POLICYGRID: Acting to Understand, Understanding to Act

Anonymous Author(s)
Affiliation
Address
email

Abstract

Embodied agents require internal models that support interventional reasoning, not merely correlational prediction. We present POLICYGRID, an embodied world model that learns causal structure online through its own actions. Unlike traditional approaches that treat causal discovery as preprocessing, POLICYGRID integrates causal learning directly into the policy loop: agents actively probe the environment to resolve causal uncertainty while simultaneously optimizing for competing objectives. This enables agents to adapt their causal understanding as they act, expanding their behavioral repertoire beyond correlation-driven policies. The framework addresses a fundamental challenge in embodied AI: how can agents maintain reliable world models when their own interventions continuously change the data distribution? To validate this approach, we evaluate POLICYGRID in building control across synthetic simulations, public datasets, and real deployment, achieving F1 = 0.89 under real-world conditions and 2.8× higher policy performance than baselines, demonstrating that embedding causal reasoning directly into the policy loop yields more robust, adaptive behavior than correlation-driven world models.

1 Introduction

Embodied agents must reason causally about how their actions affect the environment. Unlike passive observers, agents that act in the world require internal models that predict the consequences of interventions [Ha and Schmidhuber, 2018, Hafner et al., 2020]. Correlational models fail when agents intervene because correlation does not imply causation. This fundamental limitation constrains current embodied AI systems to reactive behaviors rather than principled, goal-directed action, despite theoretical work showing that general agents require world models for multi-step generalization [Richens et al., 2025].

The challenge is acute in cyber-physical systems where agents must balance competing objectives under uncertainty. Standard world models capture statistical dependencies but provide no mechanism for reasoning about interventions. When an agent acts, these models cannot distinguish between spurious correlations and genuine causal relationships. The result is brittle policies that fail when the environment shifts.

Current approaches to embodied decision-making either ignore causal structure entirely or treat causal discovery as a separate preprocessing step. This disconnect prevents embodied agents from integrating causal reasoning directly into their decision-making processes. Discovering causal structure alone is insufficient; the structure must be leveraged for principled policy generation.

We address this gap through POLICYGRID, a unified framework that integrates interventional causal discovery with policy generation for embodied agents. Building on a causal discovery core, POLICYGRID extends beyond traditional approaches [Buesing et al., 2019, Ding et al., 2022]

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

that treat causal modeling as preprocessing. Instead, it learns causal structure through targeted interventions, combining constraint-based search, neural structural equation modeling, and language model priors. The framework then leverages these validated causal graphs to generate policies with explicit trade-offs between competing objectives. This produces interpretable multi-objective optimization.

The framework addresses three key requirements for embodied causal reasoning: (1) discovering causal structure from observational and interventional data, (2) validating causal relationships through targeted experiments, and (3) translating causal knowledge into operational policies. By integrating these components, POLICYGRID enables agents to reason about the consequences of their actions rather than merely react to correlations. This work extends the embodied world model discourse by demonstrating how causal reasoning can be embedded directly within the policy loop, moving beyond the latent dynamics models of Hafner et al. [2020] and the meta-learning approaches of Finn et al. [2017] toward principled interventional reasoning [Battaglia et al., 2016, Wu et al., 2015].

We evaluate POLICYGRID in domains where agents act, measure consequences, and balance competing objectives. Cyber-physical systems with rich sensors, defined actuation, and quantifiable trade-offs provide suitable testbeds. Building control exemplifies this class: abundant sensors, clear actuation channels, and energy-comfort trade-offs. We validate across synthetic simulations, the ASHRAE Great Energy Predictor III dataset, and live office deployment. POLICYGRID outperforms correlation-based approaches in both policy performance and interpretability.

Contributions. We establish three insights for embodied causal reasoning. First, interventions serve dual purposes: control actions simultaneously manipulate the environment and refine the agent's causal understanding. POLICYGRID treats each action as both a policy decision and an experiment that updates internal world models. Second, causal discovery need not precede control but can occur within it. Agents probe their environment to resolve structural uncertainty while optimizing for objectives, collapsing the traditional separation between learning and acting. Third, policies grounded in validated causal structure outperform correlation-based alternatives in multi-objective settings. Empirical validation demonstrates F1 = 0.89 for causal recovery and 2.8$\times$ hypervolume improvement over correlation-based methods, confirming that causal world models provide a more reliable foundation for multi-objective control than statistical dependencies alone.

2 Related Work

2.1 Causal Discovery Methods

Causal discovery methods recover structural relationships from data. Constraint-based approaches like PC [Spirtes et al., 2000] test conditional independencies but fail under noise and hidden confounders [Colombo et al., 2012, Glymour et al., 2019]. Structural equation models [Kalainathan et al., 2018, Rosseel and Loh, 2022, Monti et al., 2020] capture nonlinearities but require careful specification. NOTEARS [Zheng et al., 2020] reformulates structure learning as continuous optimization. Recent work incorporates language model priors [Sun and Li, 2024, Kıcıman et al., 2023] but sacrifices robustness. These methods treat causal discovery as offline preprocessing. Graphs remain fixed during deployment. Embodied agents require adaptive causal understanding through interaction.

2.2 Embodied Agents and World Models

Embodied agents require internal models to predict intervention consequences [Ha and Schmidhuber, 2018, Hafner et al., 2020, Richens et al., 2025]. Robotics approaches use meta-learning [Finn et al., 2017] or physics priors [Wu et al., 2015, Battaglia et al., 2016] for adaptation. Cyber-physical systems employ correlation-based models [Kleissl and Agarwal, 2010, Kathirgamanathan et al., 2021, Czekster et al., 2022] that capture statistical dependencies but lack causal structure. Existing world models predict correlations, not causal effects. Correlational models fail under intervention because correlation does not imply causation [Glymour et al., 2019, Zhang et al., 2022]. Causal world models remain valid under intervention.

2.3 Causal Reasoning in Control

Control systems benefit from causal reasoning. Invariant prediction [Peters et al., 2016, 2017] ensures stability across environments. Active discovery methods [Hauser and Bühlmann, 2012, Mooij et al., 2020, Zhang et al., 2023] use interventions to resolve causal orientation. Causal bandits [Lattimore et al., 2016] optimize intervention selection. Causal reinforcement learning explores counterfactual policies [Buesing et al., 2019] and goal-conditioned reasoning [Ding et al., 2022]. These approaches separate causal discovery from control. Discovery methods assume fixed environments. Control methods assume known causal structure. Embodied agents must learn causal structure through control actions while optimizing objectives. POLICYGRID eliminates this separation. Actions optimize objectives and refine causal understanding simultaneously. Each intervention serves as both a control decision and a causal experiment.

3 Problem Formulation

We frame embodied control as the task of an agent acting in a dynamic environment where decisions must be guided by causal structure rather than correlations. To formalize this setting, we begin with the set of observable variables $\mathcal{V} = \{V_1, \ldots, V_n\}$ measured by sensors. These variables capture the system's state and are the quantities the agent must reason about. At each time step t, the environment also presents exogenous context C_t, such as weather or occupancy, which influences outcomes but cannot be controlled. The agent selects an action A_t from a feasible intervention set $\mathcal{C}$, representing its direct ability to affect the environment.

The dynamics linking these elements are unknown but assumed to follow a structural causal model (SCM) $G = (\mathcal{V}, \mathcal{E})$. Representing the environment in this way is necessary because we are interested not only in prediction but in understanding how interventions propagate. Each variable evolves according to

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

where $\mathrm{Pa}_G(V_i(t))$ are the parents of V_i in G, f_i is an unknown structural function, and $\epsilon_i(t)$ is noise. This formalism makes explicit that trajectories depend on endogenous interactions, exogenous context, and the agent's actions.

The difficulty is that G is unobserved. Without it, the agent cannot distinguish true causal influence from spurious correlation, making policies fragile under shifts in context. To address this, POLICYGRID integrates causal discovery directly into the control process via the *discovery module*. The agent does not treat discovery as a preliminary offline task; instead, it iteratively builds $\hat{G}$ by interacting with the environment. Observational data $\mathcal{D}_{obs} = \{V(t), C_t\}_{t=1}^{T}$ provide a baseline model of dependencies, but these alone are insufficient for causal identification. Interventional data $\mathcal{D}_{int} = \{V(t), A_t, C_t\}_{t=1}^{T}$ are therefore used to test competing hypotheses about system structure. Combining both sources produces a working graph

$$\hat{G} = \text{discovery\_module}(\mathcal{D}_{obs}, \mathcal{D}_{int}), \tag{2}$$

which the agent treats as its current world model.

Constructing $\hat{G}$ is only part of the problem: the ultimate goal is to act. Policies must be computed with respect to the discovered structure so that interventions are chosen for their causal effect rather than their observed association. Given context C_t and candidate actions $\mathcal{C}$, the policy engine uses $\hat{G}$ to compute

$$A_t^* = \pi(\hat{G}, C_t) \in \mathcal{C}, \tag{3}$$

where π denotes a policy that optimizes multiple objectives under uncertainty. This formulation makes clear why discovery and policy must be coupled: without $\hat{G}$, the agent cannot anticipate intervention effects; without policy, discovery has no operational value.

