LEARNING IMPLICIT CAUSAL WORLD MODELS FROM MULTI-AGENT DEMONSTRATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

In model-based reinforcement learning, world models exist as internal simulators, but their training often conflates statistical correlations with causal mechanisms. This problem is exacerbated in multi-agent systems where physical transitions are intertwined with strategic agent intents, causing world models to fail under distribution shift. We introduce Implicit Causal World Models to recover environmental dynamics from offline demonstrations without requiring pre-defined causal graphs. By incorporating policy variance, we render world models discoverable via the multi-agent system sequential backdoor condition. Evaluations across coordination tasks (Two-Door, Navigation, and Giveway) demonstrate that these models provide interpretable causal representations under both full and partial observability, with model accuracy scaling directly with interventional strength.

1 INTRODUCTION

In the evolution of model-based reinforcement learning (Moerland et al., 2022), world models have progressed from early predictive systems (Sutton, 1991) to sophisticated generative architectures (Hafner et al., 2024; 2019; LeCun et al., 2022) that utilize latent state-space models to simulate future trajectories. We frame these world models as an agent's "internal simulators," representing its belief about the environment. While world models improve sample efficiency (Ha & Schmidhuber, 2018b) in controlled settings , they remain fundamentally limited by their reliance on statistical correlations rather than causal underlying mechanisms (Kipf et al., 2020). In high-stakes environments, this leads to the spurious correlation problem, where models incorrectly attribute state transitions to non-causal features present during expert demonstrations. This challenge is compounded in multi-agent systems, where environmental physics are intertwined (Lowe et al., 2020; Dasgupta et al., 2019) with the strategic, often confounded, intents of other agents. For example, an autonomous vehicle model might incorrectly learn to accelerate because an adjacent car moves, rather than because the traffic light turned green. In this scenario, the physics of the green light correlates perfectly with another agent's intent to move, leading to causal confusion.

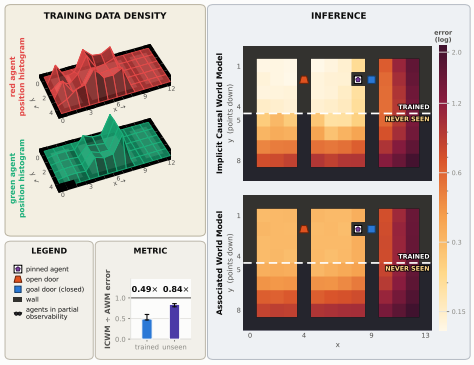


Figure 1: Trained on limited data (agents have a narrow subset of positions), **ICWM maintains low regularized transfer error, per cell** (Layout shift: Two-Door → Stretched Two-Door) on entirely unseen (0.84x) and seen (0.49x) states, outperforming baseline AWM. Details in Appendix V

Causal World Models (Ahmed et al., 2020; Nam et al., 2026) redefine internal simulators by replacing black-box transition functions with a structured network of causal links representing the environment's governing mechanisms. During policy evaluations, prior work (Richens & Everitt, 2024) highlights the theoretical robustness of agents that already know causal models, under surprise from distorted dynamics (Xia et al., 2022) introduced by selection bias or unobserved confounders. Subject to distributional shifts (Deng et al., 2025; Arjovsky et al., 2020) in the observational space, the following world models showcase invariant (Peters et al., 2015) transition dynamics for being grounded in environment physics. Wang et al. (2022) introduces a causally informed model with a known graph prior.

We propose Implicit Causal World Models (ICWM), which leverage neural architectures to approximate structural causal models (Schölkopf et al., 2021; Pawlowski et al., 2020) without requiring the

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

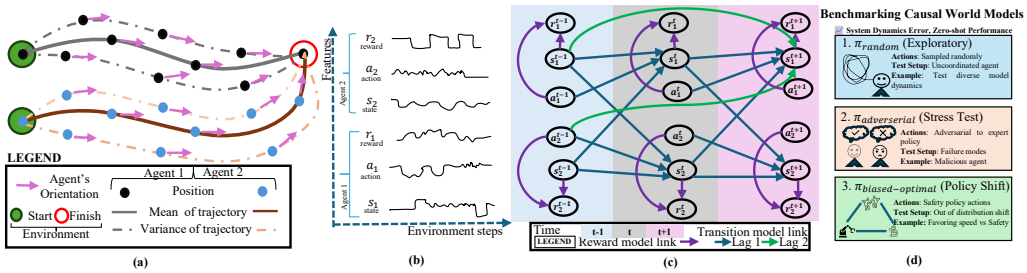


Figure 2: Overview of the ICWM Framework: (a) **Stochastic Demonstrations**: Expert multi-agent trajectories collected via soft interventions. (b) **Feature Extraction**: Temporal features (r, a, s) processed for causal discovery. (c) **Causal Structure Discovery**: Learned world model showing reward and transition physics across temporal lags (Lag 1: $t-1 \to t$, Lag 2: $t-2 \to t$). (d) **Benchmarking**: Stress testing under exploratory (π_{random}), adversarial ($\pi_{adversarial}$), and distributional ($\pi_{biased-optimal}$) policy shifts.

hand-engineered directed acyclic graphs often rendered intractable by the combinatorial complexity of multi-agent systems. Discovering true environment dynamics from policy rollouts requires overcoming temporal dependencies that violate the independent and identically distributed assumptions central to standard causal discovery (Peters et al., 2017). In a novel integration of momentary conditional independence (Runge, 2022) into reinforcement learning, we model sequential trajectories as a vector autoregressive process (Runge et al., 2019) and embed the extracted causal graph as a differentiable mask, which, under the sequential backdoor condition (Zhang & Bareinboim, 2016), forces the deep generative model to recover true environmental physics. Figure 1 introduces ICWM's generalization capability, which outperforms baselines.

Designing a true causal world model requires identifiability at Layer 2 (Intervention) of the Pearl Causal Hierarchy (Pearl, 2019). Because the conventional approach of active hard interventions, which force specific features of the environment to fixed values, is prohibitively costly in MAS (Kallus & Zhou, 2020), we instead utilize soft interventions (Correa & Bareinboim, 2020). By injecting stochastic noise to perturb the agents' policies, we alter their action probabilities rather than dictating deterministic outcomes. This forces the agents to break out of their habitual routines and severs spurious correlations, enabling the causal world model to discover the true environment dynamics.

In the framework of Causal Imitation Learning (Kumor et al., 2022), agent actions in multi-agent systems are treated as sequential interventions upon the environment by the expert policy. While behavioral cloning (Pomerleau, 1988) is the standard approach for imitating expert policies, it fails to discover the underlying environmental physics when demonstrations are influenced by unobserved confounders (de Haan et al., 2019). For the learned world model to be identifiable, the sequential backdoor condition must be satisfied. This requires that the observed state history effectively d-separates agent actions from future states along all non-causal backdoor paths. Approaches that unmix latent variables (Zhang et al., 2023; Bagi et al., 2024) via mechanism switching to achieve causal identifiability from unpaired data fail in MAS due to joint agent interventions and non-i.i.d. sequential dynamics. As demonstrated in the social board game `Tic-Tac-Toe` (Figure 4), interventional data is the critical bottleneck for bounding world model uncertainty.

In this paper (Figure 2), we address the limitations of model-based reinforcement learning in multi-agent settings by introducing a framework for discovering causal mechanisms from biased offline data. Our contributions as follows:

1. **Stochastic Interventional Data Collection (C1)**: We introduce a demonstration collection strategy that applies soft interventions through policy variance, satisfying the Pearl Causal Hierarchy requirements for identifiability without costly hard interventions.
2. **Multi-Agent Sequential Backdoor Condition (C2)**: We formalize and prove the Multi-Agent Sequential Backdoor Condition, establishing that interventional dynamics are recoverable from observational multi-agent data despite unobserved strategic confounding, and use it to decouple environmental physics from agent intent.
3. **Interventional Strength Recovers Causal Structure (C3)**: Across three coordination domains (Two-Door (Fickinger, 2020), Navigation (Bettini et al., 2022), and Giveway (Bettini et al.,

2022)), under both full and partial observability and scaling to $2/3/4$ agents, we show that discovery error falls monotonically as the interventional strength σ grows.
4. **Generalization Error Bound (C4)**: We establish, and confirm empirically, that the out-of-distribution policy evaluation error of the learned world model scales inversely with the interventional strength, falling by up to two orders of magnitude under adversarial policies.
5. **Intervention, Not Graph Quality, Is the Lever (C5)**: Pooling $1{,}240$ trained models, we show that the association between recovered-graph quality and predictive accuracy is a common-cause artifact (Figure 3) of σ: conditioning on σ removes it, and a graph-free model matches the masked model at matched data. The causal mask earns its keep specifically under partial observation and layout extrapolation, and as a structural prior whose corruption is worse than no prior at all.
6. **Calibrated Out-of-Distribution Uncertainty (C6)**: We show that the learned models localize their predictive uncertainty to a minority of state dimensions under distribution shift, rather than degrading uniformly.

Impact of this work (Figure 9) seen as advances Model-Based Reinforcement Learning toward out-of-distribution robustness and structural interpretability, ensuring autonomous systems remain reliable even when surrounding agent behaviors deviate from training distributions. This framework facilitates verifiable safety, fault attribution, agent alignment, and making precise credit assignment to agent actions feasible.

2 Related Work

In our literature review, we identify practical bottlenecks in related problems and describe the gaps.

Limitations of Associative World Models: Model-Based Reinforcement Learning often relies on associative world models (DreamerV3, PlaNet, MuZero (Schrittwieser et al., 2020)) that predict future states by matching statistical distributions. Operating at Layer 1 of the Pearl Causal Hierarchy, these models are prone to causal confusion (de Haan et al., 2019), failing to distinguish between functional environment transitions and spurious correlations. (Li et al., 2020) bounds confounding bias via counterfactual risk minimization, but strict sensitivity bounds can induce overly conservative dynamics (e.g., predicting robot control loss from standard acceleration to hedge against unobserved friction).

In MAS world models (Zhu et al., 2026; Zhang et al., 2025b;a), this manifest as an objective mismatch where models conflate invariant physical laws with the strategic intents of other agents. When dynamics are "intertwined," the world model develops a policy-dependent bias; consequently, shifting from expert to exploratory or adversarial policies causes catastrophic failure because the model expects environment physics to follow the expert's specific behavior. We address this by treating agent intent as a formal confounder U, adopting an interventional framework (Schölkopf et al., 2021) to ensure out-of-distribution robustness.

FOCUS (Zhu et al., 2022) prunes transitions using Kernel-based Conditional Independence tests (Zhang et al., 2012). These methods assume causal sufficiency. Multi-agent systems violate this assumption. Unobserved agent intents act as persistent confounders. Observational discovery yields fully-connected graphs in dense settings. We address this by treating intent as a formal confounder U. We use an interventional framework (Schölkopf et al., 2021).

Scalability of Neural Causal Architectures: GNNs (Kipf et al., 2020; Scarselli et al., 2009), C-VAEs (Yang et al., 2023), and Neural SCMs (Pawlowski et al., 2020) approximate causal structures. Their message-passing requirements scale as $O(N^2)$. This requirement makes them computationally prohibitive for dense multi-agent dynamics. Furthermore, true counterfactual reasoning remains intractable (Schölkopf et al., 2021).

Temporal Gaps in Causal Discovery: Causal discovery is divided into constraint-based (Spirtes et al., 2000) methods, which use statistical independence tests, and score-based (Chickering, 2002) methods, which maximize a fit criterion (Verma & Pearl, 1990). Both paradigms are complicated by agent demonstrations that violate i.i.d. assumptions. While PCMCI+ (Runge, 2022) addresses autocorrelated and contemporaneous dependencies in time series, it has not been integrated into reinforcement learning transition models. We bridge this gap by using these techniques to inform the structural constraints of our world model.

The Intervention-Observation Paradox: The reliance on costly hard interventions bottlenecks causal discovery (Eberhardt, 2008) and causal reinforcement learning algorithms (Zhang & Barein-

boim, 2016). We bypass this by demonstrating that multi-agent policy variance provides sufficient interventional strength for Layer 2 identifiability. This parallels Causal Imitation Learning (CIL) theory, where the sequential backdoor condition (Zhang & Bareinboim, 2016) (SBC) is used to identify robust policies. We shift this focus, applying the SBC directly to the world model to ensure transition dynamics are identifiable even when agent intents act as unobserved confounders.

Causal Credit Assignment in MARL: ICWM's identifiable simulator grounds credit assignment in true physical consequences rather than spurious policy correlations, ensuring downstream learning (Yu et al., 2022; Williams, 1992) success.

Comparison to Existing MAS Models: Current multi-agent system frameworks (Ahmed et al., 2020; Madumal et al., 2019) often focus on emergent representations or communication protocols. Our framework differentiates itself by explicitly discovering environmental physics from agent strategy. Models that treat transitions as deterministic finite automata (Hopcroft et al., 2001) (DFA) to learn the system transition (Li et al., 2024; Vafa et al., 2024; 2025b), and our work overlaps with these approaches in how structural constraints are integrated and learned alongside a generative model.

Generative Video World Models: High-fidelity video models use diffusion and transformer processes to synthesize visually coherent states (Ha & Schmidhuber, 2018a; Bruce et al., 2024). By prioritizing visual plausibility over causal grounding, these architectures are prone to physical hallucinations under policy shifts (Li et al., 2024; Tragoudaras et al., 2026). Over pixel space based (Assran et al., 2023; Zhou et al., 2025) or object-centric (Nam et al., 2026) world modeling, our framework instead prioritizes the discovery of invariant structural mechanisms to ensure robustness against spurious correlations inherent in training demonstrations.

3 BACKGROUND

3.1 STRUCTURAL CAUSAL MODEL FOR MAS

We formalize multi-agent transition dynamics a Structural Causal Model (Pearl, 2009) (SCM) $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$, with exogenous variables $\mathbf{U}$ (environment noise u_{env}, latent agent intents u^i_{intent}) and endogenous variables $\mathbf{V}$ (state s_t, joint actions $\mathbf{a_t}$, reward r_t, successor s_{t+1}). The structural mappings $\mathcal{F}$ are:

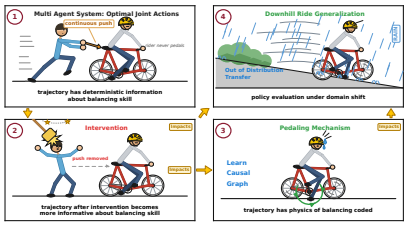


Figure 3: σ impacts graph *and* transfer.

$$\textbf{Environment Physics:} \quad s_{t+1} \leftarrow f_s(s_t, \mathbf{a}_t, u_{env}) \tag{1}$$

$$\textbf{Reward Model:} \quad r_t \leftarrow f_r(s_t, \mathbf{a}_t, u_{env}) \tag{2}$$

$$\textbf{Agent Policies:} \quad a^i_t \leftarrow f_{a^i}(s_t, u^i_{intent}) \tag{3}$$

In this formulation, f_s represents the invariant physical laws we aim to recover. The dependence of actions a^i_t on latent intents u^i_{intent} introduces persistent confounding, as u^i_{intent} typically correlates with s_t in expert demonstrations, necessitating an interventional approach for discovery. Equation 3 has basis in prior work (Lowe et al., 2020; Dasgupta et al., 2019) (See Appendix B for details.)

3.2 LONGITUDINAL CAUSAL DECISION GRAPH FOR MAS

We represent sequential multi-agent system using a Causal decision graph $\mathcal{G}$ (Figure 10) (Zhang & Bareinboim, 2021), which encodes the structural dependencies of Eqs. 1, 2, 3 over time. (See Appendix C for a detailed analysis of these topological structures.)

In multi-agent demonstrations, unobserved strategic intent acts as a confounder U that induces non-causal backdoor paths. Under full observability, introducing U makes it a common cause of both actions and transitions, producing causal confusion; under partial observability the same intent-driven actions are additionally filtered through an information bottleneck for agent's limited perspective (Fig. 10a–d). Appendix C analyzes all four topological regimes.

3.3 STOCHASTIC SOFT INTERVENTIONS

Hard interventions, such as $do(\mathbf{a}_t = \hat{\mathbf{a}})$ or $do(\mathbf{s}_t = \hat{\mathbf{s}})$, isolate causal effects by entirely severing incoming edges to a target variable (Pearl, 2009). We execute stochastic soft interventions. We perturb the underlying mechanisms by injecting controlled variance directly into the behavioral data (Appendix J).

3.3.1 Stochastic Soft Intervention via Policy

In observational data, latent intent u^i_{intent} persistently confounds agent actions. We break this confounding by applying a soft intervention to the agent's policy (**claim C1**). We replace the original (Eqn. 3) conditional distribution $P(a^i_t \mid s_t, u_{intent})$ with an interventional distribution $P^\sigma(a^i_t \mid s_t)$. Let π^i_{exp} denote the deterministic expert policy yielding demonstrations $\mathcal{D}$. We enforce a stationary mixture distribution:

$$P^\sigma(a^i_t \mid s_t) = (1-\sigma)\pi^i_{exp}(a^i_t \mid s_t) + \sigma\zeta(a^i_t) \quad (4)$$

where $\sigma \in [0,1]$ dictates the interventional strength, and ζ is a strictly independent stochastic process (e.g., a uniform distribution over $\mathcal{A}^i$). Structurally, this operates as a definitive causal soft intervention (Eberhardt, 2008) by fundamentally altering the data-generating mechanism (see Appendix D).

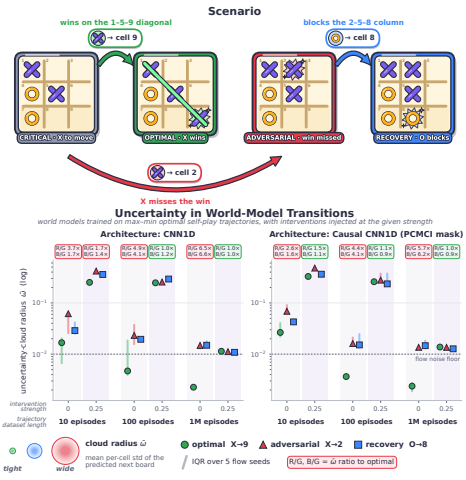


Figure 4: Tic-tac-toe scenario: world-model uncertainty against foul play, per architecture and intervention strength (Details in Appendix G).

3.3.2 Interventional Signal via Policy Entropy

We quantify this interventional capacity via the Shannon Entropy of P^σ at state s:

$$H(P^\sigma \mid s) = -\int_{\mathcal{A}^i} P^\sigma(a \mid s) \log P^\sigma(a \mid s) da \quad (5)$$

The identifiability of the transition function f_s demands aggregate entropy in $\mathcal{D}$. Deterministic demonstrations ($\sigma = 0$, $H = 0$) permanently conflate invariant physics with hidden intent. High entropy forces natural experiments. By varying actions in identical states, we break backdoor paths and isolate the environmental physics.

3.4 Causal Identifiability and the Sequential Backdoor Condition

A causal query $P(y \mid do(x))$ is identifiable if it is computable from the observed joint distribution given the causal graph (Pearl, 2009). We apply this to sequential decision-making to extract invariant dynamics from biased demonstrations.

Under the manipulated causal decision graph, when the Sequential Backdoor Criterion (SBC) (Zhang & Bareinboim, 2016) is satisfied, we estimate the causal effect of the action history $\mathbf{a}_{1:t}$ on the subsequent state s_{t+1} as $P(s_{t+1} \mid do(\mathbf{a}_{1:t}))$. is reducible to an observational expression. Therefore, the causal identifiability of the world model is asserted directly from the demonstrations.While this theorem guarantees identifiability for an isolated agent, the introduction of contemporaneous agents structurally violates the standard SBC. We detail the foundational criteria, the sequential graph surgery, and the strict necessity for a multi-agent extension in Appendix E.

3.5 Recovering a Causal World model

Following Vafa et al. (2024), we formalize environmental physics as a DFA $\mathcal{W} = (Q, \Sigma, \delta, q_0, F)$, where Q is the state space, Σ is the action alphabet, and $\delta : Q \times \Sigma \to Q$ is the transition function. We define q_{reject} as a sink state for impossible transitions, with $F = Q \setminus \{q_{\text{reject}}\}$ as the set of valid states. A generative model (proxy for world model) $m(\cdot) : \Sigma^* \to \Delta(\Sigma)$ recovers $\mathcal{W}$ if it satisfies exact next-token prediction for any sequence $s \in S(q)$ leading to state q:

$$\forall q \in F, \forall s \in S(q), \forall a \in \Sigma : m(a \mid s) > 0 \iff \delta(q,a) \neq q_{\text{reject}} \quad (6)$$

While DFA recovery ensures a model honors environmental rules via their support, it does not guarantee recovery of the invariant causal mechanism f_s in MAS, where u_{intent} acts as a confounder.

To bridge this, we assert that robust recovery requires satisfying the SBC. By conditioning on a history H_k that d-separates the current action $\mathbf{a}_k$ from the intent u_{intent} within the manipulated graph $\mathcal{G}_{\underline{\mathbf{a}_{1:k-1}}\overline{\mathbf{a}_{k+1:t}}}$, the model identifies the transition physics f_s independently of the expert policy or demonstrations.

4 FRAMEWORK AND METHODOLOGY

4.1 THE MULTI-AGENT SEQUENTIAL BACKDOOR CONDITION

To establish the structural identifiability of our Implicit Causal World Model (**claim C2**), we must demonstrate that the interventional distribution of the environmental transition can be uniquely computed from observational multi-agent data. Let $H_t = (\mathbf{s}_{0:t}, \mathbf{a}_{0:t-1})$ represent the observed history, U the unobserved latent intent, A_t^i the action of the target agent, and A_t^{-i} the joint actions of all other agents.

We formalize the Multi-Agent Sequential Backdoor Condition (MAS-SBC) under the premise that agents operate via decentralized policies and transitions are Markovian with respect to the state-action space. The following condition breaks the statistical collinearity of demonstrations necessary to resolve.

Theorem 1 (MAS-SBC). *Given the multi-agent causal decision graph $\mathcal{G}$, the interventional transition dynamics $P(s_{t+1} \mid do(a_t^i), H_t)$ can be resolved strictly through observational distributions via the sequential backdoor adjustment:*

$$P(s_{t+1} \mid do(a_t^i), H_t) = \sum_{a_t^{-i}} P(s_{t+1} \mid a_t^i, a_t^{-i}, H_t) P(a_t^{-i} \mid H_t) \tag{7}$$

The complete formal proof (**claim C2**) using Pearl's do-calculus on the mutilated graph $\mathcal{G}_{\overline{A^i_{1:t-1}}\underline{A^i_{t+1:\dots}}}$ is provided in Appendix F.

