Taqiya Ehsan

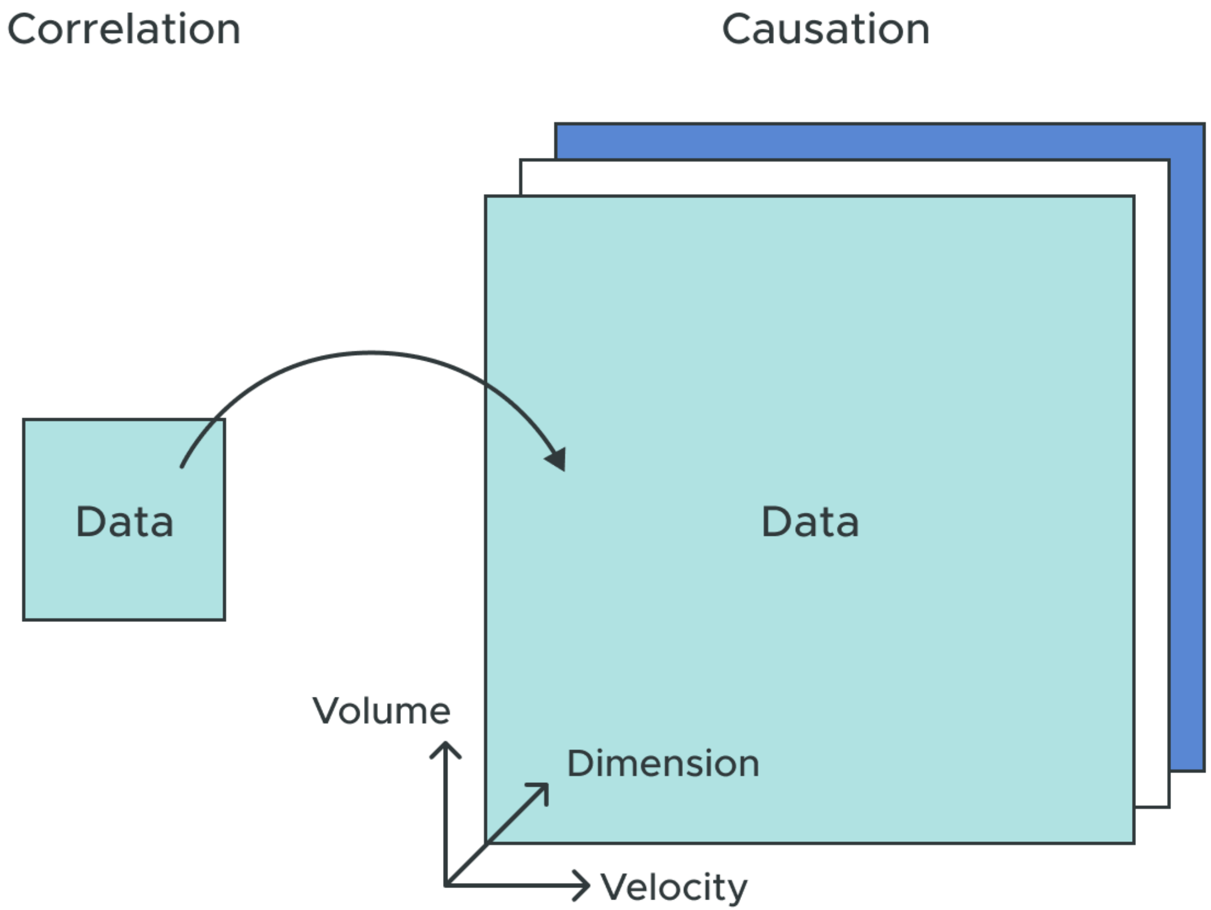
11/04/2025

Defense Against the Dark Correlations

Introduction to Causal Reasoning and the `PolicyGRID` Framework

Correlation vs Causation

Correlation vs Causation

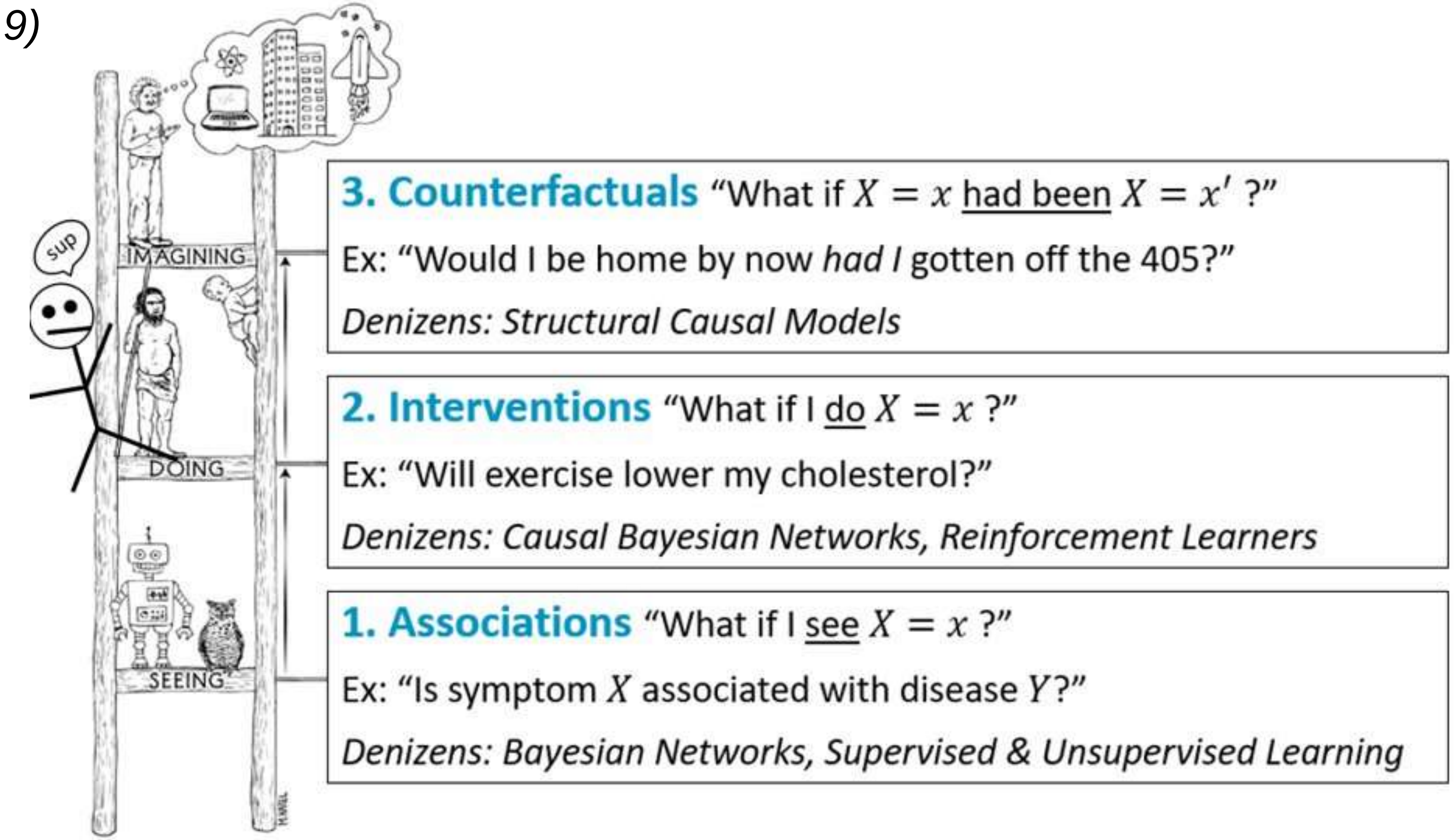


- correlation != causation
- models fail when agents intervene
- current systems either:
  - ignore causal structures
  - treat causality as an offline pre-processing step

The Causal Ladder

Levels of Reasoning *(Pearl, 2019)*

- Causal reasoning is the ability to uncover cause-and-effect relationships to explain events and predict outcomes.



Pearl, Judea. "*Models, reasoning and inference*." Cambridge, UK: Cambridge University Press 19.2 (2000): 3.
Pearl, Judea, and Dana Mackenzie. "*The book of why: the new science of cause and effect*." Basic books, 2018.

Structural Causal Model (SCM)

