Discovering What You Can Control: Interventional Boundary Discovery for Reinforcement Learning

Jiaxin Liu
University of Illinois Urbana-Champaign
jiaxin26@illinois.edu

Anzhe Cheng
University of Southern California
anzheche@usc.edu

Paul Bogdan
Ming Hsieh Department of Electrical and Computer Engineering
Thomas Lord Department of Computer Science
Center for Complex Particle Systems (COMPASS)
University of Southern California, Los Angeles, USA
pbogdan@usc.edu

github.com/jiaxin26/IBD-RL

Abstract

When an RL agent's observations contain distractors driven by the same confounders as its true state, observational data alone cannot identify which dimensions the agent controls. In our benchmarks, even state-conditioned observational selectors can collapse when distractors mimic controllable state variables. We propose Interventional Boundary Discovery (IBD), which treats the agent's own action channel as a source of randomized interventions: randomizing actions implements an interventional contrast, and per-dimension two-sample tests with FDR correction produce a binary mask over observation dimensions. Across 12 continuous-control settings with up to 100 distractors, IBD matches oracle return in 11 of 12 settings, while observational baselines including mutual information, state-conditioned forward models, and gradient-based sensitivity often underperform simply passing the full observation to SAC.

1 Introduction

Reinforcement Learning (RL) has achieved strong results in continuous control and sequential decision making [1, 2], with modern agents solving increasingly challenging tasks from robotics to simulated benchmarks [3–5]. A common assumption behind this progress is that the agent receives Markov or near-Markov observations that are sufficiently informative about the underlying state [6–8]. In practice, observations often mix truly controllable state variables with many irrelevant dimensions (nuisance signals, background dynamics, sensor artifacts, and exogenous processes), a problem widely recognized in both visual RL [9–12] and state-space RL [13, 14]. Such distractors are especially harmful when they are not merely random noise, but statistically resemble the true state or correlate with the agent's behavior. In this case, a policy trained on the full observation can spend capacity modeling non-causal variation, while a passive feature selector can retain dimensions that look predictive but are outside the agent's control.

This failure is not merely a consequence of high-dimensional inputs. The more difficult case arises when irrelevant dimensions are statistically entangled with the agent's behavior. For example, an

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

exogenous coordinate may change over the same time scale as the task state, correlate with episode progress, or share latent causes with the agent's actions [14]. Such a dimension can therefore appear useful under passive observation, even though intervening on the agent's action would not change its distribution. Conversely, a truly action-reachable dimension may have a weaker observational association if its effect is delayed, stochastic, or only visible under particular action sequences. Thus, passive feature selection can confuse correlation with controllability: it may retain dimensions that co-vary with actions while discarding dimensions that are actually downstream of actions. This creates an identifiability gap for observation selection in RL, where the goal is not only to predict the next observation but to identify which parts of the observation lie within the agent's causal reach.

Existing methods do not directly resolve this distinction (Section 2). Visual-RL methods such as SVEA, DrQ-v2, and CURL [10–12] suppress pixel-level nuisance variation but do not test whether a retained feature is causally influenced by actions, and exogenous-block and Denoised MDPs [13–16] still compute their selection signal from observational trajectories, where a distractor that co-varies with the true state can look like a controllable dimension. Empowerment and causal representation learning [17–19] either aggregate action influence or target richer latent structure, and neither answers which observed dimensions actions can reach.

We propose **Interventional Boundary Discovery (IBD)**, a model-free procedure for identifying the agent's *Causal Sphere of Influence* (SoI): the set of observed dimensions whose future distribution changes under an intervention on the agent's actions. IBD uses the agent's own action channel as an intervention mechanism. During a short probing phase, it randomizes actions to approximate an action intervention, compares baseline and intervention trajectory statistics with per-dimension two-sample tests, and applies false-discovery-rate (FDR) control [20]. The result is a binary mask over observation dimensions. This mask can preprocess the input to downstream RL algorithms such as SAC [3] and TD3 [4], without requiring a learned world model, latent graph recovery, or changes to the policy optimizer. IBD also provides a diagnostic view of when feature selection matters. Comparing downstream performance under the Full State, IBD mask, and Oracle mask separates different failure modes. If IBD closes the Full State–Oracle gap, the main bottleneck is distractor-induced; if the gap is small, feature selection is not the limiting factor; if all three fail, the bottleneck lies beyond observation masking. Across 12 continuous-control settings, IBD recovers near-oracle masks and improves downstream learning when observational baselines are misled by spuriously correlated distractors, and it stays accurate when a genuine latent confounder is added (Section 4.6).

The primary contributions of this work are summarized as follows:

- We formalize the *Causal Sphere of Influence* in RL and show that observational feature selection has a fundamental identifiability gap: confounded distractors can rank above truly action-reachable dimensions.
- We introduce **Interventional Boundary Discovery**, a model-free procedure using randomized action interventions, per-dimension hypothesis testing, and FDR control to recover a binary mask over controllable dimensions.
- Across 12 continuous-control settings, IBD delivers near-oracle downstream RL performance where observational baselines often fail, while also serving as a diagnostic for separating distractor-induced failures from other RL bottlenecks.

2 Related Work

Distractor-Robust Reinforcement Learning. Distractors degrade RL agents through both representation learning and optimization difficulty [9]. In visual RL, CURL [12], SVEA [10], and DrQ-v2 [11] reduce sensitivity to pixel-level nuisance variation through contrastive objectives and strong augmentation. In state-space RL, Denoised MDPs [14] and the Exogenous Block MDP framework [13], with extensions to single-trajectory and action-free settings [15, 16], separate endogenous from exogenous factors via inverse dynamics or latent decomposition. These methods improve robustness, but their selection signal is computed from observational trajectories: a distractor whose dynamics co-vary with action-reachable state can still appear useful for inverse dynamics or forward prediction even when no intervention on the action affects it. IBD replaces this passive signal with an active interventional contrast that severs confounding paths at the action node.

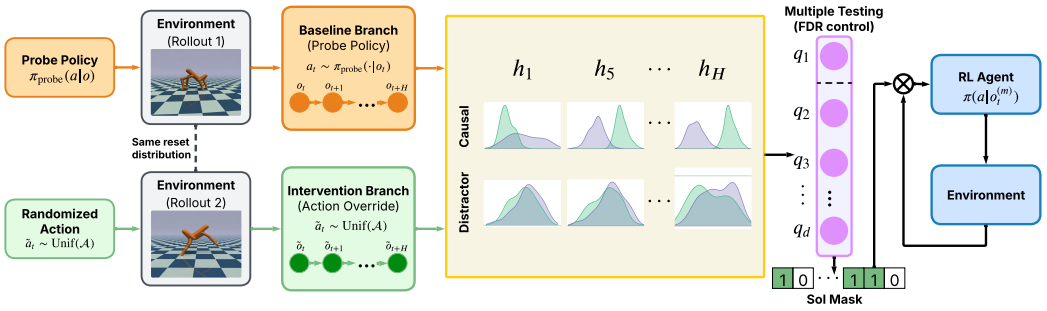


Figure 1: **Overview of Interventional Boundary Discovery (IBD).** From the same reset distribution, IBD collects two independent rollouts: a baseline branch under a probe policy π_{probe}, and an intervention branch in which the action mechanism is replaced by a confounder-independent randomized policy, $\tilde{a}_t \sim \text{Unif}(\mathcal{A})$. For each observation dimension and horizon, IBD compares trajectory statistics across the two branches using a two-sample test. Benjamini–Hochberg FDR correction then yields a binary SoI mask, which filters observations before downstream RL training.

Causal Representation Learning. Causal representation learning aims to recover latent causal variables from temporal observations [17]. CITRIS [21] and iCITRIS [22] identify latent variables under known intervention targets, with extensions to invariance principles [18] and broader intervention classes [19, 23]. These methods recover a richer object than feature selection requires (a latent graph rather than a mask over observed dimensions), typically under assumptions on intervention labels, temporal variation, or latent transition structure. IBD targets the narrower question of dimension-level action reachability in observed space, requires no latent recovery, and can serve as a pre-filter before more detailed graph-level analysis.

Control, Empowerment, and Feature Selection. Empowerment [24, 25] measures channel capacity between actions and future states, but as an aggregate scalar rather than a dimension-level criterion: high $I(A; S' \mid S)$ does not imply that a specific observed dimension is causally affected by actions. Classical controllability analysis [26] provides a precise model-based notion of reachability for linear systems with known dynamics; recent work on interpretable controllability features [27] routes through learned latent representations rather than direct observed-space testing. Classical feature-selection criteria (mutual information, variance, forward-model sensitivity) form natural baselines but conflate statistical and causal relevance: a distractor highly correlated with actions through shared latent causes or behavioral progress can rank highly without being causally influenced. State abstraction [28] groups states by optimal-policy equivalence, a different objective. IBD addresses dimension-level action reachability directly through interventional contrasts in the observed space, requiring neither a known model nor a learned latent representation.