4.2 IDENTIFIABILITY VIA THE MAS-SBC

Decoupling f_s from U requires two things. **Structural identifiability**: Theorem 1 extends sequential adjustment by marginalizing over contemporaneous peer actions, blocking the multi-agent backdoor path $A_t^i \leftarrow U \rightarrow A_t^{-i} \rightarrow s_{t+1}$. **Causal support**: the adjustment cannot be evaluated on deterministic expert data, where a_t^i is a deterministic function of U and the two are collinear. Stochastic soft interventions ($\sigma > 0$) break this collinearity and guarantee the positivity needed to compute interventional queries. In practice we apply these interventions independently and simultaneously to all agents, with $\zeta(a_t^i) \perp \zeta(a_t^j)$.

4.3 GENERALIZATION ERROR BOUND

To formalize the necessity of stochastic soft interventions, we establish a theoretical bound on the out-of-distribution policy evaluation error. We mathematically prove that the generalization error of the learned causal world model scales inversely with the square root of the interventional strength (σ) applied during data collection. The formal theorem (**claim C4**), the step-by-step derivation using the Simulation Lemma, and the discussion of its theoretical implications are provided in Appendix H.

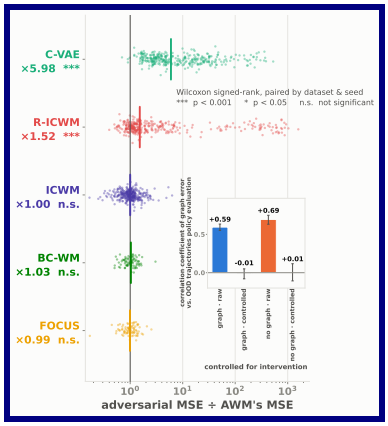


Figure 5: MSE of adversarial policy evaluation relative to AWM, per architecture. Good graph recovery and adversarial MSE comes from intervention.

4.4 CAUSAL STRUCTURE DISCOVERY

Assumptions. Our MAS formulation violates causal sufficiency by construction, since u_{intent} is unobserved. We therefore assume: **(i) Strict temporal lag** : the physical mechanism f_s acts with lag $\tau \geq 1$, so no true causal interaction is instantaneous (Figure 10); and **(ii) Contemporaneous confounding isolation** : u_{intent} coordinates actions strictly at time t and has no direct lagged effect on s_{t+1} outside $\mathbf{a}_t$.

Structure Discovery. Under above assumptions, we treat the MAS trajectories in $\mathcal{D}_\sigma$ as a Vector Autoregressive (VAR) process. Since the sampled histories $\mathcal{H}_t^\tau$ are almost IID across the dataset, we define the conditioning set as the state-action-reward history ($\mathcal{H}_t^\tau$) up to a maximum lag τ:

$$\mathcal{H}_t^\tau = \{(s_{t-k}, \mathbf{a}_{t-k}, r_{t-k}) \mid 1 \leq k \leq \tau\} \tag{8}$$

We utilize the PCMCI+ algorithm to evaluate the Conditional Mutual Information (CMI) (Wyner, 1978) and identify the directed edge set E of the skeleton $\mathcal{G}_{full}$. The discovery yields two sub-graphs: Lagged discovery (E_{lag}: edges $V_{t-k} \rightarrow V_t$

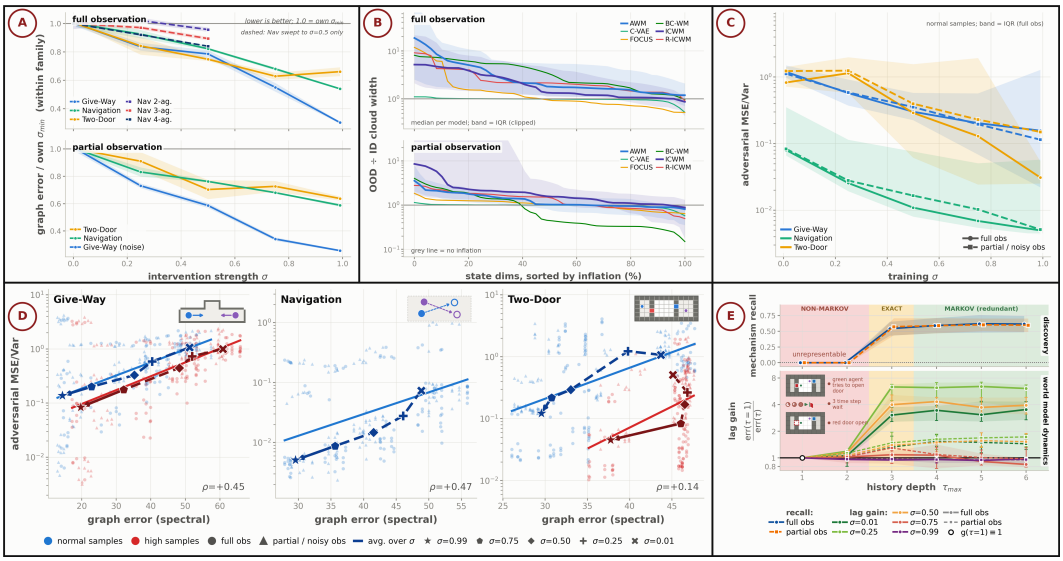


Figure 6: **Interventional strength drives both causal-structure recovery and out-of-distribution transfer.** **(A)** Graph error vs σ (dashed: navigation $2/3/4$ agents). **(B)** Per-axis OOD/ID density-width inflation, dimensions sorted. **(C)** Adversarial error (MSE/Var) vs σ. **(D)** Adversarial error vs graph error; broken lines trace the $\sigma = 0.99 \rightarrow 0.01$ trajectory. **(E)** Non-Markov $\rightarrow$ Markov at the true delay: ICWM diagnostics

spanning time steps) and contemporaneous discovery ($E_{contemp}$: edges $V_{i,t} \rightarrow V_{j,t}$ within the same step). Under Assumption **(i)**, invariant physical transitions lie in E_{lag}. Assumption **(ii)** restricts spurious intent correlations to $E_{contemp}$. We recover the unconfounded causal skeleton $\mathcal{G}_{skeleton} = (V, E_{skeleton})$ by discarding contemporaneous dependencies: $E_{skeleton} = \{(V_{t-k} \rightarrow V_t) \in E_{lag} \mid 1 \leq k \leq \tau\}$ Conditional independence tests in PCMCI+ assign a p-value to each candidate edge. This yields an adjacency tensor $\mathcal{C} \in \mathbb{R}^{|V| \times |V| \times \tau}$ encoding these significance values (see Appendix I):

$$\mathcal{D}_\sigma \xrightarrow{\text{PCMCI}^+} \mathcal{G}_{skeleton} \implies \mathcal{C} \in \mathbb{R}^{|V| \times |V| \times \tau}$$

4.5 Causal Masked World Model Architecture

We instantiate the Implicit Causal World Model (ICWM) by translating $\mathcal{G}_{skeleton}$ edges into a tensor mask $\mathbf{C}$ that structurally enforces the learned SCM.

Neural Masking. Let transition function f_s be parameterized by neural network Φ with parameters θ. Given input tensor $\mathbf{x}_t = \mathcal{H}_t^\tau$ (history up to lag τ), we satisfy the `SBC/MAS-SBC` by applying mask $\mathbf{C}$ via the per feature Hadamard product: $\hat{s}_{t+1} = \Phi(\mathbf{x}_t \odot \mathbf{C}; \theta)$. This restricts the receptive field so each prediction $\hat{s}_{i,t+1}$ depends strictly on its unconfounded parents $Pa(s_{i,t+1}) \in \mathcal{G}_{skeleton}$, isolating invariant physics.

5 Experimentation

Setup. We evaluate ICWM against five baselines (Appendix L): the associative world model AWM, the causal-pruning method FOCUS, a behavior-cloning world model BC-WM, a recurrent variant R-ICWM, and a generative C-VAE. AWM is the only *graph-free* architecture; every other model, ICWM included, receives the same soft PCMCI mask, a differentiable gate that weighs each input by the significance of its discovered causal links (Appendix L.2). Throughout, σ is the *interventional strength* of a dataset: $\sigma = 0$ is purely observational expert data, $\sigma \rightarrow 1$ is near-random exploration. We use two metrics (Appendix M): *prediction error* (variance-normalized next-state MSE across environments) and *graph error* (normalized spectral distance to the ground-truth graph). Models train on datasets $\mathcal{D}_\sigma^N$ of N episodes at interventional strength σ (Appendix N,O), and are evaluated on held-out biased-optimal ($\sigma = 0.5$), random ($\sigma = 0.9$), and adversarial policies.

Domains. We use three multi-agent coordination domains (Figure 16) featuring high inter-agent confounding: `Two-Door`, where agents must sequentially open a red then blue door for a fast joint reward; `Navigation`, a pathfinding task strategically coupling agents via collision avoidance and



Figure 7: **What actually breaks a world model under shift, what the mask is worth, and how to scale and improve. (I)** Normalized error $\rho \downarrow$ for shifted layouts on the dynamics/position plane ($\rho = 1$ is lossless) across intervention strengths σ. **(II)** Corrupting the causal mask inflates error. **(III Top):** Proposed transition to a canonical, directional composite sector state-space representation for identifiably over raw sensors. **(III Bottom):** Adversarial trajectory error localizes to low-level sensor dimensions, scaling proportionally with agent count. **(IV)** Adversarial error on $2/3/4$-agent navigation (log axis): the line follows each model's geometric mean across team sizes. **(V)** Higher training σ exacerbates transfer degradation under causal mask corruption. **(VI)** Relative per-cell error of the ICWM against the training layout for fixed-geometry shifts. Orange borders indicate cells with significant error deviation.

shared goals; and `Giveway`, where agents in a narrow corridor must negotiate a recessed notch to pass each other without colliding. Shifted environments can be found in Appendix K.2.

We organize the results around six research questions: **(i)** Do soft interventions recover the true causal mechanism (**claim C3**)? **(ii)** Does the learned world model yield positive transfer in surprise settings (**claims C3 & C6**)? **(iii)** What mechanisms are responsible for this transfer (**claim C5**)? **(iv)** When does positive transfer fail, and what breaks? **(v)** How does the framework handle non-Markovian versus Markovian dynamics? **(vi)** Does memory help, and where?

Q(i)**.** Figure 6A evaluates PCMCI discovery against the ground-truth graph as σ increases. Graph error falls steadily in *every* environment under both full and partial observability. This trend survives scaling: with 2, 3, and 4 navigation agents, the recovered graphs improve identically. Thus, the identifiability of the causal mechanism improves directly with interventional strength (see Appendix P for details). **Q**(ii)**.** We evaluate every trained model on an *adversarial* policy far outside the demonstration distribution. In all three environments, variance-normalized error falls monotonically as training σ increases (Figure 6C) by up to two orders of magnitude on navigation matching our theoretical bound that OOD error scales inversely with intervention. The causal world models (ICWM, AWM) significantly outperform the generative C-VAE baseline. Extended zero shot policy evaluations and architectural variation experiments are detailed in Appendix Q. To evaluate reasoning under unexpected state shifts, we freeze each world model, fit predictive density models, and report the per-feature inflation of uncertainty widths (comparing OOD states to their ID equivalents). Figure 6B demonstrates that uncertainty is localized to a few specific dimensions rather than degrading the model output uniformly. The full point cloud study is detailed in Appendix S. **Q**(iii)**.** Figure 6D shows adversarial evaluation error and graph error improve jointly

as σ grows (Figure 3). However, conditioning on σ to compare graph-informed and graph-free models on matched data reveals no direct effect from the graph itself (Figure 5); under full observability, AWM matches ICWM's out-of-distribution error. Thus, soft interventions are the true mechanism for transfer, driving generalization without inducing policy-specific bias. Nevertheless, the graph is not inert. Mask ablation studies demonstrate that perturbing the causal mask severely degrades (Figure 7-II), transfer performance (Appendix R), and downgrade is higher (Figure 7-V) at stronger intervention strengths. The correct causal mask acts as a vital conditional prior, limiting error propagation under partial observability (Figure 1) and providing necessary structure during environmental layout shifts. **Q**(iv)**.** We evaluate trained models on seven altered environments, measuring relative error to quantify the transfer gap (Figure 7-I). We establish that positive transfer is governed by two conditions: (1) the underlying causal mechanisms must be preserved under the shift (ensuring the learned manifold accurately captures the dynamics), and (2) the interventional strength must remain bounded; moderate interventions apply generalizations to learnt model, excessively high interventions corrupt (backed by our proof). Notably, under the `Anti-Causal` Two-Door shift with reversed dynamics, ICWM achieves lossless transfer (Figure 7-VI), reducing the median error to approx. $\frac{1}{6}^{th}$ of the baseline environment. Figure 7-IV illustrates how interventions help multi-agent causal world models generalize to increasing agent counts (note that these results exclude LiDAR observations). As Figure 7-III bottom shows, dropping raw LiDAR inputs reduces adversarial error by up to $157\times$ as agent counts scale. This exposes a fundamental limitation: learning robust causal representations from correlated, low-level sensors under constrained intervention budgets is practically intractable. To resolve this for downstream agentic system design, Figure 7-III top proposes future work replacing raw low level LiDAR beams with a canonical, rank-sufficient minimal representation, as composite sectors, featuring an identifiable causal schema.

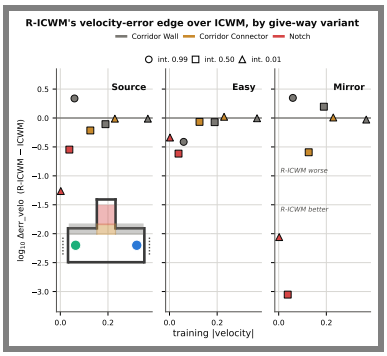


Figure 8: Velocity error, R-ICWM minus ICWM ($\log_{10}$). Below 0 means R-ICWM is better.

Q(v)**.** On a Two-Door variant where the door opens three steps after a toggle (Figure 6-E), mechanism recall is *exactly* 0 for history depths $\tau_{\max} < 3$, but jumps to ≈ 0.6 once $\tau_{\max} \geq 3$ matches the true delay. In-distribution policy evaluation error improves in lockstep, yielding up to a $7\times$ gain under full observation. However, this benefit is highly sensitive to interventional strength, but offers almost no advantage at extreme interventions ($\sigma \geq 0.75$), where the mechanism is already approximately identifiable with high confidence from short histories. **Q**(vi)**.** R-ICWM's recurrent memory excels where single frames lack information. Segmenting the give-way environment by speed (fast `Wall`, slower `Connector`, paused `Notch`) reveals the performance gap tracks speed rather than layout (Figure 8). Both models tie in the fast `Wall`, but R-ICWM improves by one to three orders of magnitude in the stop-start `Notch`. Memory strongly aids slow, discontinuous motion but adds little to fast transitions. This nuance appears only in velocity error because mixed dynamics error averages it away (Appendix U).

6 CONCLUSION AND LIMITATIONS

We introduce ICWM, which learns causal world models from multi-agent demonstrations collected under soft interventions. Our central finding is that interventional strength drives both causal mechanism recovery and out-of-distribution accuracy. By limiting model uncertainty, the discovered causal graphs prevent uniform model collapse under covariate shift. Both established causal theory and our experimental results confirm that positive transfer is a direct byproduct of `mechanism preserving shifts and data fusion` (Bareinboim & Pearl, 2016). Our generalization error bounds (Section 4.3) establish a foundation for runtime verification in self-correcting agents. A world model that accurately isolates causal state transitions can evaluate actions prior to execution for `shielding` (Alshiekh et al., 2017) and provide exact error signals for gradient-free `self-revision` (Shinn et al., 2023). **As limitation**, establishing MAS-SBC through `targeted interventions` where agent coordination might break will require updated proofs to maintain recovery guaranties. **Future work** will connect these justified world models to causal Q-learning (Méndez-Molina et al., 2020) and explore the inherent `steerability gap` in associative models (Vafa et al., 2025a).

Reproducibility Statement

All of our results come from simulated environments with seeded pipelines. We describe the environments and their state and action representations in Appendices K, and the expert policies used to generate demonstrations in Appendix N . We give the causal discovery procedure and its theoretical grounding in Appendix I, and the soft-mask rule in Appendix L.2. We describe the architectures and all baselines in Appendix L, and our evaluation metrics in Appendix M. Every reported number is averaged over multiple seeds. For the cell-paired comparisons behind our main claims, we hold the model, environment, observability, σ, and seed fixed and vary only the factor we are studying. We specify the shifted layouts in Appendix K.2 and the corresponding transfer results in Appendix U, and the mask-corruption procedure in Appendix R.

Ethics Statement

This work studies learned world models in simulated multi-agent environments. It involves no human subjects, no personally identifiable information, and no real-world data, since all demonstrations are generated programmatically in open-source simulators. One point is worth stating clearly. Our method collects data under soft interventions, which means deliberately adding randomness to an agent's actions. This is harmless in simulation, but in a robotic or safety-critical setting it would mean perturbing a system whose failures carry real cost. We therefore do not recommend applying uniform interventions outside simulation without a separate safety layer, and we treat reducing the required intervention budget as a safety concern and not only a question of efficiency. We declare no conflicts of interest or sponsorship that influenced this work.

AI Usage Statement

We used large language models only as assistive tools, for editing prose that we wrote ourselves and for help with implementation and plotting code. We did not use them to generate research ideas, design experiments, produce or analyse results, or find citations. We reviewed all text and code written with LLM assistance. Every reported number comes from the executed pipeline, and we checked every citation against its primary source. The authors take full responsibility for the contents of this paper.

References

Ossama Ahmed, Frederik Träuble, Anirudh Goyal, Alexander Neitz, Yoshua Bengio, Bernhard Schölkopf, Manuel Wüthrich, and Stefan Bauer. Causalworld: A robotic manipulation benchmark for causal structure and transfer learning, 2020. URL https://arxiv.org/abs/2010.04296.

Mohammed Alshiekh, Roderick Bloem, Ruediger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding, 2017. URL https://arxiv.org/abs/1708.08611.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020. URL https://arxiv.org/abs/1907.02893.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture, 2023. URL https://arxiv.org/abs/2301.08243.

Atari, Inc. Gotcha, 1973. Arcade video game.

Shayan Shirahmad Gale Bagi, Zahra Gharaee, Oliver Schulte, and Mark Crowley. Implicit causal representation learning via switchable mechanisms. *arXiv preprint arXiv:2402.11124*, 2024.

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

Matteo Bettini, Ryan Kortvelesy, Jan Blumenkamp, and Amanda Prorok. Vmas: A vectorized multi-agent simulator for collective robot learning, 2022. URL https://arxiv.org/abs/2207.03530.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments, 2024. URL https://arxiv.org/abs/2402.15391.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Juan Correa and Elias Bareinboim. General transportability of soft interventions: Completeness results. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 10902–10912. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/7b497aa1b2a83ec63d1777a88676b0c2-Paper.pdf.

Ishita Dasgupta, Jane Wang, Silvia Chiappa, Jovana Mitrovic, Pedro Ortega, David Raposo, Edward Hughes, Peter Battaglia, Matthew Botvinick, and Zeb Kurth-Nelson. Causal reasoning from meta-reinforcement learning, 2019. URL https://arxiv.org/abs/1901.08162.

Pim de Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning, 2019. URL https://arxiv.org/abs/1905.11979.

Zeyun Deng, Jasorsi Ghosh, Fiona Xie, Yuzhe Lu, Katia Sycara, and Joseph Campbell. Energy-based transfer for reinforcement learning, 2025. URL https://arxiv.org/abs/2506.16590.

Frederick Eberhardt. Almost optimal intervention sets for causal discovery. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, UAI'08, pp. 161–168, Arlington, Virginia, USA, 2008. AUAI Press. ISBN 0974903949.

Arnaud Fickinger. Multi-agent gridworld environment for openai gym. https://github.com/ArnaudFickinger/gym-multigrid, 2020.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution, 2018a. URL https://arxiv.org/abs/1809.01999.

David Ha and Jürgen Schmidhuber. World models. 2018b. doi: 10.5281/ZENODO.1207631. URL https://zenodo.org/record/1207631.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels, 2019. URL https://arxiv.org/abs/1811.04551.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models, 2024. URL https://arxiv.org/abs/2301.04104.

John E. Hopcroft, Rajeev Motwani, and Jeffrey D. Ullman. Introduction to automata theory, languages, and computation, 2nd edition. *SIGACT News*, 32(1):60–65, March 2001. ISSN 0163-5700. doi: 10.1145/568438.568455. URL https://doi.org/10.1145/568438.568455.

Toru Iwatani and Namco. Pac-man, 1980. Arcade video game.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax, 2017. URL https://arxiv.org/abs/1611.01144.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1):99–134, 1998. ISSN 0004-3702. doi: https://doi.org/10.1016/S0004-3702(98)00023-X. URL https://www.sciencedirect.com/science/article/pii/S000437029800023X.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, pp. 267–274, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. ISBN 1558608737.

Sham Machindranath Kakade. *On the sample complexity of reinforcement learning*. PhD thesis, University College London, 2003. URL https://homes.cs.washington.edu/~sham/papers/thesis/sham_thesis.pdf.

Nathan Kallus and Angela Zhou. Confounding-robust policy evaluation in infinite-horizon reinforcement learning, 2020. URL https://arxiv.org/abs/2002.04518.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2):209–232, 2002. doi: 10.1023/A:1017984413808. URL https://link.springer.com/article/10.1023/A:1017984413808.

Thomas Kipf, Elise van der Pol, and Max Welling. Contrastive learning of structured world models, 2020. URL https://arxiv.org/abs/1911.12247.

Daniel Kumor, Junzhe Zhang, and Elias Bareinboim. Sequential causal imitation learning with unobserved confounders, 2022. URL https://arxiv.org/abs/2208.06276.

Yann LeCun et al. A path towards autonomous machine intelligence. *Open Review*, 62(1):1–62, 2022.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task, 2024. URL https://arxiv.org/abs/2210.13382.

Minne Li, Mengyue Yang, Furui Liu, Xu Chen, Zhitang Chen, and Jun Wang. Causal world models by unsupervised deconfounding of physical dynamics, 2020. URL https://arxiv.org/abs/2012.14228.