formalizes how variables in a system influence one another.

SCM = Equations + DAG + noise = $\langle U, V, F, P(U) \rangle$

U: Exogenous (external) variables

V: Endogenous (system) variables

F: Structural equations defining each V_i as a function

of its parents (PA) and

its own noise (U): $V_i = f_i(PA_i, U_i)$

P(U): Noise distribution.

Structural Causal Model (SCM)

formalizes how variables in a system influence one another.

SCM = Equations + DAG + noise = *⟨U,V,F,P(U)⟩*

U: Exogenous (external) variables

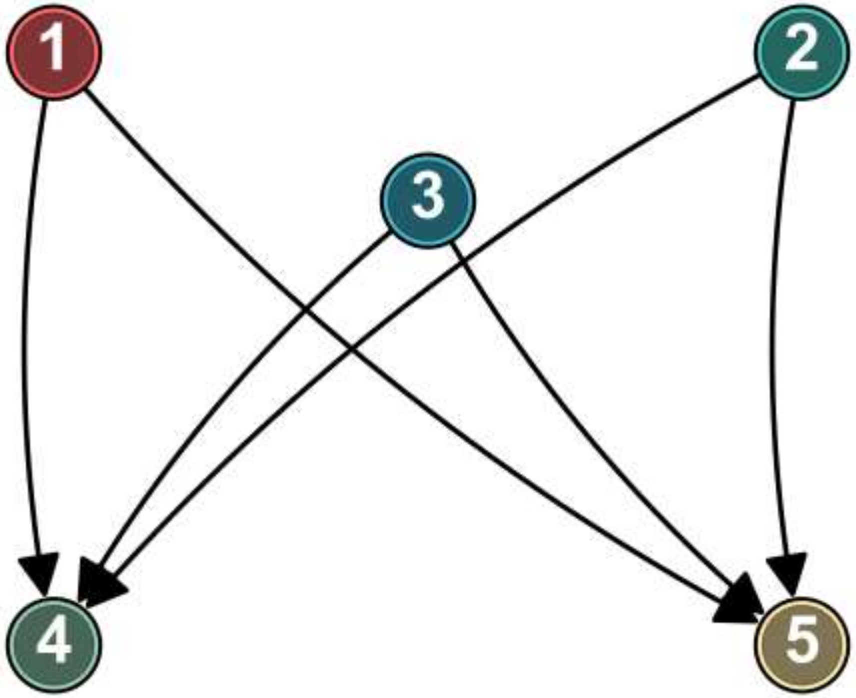
V: Endogenous (system) variables

F: Structural equations defining each V_i as a function

of its parents (PA) and

its own noise (U): $V_i = f_i(PA_i, U_i)$

P(U): Noise distribution.



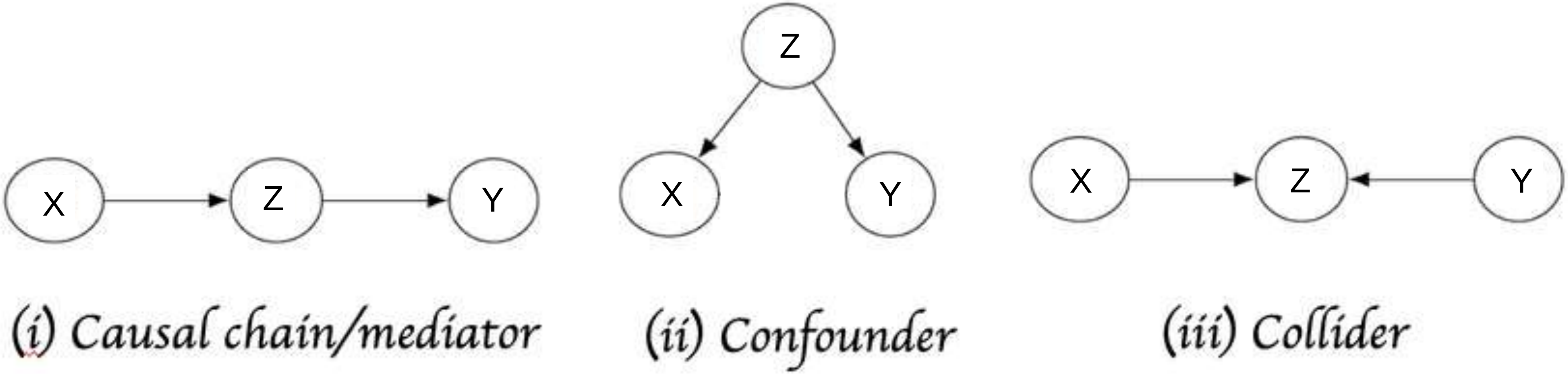
$X_1=U_1$; $X_2=U_2$; $X_3=U_3$

$X_4 = 0.4\ X_1 + 0.3X_2 + 0.3\ X_3 + U_4$;