The overall problem is therefore to jointly infer a causal model and optimize actions over it. Formally,

$$(\hat{G}, \{A_t^*\}_{t=1}^{T}) = \textsc{PolicyGRID}(\mathcal{D}_{obs}, \mathcal{D}_{int}, \mathcal{C}), \tag{4}$$

where both components are solved within a unified embodied framework.

Although we demonstrate this framework in building control—where dense sensing, clear actuation, and energy–comfort trade-offs provide a concrete setting—the formulation is not specific to that

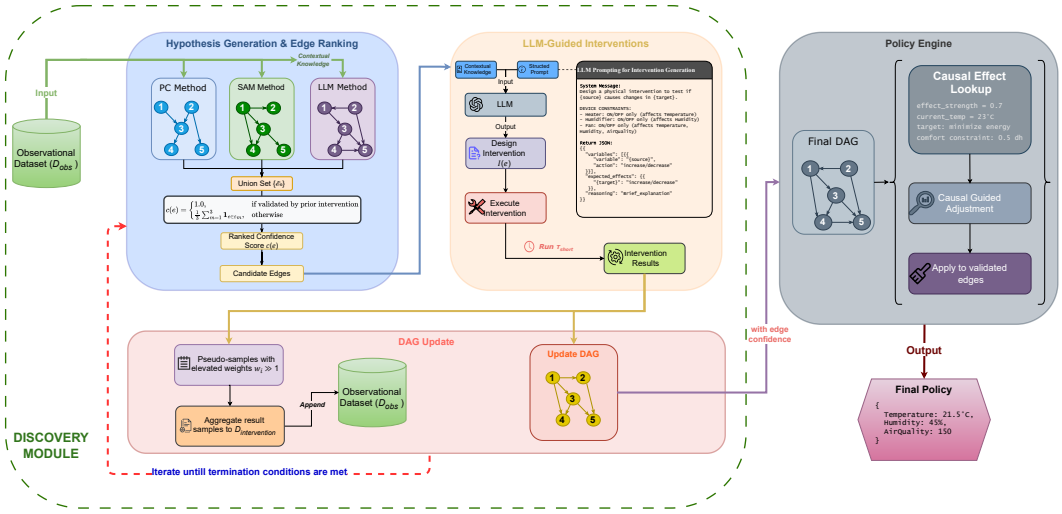


Figure 1: Architecture of POLICYGRID. The discovery module (left) iteratively generates, validates, and refines causal hypotheses; the policy engine (right) leverages the validated DAG to generate multi-objective policies.

domain. Any embodied agent facing coupled dynamics, exogenous context, and the need for multi-objective control can be expressed within the same problem structure.

4 Methodology

POLICYGRID operationalizes embodied agents by iteratively learning a causal world model from both observational and interventional data, and using that model to generate interpretable policies under competing objectives. The framework consists of two tightly coupled modules: (i) a causal *discovery module* that constructs and validates a directed acyclic graph (DAG) over observed variables; and (ii) a causal *policy engine* that queries the validated DAG to evaluate and recommend interventions.

4.1 Embodied World-Model Learning

At the core of POLICYGRID is its *discovery module*, which closes the perception–action loop by refining a world model of system dynamics. Let $\mathcal{V} = \{V_1, \ldots, V_n\}$ denote the set of n observed variables (e.g., temperature, humidity, device states in a building example), and let $D_{\text{obs}} = \{v_t\}_{t=1}^{T}$ denote a dataset of T observational measurements. The discovery module incrementally learns a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each edge $e = i \to j \in \mathcal{E}$ encodes a candidate causal dependency. The process iterates over three stages: candidate edge generation, interventional validation, and refinement.

Multi-method DAG construction. Candidate causal edges are proposed using three complementary methods: *Constraint-based PC* [Spirtes et al., 2000], which builds a skeleton from conditional independence tests (Fisher's z-test with significance level $\alpha = 0.05$) and orients edges using PC rules; *neural SEM* via *SAM* [Kalainathan et al., 2018] learns a weighted adjacency matrix $W \in \mathbb{R}^{n \times n}$ by minimizing a reconstruction loss with L_1, L_2, and acyclicity penalties, optimized with hyperparameters $\lambda_1 = 0.1$, $\lambda_2 = 0.01$, generator learning rate $\eta_g = 0.01$, discriminator learning rate $\eta_d = 0.005$, for 200 epochs and three random restarts; an *LLM (GPT-3.5-turbo)* [Ye et al., 2023], prompted with domain physics and actuator constraints to propose acyclicity-consistent edges.

The union of the three outputs forms the candidate edge set. Each edge e is assigned a confidence score

$$c(e) = \begin{cases} 1.0, & \text{if } e \text{ was previously validated,} \\ \frac{1}{3}\sum_{m=1}^{3} \mathbf{1}\{e \in E_m\}, & \text{otherwise,} \end{cases}$$

where E_m is the edge set from method m. Low-confidence edges are prioritized for testing.

Interventional validation. Let $X_i \in \mathcal{V}$ be a candidate parent of $X_j \in \mathcal{V}$. An intervention $do(X_i = x_i')$ sets X_i to a value x_i' under actuator constraints, producing interventional data $D_{\text{int}} = \{x_k^{\text{int}}\}_{k=1}^{N_{\text{int}}}$. The causal effect is estimated via truncated factorization:

$$\Delta_{ij} = \mathbb{E}[X_j \mid do(X_i = x_i')] - \mathbb{E}[X_j \mid do(X_i = x_i)].$$

The edge $i \to j$ is validated if $|\Delta_{ij}| > \epsilon$ in at least one of $n = 3$ repeats (with threshold $\epsilon = 0.1$); otherwise it is pruned. The combined dataset is

$$D_{\text{combined}} = D_{\text{obs}} \cup \{(x_k^{\text{int}}, w_k)\}_{k=1}^{N_{\text{int}}},$$

where $w_k = 2.0$ upweights interventions to reflect their higher evidential value.

Iterative refinement. The loop repeats until (i) all candidate edges are evaluated, (ii) a budget of $T_{\max}$ interventions is reached, or (iii) the learned DAG converges. Each cycle produces an auditable log of tested edges, executed actions, measured effects Δ_{ij}, and graph updates. Historical intervention cost and risk are tracked to ensure the learned world model remains interpretable and accountable.

4.2 Causal Policy Engine

The validated DAG $\hat{\mathcal{G}}$ serves as a causal world model that the policy engine uses to generate control strategies. Using the DAG and its associated structural information, the engine evaluates candidate actions to predict their expected effects on relevant objectives. For illustration in building control, these objectives include occupant comfort and energy use, with metrics such as *degree-hours (DH)* and *kilowatt-hours (kWh)*. More generally, objectives can be defined for any cyber-physical domain with measurable trade-offs.

Formally, over a prediction horizon H,

$$\text{DH} = \sum_{t=1}^{H} \Delta t \, \max\big(0, |T_{\text{zone}}(t) - T_{\text{sp}}| - \delta\big), \quad \text{kWh} = \sum_{t=1}^{H} P_{\text{HVAC}}(t)\, \Delta t,$$

where $T_{\text{zone}}(t)$, T_{sp}, δ, and $P_{\text{HVAC}}(t)$ illustrate domain-specific instantiations of the general framework.

Unlike a closed-loop controller, the engine does not execute actions directly. Instead, it evaluates them offline using $\hat{\mathcal{G}}$, selecting policies that strike a balance across objectives. Past intervention data act as regularizers: actions that are likely to cause excessive cost or risk due to spurious edges are penalized. To ensure robust policy selection, the engine applies thresholds on the estimated causal effects, requiring that the magnitude $|\Delta| > 0.05$ with $p < 10^{-3}$.

Each recommended action is directly traceable to causal pathways in $\hat{\mathcal{G}}$ and to the interventions that validated those edges. Cost and risk summaries provide transparent feedback on operational consequences, supporting accountable decision-making in safety-critical domains.

5 Experiments

We evaluate POLICYGRID in two stages: (i) we assess the framework's ability to recover causal structure across a spectrum of controlled setups, and (ii) we benchmark its performance in leveraging this structure for embodied decision-making under uncertainty. While we collected a wide range of evaluation results, for clarity and due to space constraints, the main text focuses on Structural Hamming Distance (SHD) for causal recovery and hypervolume/Pareto metrics for control; the complete set of results and comparisons is provided in Appendix A.4.

5.1 Causal Structure Recovery

To assess the causal *discovery module* within POLICYGRID, we designed six progressively complex setups (Appendix A.1): a *base simulation* with low noise and full observability; a *noisy simulation* with high-variance Gaussian perturbations; a *hidden-variable simulation* where selected confounders are latent; a real-world dataset with labeled dependencies (ASHRAE HVAC operations [Howard et al., 2019]); a *physical testbed* with office-scale equipment where controlled interventions were performed; and a *large-scale emulator* for testing scalability and edge cases. While several of these

environments involve building systems, they are intended primarily as challenging benchmarks for testing causal recovery in embodied, sensor-rich systems.

