When Do Learned State Representations Break Sensitivity Analysis?

Siyu Wang[1] **Xiaocong Chen**[2] **Quan Z. Sheng**[3] **Lina Yao**[1]
[1]UNSW Sydney [2]Independent Researcher [3]Beijing Normal–Hong Kong Baptist University
siyu.wang5@unsw.edu.au

Abstract

Causal sensitivity analysis bounds the policy value of off-policy evaluation under unobserved confounding. In offline reinforcement learning, however, sensitivity bounds are typically applied not to raw states but to learned representations. We show that this two-stage pipeline can silently break the coverage guarantee: state aggregation under hidden confounding amplifies the effective pairwise action-sensitivity parameter in the latent space. Under outcome sufficiency and a common target, separate marginal-model constructions show that bounds computed at the nominal confounding level may exclude the true policy value. We characterize the amplification mechanism, give a sufficient condition for preservation, and prove a continuity result showing that the safety condition is not an isolated point: amplification grows continuously with the conditional-independence violation. Neither unsupervised nor standard task-aware representation objectives target the right quantity. We propose Sensitivity-Preserving Representation Learning (SPRL), which augments latent dynamics prediction with a kernel-weighted within-cell propensity homogeneity penalty, an observable proxy for exact population propensity homogeneity, which preserves action sensitivity under binary hidden confounding, and show across synthetic and semi-synthetic benchmarks that SPRL reduces fitted amplification relative to LDM, with interval-width and reward-prediction costs. Code: `https://github.com/sy-siyu/SPRL`.

1 Introduction

Off-policy evaluation (OPE) of sequential decision policies from observational data [Precup et al., 2000, Thomas and Brunskill, 2016] is essential in applications of offline reinforcement learning (RL) such as healthcare, education, and robotics [Kallus and Zhou, 2020, Namkoong et al., 2020, Bennett and Kallus, 2024]. When unobserved confounders influence both the logged actions and the outcomes, exact evaluation is impossible and any estimator is subject to hidden bias. Causal sensitivity analysis addresses this by bounding the plausible influence of unobserved confounding through a sensitivity model, yielding partial identification intervals that are guaranteed to contain the true policy value under the assumed confounding strength [Rosenbaum, 2002, Tan, 2006, Jin et al., 2023, Frauen et al., 2024]. In modern offline RL pipelines, however, evaluation is rarely performed on the raw state: state representations learned by autoencoders, variational models, contrastive objectives, or latent dynamics models are standard practice [Kingma and Welling, 2013, Chen et al., 2020, Hafner et al., 2023]. This raises a question that the sensitivity analysis literature has not addressed directly: does the sensitivity model's coverage guarantee survive the standard two-stage pipeline of learning a representation and then applying sensitivity bounds in the latent space?

We show that the answer is in general no. State aggregation under hidden confounding can silently amplify the effective pairwise action-sensitivity parameter in the latent space, invalidating the nominal pairwise class. Under outcome sufficiency and a common target, separate marginal-model constructions show that bounds at the same numerical input can exclude the true policy value. The failure

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

arises from a compounding of two confounding sources: per-state action confounding (bounded by the nominal parameter) and a state-selection effect whereby the confounder alters which original state is active within a given latent cell. Melnychuk et al. [2024] study representation-induced confounding bias under dimensionality reduction (loss of observed confounders); we study amplification of hidden-variable action sensitivity through within-cell state-selection by the unobserved confounder, a mechanism that operates regardless of the encoder's training objective. Both unsupervised and standard task-aware objectives can amplify the latent action-sensitivity level beyond its nominal value.

We develop both theoretical characterization and a practical remedy. Our main contributions are:

- We prove that deterministic compression can invalidate the nominal pairwise action-sensitivity class (Theorem 1), identify the underlying state-selection mechanism, and demonstrate value noncoverage under outcome sufficiency (Appendix D.5).
- We prove a sufficient condition for action-sensitivity preservation (Proposition 2) and a continuity result (Theorem 3) showing that amplification grows continuously with the degree of conditional-independence violation, with a rate controlled by behavioural-policy positivity. For binary hidden confounding, exact population propensity homogeneity is also sufficient (Proposition 6).
- We propose Sensitivity-Preserving Representation Learning (SPRL), which augments latent dynamics prediction with a kernel-weighted within-cell propensity homogeneity penalty: an observable proxy for population propensity homogeneity targeting the amplification mechanism.
- Across Setting B, IHDP, and HalfCheetah, SPRL reduces fitted amplification relative to LDM. Its intervals include the known targets in Setting A and IHDP, with measured width and reward-prediction costs.