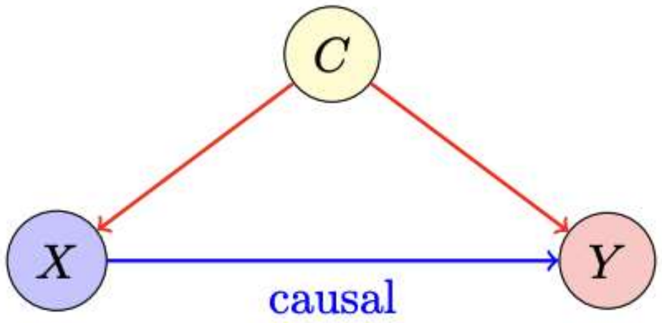
$X_5 = 0.5\ X_1 + 0.2X_2 + 0.3\ X_3 + U_5$

d-Separation: Causal Graphs Intuition

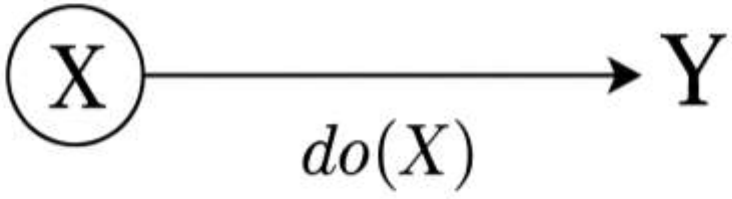
When are X and Y independent, given Z?



Back Door Criterion: do Operator



back-door path: X ← C → Y (confounding)

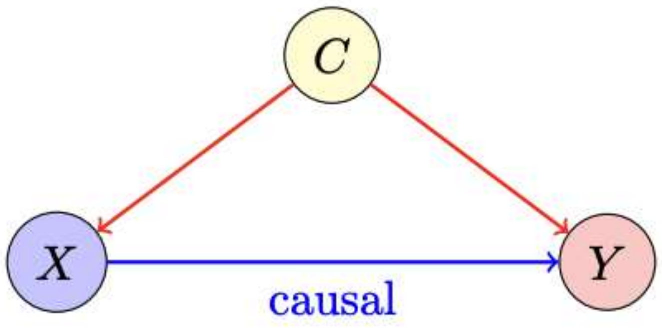


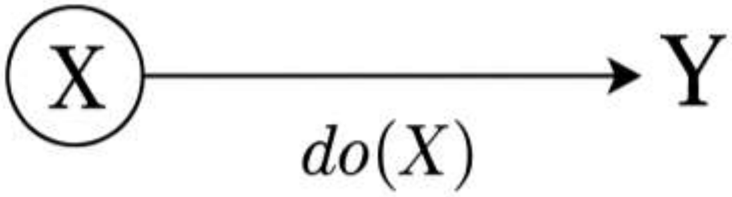
block by conditioning on C → identify causal effect.

Two Questions:

- *Observation:* “What happens when X is seen?” → $P(Y \mid X)$
- *Intervention:* “What happens when X is done?” → $P\big(Y \mid do(X)\big)$

Back Door Criterion: do Operator





Two Questions:

- *Observation:* “What happens when X is seen?” → $P(Y \mid X)$
- *Intervention:* “What happens when X is done?” → $P\big(Y \mid do(X)\big)$

The Back-Door Adjustment Formula

$$P\big(Y \mid do(X)\big) = \sum_{c} P\,(Y, \mid XC = c)\,P(C = c)$$

This expression removes the influence of confounder C, isolating the causal impact of X on Y.

Data-driven Causal Discovery Algorithms

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.	Kalainathar 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]

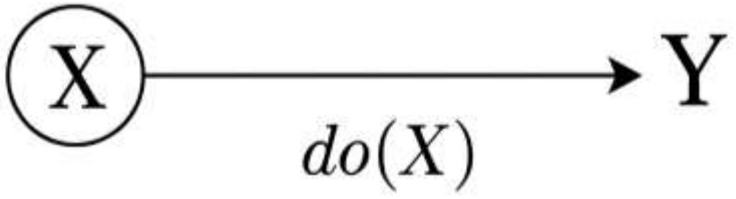
BUT…

1. **Sparse Data:** Real-world data is often insufficient to capture true probability distributions accurately.
2. **Static Data:** Reliance on static data risks accuracy, especially in dynamic environments.
3. **Edge Orientation:** Difficulty in inferring complete edge orientations due to equivalent DAGs.
4. **Expert Dependence:** Requires domain experts, who are scarce, costly, and variable in quality.
5. **Interpretability:** Traditional algorithms lack interpretability, limiting their ability to answer targeted causal questions.

Why Causality Matters

- **Prediction ≠ Understanding**
  - Correlations describe *what tends to happen*
  - But actions change the world — correlations break under intervention
  - Real-world systems need *stability*, *transferability*, and *control*
- **Goal:** move from *seeing* patterns → *understanding* mechanisms

From Observation to Intervention



Two Questions:

- *Observation:* “What happens when X is seen?” → $P(Y \mid X)$
- *Intervention:* “What happens when X is done?” → $P(Y \mid do(X))$