We compared POLICYGRID's discovery module against ten representative methods (Appendix A.2) spanning constraint-based, score-based, invariant prediction, bandit-based, and neural approaches, including PC, SAM, ICP, JCI, Causal Bandits, ABCD, GIES, IID, NOTEARS, and LLM-based priors. Structural recovery was evaluated using standard metrics (Appendix A.3)—Structural Hamming Distance (SHD), F1, and precision–recall—with intervention cost and operational risk additionally reported for real and physical setups. Only SHD results are shown in the main text; the remaining metrics appear in Appendix A.4.

5.2 Embodied Control Performance

We then evaluated the full POLICYGRID framework in four embodied control scenarios: the *base simulation*, the *noisy simulation*, the *hidden-variable setting*, and the *large-scale emulator*. The framework was compared against three baselines: (i) a variant without causal structure (POLICYGRID W/O DAG); (ii) a proportional–integral–derivative controller (ASHRAE-PID) tuned to standard setpoints [Standard, 1992]; and (iii) a correlation-based heuristic controller.

Evaluation considered two complementary perspectives. From an efficiency standpoint, we measured paired differences in mean resource consumption across policy pairs, with confidence intervals and significance testing (Appendix A.4.5). From a robustness standpoint, we examined operational performance via violation rate of imposed constraints and the hypervolume of the Pareto frontier. Only hypervolume and Pareto frontier results are presented in the main text; the remaining results are reported in Appendix A.4.

While the discovery module has been validated on a physical testbed, the full POLICYGRID framework was not deployed on hardware due to feasibility and safety constraints: extended embodied interventions can pose risks to both equipment and occupants. Safe deployment strategies—such as shadow-mode testing, staged rollouts, digital twins, and human-in-the-loop safeguards—remain important directions for future work.

6 Results

We evaluate POLICYGRID on our experimental setup, emphasizing: (i) fidelity of learned causal world models and (ii) policy effectiveness and operational performance across simulation scenarios under multi-objective constraints. Detailed robustness, observation-only, and ablation analyses are provided in Appendix A.4.

6.1 Causal World Model Fidelity

Figure 2 and Table 1 report Structural Hamming Distance (SHD) between learned and ground truth graphs across six setups. POLICYGRID achieves the lowest SHD in all cases, with exact recovery in *Base*, *Hidden-Variable*, and *Physical* (SHD = 0), and low error in *Noisy* (2), *ASHRAE* (1), and *Large-Sim* (13).

Baselines varied in accuracy. PC achieved relatively low SHD in simpler setups (*Base*: 4, *Physical*: 2) but degraded under complexity (*Large-Sim*: 49). IID and ABCD followed similar trends, reaching SHD of 56 and 53 in *Large-Sim*. SAM, GIES, and NOTEARS generally produced higher errors across all setups.

Overall, SHD increased with setup complexity, especially beyond the *Physical* case, reflecting the difficulty of recovering structure in larger and noisier systems. By comparison, POLICYGRID maintained lower SHD throughout, indicating that iterative interventions and physical priors improved robustness across conditions.

6.2 Policy Performance and Operational Metrics

We evaluated four controllers: ASHRAE-PID, Correlation-based, POLICYGRID without causal DAG, and full POLICYGRID on the *Base*, *Noisy*, *Hidden-Variable*, and *Large-Sim* setups. Table 2 reports two primary metrics: hypervolume (hv), which summarizes the area of the Pareto front over

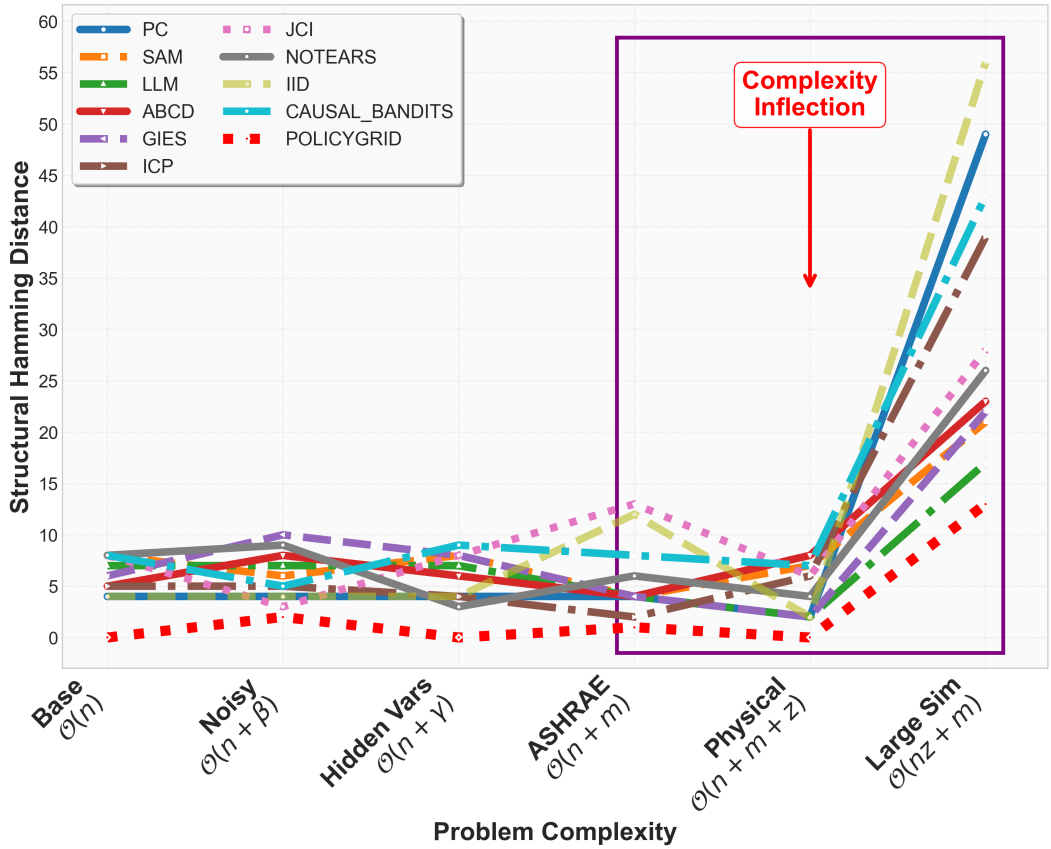


Figure 2: Structural Hamming Distance (SHD) performance comparison across eleven causal discovery methods over six experimental setups of increasing complexity. **The plot demonstrates how structural accuracy degrades as problem complexity increases, with the annotated inflection region highlighting where most traditional methods begin to fail.** PolicyGRID (red dotted line) maintains consistently low SHD values across all complexity levels, demonstrating superior robustness to challenging conditions including noise, hidden confounders, and large-scale scenarios.

energy and comfort trade-offs (higher values indicate better trade-offs), and violation rate (V%), defined as the percentage of time comfort bounds are exceeded (lower values are preferred).

Across all setups, POLICYGRID obtains the highest hypervolume and the lowest violation rates. Removing the causal DAG leads to lower hypervolume and higher violation rates, showing the importance of causal structure for policy quality. Both ASHRAE-PID and Correlation baselines remain below 10 hv and exhibit higher violations.

Figure 3 and 4 illustrate Pareto frontier points across controllers. POLICYGRID consistently places operating points in the low energy, low comfort violation region. Even at its least favorable configurations, its comfort violation levels remain below those of the best baselines. ASHRAE-PID and Correlation cluster in the high-energy, high-discomfort regime, while the DAG ablation tends to reduce energy at the cost of high comfort violations.

7 Discussions

Our results highlight the advantages of integrating causally structured world models into embodied agents operating in dynamic, sensor-rich environments. The discovery module of POLICYGRID reliably recovers DAGs across varying noise levels, latent confounders, and increasing system complexity, outperforming baseline methods that either overfit to noise or fail under hidden variables. These validated causal graphs provide a principled foundation for decision-making under uncertainty.

Table 1: Structural Hamming Distance (SHD) results for eleven causal discovery methods across six benchmark experimental setups. Bold values indicate the best (lowest) SHD performance for each setup. **Lower SHD values indicate better structural alignment with ground truth DAGs.** PolicyGRID achieves perfect or near-perfect reconstruction (SHD ≤ 2) across all setups, significantly outperforming existing methods, particularly in complex scenarios like Physical and Large-Scale simulations where traditional approaches show substantial degradation.

Method	Base	Noisy	Hidden	ASHRAE	Physical	Large-Scale
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
PolicyGRID	**0**	**2**	**0**	**1**	**0**	**13**

Table 2: Operational performance of POLICYGRID across four simulation setups. Each policy is evaluated on two metrics: (i) *violation rate*, the fraction of time comfort constraints are violated (**lower is better**), and (ii) *hypervolume*, the dominated area in the energy–comfort objective space (**higher is better**). These metrics quantify the quality of multi-objective trade-offs achieved by each policy.

