When Does Structure Help? Inductive Bias under Structural and Semantic Distribution Shift in Learned Control

Anonymous authors
Paper under double-blind review

Abstract

Learned controllers for structured dynamical systems must often generalize after deployment, when either the system's interaction structure changes or the processes acting on that structure change. These shifts are distinct, yet they are usually evaluated together. As a result, it is unclear whether an architectural inductive bias improves out-of-distribution (OOD) control because it matches the source of the shift, or merely because it matches the visible structure of the system. We separate OOD generalization into two axes: structural shift, which changes system topology, and semantic shift, which changes exogenous workload distributions while preserving topology. Crossing them yields four controlled settings: in-distribution, structural OOD, semantic OOD, and combined OOD. In a physics-grounded networked-control testbed, we compare a Transformer with fixed component slots, the same Transformer with slot identity removed to enforce permutation invariance, and a graph neural controller that explicitly encodes the physical interaction graph. All policies are trained under a common proximal policy optimization (PPO) protocol and evaluated over 20 training seeds, five held-out topologies, and measured AI-workload traces. In a weakly constraint-active regime, explicit graph structure provides no statistically supported advantage over permutation invariance under structural OOD, while domain randomization does not reliably improve generalization to the measured semantic shift. We then test whether graph structure becomes useful when topology is made directly relevant to physical safety. Using a pre-specified, model-independent calibration procedure, we construct a feasible constraint-active regime without using learned-policy outputs or screening the final OOD panel. Even in this regime, the physical-graph GNN does not significantly outperform the permutation-invariant Transformer, and the correct graph does not significantly outperform a rewired graph. These results show that matching an architecture to visible physical structure is not, by itself, sufficient to establish robust OOD control. Our contribution is a factorized framework for structural and semantic OOD, together with controlled evidence that structural inductive bias should be evaluated rather than assumed to provide robustness.

1 Introduction

Learned controllers are increasingly used in systems whose behavior is shaped by both dynamics and structure. Generalization beyond the training environment is already a recognized challenge in deep reinforcement learning (Packer et al., 2019; Cobbe et al., 2019), and structured physical systems add another source of variation through their interaction topology. Examples include networked infrastructure, multi-agent systems, transportation networks, and distributed energy systems. In these settings, deployment rarely reproduces the training environment exactly. The interaction graph may change, components may appear or disappear, or the statistical behavior of exogenous inputs may shift. A controller that performs well in-distribution can therefore face several qualitatively different forms of out-of-distribution (OOD) change after deployment.

These shifts are often treated as a single generalization problem. We argue that this obscures an important distinction. A *structural shift* changes the organization of the system itself, such as its

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

topology or component configuration. A *semantic shift* changes the processes acting on that structure, such as demand or workload distributions, while leaving the interaction graph fixed. These shifts stress different properties of a learned controller. A representation that is invariant to component ordering may help when identities or configurations change (Zaheer et al., 2017; Lee et al., 2019), while an explicit graph representation introduces a stronger relational inductive bias tied to interactions among components (Battaglia et al., 2018). Neither property, however, necessarily addresses a change in the semantics of the exogenous process.

This distinction raises a more basic question: *when does an inductive bias actually help OOD generalization?* In structured control, it is tempting to assume that graph-structured systems should benefit from graph neural policies. But the existence of a physical graph does not imply that graph information is decision-relevant to the learned control problem. If topology has little effect on the active constraints or the action consequences encountered during training and evaluation, a graph prior may add little beyond a permutation-invariant representation. Conversely, even when topology matters physically, explicit graph encoding may still fail to provide a measurable advantage if the learned policy does not exploit that information in a useful way.

We study this question through a factorized OOD evaluation framework. We cross structural shift and semantic shift to form four conditions: in-distribution, structural OOD, semantic OOD, and combined OOD. We compare three policy classes that encode progressively stronger structural assumptions: a fixed-slot Transformer (Fixed-TF), a permutation-invariant Transformer (PI-TF), and a graph neural network (GNN) controller that explicitly uses the physical interaction graph (Grid-GNN). Here, TF denotes Transformer and PI denotes permutation-invariant. All learned policies are trained with the same proximal policy optimization (PPO) procedure (Schulman et al., 2017) so that the main difference is the representation and inductive bias rather than the optimization algorithm.

We instantiate the framework in a physics-grounded networked-control testbed based on three-phase distribution-grid dynamics. The domain is useful because it provides explicit physical topology, nonlinear interactions, measurable safety constraints, and realistic opportunities to vary structure and exogenous demand independently. For semantic shift, we use measured AI-workload traces to alter workload behavior while keeping physical topology fixed. The goal is not to propose a domain-specific microgrid controller, but to use the environment as a controlled setting for diagnosing OOD generalization in structured learned control.

Our empirical results are deliberately mixed. In a nominal, weakly constraint-active regime, the explicit physical-graph GNN does not obtain a statistically supported advantage over the permutation-invariant Transformer on structural OOD, and domain randomization does not reliably improve generalization to the measured semantic shift. This motivates a stronger test of the graph hypothesis. Rather than selecting a difficult regime based on model performance, we introduce a pre-specified, model-independent physics calibration procedure that creates a constraint-active regime while keeping the final OOD topology panel untouched during benchmark construction.

The resulting constraint-active experiment provides the central diagnostic result of the paper. Across 20 training seeds, Grid-GNN-full has a voltage-violation rate 1.43 percentage points higher than the permutation-invariant Transformer in the registered paired comparison; this difference is not statistically supported after Holm correction. Relative to a deterministically rewired graph control, Grid-GNN-full is numerically better by 0.68 percentage points, but this comparison is also not statistically supported. Thus, even after constructing a regime specifically designed to make topology consequential to physical safety, explicit physical graph encoding is not sufficient to establish improved OOD control.

Our contributions are:

1. **A factorized OOD framework for structured learned control.** We separate structural and semantic distribution shift and evaluate them independently and jointly rather than treating OOD generalization as a single axis.
2. **A controlled inductive-bias comparison.** We compare fixed-slot, permutation-invariant, and graph-structured policies under a common PPO protocol, including graph controls that distinguish physical adjacency, edge information, and rewired policy graphs.
3. **A model-independent test of structural decision relevance.** We introduce a pre-specified physics calibration procedure that activates topology-dependent safety constraints without

using learned-policy outputs or screening the final evaluation panel. The resulting evidence shows that matching a policy architecture to visible physical structure does not by itself guarantee OOD robustness.