2 Preliminaries

Confounded decision problem. We consider a one-step confounded decision problem with state space $\mathcal{S}$, binary action space $\mathcal{A} = \{0, 1\}$, hidden confounder space $\mathcal{U}$, joint distribution $P(S, U)$, behavioural policy $\pi_b(a \mid s, u)$, and reward distribution $P(R \mid S, A, U)$ with conditional mean $r(s, a, u) := \mathbb{E}[R \mid S = s, A = a, U = u]$. The analyst observes only (S, A, R); U is hidden, so $\pi_b(a \mid s, u)$ is not identifiable from observed data. As shown in Figure 1(a), the edges $U \to S$ and $U \to A$ create confounding through both state occupancy and action selection. A deterministic state representation $\phi : \mathcal{S} \to \mathcal{Z}$ maps states to a compact latent space; we use *compression* for any such map, learned or hand-designed. The standard two-stage pipeline (Figure 1(b)) first learns ϕ, then applies sensitivity analysis in Z-space. The induced latent behavioural policy is

$$\tilde{\pi}_b(a \mid z, u) \;=\; \sum_{s:\, \phi(s)=z} P(s \mid z, u)\; \pi_b(a \mid s, u), \tag{1}$$

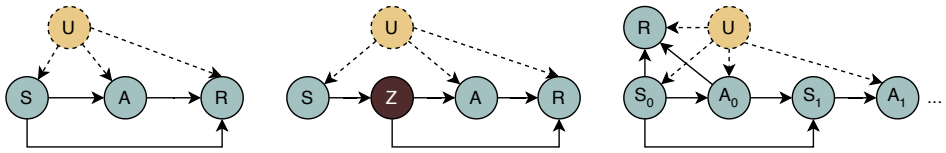
Its pairwise latent action-sensitivity level, averaged over discarded S at fixed U, is $\Gamma_{\mathcal{Z}} := \max_{z,u,u'} \mathrm{odds}(\tilde{\pi}_b(1 \mid z, u))/\mathrm{odds}(\tilde{\pi}_b(1 \mid z, u'))$, where the max is taken over z with positive support under both u and u'.

In the multi-step setting (Figure 1(c)), the agent interacts with a confounded Markov decision process (MDP) over horizon H under a stationary confounder U drawn once per trajectory. For the action-sequence result, we assume *dynamics-confounding separation* [Kallus and Zhou, 2020]: U does not directly affect the raw transition kernel given (S_t, A_t), so $P(S_{t+1} \mid S_t, A_t, U) = P(S_{t+1} \mid S_t, A_t)$. Rewards may still depend on U. The confounder enters through $P(S_0 \mid U)$ and $\pi_b(A_t \mid S_t, U)$, and possibly the reward kernel.

Causal sensitivity analysis. For target policy π_e, the expected return under data-generating process (DGP) P is, in the one-step case,

$$V(\pi_e; P) = \mathbb{E}_{(S,U)\sim P_{S,U}}\left[\sum_a \pi_e(a \mid S)\, r(S, a, U)\right]; \tag{2}$$

Fix a ϕ-measurable target $\pi_e(a \mid s) = \bar{\pi}_e(a \mid \phi(s))$, so raw and latent analyses target the same original-DGP value. The displayed $\Gamma_{\mathcal{Z}}$ averages discarded S and thus characterizes action sensitivity



(a) One-step confounded decision problem.

(b) Two-stage pipeline with representation ϕ.

(c) Multi-step confounded MDP.