Why this matters

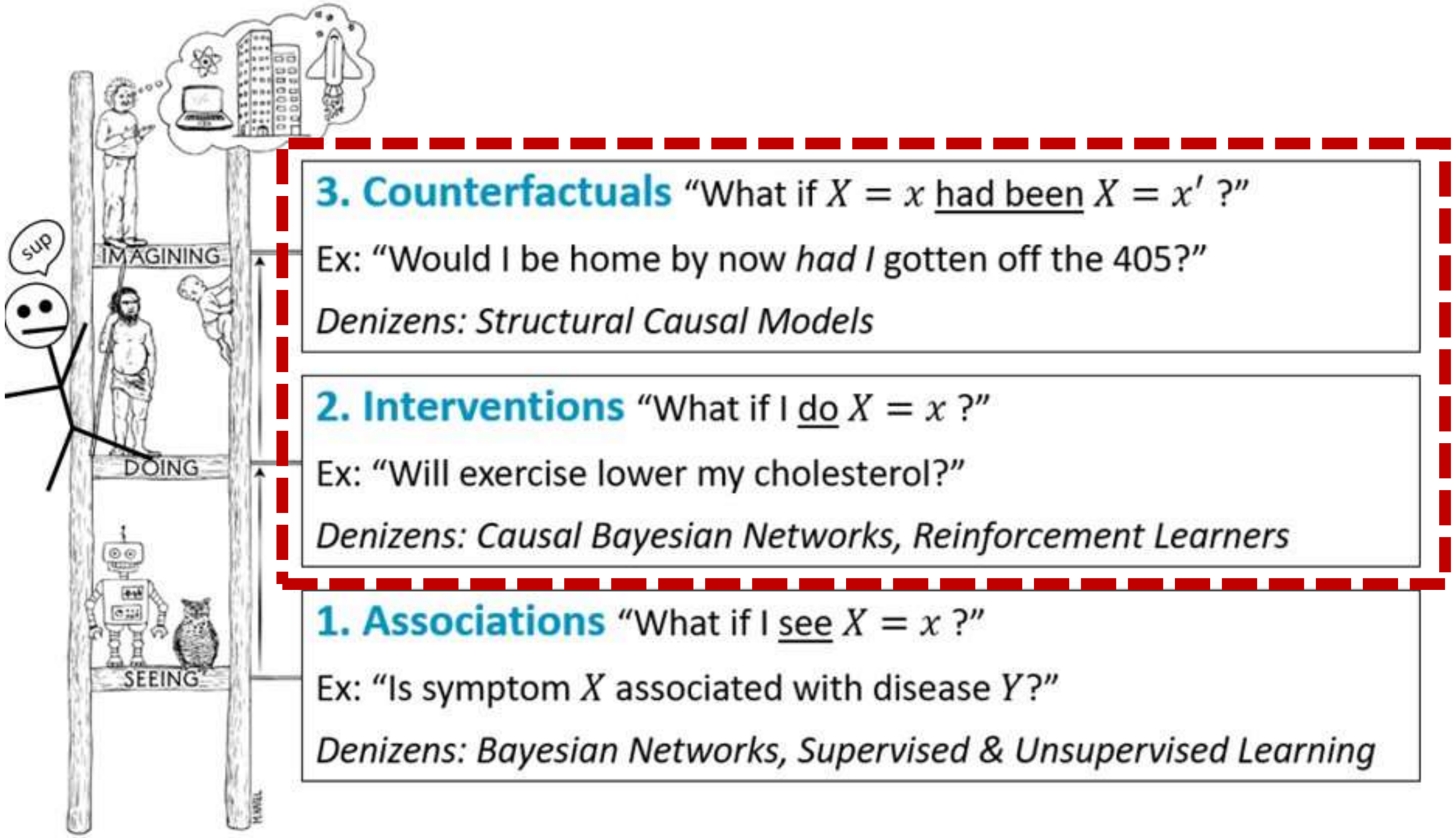
- $P(Y \mid X)$ may fail when environment shifts
- $P(Y \mid do(X))$ lets agents *predict consequences* of their actions

PolicyGRID: Acting to Understand, Understanding to Act

NeurIPS 2025 Workshop on Embodied World Models for Decision Making

The Causal Ladder

Levels of Reasoning *(Pearl, 2019)*



`PolicyGRID` operates at 2–3

- Learns *how actions change outcomes*

- Uses causal understanding to design better control policies

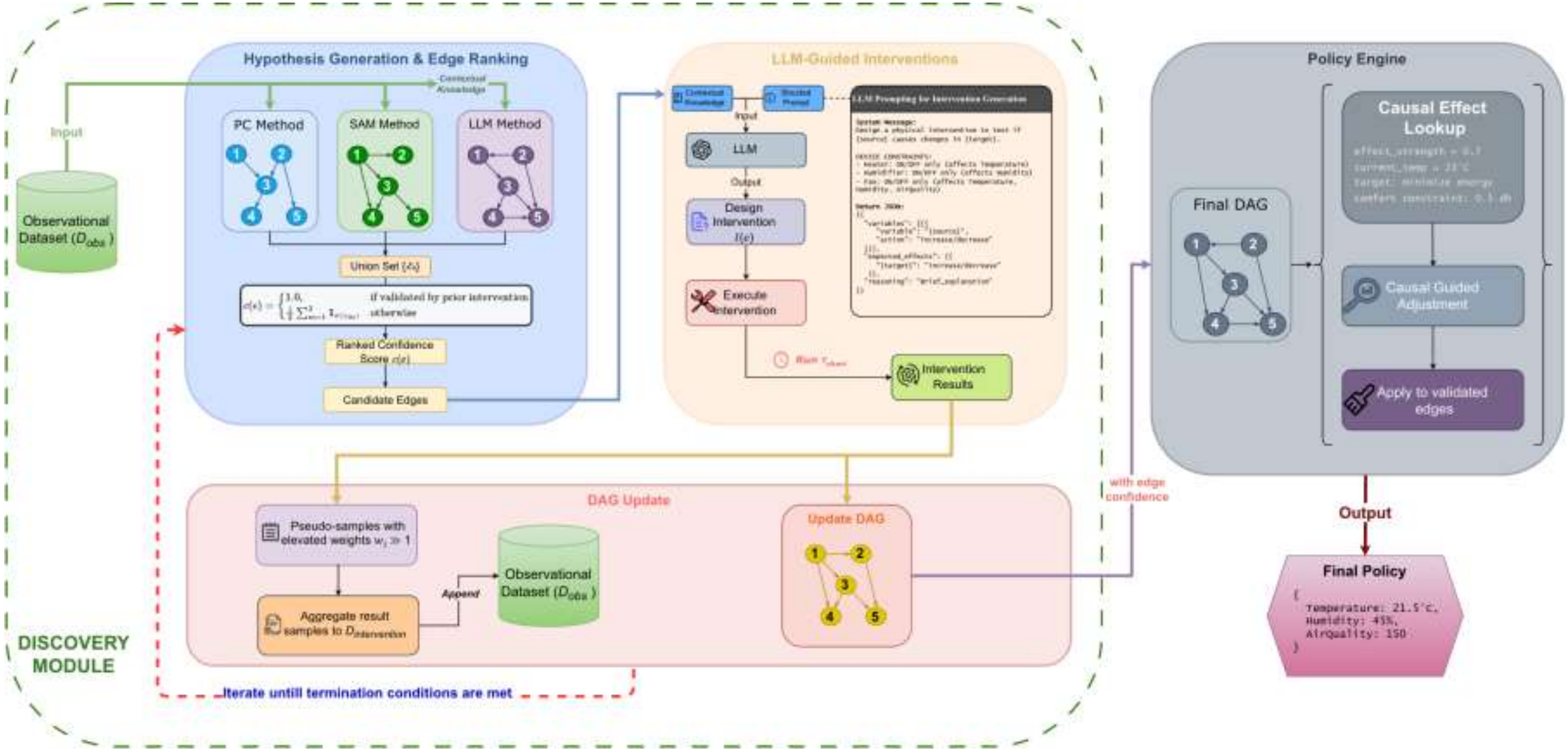
Why `PolicyGRID`

Bridging Understanding and Action

- Existing agents → *ignore causality* or treat it as *offline pre-processing*
- `PolicyGRID`:
  → *discovers* causal structure
  → *validates* interventions
  → *translates* causal knowledge into adaptive policies