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 $\mathbf{o}_t = [\mathbf{s}_t^{\text{c}}, \mathbf{s}_t^{\text{d}}]$: a causal component $\mathbf{s}^{\text{c}} \in \mathbb{R}^{d_c}$ that is causally downstream of the agent's actions $\mathbf{a}_t \in \mathcal{A} \subseteq \mathbb{R}^{d_a}$, and a distractor component $\mathbf{s}^{\text{d}} \in \mathbb{R}^{d_d}$ 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 $o_{t+h}^{(i)}$:

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

for some intervention value u, where $\text{do}(\cdot)$ denotes Pearl's do-operator [29].

The definition says that 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. Note that (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 $\mathbb{P}(o^{(i)}|\mathrm{do}(a = u)) \neq \mathbb{P}(o^{(i)})$, because the do-operator severs the confounding path.

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

Assumptions. IBD relies on four assumptions: (i) decomposability: dimensions are either fully causal or fully exogenous (relaxed in Section 4.8 to partially controllable dimensions); (ii) action overrideability during probing, which is natural in simulation and requires a safe-exploration protocol on hardware; (iii) stationary causal structure during probing; (iv) faithfulness and statistic sensitivity: a directed action-to-dimension path produces a non-trivial distributional shift that is detectable by the chosen summary Δ_i^h for some $h \in H$ (Proposition 3.4). The default Δ_i^h is insensitive to dimensions that are action-reachable only through a shift in mean level; taking the union of Δ_i^h tests and per-trajectory mean tests recovers these (Appendix J).

3.2 Algorithm

IBD estimates the SoI by an interventional contrast: trajectories collected under a probe policy are compared, dimension by dimension, against trajectories in which the action channel is randomized. Dimensions outside the SoI remain distributionally invariant under this contrast up to sampling noise, while descendants of actions exhibit a measurable distributional shift at some horizon.

From the same reset distribution we collect N baseline trajectories $\{\tau_k^{\mathrm{b}}\}_{k=1}^N$ under probe policy π_{probe}, and N intervention trajectories $\{\tau_k^{\mathrm{int}}\}_{k=1}^N$ in which each action is replaced by an i.i.d. draw from $\mathrm{Unif}(\mathcal{A})$. This implements a stochastic intervention by replacing the behavior-policy mechanism with the confounder-independent distribution $q(\mathbf{a}) = \mathrm{Unif}(\mathcal{A})$; Uniform is maximum-entropy on $\mathcal{A}$ and maximally excites action-reachable dimensions, but any confounder-independent distribution is valid. Our default π_{probe} is a structured random policy with sinusoidal actions and weak state feedback, requiring no RL training (Section 3.5).

For each dimension i and horizon $h \in \mathcal{H}$ (we use $\mathcal{H} = \{1, 5, 10\}$), each trajectory is summarized by the mean absolute h-step difference,

$$\Delta_i^h(\tau) \;=\; \frac{1}{K_h} \sum_{r=0}^{K_h-1} \left| o^{(i)}_{(r+1)h} - o^{(i)}_{rh} \right|, \qquad K_h = \lfloor T/h \rfloor, \tag{2}$$

yielding N scalars per branch that are i.i.d. across trajectories. Under this randomized-action intervention, an exogenous dimension has an action-invariant transition kernel and Δ_i^h has the same distribution under both branches; an action-reachable dimension exhibits a shifted distribution. Multiple horizons are tested because a causal chain of length k may produce no detectable effect until $h \geq k$. Since the final object is a dimension-level mask rather than a list of (i, h) discoveries, we aggregate the $|\mathcal{H}|$ horizon-level p-values within each dimension via the Bonferroni minimum $\tilde{p}_i = \min\{1,\ |\mathcal{H}| \min_{h \in \mathcal{H}} p_{i,h}\}$, then apply Benjamini–Hochberg correction [20] across the d dimension-level p-values at level $\alpha = 0.05$. The two samples are compared with a Welch t-test. The output is a binary mask $\mathbf{m} \in \{0, 1\}^d$ with $m_i = \mathbf{1}[q_i < \alpha]$, computed once at a cost of $2NT$ environment steps and reused throughout downstream training. Algorithm 1 summarizes the procedure.

Algorithm 1 Interventional Boundary Discovery (IBD)

Require: Environment $\mathcal{M}$, probe policy π_{probe} (default: structured random); trajectory count N and length T (default N=80, T=200); horizons $\mathcal{H}$ (default $\{1, 5, 10\}$); significance α (default 0.05)

Collect baseline $\{\tau_k^{\text{b}}\}_{k=1}^N$ using π_{probe} and intervention $\{\tau_k^{\text{int}}\}_{k=1}^N$ using $\text{do}(\mathbf{a} \sim \text{Unif}(\mathcal{A}))$
for each dimension $i \in [d]$ **do**
  for each horizon $h \in \mathcal{H}$ **do**
    Compute $\{\Delta_i^h(\tau_k^{\text{b}})\}_{k=1}^N$ and $\{\Delta_i^h(\tau_k^{\text{int}})\}_{k=1}^N$ via Eq. 2
    Obtain $p_{i,h}$ from a Welch t-test on the two samples
  end for
  $\tilde{p}_i \leftarrow \min\{1,\ |\mathcal{H}| \min_{h \in \mathcal{H}} p_{i,h}\}$ (Bonferroni over horizons)
end for
Apply Benjamini–Hochberg to $\{\tilde{p}_i\}_{i=1}^d$ at level α, yielding $\{q_i\}_{i=1}^d$
return Binary mask $\mathbf{m}$ with $m_i = \mathbf{1}[q_i < \alpha]$

3.3 Failure of Observational Selection

Before presenting the theoretical guarantees of IBD, we give a construction illustrating a failure mode of observational criteria. Natural dependence-based criteria, mutual information in particular, can fail to separate a confounded distractor from the true causal dimension it tracks, regardless of how well the dependence is estimated.

Proposition 3.2 (Observational dependence can rank confounded distractors above causal dimensions)**.** *A distractor can have higher mutual information with behavior actions than a causal state dimension, even though actions have zero causal effect on it.*

Construction. Consider the linear-Gaussian SCM

$$C_t \sim \mathcal{N}(0,1), \qquad a_{1,t} = C_t + \xi_t, \qquad s_{1,t+1} = \beta a_{1,t} + \eta_t, \qquad d_{k,t} = C_t + \varepsilon_t,$$

where $\beta \neq 0$, with $\xi_t, \eta_t, \varepsilon_t$ independent noises. The distractor has no incoming edge from $\mathbf{a}_t$ or s_t; future distractors are generated autonomously by the same exogenous mechanism. Since both $a_{1,t}$ and $d_{k,t}$ share the latent cause C_t, $\text{MI}(d_{k,t}; \mathbf{a}_t)$ can exceed $\text{MI}(s_{1,t+1}; \mathbf{a}_t)$ by taking $\sigma_\xi, \sigma_\varepsilon$ small and the transition noise in s_1 larger. Thus an MI selector can rank d_k above the causal dimension.

However, intervening on the action severs the edge $C_t \to a_{1,t}$ but leaves the autonomous mechanism for d_k unchanged. Hence, for every $h \geq 1$ and intervention value u,

$$P(d_{k,t+h} \mid \text{do}(\mathbf{a}_t = u)) = P(d_{k,t+h}),$$

whereas $P(s_{1,t+1} \mid \text{do}(a_{1,t} = u))$ changes with u whenever $\beta \neq 0$. Therefore observational dependence alone does not identify the causal Sphere of Influence. □

This construction isolates the key failure mode: a dimension can share a latent cause with the behavior action and therefore appear highly action-informative in observational data, while remaining outside the causal downstream of any action intervention. Our benchmark (Section 4.1) instantiates related patterns: mimicking dimensions whose marginal statistics match true proprioceptive channels, and reward-correlated dimensions that drift with episode progress. In both cases, the distractors can be statistically predictive under the behavior distribution without lying on a causal path from actions to future observations. Section 4 shows that observational baselines, including a state-conditioned forward model, drop to near-random return on the harder regimes, while IBD does not. Section 4.6 instantiates the construction itself with a genuine latent confounder.

3.4 Identification Properties

The two-sample test in IBD inherits standard properties from causal inference and multiple testing: the confounding path is severed at the action node, a causal path still produces a detectable shift under faithfulness, and FDR control bounds the rate of spurious inclusions.