Figure 1: Causal structure. (a) One-step confounded decision problem. (b) Two-stage pipeline: ϕ maps S to Z; U-information enters Z only through compression. (c) Multi-step confounded MDP under dynamics-confounding separation: U affects S_0 and all A_t but not the raw transition kernel directly; rewards may also depend on U. Solid arrows denote causal dependence; dashed arrows from U indicate the unobserved confounder's effect on observed variables.

relative to the original U. A sufficient bridge to mean-value coverage is conditional-mean outcome sufficiency: $r(s,a,u) = \bar{r}(\phi(s),a,u)$ for all (s,a,u); otherwise the outcome-complete latent variable is $U^\dagger = (S,U)$.

We write $V(\pi_e) := V(\pi_e; P^*)$ for the value under the true DGP. When U affects outcomes, $V(\pi_e)$ is generally not point-identifiable. Under reward separation $r(s,a,u) = r(s,a)$ and overlap, the one-step raw value is point-identified, as in Setting A and Appendix D. Causal sensitivity analysis restricts the influence of U through a *sensitivity model* $\mathcal{M}$: a family of distributions over (S,U,A,R) consistent with the observed marginal $P(S,A,R)$ [Rosenbaum, 2002, Tan, 2006]. The partial identification problem is

$$V^+_{\mathcal{M}}(\pi_e) = \sup_{P\in\mathcal{M}} V(\pi_e;P), \qquad V^-_{\mathcal{M}}(\pi_e) = \inf_{P\in\mathcal{M}} V(\pi_e;P), \tag{3}$$

with $[V^-_{\mathcal{M}}, V^+_{\mathcal{M}}]$ the tightest interval guaranteed to contain $V(\pi_e)$ under the constraints.

Our action-sensitivity results use a pairwise, Rosenbaum-type odds constraint [Rosenbaum, 2002, Zhao et al., 2019]. For $\Gamma \geq 1$, $\mathcal{R}(\Gamma,\mathcal{S})$ contains observed-compatible DGPs satisfying

$$\frac{1}{\Gamma} \;\leq\; \frac{\mathrm{odds}\big(\pi_b(1 \mid s,u)\big)}{\mathrm{odds}\big(\pi_b(1 \mid s,u')\big)} \;\leq\; \Gamma \qquad \text{for supported } (s,u),(s,u'), \tag{4}$$

where $\mathrm{odds}(p) := p/(1-p)$ and $\Gamma_{\mathcal{S}}$ is the minimal pairwise level for the true DGP. All conditional comparisons use supported events and essential extrema where needed. Off-support policy values used in mixture proofs follow Appendix A.1. The marginal sensitivity model (MSM) [Tan, 2006, Zhao et al., 2019] instead compares the confounded propensity with the true observed propensity $e(s) := P(A=1 \mid S=s)$:

$$\frac{1}{\Lambda} \leq \frac{\mathrm{odds}(\pi_b(1 \mid s,u))}{\mathrm{odds}(e(s))} \leq \Lambda \quad \text{for all supported } (s,u).$$

Write $\mathcal{M}(\Lambda,\mathcal{S})$ for this observed-compatible class; it shares all other DGP restrictions with $\mathcal{R}(\Gamma,\mathcal{S})$. With the same conditioning and hidden variables, the minimal population levels obey $\Lambda_{\mathcal{S}} \leq \Gamma_{\mathcal{S}} \leq \Lambda^2_{\mathcal{S}}$ and $\mathcal{R}(g,\mathcal{S}) \subseteq \mathcal{M}(g,\mathcal{S}) \subseteq \mathcal{R}(g^2,\mathcal{S})$ (Appendix A.1). Thus $\Gamma_{\mathcal{S}}$ is a conservative population MSM input; fitted propensities are separate estimates. Under outcome sufficiency, $\Gamma_{\mathcal{Z}} \leq \Gamma_{\mathcal{S}}$ also gives latent membership in both classes at input $\Gamma_{\mathcal{S}}$; otherwise the relevant hidden variable for value coverage is $U^\dagger = (S,U)$.

3 Learned Representations Can Break Sensitivity Analysis

We work in the one-step setting, which isolates the representation-induced failure without the additional complications of long-horizon dynamics, and return to the multi-step setting in our method (Section 5) and experiments (Section 6).

