
DISCOVERING WHAT YOU CAN CONTROL: INTERVENTIONAL BOUNDARY DISCOVERY FOR REINFORCEMENT LEARNING

Jiaxin Liu

University of Illinois Urbana-Champaign
jiaxin26@illinois.edu

ABSTRACT

Selecting relevant state dimensions in the presence of confounded distractors is a causal identification problem: observational statistics alone cannot reliably distinguish dimensions that correlate with actions from those that actions cause. We formalize this as discovering the agent’s *Causal Sphere of Influence* and propose Interventional Boundary Discovery (IBD), which applies Pearl’s do-operator to the agent’s own actions and uses two-sample testing to produce an interpretable binary mask over observation dimensions. IBD requires no learned models and composes with any downstream RL algorithm as a preprocessing step. Across 12 continuous control settings with up to 100 distractor dimensions, we find that: (1) observational feature selection can actively select confounded distractors while discarding true causal dimensions; (2) full-state RL degrades sharply once distractors outnumber relevant features by roughly 3:1 in our benchmarks; and (3) IBD closely tracks oracle performance across all distractor levels tested, with gains transferring across SAC and TD3.

1 Introduction

Consider a robotic arm learning to reach a target. Its sensory stream contains joint angles, velocities, and target position, all causally influenced by the arm’s motor commands. But the same stream also contains readings from ambient temperature sensors, background object positions, lighting intensity, and other agents’ movements. These distractor dimensions correlate with the robot’s state in complex, time-varying ways: a nearby human’s motion may be temporally correlated with the arm’s trajectory, and lighting changes may co-vary with the target’s visibility. Yet the robot’s actions have no causal effect on any of them. This distinction—between what an agent observes to correlate with its behavior and what it actually influences—is a causal question, not a statistical one. When a confounder C drives both the agent’s state S and a distractor D , the distractor can exhibit high mutual information with actions even though actions have no causal effect on D [Pearl, 2009]. Observational data, no matter how abundant, cannot distinguish this confounded dependence from a genuine causal link.

Our starting observation is simple: an RL agent that can randomize its own actions already has access to a natural intervention mechanism. This reframes the distractor problem as one of causal identification: the agent’s action space already provides the tool needed to separate causal dimensions from confounded ones.

Despite the causal nature of this problem, existing approaches have predominantly tackled it through representation learning. Methods like CURL [Laskin et al., 2020], DrQ-v2 [Yarats et al., 2022], and SVEA [Hansen et al., 2021] learn visual encoders that are robust to pixel-level distractors through augmentation and contrastive objectives. In the causal representation learning literature, methods like CITRIS [Lippe et al., 2022] learn disentangled latent causal variables, a more ambitious goal. What has received comparatively little attention is the simplest causal tool available to any agent that can act in its environment: intervention. An agent that can randomize its own actions and observe the consequences has access to Pearl’s do-operator [Pearl, 2009], without needing to learn any model of the world.

In this paper, we formalize the problem of discovering an agent’s *Causal Sphere of Influence* (SoI)—the set of observation dimensions that are causally downstream of at least one action dimension—and propose *Interventional Boundary Discovery* (IBD) to solve it. IBD randomizes the agent’s actions, compares the resulting trajectory statistics

against a baseline using two-sample tests, and applies Benjamini-Hochberg correction for multiple testing. Because the do-operator severs all confounding paths by construction, the test is not confounded by spurious correlations between actions and observations.

The output of IBD is a binary mask over observation dimensions: 1 for causally influenced, 0 otherwise. This mask can be applied to any downstream RL algorithm as a preprocessing step. We demonstrate this with both SAC [Haarnoja et al., 2018] and TD3 [Fujimoto et al., 2018]. The mask also serves a diagnostic function: by revealing which dimensions the agent can and cannot influence, IBD decomposes an environment’s difficulty into distinct failure modes. If the oracle (perfect mask) performs well but full-state does not, the bottleneck is representational confusion from distractors; if the oracle also fails, the bottleneck is exploration or reward shaping. This decomposition helps practitioners decide where to invest effort.

We evaluate IBD across 12 continuous control settings spanning 6 DeepMind Control Suite [Tassa et al., 2018] tasks with three tiers of distractor complexity. Our distractor design goes beyond Gaussian noise: it includes mimicking distractors calibrated to match the statistical fingerprint of true state dimensions, and reward-correlated distractors that track episode progress without being causally influenced by actions. These ensure that observational methods face a difficult discrimination problem.

IBD is lightweight: once the problem is cast as causal identification, the agent’s actions already furnish the needed interventions, and the solution reduces to a two-sample test.

Concretely, our contributions are:

1. **Problem reformulation.** We formalize the *Causal Sphere of Influence* and construct an MDP in which mutual information ranks a confounded distractor above the true causal dimension (Section 3.3), illustrating that the failure of observational feature selection under confounding is structural and persists regardless of estimator quality.
2. **Minimal interventional solution.** We introduce IBD, which treats the agent’s actions as interventions and applies a two-sample test with multiple-testing correction to produce a binary mask that plugs into any RL algorithm (Section 3.2).
3. **Empirical characterization.** We identify a *distractor scaling effect*: in our benchmarks, full-state RL degrades sharply once the distractor-to-signal ratio exceeds roughly 3:1, while IBD closely tracks oracle performance across all distractor levels tested (Sections 5.1 and 5.2).
4. **Diagnostic framework.** Comparing Full State, IBD, and Oracle performance decomposes an environment’s difficulty into representational confusion (distractors) versus exploration bottleneck, giving practitioners actionable guidance on where to invest effort (Section 5.7).

We additionally show that IBD transfers across RL algorithms (Section 5.5) and detects partially controllable dimensions down to $\sim 5\%$ causal variance (Section 5.6).

Scope. IBD operates on structured observation vectors, where interventional testing is well-defined. In modern visual RL pipelines, a learned encoder (e.g., from DrQ-v2 or CURL) first maps pixels to a feature vector, and downstream components operate on that vector. IBD slots naturally into this pipeline as a post-encoder selector: its two-sample test applies to any flat feature space regardless of how that space was produced. We demonstrate IBD on state-space tasks where ground-truth causal labels enable rigorous evaluation; the encoder-then-IBD pipeline is a natural extension that requires no algorithmic changes. The method assumes the agent can override its actions during a short probing phase and that the causal structure is stationary during probing; these conditions hold in simulation and in physical systems with safe-exploration protocols.

2 Related Work

RL with distractors. The challenge of learning from observations with task-irrelevant content has been studied extensively in the visual RL setting. The Distracting Control Suite [Stone et al., 2021] introduces visual distractors (dynamic backgrounds, camera jitter, color perturbation) to DeepMind Control tasks, and methods such as SVEA [Hansen et al., 2021], DrQ-v2 [Yarats et al., 2022], and CURL [Laskin et al., 2020] address this through data augmentation and contrastive learning at the pixel level. These approaches are complementary to ours: they handle pixel-level distractors through learned robustness, while we address state-level distractors through explicit causal identification. When a learned encoder extracts a feature vector from pixels, IBD could serve as a downstream selector on that feature space. In the state-space setting, the Exogenous Block MDP (EX-BMDP) framework [Efroni et al., 2022] formalizes exogenous distractors and provides provable sample complexity bounds via multistep inverse dynamics; recent extensions handle

single-trajectory settings [Levine et al., 2025a] and action-free representation learning [Levine et al., 2025b]. Separated world models have also been proposed to disentangle endogenous and exogenous factors for visual RL [Huang et al., 2024], while Denoised MDPs [Wang et al., 2022] categorize information by controllability and reward-relevance. These approaches rely on model estimation or inverse dynamics learning, whereas IBD operates directly via nonparametric statistical testing under intervention, making fewer structural assumptions.