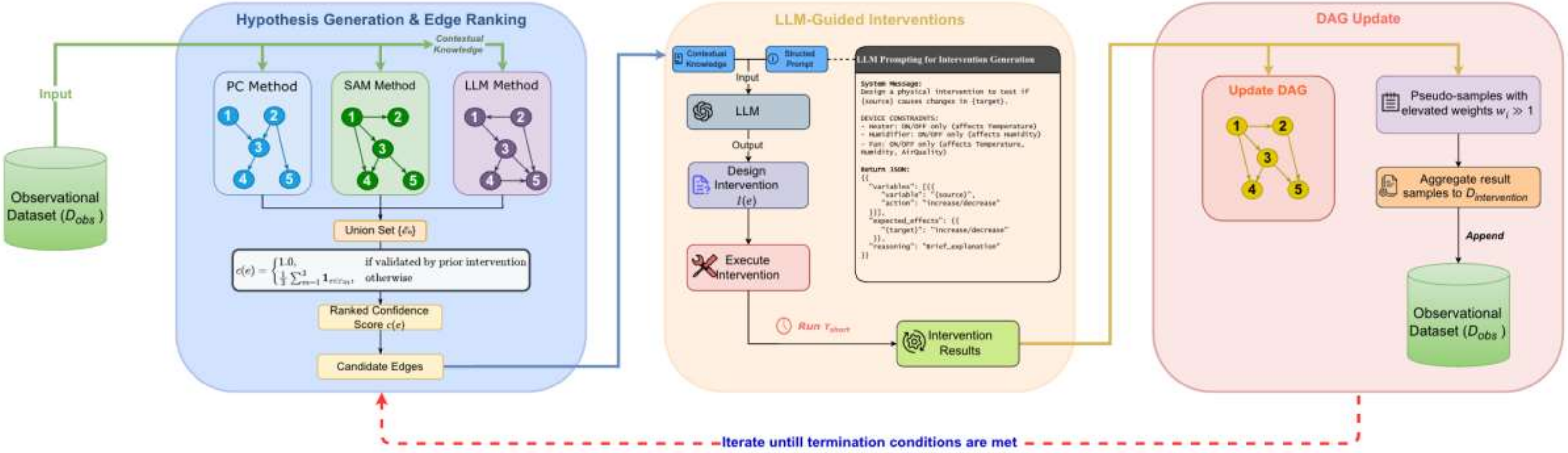
Acting to Understand, Understanding to Act

PolicyGRID

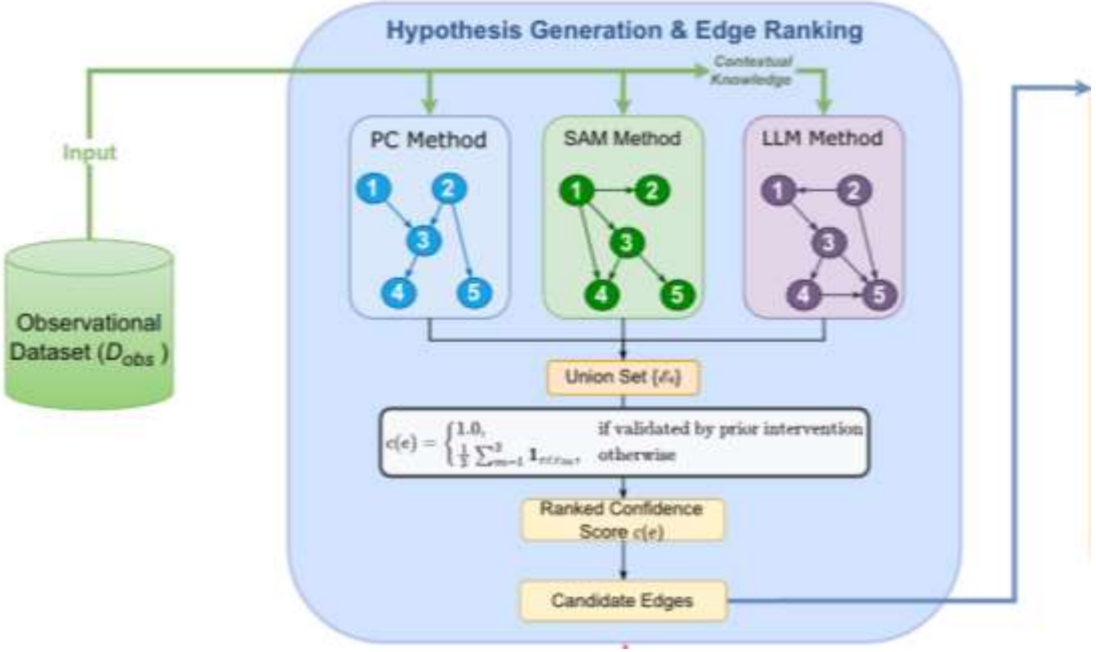


1. Discovering causal structure from data
2. Validating causal relationships
3. Translating causal knowledge into policies.

