POLICYGRID: Causal Discovery for Adaptive Policy Optimization in Embodied Agents (Student Abstract)

Taqiya Ehsan, Shuren Xia, Jorge Ortiz

Rutgers University, New Brunswick, New Jersey, USA
{taqiya.ehsan, shuren.xia, jorge.ortiz}@rutgers.edu

Abstract

Embodied agents must reason causally, as correlation-based models fail under intervention and distribution shift. This challenge arises in domains like robotics and cyber-physical systems, where agents balance efficiency and comfort under uncertainty. We introduce POLICYGRID, unifying causal discovery and control by treating each action as both decision and experiment. Leveraging constraint-based search, neural causal models, and language model priors with interventional validation, POLICYGRID yields adaptive, interpretable policies. Across synthetic, real-world, and live deployments, it achieves superior causal recovery (F1 = 0.89) and 2.8× better multi-objective performance than correlation-based baselines, demonstrating safe, generalizable decision-making.

Introduction

Agents in embodied environments must reason about how actions causally influence outcomes, as correlation-based policies fail under intervention and distribution shift (Glymour, Zhang, and Spirtes 2019; Richens et al. 2025). This challenge is acute in cyber-physical domains such as smart infrastructure and robotics, where safety, uncertainty, and competing objectives demand causal reasoning (Hafner et al. 2020).

Existing causal discovery methods (e.g., PC (Spirtes et al. 2000)) provide theoretical guarantees but yield static graphs, while control approaches such as causal bandits (Lattimore, Lattimore, and Reid 2016) and causal RL (Buesing et al. 2019) assume known causal structure. Effective agents instead require *online causal discovery* integrated directly with policy optimization (Finn, Abbeel, and Levine 2017).

POLICYGRID unifies learning and acting by treating each action as both control and experiment. The framework combines constraint-based testing, neural structural equation models (SEM), and large language model (LLM) priors with interventional validation to learn interpretable causal models while optimizing control objectives (Kalainathan et al. 2018; Sun and Li 2024). Our contributions are:

1. Dual-purpose interventions for simultaneous discovery and control;
2. Causal discovery embedded within policy optimization;
3. Validation across synthetic, real-world, and physical environments, demonstrating improved causal recovery and control performance.

Framework Architecture

We model embodied control as an agent interacting with a dynamic environment governed by an unknown structural causal model (SCM) $G = (V, E)$. Let $V = \{V_1, \ldots, V_n\}$ denote observed variables, E directed edges, C_t contextual inputs, and A_t the action at time t. The system evolves as

$$V_i(t) = f_i(\text{Pa}_G(V_i(t)), A_t, C_t, \epsilon_i(t)), \tag{1}$$

where $\text{Pa}_G(V_i)$ are causal parents and ϵ_i exogenous noise. In the absence of G, correlation-driven policies are brittle and unsafe.

POLICYGRID addresses this by jointly performing causal discovery and policy optimization within a single reasoning loop.

Discovery Module. Candidate edges are generated via constraint-based tests, neural SEM learning, and LLM priors. Uncertain edges are probed using targeted interventions; edges with significant downstream effects are retained, while others are pruned. This iterative process continues until convergence or budget exhaustion, producing an auditable DAG $\hat{G}$.

Policy Engine. The learned DAG $\hat{G}$ enables multi-objective control. Conditioned on C_t and task objectives, the engine estimates causal effects to evaluate trade-offs and optimize objectives such as comfort and energy efficiency:

$$(\hat{G}, \{A_t^*\}) = \arg\max_{\hat{G}, \pi} E[O(\pi, \hat{G}, C_t)] - \lambda \,\text{Uncertainty}(\hat{G}). \tag{2}$$

Overall, the framework jointly infers a causal model and optimizes actions:

$$(\hat{G}, \{A_t^*\}_{t=1}^T) = \text{POLICYGRID}(\mathcal{D}_{obs}, \mathcal{D}_{int}, \mathcal{C}), \tag{3}$$

where $\mathcal{D}_{obs}$ are observational samples, $\mathcal{D}_{int}$ interventional outcomes, and $\mathcal{C}$ contextual objectives. Both components operate within a unified embodied framework that refines causal structure through action while maintaining safe, interpretable control.