Policy	Base		Noisy		Hidden-Vars		Large-Sim	
	hv↑	V↓	hv↑	V↓	hv↑	V↓	hv↑	V↓
ASHRAE	8.81	8.85	8.87	8.86	9.12	9.34	8.93	8.95
Correlation	8.81	19.87	8.76	20.81	8.79	21.13	8.82	20.78
PolicyGRID (w/o DAG)	18.72	24.13	20.42	24.21	19.87	23.98	20.41	24.24
PolicyGRID	**24.55**	**6.82**	**21.90**	**7.37**	**20.91**	**7.41**	**24.06**	**7.53**

hv=Hypervolume, V=Violation %

When incorporated into the full framework, the learned causal backbone enables adaptive trade-off management between competing objectives. Unlike correlation-driven heuristics or conservative baseline controllers, which maintain relatively fixed trade-offs, POLICYGRID actively explores the objective frontier. Methods lacking causal directionality and confounder awareness often misattribute relationships, resulting in unintended constraint violations despite apparently stable performance metrics. By contrast, the causal backbone guides policies that balance objectives reliably and predictably.

Ablation studies without the causal DAG further emphasize the importance of structural guidance: while certain performance gains may still be achieved, the absence of validated causal relationships leads to higher constraint violations and reduced reliability. Overall, these findings demonstrate that embedding causal reasoning directly into the policy loop enhances adaptability, robustness, and safety for embodied agents. Although evaluated in the context of building-like environments, the principles extend broadly to any cyber-physical or embodied system where interventions influence the environment and reliable multi-objective decision-making is critical.

8 Limitations and Future Work

While POLICYGRID demonstrates strong performance across our benchmarks, several directions remain open. Iterative interventional validation in the discovery module can be computationally demanding at scale, and approximate discovery methods or hierarchical strategies may improve efficiency without sacrificing fidelity; because POLICYGRID builds directly on the world models produced by the discovery module, advances in scalable discovery will translate to more responsive and adaptive control. Accurate sensing and intervention logs are also essential: although the framework

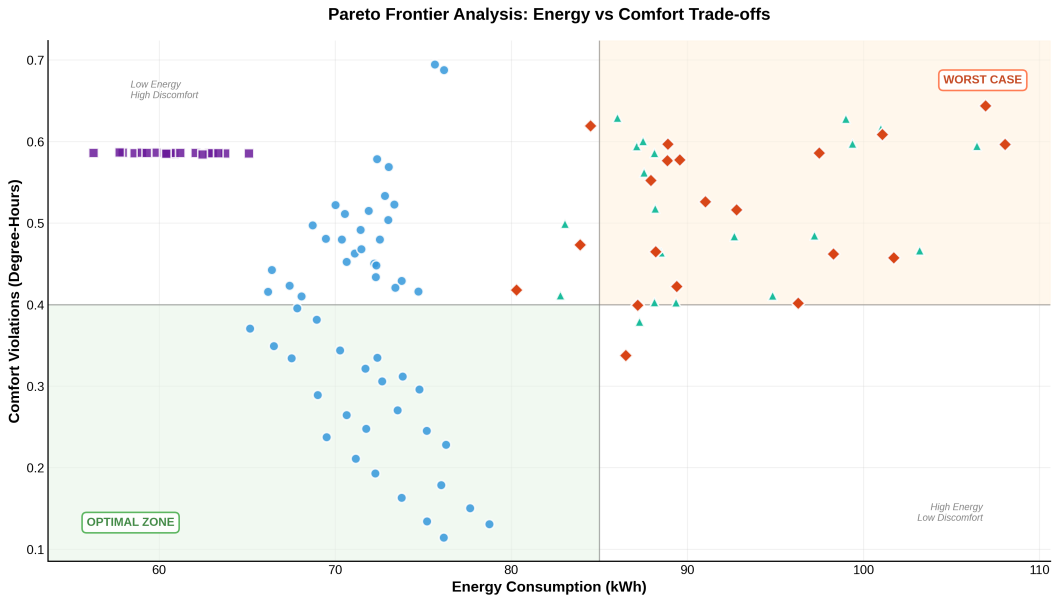


Figure 3: Pareto frontier analysis for *Large-Sim* building control scenario comparing energy consumption versus comfort violations across four control policies. PolicyGRID (blue circles) leverages causal DAG structure to achieve optimal energy-comfort trade-offs, significantly outperforming industry standard ASHRAE baseline (teal triangles), correlation-based control (orange diamonds), and ablated PolicyGRID without DAG structure (purple squares). **Lower-left region represents optimal performance zone.**

shows robustness under noise, uncertainty-aware inference and automated fault detection will further enhance reliability in real deployments. Similarly, the current approach assumes relative stationarity, and incorporating online adaptation and continual causal learning would allow POLICYGRID to adjust under non-stationary dynamics.

Finally, while our evaluation focused on energy-comfort trade-offs in building-like environments, the framework is broadly applicable to other embodied systems where interventions influence the environment. Natural next steps include extending to robotics, autonomous transportation, assistive technologies, and other cyber-physical domains, as well as broadening the policy layer to support richer multi-objective criteria, safety constraints, and human-in-the-loop feedback.

9 Conclusion

We presented POLICYGRID, a framework for embodied agents that integrates iterative causal discovery with policy generation, enabling adaptive, interpretable, and robust decision-making under uncertainty. Across a range of simulation and real-world benchmarks, POLICYGRID demonstrates that embedding causal structure directly into the policy loop improves multi-objective performance while maintaining alignment with system dynamics. These results underscore the value of causally structured world models for reasoning about interventions in partially observable, noisy, and complex environments. Future directions include multi-agent coordination, continual causal learning, and integration with richer perceptual modalities, extending POLICYGRID beyond building-like domains to general embodied AI systems that interact safely and effectively with humans and their environments.

References

Sensirion AG. Shtc3 – digital temperature and humidity sensor. https://sensirion.com/products/catalog/SHTC3, 2022. Typical accuracy: plus/minus 0.2 degrees C temperature; plus/minus 2 percent RH humidity.

Ashrae. *ASHRAE Standard 62-1989: Ventilation for Acceptable Indoor Air Quality*. Atlanta, GA, 1989.

Peter W. Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, and Koray Kavukcuoglu. Interaction networks for learning about objects, relations and physics. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4502–4510, 2016. URL https://arxiv.org/abs/1612.00222.

Lars Buesing, Theophane Weber, Yori Zwols, Sebastien Racaniere, Arthur Guez, Jean-Baptiste Lespiau, and Nicolas Heess. 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*, pages 1231–1240. PMLR, 2019. URL https://arxiv.org/abs/1811.06272.

Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pages 294–321, 2012.

Drury B Crawley, Linda K Lawrie, Frederick C Winkelmann, Walter F Buhl, Y Joe Huang, Curtis O Pedersen, Richard K Strand, Richard J Liesen, Daniel E Fisher, Michael J Witte, et al. Energyplus: creating a new-generation building energy simulation program. *Energy and buildings*, 33(4): 319–331, 2001.

Ricardo M Czekster, Roberto Metere, and Charles Morisset. Incorporating cyber threat intelligence into complex cyber-physical systems: A stix model for active buildings. *Applied Sciences*, 12(10): 5005, 2022.

Marc J Diener. Cohen's d. *The Corsini encyclopedia of psychology*, pages 1–1, 2010.

Mucong Ding, Fan Yang, Jianye Wang, Jianye Hao, Jun Zhang, and Yang Yu. Generalizing goal-conditioned reinforcement learning with variational causal reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2207.09081.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. URL https://arxiv.org/abs/1803.10122.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/1912.01603.

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

Addison Howard, Chris Balbach, Clayton Miller, Jeff Haberl, Krishnan Gowri, and Sohier Dane. Ashrae - great energy predictor iii, 2019. URL https://kaggle.com/competitions/ashrae-energy-prediction. Kaggle.

Veris Industries. Veris cw2xp2av air quality sensor datasheet. https://www.veris.com/, 2020. VOC accuracy: plus/minus 15 percent (analogous to plus/minus 15 AQI noise level).

International Organization for Standardization. Iso 7730: Ergonomics of the thermal environment. *International Standard*, 7730:1–52, 2005.

Diviyan Kalainathan, Olivier Goudet, Isabelle Guyon, David Lopez-Paz, and Michèle Sebag. Structural agnostic modeling: Adversarial learning of causal graphs. *arXiv preprint arXiv:1803.04929*, 2018.

Anjukan Kathirgamanathan, Mattia De Rosa, Eleni Mangina, and Donal P Finn. Data-driven predictive control for unlocking building energy flexibility: A review. *Renewable and Sustainable Energy Reviews*, 135:110120, 2021.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. arxiv. *arXiv preprint arXiv:2305.00050*, 2023.