Together, these results suggest that structural inductive bias should be treated as an empirical hypothesis about the source of deployment shift, not as a default guarantee of generalization.

2 Related Work

Generalization and robustness in reinforcement learning. Deep-RL generalization work has emphasized the need to distinguish training performance from performance on unseen environments. Packer et al. (2019) propose a controlled protocol covering both in-distribution and OOD generalization, while Cobbe et al. (2019) use procedurally generated train/test sets to measure overfitting in RL. Domain randomization exposes a learner to a distribution of environments during training (Tobin et al., 2017); in microgrid control, Halev et al. (2024) combine RL with domain randomization under battery-model uncertainty. These approaches motivate training across variation, but they do not by themselves distinguish a change in interaction structure from a change in the semantics of exogenous inputs. We make that distinction explicit and evaluate zero-shot structural, semantic, and combined shift separately.

Permutation invariance and graph inductive bias. Set-based models encode invariance to input ordering without requiring an explicit interaction graph (Zaheer et al., 2017; Lee et al., 2019). Graph networks instead encode relational structure directly, providing a stronger inductive bias when neighborhood interactions are relevant (Battaglia et al., 2018). Our Fixed-TF, PI-TF, and Grid-GNN comparison forms an inductive-bias ladder: fixed component identity, permutation invariance, and explicit physical graph structure. The goal is not a universal ranking of these representations, but a controlled test of whether the additional structural prior yields measurable OOD benefit under the shift it is intended to address.

Graph reinforcement learning for power-system control. Graph-based RL and topology-aware voltage control are already active areas, so our contribution is not the first use of a GNN in a power-system or microgrid controller. Donon et al. (2024) train a topology-aware graph policy for tertiary voltage control under changing grid configurations. Luo et al. (2025) incorporate topology, electrical characteristics, and source–load relationships into graph-attention MARL for Volt/Var control. Recent studies further report topology-aware graph policies for microgrid voltage-reactive-power control (Zhang et al., 2026) and topology robustness in GNN-based multi-agent active-voltage control (Eze et al., 2026). These works primarily ask how to exploit network structure for control. We ask a diagnostic question: when does access to the correct physical graph provide an advantage over a permutation-invariant controller, and does that advantage survive a deliberately corrupted policy graph?

Energy-control benchmarks and adaptation. PowerGym provides physics-based RL environments for Volt/Var control across multiple distribution systems (Fan et al., 2022), while CityLearn provides a widely used benchmark for learning-based control in grid-interactive buildings and communities (Nweye et al., 2022). Recent energy-management work also studies adaptation across unseen buildings through meta-RL (Rajendran & Gebremedhin, 2026). Adaptation and zero-shot generalization answer different questions: meta-RL can update a policy using target-task experience, whereas our evaluation keeps the trained controller fixed when structure or workload semantics change. Our contribution is therefore a factorized zero-shot evaluation protocol, not a claim that existing energy benchmarks lack generalization capability.

3 Problem Formulation

3.1 Structured learned control under distribution shift

We consider a family of control environments indexed by two factors: a structural variable G and a semantic variable ξ. The structural variable specifies the interaction structure of the system, such

as graph connectivity or component configuration. The semantic variable specifies exogenous processes acting on the system, such as workload or demand distributions.

A controller is trained on a source environment (G_s, ξ_s) and evaluated without adaptation under environments in which either or both factors may change. This produces four evaluation conditions:

- **In-distribution (ID):** (G_s, ξ_s)
- **Structural OOD:** (G_t, ξ_s), where $G_t \neq G_s$
- **Semantic OOD:** (G_s, ξ_t), where $\xi_t \neq \xi_s$
- **Combined OOD:** (G_t, ξ_t)

This factorization is an evaluation device rather than an assumption that structure and semantics are statistically independent. Its purpose is to isolate which form of deployment change a controller is robust to.

3.2 Inductive biases

We compare policies with different assumptions about component identity and interaction structure.

Fixed-TF. A Transformer receives component observations together with fixed slot identity. The representation is tied to the source component ordering.

PI-TF. The same Transformer architecture is used without slot identity. Its component wise actions are permutation equivariant, and its pooled value is invariant when observations and action-type masks are permuted consistently.

Grid-GNN. A graph neural policy receives the physical interaction graph and performs message passing over the registered network structure. The full model uses physical adjacency and line attributes. Two controls are used in the constraint-active study: an adjacency-only model that retains physical adjacency and sets both edge-attribute channels to zero, and a rewired-graph model that permutes graph endpoints while leaving node observations, action mappings, and simulator topology unchanged.

The comparison therefore separates three increasingly structured assumptions:

$$\text{fixed identity} \;\rightarrow\; \text{permutation invariance} \;\rightarrow\; \text{explicit physical graph.}$$

Figure 1 summarizes the architectures; Appendix A gives implementation details.

3.3 Evaluation objective

Our primary question is not whether one policy class has the highest average return in a single environment. We ask whether a structural inductive bias improves zero-shot generalization under a shift that changes the corresponding structure.

In the constraint-active study, the preregistered primary outcome is voltage-violation rate: the fraction of 30 building-phase voltages outside [0.95, 1.05] p.u., averaged over each 96-step episode. Return, energy cost, equity, physics penalty, voltage excursion, voltage unbalance, severe violations, and power-flow failures are reported as secondary outcomes so that safety improvements are not interpreted without their control-performance consequences.