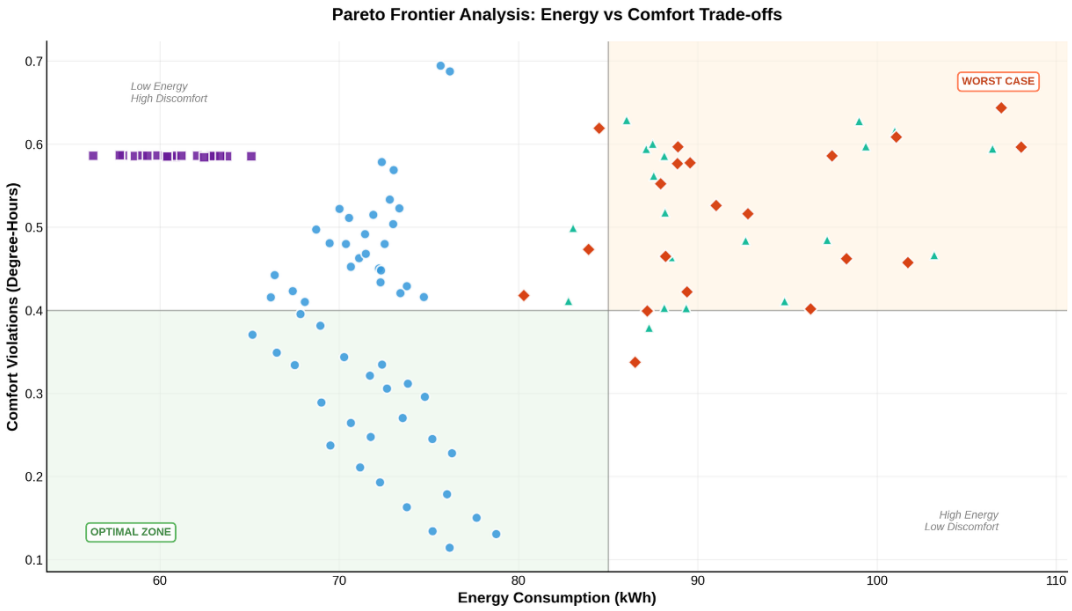

Figure 1: Pareto frontier in the *Large-Sim* scenario comparing energy use and comfort violations. PolicyGRID (blue circle) achieves superior trade-offs by leveraging causal structure, outperforming ASHRAE (teal triangle), correlation-based control (orange diamond), and ablated PolicyGRID without DAGs (purple square). Lower-left denotes optimal performance.

Experiments and Results

We evaluate POLICYGRID across six setups: four synthetic benchmarks (Base, Noisy, Hidden-Vars, Large-Sim), a real-world dataset (ASHRAE), and a physical deployment (Physical). We report causal recovery (SHD, F1), intervention efficiency (risk, cost), and multi-objective policy metrics (hypervolume, Pareto frontier, comfort violations).

Causal Recovery. Table 1 summarizes Structural Hamming Distance (SHD) across all setups. POLICYGRID achieves near-perfect recovery throughout, substantially outperforming a broad range of baseline approaches, including constraint-based methods (PC (Spirtes et al. 2000)), score-based methods (GIES (Hauser and Bühlmann 2012)), neural approaches (SAM (Kalainathan et al. 2018)), invariant and interventional methods (ICP (Peters, Bühlmann, and Meinshausen 2016), Causal Bandits (Lattimore, Lattimore, and Reid 2016)), hybrid approaches (ABCD (Toth et al. 2022), JCI (Mooij, Magliacane, and Claassen 2020)), and LLM-guided discovery (Sun and Li 2024). Across the full suite of metrics—including F1, intervention cost, and risk—POLICYGRID consistently demonstrates robust causal recovery under noise, hidden confounding, and large-scale environments. These gains arise from its integration of active, goal-directed interventions within the discovery–control loop, enabling more accurate causal identification and safer downstream policy performance.