Proposition 3.3 (Invariance to confounders)**.** *Because* IBD *collects interventional data by replacing the action mechanism with a confounder-independent randomized policy, all edges $C \to \mathbf{a}$ from confounders C to actions are severed under the intervention branch. In the population limit, the distribution of the test statistic on non-SoI dimensions is therefore invariant to the presence or absence of confounders, and* $\widehat{\text{SoI}}$ *converges to the same set whether confounders are active or removed (up to sampling noise in finite samples).*

Proposition 3.4 (Interventional detectability)**.** *Assume (i) faithfulness: 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 $\mathrm{do}(a_j = u)$ depends non-trivially on u; and (ii) statistic sensitivity: this distributional change is not annihilated by the chosen summary Δ_i^h. Then a directed causal path of length $\leq h$ from a_j to o_i is a sufficient condition for the distribution of Δ_i^h to differ between the baseline and intervention branches. Conversely, in the absence of any such path, Δ_i^h has the same distribution under both branches.*

Proposition 3.5 (Type I and FDR control)**.** *The Welch t-test on trajectory-level summaries $\{\Delta_i^h(\tau_k^b)\}, \{\Delta_i^h(\tau_k^{\mathrm{int}})\}$ controls per-pair Type I error asymptotically under the central limit theorem. After Bonferroni aggregation over horizons within each dimension, $\tilde{p}_i = \min\{1, |H| \min_{h \in H} p_{i,h}\}$, the resulting dimension-level p-values are valid under the null. Applying Benjamini–Hochberg correction to $\{\tilde{p}_i\}_{i=1}^d$ controls the dimension-level false discovery rate at level α under the standard independence or PRDS dependence conditions.*

The PRDS condition underlying Benjamini–Hochberg need not hold when mimicking distractors are statistically coupled to true dimensions; replacing BH with the Benjamini–Yekutieli procedure [30], which is valid under arbitrary dependence, costs no F1 on our benchmark (Appendix K).

These results address the two failure modes of selection: distractor leakage is controlled by FDR, while missed causal dimensions vanish asymptotically under faithfulness when the tested horizons cover the relevant causal delay and the intervention effect is detectable.

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 is the structured random policy of Section 3.2, which sufficed for all tasks in our experiments. A scout-budget sweep (Appendix C) shows identical boundary accuracy on `cheetah_run` and `reacher_hard` between the zero-budget structured probe and an 80K-step SAC scout, and higher F1 for the untrained probe on `walker_walk` (0.94 vs. 0.90), indicating that IBD does not require a learned exploratory policy as a prerequisite. With our default settings ($N = 80$, $T = 200$), the probing phase requires approximately 32K environment steps and is a one-time cost: the mask is computed once and reused for the entire downstream training run.

Table 1: Main results (episode return, mean ± 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±139	854±127	785±128	705±98	750±103	**884±70**
walker_walk	medium	**902±26**	792±112	785±128	129±79	84±6	220±135
walker_walk	hard	414±144	**842±121**	785±128	87±34	46±10	81±23
cheetah_run	easy	**508±39**	429±80	472±35	161±42	85±19	270±112
cheetah_run	medium	113±34	**479±95**	472±35	38±11	62±25	70±37
cheetah_run	hard	63±12	**436±70**	472±35	33±6	39±3	62±41
reacher_hard	easy	748±125	**939±38**	925±75	197±375	4±2	476±289
reacher_hard	medium	12±7	**929±76**	925±75	12±5	7±5	7±6
reacher_hard	hard	8±4	**907±72**	925±75	13±9	10±5	9±5
cartpole_swingup	medium	725±48	**834±26**	821±35	78±25	83±23	161±60
finger_spin	medium	365±13	**587±178**	775±181	1±1	1±1	4±6
hopper_hop	medium	0±1	**6±12**	6±8	0±0	0±0	0±0

4 Experiments

We evaluate whether randomized action interventions recover controllable observation dimensions when passive correlations are misleading. The benchmark appends distractors to continuous-control observations without changing the true state, reward, action space, or transition dynamics. This design lets us separate two questions: whether distractors degrade downstream RL, and whether the selected mask recovers the action-reachable boundary.

4.1 Benchmarks and Evaluation Protocol

We build a controlled state-space benchmark on six DeepMind Control Suite tasks using proprioceptive observations: `walker_walk` (24 true dimensions), `cheetah_run` (17), `hopper_hop` (15), `finger_spin` (9), `reacher_hard` (6), and `cartpole_swingup` (5). The original state dimensions define the ground-truth action-reachable set; appended distractors do not alter the transition, reward, or action dynamics, so the state-only observation is a valid oracle reference. We use three distractor

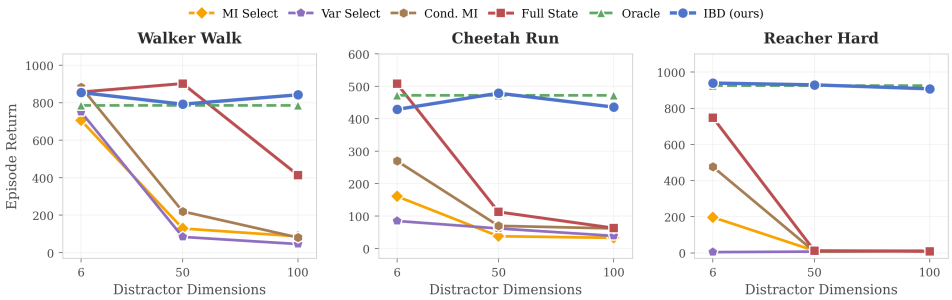


Figure 2: **Distractor scaling curve.** Episode return as a function of distractor dimensionality (6, 50, 100) for `walker_walk` (left), `cheetah_run` (center), and `reacher_hard` (right). Full State (red) degrades as distractors increase; IBD (blue) tracks oracle performance across all distractor counts.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

regimes: autonomous OU/oscillator processes, mimicking processes whose scale, variance, and autocorrelation match true proprioceptive channels, and reward-correlated drift processes that track episode progress without being caused by actions. Easy adds 6 autonomous distractors; medium adds 50 and hard adds 100, with all three types present in medium and hard.

After that, we compare Full State, Oracle, MI Select, Variance Select, Cond. MI, and IBD. To favor observational selection, MI/Variance/Cond. MI receive the true number of causal dimensions d_c as their selection budget; IBD receives no budget hint and selects via its FDR-controlled significance threshold. Unless stated otherwise, all downstream agents use SAC with identical hyperparameters (lr 3×10^{-4}, batch 256, replay buffer 300K, MLP [256,256], 300K environment steps). We report final episodic return as mean $\pm$ std over 5 seeds, with 10 evaluation episodes per seed, and report mask quality using precision, recall, and F1.

Table 2: 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 spurious correlation is not resolved by gradient-based sensitivity analysis.

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

4.2 Main Control Results

Table 1 gives the main downstream control results. IBD reaches oracle-level performance in 11 of 12 evaluated settings, while passive feature selection often performs worse than training on the full noisy observation. This pattern is strongest in the medium and hard regimes, where distractors share state-like temporal statistics or progression-linked drift. These results support the central claim: the failure mode is not high dimensionality alone, but high-dimensional input whose passive correlations make non-causal dimensions look relevant. For example, for `reacher_hard` medium, Full State reaches only 12 ± 7, and the three passive selectors remain near random: MI Select obtains 12 ± 5, Variance Select obtains 7 ± 5, and Cond. MI obtains 7 ± 6. In contrast, IBD reaches 929 ± 76, matching Oracle at 925 ± 75. The gap is notable because Cond. MI uses a state-conditioned forward model to estimate whether actions add predictive information beyond the current state. Its collapse shows that conditioning can reduce marginal association but still cannot remove spurious dependence created by action-correlated distractor dynamics.

The same pattern appears beyond a single task. On `cheetah_run` medium, Full State drops to 113 ± 34, while IBD reaches 479 ± 95 and Oracle reaches 472 ± 35. On `finger_spin` medium, all passive baselines collapse to near-zero return, while IBD recovers 587 ± 178. These gains occur even though the passive selectors know the correct selection budget. Thus, the budget is not the main issue: the ranking signal is wrong when non-causal dimensions share passive statistics with causal state dimensions. IBD avoids this failure by asking whether each dimension changes under randomized action intervention.

Table 3: Multistep inverse dynamics baseline: episode return (mean ± std over 5 seeds, medium/hard distractors) and boundary discovery accuracy (precision / recall / F1). **Bold:** best non-oracle method on return.

			Inverse Dyn.		IBD (ours)		
Environment	Distr.	Full State	Return	P / R / F1	Return	P / R / F1	Oracle
`reacher_hard`	medium	12±7	28±53	0.67 / 0.67 / 0.67	**929**±76	**0.97 / 1.00 / 0.98**	925±75
`walker_walk`	hard	414±144	624±26	0.88 / 0.88 / 0.88	**842**±121	**0.97 / 0.85 / 0.91**	785±128
`cheetah_run`	medium	113±34	285±74	0.95 / 0.95 / 0.95	**479**±95	**0.94 / 1.00 / 0.97**	472±35
`finger_spin`	medium	365±13	21±26	0.44 / 0.44 / 0.44	**587**±178	**0.96 / 1.00 / 0.98**	775±181

4.3 Scaling with Distractor Dimension

As shown in Figure 2, we measured how return changes as the number of distractors increases from 6 to 50 to 100. Full-State SAC remains competitive when the distractor-to-signal ratio is small, but degrades once distractors dominate the observation vector. The failure point is better predicted by this ratio than by the absolute distractor count. For `walker_walk`, which has 24 true dimensions, Full State tolerates 50 distractors but drops sharply at 100. For `reacher_hard`, which has only 6 true dimensions, Full State collapses already at 50 distractors.