Jan Kleissl and Yuvraj Agarwal. Cyber-physical energy systems: Focus on smart buildings. In *Design Automation Conference*, pages 749–754, 2010. doi: 10.1145/1837274.1837464.

Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: Learning good interventions via causal inference. *Advances in neural information processing systems*, 29, 2016.

What Is Data Mining. Data mining: Concepts and techniques. *Morgan Kaufinann*, 10(559-569):4, 2006.

Ricardo Pio Monti, Kun Zhang, and Aapo Hyvärinen. Causal discovery with general non-linear relationships using non-linear ica. In *Uncertainty in artificial intelligence*, pages 186–195. PMLR, 2020.

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

Chitra Nambiar, Michael Rosenberg, and Samuel Rosenberg. End use analysis of ansi/ashrae/ies standard 90.1-2019. *ASHRAE Journal*, 65(4):34–42, 2023.

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

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT press, 2017.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, AH Miller, and Sebastian Riedel. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 2463–2473, Hong Kong, China, 01 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250.

Jonathan Richens, David Abel, Alexis Bellot, and Tom Everitt. 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, 2025. URL https://arxiv.org/abs/2506.01622.

Yves Rosseel and Wen Wei Loh. A structural after measurement approach to structural equation modeling. *Psychological Methods*, 2022.

Sasan Sadrizadeh, Runming Yao, Feng Yuan, Hazim Awbi, William Bahnfleth, Yang Bi, Guangyu Cao, Cristiana Croitoru, Richard De Dear, Fariborz Haghighat, et al. Indoor air quality and health in schools: A critical review for developing the roadmap for the future school environment. *Journal of Building Engineering*, 57:104908, 2022.

John Ervin Seem. *Modeling of heat transfer in buildings*. PhD thesis, The University of Wisconsin-Madison, 1987.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

ASHRAE Standard. Thermal environmental conditions for human occupancy. *ANSI/ASHRAE, 55*, 5, 1992.

Zhuofan Sun and Qingyi Li. Leveraging llms for causal inference and discovery. *arXiv preprint arXiv:2410.16676*, 2024.

Christian Toth, Lars Lorch, Christian Knoll, Andreas Krause, Franz Pernkopf, Robert Peharz, and Julius Von Kügelgen. Active bayesian causal inference. *Advances in Neural Information Processing Systems*, 35:16261–16275, 2022.

Jiajun Wu, Ilker Yildirim, Joseph J. Lim, William T. Freeman, and Joshua B. Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 127–135, 2015. URL `https://arxiv.org/abs/1505.00333`.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, et al. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv preprint arXiv:2303.10420*, 2023.

Chaobo Zhang, Yazhou Zhao, Yang Zhao, Tingting Li, and Xuejun Zhang. Causal discovery and inference-based fault detection and diagnosis method for heating, ventilation and air conditioning systems. *Building and Environment*, 212:108760, 2022.

Jiaqi Zhang, Louis Cammarata, Chandler Squires, Themistoklis P Sapsis, and Caroline Uhler. Active learning for optimal intervention design in causal models. *Nature Machine Intelligence*, 5(10): 1066–1075, 2023.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*, pages 3414–3425. Pmlr, 2020.

A Appendix

A.1 Simulation Setup Details

We evaluated POLICYGRID's causal discovery module against all benchmark methods listed in Section 5 across six progressively complex setups, spanning synthetic simulations, dataset-driven benchmarks, and physical deployments. Each setup was designed to test robustness under varying levels of observability, stochasticity, latent confounding, and system scale. Synthetic and dataset-driven setups (Base, Noisy, Hidden, ASHRAE) were run for 60 iterations, while the physical setup used 15 iterations due to hardware constraints. To quantify causal complexity across setups, we define a symbolic score

$$\mathcal{C} = n + \alpha m + \beta r + \gamma h + \delta z,$$

where n is the number of observed variables, m the number of intervenable variables, and $r, h, z \in \{0, 1\}$ indicate the presence of noise, hidden confounders, and spatial coupling, respectively. The weights α–δ encode the relative difficulty contributed by each factor. This metric enables structured, quantitative comparisons of causal recovery performance across diverse embodied-system benchmarks (Table 3), reflecting realistic challenges in partially observable, noisy, and intervention-rich environments [Fisher et al., 2019, Zhang et al., 2022, Peters et al., 2017].

Table 3: Complexity Characterization of Experimental Setups

Setup	Complexity Expression	Big O Notation
Base	n	$\mathcal{O}(n)$
Noisy	$n + \beta$	$\mathcal{O}(n + \beta)$
Hidden	$n + \gamma$	$\mathcal{O}(n + \gamma)$
ASHRAE	$n + \alpha m + \beta + \gamma$	$\mathcal{O}(n + m)$
Physical	$n + \alpha m + \beta + \gamma + \delta$	$\mathcal{O}(n + m + z)$
Large-Scale	$nz + \alpha m + \beta + \gamma + \delta$	$\mathcal{O}(nz + m)$

A.1.1 Base Simulation (5 Variables)

Complexity: $\mathcal{O}(n)$

The base setup evaluated performance under ideal conditions using a fully observable smart HVAC simulation with five variables: temperature (T), humidity (H), air quality (AQ), energy consumption (E), and occupant satisfaction (S). A ground truth DAG was manually defined based on domain knowledge. Data were generated via a custom simulator enabling batch runs and direct interventions. Energy use was estimated by nearest-neighbor interpolation over EnergyPlus data [Crawley et al., 2001], with fallbacks from ASHRAE 90.1 [Nambiar et al., 2023] and Seem's part-load model [Seem, 1987]. Models incorporated temperature drift, ASHRAE 62.1 humidity control [Ashrae, 1989], and EPA-based ventilation energy [Sadrizadeh et al., 2022]. Occupant satisfaction followed ISO 7730 [International Organization for Standardization, 2005], combining PMV/PPD metrics with psychrometric inputs, and was penalized for excess energy use.

Table 4: Base Simulation tracks temperature, humidity, and air quality (inputs) alongside energy use and occupant satisfaction (outputs); each variable has set units and ranges, and the first three causally drive the last two.

Variable Type	Variable	Range and Units
Input Variables	Temperature (T)	18-30°C
	Humidity (H)	30-70%
	Air Quality (AQ)	0-500 AQI
Output Variables	Energy Consumption (E)	0-100% (normalized index)
	Overall Satisfaction (S)	0-100%

A.1.2 Noisy Simulation (5 Variables)

Complexity: $\mathcal{O}(n + \beta)$

This setup extended the base by adding Gaussian noise to sensor readings to test robustness against realistic measurement uncertainty. Noise levels matched typical sensor specs: ±0.2°C for temperature, ±2% RH for humidity [AG, 2022], and ±15 AQI for air quality [Industries, 2020]. Noise affected only observed values, keeping control variables precise to mimic real automation. The temperature range was also expanded to 18-40°C to evaluate stability under extreme conditions.

A.1.3 Simulation with Hidden Variables (5 Variables)

Complexity: $\mathcal{O}(n + \gamma)$

This setup added latent confounders in the base setup to simulate partial observability typical in real buildings. Hidden variables such as HVAC efficiency, building envelope properties, occupancy patterns, window states, and outdoor conditions were unobserved by the algorithms but influenced observed variables and outcomes. Energy consumption and occupant satisfaction reflected time-varying effects from building physics and occupancy-driven demand, including adaptive comfort and window use.

A.1.4 Real-World Dataset (ASHRAE; 5 Variables)

Complexity: $\mathcal{O}(n + m)$

We used the ASHRAE Great Energy Predictor III dataset [Howard et al., 2019] to evaluate POLICYGRID and the other benchmark methods on real-world energy data from over 1,400 buildings. Six physical variables were selected: *outdoor temperature, dew point, pressure, energy use, square footage,* and *construction year*. Preprocessing involved daily aggregation, weather-building merging, KNN imputation, outlier removal, and robust scaling. A ground truth DAG was defined using domain expertise and physical laws. As real interventions were unavailable, we used Random Forest surrogates trained on observational data to simulate interventions and predict effects on causal children.

A.1.5 Physical Deployment (5 Variables)

Complexity: $\mathcal{O}(n + m + z)$

This setup validated POLICYGRID and the benchmark methods under real-world hardware constraints using environmental sensors, power monitors, and standardized comfort tools in a controlled space.

Two Govee H5179 sensors measured temperature (±0.3°C) and humidity (±3%), while a BME680 provided additional readings including IAQ. Three Kasa KP125M plugs monitored power with 0.1 W resolution, reporting cumulative energy usage with 1% accuracy, and acted as actuators. PMV/PPD comfort scores followed ISO 7730 [International Organization for Standardization, 2005] using fixed occupant parameters. Satisfaction combined thermal and air quality metrics, following ISO 7730 standards [International Organization for Standardization, 2005]. Interventions were capped at 1000/day, spaced by $\geq$ 300s with a 600s stabilization window. Effects were quantified using Cohen's d [Diener, 2010]. Data were time-synced, Govee readings averaged, and energy consumption data from the Kasa plugs summed. Preprocessing used KNN imputation (k=5), IQR-based outlier removal, and Min-Max normalization [Mining, 2006]. While this setup demonstrates feasibility, scaling interventions in occupied buildings remains a practical constraint and is a direction for future deployment work.