Kenzo Lobos-Tsunekawa, Akshay Srinivasan, and Michael Spranger. Ma-dreamer: Coordination and communication through shared imagination. *arXiv preprint arXiv:2204.04687*, 2022.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments, 2020. URL https://arxiv.org/abs/1706.02275.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables, 2017. URL https://arxiv.org/abs/1611.00712.

Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. Explainable reinforcement learning through a causal lens. In *AAAI Conference on Artificial Intelligence*, 2019. URL https://api.semanticscholar.org/CorpusID:166228224.

Arquímides Méndez-Molina, Ivan Feliciano-Avelino, Eduardo F. Morales, and Luis Enrique Sucar. Causal based q-learning. *Res. Comput. Sci.*, 149:95–104, 2020. URL https://api.semanticscholar.org/CorpusID:229191955.

Thomas M. Moerland, Joost Broekens, Aske Plaat, and Catholijn M. Jonker. Model-based reinforcement learning: A survey, 2022. URL https://arxiv.org/abs/2006.16712.

Heejeong Nam, Quentin Le Lidec, Lucas Maes, Yann LeCun, and Randall Balestriero. Causal-jepa: Learning world models through object-level latent masking. *arXiv preprint arXiv:2602.11389*, 2026.

Nick Pawlowski, Daniel C. Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference, 2020. URL https://arxiv.org/abs/2006.06485.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009. ISBN 978-0-521-89560-6.

Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Commun. ACM*, 62(3):54–60, February 2019. ISSN 0001-0782. doi: 10.1145/3241036. URL https://doi.org/10.1145/3241036.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference using invariant prediction: identification and confidence intervals, 2015. URL https://arxiv.org/abs/1501.01332.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Elements of causal inference: Foundations and learning algorithms. 2017. URL https://api.semanticscholar.org/CorpusID:86533208.

Dean A. Pomerleau. Alvinn: An autonomous land vehicle in a neural network. In D. Touretzky (ed.), *Advances in Neural Information Processing Systems*, volume 1. Morgan-Kaufmann, 1988. URL https://proceedings.neurips.cc/paper_files/paper/1988/file/812b4ba287f5ee0bc9d43bbf5bbe87fb-Paper.pdf.

Christian Reiser. Causal discovery for time series with latent confounders, 2022. URL https://arxiv.org/abs/2209.03427.

Thomas Richardson and Peter Spirtes. Ancestral graph Markov models. *The Annals of Statistics*, 30 (4):962 – 1030, 2002. doi: 10.1214/aos/1031689015. URL https://doi.org/10.1214/aos/1031689015.

Jonathan Richens and Tom Everitt. Robust agents learn causal world models, 2024. URL https://arxiv.org/abs/2402.10877.

Douglas Ross and John Ward. Mouse in the maze, 1959. MIT TX-0 computer simulation.

Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets, 2022. URL https://arxiv.org/abs/2003.03685.

Jakob Runge, Sebastian Bathiany, Erik M. Bollt, Gustau Camps-Valls, Dim Coumou, Ethan R. Deyle, Clark Glymour, Marlene Kretschmer, Miguel D. Mahecha, Jordi Muñoz-Marí, Egbert H. van Nes, J. Peters, Rick Quax, Markus Reichstein, Marten Scheffer, Bernhard Scholkopf, Peter Spirtes, George Sugihara, Jie Sun, Kun Zhang, and Jakob Zscheischler. Inferring causation from time series in earth system sciences. *Nature Communications*, 10, 2019. URL https://api.semanticscholar.org/CorpusID:189819550.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *Trans. Neur. Netw.*, 20(1):61–80, January 2009. ISSN 1045-9227. doi: 10.1109/TNN.2008.2005605. URL https://doi.org/10.1109/TNN.2008.2005605.

Matthew Schlegel, Wesley Chung, Daniel Graves, Jian Qian, and Martha White. Importance resampling for off-policy prediction, 2019. URL https://arxiv.org/abs/1906.04328.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588 (7839):604–609, December 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-03051-4. URL http://dx.doi.org/10.1038/s41586-020-03051-4.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Towards causal representation learning, 2021. URL https://arxiv.org/abs/2102.11107.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning, 2023. URL https://arxiv.org/abs/2303.11366.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

Roni Stern, Nathan Sturtevant, Ariel Felner, Sven Koenig, Hang Ma, Thayne Walker, Jiaoyang Li, Dor Atzmon, Liron Cohen, T. K. Satish Kumar, Eli Boyarski, and Roman Bartak. Multi-agent pathfinding: Definitions, variants, and benchmarks, 2019. URL https://arxiv.org/abs/1906.08291.

Richard S. Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *SIGART Bull.*, 2(4):160–163, July 1991. ISSN 0163-5719. doi: 10.1145/122344.122377. URL https://doi.org/10.1145/122344.122377.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 2nd edition, 2018. ISBN 978-0-262-03924-6.

Antonios Tragoudaras, Chenyu Zhang, Daniil Cherniavskii, Antonios Vozikis, Thijmen Nijdam, Derck W. E. Prinzhorn, Mark Bodracska, Nicu Sebe, Andrii Zadaianchuk, and Efstratios Gavves. Evaluating newtonian mechanics in video generative models with real physical systems, 2026. URL https://arxiv.org/abs/2504.02918.

Keyon Vafa, Justin Y. Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model, 2024. URL https://arxiv.org/abs/2406.03689.

Keyon Vafa, Sarah Bentley, Jon Kleinberg, and Sendhil Mullainathan. What's producible may not be reachable: Measuring the steerability of generative models, 2025a. URL https://arxiv.org/abs/2503.17482.

Keyon Vafa, Peter G. Chang, Ashesh Rambachan, and Sendhil Mullainathan. What has a foundation model found? using inductive bias to probe for world models, 2025b. URL https://arxiv.org/abs/2507.06952.

Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, UAI '90, pp. 255–270, USA, 1990. Elsevier Science Inc. ISBN 0444892648.

Zizhao Wang, Xuesu Xiao, Zifan Xu, Yuke Zhu, and Peter Stone. Causal dynamics learning for task-independent state abstraction. *arXiv preprint arXiv:2206.13452*, 2022.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdú, and Marcelo J. Weinberger. Inequalities for the l1 deviation of the empirical distribution. 2003. URL https://api.semanticscholar.org/CorpusID:12164823.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

A.D. Wyner. The rate-distortion function for source coding with side information at the decoder 3-ii: General sources. *Information and Control*, 38(1):60–80, 1978. ISSN 0019-9958. doi: https://doi.org/10.1016/S0019-9958(78)90034-7. URL https://www.sciencedirect.com/science/article/pii/S0019995878900347.

Kevin Xia, Yushu Pan, and Elias Bareinboim. Neural causal models for counterfactual identification and estimation, 2022. URL https://arxiv.org/abs/2210.00035.

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Structured causal disentanglement in variational autoencoder, 2023. URL https://arxiv.org/abs/2004.08697.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative, multi-agent games, 2022. URL https://arxiv.org/abs/2103.01955.

Hongxin Zhang, Zeyuan Wang, Qiushi Lyu, Zheyuan Zhang, Sunli Chen, Tianmin Shu, Behzad Dariush, Kwonjoon Lee, Yilun Du, and Chuang Gan. Combo: Compositional world models for embodied multi-agent cooperation, 2025a. URL https://arxiv.org/abs/2404.10775.

Jiaqi Zhang, Kristjan Greenewald, Chandler Squires, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. *Advances in Neural Information Processing Systems*, 36:50254–50292, 2023.

Junzhe Zhang and Elias Bareinboim. Markov decision processes with unobserved confounders : A causal approach. 2016. URL https://api.semanticscholar.org/CorpusID:5195298.

Junzhe Zhang and Elias Bareinboim. Bounding causal effects on continuous outcome. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13):12207–12215, May 2021. doi: 10.1609/aaai.v35i13.17449. URL https://ojs.aaai.org/index.php/AAAI/article/view/17449.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schoelkopf. Kernel-based conditional independence test and application in causal discovery, 2012. URL https://arxiv.org/abs/1202.3775.

Weinan Zhang, Xihuai Wang, Jian Shen, and Ming Zhou. Model-based multi-agent policy optimization with adaptive opponent-wise rollouts, 2022. URL https://arxiv.org/abs/2105.03363.

Yang Zhang, Xinran Li, Jianing Ye, Shuang Qiu, Delin Qu, Xiu Li, Chongjie Zhang, and Chenjia Bai. Revisiting multi-agent world modeling from a diffusion-inspired perspective, 2025b. URL https://arxiv.org/abs/2505.20922.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. Dags with no tears: Continuous optimization for structure learning, 2018. URL https://arxiv.org/abs/1803.01422.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning, 2025. URL https://arxiv.org/abs/2411.04983.

Jiayi Zhu, Jianing Zhang, Yiying Yang, Wei Cheng, and Xiaoyun Yuan. Shareverse: Multi-agent consistent video generation for shared world modeling, 2026. URL https://arxiv.org/abs/2603.02697.

Zheng-Mao Zhu, Xiong-Hui Chen, Hong-Long Tian, Kun Zhang, and Yang Yu. Offline reinforcement learning with causal structured world models, 2022. URL https://arxiv.org/abs/2206.01474.

Learning Implicit Causal World Models from Multi-Agent Demonstrations (Supplementary Material)

INDEX OF CONTENTS (SUPPLEMENTARY MATERIAL)

This appendix provides extended formalisms, mathematical proofs, and detailed experimental configurations supporting the findings in the main text. Please access the supplementary material for accessing the **[Full Appendix]**.

- **Appendix A: Impact of the Paper**
- **Appendix B: Structural Causal Model Formalization**
  - Exogenous and Endogenous Partitioning
  - Structural Functions and Dependencies
  - The Role of $P(U)$
- **Appendix C: Causal Decision Graph**
  - MDP Regimes (Full Observability)
  - POMDP Regimes (Partial Observability)
  - The Observation Space and Information Bottleneck
- **Appendix D: Formalizing Policy Mixtures as Causal Interventions**
  - Mapping the Mixture to a Structural Mechanism Shift
  - The Role of Entropy in Structural Discovery
- **Appendix E: Theoretical Foundations of the Sequential Backdoor Condition**
  - The Manipulated Causal Decision Graph
  - Formal Definition of the Sequential Backdoor Criterion
  - The Structural Violation in Multi-Agent Environments
- **Appendix F: Formal Proof of the MAS Sequential Backdoor Condition**
  - Structural Axioms of the Multi-Agent System
  - Step 1: Graphical d-Separation via Induction
  - Step 2: Algebraic Reduction via do-Calculus
- **Appendix G :[See Full Appendix] Motivation Study in Tic-Tac-Toe**
- **Appendix H: Generalization Error Bound of the ICWM**
  - Discussion of Proven Implications
- **Appendix I: Theoretical Guarantees and Algorithmic Selection for Causal Discovery**
  - Completeness of the Lagged Causal Skeleton in PCMCI
  - Algorithmic Selection: PCMCI vs. Latent Discovery (LPCMCI)
- **Appendix J:[See Full Appendix] Contextual Bandit Formulation of Decision-Making**
  - The Intractability of Expected Information Gain
  - Surrogate Objectives
- **Appendix K: Environments**
  - **Appendix K.1**: MAS Environment Visualization
  - **Appendix K.2: Shifted Environments**
  - **Appendix K.3:[See Full Appendix] State Representation of Different Environments**
    - Two-Door Coordination
    - Giveway Corridor
    - Multi-Agent Navigation
  - **Appendix K.4:[See Full Appendix] Action Space Representation**
    - Continuous Force-Based Action Space (VMAS)
    - Discrete Categorical Action Space (MiniGrid)
    - Comparative Summary
  - **Appendix K.5:[See Full Appendix] Ground Truth Causal Adjacency Matrices**
- **Appendix L: Neural Architecture Design and Baselines**
  - Summary of Model Properties

  - Causal Structural Masking and AMLS
  - Baseline Architectures
  - Optimization and Loss Functions
- **Appendix M:[See Full Appendix] Evaluation Metrics**
  - Toy Study: Validating Δ_{variance}
  - Toy Evaluation: Density Metrics vs. the Variance Metric
  - Causal Metric Stress Tests
- **Appendix N:[See Full Appendix] MAPPO Training and Evaluation Performance of Expert Policy**
- **Appendix O:[See Full Appendix] Dataset Cardinality and Evaluation**
  - Scaling Dynamics and Cardinality Bounds
  - Evaluation Dataset Configuration
  - Quality of the Collected Datasets ($\mathcal{D}_{\sigma}^{N}$)
  - Evaluation Dataset Configuration: Two, Three, Four Agent Navigation
- **Appendix P: Causal Discovery Setup and Results**
  - Causal Discovery Experimental Parameters
  - Evaluation Results: Three Primary Environments
  - Evaluation Results: High Sample Complexity Experiments
  - Evaluation Results: Two, Three, Four Agent Navigation
- **Appendix Q: World Model Setup and Results**
  - Evaluation Results: Three Primary Environments
  - Evaluation Results: High Sample Complexity
  - Evaluation Results: Two, Three, Four Agent Navigation
- **Appendix R:[See Full Appendix] Causal-Mask Ablation Study**
- **Appendix S:[See Full Appendix] Uncertainty and Neural Density Measurement**
  - Grid World: Two-Door
  - Continuous Control: Give-Way
  - Two-Agent Navigation
  - Multi-Agent Navigation without LIDAR
  - Quantitative Uncertainty Metrics
- **Appendix T:[See Full Appendix] World-Model Spatial Error Density**
  - Setup, Figures, Quantitative Summary, Discussion
- **Appendix U:[See Full Appendix] Generalization under Layout Shift**
  - Setup, Figures, Quantitative Summary, Discussion
  - Where the Shift Actually Bites: Subtracting the Source Error Surface
  - Splitting the Velocity Channel: Where Recurrence Pays Off
  - Overall Takeaway
- **Appendix V: Towards Pearl's Framework for Counterfactual Queries**
  - The Pooled Correlation is Confounded
  - A Localized, Conditional Exception
  - Why Stretch Benefits and Rotation Does Not

NOTATION GLOSSARY

The following tables summarize the mathematical notation used throughout the paper, categorized by their respective methodological sections (see Tables 1, 2, 3, 4, and 5).

Symbol	Definition
Evaluation Metrics and Generalization Bounds	
D_{dim}	Dimensionality of state vector
$\mathcal{D}_\sigma$	Demonstration dataset
MSE	Mean Squared Error
RSE	Relative System Error
SHD	Structural Hamming Distance
Δ_{inv}	Policy Invariance Gap
Δ_{spec}	Spectral Distance
Δ_{variance}	Variance-normalized relative error, MSE/$\sigma^2_{\mathcal{D}}$
$\sigma^2_{\mathcal{D}}$	Target-variance floor
$\bar{s}$	Input-agnostic mean successor state
R^2	Coefficient of determination
$\mathcal{C}_{pred}, \mathcal{C}_{true}$	Predicted and ground-truth causal tensors
$\lambda_i(\cdot)$	i-th eigenvalue of an adjacency representation
$\widehat{\text{MMD}}^2$	Unbiased squared MMD (RBF kernel)
$k(u, v)$	RBF kernel, $\exp(-\gamma\|u - v\|^2)$
γ^{-1}	MMD kernel bandwidth (median heuristic)
$\hat{p}(z)$	Kernel density estimate at query point z
h	KDE bandwidth
$\mathcal{G}$	Query grid for density evaluation
$\hat{p}_{\text{norm}}, \hat{p}_{\text{adv}}$	Normal and adversarial next-state densities
Dist_{TV}	Total Variation distance
J	Complexity constant
ϵ_{env}	Irreducible stochasticity / error

Table 1: Evaluation Metrics and Generalization Bounds

Symbol	Definition
Causal Discovery and Neural Masking	
τ	Temporal lag limit
$E_{lag}, E_{contemp}$	Discovered edge sets
$\mathcal{G}_{skeleton}$	Unconfounded causal skeleton
$\mathcal{C}$	Adjacency Matrix with Lagged Structure
p_{ij}	Causal discovery p-value
α	Significance threshold
ψ_{ij}	Log-Significance Transform value
κ	Temperature parameter
Γ	Differentiable soft mask tensor
M	Binary causal mask tensor
Φ, θ	Neural transition model parameters

Table 2: Causal Discovery and Neural Masking

Symbol	Definition
Soft Interventions and Sequential Backdoor	
σ	Interventional strength ($\sigma \in [0, 1]$)
π^i_{exp}	Deterministic expert policy for agent i
ζ	Independent stochastic process
P^σ	Interventional stationary mixture policy
H_k	Observed historical trajectory up to step k
$do(\cdot)$	Pearl's intervention operator
A_t^{-i}, a_t^{-i}	Joint actions excluding agent i
$\mathcal{G}_{\overline{A}\underline{A}}$	Mutilated causal graph

Table 3: Soft Interventions and Sequential Backdoor

Symbol	Definition
Uncertainty, Error Density, and Shift	
$\sigma_{\text{ID}}, \sigma_{\text{OOD}}$	Predictive cloud width per dimension
$\sigma_{\text{OOD}}/\sigma_{\text{ID}}$	Dimensionless OOD width-inflation ratio
$\overline{\text{err}}_{\text{dyn}}$	Mean squared residual on dynamic dimensions
$\overline{\text{err}}_{\text{pos}}$	Mean squared residual on position dimensions
ρ	Shift-transfer ratio
$\mathcal{D}$	Set of dynamic state dimensions
$\Delta^{\text{norm}}_{\text{spec}}$	Normalized binary spectral distance
$\mathcal{D}^N_\sigma$	Training dataset of N episodes
$\mathcal{D}^\ddagger_\sigma$	Held-out evaluation dataset

Table 4: Uncertainty, Error Density, and Shift

Symbol	Definition
SCM and Decision Graph	
$\mathcal{M}$	SCM tuple $\langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$
$\mathbf{U}, \mathbf{V}$	Sets of exogenous/endogenous variables
u_{env}	Unobserved environmental noise
u^i_{intent}	Unobserved latent intent of agent i
s_t, S_t	Environment state at time t
$\mathbf{a}_t, \mathbf{A}_t$	Joint action of all agents at time t
a^i_t, A^i_t	Specific action of agent i at time t
r_t	Reward received at time t
o_t, O_t	Partial observation at time t
f_s, f_r, f_{a^i}	Deterministic structural functions

Table 5: SCM and Decision Graph

A IMPACT OF THE PAPER

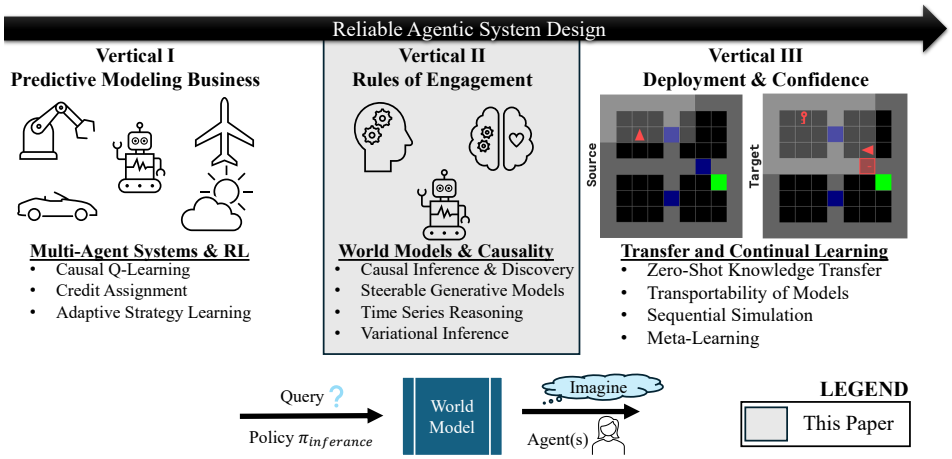


Figure 9: **The Synergistic Vertical Framework for Reliable Agentic Systems**. The architecture partitions system design into three critical pillars: behavioral coordination (Vertical I), structural understanding (Vertical II), and cross-domain scalability (Vertical III). This paper addresses Vertical II (Rules of Engagement), establishing the deconfounded causal world model as the essential physical foundation. By resolving invariant environment mechanisms from strategic agent intent, our work provides the identifiable internal simulator required for precise credit assignment in Vertical I and zero-shot knowledge transportability in Vertical III.

B STRUCTURAL CAUSAL MODEL FORMALIZATION

The transition dynamics of a Multi-Agent System are governed by the interplay between invariant environment rules and adaptive agent strategies. We formalize this via an SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$:

B.1 EXOGENOUS AND ENDOGENOUS PARTITIONING

The variables in our system are partitioned into two sets:

1. **Exogenous Variables (U)**: These are variables determined outside the model. In our framework, we distinguish between Environmental Stochasticity ($u_{env} \sim U_{env}$), which captures irreducible noise in physics (e.g., sensor noise or friction or stochastic in environment properties), and Unobserved Strategic Intents for agent i, ($u^i_{intent} \sim U_{intent}$), which represent the high-level goals or coordination signals that drive agent behaviors but are not explicitly captured in the state space.
2. **Endogenous Variables (V)**: These are the observable variables determined by the structural functions. For a single transition step (t), these include the current environment state $s_t \in \mathcal{S} = \mathcal{S}^1 \times \cdots \times \mathcal{S}^n$, the joint action $a_t \in \mathcal{A} = \mathcal{A}^1 \times \cdots \times \mathcal{A}^n$, the reward $r_t \in \mathcal{R} = \mathcal{R}^1 \times \cdots \times \mathcal{R}^n$, and the resulting next state $s_{t+1} \in \mathcal{S}$. Where $\mathcal{S}$ is the state space, $\mathcal{S}^i$ is the state space for agent i, $\mathcal{A}$ is the action space, $\mathcal{A}^i$ is the action space for agent i, $\mathcal{R}$ is the reward space, $\mathcal{R}^i$ is the reward space for agent i, and n number of agents in the multi agent system.

B.2 STRUCTURAL FUNCTIONS AND DEPENDENCIES

The set $\mathcal{F} = \{f_s, f_r f_{a^1}, \ldots, f_{a^n}\}$ defines the causal mechanisms of the world. By expressing these as deterministic functions of their parents and noise, we can model the system as a directed graph where edges represent functional dependencies:

1. **Environment Physics** (f_s): This function is the primary target of our world model. It maps the current state and joint actions to the next state. Critically, f_s is policy-invariant;

it represents the underlying "rules of the game" that remain constant regardless of whether the agents are acting optimally, randomly, or adversarially.