The amplification mechanism. Equation (1) expresses the latent behavioural policy $\tilde{\pi}_b(a \mid z,u)$ as a mixture of per-state policies with mixing weights $P(s \mid z,u)$. When these weights depend on u, two

confounding sources can interact. The first is *per-state action confounding*: within each observed state, U shifts action probabilities, bounded by $\Gamma_{\mathcal{S}}$. The second is *state-selection confounding*: across states merged into the same latent code, U shifts which state is active. After compression, the evaluator no longer observes which original state generated $Z = z$, so this within-cell state identity acts as an additional source of action variation. When both are present, the two effects compound on the odds scale, and the compounding is the mechanism by which $\Gamma_{\mathcal{Z}}$ can exceed $\Gamma_{\mathcal{S}}$. Per-state action confounding is not necessary: heterogeneous action profiles and U-dependent within-cell composition can yield $\Gamma_{\mathcal{Z}} > \Gamma_{\mathcal{S}}$ even when $\Gamma_{\mathcal{S}} = 1$.

Theorem 1 (Compression can amplify action sensitivity)**.** *There exists a one-step confounded decision problem with sensitivity parameter $\Gamma_{\mathcal{S}}$ and a deterministic compression $\phi : \mathcal{S} \to \mathcal{Z}$ such that*

$$\Gamma_{\mathcal{Z}} \;>\; \Gamma_{\mathcal{S}}.$$

For the outcome-complete latent variable $U^\dagger = (S, U)$, its pairwise assignment-odds range satisfies $\Gamma^\dagger_{\mathcal{Z}} \geq \Gamma_{\mathcal{Z}}$. Consequently, the true compressed DGP lies outside $\mathcal{R}(\Gamma_{\mathcal{S}}, \mathcal{Z})$ defined with $U^\dagger$, and its pairwise-class coverage guarantee no longer applies. Pairwise amplification alone does not establish marginal-model violation or actual value noncoverage.

Proof sketch. We construct $|\mathcal{S}| = 3$, $|\mathcal{U}| = 2$, with compression $\phi(s_1) = \phi(s_2) = z_1$, $\phi(s_3) = z_3$. The DGP is chosen so that U is strongly predictive of which state is active within z_1 while per-state action confounding remains mild, yielding $\Gamma_{\mathcal{S}} = 12/7 \approx 1.71$. Compression merges a high-action state (s_1) with a low-action state (s_2), and because U flips which one dominates within z_1, the induced latent policy $\tilde{\pi}_b(1 \mid z_1, u)$ shifts sharply with u, giving $\Gamma_{\mathcal{Z}} = (67/23)^2 \approx 8.49$, an amplification of roughly $5\times$. Full numerical construction in Appendix A. □

Amplification from state selection alone. Compression can amplify action sensitivity even when the behavioural policy is unconfounded conditional on the raw state. Specifically, let $A \perp\!\!\!\perp U \mid S$, so $\Gamma_{\mathcal{S}} = 1$. Let $\phi(s_1) = \phi(s_2) = z$, let the behavioural action probabilities at (s_1, s_2) be $(0.9, 0.1)$, and let U reverse the within-cell state weights between $(0.9, 0.1)$ and $(0.1, 0.9)$. The induced latent propensities are $(0.82, 0.18)$, giving $\Gamma_{\mathcal{Z}} \approx 20.75$. The construction in Appendix A attains $\Gamma_{\mathcal{Z}} = \Gamma_{\mathcal{S}}\Gamma_{\text{sel}}$ in Proposition 7.

Remark 1 (Generality)**.** *Amplification requires only two ingredients: the compression merges states with different action profiles, and U shifts the within-cell state distribution, i.e., $P(S \mid Z, U) \neq P(S \mid Z)$. Any learned encoder that induces such state aggregation can exhibit the same failure, regardless of training objective.*

Section 6 demonstrates fitted action-sensitivity amplification with learned representations in Setting B and compares SPRL with task-aware baselines on IHDP and HalfCheetah.

4 When Is the Learned Representation Safe?

Theorem 1 shows that compression can amplify action sensitivity. We now identify when it does not.