A.1.6 Large-Scale Simulation (13 Variables)

Complexity: $\mathcal{O}(nz + m)$ Five inter-connected EnergyPlus zones (13 state vars each) expose control of temperature, humidity, IAQ, occupancy, HVAC set-points, and lighting. A central coordinator issues zone- and building-level interventions, captures full state, and aggregates energy, comfort, IAQ, and satisfaction metrics, reusing the single-room psychrometric and energy models while realistic schedules drive coupled dynamics for benchmarking.

A.2 Baseline Methods

To evaluate POLICYGRID, we benchmarked it against ten baseline causal discovery methods spanning constraint-based, score-based, Bayesian, invariance-based, optimization-driven, and LLM-based

approaches. Each was run under identical conditions with standardized datasets, iteration budgets, and evaluation metrics (Section A.3). Implementations came from verified sources (e.g., *causal-learn*, *CDT*) with default or grid-tuned parameters.

Table 5 summarizes the baseline methods against which POLICYGRID was benchmarked. All implementations were drawn from official repositories or verified third-party libraries (e.g., causal-learn, CDT), with default or grid-tuned parameters used consistently across runs.

Table 5: Overview of benchmark methods used for causal discovery and intervention planning. The table summarizes ten representative approaches spanning constraint-based, score-based, neural, and language model-driven techniques.

Method	Description	Ref
PC	Constraint-based method using conditional independence tests to recover causal skeleton and orientations.	Spirtes et al. [2000]
SAM	Structural agnostic model learning causal graphs via adversarial training and sparsity constraints.	Kalainathan et al. [2018]
GIES	Score-based algorithm extending Greedy Equivalence Search to interventional settings.	Hauser and Bühlmann [2012]
JCI	Unified framework treating interventions as observed variables for joint learning.	Mooij et al. [2020]
ABCD	Active Bayesian approach using expected information gain for iterative interventions.	Toth et al. [2022]
Causal Bandits	Sequential decision-making using bandit feedback to optimize interventions.	Lattimore et al. [2016]
NOTEARS-I	Continuous optimization extending NOTEARS for interventional data.	Zheng et al. [2020]
ICP	Identifies causal predictors invariant across multiple environments.	Peters et al. [2016]
IID	Active learning selecting interventions based on entropy reduction.	Zhang et al. [2023]
LLM	Large language models proposing causal edges via domain knowledge reasoning.	Sun and Li [2024]

A.3 Evaluation Metric Details

Learned graphs and policies from POLICYGRID were evaluated along two dimensions: (i) structural fidelity to a known or reference causal model, and (ii) practical implications of incorrect causal assumptions for downstream decision-making.

A.3.1 Structural Accuracy

Structural accuracy metrics quantify how closely the learned DAG $\hat{G} = (\hat{V}, \hat{E})$ approximates the ground truth DAG $G^* = (V^*, E^*)$. True positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) were computed based on edge presence and orientation. These values provide a foundation for further structural metrics.

A.3.2 Structural Hamming Distance (SHD)

SHD measures the total number of edge modifications—additions, deletions, or reversals—required to convert $\hat{G}$ into G^*. Formally,

$$\text{SHD}(G^*, \hat{G}) = \sum_{i,j} |A^*_{ij} - \hat{A}_{ij}|,$$

where A^* and $\hat{A}$ are the adjacency matrices of the ground truth and learned DAGs.

A.3.3 Cost and Risk Metrics

To evaluate the operational impact of false positive edges, we define cost and risk metrics based on historical or simulated interventions. For each false positive edge e, the edge-level cost is

$$\text{edge\_cost}(e) = \frac{1}{|I_e|} \sum_{i \in I_e} \left(\alpha \cdot \text{sat\_loss}(i) + \beta \cdot \text{energy\_increase}(i)\right),$$

where I_e is the set of interventions associated with edge e, $\alpha = \beta = 0.6$, and the components are

$$\text{sat\_loss}(i) = \max\left(0, \frac{S_{\text{pre},i} - S_{\text{post},i}}{S_{\text{pre},i}}\right) \cdot \mathbf{1}_{S_{\text{pre},i} > 0}, \quad \text{energy\_increase}(i) = \max\left(0, \frac{E_{\text{post},i} - E_{\text{pre},i}}{E_{\text{pre},i}}\right) \cdot \mathbf{1}_{E_{\text{pre},i} > 0}.$$

The aggregate cost and confidence-weighted risk for method m are defined as

$$\text{Cost(m)} = \begin{cases} \frac{1}{|\hat{E}_m \setminus E^*|} \sum_{e \in \hat{E}_m \setminus E^*} \text{edge\_cost}(e), & |\hat{E}_m \setminus E^*| > 0 \\ 0, & \text{otherwise} \end{cases}$$

$$\text{Risk(m)} = \begin{cases} \frac{1}{|\hat{E}_m \setminus E^*|} \sum_{e \in \hat{E}_m \setminus E^*} \text{confidence}(e) \cdot \text{edge\_cost}(e), & |\hat{E}_m \setminus E^*| > 0 \\ 0, & \text{otherwise} \end{cases}$$

These metrics were computed for intervention-capable methods (e.g., GIES, ABCD, JCI, ICP, Causal Bandits, IID, NOTEARS) using interventions observed in the system. Effects were filtered to include only those with magnitude $|\Delta| > 0.05$ and t-test confidence $p < 0.001$, corresponding to confidence $= 0.9$.

A.3.4 Pareto Frontiers

For multi-objective evaluation, we construct Pareto frontiers over the set of candidate policies Π with objective vectors $\{\mathbf{f}_\pi\}_{\pi \in \Pi}$. A solution $\mathbf{f}_\pi$ is Pareto optimal if no other policy improves one objective without worsening another:

$$\mathcal{P} = \{\mathbf{f}_\pi \in \{\mathbf{f}_\pi\} \,|\, \mathbf{f}_{\pi'} \text{ s.t. } \mathbf{f}_{\pi'} \leq \mathbf{f}_\pi \text{ and } \mathbf{f}_{\pi'} \neq \mathbf{f}_\pi\}.$$

A.3.5 Hypervolume

The hypervolume metric quantifies the volume of objective space dominated by the Pareto-optimal solutions relative to a reference point $\mathbf{r}$ dominated by all solutions:

$$\text{HV}(\mathcal{P}, \mathbf{r}) = \text{Volume}\Big(\bigcup_{\mathbf{f} \in \mathcal{P}} [f_1, r_1] \times \cdots \times [f_m, r_m]\Big),$$

where m is the number of objectives. Higher hypervolume indicates better overall multi-objective performance, allowing direct comparison of the effectiveness of different policy sets.

A.4 Extended Experimental Results

A.4.1 Observation-Only Performance

Table 6 summarizes observation-only results for 11 methods across six experimental setups. POLICYGRID-O, which applies the full POLICYGRID discovery module (merging PC, SAM,

and LLM outputs) without interventions or iterative refinement, performs competitively across scenarios. It achieves the best or near-best SHD and F1 scores in multiple settings, including the Noisy, Hidden Variables, and Physical setups. These observation-only findings illustrate that ensemble learning improves causal structure recovery from passive data but remains limited under noise, latent confounders, and partial observability. Cost and risk metrics are omitted here because they rely on active interventions. Subsequent sections show that full POLICYGRID discovery module, leveraging targeted interventions, consistently outperforms these passive baselines.

Table 6: Observation-only performance of 11 methods across six scenarios. For each method and scenario, we report SHD (**lower is better**) and F1 score (**higher is better**).