- Quantity: $P(s_{t+1} \mid s_t, \mathbf{a}_t)$
- Relationship: $s_{t+1} \leftarrow f_s(s_t, \mathbf{a}_t, u_{env})$

2. **Reward Model (f_r)**: This maps the state-action pair to a feedback signal. In MAS, rewards are often highly confounded because expert agents only receive high rewards by following specific, intent-driven trajectories.
   - Quantity: $P(r_{t+1} \mid s_t, \mathbf{a}_t)$
   - Relationship: $r_{t+1} \leftarrow f_r(s_t, \mathbf{a}_t, u_{env})$
3. **Agent Policies (f_{a^i})**: Unlike f_s, the policy functions are agent-dependent. They are conditioned on the state s_t and the latent intent u^i_{intent}. This structure is the root of causal confusion: in expert data, the correlation between s_t, u^i_{intent}, and s_{t+1} is often so strong that standard associative models cannot determine if s_{t+1} was caused by the physics of s_t or the specific intent-driven choice of a^i_t.
   - Quantity: $\pi^i(a^i|s_t)$, where π^i is i^{th} agents policy.
   - Relationship: $a^i_t \leftarrow f_{a^i}(s_t, u^i_{intent})$, the following relation is estabished from prior work (Lowe et al., 2020; Dasgupta et al., 2019)

B.3 The Role of $P(U)$

The joint probability distribution $P(\mathbf{U})$ induces the statistical distributions in the demonstration dataset. We distinguish between these exogenous variables as follows:

1. Environmental Stochasticity (u_{env}): Represents irreducible noise (e.g., friction, jitter, weather conditions). We assume $u_{env} \perp\!\!\!\perp u_{intent}$. This term dictates predictive variance (kernel thickness or uncertainty in the model's output), which we do not explicitly deconfound. We assume u_{env} is captured by dataset variance and offer no solution for shifts in its distribution, focusing solely on structural invariance. Our position is that under a significant distribution shift of u_{env}, the world dynamics themselves have likely shifted so fundamentally that knowledge reuse from the previous model would yield diminishing returns.
2. Strategic Confounding (u^i_{intent}): Represents unobserved coordination signals that correlate with s_t in expert data, creating backdoor paths ($s_t \leftarrow u^i_{intent} \rightarrow a^i_t \rightarrow s_{t+1}$) that cause associative models to mistake agent choice for physical law. We address this by leveraging policy variance (soft interventions) to satisfy the Sequential Backdoor Condition discovering true transition rules f_s.

Physical Justification of the Exogenous Partition

Toy Example 1: Consider an autonomous vehicle with a latent intent (u_{intent}) to run a yellow light. The environment updates the vehicle's future coordinates (s_{t+1}) strictly based on the applied wheel torque (a_t). Software intent cannot alter physical coordinates without passing through the throttle; assuming a direct structural edge $u_{intent} \rightarrow s_{t+1}$ that bypasses a_t implies the agent possesses telekinesis. Thus, strategic intent strictly bottlenecks through physical actions.

Toy Example 2: Consider a drone flying in gusting wind. The wind directly alters the drone's position (s_{t+1}) regardless of the rotor speeds chosen by the policy (a_t), creating a direct lagged path. In our SCM, persistent physical forces like wind are explicitly partitioned as environmental stochasticity (u_{env}). The MAS-SBC adjustment is designed to block non-causal paths originating from strategic coordination (u_{intent}), whereas direct physical confounders (u_{env}) are naturally absorbed by the transition function $f_s(s_t, a_t, u_{env})$.

C Causal decision graph

The Causal Decision Graph (CDG) allows us to visualize how structural dependencies evolve and where unobserved variables interfere with transition discernibility.

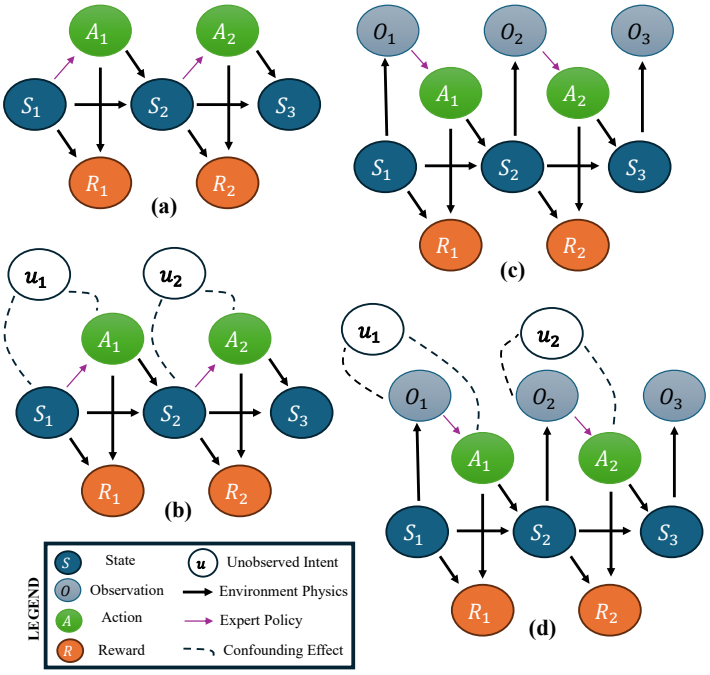


Figure 10: Causal Decision Graph for MAS (a) MDP (b) MDP with Confounding (c) Partially Observable MDP (d) Partially Observable MDP with Confounding

C.1 MDP Regimes (Full Observability)

- **Standard MDP** (Fig. 10a): (Sutton & Barto, 2018) : Actions are a direct functional child of the observed state ($S_t \to A_t$). The transition $S_{t+1} \leftarrow f_s(S_t, A_t, u_{env})$ is identifiable because no unobserved common causes exist between A_t and S_{t+1}.
- **Confounded MDP** (Fig. 10b): We introduce unobserved intent U. As shown by the dashed lines, U influences A_t while being correlated with the state sequence. This creates a backdoor path ($S_t \leftarrow U \to A_t \to S_{t+1}$). Associative models fail here because they cannot distinguish if S_{t+1} resulted from the physical action A_t or the latent intent U.

C.2 POMDP Regimes (Partial Observability)

- **Standard POMDP** (Fig. 10c): (Kaelbling et al., 1998) : The state S_t is latent. Agents perceive O_t, creating an information bottleneck ($S_t \to O_t \to A_t$). While the physics f_s still depend on S_t, the policy is constrained by the observation.
- **Confounded POMDP** (Fig. 10d): This represents the most realistic MAS scenario. The unobserved intent U influences the action A_t, which is already limited by the partial observation O_t. Recovering the invariant physics f_s in this regime requires ICWM to utilize the Sequential Backdoor Condition to d-separate the interventional influence of actions from the latent confounding of U.

C.3 The Observation Space and Information Bottleneck

In partially observable settings, we formally introduce the Observation Space $\mathcal{O}$. This space acts as a non-invertible, lossy mapping from the latent state:

- Formal Mapping: The observation $o_t \in \mathcal{O}$ is generated by an observation function f_o:
$$o_t \leftarrow f_o(s_t, u_{env}) \tag{9}$$
- The Information Bottleneck: This regime imposes a critical constraint: while the environment physics f_s remain dependent on the full physical configuration s_t, the agent's policy is restricted to the information contained in o_t. This creates a bottleneck where the agent lacks the state-transparency needed to determine its own causal influence.

- Structural Challenge: Because f_o is often a many-to-one mapping, multiple physical states s_t can map to the same o_t. This aliasing exacerbates causal confusion, as the world model must not only deconfound strategic intent but also infer the underlying state s_t from the observation history to correctly isolate the physical laws of the transition.

D Formalizing Policy Mixtures as Causal Interventions

D.1 Mapping the Mixture to a Structural Mechanism Shift

In a Structural Causal Model (f as process equation), variables (A) are deterministic functions of causal parents (PA_A) and exogenous noise (U_A): $A \leftarrow f(PA_A, U_A)$. A hard intervention replaces f with a constant. A soft intervention modifies the equation to a new mechanism f' without fully severing parental dependence.

Our stochastic policy mixture operates strictly as a soft intervention. Probabilistically, we define this mixture as the interventional distribution:

$$P^{\sigma}(a_t^i \mid s_t) = (1-\sigma)\pi_{exp}^i(a_t^i \mid s_t) + \sigma\zeta(a_t^i) \tag{10}$$

where π_{exp}^i is the deterministic expert policy, $\sigma \in [0,1]$ dictates the interventional strength, and ζ is an independent stochastic process defined over the agent's discrete or continuous action space $\mathcal{A}^i$.

Structurally, in purely observational data, actions follow the expert mechanism: $a_t^i \leftarrow f_{a^i}(s_t, u_{intent}^i)$. The mixture replaces f_{a^i} with a modified function f'_{a^i} by introducing an independent exogenous noise variable $u_{rand} \sim \zeta$:

$$a_t^i \leftarrow f'_{a^i}(s_t, u_{intent}^i, u_{rand}) = \begin{cases} f_{a^i}(s_t, u_{intent}^i) & \text{with probability } 1-\sigma \\ u_{rand} & \text{with probability } \sigma \end{cases} \tag{11}$$

Because $u_{rand} \perp\!\!\!\perp u_{intent}$, this piecewise assignment redefines the data-generating mechanism. The target variable a_t^i retains its functional dependence on s_t, fulfilling the soft intervention criteria, while the structural influence of the confounding parent u_{intent}^i degrades proportionally to σ.

D.2 The Role of Entropy in Structural Discovery

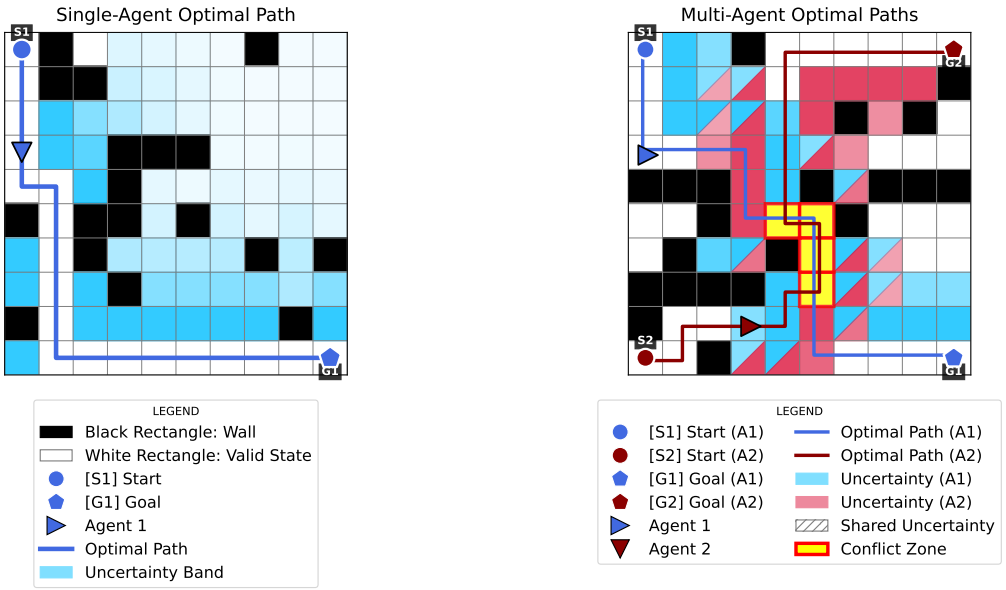


Figure 11: Navigation trajectories from source to goal for a single agent (left) and two agents (right). Optimal path and uncertainty bands are highlighted separately. Only deterministic optimal actions create spurious correlations between agent trajectories, and this confounding effect is strictly higher in multi-agent settings.

Figure 11 illustrates the formal research problem of structural discovery in maze navigation games (Stern et al., 2019; Iwatani & Namco, 1980; Atari, Inc., 1973; Ross & Ward, 1959). The Shannon entropy H of the policy quantifies the interventional strength introduced by the mixture parameter σ, defining the identifiability of the transition function f_s. In a purely deterministic expert dataset ($\sigma = 0$, $H = 0$), the absolute lack of action diversity perfectly aligns the unobserved intent u^i_{intent} with the selected action a^i_t. This zero-entropy regime creates a persistent backdoor path ($s_t \leftarrow u^i_{intent} \rightarrow a^i_t \rightarrow s_{t+1}$) where environment dynamics remain entirely conflated with agent strategy, rendering f_s structurally unidentifiable. As σ increases toward 1, the injected independent variance raises the dataset entropy. This variance guarantees causal positivity by ensuring all legally valid actions retain a non-zero probability of selection, which inherently generates the natural experiments required to observe alternative state transitions.

The primary motivation for resolving this zero-entropy confounding is to enable the extraction of invariant environmental laws from biased observational data. While a partial policy mixture ($0 < \sigma < 1$) does not completely sever the structural backdoor path, the induced entropy provides the unconfounded state and action support strictly necessary to compute unbiased sequential backdoor adjustments. By achieving causal positivity, the model can structurally decouple the agents' strategic intents from the physical environment. Isolating the true physical gradient of the transition function f_s ensures that the learned internal simulator remains accurate and robust even when surrounding agents deviate from their optimal training distributions.

E Theoretical Foundations of the Sequential Backdoor Condition

To establish the formal requirement for our Multi-Agent Sequential Backdoor Condition (MAS-SBC) in Theorem 1, we first detail the structural mechanisms of the baseline single-agent SBC framework.

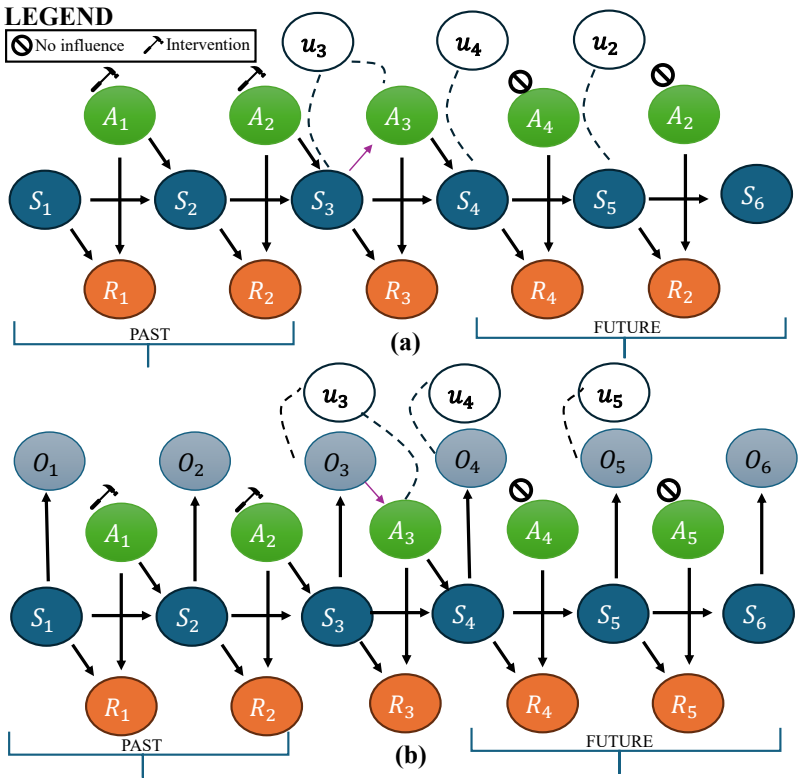


Figure 12: Manipulated Causal Decision Graph for a single agent (a) MDP with SBC at $k = 3$ (b) Partially Observable MDP with SBC at $k = 3$

E.1 The Manipulated Causal Decision Graph

To evaluate the causal impact of actions over a trajectory, we perform graph surgery of the causal decision graph $\mathcal{G}$ to create a manipulated graph $\mathcal{G}_{\overline{\mathbf{a}}_{1:k-1},\underline{\mathbf{a}}_{k+1:t}}$ (Figure 12). For a specific time step k, the graph is modified as follows:

- **Past Actions ($\overline{\mathbf{a}}_{1:k-1}$)**: The incoming edges to previous actions are removed. This isolates past actions from their historical confounders (such as latent intent U), simulating a strict sequence of prior hard interventions.
- **Future Actions ($\underline{\mathbf{a}}_{k+1:t}$)**: The outgoing edges from future actions are removed. This renders future decisions independent of their causal descendants in the graph, preventing downstream planned actions from statistically polluting the current transition gradient.

This graph surgery isolates the forward-directed causal effect of the current action on the environment from backward-flowing statistical correlations.

E.2 Formal Definition of the Sequential Backdoor Criterion

The Sequential Backdoor Criterion (SBC) provides the exact graph-theoretic test to determine if an interventional query can be resolved using purely observational data. The interventional distribution for an outcome $Y \in \{s_{t+1}, r_t\}$ given a sequence of actions $do(\mathbf{a}_1, \dots, \mathbf{a}_t)$ is identifiable if, for every $k \in \{1, \dots, t\}$, there exists a set of covariates $\mathbf{Z}_k \subseteq \mathbf{V}$ satisfying:

- $\mathbf{Z}_k$ are non-descendants of any future action in $\{\mathbf{a}_k, \dots, \mathbf{a}_t\}$.
- $(Y \perp \mathbf{a}_k \mid \mathbf{a}_1, \dots, \mathbf{a}_{k-1}, \mathbf{Z}_1, \dots, \mathbf{Z}_k)$ in the manipulated graph $\mathcal{G}_{\overline{\mathbf{a}}_{1:k-1},\underline{\mathbf{a}}_{k+1:t}}$.

To bridge this foundational theory to our specific world model formulation, the outcome Y represents our fundamental learning targets: the physical state transition (s_{t+1}) and the environment reward (r_t). The theoretical sequence of covariates $\mathbf{Z}_{1:k}$ maps directly to our observable trajectory history $H_k = (\mathbf{s}_{0:k}, \mathbf{a}_{0:k-1})$.When these structural conditions hold, the interventional distribution for the outcome mechanism is mathematically identified by:

$$P(Y \mid do(\mathbf{a}_{1:t})) = \sum_{\mathbf{z}_{1:t}} P(Y \mid \mathbf{z}_{1:t}, \mathbf{a}_{1:t}) \prod_{k=1}^{t} P(\mathbf{z}_k \mid \mathbf{z}_{1:k-1}, \mathbf{a}_{1:k-1}) \tag{12}$$

This formulation confirms that estimating the causal effect of the action history $\mathbf{a}_{1:t}$ on the transition to state s_{t+1} or reward r_t is possible. The history H_k acts as a sufficient statistic to block confounding paths from the unobserved intent U, rendering the physical mechanisms structurally identifiable for a single agent.

Intuitively, the Sequential Backdoor Condition ensures we have observed enough history ($\mathbf{Z}_k$) to decouple an agent's actions from their hidden strategic intent. Equation 12 then acts as a re-weighting formula that uses this history to simulate a "fair experiment" (Schlegel et al., 2019). This allows us to calculate the environment's true response as if we were directly controlling the agents, rather than simply observing their coordinated behavior. The conditions of the SBC are illustrated in Figure 12, which demonstrates that past actions are fixed and future actions cannot be influenced by the current decision.

E.3 The Structural Violation in Multi-Agent Environments

The baseline SBC fundamentally assumes an isolated decision process. When applied to a multi-agent system, the structural conditions formally fail (Figure 13). The actions of other agents a_k^{-i} exist contemporaneously with the target agent's action a_k^i. Both sets of actions are driven by a shared or correlated latent intent U. This topology creates a secondary, dynamic backdoor path: $a_k^i \leftarrow U \rightarrow a_k^{-i} \rightarrow Y$. Because the contemporaneous actions a_k^{-i} occur exactly at time k, they cannot be included in the historical covariate set $\mathbf{Z}_k$ (our history H_k) prior to the action a_k^i. Consequently, the observed history fails to d-separate a_k^i from U, directly violating Condition 2 of the SBC definition within the manipulated graph $\mathcal{G}_{\overline{\mathbf{a}}_{1:k-1},\underline{\mathbf{a}}_{k+1:t}}$.

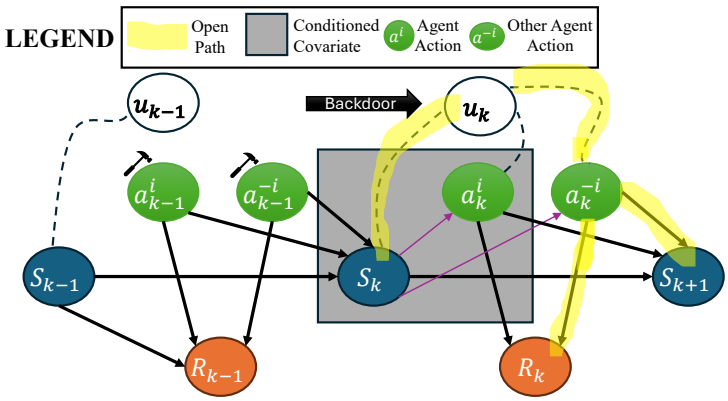


Figure 13: The dynamic confounding path in multi-agent systems. Contemporaneous actions (a_k^{-i}) driven by shared intent (U) create an unblocked backdoor path ($a_k^i \leftarrow U \rightarrow a_k^{-i} \rightarrow Y$). The standard historical conditioning set (H_k) fails to d-separate the target action from the latent intent, necessitating the MAS-SBC.

The path through U and a_k^{-i} remains open, meaning the interventional distribution fails to reduce to the observational expression. The environment physics remain permanently entangled with the intent of the surrounding agents. This mathematical failure establishes the strict necessity for extending the framework to the Multi-Agent Sequential Backdoor Condition (MAS-SBC), which executes the backdoor adjustment over the complete contemporaneous action profile.