Proposition 2 (Sufficient condition for sensitivity preservation)**.** *Let $Z = \phi(S)$ be a deterministic compression and assume positivity ($\pi_b(a \mid s, u) \in (0, 1)$ for supported (s, u) and each a). If*

$$U \;\perp\!\!\!\perp\; S \;\mid\; Z, \tag{5}$$

then $\Gamma_{\mathcal{Z}} \leq \Gamma_{\mathcal{S}}$.

Proof sketch (full proof in Appendix B). Under $U \perp\!\!\!\perp S \mid Z$, the mixing weights in Equation 1 reduce to $P(s \mid z)$, and Jensen's inequality applied to the concave function $h_\Gamma(x) := \Gamma x/(1 + (\Gamma - 1)x)$ (for $\Gamma \geq 1$) yields the bound. □

The condition $U \perp\!\!\!\perp S \mid Z$ says that, within each latent code, U provides no information about which original state is active, blocking the state-selection mechanism of Section 3. Proposition 2 is sufficient but not necessary. Appendix E gives a multiplicative decomposition characterizing when violations cause amplification.

Binary hidden confounding. A complementary sufficient condition is observable at the population level. For binary A, U, let $e(s) = P(A = 1 \mid S = s)$ and define the within-cell odds range $\Gamma_e(z) := \max_{s,s' \in \mathcal{S}_z} \text{odds}(e(s))/\text{odds}(e(s'))$, where $\mathcal{S}_z = \{s : \phi(s) = z,\ P(S = s) > 0\}$. Under positivity, for cells supported under both values of U,

$$\Gamma_{\mathcal{Z}}(z) \leq \Gamma_{\mathcal{S}} \Gamma_e(z).$$

Thus exact propensity homogeneity in every cell gives $\Gamma_{\mathcal{Z}} \leq \Gamma_{\mathcal{S}}$ without requiring $U \perp\!\!\!\perp S \mid Z$. Proposition 6 and its proof are in Appendix B.1, together with the limits for multivalued U.

For a common ϕ-measurable target and conditional-mean outcome sufficiency, either preservation condition gives population value coverage at pairwise or marginal input $\Gamma_{\mathcal{S}}$ (Corollary 5).

4.1 Continuity of Sensitivity Preservation

The exact preservation condition $U \perp\!\!\!\perp S \mid Z$ cannot be checked directly because U is unobserved, and a learned representation may satisfy it only approximately. The next result establishes that the condition is not an isolated point: amplification grows continuously with the conditional-independence violation Δ, so representations that are approximately safe incur bounded amplification. The rate of growth depends on the behavioural-policy positivity constant η and is informative as a quantitative bound only when η is bounded away from zero; we discuss its regime of validity after the statement.

Theorem 3 (Approximate sensitivity preservation)**.** *Let $(\mathcal{S}, \mathcal{A}, \mathcal{U}, P, \pi_b, r)$ be a confounded decision problem with discrete state space $\mathcal{S}$, and assume $\pi_b(a \mid s, u) \in [\eta, 1-\eta]$ for supported (s, u) and each a and some $\eta > 0$.* [1] *Let $Z = \phi(S)$ be a deterministic compression, and define the conditional independence violation*

$$\Delta \ := \max_{(z,u)\ \textit{supported}} \big\| P(S \mid Z{=}z,\, U{=}u) \ - \ P(S \mid Z{=}z) \big\|_1.$$

If $\Delta < \eta^3/(1-\eta)$,

$$\Gamma_{\mathcal{Z}} \ \leq \ \Gamma_{\mathcal{S}} \ + \ \frac{(1+\Gamma_{\mathcal{S}})(1-\eta)}{\eta^3 - (1-\eta)\Delta} \cdot \Delta. \tag{6}$$

A proof, decomposing the induced latent policy into a u-independent safe part and a Δ-bounded perturbation, is in Appendix C. **Regime of validity.** The bound Equation 6 is informative only when $\Delta < \eta^3/(1-\eta)$, a restrictive regime that tightens as $\eta \to 0$ (for example, 0.0011 at $\eta = 0.1$). Theorem 3 is therefore not a practical tool for certifying a given encoder; it is a continuity statement, with $\Delta = 0$ recovering Proposition 2 exactly and approximately safe representations incurring bounded amplification. Appendix C gives the full discussion.