Discovery Module

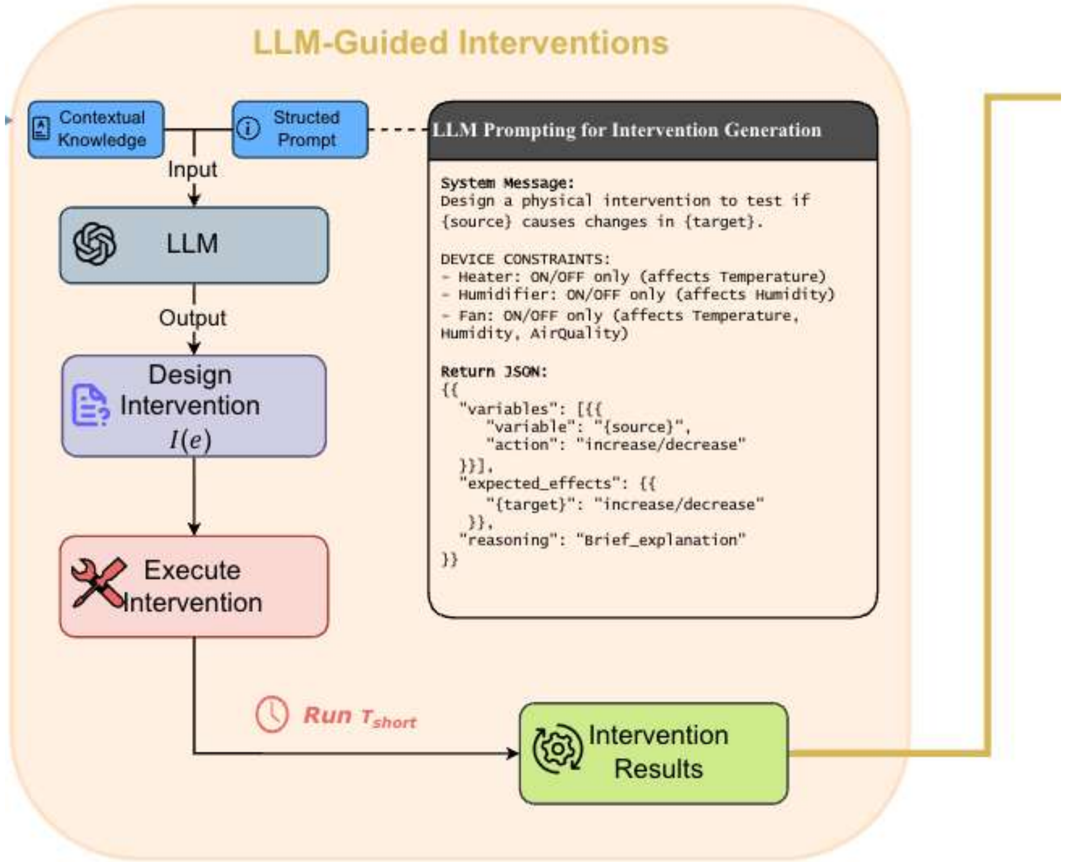


Discovery Module: **Hypothesis Generation**



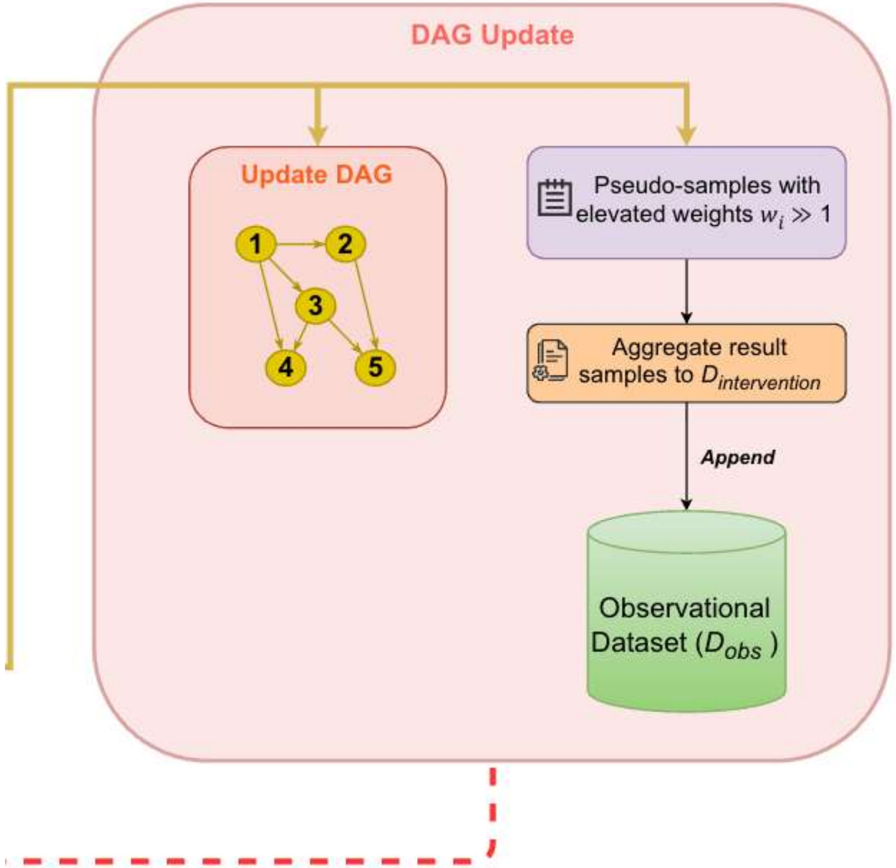
- 3 generators: PC, SAM, LLM
- Hypothesis edge union + scored ranking
- Start iterative testing in ascending order of confidence scores