F Formal Proof of the MAS Sequential Backdoor Condition

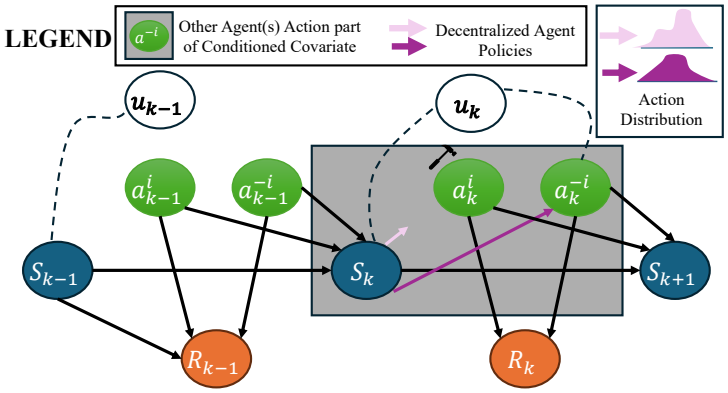


Figure 14: **Mutilated Causal Decision Graph $\mathcal{G}_{\overline{A^i}}$ for MAS-SBC Identifiability**. The hammer indicates the intervention $do(a_k^i)$, which severs incoming edges from H_k and U. Conditioning on the intact peer actions (a_k^{-i}) blocks the contemporaneous backdoor path.

To establish the structural identifiability of the world model, we utilize Pearl's do-calculus to evaluate the causal dynamics within the mutilated graph $\mathcal{G}_{\overline{A^i_{1:k-1}}\underline{A^i_{k+1:t}}}$, isolating the specific intervention on A_k^i by severing its relevant sequential dependencies.

F.1 Structural Axioms of the Multi-Agent System

We establish the true Data Generating Process using three strict structural axioms:

- **Axiom 1: Transition Markovianity:** $S_{k+1} \perp\!\!\!\perp U \mid (H_k, A_k^i, A_k^{-i})$. The physical transition mechanism f_s depends exclusively on the observed history and realized joint actions.

- **Axiom 2: Decentralized Policy:** $A_k^i \perp\!\!\!\perp A_k^{-i} \mid (H_k, U)$. Agents act simultaneously without direct causal influence on one another ($A_k^i \not\rightsquigarrow A_k^{-i}$). The only connections between contemporaneous actions are backdoor paths through shared history H_k and overarching strategic intent U.
- **Axiom 3: Confounding Structure:** $U \to A_k^i$ and $U \to A_k^{-i}$. The latent intent acts as an unobserved contemporaneous confounder, coordinating all agent actions.

F.2 Step 1: Graphical d-Separation via Induction

To apply do-calculus, we must first prove that our chosen covariate set $Z_k = \{H_k, a_k^{-i}\}$ successfully blocks all non-causal paths pointing into a_k^i. We require the assumption of **Sufficiency of History**: The observed history H_k acts as a sufficient statistic for the latent intent U, meaning $s_{k+1} \perp\!\!\!\perp U \mid (H_k, a_k^i, a_k^{-i})$.

Inductive Proof of MAS d-Separation. **Base Case (Single-Agent):** Assume peer agents a_k^{-i} are removed. The only backdoor path is historical: $a_k^i \leftarrow U \to H_k \to s_{k+1}$. Under the sufficiency of history, conditioning on $Z_k = \{H_k\}$ screens off U and blocks this path, satisfying the single-agent Sequential Backdoor Criterion.

Inductive Step (Multi-Agent Extension): Reintroducing peer agents a_k^{-i} creates a new, contemporaneous backdoor path: $a_k^i \leftarrow U \to a_k^{-i} \to s_{k+1}$. The base case set $\{H_k\}$ is historical and cannot block this instantaneous path. We expand the set to $Z_k = \{H_k, a_k^{-i}\}$. By Axiom 2, peer action a_k^{-i} acts as a chain node on the confounding path, not a collider. Therefore, explicitly conditioning on a_k^{-i} mathematically blocks the new multi-agent backdoor path.

Conclusion: Conditioning on a_k^{-i} reduces the unblocked topology back to the single-agent base case. Conditioning on H_k subsequently clears the remaining historical path. Because no unblocked paths remain, the conditional independence $(s_{k+1} \perp\!\!\!\perp a_k^i \mid H_k, a_k^{-i})$ holds in the mutilated graph $\mathcal{G}_{\overline{a_k^i}}$. □

F.3 Step 2: Algebraic Reduction via do-Calculus

With graphical d-separation proven, we resolve the interventional query to purely observational data.

Proof of Theorem 1. We expand the target interventional distribution $P(s_{k+1} \mid do(a_k^i), H_k)$ by marginalizing over the peer actions A_k^{-i} and the unobserved confounder U:

$$= \sum_{a_k^{-i}} \sum_{U} P(s_{k+1} \mid do(a_k^i), a_k^{-i}, H_k, U) P(a_k^{-i} \mid do(a_k^i), H_k, U) P(U \mid do(a_k^i), H_k)$$

Applying the rules of do-calculus based on our established axioms and d-separation:

1. **Isolating the Intervention:** $do(a_k^i)$ removes edges into A_k^i. It cannot backward-influence U or H_k. Thus, $P(U \mid do(a_k^i), H_k) = P(U \mid H_k)$.
2. **Applying Axiom 2:** A_k^{-i} is not a descendant of A_k^i. Intervening on A_k^i does not trigger a simultaneous reaction. Thus, $P(a_k^{-i} \mid do(a_k^i), H_k, U) = P(a_k^{-i} \mid H_k, U)$.
3. **Applying Axiom 1:** The transition is fully determined by observed variables. Because A_k^i is now a fixed parent, we drop the do-operator. S_{k+1} is independent of U given the actions. Thus, $P(s_{k+1} \mid do(a_k^i), a_k^{-i}, H_k, U) = P(s_{k+1} \mid a_k^i, a_k^{-i}, H_k)$.

Substituting these reductions back yields:

$$= \sum_{a_k^{-i}} P(s_{k+1} \mid a_k^i, a_k^{-i}, H_k) \sum_{U} P(a_k^{-i} \mid H_k, U) P(U \mid H_k)$$

By the Law of Total Probability, the inner summation over U reduces exactly to the marginal observational probability of the other agents' actions:

$$P(s_{k+1} \mid do(a_k^i), H_k) = \sum_{a_k^{-i}} P(s_{k+1} \mid a_k^i, a_k^{-i}, H_k) P(a_k^{-i} \mid H_k)$$

This concludes the proof, demonstrating structural identifiability. □

H GENERALIZATION ERROR BOUND OF THE ICWM

To formalize the necessity of stochastic soft interventions, we establish a theoretical bound on the out-of-distribution policy evaluation error. We explicitly scope our theoretical analysis to tabular Markov Decision Processes with discrete state space $\mathcal{S}$ and joint action space $\mathcal{A}$, noting that adaptations for continuous control domains require topological extensions such as covering numbers. Generalization error is driven by the divergence between the true transition model $\mathcal{M}^*$ and the learned model $\mathcal{M}_\theta$. Because $\mathcal{M}_\theta$ relies on the MAS-SBC adjustment to decouple invariant physics from strategic intent, its accuracy is bottlenecked by the causal support in the demonstration dataset $\mathcal{D}_\sigma$, quantified by the interventional strength $\sigma \in (0, 1]$.

Lemma 1 (Transition Estimation Error)**.** *Let $\mathcal{D}_\sigma$ be a dataset of N transitions collected under the soft interventional policy mixture P^σ. The worst-case total variation (TV) distance ($Dist_{TV}$) between the learned causal transition $\mathcal{M}_\theta$ and the true mechanism $\mathcal{M}^*$ is bounded by:*

$$\max_{s,\mathbf{a}} Dist_{TV}(\mathcal{M}_\theta(\cdot \mid s, \mathbf{a}) \parallel \mathcal{M}^*(\cdot \mid s, \mathbf{a})) \leq \frac{J}{\sqrt{N\sigma^n}} + \epsilon_{env} \tag{13}$$

where n is agent count, J is a constant dependent on the state-action space complexity, and ϵ_{env} represents irreducible environmental stochasticity and neural network approximation error.

Proof. The learning of $\mathcal{M}_\theta$ relies on the finite observational dataset $\mathcal{D}_\sigma$ generated by the mixture policy $P^\sigma(\mathbf{a} \mid s) = \prod_{i=1}^n ((1-\sigma)\pi^i_{exp}(a^i \mid s) + \sigma\zeta^i(a^i))$. We assume the independent stochastic process ζ^i establishes a uniform distribution over each individual agent's discrete action space $\mathcal{A}^i$. We first prove that the minimum probability of observing any specific joint action $\mathbf{a} \in \mathcal{A}$ in any visited state is strictly bounded from below. Because soft interventions are applied independently and simultaneously to all n agents, the probability of selecting an exploratory joint action scales as the product of their individual probabilities. Since $\zeta^i(a^i) = \frac{1}{|\mathcal{A}^i|}$, dropping the non-negative expert policy contribution yields:

$$\min_{\mathbf{a},s} P^\sigma(\mathbf{a} \mid s) \geq \prod_{i=1}^n \left(\sigma \frac{1}{|\mathcal{A}^i|}\right) = \frac{\sigma^n}{|\mathcal{A}|} \tag{14}$$

where $|\mathcal{A}| = \prod_{i=1}^n |\mathcal{A}^i|$ is the cardinality of the joint action space.

Let $N(s, \mathbf{a})$ represent the exact number of visits to a specific state and joint-action tuple. To ensure the dataset $\mathcal{D}_\sigma$ has a non-zero probability of covering physically reachable states, we require the underlying Markov Decision Process to be ergodic. An ergodic MDP guarantees that every state is visited infinitely often in the limit, establishing a strictly positive stationary state distribution $d(s)$. We define this lower bound as $d_{min} = \min_s d(s) > 0$. Under this ergodicity condition, the expected dataset support for any valid transition is bounded by (Kakade, 2003):

$$\mathbb{E}[N(s, \mathbf{a})] \geq N d_{min} \left(\frac{\sigma^n}{|\mathcal{A}|}\right) \tag{15}$$

Substituting an expectation directly into a non-linear probabilistic tail bound invalidates high-probability confidence guarantees. Therefore, we first apply a Chernoff bound to establish a high-probability lower bound on the empirical count. With probability at least $1 - \frac{\delta}{2}$, the empirical counts satisfy $N(s, \mathbf{a}) \geq \frac{1}{2}\mathbb{E}[N(s, \mathbf{a})]$.

Conditioning on this high-probability event, we apply Weissman's L_1 concentration inequality (Weissman et al., 2003). For a discrete probability distribution over an alphabet of size $|\mathcal{S}|$, the theoretical coefficient is exponential, $2^{|\mathcal{S}|}$, bounding the deviation probability by:

$$P(||\hat{P} - \mathcal{M}^*||_1 \geq \epsilon) \leq 2^{|\mathcal{S}|} \exp\left(-\frac{N(s, \mathbf{a})\epsilon^2}{2}\right) \tag{16}$$

Converting the L_1 norm to Total Variation distance ($\text{Dist}_{TV} = \frac{1}{2}||\cdot||_1$) and isolating the error threshold via a union bound over all $|\mathcal{S}||\mathcal{A}|$ state-action pairs with the remaining failure probability $\frac{\delta}{2}$, the cardinalities are correctly placed in the numerator:

$$\max_{s,\mathbf{a}} \text{Dist}_{TV}(\hat{P}(\cdot \mid s,\mathbf{a}) \parallel \mathcal{M}^*(\cdot \mid s,\mathbf{a})) \leq \sqrt{\frac{1}{2N(s,\mathbf{a})} \log\left(\frac{2.2^{|\mathcal{S}|}|\mathcal{S}||\mathcal{A}|}{\delta}\right)} \tag{17}$$

Substituting the expected support $\mathbb{E}[N(s,\mathbf{a})]$ (Equation 15) into the concentration bound yields the purely statistical error. To account for optimization and capacity limitations of the neural network $\mathcal{M}_\theta$, we use the triangle inequality:

$$\text{Dist}_{TV}(\mathcal{M}_\theta \parallel \mathcal{M}^*) \leq \underbrace{\text{Dist}_{TV}(\mathcal{M}_\theta \parallel \hat{P})}_{\substack{\text{Neural Architecture} \\ \text{\& Optimization Error}}} + \underbrace{\text{Dist}_{TV}(\hat{P} \parallel \mathcal{M}^*)}_{\substack{\text{Empirical Dataset vs.} \\ \text{True Model Distance}}} \tag{18}$$

Defining $\epsilon_{env} = \text{Dist}_{TV}(\mathcal{M}_\theta \parallel \hat{P})$ as the irreducible approximation error, and aggregating the dimensionality constraints and confidence bounds into a complexity constant $J = \sqrt{\frac{|\mathcal{A}|}{d_{min}} \log\left(\frac{2^{|\mathcal{S}|+1}|\mathcal{S}||\mathcal{A}|}{\delta}\right)}$, we recover the final bound:

$$\max_{s,\mathbf{a}} \text{Dist}_{TV}(\mathcal{M}_\theta(\cdot \mid s,\mathbf{a}) \parallel \mathcal{M}^*(\cdot \mid s,\mathbf{a})) \leq \frac{J}{\sqrt{N\sigma^n}} + \epsilon_{env} \tag{19}$$

□

Theorem 2 (ICWM Generalization Bound)**.** *Given a reward function bounded by R_{max} and discount factor $\gamma \in [0,1)$, the error in evaluating any arbitrary target policy π_{test} using the learned ICWM $\mathcal{M}_\theta$ compared to the true environment $\mathcal{M}^*$ is bounded by:*

$$\left|V_{\pi_{test}}^{\mathcal{M}_\theta} - V_{\pi_{test}}^{\mathcal{M}^*}\right| \leq \frac{2\gamma R_{max}}{(1-\gamma)^2}\left(\frac{J}{\sqrt{N\sigma^n}} + \epsilon_{env}\right) \tag{20}$$

Proof. We define the value function $V_\pi^{\mathcal{M}}(s)$ as the expected discounted cumulative reward starting from state s and following policy π under transition dynamics $\mathcal{M}$:

$$V_\pi^{\mathcal{M}}(s) = \mathbb{E}_{\pi,\mathcal{M}}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s\right] \tag{21}$$

We bound the difference in the expected discounted return of an arbitrary target policy π_{test} evaluated in the true environment $\mathcal{M}^*$ versus the learned causal world model $\mathcal{M}_\theta$. We invoke the Simulation Lemma (Kearns & Singh, 2002), which systematically bounds multi-step value function discrepancies using the maximum single-step transition divergence.

By applying the Performance Difference Lemma (Kakade & Langford, 2002; Kakade, 2003), the difference in value functions can be expanded as a telescoping sum over the state-action distribution induced by the policy:

$$V_{\pi_{test}}^{\mathcal{M}_\theta}(s) - V_{\pi_{test}}^{\mathcal{M}^*}(s) = \frac{1}{1-\gamma}\mathbb{E}_{(s,\mathbf{a})\sim d_{\pi_{test}}^{\mathcal{M}_\theta}}\left[\gamma \sum_{s'} \left(\mathcal{M}_\theta(s' \mid s,\mathbf{a}) - \mathcal{M}^*(s' \mid s,\mathbf{a})\right) V_{\pi_{test}}^{\mathcal{M}^*}(s')\right] \tag{22}$$

Because the maximum single-step reward is bounded by R_{max}, the maximum possible accumulated value of any state over an infinite horizon is the geometric sum $R_{max}\sum_{t=0}^{\infty}\gamma^t$. For $\gamma \in [0,1)$, this sum converges exactly to $\frac{R_{max}}{1-\gamma}$. We bound the inner summation using this maximum value constraint and the Total Variation distance (where $\sum_{s'}|P(s') - Q(s')| \cdot V(s') \leq 2\text{Dist}_{TV}(P \parallel Q)\max V$):

$$\left|V_{\pi_{test}}^{\mathcal{M}_\theta} - V_{\pi_{test}}^{\mathcal{M}^*}\right| \leq \frac{2\gamma R_{max}}{(1-\gamma)^2} \max_{s,\mathbf{a}} \text{Dist}_{TV}(\mathcal{M}_\theta(\cdot \mid s,\mathbf{a}) \parallel \mathcal{M}^*(\cdot \mid s,\mathbf{a})) \tag{23}$$

Substituting the bounded transition error established in Lemma 1 directly into this inequality yields:

$$\left|V_{\pi_{test}}^{\mathcal{M}_\theta} - V_{\pi_{test}}^{\mathcal{M}^*}\right| \leq \frac{2\gamma R_{max}}{(1-\gamma)^2}\left(\frac{J}{\sqrt{N\sigma^n}} + \epsilon_{env}\right) \tag{24}$$

This concludes the proof. □

H.1 Discussion of Proven Implications

The mathematical derivations in Theorem 2 establish three core theoretical implications regarding the identifiability of causal world models:

1. **Unidentifiability in Deterministic Regimes:** The presence of σ in the denominator of the bound dictates that as $\sigma \to 0$ (a purely deterministic expert), the estimation error mathematically approaches infinity. In this zero-entropy regime, expert actions are perfectly collinear with the unobserved latent intent U, rendering the true physical transition function f_s structurally unidentifiable due to the lack of counterfactual support.
2. **Multi-Agent Scalability Penalty:** The error bound scales inversely with $\sqrt{\sigma^n}$, formally penalizing sample complexity exponentially with respect to the number of agents n. This mathematically demonstrates why targeted rather than uniform independent interventions are required to maintain scalability in dense multi-agent configurations.
3. **Bounding the Planning Snowball Effect:** Multi-step planning algorithms compound single-step transition errors over time, represented by the $\frac{\gamma}{(1-\gamma)^2}$ multiplier. By ensuring $\sigma > 0$, the soft intervention guarantees sufficient causal positivity across the joint state-action space, successfully constraining the maximum single-step deviation Dist_{TV} and capping the compounding multi-step error.
4. **Diminishing Returns on Stochasticity:** The bound shrinks proportionally to the square root of the interventional strength ($\frac{1}{\sqrt{\sigma^n}}$) but is ultimately floored by ϵ_{env}. This indicates that while small amounts of initial policy variance provide massive reductions in generalization error by breaking causal confounding, adding excessive noise eventually yields diminishing returns against the irreducible approximation limits of the neural architecture.

These theoretical implications are directly validated by our empirical results in Section 5. Specifically, the observed system error rapidly decreases as σ is introduced but plateaus or degrades at extreme values (e.g., $\sigma \geq 0.75$), confirming the mathematical dynamics established in this bound.

I Theoretical Guarantees and Algorithmic Selection for Causal Discovery

In this section 4.4, we provide the theoretical justifications for our causal structure discovery methodology. First, we provide a formal proof demonstrating that the elimination of contemporaneous edges ($E_{contemp}$) is a principled operation that does not violate the completeness guarantees of the true physical dynamics. Second, we rigorously justify the selection of PCMCI+ over latent causal discovery algorithms such as LPCMCI (Reiser, 2022), grounding this decision in the architectural constraints of our Implicit Causal World Model (ICWM).

I.1 Completeness of the Lagged Causal Skeleton in PCMCI+

In Section 4.4, we define the unconfounded causal skeleton $\mathcal{G}_{skeleton}$ by strictly discarding the contemporaneous edge set $E_{contemp}$ generated by the PCMCI+ algorithm. A critical theoretical concern is whether the exclusion of $E_{contemp}$ inadvertently drops true physical dependencies, particularly if the algorithm's minimality principle prioritizes contemporaneous correlations over lagged physical mechanisms. We demonstrate that under the structural assumptions of our multi-agent framework, the lagged structural edges (E_{lag}) are preserved and independently satisfy the completeness guarantee for the invariant transition function f_s.

I.1.1 Theoretical Setup and the MCI Test

Let $\mathbf{V}$ be the set of observable endogenous variables evolving as a Vector Autoregressive (VAR) process. PCMCI+ identifies the causal graph through the Momentary Conditional Independence

(MCI) test. For a putative lagged edge $V^i_{t-\tau} \to V^j_t$ with $\tau \geq 1$, the MCI test evaluates:

$$V^i_{t-\tau} \perp\!\!\!\perp V^j_t \mid (\widehat{Pa}(V^j_t) \setminus \{V^i_{t-\tau}\}) \cup \widehat{Pa}(V^i_{t-\tau}) \tag{25}$$

where $\widehat{Pa}(\cdot)$ denotes the set of preliminary parents identified in the initial PC phase.

In our MAS formulation, the true Data Generating Process or SCM $\mathcal{M}$ is governed by two distinct mechanisms:

1. **Physics (f_s):** Governs invariant, lagged structural edges ($\tau \geq 1$).
2. **Strategic Intent (u_{intent}):** Acts as an unobserved contemporaneous confounder ($\tau = 0$), inducing spurious correlations between V^i_t and V^j_t.

I.1.2 Proof of Lagged Skeleton Completeness

Claim: If a true physical mechanism $V^i_{t-\tau} \to V^j_t$ exists, the presence of the unobserved contemporaneous confounder u_{intent} will not cause PCMCI+ to d-separate and drop the lagged edge, ensuring E_{lag} is structurally complete for f_s.

Proof. Suppose a true invariant physical edge $V^i_{t-\tau} \to V^j_t$ exists. For PCMCI+ to incorrectly drop this edge, it must find a separating set $\mathbf{S} \subset \mathbf{V}$ such that $V^i_{t-\tau} \perp\!\!\!\perp V^j_t \mid \mathbf{S}$.

Under the minimality principle of causal discovery, if a contemporaneous variable V^k_t perfectly mediated the relationship (i.e., $V^i_{t-\tau} \to V^k_t \to V^j_t$), conditioning on V^k_t would correctly d-separate the lagged cause from the effect. However, we must evaluate the structural paths created by the unobserved intent.

By Assumption 4.4, u_{intent} coordinates actions strictly at time t. This creates a spurious contemporaneous path: $V^i_t \leftarrow u_{intent} \to V^j_t$. Because u_{intent} is unobserved, PCMCI+ detects this correlation and instantiates a contemporaneous edge in $E_{contemp}$ (e.g., $V^i_t \leftrightarrow V^j_t$ or $V^i_t \to V^j_t$). We must determine if conditioning on this contemporaneous relationship d-separates the lagged physical edge. It does not, due to two structural axioms:

1. **Independence of Physics and Intent:** The physical transition $V^i_{t-\tau} \to V^j_t$ is an autonomous mechanism governed by f_s. The strategic intent u_{intent} at time t does not alter the physical laws defined at time $t - \tau$. Therefore, the conditional mutual information $I(V^i_{t-\tau}; V^j_t \mid \mathbf{S})$ remains strictly positive for any subset $\mathbf{S}$ containing contemporaneous variables.
2. **Collider Bias Avoidance:** If $V^i_{t-\tau}$ and u_{intent} both causally influence V^j_t, then V^j_t acts as a collider. Conditioning on V^j_t (or its contemporaneous siblings) does not block the path from $V^i_{t-\tau}$; rather, it opens spurious paths. The MCI test (Equation 25) is mathematically designed to condition on the parents of both the cause and effect, isolating the specific structural link and preventing collider bias.