Multi-step extension. Under dynamics-confounding separation, the per-step amplification bound (Proposition 7) extends across the horizon: the latent action-sequence likelihood ratio compounds as $\overline{\Gamma}_{\mathcal{S}}^{H} \cdot \prod_t \widetilde{\Gamma}_{\text{sel}}^{(t)}$ (Corollary 8), where $\widetilde{\Gamma}_{\text{sel}}^{(t)}$ is the history-conditional state-selection factor at step t that generalises $\Gamma_{\text{sel}}^{(t)}$ to the non-Markovian latent process. SPRL applies its observable propensity-homogeneity penalty at each step. The marginal factor bounds the history factor when $P(S_t \mid Z_{0:t}, A_{0:t-1}, U) = P(S_t \mid Z_t, U)$ (Remark 4, Appendix F). A policy-value statement additionally uses the sequential outcome-sufficiency analogue of Corollary 5.

5 Sensitivity-Preserving Representation Learning

Section 4 establishes that $U \perp\!\!\!\perp S \mid Z$ is sufficient for action-sensitivity preservation, and that Δ controls the rate of amplification (Theorem 3). Since U is unobserved, both conditions involve unmeasurable quantities. We therefore develop an observable proxy.

5.1 From the Safety Condition to an Observable Proxy

The safety condition $U \perp\!\!\!\perp S \mid Z$ cannot be checked from logged data because U is unobserved. We therefore target within-cell marginal-propensity homogeneity: for binary U, the odds range of

[1] For binary actions, two-sided positivity is equivalent to bounded log-propensity, a standard assumption in confounding-robust OPE [Kallus and Zhou, 2020]; see Appendix C for discussion.

$e(s) := P(A = 1 \mid S = s)$ bounds the extra action-sensitivity factor (Proposition 6). Penalising within-cell propensity heterogeneity targets the state-selection mechanism of Section 3, encouraging a propensity-level sufficiency property in the spirit of information-bottleneck representation learning [Tishby and Zaslavsky, 2015].

5.2 The SPRL Objective

Sensitivity-Preserving Representation Learning (SPRL) combines a dynamics prediction objective with a within-cell propensity homogeneity penalty.

Dynamics prediction loss. The encoder $\phi : \mathcal{S} \to \mathcal{Z}$ is trained jointly with a reward predictor g and, in the multi-step setting, a transition predictor f; write $d_Z = \dim \mathcal{Z}$:

$$\mathcal{L}_{\mathrm{dyn}}(\phi, f, g) = \mathbb{E}\Big[\tfrac{1}{d_Z}\big\|f(\phi(S_t), A_t) - \mathrm{sg}[\phi(S_{t+1})]\big\|_2^2 + \big(g(\phi(S_t), A_t) - R_t\big)^2\Big], \tag{7}$$

where $\mathrm{sg}[\cdot]$ denotes stop-gradient on the target branch [LeCun et al., 2022, Assran et al., 2023]. The transition squared error is averaged over d_Z coordinates; both prediction terms have unit weight. Predicting $\phi(S_{t+1})$ rather than S_{t+1} avoids incentivising reconstruction-level detail. Under dynamics-confounding separation (Section 2), the next raw state is conditionally independent of U given (S_t, A_t). Compression can still induce U-dependent, non-Markovian latent dynamics (Remark 4), so latent prediction supplies a task signal rather than a sensitivity guarantee. In the one-step setting, $\mathcal{L}_{\mathrm{dyn}}$ reduces to reward prediction.