4 Experimental Framework

4.1 Physics-grounded networked-control testbed

We use a centralized learned controller operating a heterogeneous distribution-grid environment with three-phase asymmetric AC power flow. Existing benchmark efforts such as PowerGym (Fan et al., 2022) and CityLearn (Nweye et al., 2022) demonstrate the value of standardized learning

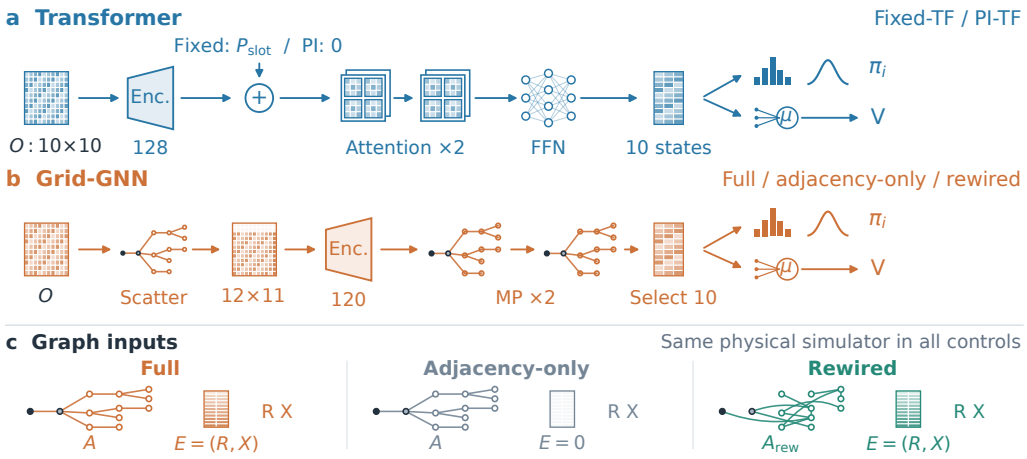

Figure 1: Policy architectures and graph controls. Fixed-TF and PI-TF differ only in slot embeddings. Grid-GNN applies two message-passing layers to bus-mapped observations. Full uses physical adjacency and line attributes; adjacency-only zeros the attributes; rewired permutes edge endpoints. The simulator is unchanged. Fixed-TF is not evaluated in the constraint-active study.

Table 1: Core frozen environment and training settings.

Setting	Frozen value
Controller	Centralized learned controller
Episode length	96 steps
Controlled quantity	Flexible-demand modulation
Component state	10 features
Voltage band	0.95–1.05 p.u.
VUF reporting threshold	2%
Equity reward weight	1
Physics reward weight	100
Main train seeds	20

environments for power and building control; our testbed is constructed specifically to isolate structural and semantic OOD axes. The controller acts on flexible demand at controllable nodes while the simulator determines the resulting electrical state.

The environment uses 96-step episodes, a 10-feature component state, and a common centralized action interface across model classes. The control variables in the present experiments are flexible-demand modulation actions; we do not claim battery or HVAC dynamics that are not represented in the frozen environment. Voltage safety is defined by the 0.95–1.05 p.u. band and the voltage unbalance factor (VUF) reporting threshold is 2%. The frozen reward combines energy cost, equity, and a physics penalty, with equity weight 1 and physics weight 100.

4.2 Semantic shift from measured workload traces

Semantic OOD changes workload behavior without altering physical topology. We use the public National Laboratory of the Rockies (NLR) Dataset of Generative AI Workload Power Profiles (Vercellino et al., 2026), which contains high-resolution measured power traces from generative-AI workloads executed on the Kestrel HPC platform. The frozen semantic panel uses measured Llama-family inference/training and Stable Diffusion training traces. Each selected trace is pooled into 96 equal workload-progress bins, converted from watts to kilowatts, and energy-normalized to the source daily energy before being used as a workload-shape intervention.

The resulting 96-point series are therefore *normalized workload-progress templates*, not measured quarter-hour campus-meter data. They provide external evidence for workload-family shape varia-

tion while preserving a clear boundary between measured compute-power behavior and the simulated campus-control environment.

4.3 Common PPO training

All learned controllers use the same PPO implementation (Schulman et al., 2017) and centralized action interface. The frozen configuration uses learning rate 3×10^{-4}, discount factor 0.99, generalized advantage estimation (GAE) parameter 0.95, clipping coefficient 0.20, value clipping 0.20, at most 10 optimization epochs, minibatch size 32, entropy coefficient 0.01, value coefficient 0.5, gradient norm 1, and target Kullback-Leibler (KL) 0.03. Continuous actions use a tanh-transformed Normal policy.

4.4 Nominal factorized-OOD study

The nominal study, referred to as V3, evaluates structural and semantic shift separately and jointly. Structural OOD uses five held-out electrical topologies. Semantic OOD changes workload behavior while preserving topology, including the measured workload panel above. The study also includes domain randomization and a rewired-graph control.

The registered V3 analysis uses the training seed as the inferential unit, paired bootstrap confidence intervals, exact two-sided paired sign-flip tests, and Holm correction across three hypotheses. H1 compares Grid-GNN-full with PI-TF under topology OOD; H2 compares Grid-GNN-full with the rewired graph control; H3 compares domain-randomized PI-TF with PI-TF under the measured semantic shift. Voltage violations and power-flow failures are essentially absent in the nominal regime and VUF remains below threshold, so V3 tests generalization in a setting where the physical graph is available but the principal safety constraints are weakly active.

4.5 Model-independent constraint activation

A potential explanation for the nominal null result is that topology is not sufficiently consequential to the control objective. We therefore design a separate constraint-active study in which benchmark construction is completed before any learned-policy comparison is opened.

The final protocol separates three sets of topologies:

1. source candidates used to select a feasible training source;
2. calibration-only topologies used to determine whether a stress level activates topology-sensitive voltage constraints;
3. an untouched final evaluation panel reserved for the OOD claim experiment.

The calibration uses neutral and maximum-curtailment open-loop actions and does not use learned checkpoints, policy outputs, or the final evaluation panel. The first admissible source/stress pair under the frozen selection rule is then fixed for training. The procedure selects source topology `V6_S2` at stress multiplier 8.0. At this stress level, the calibration panel is constraint-active while remaining controllable under maximum curtailment. The final panel is checked structurally but is not power-flow screened during calibration; learned-policy evaluation follows the source-only pilot.

Two earlier pre-specified calibration attempts stop before training because no admissible regime is found. We retain these stops in the experimental record rather than modifying their gates after observing the calibration outcomes.

4.6 Constraint-active claim experiment

The final constraint-active study, referred to as V6, uses 20 train seeds, 100 training episodes per seed, source topology `V6_S2`, stress multiplier 8.0, five untouched final evaluation topologies, and ten deterministic evaluation scenarios per topology. Four frozen policy arms are evaluated: PI-TF, Grid-GNN-full, Grid-GNN-adjacency-only, and Grid-GNN-rewired. The synthetic in-distribution workload process is held fixed across source and held-out topologies, so this study tests structural OOD.