Causal representation learning. A growing body of work aims to learn causal representations from observational or interventional data. Schölkopf et al. [2021] articulate the broader agenda of discovering causal variables from raw observations, and methods like CITRIS [Lippe et al., 2022] and iCITRIS [Lippe et al., 2023] learn disentangled latent causal variables from temporal data with interventions. Recent work has unified CRL under the invariance principle [Yao et al., 2025], extended identifiability to unknown multi-node interventions [Varici et al., 2024], and established score-based identifiability for general nonlinear models [Varici et al., 2025]. Concurrently, causal approaches to confounding in deep RL have been formalized [Li et al., 2025]. These methods solve a more ambitious problem, discovering the latent causal graph, whereas IBD operates at a simpler level of abstraction: a binary mask over observed dimensions, identifying which dimensions to attend to rather than how they causally relate.

Empowerment and controllability. The concept of *empowerment* [Klyubin et al., 2005, Mohamed and Rezende, 2015]—the channel capacity between an agent’s actions and its future states—is the closest conceptual relative to our notion of the Sphere of Influence. Empowerment quantifies how much influence an agent can exert, aggregated as a scalar over all dimensions. IBD answers a complementary question: which specific dimensions does the agent causally influence at all? This difference in abstraction level has practical consequences. Empowerment requires density estimation in high-dimensional state-action spaces and, when estimated per-dimension, measures statistical dependence: a confounded distractor that correlates with the agent’s state can exhibit high mutual information $I(A; S'|S)$ without being causally influenced. IBD replaces density estimation with a hypothesis test under the do-operator, ensuring that only causal influence is detected. In linear systems, controllability analysis [Kalman, 1960] can determine which state dimensions are reachable from actions. IBD can be viewed as a nonparametric, nonlinear generalization that determines causal reachability without a dynamics model. Recent work on interpretable controllability features in MDPs [Kooi et al., 2023] also aims to distinguish controllable from uncontrollable dimensions, though via learned latent representations rather than direct interventional testing.

Feature selection and state abstraction. Classical feature selection methods (mutual information, variance-based ranking, forward-model prediction error) are widely used in supervised learning and have natural extensions to RL state spaces. Our Cond. MI baseline, which measures the R^2 gain from including actions in a learned forward model, represents a stronger observational approach than raw MI or variance. State abstraction methods [Li et al., 2006] aim to group states that are equivalent under the optimal policy, a different objective from identifying the agent’s causal reach. We focus on the causal identification aspect of feature selection, which is complementary to these approaches.

3 Method

3.1 The Causal Sphere of Influence

We consider a Markov Decision Process $\mathcal{M} = (\mathcal{O}, \mathcal{A}, T, R, \gamma)$ where the observation space $\mathcal{O} \subseteq \mathbb{R}^d$ decomposes (unknown to the agent) as $\mathbf{o}_t = [\mathbf{s}_t^c, \mathbf{s}_t^d]$: a causal component $\mathbf{s}^c \in \mathbb{R}^{d_c}$ that is causally downstream of the agent’s actions $\mathbf{a}_t \in \mathcal{A} \subseteq \mathbb{R}^{d_a}$, and a distractor component $\mathbf{s}^d \in \mathbb{R}^{d_d}$ that evolves autonomously. The total observation dimension is $d = d_c + d_d$, and the agent does not know which dimensions are causal.

Definition 3.1 (Causal Sphere of Influence). Observation dimension i belongs to the agent’s *Sphere of Influence* (SoI) if and only if there exists an action dimension j and a horizon $h \geq 1$ such that the intervention $\text{do}(a_t^{(j)} = u)$ changes the marginal distribution of $o_{t+h}^{(i)}$:

$$\mathbb{P}\left(o_{t+h}^{(i)} \mid \text{do}(a_t^{(j)} = u)\right) \neq \mathbb{P}\left(o_{t+h}^{(i)}\right) \quad (1)$$

for some intervention value u , where $\text{do}(\cdot)$ denotes Pearl’s do-operator [Pearl, 2009].

This definition captures a precise intuition: dimension i is inside the SoI if and only if there exists a directed causal path from some action to dimension i in the environment’s causal graph. Crucially, (1) involves the do-operator, not conditional probability. A dimension correlated with actions due to a common confounder satisfies $\mathbb{P}(o^{(i)}|a) \neq \mathbb{P}(o^{(i)})$ but does *not* satisfy $\mathbb{P}(o^{(i)}|\text{do}(a = u)) \neq \mathbb{P}(o^{(i)})$, because the do-operator severs the confounding path.

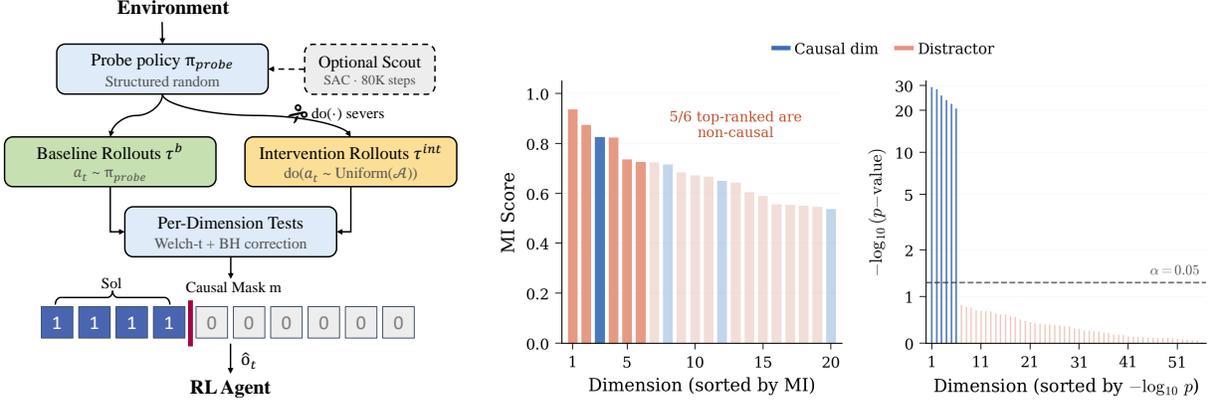


Figure 1: **IBD pipeline and core result.** *Left:* Phase 1: a structured random probe policy collects baseline trajectories and interventional trajectories where actions are replaced by random noise. Phase 2: per-dimension Welch t -tests with BH correction produce a binary causal mask applied to a downstream RL algorithm. *Right:* Feature ranking on `reacher_hard` (6 true dims, 50 distractors). Under MI-based ranking, causal and distractor dimensions are interleaved; under IBD ranking, all causal dimensions fall above the $\alpha=0.05$ threshold and all distractors fall below.

The goal of IBD is to recover SoI from interaction data, producing a binary mask $\mathbf{m} \in \{0, 1\}^d$ with $m_i = \mathbf{1}[i \in \widehat{\text{SoI}}]$. The downstream RL algorithm then receives only the masked observation $\mathbf{o}_t \odot \mathbf{m}$.

Assumptions. IBD relies on the following assumptions, which we state explicitly to clarify its scope.

1. *Decomposability.* The observation space decomposes into causal and distractor components as above; dimensions are either fully causal or fully exogenous. We relax this in Section 5.6 to handle partially controllable dimensions.
2. *Action overrideability.* The agent can replace its actions with random noise during the probing phase. This is natural in simulation; on physical systems it requires a safe exploration protocol.
3. *Stationary causal structure.* The causal graph does not change during the probing phase. Non-stationary environments would require periodic re-probing.
4. *Faithfulness.* If a directed causal path from an action to a dimension exists, the intervention produces a non-trivial distributional shift (Proposition 3.4).

3.2 Algorithm

IBD operates in two phases (Figure 1).

Phase 1: Data collection. A probe policy π_{probe} collects trajectories that induce non-degenerate state coverage so that causal effects are detectable. By default, we use a structured random policy (sinusoidal actions with weak state feedback and exploration noise) which requires no RL training (see Section 3.5). Using π_{probe} , we collect two sets of trajectories. Baseline trajectories $\{\tau_k^b\}_{k=1}^N$ are collected under the probe policy’s normal behavior. Intervention trajectories $\{\tau_k^{int}\}_{k=1}^N$ are collected with all action dimensions replaced by i.i.d. draws from $\text{Uniform}(\mathcal{A})$ at every step, implementing $do(\mathbf{a}_t = \text{noise})$ and severing all incoming edges to the action node in the causal graph.