Within-cell propensity homogeneity penalty. Let $\hat{e}(s)$ be an estimate of $e(s)$, obtained by pre-training a neural classifier on the logged (S, A) pairs and freezing it during encoder training. The penalty

$$\mathcal{L}_{\mathrm{prop}}(\phi) \;=\; \frac{\sum_{i \neq j} w_{ij}\,\big(\hat{e}(s_i) - \hat{e}(s_j)\big)^2}{\sum_{i \neq j} w_{ij}}, \qquad w_{ij} = \exp\Big(-\tfrac{\|\phi(s_i) - \phi(s_j)\|^2}{2\sigma^2}\Big), \tag{8}$$

uses Gaussian weights to emphasize pairs with nearby codes, excluding self-pairs. In each minibatch, σ is the median off-diagonal code distance, lower-bounded at 0.1 (Appendix H).

Full objective. The SPRL objective is

$$\min_{\phi,\, f,\, g} \;\mathcal{L}_{\mathrm{dyn}}(\phi, f, g) \;+\; \lambda_{\mathrm{prop}} \cdot \mathcal{L}_{\mathrm{prop}}(\phi), \tag{9}$$

where $\lambda_{\mathrm{prop}} \geq 0$ trades off dynamics prediction quality against fitted-propensity similarity. At $\lambda_{\mathrm{prop}} = 0$, SPRL reduces to a standard latent world model; as λ_{prop} grows, the encoder progressively avoids merging states with different fitted propensities.

The link between small $\mathcal{L}_{\mathrm{prop}}$ and population preservation is heuristic: the loss averages discrepancies in a frozen fitted teacher, whereas Proposition 6 uses the worst within-cell odds range of the true propensity. Appendix G gives the population pair-ratio identity and the limitations of the empirical loss. The implementation details are in Appendix H.

6 Experiments

We validate the theoretical results of Sections 3 and 4 and evaluate SPRL across two synthetic settings that isolate progressively harder failure modes. Setting A is a one-step problem with disjoint confounded and task-relevant features ($S_\Delta = \varnothing$), testing how state compression affects empirical target inclusion. Setting B is a multi-step confounded MDP (H=5) with overlapping features ($S_\Delta \neq \varnothing$), where encoding S_Δ improves dynamics prediction but admits the amplification mechanism of Theorem 1. Section 6.5 extends to IHDP and HalfCheetah; full tables can be found in Appendix J.

6.1 Experimental Setup

State feature structure. Following the covariate decomposition of Melnychuk et al. [2024], we partition states into noise $S_\varnothing$, action-predictive nuisance S_a (shifted by U), clean reward features S_r, and confounded dynamics S_Δ (shifted by U and used in transitions/rewards). Settings A and B set

Table 1: Setting A empirical target inclusion and samplewise interval width (K=100 seeds per strength, analytic target 0.700, nominal $\Lambda = e^{|w_u|}$). Incl is the number of seed intervals including the target; W/R is mean width divided by the samplewise Raw-State mean width. Widths remain descriptive when inclusion is low. SPRL is shown in bold.

		w_u=0.3		w_u=0.5		w_u=0.8	
Method	Type	Incl	W/R	Incl	W/R	Incl	W/R
Raw State	Reference	100/100	1.000	100/100	1.000	100/100	1.000
AE	Unsupervised	19/100	1.527	45/100	1.718	64/100	2.193
VAE	Unsupervised	0/100	1.177	39/100	1.182	100/100	1.190
Contrastive	Unsupervised	7/100	1.221	48/100	1.249	97/100	1.304
Balanced	Task-aware	100/100	0.997	100/100	0.999	100/100	1.001
LDM	Task-aware	100/100	0.988	100/100	0.991	100/100	0.994
SPRL (λ=1)	**Ours**	**100/100**	**1.001**	**100/100**	**1.016**	**100/100**	**1.053**

$S_\Delta = \varnothing$ and $S_\Delta \neq \varnothing$ respectively. The learner is never given the partition; our theoretical results hold without any such assumption. Settings A/B use $P(R \mid S, A)$ to isolate representation-induced confounding; Appendix D.5 gives a finite $H = 2$ outcome-confounded construction.