IBD remains close to Oracle across the scaling curve because the per-dimension test asks whether a coordinate responds to action intervention, not whether it resembles other coordinates passively. A denser five-point sweep ($d_d \in \{6, 24, 50, 100, 150\}$, 3 seeds per cell) confirms this on the distractor-to-signal ratio axis: in the degradation regime $d_d \geq d_c$, a descriptive log–log fit gives $R/R_{\text{oracle}} \propto (d_d/d_c)^{-1.71}$ for Full State, while IBD's exponent is 0.002, consistent with no degradation. Fit details, confidence intervals, and per-task curves are in Appendix H.

4.4 Causal Boundary Recovery

We next test whether the downstream gains correspond to accurate recovery of the action-reachable boundary. Across all 12 settings, IBD achieves mean precision, recall, and F1 of 0.95, with precision at least 0.93 in 11 of the 12 settings, so FDR-controlled intervention testing rarely readmits the exogenous distractors that feature selection is meant to remove. Recall drops only where some true dimensions have a weak intervention response (`walker_walk` 0.85–0.92, `hopper_hop` 0.75), and downstream return stays close to Oracle in those cases, suggesting that the missed dimensions are redundant with selected state variables or weakly tied to the reward. Boundary quality is also stable as distractor count grows: on `cheetah_run`, F1 is 0.99, 0.97, and 0.98 from easy to medium to hard. Per-setting numbers are in Table 6 (Appendix D); the mask-level results agree with the control results, indicating that IBD improves return by excluding non-causal dimensions rather than by overfitting a specific learner.

4.5 Stronger Observational Dynamics Baselines

We further compare against two model-based observational baselines. The *gradient-attribution* baseline trains a joint forward model $f(\mathbf{s}, \mathbf{a}) \to \mathbf{s}'$ and ranks each output dimension by $\mathbb{E}[\|\partial f_i / \partial \mathbf{a}\|_1]$. The *multistep inverse dynamics* baseline adapts the action-recovery signal used in exogenous-block MDP methods [13, 15]: it trains $g(\mathbf{o}_t, \mathbf{o}_{t+k}) \to \mathbf{a}_t$ and ranks dimensions by the validation MSE increase caused by ablating each dimension. Both baselines use natural action variation in collected rollouts rather than action interventions. (Implementation details are given in Appendix F and G.)

Neither model-based observational baseline closes the gap. In Table 2, gradient attribution reaches only 5 ± 10 on `reacher_hard`, compared with 929 ± 76 for IBD, and reaches 56 ± 32 on `cartpole_swingup`, below the simpler passive selectors in Table 1. Its boundary F1 is only 0.44–0.50, meaning that roughly half of the selected dimensions are distractors. A learned forward model can assign action sensitivity to mimicking distractors when they co-vary with action-influenced state under the observational rollout distribution.

Multistep inverse dynamics is stronger than raw MI or variance and improves over Full State in 3 of 4 settings in Table 3. However, it still does not match IBD or Oracle. On `reacher_hard` medium, inverse dynamics obtains F1 0.67, so two of the six selected dimensions are distractors; this small leakage reduces return from 929 to 28. On `finger_spin` medium, F1 falls to 0.44 and return drops below Full State. These results show that action-recovery objectives can still confuse causal influence with passive co-variation unless the action mechanism is intervened on.

Table 4: **Confounded distractor regime** (boundary F1, mean over `walker_walk`, `cheetah_run`, and `reacher_hard`, 5 seeds each). A latent AR(1) process C_t enters the action channel with strength κ and drives 12 distractor dimensions with strength λ; the backdoor path is open only when both edges are active (κ=0.5, λ=1). Per-task results and false-positive breakdowns are in Appendix I.

Method	κ=0, λ=0	κ=0.5, λ=0	κ=0, λ=1	κ=0.5, λ=1
IBD (ours)	**0.934**	**0.934**	**0.949**	**0.949**
Inverse Dyn.	0.870	0.833	0.872	0.733
MI Select	0.919	0.724	0.919	0.403
Grad. Attr.	0.563	0.531	0.559	0.457
Cond. MI	0.498	0.456	0.499	0.455
Var Select	0.329	0.356	0.359	0.368

4.6 Confounded Distractors

The distractors used so far are appended exogenous processes, so Proposition 3.2 is illustrated by construction rather than instantiated in the benchmark. We therefore add a regime in which a genuine latent confounder enters the action channel, realizing the backdoor path of Proposition 3.2 in a continuous-control environment. A latent AR(1) process C_t perturbs the probe policy's actions with strength κ and independently drives 12 appended distractor dimensions with strength λ, with no edge from any action or state to those dimensions. The 2×2 design over (κ, λ) isolates the effect of opening the backdoor path: the ground-truth action-reachable set is identical in all four cells, so F1 is comparable across the columns of Table 4. Details of the generative process are in Appendix I.

The control cell κ=λ=0 confirms that the observational baselines are not simply mis-implemented: MI Select reaches F1 0.919 there, and multistep inverse dynamics 0.870. Opening a single edge does not create a backdoor path, since C_t then influences either the actions or the distractors but not both. Only when both edges are active is the backdoor path open, and the observational selectors degrade sharply: MI Select falls to 0.403, admitting 8.5 of the 12 confounded dimensions, while inverse dynamics falls to 0.733 with 3.4 confounded false positives. IBD is essentially unchanged across all four cells (F1 0.934–0.949, identical for κ=0 and κ=0.5), because randomizing the action mechanism severs $C \to \mathbf{a}$ while leaving the autonomous mechanism of the distractors intact, exactly as Proposition 3.3 predicts.

4.7 Transfer across Downstream RL Algorithms

To test whether the benefit depends on SAC, we repeat the hard-distractor experiments on `walker_walk` and `cheetah_run` with TD3 [4] as the downstream algorithm. The same mask transfers without change. On `walker_walk`, TD3 with Full State reaches 142 ± 131, while TD3 with IBD reaches 734 ± 208 against Oracle at 781 ± 116. On `cheetah_run`, Full State reaches 67 ± 12, while IBD reaches 356 ± 80 against Oracle at 425 ± 24. Thus, IBD reaches 94% and 84% of oracle return on the two tasks.

The gain is larger for TD3 than for SAC on `walker_walk`: IBD improves TD3 Full State by more than $5\times$, compared with about $2\times$ for SAC. This is consistent with TD3 being more sensitive to irrelevant high-dimensional inputs because it lacks SAC's entropy-regularized training objective. Since the mask is computed before downstream RL training and is unchanged across learners, the result indicates that the main benefit comes from causal feature filtering rather than from a SAC-specific effect.

4.8 Partially Controllable Dimensions

In the real system, dimensions might contain both action-dependent and autonomous components. Thus, we test this setting by adding partially controllable dimensions with dynamics

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

where g is a nonlinear function of state and action, z_t is an exogenous OU process, and $\alpha \in [0, 1]$ controls the strength of causal mixing.

IBD detects weak but real action influence. At $\alpha = 0$, the added dimensions are purely exogenous and recall is correctly zero. By $\alpha \approx 0.05$, corresponding to only about 5% causal variance, recall reaches 1.0 on both `cheetah_run` and `walker_walk`. Precision remains at least 0.92 across all tested values of α. The intervention test therefore behaves as a graded detector of action influence rather than as a brittle binary test.

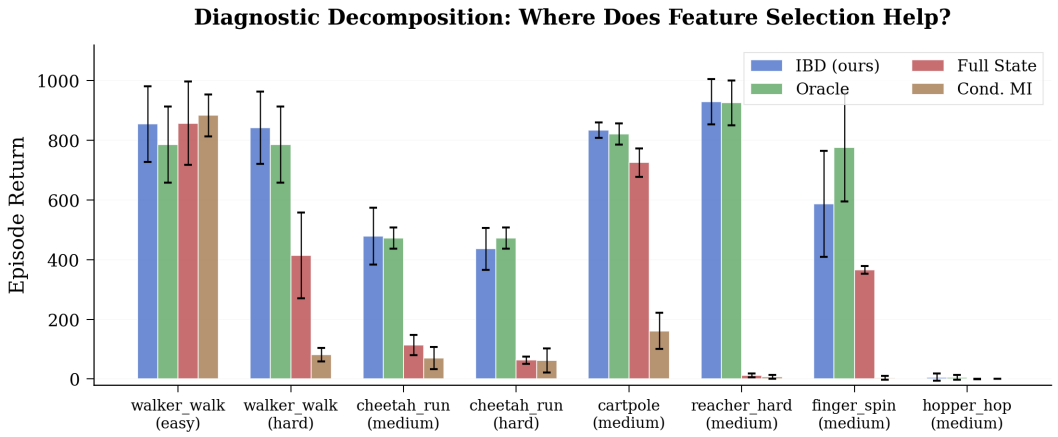


Figure 3: **Where does feature selection help?** Episode return across 8 representative settings. Three regimes emerge. (1) `walker_walk` (easy): all methods perform similarly, so no selection is needed. (2) Most medium/hard settings: Oracle $\gg$ Full State but IBD $\approx$ Oracle, so distractors are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

4.9 Diagnostic Decomposition

Figure 3 turns the comparison among Full State, IBD, and Oracle into a diagnostic. The three regimes it identifies map onto distinct remedies: when Full State already matches Oracle no selection is needed; when Oracle dominates and IBD closes the gap, as in `reacher_hard` medium and `cheetah_run` medium/hard, the bottleneck is distractor confusion; and when all three fail the problem lies in exploration or reward design rather than in the observation space.