Phase 2: Statistical testing. For each observation dimension i and each horizon $h \in \mathcal{H}$ (we use $\mathcal{H} = \{1, 5, 10\}$), we extract a trajectory-level summary statistic: the mean absolute h -step difference,

$$\Delta_i^h(\tau) = \frac{1}{K} \sum_{k=1}^K \left| o_{i+k}^{(i)} - o_{i+(k-1)h}^{(i)} \right|, \quad (2)$$

where $K = \lfloor T/h \rfloor$ and windows are non-overlapping. This yields one scalar per trajectory, ensuring sample independence across the test.

We then perform a Welch t -test comparing the baseline and intervention samples $\{\Delta_i^h(\tau_k^b)\}$ versus $\{\Delta_i^h(\tau_k^{int})\}$. Across all $m = d \times |\mathcal{H}|$ tests, Benjamini-Hochberg (BH) correction [Benjamini and Hochberg, 1995] controls the false

Algorithm 1 Interventional Boundary Discovery (IBD)

Require: Environment \mathcal{M} , probe policy π_{probe} (default: structured random), trajectory count N , length T , horizons \mathcal{H} , significance α

- 1: Collect N baseline trajectories $\{\tau^{\text{b}}\}$ using π_{probe}
- 2: Collect N intervention trajectories $\{\tau^{\text{int}}\}$ using $\text{do}(\mathbf{a} = \text{noise})$
- 3: **for** each dimension $i \in \{1, \dots, d\}$ and horizon $h \in \mathcal{H}$ **do**
- 4: Compute $\Delta_i^h(\tau_k^{\text{b}})$ and $\Delta_i^h(\tau_k^{\text{int}})$ for all k
- 5: Obtain p -value $p_{i,h}$ from Welch t -test
- 6: **end for**
- 7: Apply Benjamini-Hochberg correction to $\{p_{i,h}\}$ at level α
- 8: $\widehat{\text{SoI}} \leftarrow \{i : \exists h \in \mathcal{H} \text{ s.t. adjusted } p_{i,h} < \alpha\}$
- 9: **return** Binary mask \mathbf{m} with $m_i = \mathbf{1}[i \in \widehat{\text{SoI}}]$

(Eq. 2)

discovery rate at level $\alpha = 0.05$. Dimension i is classified as $i \in \widehat{\text{SoI}}$ if any horizon h yields a significant adjusted p -value. The complete procedure is summarized in Algorithm 1.

Multi-horizon testing. A causal chain $a \rightarrow x_1 \rightarrow x_2 \rightarrow o_i$ of length k may only produce a detectable distributional shift at horizon $h \geq k$, because the effect needs k time-steps to propagate. Testing at multiple horizons $\mathcal{H} = \{1, 5, 10\}$ ensures that both immediate and delayed causal effects are captured. The BH correction accounts for the additional tests without inflating the false discovery rate.

3.3 Structural Limitations of Observational Selection

The difficulty that observational feature selection faces under confounded distractors is not a failure of estimation; it is structural. We illustrate this with a concrete construction.

Proposition 3.2 (MI can rank distractors above causal dimensions). *There exists an MDP in which a distractor dimension d_k achieves the highest mutual information with actions among all observation dimensions, yet $\text{do}(\mathbf{a})$ has zero causal effect on d_k .*

Construction. Let s_1 be a true state dimension influenced by action a_1 : $s_{1,t+1} = f(s_{1,t}, a_{1,t}) + \eta$. Define a distractor $d_k = s_{1,t} + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ with small σ_ε . Because d_k tracks s_1 closely, $\text{MI}(d_k; \mathbf{a}) \approx \text{MI}(s_1; \mathbf{a})$; so both rank equally under MI-based selection. With sufficiently small σ_ε , d_k can even exceed s_1 in MI due to noise reduction effects. However, d_k is not causally downstream of \mathbf{a} : performing $\text{do}(\mathbf{a} = \text{noise})$ changes the future of s_1 but has no effect on d_k , because d_k 's value at time t is already determined. The dependence of d_k on actions is entirely mediated by the confounding path through s_1 , which the do -operator severs. \square

This construction uses a distractor that is a deterministic function of the true state for simplicity; the key insight, that confounding inflates observational scores, extends to truly exogenous distractors. Our benchmark (Section 4) instantiates this principle with mimicking distractors that are independent exogenous processes calibrated to reproduce the variance, autocorrelation, and frequency content of true proprioceptive dimensions, and reward-correlated distractors whose drift rate tracks episode progress. In experiments, the MI baseline achieves a return of 12 on `reacher_hard` with 50 distractors, compared to IBD's 929 (Table 1). Even with a learned forward model that conditions on state (Cond. MI), the observational approach degrades from 476 (easy) to 7 (medium) on the same task, suggesting that the underlying identifiability gap persists even with state-conditioned models.

3.4 Theoretical Properties

We state three properties of IBD that together characterize its reliability guarantees. Each follows from standard results in causal inference and nonparametric statistics; we include them to make explicit why an interventional approach avoids the failure modes of observational selection.

Proposition 3.3 (Confounding immunity). *Because IBD collects interventional data under $\text{do}(\mathbf{a} = \text{noise})$, all edges $C \rightarrow \mathbf{a}$ from confounders C to actions are severed. The two-sample test statistic is therefore invariant to the presence or absence of confounders: $\widehat{\text{SoI}}$ is identical whether confounders are active or removed.*

This is the property that distinguishes IBD from observational alternatives. Confounders create correlations between actions and observations that persist under any amount of observational data. The do -operator eliminates these correlations by making actions independent of all non-descendant variables [Pearl, 2009].

Table 1: Main results (episode return, mean \pm std over 5 seeds). **Bold**: best non-oracle method (within 1 std).

Environment	Distr.	Full State	IBD (ours)	Oracle	MI Select	Var Select	Cond. MI
walker_walk	easy	857 \pm 139	854 \pm 127	785 \pm 128	705 \pm 98	750 \pm 103	884\pm70
walker_walk	medium	902\pm26	792 \pm 112	785 \pm 128	129 \pm 79	84 \pm 6	220 \pm 135
walker_walk	hard	414 \pm 144	842\pm121	785 \pm 128	87 \pm 34	46 \pm 10	81 \pm 23
cheetah_run	easy	508\pm39	429 \pm 80	472 \pm 35	161 \pm 42	85 \pm 19	270 \pm 112
cheetah_run	medium	113 \pm 34	479\pm95	472 \pm 35	38 \pm 11	62 \pm 25	70 \pm 37
cheetah_run	hard	63 \pm 12	436\pm70	472 \pm 35	33 \pm 6	39 \pm 3	62 \pm 41
reacher_hard	easy	748 \pm 125	939\pm38	925 \pm 75	197 \pm 375	4 \pm 2	476 \pm 289
reacher_hard	medium	12 \pm 7	929\pm76	925 \pm 75	12 \pm 5	7 \pm 5	7 \pm 6
reacher_hard	hard	8 \pm 4	907\pm72	925 \pm 75	13 \pm 9	10 \pm 5	9 \pm 5
cartpole_swingup	medium	725 \pm 48	834\pm26	821 \pm 35	78 \pm 25	83 \pm 23	161 \pm 60
finger_spin	medium	365 \pm 13	587\pm178	775 \pm 181	1 \pm 1	1 \pm 1	4 \pm 6
hopper_hop	medium	0 \pm 1	6\pm12	6 \pm 8	0 \pm 0	0 \pm 0	0 \pm 0

Proposition 3.4 (Interventional detectability). *Under the faithfulness assumption (if a directed causal path from a_j to o_i of length $\leq h$ exists, then the conditional distribution of $o_{t+h}^{(i)}$ given $\text{do}(a_j = u)$ depends non-trivially on u), intervening on action dimension j changes the distribution of h -step state differences on dimension i if and only if a directed causal path of length $\leq h$ from a_j to o_i exists.*

Proposition 3.5 (Finite-sample error control). *The permutation test variant of IBD controls per-test Type I error at level $\leq \alpha$ exactly, for any sample sizes N and any number of permutations B [Phipson and Smyth, 2010]. The Welch t -test variant controls Type I error asymptotically under the central limit theorem applied to trajectory-level summaries. In both cases, BH correction across all $m = d \times |\mathcal{H}|$ tests controls the overall false discovery rate at level α .*

Together, these properties provide safeguards against two failure modes: including distractor dimensions (bounded by FDR control) and missing causal dimensions (ensured by detectability under faithfulness, with multi-horizon testing capturing delayed effects).