H1. Grid-GNN-full has lower voltage-violation rate than PI-TF under held-out constraint-active topologies.

H2. Grid-GNN-full has lower voltage-violation rate than Grid-GNN-rewired.

The reporting unit is the training seed. Within each train seed, the primary metric is averaged over all five final topologies and ten evaluation scenarios. We use paired bootstrap 95% confidence intervals, exact two-sided paired sign-flip tests, and Holm correction across H1 and H2. Source evaluation episodes are excluded from these contrasts. Confidence intervals quantify training seed variation conditional on the five fixed topologies and ten scenarios

5 RESULTS

5.1 NOMINAL STRUCTURAL AND SEMANTIC OOD

In the nominal regime, none of the three registered V3 hypotheses is supported. For H1, the mean Grid-GNN-full minus PI-TF difference on topology OOD is approximately -1.25×10^4 episode-return units; the paired bootstrap interval includes zero and the Holm-adjusted p-value is 1.00. Thus, explicit physical graph encoding does not provide a supported advantage over permutation invariance in this regime.

For H2, Grid-GNN-full is directionally better than the deterministically rewired graph control by approximately 1.30×10^5 episode-return units, but the comparison remains inconclusive (Holm-adjusted $p = 0.295$). H3 likewise provides no supported improvement from domain randomization over PI-TF under the measured AI-workload semantic shift (Holm-adjusted $p = 1.00$). The registered paired bootstrap intervals for the primary V3 comparisons include zero.

These findings separate two issues. First, structural and semantic OOD are not interchangeable: graph-aware representation does not automatically address changed workload semantics. Second, a graph-structured simulator does not by itself establish a benefit from a graph-biased policy. Because the nominal safety regime is weakly active, however, V3 does not resolve whether graph information becomes useful once topology materially affects physical safety. That ambiguity motivates the separately calibrated constraint-active study.

5.2 CONSTRAINT-ACTIVE BENCHMARK CALIBRATION

The model-independent calibration freezes source `V6_S2` at stress multiplier 8.0 before learned-policy evaluation. On the calibration topology panel, the neutral-policy mean voltage-violation rate is approximately 6.16%, all five calibration topologies show nonzero violations, and the maximum-curtailment mean violation rate remains 0%. The source itself has a neutral voltage-violation rate of approximately 5.82% with 0% under maximum curtailment. The final evaluation topologies are not screened during this process.

This establishes a regime in which voltage constraints are active, topology produces heterogeneous physical consequences, and the controller retains the physical ability to reduce violations.

5.3 EXPLICIT GRAPH STRUCTURE DOES NOT OUTPERFORM PERMUTATION INVARIANCE

All 80 frozen arm-by-seed cells complete successfully, yielding 4,800 registered evaluation observations and no evaluation power-flow failures. Table 2 summarizes the primary V6 safety result.

For H1, Grid-GNN-full has a voltage-violation rate **1.43 percentage points higher** than PI-TF on average across the frozen final panel. The 95% paired bootstrap confidence interval is approximately $[-0.77, 4.15]$ percentage points. The exact two-sided paired sign-flip p-value is 0.335, and the Holm-adjusted p-value is 0.671. H1 is therefore not supported.

This result directly addresses the main ambiguity in V3. Topology now affects active physical safety constraints, yet explicit graph encoding still does not produce a statistically supported advantage over a permutation-invariant policy.

Table 2: Constraint-active V6 voltage-violation results. Arm values are means across train-seed means on the five held-out topologies; hypothesis intervals and tests are paired at the train-seed level.

Arm / comparison	Violation rate or effect	95% bootstrap CI
PI-TF	2.13%	[1.13, 3.37]%
Grid-GNN-full	3.56%	[1.89, 5.80]%
Grid-GNN-adj-only	3.13%	[1.33, 5.32]%
Grid-GNN-rewired	4.24%	[2.46, 6.30]%
H1: full − PI-TF	+1.43 pp	[-0.77, +4.15] pp
H2: full − rewired	-0.68 pp	[-2.16, +0.59] pp

5.4 Correct physical graph vs. rewired graph

For H2, Grid-GNN-full is numerically better than Grid-GNN-rewired by approximately **0.68 percentage points** in voltage-violation rate. The 95% paired bootstrap interval is approximately $[-2.16, 0.59]$ percentage points. The exact paired sign-flip p-value is 0.374, with Holm-adjusted $p = 0.671$. H2 is therefore not supported.

The direction of the effect is consistent with a possible benefit of the correct graph relative to an incorrect graph, but the evidence is insufficient to claim a reliable mechanism. We therefore avoid interpreting the nonsignificant result as either proof that topology is irrelevant or proof that the models are equivalent.

5.5 Secondary graph controls and performance trade-offs

The adjacency-only Grid-GNN is a registered secondary graph control rather than a third confirmatory hypothesis. Its mean voltage-violation rate is 3.13%, compared with 3.56% for Grid-GNN-full, 4.24% for the rewired model, and 2.13% for PI-TF. These arm-level values are descriptive; we do not promote the adjacency-only comparison to a post-hoc confirmatory test. Grid-GNN-full has numerically higher return and lower energy cost than the rewired model, but a higher equity penalty (weighted dissatisfaction; 521.42 versus 448.88). Relative to PI-TF, it has lower return and higher cost and equity penalty. These comparisons are descriptive; return combines cost and penalties and is not independent evidence of a safety benefit. Appendix B reports all secondary outcomes and per-topology results.

6 Discussion

6.1 A graph-structured environment does not guarantee a graph-model advantage

The main finding is not that graph structure is unimportant. Rather, the experiments show that *the presence of a physical graph is not sufficient to establish a benefit from a graph inductive bias*. This distinction matters because graph-based controllers are often motivated by the structural correspondence between a networked environment and a message-passing architecture. Our results show that architectural correspondence alone does not guarantee better zero-shot OOD control.

This conclusion survives a stronger test than the nominal benchmark. After the nominal graph advantage is unsupported, we construct a separate constraint-active regime without using model performance to select the benchmark. The full GNN still does not significantly outperform permutation invariance, and the correct graph still does not significantly outperform a rewired graph.

6.2 Structural and semantic OOD are distinct

The semantic-shift results reinforce the value of factorizing OOD. Permutation invariance and graph structure concern how the policy represents components and interactions. They do not directly solve a change in the statistical behavior of exogenous workloads. Conversely, domain randomization aimed at workload variability does not imply robustness to topology changes.