5 Discussion

The degradation reflects a finite-capacity RL failure under high-dimensional irrelevant inputs, not a changed control problem: distractors do not affect the true state, reward, actions, or transitions. Although sufficiently large policies could learn to ignore them, finite-budget training can overvalue non-causal dimensions correlated with state statistics or episode progress. IBD addresses this before policy learning by applying an interventional observation mask, and can also operate on learned encoder features [11, 12]. All of our experiments are in simulation, with stationary probing and appended exogenous distractors. Running IBD on a real robot, for example in long-horizon manipulation [31], would require a safe way to override actions during probing. We plan to study this setting, along with indirect action effects, policy-dependent distractors, and other model capacities and training budgets [13, 15].

Acknowledgments and Disclosure of Funding

The authors J.L., A.C., and P.B. acknowledge the support by the National Science Foundation (NSF) Award 2243104 under the Center for Complex Particle Systems (COMPASS), the Career Award CPS-1453860, CNS-1932620, and NSF Mid-Career Advancement Award BCS-2527046, the U.S. Army Research Office (ARO) under Grant No. W911NF-23-1-0111, the Defense Advanced Research Projects Agency (DARPA) Young Faculty Award and DARPA Director Fellowship Award under Grant Number N66001-17-1-4044, Intel faculty awards and a Northrop Grumman grant. P.B. is grateful to National Institute of Health (NIH) for the grants R01 AG 079957 “Interpretable machine learning to synergize brain age estimation and neuroimaging genetics” and RF1 AG 082201 “Neurovascular calcification and ADRD in two nonindustrial Native American populations”. P.B. is also grateful to Carnegie Mellon University for hosting him during his sabbatical. The views, opinions, and/or findings in this article are those of the authors and should not be interpreted as official views or policies of the Department of War, the National Institutes of Health or the National Science Foundation. The authors declare no competing interests.

References

[1] Zilin Kang, Chenyuan Hu, Yu Luo, Zhecheng Yuan, Ruijie Zheng, and Huazhe Xu. A forget-and-grow strategy for deep reinforcement learning scaling in continuous control. *arXiv preprint arXiv:2507.02712*, 2025.

[2] Chong Zhang, Wenli Xiao, Tairan He, and Guanya Shi. Wococo: Learning whole-body humanoid control with sequential contacts. *arXiv preprint arXiv:2406.06005*, 2024.

[3] 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.

[4] Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *ICML*, 2018.

[5] 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.

[6] Scott Jeen, Tom Bewley, and Jonathan M Cullen. Zero-shot reinforcement learning under partial observability. *arXiv preprint arXiv:2506.15446*, 2025.

[7] Hai Huu Nguyen, Long Dinh Van The, Christopher Amato, and Robert Platt. Leveraging mutual information for asymmetric learning under partial observability. In *8th Annual Conference on Robot Learning*, 2024.

[8] Dongchi Huang, Jiaqi Wang, Yang Li, Chunhe Xia, Tianle Zhang, and Kaige Zhang. Pigdreamer: Privileged information guided world models for safe partially observable reinforcement learning. *arXiv preprint arXiv:2508.02159*, 2025.

[9] 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`.

[10] Nicklas Hansen, Hao Su, and Xiaolong Wang. Stabilizing deep q-learning with convnets and vision transformers under data augmentation. In *NeurIPS*, 2021.

[11] Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Mastering visual continuous control: Improved data-augmented reinforcement learning. In *ICLR*, 2022.

[12] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. CURL: Contrastive unsupervised representations for reinforcement learning. In *ICML*, 2020.

[13] 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.

[14] 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.

[15] Alexander Levine, Peter Stone, and Amy Zhang. Learning a fast mixing exogenous block MDP using a single trajectory. In *International Conference on Learning Representations*, 2025.

[16] Alexander Levine, Peter Stone, and Amy Zhang. Offline action-free learning of ex-bmdps by comparing diverse datasets. *Reinforcement Learning Journal*, 2025.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. CITRIS: Causal identifiability from temporal intervened sequences. In *ICML*, 2022.

[22] 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.

[23] 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.

[24] Alexander S Klyubin, Daniel Polani, and Chrystopher L Nehaniv. Empowerment: A universal agent-centric measure of control. In *IEEE Congress on Evolutionary Computation*, 2005.

[25] Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. In *NeurIPS*, 2015.

[26] Rudolf E Kalman. On the general theory of control systems. *IFAC Proceedings Volumes*, 1(1): 491–502, 1960.

[27] Jacob E. Kooi, Mark Hoogendoorn, and Vincent François-Lavet. Disentangled (un)controllable features. In *IEEE Symposium Series on Computational Intelligence (SSCI)*, 2023.

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

[29] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009.

[30] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

[31] Yuanzhe Liu, Jingyuan Zhu, Yuchen Mo, Gen Li, Xu Cao, Jin Jin, Yifan Shen, Zhengyuan Li, Tianjiao Yu, Wenzhen Yuan, et al. Palm: Progress-aware policy learning via affordance reasoning for long-horizon robotic manipulation. *arXiv preprint arXiv:2601.07060*, 2026.

[32] Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. URL `http://jmlr.org/papers/v22/20-1364.html`.

A Distractor Environment Design

Our distractor benchmark instantiates three distractor patterns that recur in instrumented physical and software-driven systems. Each captures a different statistical regime that an observation-based selector would face in deployment.

Autonomous distractors are Ornstein-Uhlenbeck processes and coupled oscillators that evolve independently of the agent's actions and state. They model environmental noise sources (vibration, ambient processes, secondary equipment cycles) whose variance and frequency content differ from true state dimensions, and are filterable by any method that examines statistical properties beyond raw magnitude.

Mimicking distractors are exogenous OU processes whose marginal statistics are matched to those of true proprioceptive dimensions, modeling sensor cross-talk and time-correlated readout artifacts that share statistical signatures with the channels they sit alongside. The match 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`: $\sigma_{\text{ref}} = 0.75$). Because they are indistinguishable from true dimensions in their marginal distribution, methods that rely solely on marginal statistics cannot separate them from genuine state.

Reward-correlated distractors are autonomous processes whose temporal profile resembles the within-episode reward curve typical of locomotion tasks: a saturating ramp $1 - \exp(-\kappa t)$, gait-frequency sinusoidal oscillation, and OU noise. Their dynamics depend only on episode time t, not on state, reward, or actions; the resemblance to reward curves nonetheless induces spurious mutual information with the agent's action sequence (because policy actions also accumulate over the episode), even though no causal path from actions to these dimensions exists.

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. Distinguishing all three types requires reasoning about causal influence rather than statistical association, since an action intervention leaves the distribution of every type unchanged regardless of its marginal similarity to true state.

B Partial Controllability: Experimental Details

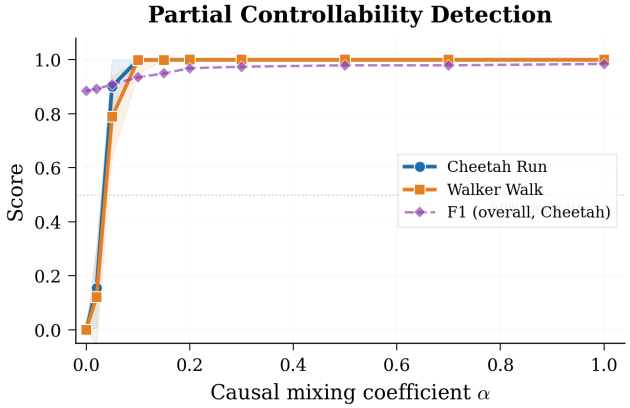


Figure 4: **Partial controllability detection.** Recall on partially controllable dimensions as a function of mixing coefficient α, for `cheetah_run` and `walker_walk`. At $\alpha = 0$ (purely exogenous), recall is correctly zero; by $\alpha \approx 0.05$ ($\sim$5% causal variance), recall reaches 1.0. 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 4.8), 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: < 1s).