3.5 Practical Considerations

The probe policy need not be a trained RL agent. Its role is to induce sufficient state coverage so that the intervention’s causal signal is detectable. Our default probe is the structured random policy described in Section 3.2, which requires no RL training and no GPU. This policy suffices for all tasks in our experiments: our scout ablation (Section C) shows that training an RL scout for up to 160K steps provides no improvement in boundary accuracy on `cheetah_run` and `reacher_hard`, and slightly decreases recall on `walker_walk` ($F1 = 0.94$ without scout vs. 0.90 with 80K-step scout), likely because the trained policy’s narrower behavioral distribution reduces state coverage. An optional lightweight scout can be trained if domain knowledge suggests that structured random actions cannot reach relevant states; in our experiments, the structured random probe was sufficient throughout.

IBD supports two intervention modes. Joint intervention randomizes all action dimensions simultaneously, requiring only one intervention dataset and producing a stronger aggregate signal. Per-action intervention randomizes one dimension at a time, yielding a finer-grained Causal Influence Matrix at the cost of d_a intervention datasets. Joint intervention suffices for feature selection and is our default.

IBD’s computational cost is: $2N$ trajectories of length T for data collection, plus negligible time for statistical testing. With our default settings ($N = 80$, $T = 200$), the probing phase requires approximately 32K environment steps and completes in under 3 minutes with no GPU requirement. This is a one-time cost: the mask is computed once and reused for the entire downstream training run.

4 Experimental Setup

4.1 Environments and Distractor Design

We evaluate on 6 tasks from the DeepMind Control Suite [Tassa et al., 2018]: `walker_walk` (24 true dims), `cheetah_run` (17), `cartpole_swingup` (5), `finger_spin` (9), `hopper_hop` (15), and `reacher_hard` (6). Each task is augmented with distractors at three levels of difficulty, designed to defeat progressively more sophisticated selection methods.

The **easy** level appends 6 autonomous distractors (4 Ornstein-Uhlenbeck processes and 2 coupled oscillators) that evolve independently and have obviously different statistical properties from true state dimensions. The medium level appends 50 distractors: 12 autonomous, 28 mimicking, and 6 reward-correlated, plus 4 oscillators. Mimicking distractors are OU processes calibrated to match the scale, variance, and autocorrelation of true proprioceptive dimensions, with calibration based on the true observation dimensionality ($\sigma_{\text{ref}} = 0.5 + 0.05 \cdot d_c$). These defeat variance-based selection, as they are statistically indistinguishable from true dimensions in their marginal distribution. Reward-correlated distractors follow drift processes whose rate tracks episode progress (distance traveled, target proximity), creating spurious mutual information with actions without being causally influenced by them. The **hard** level appends 100 distractors: 18 autonomous, 60 mimicking, 14 reward-correlated, and 8 oscillators, bringing the signal-to-noise ratio to approximately 1:4 for a task like `walker_walk` (24 true dims vs. 100 distractors).

All distractors are truly exogenous: no causal path from any action to any distractor exists. Distractors do not perturb observations or bias actions; they are appended as additional observation dimensions. This guarantees that the oracle mask (keeping only true state dims) is a valid upper bound on achievable performance. We relax this purely-exogenous assumption in Section 5.6, where we evaluate IBD on partially controllable dimensions.

4.2 Methods and Evaluation

We compare six methods, all using SAC [Haarnoja et al., 2018] as the downstream RL algorithm with identical hyperparameters (learning rate 3×10^{-4} , batch size 256, MLP [256, 256], 300K training steps, 5 seeds). **Full State** trains SAC on the complete observation $\mathbf{o} \in \mathbb{R}^{d_c+d_a}$ with no feature selection. **Oracle** trains SAC on the true causal dimensions only ($\mathbf{o}^c \in \mathbb{R}^{d_c}$), using ground-truth labels; this is an upper bound. **IBD** (ours) runs the two-phase procedure described in Section 3.2, then trains SAC on the discovered $\widehat{\text{SoI}}$ dimensions. **MI Select** computes mutual information between each observation dimension and actions from observational rollouts, then selects the top- d_c dimensions. **Variance Select** ranks dimensions by variance of action-conditioned prediction residuals and selects the top- d_c . **Cond. MI** is a stronger observational baseline that trains two small MLPs per observation dimension—one predicting s'_i from (s, a) and one from s alone—and ranks dimensions by the R^2 gain from including actions. This conditions on the current state, filtering out much of the marginal correlation that confounds raw MI, but remains observational: a distractor whose dynamics correlate with the agent’s proprioceptive state can still score highly without any causal link from actions. For the MI, Variance, and Cond. MI baselines, we provide the true number of causal dimensions d_c as the selection budget. This is a substantial advantage that IBD does not receive, since IBD must determine both which and how many dimensions to select using only the significance threshold α .

Each method is evaluated on the mean \pm standard deviation of the final episodic return (average of 10 evaluation episodes at the end of training) across 5 random seeds. For IBD, we additionally report boundary discovery accuracy: precision (fraction of selected dims that are truly causal), recall (fraction of truly causal dims that are selected), and F1.

5 Results

5.1 Main Results

Table 1 presents the full results across all 12 settings. IBD achieves the highest non-oracle return in 10 of 12 settings; in the remaining two cases (discussed in Section 5.7), full-state RL is already competitive and IBD closely matches the oracle. (In some easy settings, stochastic variation causes the oracle’s mean return to fall slightly below Full State; the differences are within one standard deviation.)

Observational baselines struggle across all difficulty levels. On `reacher_hard` with 50 distractors, Full State achieves 12, MI Select achieves 12, Variance Select achieves 7, all effectively random, while IBD achieves 929, nearly identical to the oracle’s 925. On `cartpole_swingup`, IBD (834) outperforms Full State (725) and approaches the oracle (821), while MI (78) and Variance (83) are an order of magnitude worse. On `finger_spin`, IBD (587) captures most of the oracle’s 775, while MI and Variance each achieve a return of 1.

Notably, MI and Variance often perform worse than doing nothing. On `walker_walk` with medium distractors, Full State achieves 902, but MI Select drops to 129 and Variance to 84: actively selecting features via observational statistics produces a 7–11 \times degradation relative to simply using all dimensions. On `cheetah_run` with easy distractors—only 6 autonomous distractors with no mimicking or reward-correlated components—MI (161) and Variance (85) still underperform Full State (508) by 3–6 \times . This occurs because both methods are given the true d_c as their selection budget (an advantage IBD does not receive) yet still select the wrong dimensions: mimicking distractors pass their statistical filters while true causal dimensions with lower variance or MI are excluded.