The practical implication is that "OOD robustness" should not be treated as a single scalar property of a controller. A policy can be invariant to one source of change while remaining brittle to another.

6.3 What the null results do and do not establish

The constraint-active study provides no statistically supported evidence that Grid-GNN-full improves OOD voltage safety over PI-TF under its frozen evaluation protocol. It also provides no statistically supported evidence that the correct graph improves over the rewired graph.

These findings should not be read as an equivalence result. The experiments do not prove that graph information is never useful, that topology is irrelevant, or that graph and permutation-invariant policies behave identically. The H2 point estimate is directionally favorable to the correct graph, but uncertainty remains too large for a mechanism claim.

A more cautious conclusion is that graph inductive bias should be treated as a testable hypothesis about a control problem rather than a default design rule.

6.4 Benchmark construction without model selection

A secondary methodological contribution is the calibration procedure itself. If benchmark difficulty is chosen after inspecting which model performs better, an apparent architectural advantage can be entangled with benchmark selection. We instead select the constraint-active regime using open-loop physics and reserve the final evaluation topologies until after the regime and source-learning gate are frozen.

The earlier stopped calibration attempts are useful in this context. They show that satisfying activity, feasibility, controllability, and topology sensitivity simultaneously is nontrivial. Reporting these stops makes clear that the final regime was not chosen simply because it produced the desired model ordering.

7 Limitations

The experimental system is a controlled simulation rather than a real deployed grid. Although the semantic shift uses measured compute-workload traces, those traces are not campus-meter measurements and come from a limited measurement context. The constraint-active V6 study uses a synthetic network construction designed to support controlled topology intervention.

The study also evaluates a deliberately small architecture family. The results do not imply that all GNN architectures would behave similarly, nor do they establish equivalence between graph and permutation-invariant policies. We intentionally avoid post-hoc architecture search after opening the final OOD panel.

Finally, the final statistical conclusions are specific to the frozen training budget, topology family, source regime, and evaluation protocol. The goal is diagnostic evidence about inductive bias under controlled shift, not a universal ranking of policy classes.

8 Conclusion

We study OOD generalization in structured learned control by separating two deployment changes that are often conflated: structural shift and semantic shift. This factorization lets us ask which inductive biases address which failure modes.

Across a nominal regime and a separately calibrated constraint-active regime, explicit physical graph encoding does not produce a statistically supported advantage over permutation invariance on structural OOD. In the constraint-active study, the correct graph is numerically better than a rewired graph but not significantly so. Semantic workload shift remains a distinct generalization problem and is not reliably resolved by domain randomization in our nominal study.

The resulting lesson is simple: *matching a model to the visible structure of a system does not guarantee robust OOD control*. Structural inductive bias should be evaluated against the actual source of deployment shift and the decision problem it is intended to solve.

Reproducibility Statement

We provide the experimental design, model definitions, training protocol, seed structure, statistical procedures, and benchmark-construction rules in the main paper and appendix. The supplementary material will include the frozen experiment configurations, topology manifests, execution reports, analysis outputs, and source code needed to reproduce the reported results. The constraint-active benchmark was calibrated without learned-policy outputs, and the final OOD panel was reserved until the preregistered source-learning gate had passed. Exact software versions, seeds, and protocol hashes are retained in the reproducibility archive.

Ethics Statement

This work studies learned control in a simulated networked energy-system environment and does not involve human subjects or personal data. Measured compute-workload traces are used only as aggregate workload-shape inputs to the simulator. The experiments are intended for methodological evaluation of OOD generalization rather than direct deployment guidance for safety-critical infrastructure. We therefore avoid claims of real-grid validation and report the limitations of the controlled simulation explicitly.

AI Use Statement

Generative AI tools were used to assist with research ideation and experimental methodology design, literature discovery and synthesis, creation and review of research code and experiment artifacts, and drafting and editing portions of the manuscript. AI tools also assisted with figure design and the code used to render architecture diagrams and plot the completed experimental results. AI-assisted code and experimental artifacts were checked through frozen-protocol validation, hash verification, reproducibility checks, and direct inspection of the resulting outputs. AI-assisted literature discovery is not treated as a substitute for source verification; bibliographic claims and citations included in the final manuscript are reviewed by the authors against the underlying sources. All experimental protocols, analyses, claims, and manuscript content remain the responsibility of the authors.

References

Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George E. Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks, 2018.

Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. Quantifying generalization in reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1282–1289. PMLR, 2019.

Balthazar Donon, François Cubelier, Efthymios Karangelos, Louis Wehenkel, Laure Crochepierre, Camille Pache, Lucas Saludjian, and Patrick Panciatici. Topology-aware reinforcement learning for tertiary voltage control. *Electric Power Systems Research*, 234:110658, 2024. doi: 10.1016/j.epsr.2024.110658.

Chijioke Eze, Chetan Prakash, and Antonello Monti. Graph neural network-based multi-agent reinforcement learning for active voltage control: Performance and topology robustness. *Energy and AI*, 25:100835, 2026. doi: 10.1016/j.egyai.2026.100835.

Ting-Han Fan, Xian Yeow Lee, and Yubo Wang. Powergym: A reinforcement learning environment for volt-var control in power distribution systems. In *Proceedings of The 4th Annual Learning for Dynamics and Control Conference*, volume 168 of *Proceedings of Machine Learning Research*, pp. 21–33. PMLR, 2022.

Avishai Halev, Yongshuai Liu, and Xin Liu. Microgrid control under uncertainty. *Engineering Applications of Artificial Intelligence*, 138:109360, 2024. doi: 10.1016/j.engappai.2024.109360.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753. PMLR, 2019.

Fengzhang Luo, Shengyuan Wang, Yunqiang Lv, Ranfeng Mu, Jiacheng Fo, Tianyu Zhang, Jing Xu, and Chengshan Wang. Domain knowledge-enhanced graph reinforcement learning method for volt/var control in distribution networks. *Applied Energy*, 398:126409, 2025. doi: 10.1016/j.apenergy.2025.126409.