Because no subset of contemporaneous variables generated by u_{intent} can d-separate the true lagged physical mechanism, the MCI test will return a significant dependency. The true physical edge is strictly preserved in E_{lag}. □

I.1.3 A Principled Heuristic Backed by Causal Decision Graphs

Because the true physical edges are mathematically guaranteed to remain in E_{lag} (Figure 15), the contemporaneous edges $E_{contemp}$ identified by PCMCI+ contain strictly spurious correlations driven by u_{intent} (as well as unresolvable Markov equivalence structures caused by the unobserved confounding).

We formalize the removal of $E_{contemp}$ not as arbitrary graph surgery, but as a principled heuristic backed by Causal Decision Graphs (CDGs). In the CDG framework applied to multi-agent systems, environmental mechanisms (f_s) are defined by temporally extended structural equations, whereas joint policy execution (driven by u_{intent}) induces contemporaneous dependencies exclusively between decision nodes at time t.

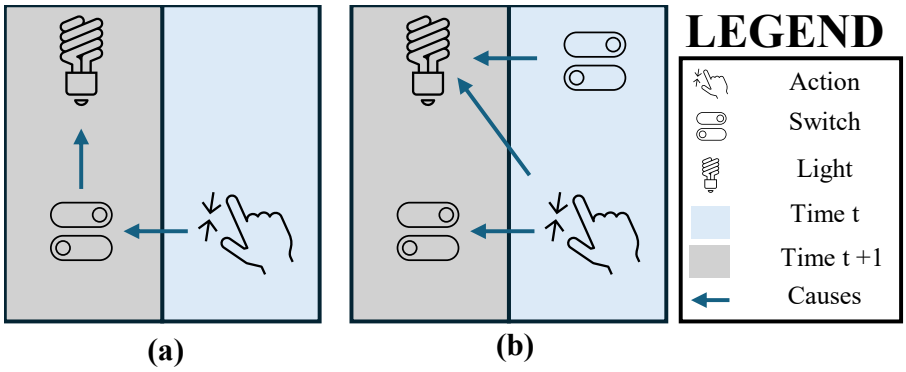


Figure 15: Resolving spurious contemporaneous correlation into temporally extended structural mechanisms. (a) A naive causal representation exhibiting a contemporaneous dependency ($E_{contemp}$). The switch appears to instantaneously cause the light to illuminate within the same time step t, a structural artifact typical of confounded multi-agent coordination. (b) The correct structural formulation derived via our CDG-backed heuristic. The invariant physical mechanism (f_s) operates with a strict temporal lag ($\tau \geq 1$), demonstrating that the state of the light at $t+1$ is fully determined by the parent state (the switch) at time t. This physical grounding illustrates why true environmental dynamics are strictly captured within the lagged edge set (E_{lag}), justifying the elimination of contemporaneous edges to recover the unconfounded causal skeleton.

Therefore, defining the final causal skeleton as $E_{skeleton} = \{(V_{t-k} \to V_t) \in E_{lag} \mid 1 \leq k \leq \tau\}$ is a domain-informed heuristic derived directly from the CDG of the multi-agent environment. This heuristic safely eliminates the unblocked backdoor paths associated with joint intent without violating the completeness guarantee of the VAR process underlying the physical environment.

I.2 Algorithmic Selection: PCMCI+ vs. Latent Discovery (LPCMCI)

While the standard PCMCI+ algorithm operates under the assumption of causal sufficiency, the multi-agent system explicitly contains an unobserved confounder. In standard causal discovery literature, the Latent PCMCI (LPCMCI) algorithm is typically deployed to handle such hidden variables. However, we actively choose PCMCI+ combined with strict structural assumptions over LPCMCI. This decision is driven by the strict architectural requirements of our neural world model and the necessity of forward reasoning using known physical priors.

I.2.1 Architectural Infeasibility of Partial Ancestral Graphs (PAGs)

The Sequential Backdoor Condition (SBC) requires a deterministic, fully resolved directed graph to execute causal adjustments. Our Implicit Causal World Model (ICWM) instantiates this graph as a causal tensor mask $\mathbf{C}$ to restrict information flow in the neural network.

LPCMCI outputs a Partial Ancestral Graph (PAG), which explicitly models latent confounding using bidirected edges ($\leftrightarrow$) and marks Markov equivalence ambiguities with circle edges (e.g., $o \to$). Translating an ambiguous PAG (Richardson & Spirtes, 2002) into a deterministic neural architecture is mathematically infeasible without relying on arbitrary, heuristic tie-breaking for unresolved edges. By utilizing PCMCI+ and filtering out $E_{contemp}$, we obtain a strictly directed graph (E_{lag}) that maps flawlessly into our causal tensor mask.

K Environments

K.1 MAS Environment Visualization

The Following figure 16 represents the three important domains chosen for evaluation.

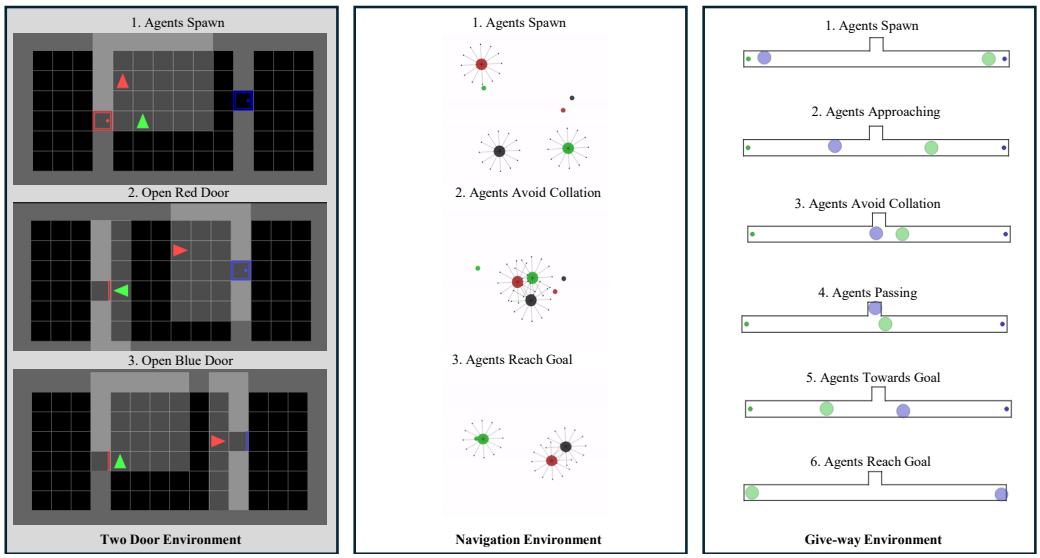


Figure 16: Environments used for empirical experiments.

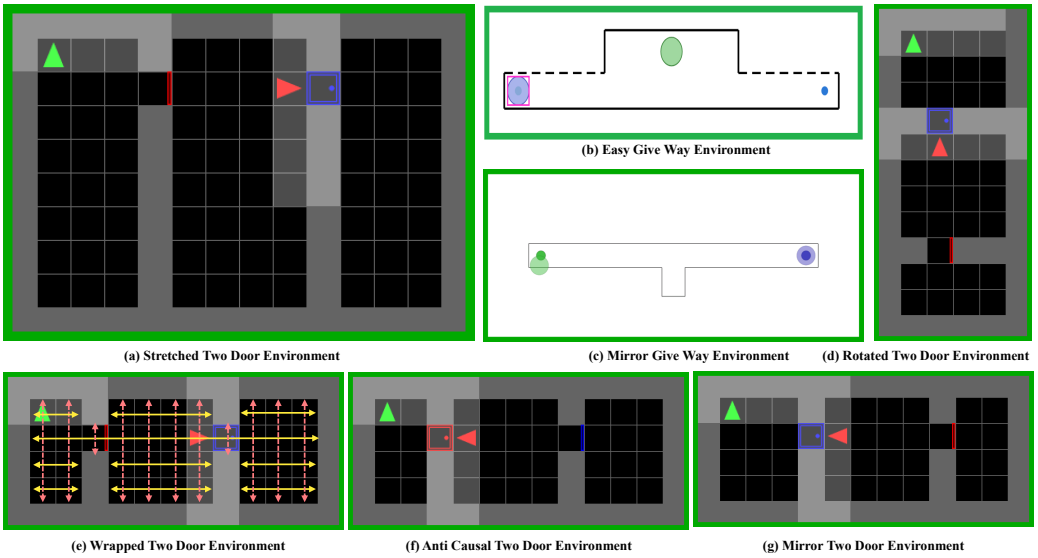


Figure 17: Example states from the seven shifted environments. Each keeps the dynamics of its default environment but changes the layout (or, for the anti-causal variant, the cue–dynamics mapping), so all seven are out of distribution for a model trained only on the defaults. Panels: (a) Stretched Two Door, (b) Easy Give Way, (c) Mirror Give Way, (d) Rotated Two Door, (e) Wrapped Two Door, (f) Anti-Causal Two Door, (g) Mirror Two Door.

K.2 Shifted environments

To test whether a world model has learned the true dynamics rather than the specific layout it was trained on, we build a set of *shifted* environments. Each one keeps the same agents, the same physics, and the same task, but changes the geometry of the arena. A model that understands how the agents move should transfer to these layouts, while a model that has memorised the training layout should break. We train only on the default environments and evaluate on the shifted ones, so every shifted layout is strictly out of distribution.

Figure 17 shows one example state from each shifted environment. The green and red triangles are the two agents in the grid (two-door) environments, the green and blue circles are the two agents in the give-way environments, the blue box marks the goal door, and the red bar marks the closed door.

How each layout differs from its default. The differences are summarised below and can be checked against Fig. 17. Panels are listed in the order they appear in the figure.

- **Stretched Two Door** (Fig. 17(a)). The two-door room is scaled wider, so the two corridors and the central room are longer than in the default. The doors, the walls, and the agents are unchanged; only the distances the agents must travel grow. This tests whether the model handles the same behaviour played out over a larger space.
- **Easy Give Way** (Fig. 17(b)). The give-way corridor is widened and the passing bay is enlarged, so the two agents have more room to resolve the give-way manoeuvre than in the default. The task and the physics are unchanged; only the spatial margins grow. This tests whether the model preserves the give-way dynamics when the coordination is made geometrically easier.
- **Mirror Give Way** (Fig. 17(c)). The give-way notch is flipped from above the corridor to below it. The corridor, the two goals, the agent radius, and the action range are all unchanged, so the only difference is that an agent now steps down into the bay instead of up. This tests whether the model has tied the give-way behaviour to a specific side of the corridor.
- **Rotated Two Door** (Fig. 17(d)). The two-door room is rotated counter-clockwise by $90°$. The wide, short default becomes tall and narrow, and the corridors that ran left and right now run top and bottom. Positions and motions that were horizontal in the default are vertical here, so the model must apply the same dynamics along a different axis.
- **Wrapped Two Door** (Fig. 17(e)). The two-door arena is made toroidal: an agent that exits one edge re-enters from the opposite edge (indicated by the wrap-around arrows), so the room's boundaries connect rather than block. The room, doors, and agents are otherwise those of the default. This tests whether the model has learned local motion rules that still hold when the global connectivity of the arena changes.
- **Anti-Causal Two Door** (Fig. 17(f)). The layout matches the default two-door room, but the cue-to-dynamics mapping is inverted: the relationship between the observed signal and the resulting transition is reversed relative to training. This is a dynamics shift rather than a geometric one, and it tests whether the model has learned the true causal mechanism rather than a cue that only correlates with it in the default distribution.
- **Mirror Two Door** (Fig. 17(g)). The whole layout is reflected left to right. The room, the two doors, and the corridors move to the opposite side of the arena, so a fixed agent that sat on the right in the default now sits on the left. The geometry is identical to the default up to a mirror, which isolates whether the model has learned a direction-specific rule rather than the true symmetric dynamics.
- **Delayed Two-Door** (Fig. 18): This variation introduces a higher-order Markov dynamics shift to the left red door. When toggled by an agent, the door's state now updates after a three-step unobserved delay rather than the standard single-step response. The environment is therefore only Markov-sufficient when modeling lags of three or greater. This evaluates the method's capacity to generalize to higher-order temporal relations.

L Neural Architecture Design and Baselines

To evaluate causal robustness under confounding, our baselines map to established architectural paradigms: Causal-VAE and R-ICWM capture the generative latent coordination of Multi-Agent Dreamer Lobos-Tsunekawa et al. (2022), the unconstrained Associative World Model serves as the dense probabilistic baseline analogous to MAMBPO Zhang et al. (2022), and the PC-derived FOCUS baseline represents the sparse structural abstraction of Causal Dynamics Learning Wang et al. (2022).

L.1 Summary of Model Properties

Table 6 categorizes the world models by their discovery mechanism, interventional strength (σ), and masking strategy.

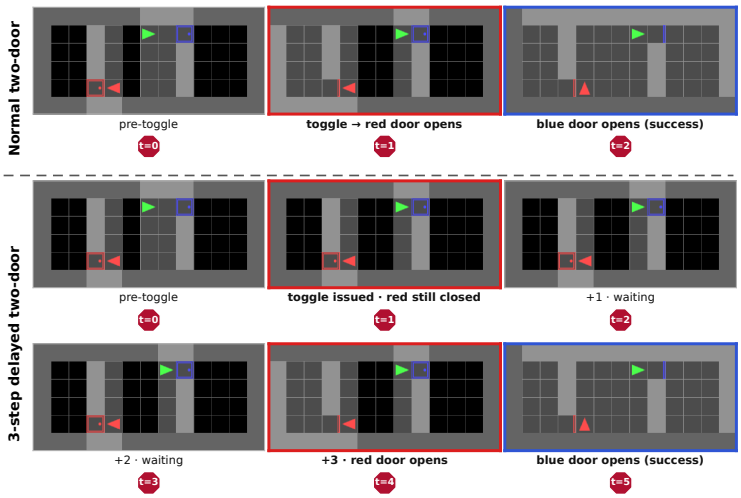


Figure 18: Optimal-policy rollouts in the normal vs. 3-step delayed two-door environments. Toggling the red door opens it instantly in the normal env, but only after a 3-step lag in the delayed one, before the blue agent succeeds. Frame timesteps are marked below each panel.

Table 6: Baseline configurations for causal transition modeling.

Abbreviation	Discovery Method	Data Entropy (σ)	Masking Operator
AWM_{σ}	N/A	σ	Dense (None)
$\text{FOCUS}^{\odot}_{\sigma=0}$	PC Algorithm	0.0	Hadamard ($\odot$)
$\text{BC-WM}^{\odot}_{\sigma=0}$	PCMCI+	0.0	Hadamard ($\odot$)
$\text{ICWM}^{\odot}_{\sigma}$	PCMCI+	σ	Hadamard ($\odot$)
$\text{R-ICWM}^{\odot}_{\sigma}$	PCMCI+	σ	Hadamard ($\odot$)
$\text{C-VAE}^{\text{inj}}_{\sigma}$	PCMCI+	σ	Latent Injection

L.2 Causal Structural Masking and AMLS

The fundamental structural constraint in our architecture is the *Adjacency Matrix with Lagged Structure* (AMLS) tensor, denoted as $\mathcal{C}$. For a system with history $\tau = 1$, $\mathcal{C}$ represents the dependencies between variables at time t and state transitions at $t + 1$.

To bridge discrete discovery with continuous neural optimization in our proposed framework, we implement *Statistical Soft Masking* via a differentiable gating mechanism (Jang et al., 2017; Maddison et al., 2017; Zheng et al., 2018). This modulates input features through a Log-Significance Transform (ψ_{ij}) relative to a significance threshold α:

$$\psi_{ij} = \log\left(\frac{\alpha}{p_{ij} + \epsilon}\right) \tag{26}$$

The functional mask weight Γ_{ij} is then computed using a sigmoid transformation scaled by a unique temperature parameter κ:

$$\Gamma_{ij} = \sigma\left(\frac{\psi_{ij}}{\kappa}\right) = \frac{1}{1 + \exp(-\psi_{ij}/\kappa)} \tag{27}$$

Specifically, $\psi_{ij} > 0$ (where $p_{ij} < \alpha$) indicates significant causal edges, driving $\Gamma_{ij} \to 1$ to permit information flow. Conversely, $\psi_{ij} < 0$ indicates non-causal edges, driving $\Gamma_{ij} \to 0$ to enforce *Causal Shielding*. The temperature κ controls the "hardness" of the gate; as $\kappa \to 0$, the mask converges to a discrete binary shutter.

L.3 Baseline Architectures

All MLP-based transition functions share a consistent depth of 3 hidden layers and a width of 32 neurons per agent in the framework.

1. **Associative World Model (AWM_σ):** A standard 3-layer MLP with dense connectivity. Every input feature in the state-action vector can influence every output state prediction.
$$\hat{s}_{t+1} = \text{MLP}_3([s_t, \mathbf{a}_t]; \theta) \tag{28}$$

2. **The FOCUS Baseline ($\text{FOCUS}^{\odot}_{\sigma=0}$):** Utilizes (Zhu et al., 2022) a binary causal tensor $\mathcal{C}$ derived from the PC algorithm. Masking is applied strictly at the input stage via the Hadamard product.
$$\hat{s}_{t+1} = \text{MLP}_3([s_t, \mathbf{a}_t] \odot \mathcal{C}; \theta) \tag{29}$$

3. **Behavioral Cloning World Model ($\text{BC-WM}^{\odot}_{\sigma=0}$):** Employs the PCMCI+ algorithm to generate a binary tensor $\mathcal{C}$ but is trained on non-interventional expert demonstrations ($\sigma = 0$).
$$\hat{s}_{t+1} = \text{MLP}_3(\mathbf{v}_t \odot \mathcal{C}; \theta) \tag{30}$$

4. **Implicit Causal World Model ($\text{ICWM}^{\odot}_{\sigma}$):** The proposed framework. Integrates PCMCI+ structural masking via the soft gate Γ with interventional data entropy σ to recover invariant environmental physics.
$$\hat{s}_{t+1} = \text{MLP}_3(\mathbf{v}_t \odot \Gamma; \theta) \tag{31}$$

5. **Recurrent Implicit Causal World Model (R-ICWM_σ):** Extends the proposed framework to natively handle temporal dependencies ($\tau \geq 1$) and historical conditioning (H_k) required by the MAS-SBC. By utilizing a Long Short-Term Memory (LSTM) architecture, the model unrolls the temporal sequence. Crucially, the statistical soft mask Γ is applied strictly to the raw historical inputs $\mathbf{v}_k$, ensuring the hidden memory states (h_k, c_k) remain unmasked to fully preserve recurrent dynamics.

$$\text{Masked Input:} \quad \tilde{\mathbf{v}}_k = \mathbf{v}_k \odot \Gamma \quad \text{for } k \in \{t - \tau, \ldots, t\} \tag{32}$$
$$\text{Recurrent Update:} \quad h_{k+1}, c_{k+1} = \text{LSTMCell}_{32}(\tilde{\mathbf{v}}_k, h_k, c_k; \theta_{\text{lstm}}) \tag{33}$$
$$\text{State Prediction:} \quad \hat{s}_{t+1} = \text{MLP}_2(h_{t+1}; \theta_{\text{head}}) \tag{34}$$

6. **Injected Causal VAE (C-VAE$_\sigma^{\text{inj}}$):** Evaluates whether latent structural disentanglement can replace explicit input-level masking. To maintain equivalent representational capacity with the baseline architectures (Yang et al., 2023), both the inference (encoder) and generative (decoder) networks are parameterized as 3-layer Multi-Layer Perceptrons (MLPs). The discovered causal tensor $\mathcal{C}$ is injected strictly to constrain the forward transition dynamics in the latent space.

$$\text{Encoder (3-layer):} \quad h_t^{(e)} = \text{MLP}_\phi^{(3)}(\mathbf{v}_t)$$
$$z_t \sim q_\phi(z \mid \mathbf{v}_t) = \mathcal{N}\big(\mu_\phi(h_t^{(e)}), \Sigma_\phi(h_t^{(e)})\big) \tag{35}$$
$$\text{Causal Transition:} \quad z_{t+1} = \text{ReLU}\big(\mathbf{W}_z(z_t \odot \mathcal{C}) + \mathbf{W}_a \mathbf{a}_t + b_z\big) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{36}$$
$$\text{Decoder (3-layer):} \quad h_{t+1}^{(d)} = \text{MLP}_\theta^{(3)}(z_{t+1})$$
$$\hat{s}_{t+1} \sim p_\theta(s \mid h_{t+1}^{(d)}) \tag{37}$$

L.4 Optimization and Loss Functions

Training configurations are partitioned into deterministic models (MLP and LSTM variants) and generative latent models (C-VAE). All baselines use the Adam optimizer with a standardized learning rate of 1×10^{-3}. Deterministic architectures minimize Mean Squared Error (MSE) for direct state prediction, whereas the C-VAE maximizes the Evidence Lower Bound (ELBO), combining an MSE reconstruction term with a Kullback-Leibler (KL) divergence penalty.

Table 7: Optimization for MLP & LSTM Models

Hyperparameter	Value
Target Models	AWM, FOCUS, BC-WM, ICWM, R-ICWM
Loss Function	MSE
Optimizer	Adam
Learning Rate (η)	1×10^{-3}
Epochs	50

Table 8: Optimization for Causal VAE

Hyperparameter	Value
Target Models	C-VAE$_\sigma^{\text{inj}}$
Loss Function	ELBO (MSE + KL Div.)
Optimizer	Adam
Learning Rate (η)	1×10^{-3}
Epochs	100

P Causal Discovery Setup and Results

P.1 Causal Discovery Experimental Parameters

Table 9 summarizes the configuration parameters used for the PC and PCMCI causal discovery algorithms across all experimental environments and observation modes.

