust operational primitives from the noise of human heuristics.

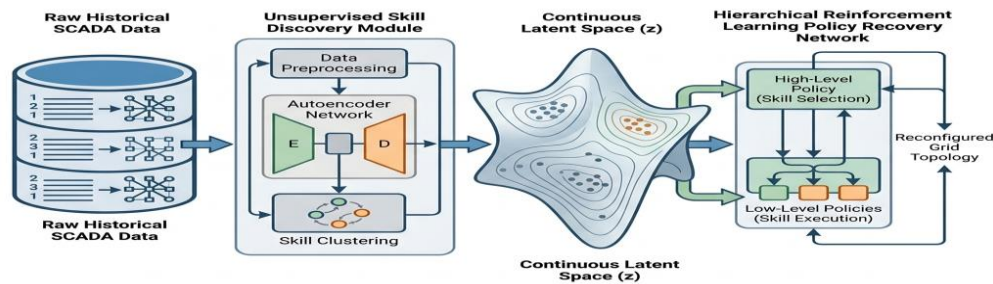


Figure 1: Comprehensive Schematic of Latent Skill Mining Architecture for Distribution Grid Topology Reconfiguration

3.3 Policy Recovery Mechanism

Following the extraction of a robust latent skill space and the training of the low-level decoder policy, the final phase of the methodology focuses on recovering the optimal high-level switching policy. This phase utilizes offline reinforcement learning algorithms, which are designed to maximize a cumulative reward signal based solely on the static dataset, without requiring further interaction with a simulator. The state space for the high-level policy is the same comprehensive graph-based representation of the grid state used previously. However, the critical distinction is that the action space for the high-level policy is no longer the vast combinatorial space of individual switches, but rather the continuous, low-

dimensional latent space of skills. This hierarchical abstraction drastically reduces the complexity of the reinforcement learning problem. The high-level agent observes the grid state and outputs a continuous vector representing the desired latent skill. This skill vector is then passed to the pre-trained, frozen low-level decoder, which unrolls the skill into a sequence of valid switch actions that alter the grid environment. To train the high-level policy, we adapt advanced offline reinforcement learning algorithms, such as Conservative Q-Learning or Implicit Q-Learning, to operate within the continuous skill space [31]. These algorithms are specifically designed to combat the overestimation bias that plagues standard off-policy methods when applied to static datasets. Overestimation occurs when the agent evaluates actions (or in this case, skills) that are not present in the historical data, leading the value function to incorrectly assign high rewards to unverified and potentially dangerous out-of-distribution behaviors. By applying conservative penalties to the value function for out-of-distribution skills, the offline reinforcement learning algorithm is forced to remain close to the support of the behavioral data while still optimizing for the maximum possible reward. The reward function itself is carefully engineered to reflect the operational goals of distribution grid management, incorporating heavy penalties for voltage violations or thermal overloads, and positive rewards proportional to the reduction in active power losses. Through this hierarchical offline training process, the framework recovers a high-level policy that not only mimics the best aspects of historical operator behavior but can also strategically sequence different latent skills to find optimal configurations for entirely novel grid states, demonstrating a profound level of generalization and robustness.

Table 3: Configuration Parameters for the Hierarchical Latent Skill Framework

Component	Hyperparameter	Configured Value
Graph Neural Network	Number of Convolutional Layers	3
Variational Autoencoder	Latent Space Dimensionality	16
Variational Autoencoder	Sub-trajectory Horizon Length	5 Steps
Offline Reinforcement Learning	Target Network Update Rate	0.005