Kingsley Nweye, Zoltan Nagy, Sharada Mohanty, Dipam Chakraborty, Siva Sankaranarayanan, Tianzhen Hong, Sourav Dey, Gregor Henze, Jan Drgona, Fangquan Lin, Wei Jiang, Hanwei Zhang, Zhongkai Yi, Jihai Zhang, Cheng Yang, Matthew Motoki, Sorapong Khongnawang, Michael Ibrahim, Abilmansur Zhumabekov, Daniel May, Zhihu Yang, Xiaozhuang Song, Han Zhang, Xiaoning Dong, Shun Zheng, and Jiang Bian. The citylearn challenge 2022: Overview, results, and lessons learned. In *Proceedings of the NeurIPS 2022 Competitions Track*, volume 220 of *Proceedings of Machine Learning Research*, pp. 85–103. PMLR, 2022.

Charles Packer, Katelyn Gao, Jernej Kos, Philipp Krähenbühl, Vladlen Koltun, and Dawn Song. Assessing generalization in deep reinforcement learning, 2019.

Sai Sasidhar Punyam Rajendran and Alemayehu Gebremedhin. Meta-reinforcement learning for multi-energy building microgrids: Curriculum design, rapid adaptation, and scalable deployment. *Energy Conversion and Management*, 363:121652, 2026. doi: 10.1016/j.enconman.2026.121652.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017. doi: 10.1109/IROS.2017.8202133.

Roberto Vercellino, Jared Willard, Gustavo Campos, Weslley da Silva Pereira, Olivia Hull, Matt Selensky, and Juliane Mueller. Dataset of generative ai workload power profiles. National Laboratory of the Rockies Data Catalog, 2026. Dataset ID 312.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yunfei Zhang, Kefan Bao, Gaige Liang, Wennan Zhuang, Longlong Qiang, Difei Tang, Xiangyu Lu, and Mingxiao Zhang. Topology-aware graph reinforcement learning for voltage-reactive power control in grid-connected microgrids. *Electricity*, 7(2):60, 2026. doi: 10.3390/electricity7020060.

A Additional Experimental Details

A.1 Policy implementation

The V6 graph contains 12 buses: an external-grid bus, a main-board bus, and ten controlled building buses. We scatter each ten-feature building observation to its bus and append a controllability indicator. The two uncontrolled buses receive zero raw inputs; their hidden states are not constrained to remain zero. A linear 11-to-120 encoder with a Gaussian error linear unit (GELU) activation is followed by two message-passing layers. Each physical line supplies two directed edges with total resistance and reactance in ohms.

For layer ℓ, messages and node updates are

$$\bar{m}_i^\ell = \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \left(W_s^\ell h_j^\ell + W_e^\ell e_{ji} \right),$$

$$h_i^{\ell+1} = \mathrm{LN}\Big(h_i^\ell + \mathrm{MLP}^\ell\big([h_i^\ell \,\|\, \bar{m}_i^\ell]\big) \Big).$$

Here, $\mathcal{N}(i)$ denotes the neighbors of node i, h_i^ℓ its hidden state at layer ℓ, and e_{ji} the resistance–reactance feature vector of edge $j \to i$. W_s^ℓ and W_e^ℓ are learned projection matrices, $\|$ denotes concatenation, LN denotes layer normalization, and MLP denotes a multilayer perceptron.

Both message projections are bias-free. The update MLP has dimensions 240→240→120 with GELU. Controlled-node states feed 120→64→5 discrete-logit and 120→64→1 continuous-mean heads. A shared learned log standard deviation parameterizes continuous sampling. The value head is 120→64→1 after mean pooling over controlled nodes. All heads use GELU. Six buildings use discrete actions and four use continuous actions.

PI-TF uses a 10-to-128 encoder with GELU, two four-head self-attention layers, and one 128→256→128 feed-forward block after the attention layers, with residual connections and LayerNorm. Fixed-TF adds learned slot embeddings before attention. PI-TF has 225,416 parameters and each GNN condition has 228,408. The GNN conditions share architecture, not learned weights. Both policies are centralized: observations include aggregate controlled demand and the critic pools building states. Fixed-TF is not an arm of the constraint-active study.

A.2 Graph controls

Full uses physical adjacency and total line resistance/reactance. Adjacency-only preserves the graph and model but zeros both edge channels. The 480 bias-free edge-projection weights remain instantiated and receive no informative edge input. Resistance and reactance are numerically equal on every frozen line; their separate effects are therefore not identified.

Rewired applies a deterministic endpoint permutation with seed 424242, fixing only the external-grid index. In bus-index order, the permutation is [0, 6, 7, 2, 9, 8, 10, 3, 5, 11, 4, 1]. The main-board index is not fixed. This preserves the unlabeled graph and degree multiset while changing alignment with physical node identities; it is not an edge-swap intervention. Edge-attribute rows, node observations, controlled indices, and simulator connectivity remain unchanged. Figure A1 shows the source graph. This is one frozen alignment control, not an average over independently sampled rewiring controls.

A.3 Observations, actions, and PPO details

The ten observation entries are current uncontrolled demand, previous controlled load, previous mean three-phase voltage, constants 0.5 and 22.0, building entropy and sparsity scores, equity weight, current tariff, and previous aggregate controlled demand. The constants do not represent simulated battery or temperature dynamics. Observations and edge attributes are not explicitly normalized. Discrete controls reduce demand by 0%, 20%, 40%, 60%, or 80%; continuous action $a \in [-1, 1]$ multiplies demand by $1 + 0.5a$. Evaluation uses discrete argmax and tanh of the continuous mean.