Parameter	Two-Door		Give-way		Navigation	
	Full	Part	Full	Noisy	Full	Part
[Causal Graph Parameters]						
discovery_algorithm	PC / PCMCI	PC / PCMCI	PC / PCMCI	PC / PCMCI	PC / PCMCI	PC / PCMCI
CI_test_name	Gsquared	Gsquared	robust_parcorr	robust_parcorr	robust_parcorr	robust_parcorr
CI_test_significance	analytic	analytic	analytic	analytic	analytic	analytic
alpha_level	0.005	0.005	0.005	0.005	0.005	0.005
pc_alpha	0.1	0.1	0.1	0.1	0.1	0.1
tau_min	1	1	1	1	1	1
tau_max	1	1	1	1	1	1
max_cond_parent_count	3	3	3	3	3	3
missing_value	-99	-99	-99	-99	-99	-99
missing_padding_length	5	5	5	5	5	5
discrete_reward	true	true	false	false	false	false

Table 9: Causal Discovery Configuration Parameters for PC and PCMCI Algorithms

P.2 Evaluation Results: on three primary enviorments

Figure 19 presents the evaluation results for the recovered causal graphs, illustrating how discovery algorithm performance varies across different interventional strengths. A crucial misinterpretation would be to view Figure 19 as evidence that the PC algorithm is superior to the PCMCI algorithm; this appearance is an artifact of the experimental design, where the number of transitions is fixed and normalized to isolate the effects of information gain relative to interventional strength.

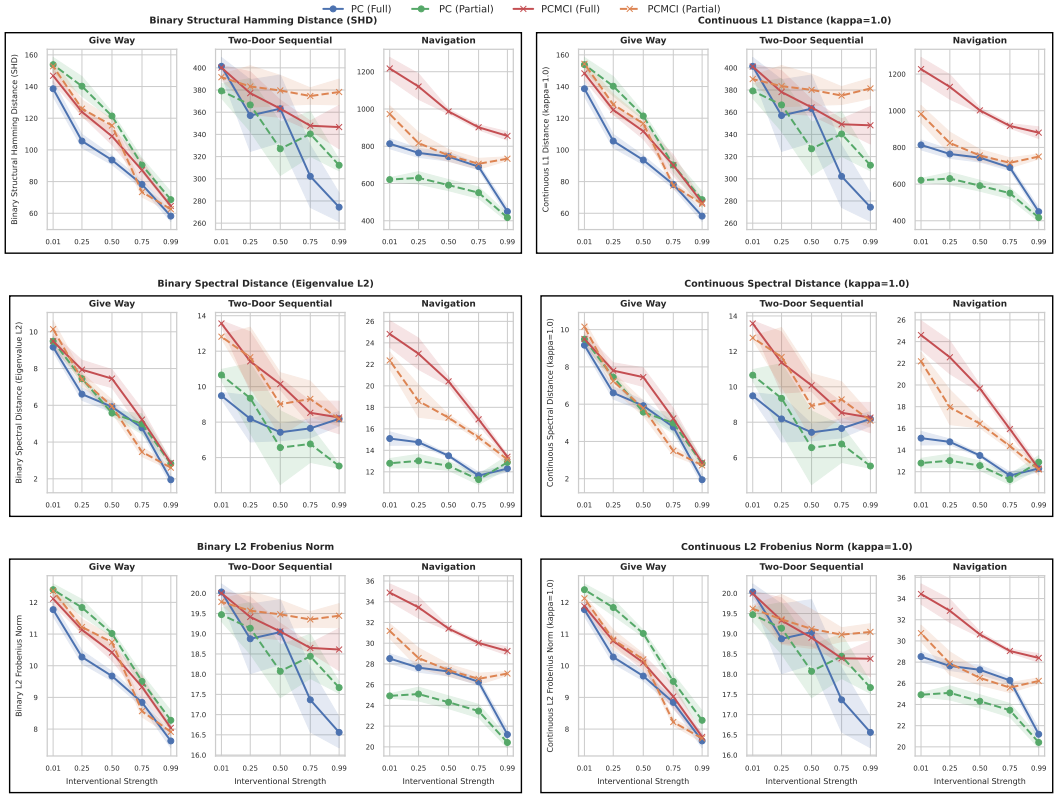


Figure 19: **Recovered causal-graph error falls as interventional strength grows, across every environment, metric, and observation mode.** Each of the six bordered panels reports one graph-error metric against the interventional strength σ of the training data, for the three coordination domains (give-way, two-door, navigation). Rows: Structural Hamming Distance (SHD), spectral distance (eigenvalue L_2), and L_2 Frobenius norm; columns: the binary adjacency and the continuous soft-mask ($\kappa = 1.0$) formulations of each. Curves compare the PC and PCMCI discovery algorithms under full and partial/noisy observation; shaded bands are ± 1 standard deviation over five seeds. Graph error decreases monotonically with σ in nearly every cell: stronger soft interventions turn observational demonstrations into an identifiable causal structure, and the effect is consistent whether error is measured locally (SHD, Frobenius) or spectrally, and whether the graph is read as a binary skeleton or a continuous mask.

P.3 Evaluation Results: High sample complexity experiments

The results in Figure 20 show that increasing sample complexity significantly improves recovery stability. The thin uncertainty bands in Figure 20 confirm that the algorithms produce consistent results across different seeds.

Under partial observation settings in Figure 20, the models successfully recover the causal structure. This is possible because both environments share the same state space dimension. According to causal discovery theory, this recovery is effective as long as the parent-child relationships remain invariant across environment steps. By relying on these fixed dependencies, the framework generalizes the causal skeleton across different tasks despite incomplete state observability.

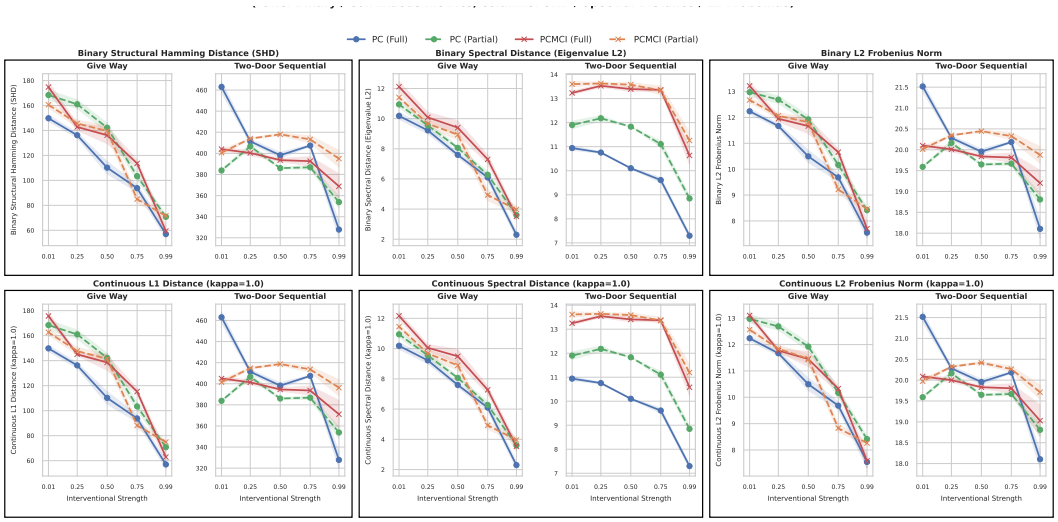


Figure 20: **At high sample complexity, the intervention–recovery trend holds and the discovered graphs sharpen further.** Recovered-graph error versus interventional strength σ for the give-way and two-door domains trained on the high-sample datasets (give-way ≈30k–120k transitions, two-door ≈84k–2M). The six bordered panels span two metric families (top row: binary adjacency; bottom row: continuous soft-mask, $\kappa = 1.0$) across three metrics per row (Structural Hamming Distance, spectral distance, and L_2 Frobenius norm). Curves compare the PC and PCMCI discovery algorithms under full and partial/noisy observation; shaded bands are ± 1 standard deviation over five seeds, and are visibly tighter than in the standard-sample regime. The downward trend with σ persists in every cell, and partial-observation recovery closely tracks full observation because both domains share the same state-space dimension—so the parent–child dependencies the algorithm must recover are invariant across the two modes.

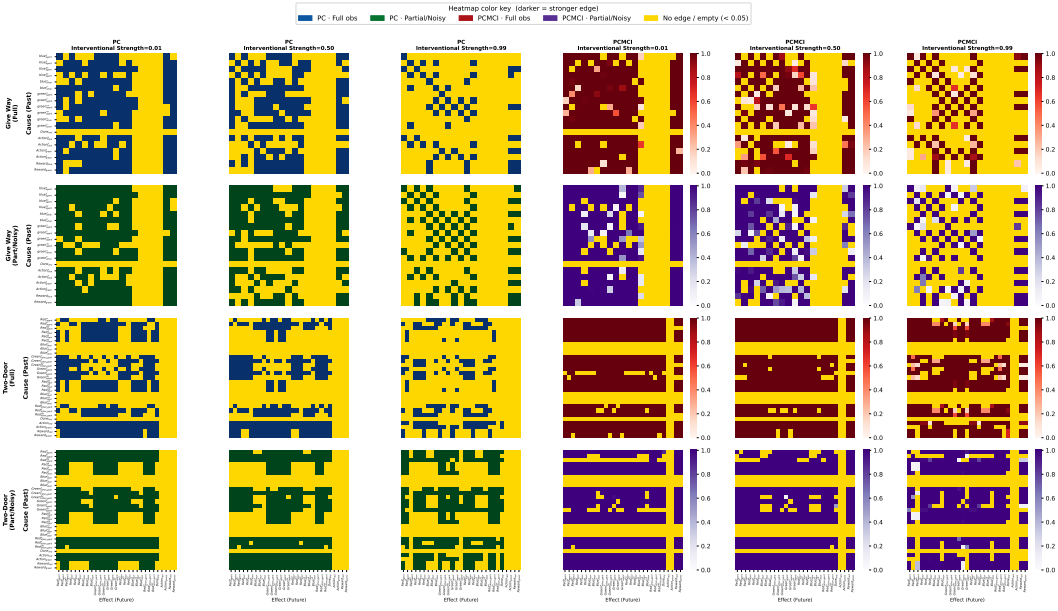
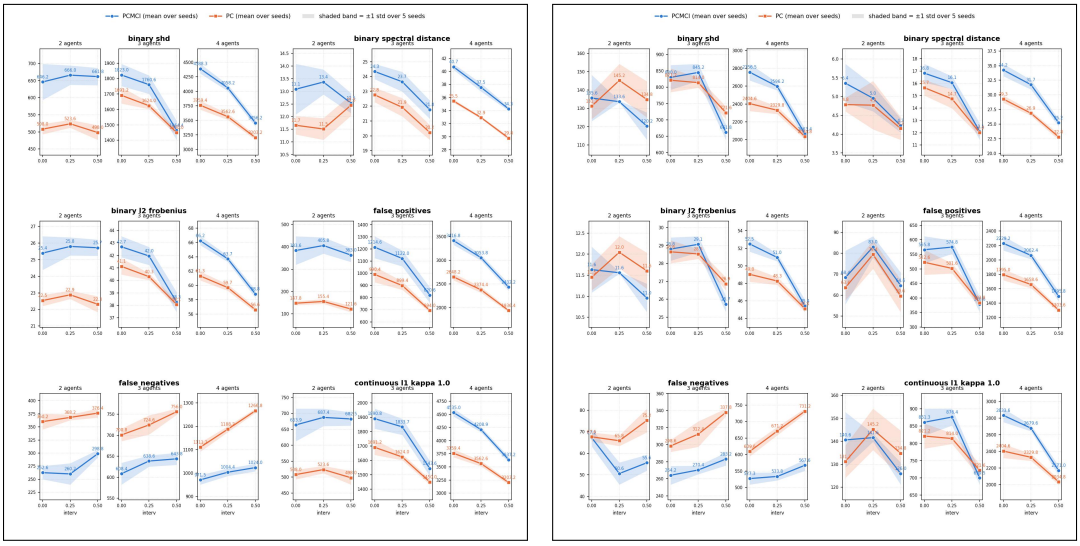


Figure 21: This figure displays recovered causal graph structures for Give-Way and Two-Door environments under full and partial/noisy observation settings. The heatmaps demonstrate algorithm performance across varying intervention strengths (0.01, 0.50, 0.99), with the color key indicating edge strength and yellow denoting empty or non-significant entries (¡ 0.05).

The recovered causal graphs in Figure 21 demonstrate that stronger interventional strength consistently leads to a graph structure that aligns more closely with the underlying environment physics. By systematically perturbing the system, we reveal causal mechanisms that remain obscured under purely observational data.

We adopt distinct interpretations for the two discovery approaches. The PC algorithm outputs are best suited for testing discrete edge hypotheses, providing a clear, binary structure of causal dependencies. In contrast, PCMCI coupled with soft masking offers a fuzzy perspective of the causal graph. This probabilistic representation is advantageous, as it allows the subsequent parameter learning phase to absorb and mitigate minor masking errors, resulting in a more robust and flexible world model. By assigning continuous weights to potential edges, the model avoids the brittleness of hard-thresholding; if a specific edge is misidentified, the model retains the ability to learn away this inaccuracy during the training of the dynamics model rather than being forced to rely on a fundamentally incorrect structural assumption.

P.4 Evaluation Results: two, three, four agent navigation



(a) Has LiDAR dimensions

(b) No LiDAR dimensions

Figure 22: Performance metrics for PCMCI and PC algorithm's recovered graph in multi-agent navigation. Metrics include Binary SHD, Spectral Distance, Frobenius Norm, and error rates testing for intervention strength.

From figure 22(a) the graph recovery metric error does not drop to zero because the environment is deterministic. Using 12 LiDAR scans per agent violates the faithfulness assumption, which requires independent noise. The tests mistake physical constraints for causal dependencies. A simpler way to describe the issue would be a near proximity agent show up across multiple LIDAR scan angles for the ego agent we are interested in. Interventions cannot fix these errors from finite data because the underlying physics remains deterministic (At all times, including following an intervention, a single LiDAR scan dimension can be represented as a linear combination of other LiDAR scan dimensions and system variables).

Despite these challenges, our modeling performs well on the remaining features. This success calls us to represent the state space of this task in a more rank-sufficient way in the future. By compressing redundant LiDAR information into a lower-dimensional representation, we can mitigate the impact of deterministic correlations on the causal discovery process (figure 22(b)).

For Future work, We formalize this rank-sufficient state space s_t at time step t as:

$$s_t = \{p_t, \phi_t, \tilde{\mathbf{L}}_t, \mathbf{O}_t\} \tag{38}$$

where:

- $p_t = (x_t, y_t)$: Agent spatial coordinates.

- ϕ_t: Agent orientation.
- $\tilde{\mathbf{L}}_t$: The compressed, rank-sufficient representation of the raw LiDAR vector.
- $\mathbf{O}_t$: Relative state information of other agents.

Q World Model Setup and Results

Q.1 Evaluation Results: on three primary environments



Figure 23: **World-model one-step prediction error across the canonical three-environment sweep.** Six architectures (AWM, C-VAE, FOCUS, BC-WM, ICWM, R-ICWM) trained on `give_way`, `navigation`, and `two_door` under full and partial observability, at intervention strengths $\sigma \in \{0.01, 0.25, 0.5, 0.75, 0.99\}$. Rows: evaluation policy (adversarial, biased-optimal, random). Columns: metric (raw MSE, relative to $\sigma = 0.5$, variance-normalized). Each panel is a 3×2 grid of half-violins over environment and observability. FOCUS and BC-WM are trained only at $\sigma \approx 0$ and shown as a flat reference. Error drops one to two orders of magnitude with increasing σ for `give_way` and `navigation`, where ICWM and R-ICWM attain the lowest error; `two_door` shows no such trend, consistent with its sparsity residing in the input to output transition matrix rather than input marginals.

Figure 23 reports one-step prediction error for six world-model architectures: AWM, C-VAE, FOCUS, BC-WM, ICWM, and R-ICWM. Each model is trained on three environments (`give_way`, `navigation`, `two_door`) under full and partial observability, at five training intervention

strengths $\sigma \in \{0.01, 0.25, 0.5, 0.75, 0.99\}$, where σ is the probability that an agent takes a random (intervention in our setting) rather than optimal action. Each panel shows a 3×2 grid of half-violin distributions, one for each environment and observability combination.

The three rows of panels correspond to the policy used to generate evaluation trajectories: adversarial $\mathcal{D}^{\ddagger\ddagger}$ (top), biased-optimal $\mathcal{D}^{\ddagger}_{\sigma=0.5}$ matched to each model's training σ (middle), and random exploration $\mathcal{D}^{\ddagger}_{\sigma=0.9}$ (bottom). The three columns correspond to the error metric: raw MSE (left), MSE relative to the same model's error at $\sigma = 0.5$ (center), and variance-normalized relative error (right).

Reading the trends. Error falls by one to two orders of magnitude as σ increases toward 0.99 for `give_way` and `navigation`. This shows that training on more strongly intervened trajectories produces better one-step dynamics models in these two environments. `two_door` does not follow this pattern. Its error stays within roughly one order of magnitude across the full σ range, and the ranking of architectures is less stable. We attribute this to `two_door`'s causal sparsity living in the input by output transition matrix rather than in individual input variables, so the soft input gate used by ICWM and R-ICWM captures little of the useful structure in this environment.

ICWM and R-ICWM achieve the lowest error at high σ under the adversarial and biased-optimal evaluation policies. This is clearest in `give_way` and `navigation`, where they fall roughly one order of magnitude below AWM by $\sigma = 0.99$. FOCUS and BC-WM are trained only near $\sigma \approx 0$, so they behave as flat, σ-invariant reference models. They are competitive only at low σ, and their relative error grows by one to three orders of magnitude as the evaluation policy moves further from the deterministic-optimal regime. C-VAE has the highest error and highest variance across nearly every panel, particularly in `navigation`. This suggests its stochastic latent variable does not improve one-step prediction accuracy in this setting.

The center and right columns confirm the same pattern under two different normalizations. Relative error against each model's own $\sigma = 0.5$ checkpoint isolates how error scales with distributional shift. Variance normalization further controls for the fact that next-state variance itself grows with σ. Both normalizations preserve the same architecture ranking and the same `two_door` exception, so the effect is not an artifact of the normalization choice. Under random exploration (bottom row), differences between architectures shrink at low to moderate σ for `give_way` and `navigation`, since random rollouts are less sensitive to the fine-grained transition structure that graph-informed gating exploits. The ICWM and R-ICWM advantage reappears at $\sigma \geq 0.75$, as random exploration increasingly resembles the high-intervention training regime.

Q.2 Evaluation Results: High sample complexity experiments

Figure 24 repeats the evaluation of Appendix Q.1 on high-sample training datasets. The `give_way` datasets grow from roughly 30k to roughly 120k transitions, and the `two_door` datasets grow from roughly 84k to roughly 2M transitions. `navigation` is not included in this sweep. All six architectures (AWM, C-VAE, FOCUS, BC-WM, ICWM, R-ICWM) are retrained from scratch at the same five intervention strengths $\sigma \in \{0.01, 0.25, 0.5, 0.75, 0.99\}$, under full and partial observability. The figure layout matches Appendix Q.1. Rows show the evaluation rollout policy: adversarial (top), biased optimal (middle), and random exploration (bottom). Columns show the error metric: raw MSE (left), MSE relative to the same model at $\sigma = 0.5$ (center), and variance-normalized relative error (right). Each panel now contains a 2×2 grid of half-violins over the four (environment, observability) groups.

The main purpose of this sweep is to test whether the `two_door` exception observed in Appendix Q.1 is a sample-complexity artifact. It largely is. With roughly 24 times more training data, `two_door` acquires the monotone trend it previously lacked. Under the adversarial evaluation, its MSE falls by about one order of magnitude from $\sigma = 0.01$ to $\sigma = 0.99$ in both observability settings, the violins tighten sharply, and the architecture ranking stabilizes, with ICWM, R-ICWM, and AWM forming the low-error group at high σ. The wide, unstable distributions seen at normal sample sizes are gone. This indicates that the earlier flat profile reflected undertrained models rather than a property of the environment, although the gap between graph-informed models and AWM remains small in `two_door`, consistent with its causal sparsity residing in the input to output transition matrix rather than in the input marginals that the soft gate can express.

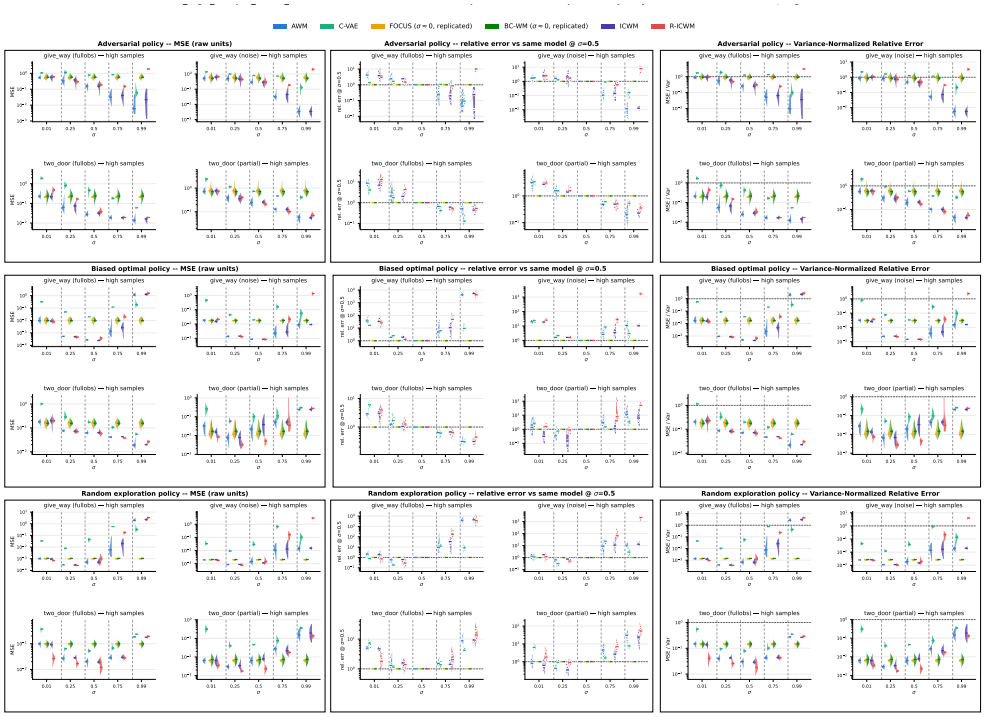


Figure 24: **World-model one-step prediction error in the high-sample regime.** Six architectures trained on enlarged datasets for `give_way` (roughly 120k transitions) and `two_door` (roughly 2M transitions) under full and partial observability, at intervention strengths $\sigma \in \{0.01, 0.25, 0.5, 0.75, 0.99\}$. Layout matches Figure 23. Rows: evaluation policy (adversarial, biased optimal, random). Columns: metric (raw MSE, relative to $\sigma = 0.5$, variance-normalized). FOCUS and BC-WM are trained only at $\sigma \approx 0$ and shown as a flat reference. With more data, `two_door` gains the monotone improvement with σ that it lacked at normal sample sizes and its distributions tighten sharply, while `give_way` reproduces the earlier trends, with ICWM lowest at high σ under adversarial evaluation.