Discovery Module: **Interventions**

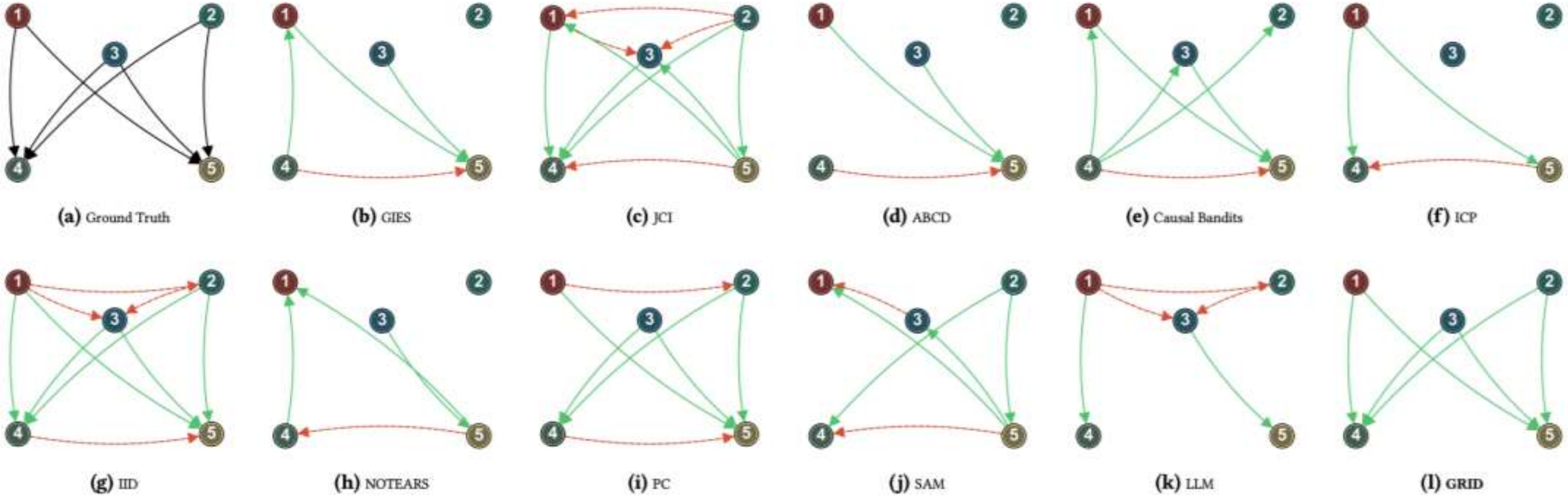


- Intervention on each edge
- LLM designs the intervention to be executed

Discovery Module: **DAG Update**



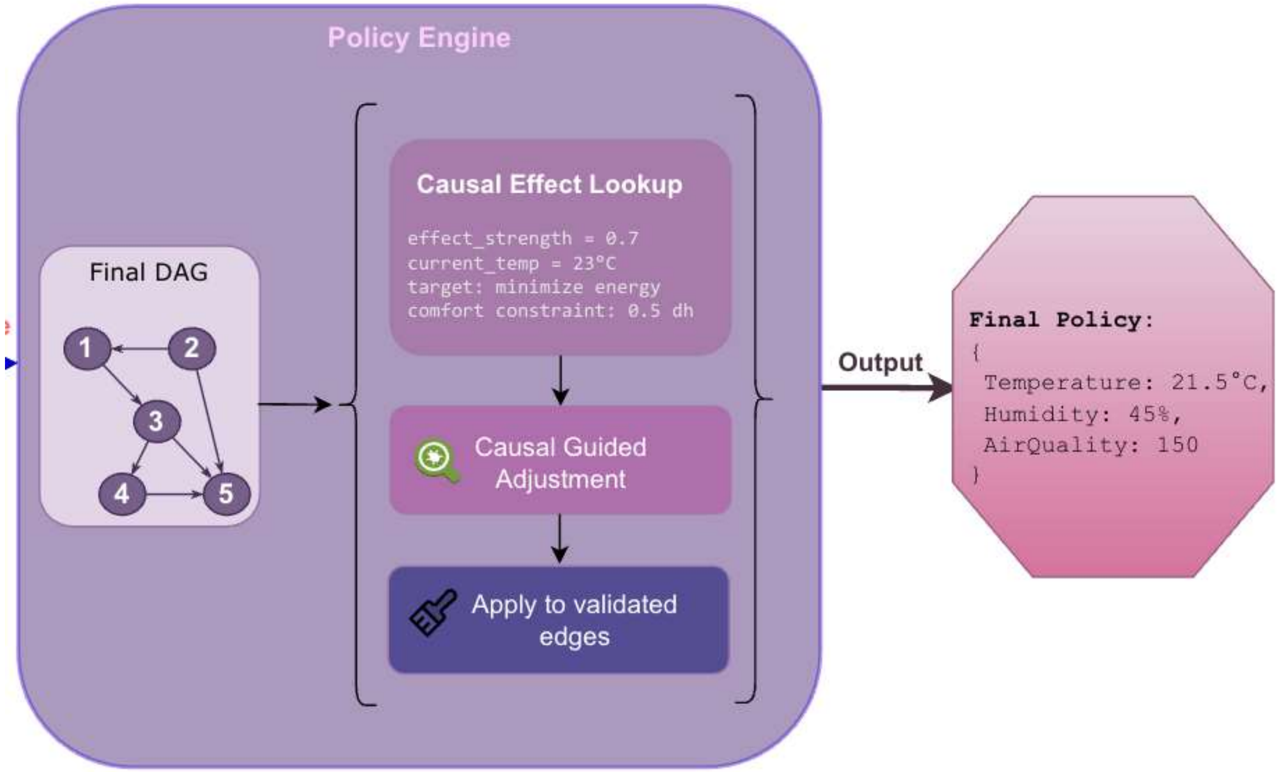
- If intervention effects > threshold:
  - Continue testing and append intervention data to dataset
  - Else, discard edge
- Update final DAG with validated edge
- Repeat until:
  - all candidate edges are evaluated
  - Tmax interventions is reached
  - learned DAG converges



(a) Ground Truth (b) GIES (c) JCI (d) ABCD (e) Causal Bandits (f) ICP

(g) IID (h) NOTEARS (i) PC (j) SAM (k) LLM (l) GRID

Policy Engine



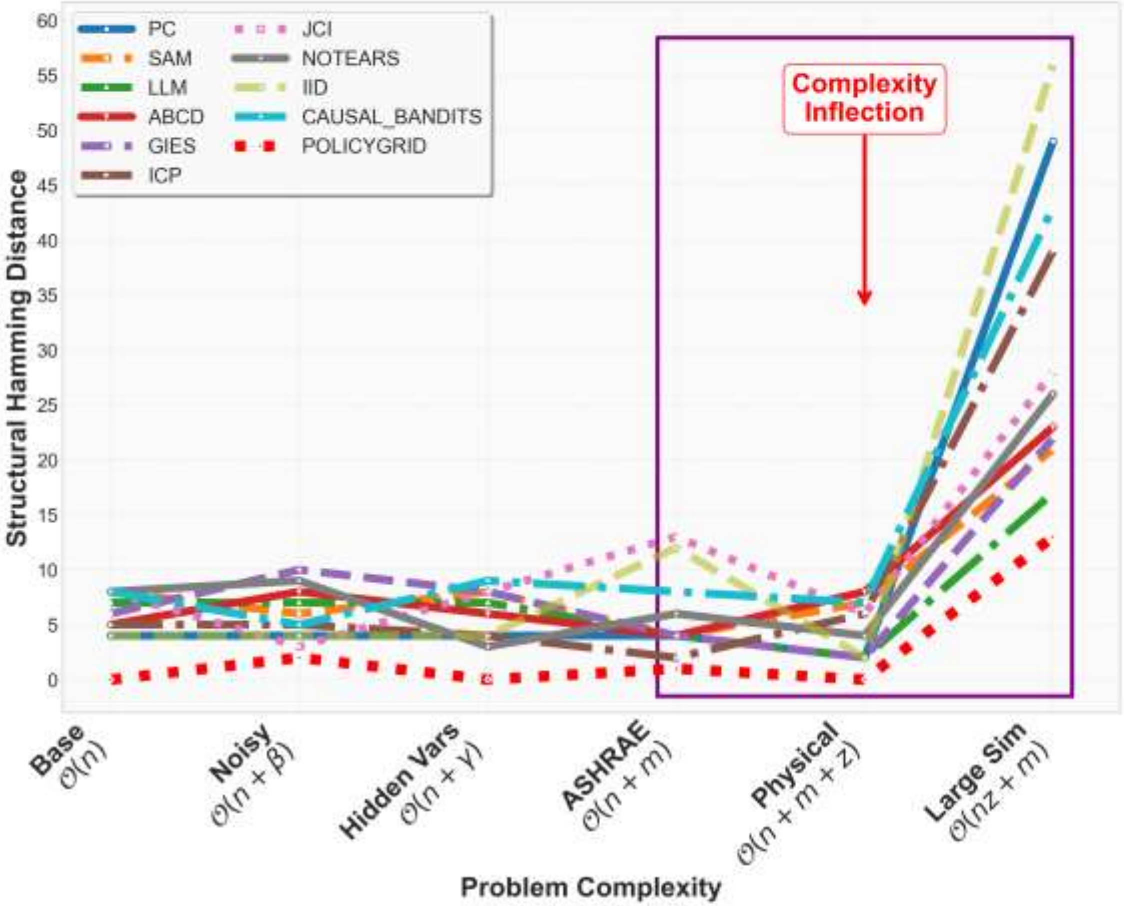
- Validated DAG → world model
- For each edge:
  - Get effect_strength from past intervention data
  - Adjust controls accordingly
  - Apply necessary constraints to all adjustments
  - Return final policy