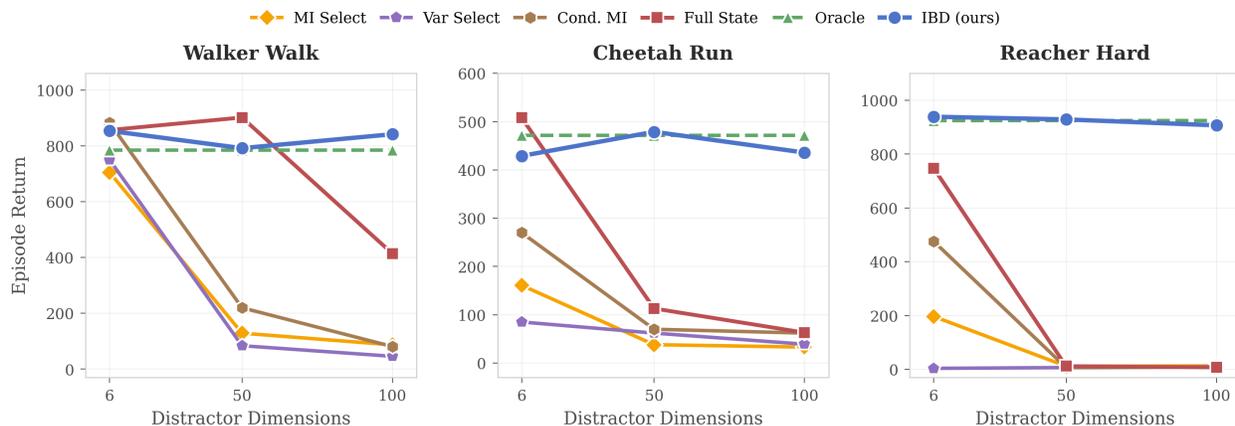


Figure 2: **Distractor scaling curve.** Episode return as a function of distractor dimensionality (6, 50, 100) for `walker_walk` (left), `cheetah_run` (center), and `reacher_hard` (right). Full State (red) degrades as distractors increase; IBD (blue) tracks oracle performance across all distractor counts. Shaded regions: ± 1 std over 5 seeds.

The conditional MI baseline (Cond. MI), which trains a learned forward model conditioned on the current state, is the most informative observational baseline we test. On easy settings it is competitive, achieving 884 on `walker_walk` (vs. IBD’s 854) and 476 on `reacher_hard` (vs. IBD’s 939), showing that state-conditioning filters out much of the marginal correlation when distractors are few and statistically distinct. However, Cond. MI degrades sharply as distractor complexity increases: on `walker_walk`, from 884 (easy) to 220 (medium) to 81 (hard); on `reacher_hard`, from 476 to 7 to 9. State-conditioning removes marginal correlation but does not remove confounding (Proposition 3.3), so distractors whose dynamics covary with the agent’s proprioceptive state can still receive high scores.

5.2 Distractor Scaling Effect

The scaling experiment (Figure 2) reveals a consistent pattern across the three tasks we examine in detail. On `cheetah_run`, Full State degrades from 508 (6 distractors) to 113 (50 distractors) to 63 (100 distractors), an 88% decline. On `walker_walk`, the collapse is delayed: Full State remains competitive at 902 (50 distractors) but drops to 414 (100 distractors), a 54% decline. Observational baselines degrade even faster: on `walker_walk`, MI Select falls from 705 (easy) to 129 (medium) to 87 (hard), while Variance Select falls from 750 to 84 to 46, performing worse than Full State at every distractor level above easy.

The collapse threshold correlates with the distractor-to-signal ratio. `cheetah_run` has 17 true dimensions, so 50 distractors yield a ratio of $\sim 3:1$, precisely where performance degrades sharply. `walker_walk` has 24 true dimensions, so the same 50 distractors yield only $\sim 2:1$, and the collapse occurs at 100 distractors ($\sim 4:1$). The most extreme case is `reacher_hard`, which has only 6 true dimensions. Here the distractor-to-signal ratio reaches $\sim 8:1$ at medium and $\sim 17:1$ at hard. Full State collapses from 748 (easy) to 12 (medium) to 8 (hard), a 98.9% degradation, while IBD maintains 939, 929, and 907, all within one standard deviation of the oracle’s 925. Variance Select is particularly striking: it achieves only 4 even with just 6 easy distractors, illustrating that observational methods can select exactly the wrong dimensions when the statistical mimicry is effective.

These results suggest a practical heuristic: in our benchmark suite, when the expected number of irrelevant features exceeds roughly $3\times$ the relevant features (with our fixed architecture and training budget), interventional feature selection offers substantial gains. In contrast, IBD’s performance is essentially flat across distractor counts, with fluctuations falling within one standard deviation and exhibiting no downward trend. The oracle is similarly stable, confirming that IBD’s mask effectively removes the dimensions responsible for full-state degradation.

Cross-task consistency of the collapse ratio. The degradation threshold is better explained by the distractor-to-signal ratio than by the absolute number of distractors. Table 2 summarizes the evidence: `walker_walk` (24 true dims) tolerates 50 distractors but collapses at 100; `cheetah_run` (17 true dims) already collapses at 50; `reacher_hard` (6 true dims) collapses at 50 as well. The absolute distractor counts at which collapse occurs differ (50–100), but the corresponding ratios cluster in the range 3–8:1. This consistency across tasks with different true dimensionalities suggests that the threshold reflects a property of the function approximator rather than of any specific task.

Table 2: Cross-task collapse analysis. ‘‘Collapse at’’ is the smallest distractor count where Full State return drops below 50% of the easy-distractor baseline. The collapse ratio is more consistent across tasks than the absolute count.

Task	d_c	Return (easy)	Collapse at	Ratio at collapse	Return at collapse
walker_walk	24	857	100	4.2 : 1	414
cheetah_run	17	508	50	2.9 : 1	113
reacher_hard	6	748	50	8.3 : 1	12

A capacity-based intuition. A [256, 256] MLP has a fixed number of parameters regardless of input dimensionality. As distractors inflate the input from d_c to $d_c + d_d$, the network must distribute its representational capacity across all dimensions, including irrelevant ones. The effective capacity per relevant input dimension scales roughly as width/ d ; when d grows by a factor of 3–5 \times from distractors alone, this per-dimension capacity drops below the threshold needed to learn a good value function or policy within the training budget. This argument predicts that the collapse ratio should shift upward with wider networks or longer training, and downward with more complex tasks that demand more capacity per dimension. We leave systematic verification of this prediction to future work, but note that it is consistent with our observation that `reacher_hard` (a simpler task with only 6 true dims) tolerates a higher ratio (8.3:1) before collapse than `cheetah_run` (2.9:1), whose 17-dimensional dynamics are more complex.

5.3 Causal vs. Observational Feature Ranking

Figure 1 (right) makes the observational–interventional distinction visually concrete. In the MI ranking, distractor dimensions are interspersed throughout the top positions; mimicking distractors that track the arm’s joint angles produce MI scores comparable to the true dimensions, so that 5 of the top-6 selected dimensions are non-causal. In the IBD ranking, every causal dimension produces a p -value above the significance threshold while every distractor falls below, consistent with the confounding immunity established in Proposition 3.3.

5.4 Boundary Discovery Accuracy

Across all 12 settings, IBD achieves mean precision 0.96, mean recall 0.94, and mean F1 0.95. Precision is consistently high (≥ 0.93), the property ensured by FDR control (Proposition 3.5), while recall is slightly lower in settings where certain causal dimensions have subtle interventional effects (0.85 for `walker_walk_hard`, 0.75 for `hopper_hop`). This asymmetry is favorable in practice: including a few extra distractor dimensions has a minor effect on downstream RL, whereas missing a critical causal dimension could be catastrophic. Empirically, IBD’s downstream return is oracle-level despite imperfect recall, suggesting that the occasionally missed dimensions carry limited task-relevant information. Boundary accuracy is also stable across distractor counts (e.g., `cheetah_run` F1: 0.99/0.97/0.98 at easy/medium/hard), because the per-dimension test is independent of how many other dimensions exist.

5.5 Algorithm-Agnostic Validation

Table 3: Algorithm-agnostic validation. IBD improves over Full State with both SAC and TD3 backends.