Method	Base		Noisy		Hidden Vars		ASHRAE		Physical		Large Sim	
	SHD	F1	SHD	F1	SHD	F1	SHD	F1	SHD	F1	SHD	F1
PC	4	0.60	4	0.60	4	0.60	4	0.6	2	0.88	49	0.33
SAM	8	0.33	6	0.25	8	0.20	4	0.33	7	0.36	21	0.16
LLM	7	0.36	7	0.36	7	0.36	4	0.6	2	0.86	17	0.45
GIES	4	0.6	4	0.60	4	0.67	5	0.29	5	0.71	19	0.34
JCI	6	0.40	6	0.50	4	0.67	2	0.75	6	0.50	48	0.37
ABCD	5	0.44	6	0.50	6	0.25	2	0.75	4	0.71	54	0.41
Causal Bandits	4	0.60	4	0.67	4	0.75	3	0.67	5	0.71	48	0.37
ICP	5	0.44	5	0.55	2	0.80	3	0.67	6	0.50	23	0.26
IID	2	0.8	6	0.57	5	0.62	2	0.80	6	0.50	28	0.15
NOTEARS	3	0.73	6	0.25	4	0.67	4	0.5	6	0.57	44	0.46
PolicyGRID-O	5	0.62	2	0.86	3	0.8	4	0.67	2	0.86	37	0.43

Note: SHD = Structural Hamming Distance, F1 = F1 Score

A.4.2 Core Metrics Results

Table 7: Precision, Recall, and F1 score for causal edge recovery across six experimental setups: *Base, Noisy, Hidden, ASHRAE, Physical*, and *Large-Sim*. Each method was evaluated on its ability to recover the true causal graph under varying data conditions. Bolded values indicate the highest F1 score per setup. *Note:* P = Precision, R = Recall, F1 = F1 Score

Method	Base			Noisy			Hidden			ASHRAE			Physical			Large-Sim		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
PC	0.75	0.50	0.60	0.75	0.50	0.60	0.67	0.67	0.67	0.67	0.40	0.50	1.00	0.75	0.88	0.24	0.55	0.33
SAM	0.33	0.33	0.33	0.50	0.17	0.25	0.25	0.17	0.20	1.00	0.20	0.33	0.67	0.25	0.36	0.67	0.09	0.16
LLM	0.40	0.33	0.36	0.40	0.33	0.36	0.40	0.33	0.36	0.60	0.60	0.60	1.00	0.75	0.86	0.78	0.32	0.45
GIES	0.50	0.33	0.40	0.25	0.33	0.29	0.38	0.50	0.43	0.67	0.40	0.50	0.80	1.00	0.89	0.50	0.45	0.48
JCI	0.40	0.67	0.50	0.67	1.00	0.80	0.38	0.50	0.43	0.10	0.20	0.13	0.75	0.38	0.50	0.43	0.86	0.58
ABCD	0.67	0.33	0.44	0.00	0.00	0.00	0.50	0.33	0.40	0.67	0.40	0.50	0.50	0.25	0.34	0.48	0.73	0.58
C. Bandits	0.33	0.33	0.33	0.57	0.67	0.62	0.20	0.17	0.18	0.25	0.20	0.22	0.67	0.25	0.36	0.11	0.14	0.12
ICP	0.67	0.33	0.44	0.67	0.33	0.44	0.75	0.50	0.60	0.80	0.80	0.80	0.75	0.38	0.50	0.16	0.18	0.17
IID	0.60	1.00	0.75	0.60	1.00	0.75	0.60	1.00	0.75	0.27	0.80	0.40	0.80	1.00	0.89	0.28	1.00	0.44
NOTEARS	0.25	0.17	0.20	0.33	0.50	0.40	0.71	0.83	0.77	0.33	0.20	0.25	0.83	0.63	0.71	0.43	0.59	0.50
POLICYGRID	1.00	1.00	**1.00**	0.75	1.00	**0.86**	1.00	1.00	**1.00**	1.00	0.80	**0.89**	1.00	1.00	**1.00**	0.80	0.55	**0.65**

Table 7 reports precision, recall, and F1 scores across all six setups. POLICYGRID discovery module consistently achieves the highest F1 scores, including perfect recovery in *Base*, *Hidden*, and *Physical*, and strong performance in more challenging scenarios like *Noisy* (0.86), *ASHRAE* (0.89), and *Large-Sim* (0.65). While POLICYGRID performs well on graphs with $\tilde{1}3$ variables, scaling to higher-dimensional or highly coupled systems will require architectural and runtime optimization.

Performance among baselines varies. PC performs well in simpler setups (F1 = 0.88 in *Physical*) but drops under scale (*Large-Sim*, 0.33). LLM-based priors peak in *Physical* (0.86) but are less consistent elsewhere. JCI shows mixed performance, with moderate scores in *Noisy* (0.80) and *Large-Sim* (0.58), but low performance in *ASHRAE* (0.13). SAM and Causal Bandits generally remain below 0.40 across most setups. IID and ICP achieve higher recall but lower precision, with F1 peaking at

0.89 (*Physical*) and 0.80 (*ASHRAE*), respectively. Overall, POLICYGRID demonstrates the most consistent and effective performance across diverse conditions.

A.4.3 Risk–Cost Analysis

Table 8 shows normalized intervention risk and cost across the six setups. POLICYGRID discovery module records zero risk and cost in *Base*, *Hidden*, *ASHRAE*, and *Physical*, and near-zero values in *Noisy* (risk = 0.036, cost = 0.022) and *Large-Sim* (risk = 0.001, cost = 0.026). Baseline methods exhibit higher risk and cost across multiple scenarios. For example, ABCD incurs the highest cost in *ASHRAE* (0.885), and GIES shows elevated risk in *Noisy* (0.452). These outcomes indicate that POLICYGRID consistently selects low-risk, low-cost interventions, even under noisy, partially observable, or complex settings.

Table 8: Normalized intervention risk and cost for each method across six experimental setups: *Base, Noisy, Hidden, ASHRAE, Physical*, and *Large-Sim*. Lower values indicate fewer risks and lower costs associated with the interventions selected by each causal discovery approach. Bold values indicate the lowest risk or cost in each setup.

Method	Base		Noisy		Hidden		ASHRAE		Physical		Large-Sim	
	Risk	Cost	Risk	Cost	Risk	Cost	Risk	Cost	Risk	Cost	Risk	Cost
GIES	0.439	0.472	0.452	0.486	0.056	0.06	0.054	0.067	0.175	0.218	0.033	0.046
JCI	0.447	0.480	0.470	0.505	0.066	0.071	0.196	0.245	0.269	0.336	0.05	0.1
ABCD	0.460	0.495	0.439	0.472	0.083	0.089	0.708	0.885	0.161	0.201	0.018	0.030
Causal Bandits	0.443	0.476	0.453	0.487	0.053	0.057	0.24	0.3	0.014	0.018	0.031	0.1
ICP	0.05	0.1	0.05	0.1	0.05	0.1	0.1	0.1	0.250	0.312	0.024	0.037
IID	0.449	0.483	0.445	0.479	0.074	0.080	0.182	0.227	0.151	0.188	0.024	0.037
NOTEARS	0.441	0.475	0.455	0.490	0.033	0.036	0.1	0.1	0.08	0.1	0.024	0.038
POLICYGRID	**0.000**	**0.000**	**0.036**	**0.022**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.001**	**0.026**

Note: Lower Risk and Cost values are better

A.4.4 Ablation Studies

We performed ablations on the *Large-Sim* setup to quantify the contribution of key components of POLICYGRID discovery module. As shown in Table 9, removing components such as the edge ranker or LLM-guided interventions consistently reduces F1, increases SHD, and raises intervention cost. The ranking mechanism focuses exploration on high-confidence edges, while LLM-guided interventions reduce unnecessary actions. A minimal system, which disables all intelligent discovery components, highlights the performance gap and demonstrates the complementary value of each design element.

A.4.5 Paired Differences with Statistical Tests

Paired comparisons at matched operating points between POLICYGRID and baseline methods are summarized in Table 10. Each entry reports the mean ΔkWh, 95% confidence interval, and p-value. Across all setups, POLICYGRID consistently reduces energy consumption relative to ASHRAE-PID, correlation-based, and ablated (POLICYGRID w/o DAG) policies. Differences are statistically significant in nearly all cases ($p \leq 0.05$), with mean reductions ranging from approximately 0.45 to 15.7 kWh. These results demonstrate the robustness and effectiveness of POLICYGRID in embodied decision-making under varied operating conditions, beyond building-specific tasks.

A.5 LLM Prompting

POLICYGRID uses GPT-3.5-turbo [Ye et al., 2023] to generate causal graphs and interventions grounded in physical building principles. A fixed system message defines the model's role as a causal discovery and intervention design expert and enforces a strict JSON output format, while dynamic user messages specify variables and constraints like acyclicity and physical plausibility (see example in A.5 for hypothesis generation prompt and A.5 for intervention design prompt). The LLM infers causal hypotheses by reasoning over variable names, descriptions, and units [Petroni et al., 2019,

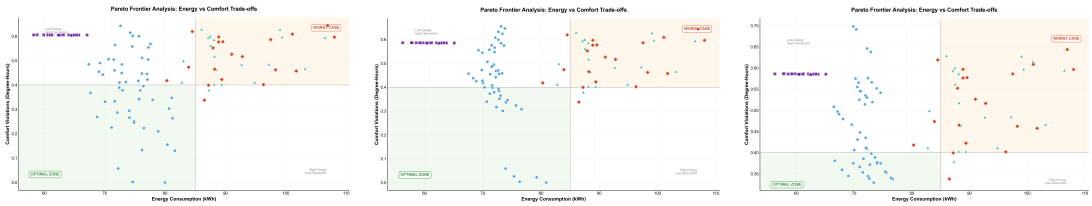
Table 9: Ablation study on *Large-Sim* for the discovery module of POLICYGRID. Top: hypothesis generator ablations. Middle: intervention component ablations. Bottom: minimal system vs. full discovery module of POLICYGRID.