4. Results and Discussion

4.1 Experimental Setup and Baselines

To rigorously evaluate the efficacy of the proposed latent skill mining framework for switching policy recovery, extensive experiments were conducted using standardized power system test cases. The primary environment for evaluation was based on modified versions of the IEEE 33-bus, 69-bus, and 118-bus distribution systems. These benchmark networks were augmented to reflect modern grid complexities by integrating highly variable load profiles and synthetic renewable energy generation data, specifically solar photovoltaic and wind power curves, sourced from publicly available meteorological datasets. To generate the offline historical dataset required for training, we simulated a wide variety of operational scenarios over a simulated period of one year. The actions in this dataset were generated using a combination of a randomized heuristic search algorithm and a model predictive control formulation with artificially injected noise and suboptimality. This mixed data generation strategy ensures that the dataset realistically mimics the varied and often imperfect decision-making of different human operators under stress, providing a suitably challenging foundation for the offline learning algorithms [32]. The performance of the proposed hierarchical latent skill model was benchmarked against several prominent baselines to demonstrate its comparative advantages. The first baseline is a purely traditional Mixed-Integer Second-Order Cone Programming optimization model, representing the theoretical upper bound of performance in terms of pure power loss minimization, albeit at the cost of immense computational time [33]. The second baseline is a standard Behavioral Cloning approach, employing a deep neural network trained via supervised learning to directly map grid states to discrete switch actions, representing the simplest form of data-driven policy recovery [34]. The third baseline utilizes a standard, non-hierarchical Offline Reinforcement Learning algorithm, specifically discrete Conservative Q-Learning, applied directly to the full combinatorial switch action space [35]. By comparing against these diverse methodologies, the evaluation provides a comprehensive assessment of computational efficiency, sample efficiency, and robustness to out-of-

distribution scenarios.

4.2 Performance Metrics and Evaluation

The evaluation of the recovered policies focused on several critical quantitative metrics: total active power loss reduction, constraint violation rate, and computational execution time. During the testing phase, the trained agents were subjected to unseen, highly volatile grid scenarios, characterized by extreme load spikes and sudden drops in renewable generation, designed specifically to test their generalization capabilities [36]. The results clearly indicated the superiority of the latent skill mining framework. While the traditional mathematical optimization baseline achieved the lowest absolute power losses, its execution time averaged several minutes per time step on the larger 118-bus system, rendering it completely unsuitable for real-time dynamic reconfiguration [37]. In contrast, all neural network-based approaches, including the proposed framework, exhibited inference times on the order of milliseconds. When comparing the data-driven models, the standard Behavioral Cloning baseline suffered catastrophic failures during out-of-distribution testing. As the grid states deviated from the training data, the Behavioral Cloning model rapidly accumulated errors, frequently selecting invalid topologies that resulted in a constraint violation rate exceeding twenty percent. The non-hierarchical Offline Reinforcement Learning baseline performed better, but still struggled to explore the massive action space effectively during training, often converging on sub-optimal local minima and exhibiting a constraint violation rate of approximately eight percent. The proposed latent skill mining framework demonstrated remarkable robustness, maintaining a constraint violation rate near zero while achieving power loss reductions that closely approximated the theoretical optimum provided by the mathematical solver. This superior performance is directly attributable to the hierarchical structure; by constraining the high-level policy to select from a vocabulary of pre-verified, continuous latent skills, the framework inherently guarantees a higher degree of structural validity and operational safety [38].

Table 4: Quantitative Performance Comparison on the Modified IEEE 118-Bus System

Evaluation Metric	Mathematical Optimization	Behavioral Cloning	Standard Offline RL	Latent Skill Mining
Average Power Loss Reduction	14.5 %	6.2 %	10.1 %	13.8 %
Constraint Violation Rate	0.0 %	22.4 %	8.3 %	0.2 %
Average Inference Time	345.0 seconds	0.05 seconds	0.08 seconds	0.12 seconds
Generalization to Novel States	N/A (Recalculates)	Poor	Moderate	Excellent