Policy Performance. Figure 1 shows the energy–comfort Pareto frontier. POLICYGRID (blue circle) dominates competing methods, reducing both energy use and comfort violations relative to correlation-based and ablated variants. These gains are primarily attributable to improved causal modeling rather than additional data or tuning, as confirmed by comparison with POLICYGRID without causal graphs (purple square). The learned policies transfer from simulation to physical deployment without retraining. Additional experimental details and ablations are provided in the NeurIPS Embodied World Models workshop paper (Ehsan, Xia, and Ortiz 2025).

Method	B	N	H	A	P	L
PC	4	4	4	4	2	49
SAM	8	6	8	4	7	21
LLM	7	7	7	4	2	17
GIES	6	10	8	4	2	22
JCI	8	3	8	13	6	28
ABCD	5	8	6	4	8	23
Causal Bandits	8	5	9	8	7	43
ICP	5	5	4	2	6	39
IID	4	4	4	12	2	56
NOTEARS	8	9	3	6	4	26
GRID	**0**	**2**	**0**	**1**	**0**	**13**

B: Base, **N**: Noisy, **H**: Hidden Vars, **A**: ASHRAE, **P**: Physical, **L**: Large-Sim

Table 1: Structural Hamming Distance (SHD) across six setups. Lower is better.

Conclusion

POLICYGRID unifies causal discovery, targeted intervention, and policy optimization within a single embodied reasoning loop, allowing agents to refine world models through action. Across synthetic, real-world, and physical systems, it delivers interpretable, robust gains in both causal recovery and multi-objective control, enabling safer generalization across domains.

Acknowledgments

This work was supported by the National Science Foundation (NSF) and Center for Smart Streetscapes (CS3) under NSF Cooperative Agreement No. EEC-2133516.

References

Buesing, L.; Weber, T.; Zwols, Y.; Racaniere, S.; Guez, A.; Lespiau, J.-B.; and Heess, N. 2019. Woulda, Coulda, Shoulda: Counterfactually-Guided Policy Search. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, 1231–1240. PMLR.

Ehsan, T.; Xia, S.; and Ortiz, J. 2025. PolicyGRID: Acting to Understand, Understanding to Act. In *NeurIPS 2025 Workshop on Embodied World Models for Decision Making*.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, 1126–1135. PMLR.

Glymour, C.; Zhang, K.; and Spirtes, P. 2019. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10: 524.

Hafner, D.; Lillicrap, T.; Ba, J.; and Norouzi, M. 2020. Dream to Control: Learning Behaviors by Latent Imagination. In *International Conference on Learning Representations (ICLR)*.

Hauser, A.; and Bühlmann, P. 2012. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research*, 13(1): 2409–2464.

Kalainathan, D.; Goudet, O.; Guyon, I.; Lopez-Paz, D.; and Sebag, M. 2018. Structural Agnostic Modeling: Adversarial Learning of Causal Graphs. *arXiv preprint arXiv:1803.04929*.

Lattimore, F.; Lattimore, T.; and Reid, M. D. 2016. Causal bandits: Learning good interventions via causal inference. *Advances in neural information processing systems*, 29.

Mooij, J. M.; Magliacane, S.; and Claassen, T. 2020. Joint causal inference from multiple contexts. *Journal of machine learning research*, 21(99): 1–108.

Peters, J.; Bühlmann, P.; and Meinshausen, N. 2016. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5): 947–1012.

Richens, J.; Abel, D.; Bellot, A.; and Everitt, T. 2025. General Agents Need World Models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*. PMLR.

Spirtes, P.; Glymour, C. N.; Scheines, R.; and Heckerman, D. 2000. *Causation, prediction, and search*. MIT press.

Sun, Z.; and Li, Q. 2024. Leveraging LLMs for Causal Inference and Discovery. *arXiv preprint arXiv:2410.16676*.

Toth, C.; Lorch, L.; Knoll, C.; Krause, A.; Pernkopf, F.; Peharz, R.; and Von Kügelgen, J. 2022. Active bayesian causal inference. *Advances in Neural Information Processing Systems*, 35: 16261–16275.