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 5: Scout ablation (P/R/F1, mean over 3 seeds). Budget $= 0$: structured random probe; $\geq$10K: 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 all match budget $= 0$ and are omitted. On `walker_walk`, the untrained policy achieves the highest F1 (0.94 versus 0.90 at 80K), likely because a trained policy converges toward a narrow behavioral mode that reduces state coverage.

D Boundary Discovery Accuracy: Full Results

Table 6: 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

Table 6 reports IBD's per-setting precision, recall, and F1. Precision is at least 0.93 in 11 of 12 settings and 0.86 on `reacher_hard` hard, confirming that FDR control keeps distractor leakage low. 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 7: 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 / 200×200
Grad. Attr.	Arch / epochs / LR / batch / data	MLP [128,128] / 80 / 10^{-3} / 2048 / 200×200
Inverse Dyn.	Arch / epochs / LR / batch / data / horizon k	MLP [128,128] / 80 / 10^{-3} / 2048 / 200×200 / 3
Robustness	α / partial dims / exo distr / seeds	{0…1.0} / 6 / 20 / {42,142,242}

Compute resources. All experiments run on CPU; no GPU is required. The IBD probing step uses only NumPy and SciPy (Welch's t-test, Benjamini–Hochberg correction) and contains no neural network: a single probe (32K environment steps with N=80, T=200) completes in 2–3 minutes on one CPU core (Appendix B). Downstream SAC/TD3 training uses Stable-Baselines3 with the default CPU device, which is faster than GPU for the MLP [256, 256] policies due to data-transfer overhead; one 300K-step training run takes approximately 27 minutes on a single CPU core (measured across 99 runs spanning all six tasks: min 25.6, max 29.3, mean 27.0 minutes).

Total compute for the experiments reported in this paper is approximately 280 CPU-hours, broken down as: Table 1 main results (12 settings $\times$ 6 methods $\times$ 5 seeds $\approx$ 360 runs, $\sim$160 CPU-h); the TD3 validation of Section 4.7 ($\sim$14 CPU-h); Tables 2 and 3 model-based baselines ($\sim$18 CPU-h); the dense scaling sweep in Appendix H ($\sim$60 CPU-h); the partial-controllability and scout-budget ablations ($\sim$6 CPU-h); and the confounded-distractor, summary-statistic, and multiplicity studies of Appendices I–K ($\sim$25 CPU-h), which are mask-level only: no downstream policy is trained and each probe costs 2–3 minutes.

F Gradient Attribution Baseline: Implementation Details

The gradient attribution baseline (results in Table 2) trains a joint forward dynamics model $f(\mathbf{s}, \mathbf{a}) \to \mathbf{s}'$ (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: $\text{score}_i = \mathbb{E}_{\text{data}}\big[\|\partial f_i(\mathbf{s}, \mathbf{a})/\partial \mathbf{a}\|_1\big]$. This directly measures how sensitive the learned dynamics are to the action input and represents a direct model-based approach to identifying action-influenced dimensions.

However, gradient attribution remains observational. When mimicking distractors co-vary with true state dimensions that also co-vary with actions, a learned forward model can assign action sensitivity to those distractors under the observational data distribution, even though the interventional effect of $\text{do}(\mathbf{a})$ on them is zero. Its boundary discovery F1 is accordingly low (0.44–0.50), indicating that gradient attribution selects roughly half distractors and half true dimensions.

G Multistep Inverse Dynamics Baseline: Implementation Details

As a further test of whether observational methods with stronger structural priors can succeed where MI, Cond. MI, and gradient attribution fail, we evaluate a continuous-control adaptation of the multistep inverse dynamics principle from the EX-BMDP line [13, 15, 16] (results in Table 3). We do not compare directly against STEEL [15], whose discrete latent state recovery, fast-mixing assumption, and PAC-style $|S|$ upper bound do not transfer to continuous state-action spaces without substantial reformulation. Instead, we isolate the multistep inverse dynamics signal that is the

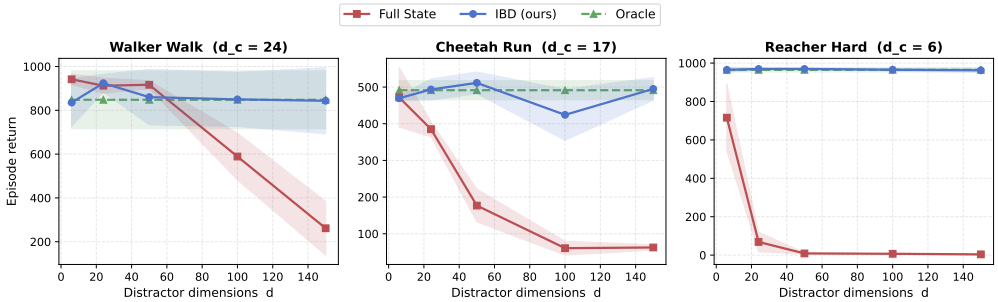


Figure 5: **Distractor scaling curves (dense sweep).** Episode return vs. distractor count d for `walker_walk` ($d_c = 24$, left), `cheetah_run` ($d_c = 17$, center), and `reacher_hard` ($d_c = 6$, right). Full State (red) collapses earlier on tasks with smaller d_c; IBD (blue) tracks Oracle across all distractor counts. Shaded regions: ± 1 std over 3 seeds.

core empirical ingredient of that line and apply it at the observation-dimension level, matching our budget-controlled feature selection protocol.

The method trains a network $g(\mathbf{o}_t, \mathbf{o}_{t+k}) \to \mathbf{a}_t$ (MLP [128, 128], 80 epochs) on rollouts collected with the same structured-random probe policy used by IBD, with horizon $k = 3$. For each observation dimension i, we score dimension i by the increase in validation MSE when dimension i is zeroed out in both $\mathbf{o}_t$ and $\mathbf{o}_{t+k}$ at test time: $\text{score}_i = \text{MSE}_{\text{ablated}}(i) - \text{MSE}_{\text{full}}$. Dimensions whose values carry action information receive high scores; purely exogenous distractors should receive low scores. The top-d_c dimensions are selected, matching the budget given to MI, Variance, and Cond. MI.

This baseline remains observational: it exploits natural action variation in the data collection policy, not interventions. It is therefore still vulnerable to confounded distractors whose dynamics co-vary with action-influenced state dimensions (Section 4.6), in contrast to IBD, which severs such confounding by replacing the action mechanism with a confounder-independent randomized policy.

H Additional Distractor Scaling Sweep

This appendix supplements Section 4.3 with a denser distractor sweep on the distractor-to-signal ratio axis d/d_c. Across this sweep, Full State return decreases as d/d_c grows, while IBD stays close to Oracle. We summarize the trend with a power-law fit; the fit is descriptive on the points we sweep, not a claim about a universal scaling exponent.

Setup. We sweep five distractor counts, $d \in \{6, 24, 50, 100, 150\}$, on `walker_walk` ($d_c = 24$), `cheetah_run` ($d_c = 17$), and `reacher_hard` ($d_c = 6$), giving distractor-to-signal ratios from 0.25 up to 25. Three downstream RL methods are evaluated: Full State (raw observation), IBD (interventional mask), and Oracle (ground-truth action-reachable dimensions). All hyperparameters match Section 4.1: SAC, 300K environment steps, MLP $[256, 256]$, 3 seeds per cell. Oracle is mask-invariant in distractor count and reported once per task.

Linear-scale curves. Figure 5 shows episode return as a function of distractor count for each task. Three failure profiles emerge. `Walker_walk` ($d_c = 24$) tolerates distractors up to $d = 50$ ($d/d_c \approx 2$) and only collapses at $d = 100$. `Cheetah_run` ($d_c = 17$) degrades monotonically from $d = 24$ onward. `Reacher_hard` ($d_c = 6$) collapses already at $d = 24$ ($d/d_c = 4$) and stays near zero return through $d = 150$. IBD tracks Oracle in all three tasks across the full sweep.

Power-law fit on the ratio axis. Figure 6 plots normalized return R/R_{oracle} against the ratio d/d_c on log–log axes. We fit a power law $R/R_{\text{oracle}} = a \cdot (d/d_c)^{-\alpha}$ on the degradation regime $d \geq d_c$, since for $d < d_c$ distractors are a minority of observation dimensions and the network has sufficient capacity to ignore them, so a power-law model is not appropriate there. Fits are done by weighted least squares in log space, with 95% confidence intervals on α from 1000 bootstrap resamples.