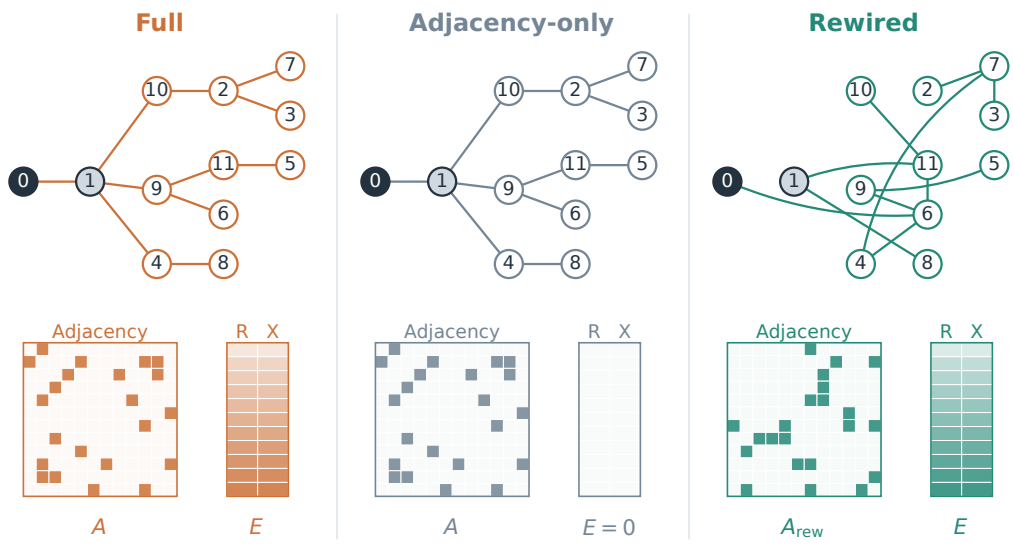

Figure A1: Frozen graph controls in the constraint-active study. Node positions and labels retain physical bus identity. Full and adjacency-only share physical adjacency; adjacency-only sets all line attributes to zero. Rewired applies an isomorphic endpoint permutation (seed 424242; bus 0 fixed), retaining ordered attribute rows. The physical simulator, observations, and controlled buses remain unchanged. Adjacency matrices use bus order 0–11; each attribute row represents one undirected line. Attribute colors share a common relative scale.

In addition to Section 4.3, PPO uses Adam, normalized advantages, reward scaling of 10^{-3}, and joint log probabilities summed over buildings. Updates stop early when approximate KL exceeds 1.5 times the target of 0.03; episode-end values are not bootstrapped. Reported returns use unscaled environment rewards. Training seeds are 0–19 and scenario seeds are $800000+1000s+e$ for training seed s and episode index e. Evaluation uses scenario seeds 52000–52009 on the source and each final topology. The selected source has topology seed 7235 and downstream line-length multiplier 8.0. Exact dependencies and frozen hashes are recorded in the accompanying reproducibility files.

A.4 Safety and utility endpoints

Voltage-violation rate averages indicators over ten buildings, three phases, and 96 steps. Voltage excursion is $\max(0, 0.95 - v) + \max(0, v - 1.05)$. The severe indicator is true if excursion exceeds 0.02 p.u. or the building VUF exceeds 2%; the VUF condition is repeated across phases. Severe events therefore need not be a subset of ordinary voltage violations. VUF-violation rate averages over buildings and time. Reported voltage minima, voltage maxima, maximum excursion, and maximum VUF are averages of episode extrema, not extrema over the entire evaluation panel.

Per-step energy cost is aggregate controlled demand times tariff times 0.25 hours. The equity penalty sums building equity weights times dissatisfaction: the curtailed fraction for discrete controls or absolute modulation for continuous controls. It measures weighted dissatisfaction rather than inequality across buildings. The physics penalty is 100 times summed voltage excursion plus 50 times summed VUF in percent, without thresholding the VUF term. Reward is $-\text{cost} - \text{equity penalty} - 100 \times \text{physics penalty}$. Return is consequently not independent evidence for a safety effect.

A.5 Statistical analysis and output verification

For H1 and H2, each seed contributes one difference after averaging the five final topologies and ten scenarios with equal weight. The registered percentile bootstrap uses 10,000 paired train-seed resamples and analysis seed 20260917. Exact two-sided sign-flip tests enumerate all 2^{20} sign assignments, under the registered symmetry/exchangeability assumption, followed by Holm correction

across the two primary tests. Adjacency-only, utility outcomes, and per-topology summaries remain descriptive. Intervals are conditional on the frozen topology/scenario panel and do not quantify uncertainty over a population of unseen topologies.

All 80 arm-by-seed cells are present. The completed outputs contain 4,000 held-out and 800 source evaluation episodes, with no evaluation power-flow failures; source episodes do not enter the primary contrasts. Verification checked frozen-file hashes, output completeness, checkpoint structures, graph arrays, and logged reward decomposition. Postprocessing used existing outputs without training, tuning, or additional policy evaluation.

A.6 Source-only pilot checks

Before final-panel evaluation, PI-TF and Grid-GNN-full passed a source-only pilot using training seeds 0–2, 100 training episodes, and three source evaluation scenarios per seed (53000–53002). All six runs completed with finite outputs, nondegenerate actions in the final training episode, and no source-evaluation power-flow failures. These checks assess operational validity rather than optimization convergence.

B Additional Results

B.1 Safety and control performance

Table B1 reports the complete registered V6 endpoint panel. Values are means with pointwise 95% bootstrap intervals over 20 training seeds; each seed averages the five final topologies and ten scenarios. Lower violation rates, costs, and penalties and higher return are preferred. These descriptive summaries do not add confirmatory tests.

Table B1: Safety and utility on the fixed final panel: means [pointwise 95% train-seed bootstrap intervals]. Cost is in simulator units; equity is weighted dissatisfaction. Extrema are averaged across episodes. The severe endpoint combines voltage excursion and VUF events (Appendix A.4). All secondary comparisons are descriptive.