Experiments

1. Causal Structure Recovery
   i. tested on 6 progressively
   ii. compared against 10 representative methods
   iii. SHD, F1, precision-recall, cost/risk
2. Embodied Control Performance
   i. tested in 4 embodied control scenarios
   ii. compared against 3 baselines
   iii. 2 complimentary evaluation perspectives

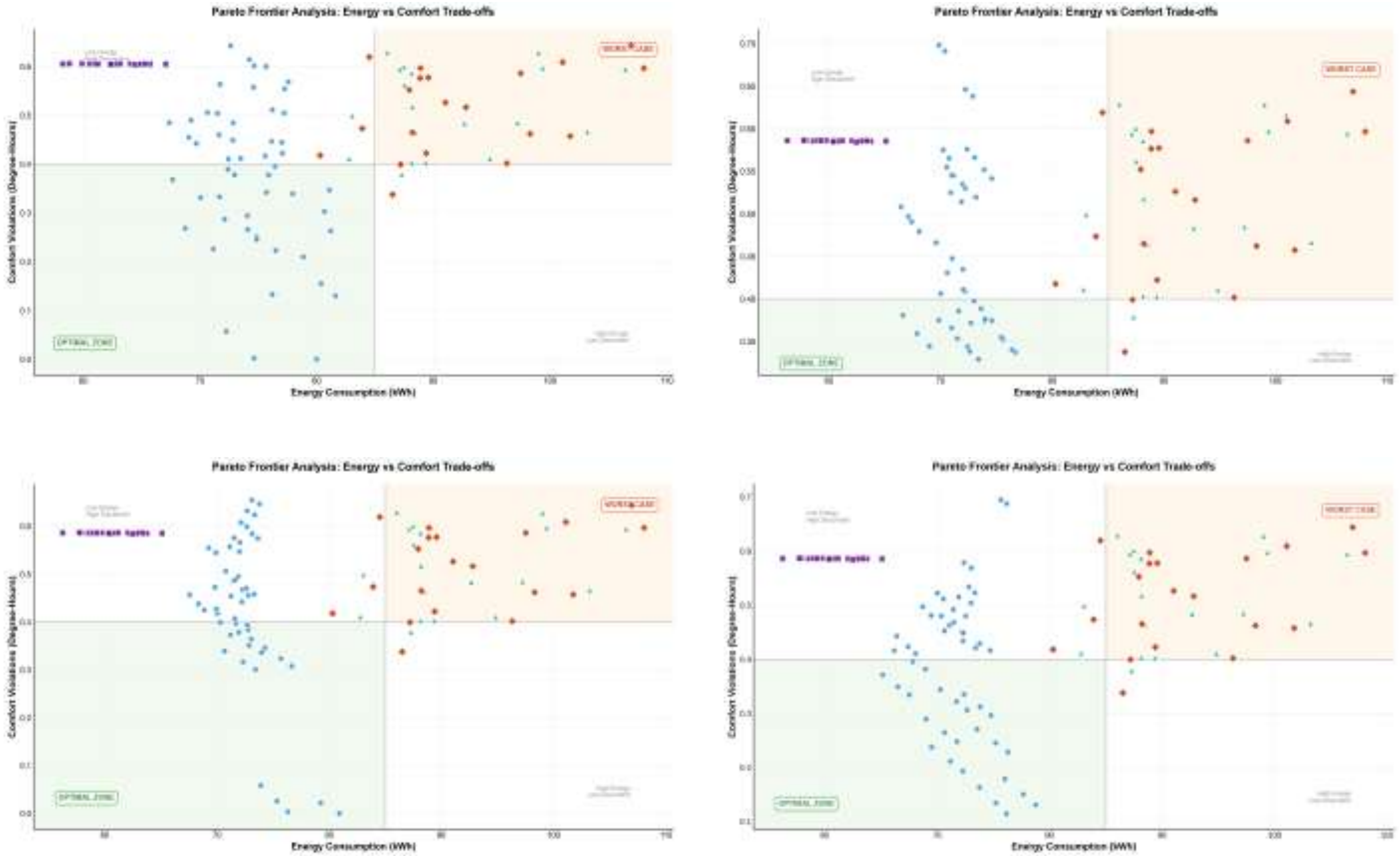
Results: **Causal World Model Fidelity**

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**



- `PolicyGRID` has lowest SHD across all setups
- Baselines degrade with complexity

Results: **Policy Performance**



- Highest hypervolume and lowest violation rates with `PolicyGRID`
- Causal DAG critical for policy quality
- Baselines (PID, Correlation) show low hv and high violations
- `PolicyGRID` dominates Pareto front with low energy and low discomfort

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 %

Takeaways + Discussion

https://openreview.net/pdf?id=SqaOyB89rE

- **Causality links understanding and action**—moving beyond pattern recognition to mechanism-based reasoning.
- **Structural Causal Models** explain *why* outcomes occur and how interventions alter them.
- **`PolicyGRID`** embeds this reasoning into control, yielding robust and explainable policies.
- From *correlation to control*, it enables systems that learn to act, not just predict.
- **Open question:** How can multimodal models better capture causal links across sensors, language, and perception?

Acting to understand — understanding to act.