Environment	Algorithm	Full State	IBD (ours)	Oracle
walker_walk	SAC	414 \pm 144	842\pm121	785 \pm 128
walker_walk	TD3	142 \pm 131	734\pm208	781 \pm 116
cheetah_run	SAC	63 \pm 12	436\pm70	472 \pm 35
cheetah_run	TD3	67 \pm 12	356\pm80	425 \pm 24

To verify that IBD’s gains are not artifacts of SAC’s entropy regularization, we repeat the hard-distractor experiments with TD3 [Fujimoto et al., 2018] as the downstream algorithm (Table 3). The IBD advantage persists and in some cases strengthens. On `walker_walk`, IBD improves over Full State by 5.2 \times with TD3 (vs. 2.0 \times with SAC). On `cheetah_run`, the improvement is 5.3 \times (vs. 6.9 \times).

Notably, TD3 without masking degrades more severely than SAC: `walker` Full State drops from 414 (SAC) to 142 (TD3). This is expected; TD3 lacks SAC’s entropy regularization, making it more sensitive to the noise injected by high-dimensional irrelevant inputs. The implication is that IBD’s causal masking is most valuable for algorithms that lack built-in robustness to irrelevant features. IBD combined with TD3 reaches 94% (`walker`) and 84% (`cheetah`) of oracle performance, closely matching the SAC results and confirming that the mask quality is independent of

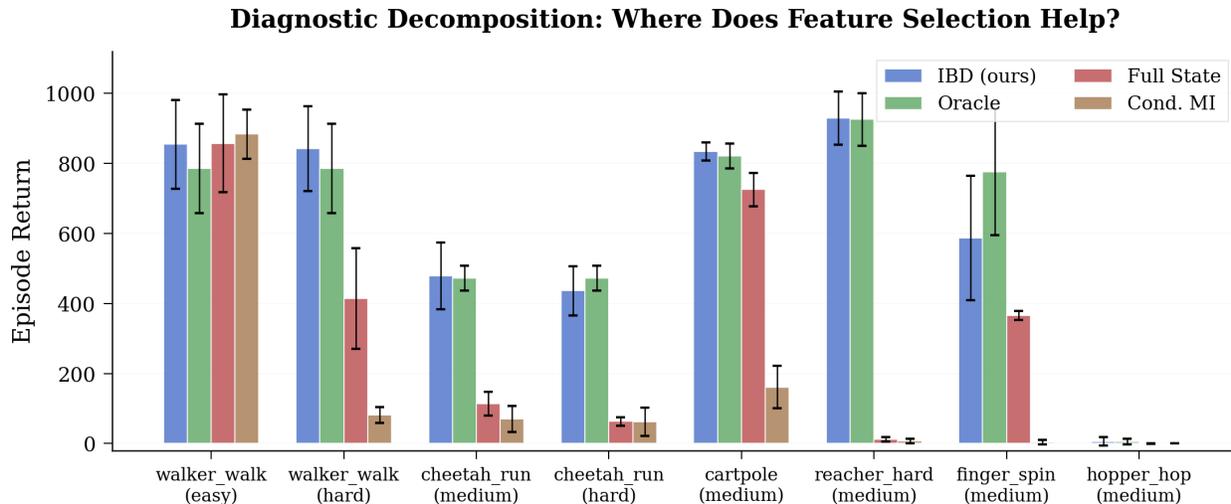


Figure 3: **Diagnostic decomposition.** Episode return across 8 representative settings. Three regimes emerge. (1) `walker_walk` (easy): all methods perform similarly, so no selection is needed. (2) Most medium/hard settings: Oracle \gg Full State but IBD \approx Oracle, indicating that distractors are the bottleneck and IBD resolves it. (3) `hopper_hop`: all methods near zero, indicating that the bottleneck is exploration, not feature selection.

the downstream algorithm. We note that this validation covers two environments under hard distractors; broader algorithm-agnostic evaluation across all settings is a natural direction for future work.

5.6 Robustness: Partial Controllability

The main experiments use purely exogenous distractors: no causal path from actions to distractors exists. A natural question is whether IBD can detect dimensions with partial controllability, i.e., dimensions whose dynamics mix action-dependent and exogenous components. We construct such dimensions with dynamics

$$x_{t+1} = \alpha \cdot g(\mathbf{s}_t, \mathbf{a}_t) + (1 - \alpha) \cdot z_t, \quad (3)$$

where g is a nonlinear function of state and action (integrated via leaky dynamics), z_t is an exogenous OU process, and $\alpha \in [0, 1]$ controls the causal mixing coefficient. We augment `cheetah_run` and `walker_walk` with 6 partially controllable dimensions and 20 exogenous distractors, and sweep $\alpha \in \{0, 0.02, 0.05, 0.1, 0.15, 0.2, 0.3, 0.5, 0.7, 1.0\}$ across 3 seeds. The results reveal a sharp detection threshold: at $\alpha = 0$, recall is correctly zero, but by $\alpha \approx 0.05$, IBD reliably detects all 6 partial dimensions. Precision remains ≥ 0.92 across all α values, confirming that this sensitivity does not inflate false positives. Full details are in Section B.

5.7 When IBD Does Not Help

We report two settings where IBD does not outperform Full State, and use these to illustrate IBD’s diagnostic value (Figure 3). On `walker_walk` with medium distractors, Full State achieves 902 ± 26 while IBD achieves 792 ± 112 . With 24 true dimensions and 50 distractors (ratio $\sim 2:1$, below the threshold identified in Section 5.2), SAC’s MLP has sufficient capacity to handle the full 74-dimensional input. IBD’s recall of 0.86 means 3–4 true dimensions are occasionally missed, causing a slight performance dip. The diagnostic interpretation: this environment does not need feature selection. On `hopper_hop` with medium distractors, all methods achieve near-zero returns: Full State 0.3, IBD 6.1, Oracle 6.1. IBD matches the oracle, so feature selection is not the bottleneck. The bottleneck is exploration; `hopper_hop` is an intrinsically hard task where even perfect state information does not yield good performance within 300K steps. The practitioner’s correct response is to invest in exploration strategies, not in better feature selection. This decomposition provides actionable information about the source of an environment’s difficulty. When Oracle \gg Full State but IBD \approx Oracle, the problem is representational confusion from distractors, and IBD solves it. When Oracle \approx Full State \approx IBD, there are few distractors and no action is needed. When all three are poor, the bottleneck lies elsewhere—exploration, reward design, or training budget—and feature selection is irrelevant.

6 Discussion

From representation learning to causal identification. Existing approaches have predominantly framed the distractor problem as requiring learned representations: invariant encoders, contrastive objectives, and causal world models. Our results suggest that when the goal is feature selection on a structured state vector, the problem can be productively reframed as causal identification, and the agent’s action space doubles as the intervention tool needed to solve it. An important nuance: our distractors are appended dimensions that do not corrupt the true state or bias the reward. In principle, a sufficiently large network with enough training could learn to ignore them. The degradation we observe reflects a practical failure of fixed-capacity function approximators under high-dimensional irrelevant input; IBD sidesteps this capacity-vs-distractor tradeoff entirely. Its value is therefore best understood as improving sample efficiency and practical robustness, rather than solving a problem that is information-theoretically impossible for full-state methods.

Complementarity with learned encoders. Recent work on causal world models [Schölkopf et al., 2021] and unified CRL frameworks [Yao et al., 2025] aims to discover the full causal structure of an environment. IBD produces a binary mask, not a graph; a suitable level of abstraction for the feature selection problem. Importantly, the two approaches compose naturally. A visual encoder (e.g., from DrQ-v2 or CURL) maps pixels to a feature vector $\mathbf{z} \in \mathbb{R}^k$, but has no explicit incentive to exclude latent dimensions that are exogenous to the agent’s actions: contrastive or reconstruction objectives preserve all predictable variation, not only action-caused variation. IBD can operate directly on \mathbf{z} as a post-encoder selector, using the same two-sample test to identify which latent dimensions respond to action interventions. This requires no changes to IBD’s algorithm: the statistical test applies to any flat feature space regardless of its origin. Analogously, EX-BMDP methods [Efroni et al., 2022, Levine et al., 2025a] separate endogenous from exogenous latent factors via inverse dynamics; IBD achieves the same separation via direct intervention, making fewer modeling assumptions. If a practitioner additionally needs the causal graph, IBD’s per-action variant can serve as a warm start by identifying the relevant dimensions before graph discovery.