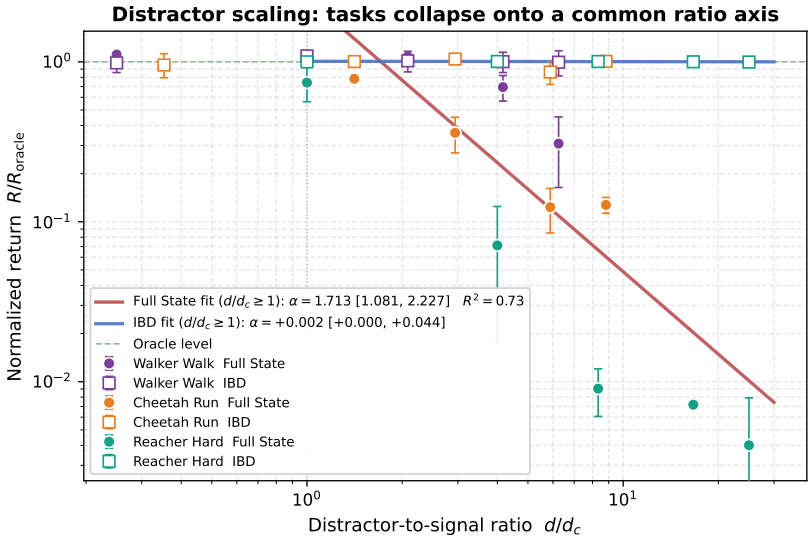


Figure 6: **Distractor scaling on the ratio axis.** Normalized return R/R_{oracle} vs. distractor-to-signal ratio d/d_c on log–log axes. Filled markers: Full State; open markers: IBD; colors distinguish tasks. Solid lines are descriptive power-law fits restricted to $d \geq d_c$ (vertical dotted line). Pooled across tasks, Full State follows $R/R_{\text{oracle}} \propto (d/d_c)^{-1.71}$ ($R^2 = 0.73$); IBD stays at the oracle level with $\alpha \approx 0$.

Pooling Full State across the three tasks ($n = 13$ ratio–task points), we obtain

$$\alpha_{\text{Full State}} = 1.71 \quad [1.08,\ 2.23], \qquad R^2 = 0.73,$$

and pooling IBD on the same set,

$$\alpha_{\text{IBD}} = 0.002 \quad [0.000,\ 0.044].$$

The two bootstrap CIs do not overlap on this sweep: IBD's exponent is consistent with no degradation across nearly two decades of d/d_c, while Full State decays roughly as $(d/d_c)^{-1.7}$ on the swept range.

Connection to the collapse threshold. The pooled exponent $\alpha \approx 1.71$ is broadly consistent with the per-task collapse points in Figure 5. Using the same normalization as the fit, Full State return first falls below 50% of oracle return between distractor-to-signal ratios of 1 and 4 on `reacher_hard`, between 1.4 and 2.9 on `cheetah_run`, and between 4.2 and 6.3 on `walker_walk`. A power-law extrapolation with slope $\alpha \approx 1.71$ predicts that R/R_{oracle} falls below 0.5 near $d_d/d_c \approx 1.7$, inside the observed interval for `reacher_hard` and `cheetah_run`; `walker_walk` tolerates a higher ratio, reflecting task-specific reward shape and policy robustness around the collapse. IBD's near-zero exponent on the same axis quantifies the qualitative claim of Section 4.3 that interventional masking decouples downstream return from the distractor-to-signal ratio.

I Confounded Distractor Regime: Design and Per-Task Results

This appendix details the confounded regime of Section 4.6. A latent process, never observed by any method, follows a discrete-time AR(1) recursion with unit stationary variance,

$$C_t = \rho_c\, C_{t-1} + \sqrt{1-\rho_c^2}\,\xi_t, \qquad \xi_t \sim \mathcal{N}(0, I),$$

and enters the system through two edges whose strengths are swept independently. The probe policy's action is perturbed by C_t with strength κ,

$$\mathbf{a}_t = \text{clip}\big(\pi_{\text{probe}}(\mathbf{o}_t) + \kappa\, K C_t\big),$$

and the confounded distractor block is driven by the same latent process with strength λ,

$$\mathbf{d}_t = \rho\, \mathbf{d}_{t-1} + \lambda\, L C_t + \varepsilon_t.$$

There is no edge from $\mathbf{a}_t$ or $\mathbf{s}_t$ into $\mathbf{d}_t$, so the confounded dimensions remain outside the SoI by construction and the oracle mask (true state dimensions only) stays valid in all four cells. When both $\kappa > 0$ and $\lambda > 0$, however, $\mathbf{a}_t$ and $\mathbf{d}_t$ share a latent cause, which opens the backdoor path of Proposition 3.2: $I(\mathbf{d}_{t+1}; \mathbf{a}_t) > 0$, a state-conditioned forward model gains predictive power from $\mathbf{a}_t$ because actions reveal C_t that $\mathbf{s}_t$ does not determine, and $\mathbf{a}_t$ becomes partially recoverable from observation pairs that include $\mathbf{d}$. Under $\mathrm{do}(\mathbf{a} \sim \mathrm{Unif}(\mathcal{A}))$ the edge $C \to \mathbf{a}$ is severed while the mechanism generating $\mathbf{d}$ is untouched, so IBD's two-sample test sees no shift on those dimensions.

Each task receives 12 confounded dimensions in addition to the medium exogenous distractor bank, giving observation dimensions 86 (`walker_walk`), 79 (`cheetah_run`), and 68 (`reacher_hard`). We use $\rho_c = 0.95$ for the latent process and sweep $\kappa \in \{0, 0.5\}$, $\lambda \in \{0, 1\}$ with 5 seeds per cell, for 360 probe–selector runs in total.

Table 8 reports the fully confounded cell per task, separating false positives on confounded dimensions ($\mathrm{FP}_{\mathrm{conf}}$, out of 12) from those on exogenous distractors ($\mathrm{FP}_{\mathrm{exo}}$). The failure of the observational selectors is specifically a confounding failure rather than general noise sensitivity: on `walker_walk`, MI Select admits 11.8 of the 12 confounded dimensions, and multistep inverse dynamics admits 8.4, whereas IBD admits 0.2. Variance Select is an instructive exception in the opposite direction: it admits almost no confounded dimensions but 15.4 exogenous ones, because marginal variance is uninformative about the latent process, so its low F1 has a different origin from the other baselines'.

Table 8: Per-task results in the fully confounded cell ($\kappa = 0.5$, $\lambda = 1$; mean $\pm$ std over 5 seeds). $\mathrm{FP}_{\mathrm{conf}}$: confounded dimensions admitted (out of 12). $\mathrm{FP}_{\mathrm{exo}}$: exogenous distractors admitted. **Bold:** best F1 per task.

Task	Method	F1	$\mathrm{FP}_{\mathrm{conf}}$	$\mathrm{FP}_{\mathrm{exo}}$
`walker_walk`	IBD (ours)	**0.917**±0.035	0.2	2.0
	Inverse Dyn.	0.650±0.020	8.4	0.0
	MI Select	0.442±0.020	11.8	1.6
	Grad. Attr.	0.308±0.020	8.4	8.2
	Cond. MI	0.350±0.020	2.6	13.0
	Var Select	0.358±0.020	0.0	15.4
`cheetah_run`	IBD (ours)	**0.961**±0.022	0.0	1.4
	Inverse Dyn.	0.882±0.000	1.8	0.2
	MI Select	0.435±0.096	9.6	0.0
	Grad. Attr.	0.529±0.000	8.0	0.0
	Cond. MI	0.682±0.080	2.6	2.8
	Var Select	0.412±0.000	4.2	5.8
`reacher_hard`	IBD (ours)	**0.969**±0.038	0.0	0.4
	Inverse Dyn.	0.667±0.000	0.0	2.0
	MI Select	0.333±0.000	4.0	0.0
	Grad. Attr.	0.533±0.067	2.8	0.0
	Cond. MI	0.333±0.000	0.6	3.4
	Var Select	0.333±0.000	0.0	4.0

J Summary-Statistic Ablation

The summary Δ_i^h of Equation (2) measures step-to-step displacement, so it is by construction insensitive to an action effect that shifts the mean level of a dimension without changing its increments. This appendix quantifies both how much that blind spot costs on the standard benchmark and how it can be removed.

We compare three statistics, all scored on the same collected trajectories so that the comparison is exactly paired. Delta is the published Δ_i^h; level is the per-trajectory mean of the raw values $o^{(i)}$; both takes the union of the two families of tests under a joint Benjamini–Hochberg correction, so the reported gains already pay the multiplicity cost of the extra family.

On four standard-benchmark settings (`cheetah_run` medium, `walker_walk` medium and hard, `hopper_hop` medium), the displacement statistic dominates the level statistic everywhere, with recall higher by 0.35–0.49, confirming that it is the appropriate default when every action-reachable

dimension carries an interventional signal in its increments. Adding the level family raises recall where some dimensions respond weakly (`hopper_hop` $0.773 \to 0.880$, `walker_walk` medium $0.875 \to 0.950$) and costs a few points of precision where delta already achieves recall 1.0 (`cheetah_run`), since the second family then only adds opportunities for false positives.