The `give_way` results reproduce the normal-sample trends. Under the adversarial evaluation, error falls by two to three orders of magnitude as σ increases, and ICWM reaches the lowest error of any model at $\sigma = 0.99$, roughly one order of magnitude below AWM. FOCUS and BC-WM remain flat σ-invariant references, competitive only at low σ, and their relative error grows by up to three orders of magnitude at high σ. C-VAE remains the weakest and most variable model overall. More data therefore does not close the gap created by training-policy coverage: what the trajectories contain matters more than how many there are.

The middle and bottom rows show a complementary policy-mismatch effect in `give_way`. When evaluation trajectories are generated by near-optimal policies, error is lowest for models trained at moderate intervention strength (σ between 0.25 and 0.5) and rises steeply at $\sigma = 0.99$, where the training data contains almost no optimal behavior. This mirror image of the adversarial trend confirms that each model is most accurate on the region of state space its training policy visited, and that no single training σ dominates across all evaluation policies. The relative-error and variance-normalized columns preserve every one of these patterns, so none of them is an artifact of the error normalization.

Q.3 Evaluation Results: two, three, four agent navigation

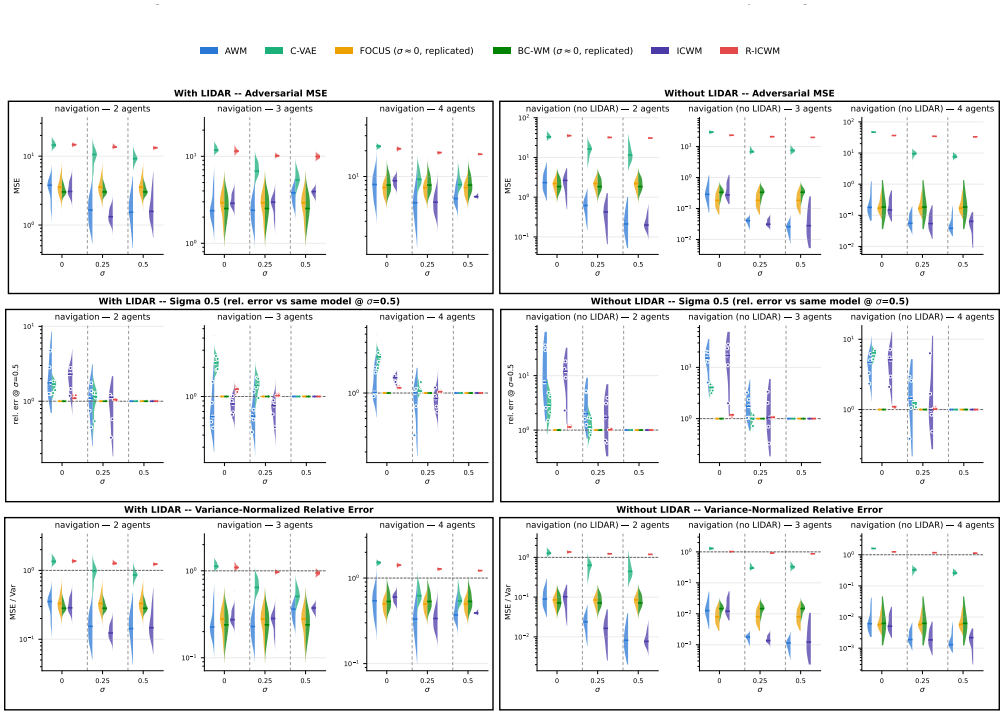


Figure 25: **Multi-agent world-model error with and without LIDAR.** Six architectures trained on `navigation` with 2, 3, and 4 agents at intervention strengths $\sigma \in \{0, 0.25, 0.5\}$, evaluated on adversarial rollouts. Each metric (raw MSE, error relative to the same model at $\sigma = 0.5$, variance-normalized error) is shown for the full observation including LIDAR and for a no-LIDAR variant with those dimensions removed from the state and the causal graph. FOCUS and BC-WM are trained only at $\sigma \approx 0$ and shown as a flat reference. With LIDAR, variance-normalized error stays near 0.1 to 1 and architectures are hard to separate. Without LIDAR it falls to 10^{-2} to 10^{-3}, the gain from interventional data grows to roughly tenfold, and ICWM matches or beats AWM at every agent count.

Figure 25 evaluates the six world-model architectures on the multi-agent `navigation` environment with 2, 3, and 4 agents, at training intervention strengths $\sigma \in \{0, 0.25, 0.5\}$. Hidden width scales with agent count. All models are evaluated on adversarial rollouts only. Each condition is shown twice: once for models trained on the full observation, which includes each agent's LIDAR returns, and once for a no-LIDAR variant in which the LIDAR dimensions are removed from both the state space and the causal graph before training. The no-LIDAR variant is motivated by the dis-

covery stage: LIDAR readings produce dense, low-quality causal masks, and they are intrinsically hard to predict. The figure reports three metrics for both variants: raw MSE, error relative to the same model at $\sigma = 0.5$, and variance-normalized relative error.

With LIDAR, prediction error is dominated by the LIDAR dimensions themselves. Variance-normalized error sits between 0.1 and 1 for every architecture and every agent count, meaning the models recover only a modest fraction of the predictable variance. Differences between architectures are compressed, and increasing σ yields only a small improvement. This is consistent with the LIDAR returns acting as high-variance, largely unpredictable targets that swamp the error signal from the agent dynamics.

Dropping LIDAR reveals the structure that the full-state results obscure. Variance-normalized error falls to between 10^{-2} and 10^{-3} for AWM and ICWM, one to two orders of magnitude below the with-LIDAR setting. The benefit of interventional training data also becomes much larger: models trained at $\sigma = 0$ are up to ten times worse than the same architecture trained at $\sigma = 0.5$, whereas the corresponding ratio with LIDAR is closer to two or three. ICWM matches or slightly outperforms AWM across agent counts in this setting, most visibly at 3 and 4 agents. The trends are stable as the number of agents grows from 2 to 4, indicating that the approach scales with agent count once the uninformative sensor dimensions are removed.

Two architectures underperform throughout this sweep. C-VAE and R-ICWM sit roughly an order of magnitude above the other models in raw MSE in both variants and remain near a variance-normalized error of 1 even without LIDAR. For R-ICWM this contrasts with its competitive performance in the single-agent sweeps of Appendix Q.1, suggesting that its recurrent history mechanism does not transfer well to the larger multi-agent state spaces. FOCUS and BC-WM, trained only at $\sigma \approx 0$, again act as flat references and are outperformed by the σ-matched models as intervention strength grows.

V Towards Pearl's framework to evaluate counterfactual queries

V.1 Setup

Five world models in Appendix T receive a discovered causal graph (ICWM, R-ICWM, FOCUS, BC-WM, and C-VAE, which injects it into its latent transition); AWM receives none. If the graph does real work, the models using it should predict better out of distribution, and that edge should grow with graph accuracy. We test this over 1,240 runs and 295 datasets across three environments, six architectures, and the full intervention sweep, joining each run to its graph's structural error.

V.2 The pooled correlation is confounded

Pooled, structural error and out-of-distribution error correlate *negatively* (Fig. 26), which would mean worse graphs make better models. That is Simpson's paradox: the environments differ by an order of magnitude in variable count and error scale. Within a cell (environment, observation mode, intervention level, architecture) the correlation flips and reaches $+0.59$ under adversarial evaluation. But AWM, which never sees a graph, shows $+0.69$ under the same evaluation, with no difference between the groups ($p = 0.29$, Mann-Whitney on per-cell correlations).

Figure 27 confirms it: no significant gain to ICWM, FOCUS, or BC-WM over AWM, R-ICWM worse, and no scaling with graph quality. Intervention strength σ explains the $+0.59$, correlating with structural error ($\rho = -0.75$) and OOD error ($\rho = -0.48$) alike; holding it fixed drops the structural-to-OOD correlation to -0.01. The null replicates on frozen models across the seven shifted layouts of Appendix U,where the ICWM/AWM ratio straddles parity and the cross-architecture spread collapses to within 1% at the highest intervention. The two-door gate keeps 21–22 of 29 inputs open at every σ, since that environment's sparsity lives in the input-by-output coupling an input-marginal gate cannot express.

V.3 A LOCALIZED, CONDITIONAL EXCEPTION

Stretched Two-Door adds a never-visited region, forcing extrapolation. Under partial observation ICWM beats AWM by $0.53\times$ (Fig. 28A; 87% of paired cells, $p = 2.8 \times 10^{-15}$), and every graph-gated regressor beats AWM, while C-VAE, given the same soft-gated PCMCI graph, is $2.66\times$ worse: structure pays off only through an architecture that can use it. Under full observation the mask *hurts*, gating away usable signal. Panel F reads this as Pearl's three stages: partial observation is abduction, where a structural prior substitutes for unseen state; σ is action, peaking at 0.5; the unseen region is prediction.

V.4 WHY STRETCH BENEFITS AND ROTATION DOES NOT

The graph is byte-identical across all five two-door shifts, so it cannot explain their differences (Fig. 29). What matters is whether the shift keeps the frozen mask *aligned* with the room's coordinates and whether a state-variable mask can act on it. Stretching adds unseen values along meaning-preserving axes (invalidity 0.0; helps, 0.68–$0.92\times$). Mirroring swaps door columns but keeps the axes (invalidity 0.5; helps on average, fails at high intervention). Rotation transposes the axes (invalidity 1.0; $p = 0.35$). Wrap changes only the transition rule (invalidity 0.0, lowest novelty; $1.03\times$, $p = 0.95$). Anti-Causal swaps the door *states*, which the mask firewalls, for the largest benefit anywhere ($0.007\times$). Novelty does not decide it: stretch and rotation share 75% novelty with opposite outcomes, wrap and anti-causal 33% with 150-fold-different outcomes. What changed does, a coordinate's meaning, the transition rule, or a variable's value; Appendix V.6 isolates the last two.

V.5 FIGURES

Figures 26 and 27 cover the pooled analysis: how structural recovery relates to out-of-distribution error, why the raw correlation is confounded, and what happens when the graph is withheld. Figures 28 and 29 cover the exception and its boundary: where a frozen causal mask does help, and what property of a shift decides that.

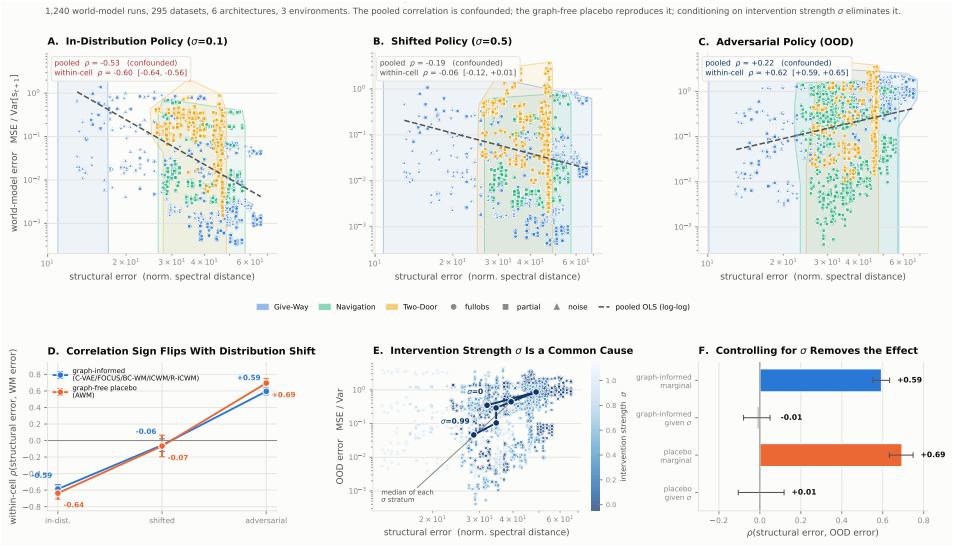


Figure 26: Structural recovery versus counterfactual world-model error, pooled over 1,240 world-model runs and 295 datasets. The naive pooled correlation is confounded by environment; the within-cell correlation rotates from negative in distribution to $+0.59$ under adversarial, out-of-distribution evaluation (panel D); conditioning on intervention strength σ removes the effect entirely (panel F).

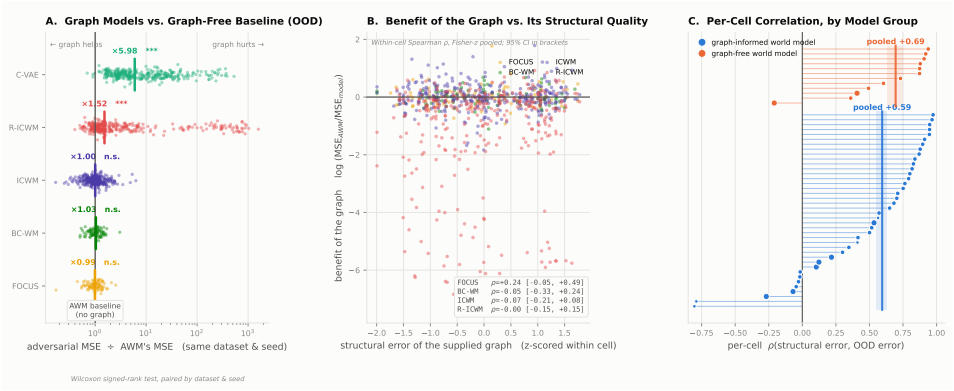


Figure 27: Does receiving a causal graph improve out-of-distribution prediction? Panel A compares graph-informed against graph-free models directly; panel B tests whether the benefit of the graph scales with the graph's own structural quality (it does not); panel C breaks the per-cell correlation down by model group, the placebo test that motivates Figure 26's common-cause reading.

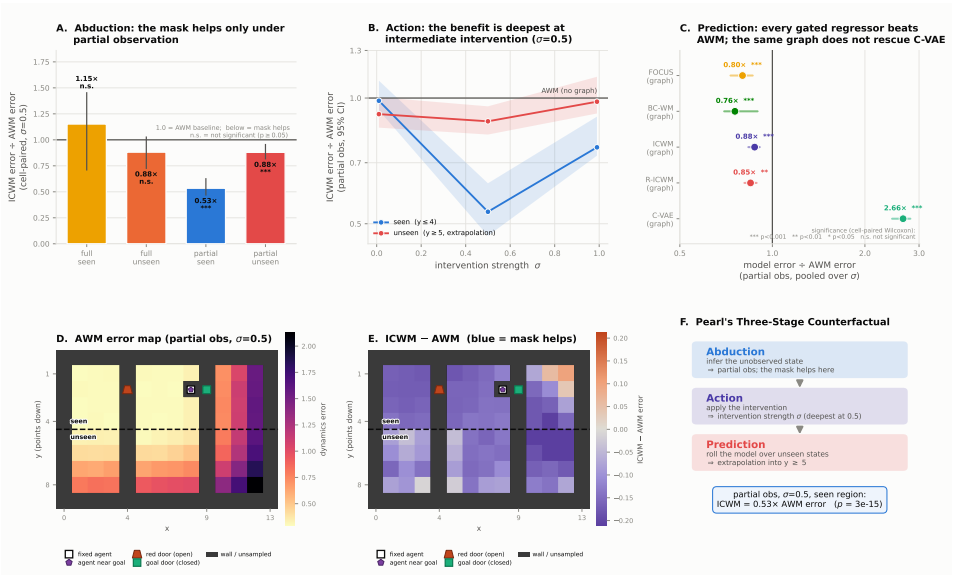


Figure 28: Where the causal mask pays off: partial observation and extrapolation on the Stretched Two-Door layout. Panel A: the mask helps only under partial observation. Panel B: the benefit is deepest at intermediate intervention strength ($\sigma = 0.5$). Panel C: every graph-gated regressor beats the graph-free AWM baseline under partial observation, while C-VAE, which receives the same soft-gated graph, is far worse. Panels D–E: the spatial error maps behind panel A's headline number. Panel F: the result read as Pearl's three-stage counterfactual query, abduction, action, prediction, with the mask assisting specifically the abduction stage.

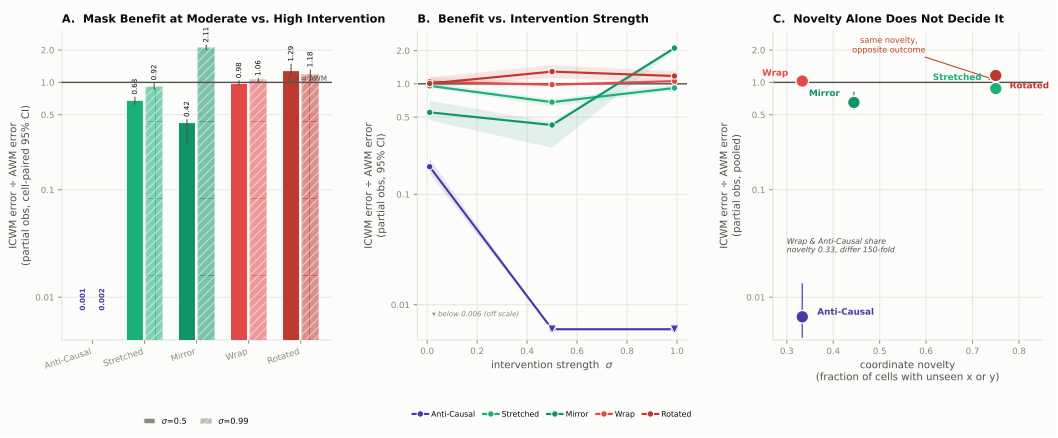


Figure 29: What decides when the mask helps: alignment and what moved, not the graph. The discovered graph is identical across all five two-door shifts; what differs is whether the shift keeps the frozen mask aligned with the room's coordinates and whether it is a shift the mask can act on at all. Panel A: benefit at moderate versus high intervention strength, by shift (Anti-Causal's sub-0.01 ratio is labelled at the axis floor). Panel B: benefit across the full intervention sweep. Panel C: Stretched and Rotated present the same 75% coordinate novelty with opposite outcomes, and Wrap and Anti-Causal the same 33% novelty with 150-fold-different outcomes, so novelty alone does not decide it.

V.6 Three controlled shifts that separate what earlier layouts confounded

The layouts above all move the geometry, confounding the room moving, the visited coordinates changing, and the state distribution changing. Three shifts perturb one factor each, with the graph and frozen checkpoints unchanged. `Wrap Two-Door` keeps the room byte-identical and changes only the transition rule: a `forward` step into a wall re-enters from the opposite side. `Anti-Causal Two-Door` keeps the geometry and every door position and swaps only the door *states* (blue open, red closed, agent moved to match), off-manifold in a task requiring red before blue. `Easy Give-Way` widens the notch five-fold in the continuous environment, where there is no grid mask to align. All ratios use the cell-paired median log-ratio, 4000-replicate bootstrap CI, and Wilcoxon test of Fig. 29A.

Wrap bounds the alignment claim. Wrap is the best case for the mask: zero invalidity, lowest novelty, doors in trained positions and states. The mask does nothing, 1.03 partial ($p = 0.95$) and 0.98 full ($p = 0.51$), against 0.88 for stretched at the same invalidity. Alignment is necessary but not sufficient: a mask constrains which variables inform a prediction, not the rule mapping them forward.

Anti-Causal reveals containment, not accuracy. Swapped doors give the largest mask effect anywhere, 0.007 on agent-state error, winning 100% of 1,960 cells ($p = 1.5 \times 10^{-70}$). It is not better door prediction: on *all* dimensions every model sits near a median error of 112, the environment being off-manifold for all of them. The gap is confined to agent-state, where ICWM (0.03) is 90× below AWM (2.37) on identical geometry, mostly because AWM blows up to 8.9× its unshifted error. The mask keeps the corrupted door signal out of the agent-position prediction; AWM conditions on everything and lets it contaminate a sub-problem it could otherwise learn. Layout shifts move every variable at once, so no earlier environment could show this.

Two caveats. The containment is learned from σ, ICWM's advantage growing from 0.18 at $\sigma = 0.01$ to 0.001 at $\sigma = 0.5$ under partial observation while the other shifts stay flat. And two families contain the error differently, separated by prediction confidence Δ_{var} (prediction error over mean-predictor variance). ICWM and R-ICWM mispredict the doors as confidently as AWM ($\Delta_{\text{var}} \approx 270$) but keep that error out of the agent sub-problem, while FOCUS and BC-WM, the $\sigma = 0$ baselines, sit in a near-static regime ($\Delta_{\text{var}} \approx 67$, all-dimension error ≈ 12): they never learned a dynamic door model, so their lower door error is a guess, not containment.

Easy Give-Way is a clean control. The mask stays within 4% of AWM at every intervention level (0.96 pooled under observation noise), against the $0.65\times$–$2.1\times$ swings of the two-door family; a widened notch gives a grid-shaped mask nothing to align to. Some points reach significance only because the continuous sweep has ~ 44k cells, so this is "no layout-shift-scale effect" rather than a strict null.

What moved, not how much. Wrap and Anti-Causal share the same 0.33 fraction of unseen coordinates, yet the mask does nothing on one and helps 150-fold on the other. The kind of change sorts them: a corrupted variable is firewalled, rearranged but meaning-preserving space helps if the mask stays aligned, and a changed movement rule or transposed axes leave the mask nothing to act on. Containment is not an observation artifact, being near-identical under full and partial observation (0.0069 vs 0.0066).

V.7 Summary

Structural recovery quality does not predict counterfactual accuracy, pooled, within cells after the control test, under layout shift, or with the graph held identical across shifts. What looks like a structure effect is σ acting on the graph and the world model separately.

One regime survives: partial observation, intermediate σ, a shift preserving each coordinate's meaning, and an architecture that can use the mask. Even there the benefit tracks alignment between the frozen prior and the query's geometry, not graph accuracy. The controlled shifts give the mechanism. An aligned mask buys nothing when the transition rule is what moved (Wrap); where it helps most, on swapped doors, it works by *containment*, keeping a corrupted variable group out of a learnable sub-problem rather than predicting those variables better. That firewall strengthens with training intervention by two orders of magnitude, the same σ lever behind every result above. Structure is an intervention-learned tool for one stage of a counterfactual question, not a general performance lever, which is what motivates the closing section.