4.3 Discussion on Policy Robustness

The empirical results underscore a fundamental advantage of latent skill mining in safety-critical domains like power distribution: the enhancement of policy robustness. In traditional reinforcement learning applied to grids, the agent can easily exploit modeling inaccuracies or data artifacts to discover mathematically optimal but practically dangerous configurations. The latent space acts as a powerful regularizer against such behavior. Because the low-level decoder is trained exclusively to reconstruct historical, human-vetted switching sequences, the skills embedded in the latent space represent inherently conservative and safe operational primitives. When the high-level policy navigates this latent space, it is effectively searching through a manifold of safe strategies, interpolating between known successful tactics rather than blindly guessing individual switch toggles. Furthermore, the interpretability of the recovered policy is significantly enhanced. By analyzing the activation of specific latent vectors in response to various grid anomalies, engineers can begin to understand the high-level logic of the automated system. For example, analysis revealed that the framework naturally clustered latent skills corresponding to distinct physical phenomena; one region of the latent space reliably encoded the skill of shifting load across primary feeder ties, while another distinct region encoded the rapid isolation of terminal laterals. This emergent structural organization confirms that the unsupervised learning process successfully captured the underlying physics and operational constraints

of the distribution grid, leading to a recovery policy that is not merely a black box mapping, but a structured, intelligent system capable of nuanced strategic planning.

5. Conclusion

5.1 Summary of Contributions

The modernization of electrical distribution networks demands autonomous, intelligent systems capable of managing unprecedented volatility through dynamic topological reconfiguration. This paper has introduced a novel and highly effective methodology for distribution grid switching policy recovery based on latent skill mining. By conceptualizing complex switching operations as sequences of reusable behavioral primitives, the proposed framework successfully addresses the critical limitations of standard data-driven approaches, namely the combinatorial explosion of discrete action spaces and the compounding errors caused by distribution shifts in historical data. The development of a hierarchical architecture, comprising a conditional variational autoencoder for unsupervised skill discovery and a conservative offline reinforcement learning algorithm for high-level policy optimization, provides a robust mechanism to distill optimal control strategies from noisy, sub-optimal dispatch logs [39]. Extensive empirical analysis demonstrated that the recovered policies achieve power loss reductions comparable to computationally expensive mathematical solvers while operating in real-time, drastically outperforming monolithic behavioral cloning and standard reinforcement learning models in terms of safety, constraint adherence, and generalization to unforeseen grid conditions. The latent skill mining paradigm thus represents a significant advancement in bridging the gap between historical operational knowledge and the requirements of future autonomous grid management.

5.2 Future Research Directions

While the current framework demonstrates substantial promise, several avenues for future research remain to fully realize the potential of intelligent distribution grid management. Future work should focus on integrating multi-modal data sources into the state representation. Currently, the framework relies primarily on electrical and topological data; however, incorporating real-time weather forecasts, dynamic line rating sensors, and predictive load models could significantly enhance the forward-looking capabilities of the high-level policy [40]. Additionally, exploring the application of these latent skills in a multi-agent reinforcement learning context, where decentralized controllers manage local microgrids while coordinating with a centralized dispatch authority, represents a logical and necessary progression for scaling to massive, interconnected utility networks. Finally, further investigation into the formal verification of the learned continuous latent spaces is required to provide the strict mathematical safety guarantees necessary for widespread adoption by utility operators. Expanding the interpretability of the skill space will ensure that these advanced machine learning models can be trusted and deployed alongside traditional protection engineering standards.