Metric	PI-TF	Full	Adjacency-only	Rewired
Voltage violation (%)	2.133 [1.130, 3.368]	3.560 [1.888, 5.802]	3.127 [1.335, 5.323]	4.244 [2.461, 6.304]
Severe violation (%)	1.654 [0.670, 2.877]	3.274 [1.311, 5.893]	2.761 [0.847, 5.102]	4.007 [2.050, 6.249]
Min. voltage (p.u.)	0.9500 [0.9452, 0.9544]	0.9432 [0.9347, 0.9498]	0.9466 [0.9379, 0.9538]	0.9411 [0.9336, 0.9480]
Max. voltage (p.u.)	0.9997 [0.9997, 0.9998]	0.9998 [0.9997, 0.9999]	0.9997 [0.9996, 0.9998]	0.9998 [0.9997, 0.9998]
Max. excursion (p.u.)	0.0067 [0.0037, 0.0102]	0.0122 [0.0065, 0.0197]	0.0100 [0.0042, 0.0170]	0.0136 [0.0081, 0.0198]
VUF violation (%)	1.544 [0.655, 2.644]	2.891 [1.281, 5.015]	2.454 [0.764, 4.500]	3.589 [1.895, 5.506]
Mean VUF (%)	0.572 [0.524, 0.627]	0.650 [0.574, 0.746]	0.614 [0.533, 0.710]	0.671 [0.593, 0.757]
Max. VUF (%)	1.612 [1.471, 1.769]	1.857 [1.624, 2.160]	1.725 [1.476, 2.024]	1.916 [1.678, 2.178]
Episode return (10^6)	-2.771 [-3.037, -2.534]	-3.154 [-3.638, -2.778]	-2.979 [-3.457, -2.580]	-3.260 [-3.687, -2.870]
Energy cost (10^3 units)	16.341 [14.906, 17.884]	17.959 [16.156, 20.140]	17.508 [15.545, 19.683]	18.668 [16.723, 20.749]
Equity penalty	492.91 [381.43, 605.15]	521.42 [424.60, 615.82]	443.59 [330.78, 558.66]	448.88 [353.10, 543.26]
Physics penalty (10^3)	27.537 [25.182, 30.182]	31.354 [27.611, 36.173]	29.608 [25.642, 34.370]	32.408 [28.524, 36.664]
PF failures / episode	0 [0, 0]	0 [0, 0]	0 [0, 0]	0 [0, 0]

The exact paired sign-flip p-values are 0.335369 for H1 and 0.373985 for H2; both Holm-adjusted values are 0.670738. Full has numerically lower severe-violation rate and higher mean episode min-

imum voltage than Rewired, but is numerically worse on these endpoints than PI-TF. Its lower cost relative to Rewired accompanies higher weighted dissatisfaction. These patterns do not establish a general safety or utility advantage.

B.2 Per-topology descriptive results

Table B2 and Figure B1 summarize voltage violations on each fixed final topology. Full has a numerically higher mean rate than PI-TF and a lower mean rate than Rewired on all five topologies. Rates are larger on E2 and E5 than on E1 and E3. These directions are descriptive: the five topologies reuse the same trained policies and do not supply five independent sets of training runs. Error bars resample seed means within each fixed topology. Figure B1 also displays the return, cost, and equity summaries used in Section 5.5.

Table B2: Per-topology voltage-violation rate (%): means [pointwise 95% train-seed bootstrap intervals]. Each seed averages ten scenarios within that topology. These are descriptive summaries of the same 20 trained policies per arm, not additional independent training runs.

Topology	PI-TF	Full	Adjacency-only	Rewired
V6_E1	0.077 [0.000, 0.231]	0.264 [0.000, 0.667]	0.350 [0.000, 0.730]	0.324 [0.000, 0.694]
V6_E2	4.561 [2.756, 6.643]	7.051 [4.581, 10.055]	5.471 [2.888, 8.406]	8.117 [5.379, 11.052]
V6_E3	0.090 [0.000, 0.270]	0.273 [0.000, 0.687]	0.371 [0.000, 0.768]	0.430 [0.081, 0.842]
V6_E4	2.064 [1.140, 3.205]	3.416 [1.765, 5.724]	3.262 [1.372, 5.569]	4.119 [2.302, 6.249]
V6_E5	3.871 [1.699, 6.622]	6.798 [3.056, 12.000]	6.183 [2.090, 11.215]	8.232 [4.265, 12.909]

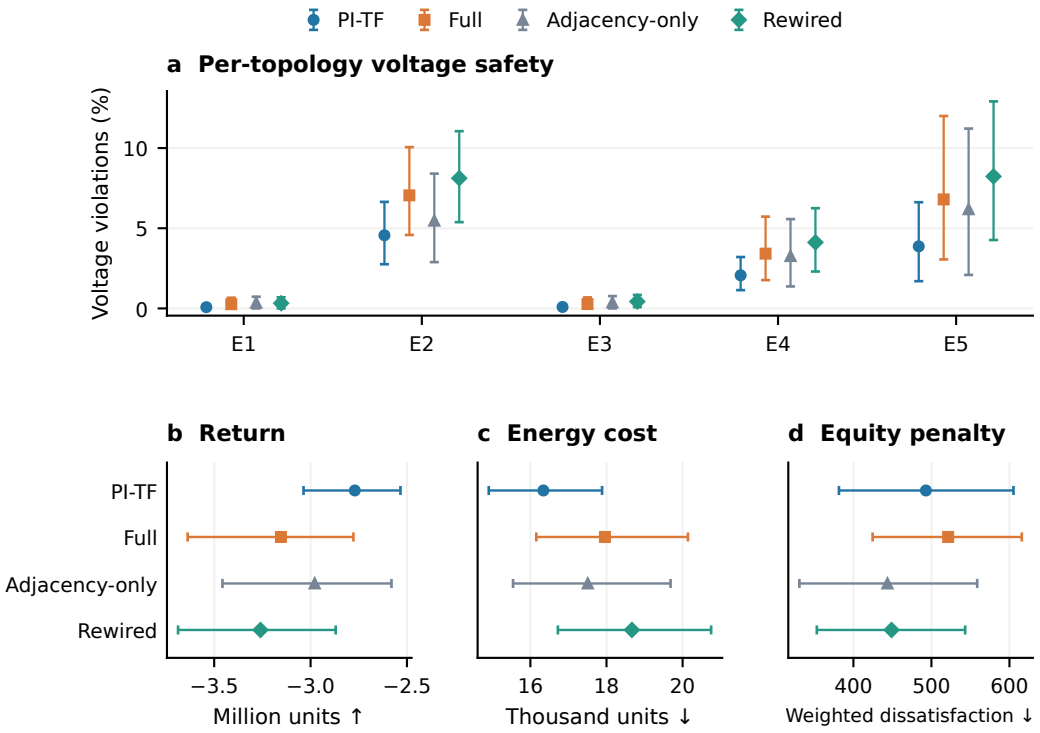

Figure B1: Descriptive results on the frozen final panel. (a) Voltage-violation rates by topology. (b–d) Episode return, energy cost, and equity penalty averaged over the five topologies. Points are means over 20 training seeds; bars are pointwise 95% seed-bootstrap intervals, conditional on the fixed topology/scenario panel. Adjacency-only and utility comparisons are secondary. These intervals do not support new topology-specific tests. Cost uses simulator units and equity denotes weighted dissatisfaction.