Limitations. Our main benchmark uses purely exogenous distractors; the partial-controllability experiment (Section 5.6) demonstrates that IBD handles mixed dynamics, but settings with more complex causal structure, such as dimensions that mediate between actions and distractors, or distractors whose statistics shift with the agent’s policy, remain open challenges. The observational baselines we compare against (MI, Variance, Cond. MI) span a range of sophistication; we also evaluate a gradient-based attribution baseline (Section F) and find that it degrades similarly under confounded distractors. The distractor-to-signal ratio at which full-state RL collapses is consistent across tasks in our benchmarks (Table 2), and we provide a capacity-based intuition for why this occurs; however, we expect the precise threshold to shift with network width, training budget, and task complexity. Systematic characterization of this dependence (for example, by sweeping architecture size jointly with distractor count) is a natural direction for future work. Our evaluation uses state-space observations where ground-truth causal labels are available; extending to the encoder-then-IBD pipeline on pixel-based tasks would further validate the compositional design described in Section 1.

7 Conclusion

This paper argued that feature selection for RL under confounded distractors is a causal identification problem, and that the agent’s own action space already provides the interventions needed to solve it. IBD operationalizes this idea with a structured random probe policy and a two-sample test, producing a binary mask that plugs into any downstream algorithm without learned models. Across 12 settings, the causal mask closely tracks oracle performance while observational baselines, including state-conditioned forward models, degrade under confounding. An unexpected finding is that a structured random probe consistently matches or outperforms a trained RL scout, suggesting that broad state coverage matters more than task-directed behavior for causal identification.

Two extensions follow naturally from this work. First, in visual RL pipelines a learned encoder maps pixels to a feature vector before any downstream processing; applying IBD’s two-sample test to that feature vector requires no algorithmic changes and would test whether interventional selection complements encoder-level robustness. Second, our analysis assumes stationary causal structure. Environments whose dynamics shift over time (due to wear, non-stationarity, or co-adaptation with other agents) would require periodic re-probing, and understanding how to amortize the probing cost across such shifts is an open problem.

References

- Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009.
- Michael Laskin, Aravind Srinivas, and Pieter Abbeel. CURL: Contrastive unsupervised representations for reinforcement learning. In *ICML*, 2020.
- Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Mastering visual continuous control: Improved data-augmented reinforcement learning. In *ICLR*, 2022.
- Nicklas Hansen, Hao Su, and Xiaolong Wang. Stabilizing deep q-learning with convnets and vision transformers under data augmentation. In *NeurIPS*, 2021.
- Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. CITRIS: Causal identifiability from temporal intervened sequences. In *ICML*, 2022.
- Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *ICML*, 2018.
- Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *ICML*, 2018.
- Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.
- Austin Stone, Oscar Ramirez, Kurt Konolige, and Rico Jonschkowski. The distracting control suite – a challenging benchmark for reinforcement learning from pixels, 2021. URL <https://arxiv.org/abs/2101.02722>.
- Yonathan Efroni, Dipendra Misra, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Provably filtering exogenous distractors using multistep inverse dynamics. In *International Conference on Learning Representations*, 2022.
- Alexander Levine, Peter Stone, and Amy Zhang. Learning a fast mixing exogenous block MDP using a single trajectory. In *International Conference on Learning Representations*, 2025a.
- Alexander Levine, Peter Stone, and Amy Zhang. Offline action-free learning of ex-bmdps by comparing diverse datasets. *Reinforcement Learning Journal*, 2025b.
- Kaichen Huang, Shenghua Wan, Minghao Shao, Hai-Hang Sun, Le Gan, Shuai Feng, and De-Chuan Zhan. Leveraging separated world model for exploration in visually distracted environments. In *Advances in Neural Information Processing Systems*, 2024.
- Tongzhou Wang, Simon S Du, Antonio Torralba, Phillip Isola, Amy Zhang, and Yuandong Tian. Denoised MDPs: Learning world models better than the world itself. In *International Conference on Machine Learning*, 2022.
- Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.
- Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. Causal representation learning for instantaneous and temporal effects in interactive systems. In *ICLR*, 2023.
- Dingling Yao, Dario Rancati, Riccardo Cadei, Marco Fumero, and Francesco Locatello. Unifying causal representation learning with the invariance principle. In *International Conference on Learning Representations*, 2025.
- Burak Varici, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. Linear causal representation learning from unknown multi-node interventions. In *Advances in Neural Information Processing Systems*, 2024.
- Burak Varici, Emre Acartürk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based causal representation learning: Linear and general transformations. *Journal of Machine Learning Research*, 26(112):1–90, 2025.
- Mingxuan Li, Junzhe Zhang, and Elias Bareinboim. Confounding robust deep reinforcement learning: A causal approach. In *Advances in Neural Information Processing Systems*, 2025.
- Alexander S Klyubin, Daniel Polani, and Christopher L Nehaniv. Empowerment: A universal agent-centric measure of control. In *IEEE Congress on Evolutionary Computation*, 2005.
- Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. In *NeurIPS*, 2015.
- Rudolf E Kalman. On the general theory of control systems. *IFAC Proceedings Volumes*, 1(1):491–502, 1960.
- Jacob E. Kooi, Mark Hoogendoorn, and Vincent François-Lavet. Disentangled (un)controllable features. In *IEEE Symposium Series on Computational Intelligence (SSCI)*, 2023.

Lihong Li, Thomas J Walsh, and Michael L Littman. Towards a unified theory of state abstraction for MDPs. In *ISAIM*, 2006.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995.

Belinda Phipson and Gordon K Smyth. Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1), 2010.

A Distractor Environment Design

Our distractor benchmark uses three types of distractors, each designed to defeat a specific class of feature selection methods.

Autonomous distractors are Ornstein-Uhlenbeck processes and coupled oscillators that evolve independently of the agent’s actions and state. Their variance and frequency content differ from true state dimensions, making them easy to filter by any method that examines statistical properties beyond raw magnitude.

Mimicking distractors are exogenous OU processes calibrated to reproduce the statistical fingerprint of true proprioceptive dimensions. The calibration targets scale, variance, autocorrelation, and frequency content, using $\sigma_{\text{ref}} = 0.5 + 0.05 \cdot d_c$ so that each task receives appropriately scaled distractors (*walker*: $\sigma_{\text{ref}} = 1.7$; *cartpole*: $\sigma_{\text{ref}} = 0.75$). These defeat variance-based selection because they are indistinguishable from true dimensions in their marginal distribution.

Reward-correlated distractors follow drift processes whose rate is correlated with episode progress (distance traveled, target proximity) but are not causally influenced by actions. They defeat MI-based selection because the correlation with reward creates spurious mutual information between these distractors and the agent’s action sequence.

The composition at each difficulty level is as follows. Easy (6 total): 4 OU + 2 oscillator. Medium (50 total): 12 autonomous + 28 mimicking + 6 reward-correlated + 4 oscillator. Hard (100 total): 18 autonomous + 60 mimicking + 14 reward-correlated + 8 oscillator. Only interventional methods are robust to all three types, because the do-operator reveals that no causal path from actions to any distractor exists regardless of statistical similarity.