The picture changes when the blind spot is actually exercised. We add 6 dimensions that are action-reachable only through their mean level, leaving their increment distribution unchanged under intervention. Table 9 reports the detection rate on those constructed dimensions alongside overall boundary quality. The displacement statistic misses them almost entirely, recovering none of the six on `reacher_hard`; the level statistic recovers all of them but loses the dimensions whose signal is carried by increments; the union recovers both families and attains the best overall F1 in all three tasks.

Table 9: Detection of level-shift dimensions (mean $\pm$ std over 5 seeds, medium distractors, 6 constructed dimensions per task). Level recall is the detection rate on the constructed dimensions; precision, recall, and F1 are over the full boundary.

Setting	Statistic	Level recall	Precision	Recall	F1
`cheetah_run`	delta	0.067±0.082	0.956±0.022	0.757±0.021	0.845±0.020
	level	1.000±0.000	0.958±0.055	0.548±0.044	0.695±0.039
	both	1.000±0.000	0.936±0.030	1.000±0.000	**0.967**±0.016
`reacher_hard`	delta	0.000±0.000	0.893±0.096	0.500±0.000	0.639±0.025
	level	1.000±0.000	0.931±0.090	0.583±0.000	0.715±0.028
	both	1.000±0.000	0.905±0.090	1.000±0.000	**0.948**±0.050
`walker_walk`	delta	0.033±0.067	0.911±0.028	0.740±0.033	0.816±0.022
	level	1.000±0.000	0.962±0.047	0.327±0.013	0.488±0.019
	both	1.000±0.000	0.919±0.054	0.947±0.027	**0.932**±0.034

Since the union costs at most a few points of precision when the blind spot is absent and is necessary when it is present, we recommend it as the default in deployments where the mechanism generating each dimension is unknown, and retain the displacement statistic alone for the results reported in the main text.

K Multiple-Testing and Test-Statistic Variants

Proposition 3.5 invokes the standard Benjamini–Hochberg guarantee, which requires independence or PRDS among the dimension-level p-values. Mimicking distractors are calibrated against true proprioceptive channels and are therefore statistically coupled to them, so this condition cannot be taken for granted. We evaluate two axes on the confounded benchmark of Section 4.6: the two-sample test (Welch t-test versus a two-sample Kolmogorov–Smirnov test, which is sensitive to distributional changes beyond the mean) and the multiplicity correction (BH versus Benjamini–Yekutieli [30], which is valid under arbitrary dependence at the cost of a $\sum_{i=1}^{d} 1/i$ factor). All four variants are computed from the same probe trajectories, so the comparison is exactly paired.

Table 10: Test-statistic and multiplicity variants, pooled over the four confounding cells, 3 tasks, and 5 seeds ($n = 60$). FDR is the realized false discovery proportion at nominal level $\alpha = 0.05$.

Variant	Precision	Recall	F1	Realized FDR	$\mathrm{FP}_{\mathrm{conf}}$
Welch + BH (default)	0.913	**0.975**	0.942	0.087	0.23
Welch + BY	**1.000**	0.954	**0.975**	**0.000**	**0.00**
KS + BH	0.946	0.969	0.956	0.054	0.07
KS + BY	0.993	0.938	0.963	0.007	**0.00**

Two observations follow from Table 10. First, the realized false discovery proportion under BH is 0.087 against a nominal $\alpha = 0.05$, so the dependence among dimension-level p-values is strong enough to matter, consistent with the concern that the PRDS condition may not hold. Second, the remedy is inexpensive: the Benjamini–Yekutieli correction, which is valid under arbitrary dependence, drives the realized FDR to zero and admits no confounded dimension at all, while its

more conservative threshold costs only 0.021 recall. Its F1 is accordingly higher than the default (0.975 versus 0.942). The KS test behaves similarly to Welch on this benchmark, indicating that the results do not lean on the specific choice of two-sample test.

We retain Welch + BH as the default so that the main results correspond to a single fixed configuration, and note that BY is the appropriate choice when a conservative guarantee is required and a small loss of recall is acceptable.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and Section 1 state three contributions—formalizing the Causal Sphere of Influence (Definition 3.1), proposing IBD as a model-free interventional procedure (Algorithm 1), and demonstrating near-oracle downstream performance across 12 continuous-control settings—each of which is supported by Section 3 (Propositions 3.2–3.5) and the experiments in Section 4 (Tables 1 and 6).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 3.1 states the four core assumptions (decomposability, action overrideability, stationary causal structure, faithfulness/statistic sensitivity). Section 5 notes that the setting assumes stationary probing with appended exogenous distractors and identifies indirect action effects, policy-dependent distractors, and different capacity–budget regimes as future directions. Robustness to the choice of summary statistic and multiplicity correction is examined in Appendices J and K, and per-setting precision and recall are reported in Table 6.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: All theoretical claims (Propositions 3.2–3.5) are numbered and stated with explicit assumptions. Proposition 3.2 is supported by an explicit linear-Gaussian SCM construction in Section 3.3. Propositions 3.3–3.5 are accompanied by inline proof arguments invoking do-operator semantics [29], faithfulness, the central limit theorem for the Welch statistic, and the standard Benjamini–Hochberg guarantee under independence/PRDS dependence [20].

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4.1 specifies the benchmark, distractor regimes, and evaluation protocol; Algorithm 1 fully specifies IBD; Appendix A details the distractor environment design; Appendix E (Table 7) lists every hyperparameter for IBD, downstream RL (SAC and TD3), and all baselines (Cond. MI, Gradient Attribution, Multistep Inverse Dynamics). The public code release (`https://github.com/jiaxin26/IBD-RL`) reproduces all reported tables and figures.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: Code is publicly available at `https://github.com/jiaxin26/IBD-RL` and contains the IBD library, the state-space distractor benchmark, all baseline implementations, run scripts for the main results (`run_dmcontrol.py`), the distractor scaling sweep (`run_scaling.py`), partial controllability (`run_robustness.py`), the scout ablation (`run_scout_ablation.py`), the confounded regime and test variants (`run_confounding_sweep.py`, `analyze_confounding.py`), and the summary-statistic study (`run_summary_stat.py`, `run_levelshift.py`), together with `requirements.txt` and `pyproject.toml`. The DeepMind Control Suite is publicly available and no proprietary data is required.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

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

   Answer: [Yes]

Justification: Section 4.1 specifies the learning rate, batch size, replay buffer, MLP architecture, training steps, number of seeds, and evaluation protocol. Appendix E (Table 7) lists the seed values and gives the same details for every baseline (Cond. MI, Gradient Attribution, Multistep Inverse Dynamics) as well as the IBD probing budget. Hyperparameters are not tuned per task.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: All return values in Tables 1 to 3 and the TD3 results in Section 4.7 are reported as mean $\pm$ 1 standard deviation over 5 random seeds. The partial-controllability study, the scout-budget ablation, and the dense scaling sweep use 3 seeds (Appendices B, C, and H). Mask-level results in Tables 8 and 9 report mean $\pm$ 1 std over 5 seeds. Shaded regions in all figures denote ± 1 std over seeds. Power-law fits in Appendix H report 95% confidence intervals from 1000 bootstrap resamples in log space.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

Answer: [Yes]

Justification: The "Compute resources" paragraph in Appendix E reports per-run wall-clock time ($\sim$27 minutes per 300K-step SAC/TD3 run, $\sim$2–3 minutes per IBD probe, both on a single CPU core) and the total $\sim$280 CPU-hours used for all experiments in the paper. All experiments run on CPU; no GPU is required.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper introduces a feature-selection method for state-space RL evaluated entirely in simulation on DeepMind Control Suite tasks. It does not involve human subjects, personal data, or models with foreseeable misuse potential, and conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: This is foundational methodological research on causal feature selection for RL, evaluated entirely in simulation. There is no direct path to a deployed application or to specific positive or negative societal impacts beyond those generic to improvements in RL sample efficiency.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The released artifact is a feature-selection algorithm and a state-space distractor benchmark built on DeepMind Control Suite. It contains no pretrained generative models, scraped data, or other components with high misuse risk.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: DeepMind Control Suite [5] is used under its Apache-2.0 license; SAC [3] and TD3 [4] are used as downstream RL backbones via Stable-Baselines3 [32] (MIT license). All assets are cited in the references, and no asset is redistributed in modified form except through standard wrapper interfaces.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: The IBD library and the state-space distractor benchmark (autonomous OU/oscillator, mimicking, and reward-correlated regimes; see Appendix A) are released at `https://github.com/jiaxin26/IBD-RL`, together with `requirements.txt`, `pyproject.toml`, and run scripts that reproduce Tables 1 to 3 and 6, the TD3 results of Section 4.7, and the figures.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects; all experiments are conducted in simulation on DeepMind Control Suite.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human-subjects research, so IRB approval (or equivalent) is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

Justification: LLMs are not used as a component of the core methodology. The IBD algorithm, theoretical analysis, benchmark design, and all experiments do not depend on LLMs.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.