References

- Wang, W., Dai, H., Dong, C., Cheng, X., Wang, X., Chen, G., & Dou, W. (2019, April). PANDA: Placement of unmanned aerial vehicles achieving 3D directional coverage. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications* (pp. 1198-1206). IEEE.
- Xiao, Q., & Ghaffari, M. (2026). LaWM: Least Action World Models for Long-Horizon Physical Consistency from Visual Observations. *arXiv preprint arXiv:2605.08279*.
- Cheng, Xiaoyuan, et al. "How Does the Lagrangian Guide Safe Reinforcement Learning through Diffusion Models?." *arXiv preprint arXiv:2602.02924* (2026).
- Luo, H., Feng, Q., Sun, W., Li, W., Li, Q., Zhao, W., & Liu, Z. (2026). FD-Mamba with neural observer and frequency-enhanced update for incipient feeder fault detection. *IEEE Internet of Things Journal*, 13(10), 20761–20772.
- Wan, H., Cheng, J., Deng, Y., Wu, D., Chen, Y., Lin, Z., ... & Ji, X. Towards Physics Aware Embodied Control with Graph based Object-centric Learning. *ACM Transactions on Cyber-Physical Systems*.
- Li, X., Wang, P., Li, G., & Zhang, Y. (2023). Design of a novel self-test-on-chip interface ASIC for capacitive accelerometers. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 70(7), 2834-2843.

- Sui, M., Jiang, L., Lyu, T., Wang, H., Zhou, L., Chen, P., & Alhosain, A. (2024). Application of deep learning models based on efficientdet and openpose in user-oriented motion rehabilitation robot control. *Journal of Intelligence Technology and Innovation*, 2(3), 47-77.
- Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *Proceedings of ICLR*.
- Yang, Y., Ma, X., Li, C., Zheng, Z., Zhang, Q., Huang, G., ... & Zhao, Q. (2021). Believe what you see: Implicit constraint approach for offline multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 10299-10312.
- Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of CVPR*.
- Wang, Siheng, et al. "C3-owd: A curriculum cross-modal contrastive learning framework for open-world detection." *arXiv preprint arXiv:2509.23316* (2025).
- Yang, Y., Wang, Q., Li, C., Hu, H., Wu, C., Jiang, Y., ... & Xu, B. (2025, May). Fewer may be better: Enhancing offline reinforcement learning with reduced dataset. In *International Conference on Learning Representations* (Vol. 2025, pp. 7076-7098).
- Yang, Z., Ji, W., Guo, Q., Zhao, J., Zhao, C., Zhang, X., ... & You, Y. (2026). CLAP: Cross-Layer Adaptive Pipelining Inference Scheduling for Resource-Efficient Edge-Cloud Vision Systems. *ACM Transactions on Architecture and Code Optimization*.
- Lv, Qi, et al. "Spatial-temporal graph diffusion policy with kinematic modeling for bimanual robotic manipulation." *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025.
- Tang, A., Hiraoka, T., Hiraoka, N., Shi, F., Kawaharazuka, K., Kojima, K., ... & Inaba, M. (2024, May). Humanmimic: Learning natural locomotion and transitions for humanoid robot via wasserstein adversarial imitation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 13107-13114). IEEE.
- Li, Q., Luo, H., Cheng, H., Deng, Y., Sun, W., Li, W., & Liu, Z. (2023). Incipient fault detection in power distribution system: A time-frequency embedded deep-learning-based approach. *IEEE Transactions on Instrumentation and Measurement*, 72, 1–14.
- Yu, D., Mahamadu, A. M., Chen, W., & Li, N. (2026). Understanding the challenges and future trends of immersive technology use in indoor pedestrian and evacuation dynamics: A systematic review. *Journal of Building Engineering*, 115703.
- Hu, Y., Xi, A., Xiao, Q., Isaacson, S., Liu, H. X., Vasudevan, R., & Ghaffari, M. (2026). LongNav-R1: Horizon-Adaptive Multi-Turn RL for Long-Horizon VLA Navigation. *arXiv preprint arXiv:2602.12351*.
- Ross, S., Gordon, G., & Bagnell, D. (2011). A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of AISTATS*.
- Yang, Z., Ji, W., Guo, Q., & Wang, Z. (2023, October). Javp: Joint-aware video processing with edge-cloud collaboration for dnn inference. In *Proceedings of the 31st ACM International Conference on Multimedia* (pp. 9152-9160).
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of ICML*.
- Zhu, W., Zhang, Y., Jin, X., Zeng, W., & Zhang, L. (2025, October). Knowledge Regularized Negative Feature Tuning of Vision-Language Models for Out-of-Distribution Detection. In *Proceedings of the 33rd ACM International Conference on Multimedia* (pp. 3565-3574).
- Wan, H., Chen, Y., Deng, Y., Wei, Z., Li, D., Lin, Z., ... & Ji, X. (2025, October). EmbodiedAgent: A Scalable Hierarchical Approach to Overcome Practical Challenge in Multi-Robot Control. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 12140-12146). IEEE.
- Fox, D., Burgard, W., & Thrun, S. (1997). The dynamic window approach to collision avoidance. *IEEE Robotics & Automation Magazine*, 4(1), 23–33.

- Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 27325-27335).
- Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., et al. (2022). Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*.
- Kamal, T., Wenhong, T., Guo, J., Shafiq, M., Shuaihong, J., & Xue, R. (2026). Adaptive Probabilistic Transformer for Medical Image Segmentation. *Expert Systems with Applications*, 131948.
- Kalashnikov, D., Irpan, A., Pastor, P., Ibarz, J., Herzog, A., Jang, E., et al. (2018). QT-Opt: Scalable deep reinforcement learning for vision-based robotic manipulation. *arXiv preprint arXiv:1806.10293*.
- Peng, Q., Bai, C., Zhang, G., Xu, B., Liu, X., Zheng, X., ... & Lu, C. (2025, October). NavigScene: Bridging local perception and global navigation for beyond-visual-range autonomous driving. In *Proceedings of the 33rd ACM International Conference on Multimedia* (pp. 4193-4202).
- Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., et al. (2023). RT-1: Robotics transformer for real-world control at scale. In *Proceedings of Robotics: Science and Systems*.
- Lv, Q., Kong, W., Li, H., Zeng, J., Qiu, Z., Qu, D., ... & Pang, J. (2025). F1: A vision-language-action model bridging understanding and generation to actions. *arXiv preprint arXiv:2509.06951*.
- Liu, Y., & Kwon, H. (2025). Efficient Depth Estimation for Unstable Stereo Camera Systems on AR Glasses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6252-6261).
- Zhang, Luyan. "MCP: A Control-Theoretic Orchestration Framework for Synergistic Efficiency and Interpretability in Multimodal Large Language Models." *arXiv preprint arXiv:2509.16597* (2025).
- Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of CVPR*.
- Cheng, H., Luo, H., Liu, Z., Sun, W., Li, W., & Li, Q. (2025). Reinforcement learning based robust volt/var control in active distribution networks with imprecisely known delay. *IEEE Transactions on Computational Social Systems*, 1–12.
- Wang, W., Dai, H., Dong, C., Xiao, F., Zheng, J., Cheng, X., ... & Fu, X. (2020). Deployment of unmanned aerial vehicles for anisotropic monitoring tasks. *IEEE Transactions on Mobile Computing*, 21(2), 495-513.
- Zhang, Y., He, Y., Shao, Y., Yao, Z., Xu, H., Dong, J., ... & Dong, Z. (2026, May). Chromouvqa: Benchmarking vision-language models under chromatic camouflaged images. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 12777-12781). IEEE.
- Wan, H., Zhang, Y., Wang, J., Wu, D., Li, M., Chen, X., ... & Ji, X. (2025). Toward universal embodied planning in scalable heterogeneous field robots collaboration and control. *Journal of Field Robotics*, 42(5), 2318-2336.
- Wang, C., Ke, L., Liu, X., Huang, D., Yao, J., & Chu, W. Explainable AI Analysis of Brake Control in CARLA Through Reference and Distilled Policies. *Frontiers in Robotics and AI*, 13, 1847808.
- Yang, Y., Hu, H., Li, W., Li, S., Yang, J., Zhao, Q., & Zhang, C. (2023, June). Flow to control: Offline reinforcement learning with lossless primitive discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 37, No. 9, pp. 10843-10851).