B Partial Controllability: Experimental Details

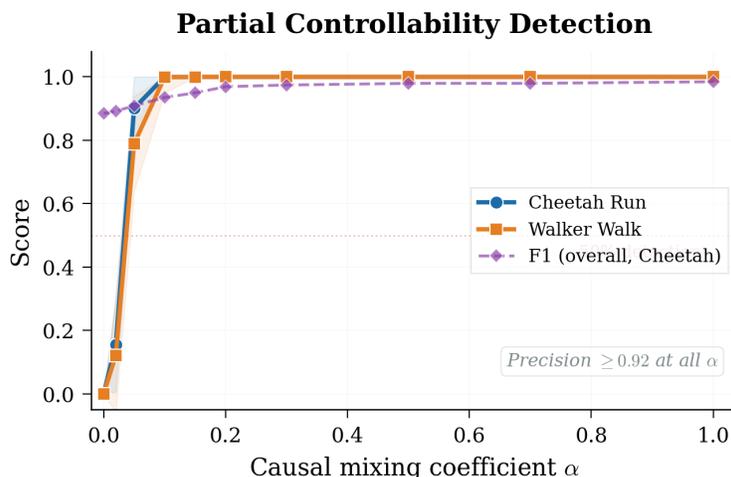


Figure 4: **Partial controllability detection.** Recall on partially controllable dimensions as a function of mixing coefficient α , for *cheetah_run* and *walker_walk*. At $\alpha = 0$ (purely exogenous), recall is correctly zero; by $\alpha \approx 0.05$ ($\sim 5\%$ causal variance), recall reaches 1.0. The dashed line shows overall F1 on *cheetah_run*, confirming that boundary quality remains high. Precision ≥ 0.92 at all α values. Shaded regions: ± 1 std over 3 seeds.

For the robustness study (Section 5.6), each partially controllable dimension follows the dynamics in Equation (3), where the causal component g maps a randomly-assigned action dimension through a nonlinear transformation (\tanh

with per-dimension gain and bias), integrated via leaky dynamics with rate constant $50 \times dt$ to ensure the causal signal is on the same scale as the exogenous OU noise. The exogenous component z_t is an independent OU process with $\tau \in [1, 4]$ and $\sigma \in [0.2, 0.6]$, matching the scale of DMControl proprioceptive dimensions.

Each configuration adds 6 partially controllable dimensions and 20 exogenous distractors (10 autonomous OU + 10 mimicking) to the base DMControl task. We sweep 10 values of α across 3 seeds per task, for a total of 60 IBD probes. Each probe takes approximately 2–3 minutes (trajectory collection: 160 trajectories \times 200 steps; statistical testing: < 1 s).

C Scout Policy Ablation

We ablate the scout training budget across $\{0, 10\text{K}, 20\text{K}, 40\text{K}, 80\text{K}, 160\text{K}\}$ steps on three tasks with medium distractors, using 3 seeds per configuration. Budget = 0 corresponds to IBD’s structured random probe policy, which requires no RL training and no GPU.

Table 4: Scout ablation (P/R/F1, mean over 3 seeds). Budget = 0: structured random probe; $\geq 10\text{K}$: SAC scout. Boundary accuracy is identical or better without RL pre-training.

Task	Scout Budget	Precision	Recall	F1
cheetah_run	0	0.944	1.000	0.971
	80K	0.944	1.000	0.971
reacher_hard	0	0.952	1.000	0.974
	80K	0.952	1.000	0.974
walker_walk	0	0.910	0.972	0.940
	80K	0.902	0.889	0.895

On cheetah_run and reacher_hard, boundary accuracy is identical across all budgets (intermediate values omitted for space; all match budget = 0). On walker_walk, the untrained policy achieves the highest F1 (0.94 vs. 0.90 at 80K), likely because a trained policy converges toward a narrow behavioral mode that reduces state coverage. These results confirm that IBD’s practical overhead reduces to trajectory collection alone; approximately 32K environment steps and under 3 minutes, with no GPU requirement.

D Boundary Discovery Accuracy: Full Results

Table 5 reports IBD’s per-setting precision, recall, and F1. Precision is ≥ 0.93 in every setting, confirming that FDR control effectively prevents distractor leakage. Recall reaches 1.00 on most tasks; the two exceptions are walker_walk (0.85–0.92, where some proprioceptive dimensions have weak interventional signal) and hopper_hop (0.75, an intrinsically difficult exploration task). Boundary accuracy is stable across distractor counts: for reacher_hard, F1 is 0.97 (easy), 0.98 (medium), and 0.93 (hard).

E Hyperparameters

Table 6: Hyperparameters used in all experiments.

Component	Parameter	Value
SAC / TD3	LR / batch / buffer / arch	3×10^{-4} / 256 / 300K / MLP [256, 256]
TD3 only	Action noise σ	0.1
IBD Probe	Trajectories N / length T / horizons \mathcal{H} / α	80 / 200 / {1,5,10} / 0.05
Training	Steps / eval freq / eval eps / seeds	300K / 50K / 10 / {42,142,242,342,442}
Cond. MI	Arch / epochs / LR / batch / data	MLP [64,64] / 50 / 10^{-3} / 2048 / 200 \times 200
Grad. Attr.	Arch / epochs / LR / batch / data	MLP [128,128] / 80 / 10^{-3} / 2048 / 200 \times 200
Robustness	α / partial dims / exo distr / seeds	{0...1.0} / 6 / 20 / {42,142,242}

Table 5: IBD boundary discovery accuracy (P/R/F1, mean over 5 seeds). High precision = few distractor dims leaked; high recall = few true dims missed.

Environment	Distr.	Precision	Recall	F1
walker_walk	easy	0.98	0.92	0.95
walker_walk	medium	0.95	0.86	0.90
walker_walk	hard	0.97	0.85	0.91
cheetah_run	easy	0.98	1.00	0.99
cheetah_run	medium	0.94	1.00	0.97
cheetah_run	hard	0.96	1.00	0.98
reacher_hard	easy	0.94	1.00	0.97
reacher_hard	medium	0.97	1.00	0.98
reacher_hard	hard	0.86	1.00	0.93
cartpole_swingup	medium	0.93	1.00	0.96
finger_spin	medium	0.96	1.00	0.98
hopper_hop	medium	0.96	0.75	0.84

F Gradient Attribution Baseline

To test whether a stronger model-based observational method can succeed where MI and Cond. MI fail, we evaluate a gradient attribution baseline. This method trains a joint forward dynamics model $f(\mathbf{s}, \mathbf{a}) \rightarrow \mathbf{s}'$ (MLP [128, 128], 80 epochs) and scores each observation dimension i by the mean absolute gradient of the predicted next state with respect to actions: $\text{score}_i = \mathbb{E}_{\text{data}}[\|\partial f_i(\mathbf{s}, \mathbf{a}) / \partial \mathbf{a}\|_1]$. This directly measures how sensitive the learned dynamics are to the action input; a direct model-based approach to identifying action-influenced dimensions.

However, gradient attribution remains observational: when mimicking distractors co-vary with true state dimensions (which co-vary with actions), the learned model develops spurious gradient paths $\mathbf{a} \rightarrow \mathbf{s} \rightarrow \text{distractor}$, assigning non-zero sensitivity scores to dimensions that $\text{do}(\mathbf{a})$ does not causally affect.

Table 7: Gradient attribution baseline (episode return, mean \pm std over 5 seeds, medium distractors). Grad. Attr. performs comparably to or worse than Cond. MI, confirming that the failure of observational methods under confounding is not resolved by gradient-based sensitivity analysis.

Environment	Full State	Cond. MI	Grad. Attr.	IBD (ours)	Oracle
cartpole_swingup	725 \pm 48	161 \pm 60	56 \pm 32	834 \pm 26	821 \pm 35
finger_spin	365 \pm 13	4 \pm 6	62 \pm 113	587 \pm 178	775 \pm 181
hopper_hop	0 \pm 1	0 \pm 0	0 \pm 0	6 \pm 12	6 \pm 8
reacher_hard	12 \pm 7	7 \pm 6	5 \pm 10	929 \pm 76	925 \pm 75

Table 7 shows results on four tasks with medium distractors (50 distractor dimensions). Gradient attribution performs comparably to or worse than Cond. MI across all settings. On `reacher_hard`, it achieves a return of only 5 (vs. IBD’s 929), and on `cartpole_swingup` it scores 56, below even the naive MI and Variance baselines (78 and 83, respectively; Table 1). The boundary discovery F1 scores are also low (0.44–0.50), indicating that gradient attribution selects roughly half distractors and half true dimensions. These results confirm that the difficulty observational methods face under confounded distractors persists across model classes and sensitivity measures, reinforcing the case for interventional identification via the do-operator.