Configuration	Prec. ↑	Rec. ↑	F1 ↑	SHD ↓	Cost ↓	Risk ↓
PC + LLM	0.778	0.318	0.452	17	0.222	0.049
PC + SAM	0.500	0.318	0.389	22	0.500	0.250
PC only	0.667	0.273	0.387	19	0.333	0.111
LLM only	0.800	0.182	0.296	19	0.200	0.040
SAM + LLM	0.500	0.136	0.214	22	0.500	0.250
SAM only	0.000	0.000	0.000	24	1.000	1.000
No edge ranking	0.609	0.636	0.622	17	0.391	0.153
No LLM interventions	0.522	0.545	0.533	21	0.478	0.229
No edge validation	0.579	0.500	0.537	19	0.421	0.177
No dataset update	0.636	0.318	0.424	19	0.364	0.132
No ranking + intervention	0.571	0.545	0.558	19	0.429	0.184
No intervention + update	0.684	0.591	0.634	15	0.316	0.100
Minimal system	0.583	0.318	0.412	20	0.417	0.174
PolicyGRID (discovery module)	**0.800**	**0.550**	**0.650**	**13**	**0.026**	**0.001**

Table 10: Paired differences in energy consumption (ΔkWh) between POLICYGRID and baseline methods across four simulation setups. Each entry reports the mean ΔkWh, 95% confidence interval (CI), and p-value, highlighting statistically significant improvements of POLICYGRID over alternative policies.

Setup	Policy Pair	Mean ΔkWh	95% CI	p-value
Base	POLICYGRID vs ASHRAE	15.624	[15.574, 15.674]	0.050
	POLICYGRID vs POLICYGRID (w/o DAG)	5.836	[5.786, 5.886]	0.050
	POLICYGRID vs Correlation	15.729	[15.679, 15.779]	0.050
	ASHRAE vs POLICYGRID (w/o DAG)	-9.788	[-9.838, -9.738]	0.050
	ASHRAE vs Correlation	0.105	[0.055, 0.155]	0.050
	POLICYGRID (w/o DAG) vs Correlation	9.893	[9.843, 9.943]	0.050
Noisy	POLICYGRID vs ASHRAE	12.936	[12.886, 12.986]	0.050
	POLICYGRID vs POLICYGRID (w/o DAG)	1.459	[1.409, 1.509]	0.050
	POLICYGRID vs Correlation	13.041	[12.991, 13.091]	0.050
	ASHRAE vs POLICYGRID (w/o DAG)	-11.477	[-11.527, -11.427]	0.050
	ASHRAE vs Correlation	0.105	[0.055, 0.155]	0.050
	POLICYGRID (w/o DAG) vs Correlation	11.582	[11.532, 11.632]	0.050
Hidden-Vars	POLICYGRID vs ASHRAE	11.924	[11.874, 11.974]	0.050
	POLICYGRID vs POLICYGRID (w/o DAG)	0.447	[0.397, 0.497]	0.050
	POLICYGRID vs Correlation	12.029	[11.979, 12.079]	0.050
	ASHRAE vs POLICYGRID (w/o DAG)	-11.477	[-11.527, -11.427]	0.050
	ASHRAE vs Correlation	0.105	[0.055, 0.155]	0.050
	POLICYGRID (w/o DAG) vs Correlation	11.582	[11.532, 11.632]	0.050
Large-Sim	POLICYGRID vs ASHRAE	11.423	[11.373, 11.473]	0.050
	POLICYGRID vs POLICYGRID (w/o DAG)	-0.053	[-0.103, -0.003]	0.150
	POLICYGRID vs Correlation	11.529	[11.479, 11.579]	0.050
	ASHRAE vs POLICYGRID (w/o DAG)	-11.477	[-11.527, -11.427]	0.050
	ASHRAE vs Correlation	0.105	[0.055, 0.155]	0.050
	POLICYGRID (w/o DAG) vs Correlation	11.582	[11.532, 11.632]	0.050

Kıcıman et al., 2023], producing a causal graph $\mathcal{G}_{\text{LLM}}$. It is configured with temperature 1.0 and top-p 0.8 to balance creativity and domain accuracy.



(a) Base **(b)** Noisy **(c)** Hidden-Vars

Figure 4: Pareto frontier comparison across Base, Noisy, and Hidden-Vars setups with same control policies and legend as Figure 3.

LLM Prompting for Base Experimental Setup

System Message:

```
You are an expert in causal discovery analyzing a smart room environment with
     5 variables.
Generate a causal DAG based on physical principles and environmental systems.
Return your answer as a JSON object with this format exactly:
{"nodes": ["Temperature", "Humidity", "AirQuality", "EnergyConsumption", "
    OverallSatisfaction"],
 "edges": [["source", "target"], ...]}
```

User Message:

```
Analyze this dataset with variables: Temperature, Humidity, AirQuality,
    EnergyConsumption, OverallSatisfaction

Rules:
1. Include directed edges based on likely causal mechanisms
2. No cycles or self-loops allowed
3. Focus on primary physical relationships
```

LLM Prompting for Intervention Design

System Message:

```
Design a physical intervention to test if {source} causes changes in {target}.


DEVICE CONSTRAINTS:
- Heater: ON/OFF only (affects Temperature)
- Humidifier: ON/OFF only (affects Humidity)
- Fan: ON/OFF only (affects Temperature, Humidity, AirQuality)

Return JSON:
{{
    "variables": [{{
        "variable": "{source}",
        "action": "increase/decrease"
    }}],
    "expected_effects": {{
        "{target}": "increase/decrease"
    }},
    "reasoning": "Brief_explanation"
}}
```

User Message:

```
Design intervention to test {source} --> {target} causality using binary
    device control.
```

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 clearly state the central contributions: introducing POLICYGRID, integrating causal discovery (GRID) with embodied control, and evaluating across six benchmarks. The claims align with theoretical framing and experimental results.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 8 discusses computational intensity of iterative interventions, sensitivity to sensing quality, assumptions of stationarity, and the scope of evaluation.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not provide new theoretical proofs; instead it develops and evaluates a practical causal discovery and control framework. Formalism (e.g., causal DAG definitions, intervention cost functions) is stated in Section 3, but no theorems or proofs are claimed.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper?

   Answer: [Yes]

   Justification: Section 5 and Appendix A provide dataset descriptions, simulation setups, evaluation metrics (e.g., SHD, hypervolume, violation rate), and baseline comparisons. Experimental setups (base, noisy, hidden variable, large-scale simulations, ASHRAE dataset, physical testbed) are described in sufficient detail.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results?

   Answer: [Yes]

   Justification: An anonymized code and data repository link will be made available upon acceptance and is planned for the camera-ready version. Public datasets (e.g., ASHRAE Great Energy Predictor III) are already openly accessible, and full instructions for reproducing results will be provided in the repository.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.)?

   Answer: [Yes]

   Justification: Hyperparameters for neural SEMs, intervention budgets, and optimizer choices are reported in Section 4. For public datasets, the train/test splits follow standard protocols (e.g., ASHRAE competition splits).

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Results include 95% confidence intervals and p-values for paired comparisons (see Table 10). Error bars are defined over repeated runs with random seeds. Figure 2 which reports Structural Hamming Distance (SHD) between learned and ground truth graphs across six setups, does not include error bars because they would obfuscate the main takeaway from the figure.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: All experiments were run on standard desktop and workstation machines without requiring specialized compute (e.g., cloud clusters or large GPU farms). Since results can be reproduced on regular hardware, we did not include detailed compute resource reporting.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics?

   Answer: [Yes]

   Justification: The work complies with the NeurIPS Code of Ethics. Experiments are on simulated and publicly available datasets, plus a controlled office testbed with no human subject data.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Section 7 highlights positive impacts (energy efficiency, interpretability, robustness in embodied systems) and potential risks (over-reliance on faulty sensors, privacy concerns if extended to occupant modeling). Mitigation strategies are discussed.

11. **Safeguards**

    Question: Does the paper describe safeguards for responsible release of data or models that have a high risk for misuse?

    Answer: [NA]

    Justification: The released models and code would not pose dual-use risks such as generative misuse. Only causal discovery and control modules for physical systems are to be released.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models) properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Public datasets (ASHRAE) are cited with original licenses, and baselines (PC, NOTEARS, ICP, SAM, GIES) are properly referenced with software citations.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The live office testbed setup is detailed in Appendix A.4 and the dataset will be released with metadata, sensor specifications, and intervention protocols as noted earlier.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and details about compensation?

    Answer: [NA]

Justification: The work does not involve crowdsourcing or human subject experiments. Only building sensor data and simulations are used.

15. **Institutional review board (IRB) approvals**

Question: Does the paper describe potential risks incurred by study participants, and whether IRB approvals were obtained?

Answer: [NA]

Justification: The work does not involve human subjects. All experiments are with simulated environments or non-identifiable building data.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

Answer: [Yes]

Justification: Section 4 and Appendix A.5 note that domain-informed priors are incorporated using LLM-guided constraints. This usage is central to GRID's discovery loop and is fully described. No LLMs were used for writing beyond editing assistance.