Baselines. We compare seven representations in Settings A/B, and four (Raw State, Balanced, LDM, SPRL) on IHDP/HalfCheetah. The learned encoders share a three-layer multi-layer perceptron (MLP) backbone with matched latent dimensions (d_Z=2 for Setting A and IHDP, d_Z=5 for Setting B, d_Z=4 for HalfCheetah) and the architectural details are given in Appendix H.2. **Raw State** (no compression) is the reference. **AE**, **VAE** [Kingma and Welling, 2013], and **Contrastive** [Chen et al., 2020] are unsupervised baselines. **Balanced** [Shalit et al., 2017] adds a Maximum Mean Discrepancy (MMD) penalty [Gretton et al., 2012] matching $P(Z \mid A{=}0)$ and $P(Z \mid A{=}1)$ to reward prediction. **LDM** learns a latent transition and reward model [Hafner et al., 2023] and is equivalent to SPRL with λ_{prop}=0. **SPRL** (ours) adds the propensity homogeneity penalty $\mathcal{L}_{\text{prop}}$ to LDM.

Evaluation metrics. In Setting A, we evaluate empirical target inclusion directly: encode states $Z = \phi(S)$, compute sign-aware samplewise Horvitz–Thompson inverse probability weighting (HT-IPW) intervals at nominal marginal input $\Lambda = \exp(|w_u|)$, and check whether the analytic target lies within the bounds (Appendix H.3). Because latent propensities are fitted and the nominal input need not hold after compression, these counts are not guaranteed confidence coverage. In Setting B, multi-step IPW bounds compound conservatively over the horizon and cannot discriminate methods, so we directly measure the *amplification ratio* $\Gamma_{\mathcal{Z}}/\Gamma_{\mathcal{S}}$ via logistic regression of A on $(Z, U, Z \times U)$ with oracle access to U. We summarize each seed by its median per-sample ratio (typical fitted amplification) and 95th percentile (upper tail); details in Appendix H.4. We additionally report **U-AUC**, the linear-probe AUC for predicting U from Z, as a diagnostic for whether a representation suppresses U-information globally (low U-AUC) or retains it (high U-AUC); Proposition 2 implies that action-sensitivity preservation does not require low U-AUC. We report the means of seed-level summaries across K=100 seeds, with paired uncertainty for the main comparisons.

6.2 Setting A: Empirical Target Inclusion Under Disjoint Features

Setting A uses a one-step confounded decision problem ($d_S = 12$) with confounder $U \sim \text{Bernoulli}(0.5)$ shifting S_a and biasing the behavioural policy; the reward depends only on S_r and A. We vary $w_u \in \{0.3, 0.5, 0.8\}$, giving true confounding strengths $\Gamma_{\mathcal{S}} \in \{1.35, 1.65, 2.23\}$. Full DGP parameters are presented in Appendix H.5.

Table 1 reports empirical target inclusion and width ratios. Raw State, Balanced, LDM, and SPRL each include 0.700 in all 100/100 runs at every strength. The unsupervised results vary with w_u: for example, VAE increases from 0/100 at 0.3 to 100/100 at 0.8. SPRL's mean width is 1.001/1.016/1.053 times Raw State, but exceeds LDM by 1.3%, 2.6%, and 5.9%, respectively. Thus SPRL matches LDM's empirical target inclusion with a measured width cost. Figure 3 displays the mean endpoints.

Table 2: Setting B fitted per-step action-sensitivity amplification $\Gamma_{\mathcal{Z}}/\Gamma_{\mathcal{S}}$ and U-AUC across confounding strengths (K=100 seeds). For each seed, we compute the median and 95th-percentile of per-sample odds ratios; Med and p95 are the means of these seed-level statistics across K=100 seeds. Raw State is fixed at 1 as the population reference; SPRL is shown in bold.

		w_u=0.3 ($\Gamma_{\mathcal{S}}$=1.35)			w_u=0.5 ($\Gamma_{\mathcal{S}}$=1.65)			w_u=0.8 ($\Gamma_{\mathcal{S}}$=2.23)		
Method	Type	Med	p95	U-AUC	Med	p95	U-AUC	Med	p95	U-AUC
Raw State	Reference	1.00	1.00	0.990	1.00	1.00	0.990	1.00	1.00	0.991
AE	Unsupervised	1.37	2.25	0.986	1.32	2.19	0.986	1.29	2.16	0.987
VAE	Unsupervised	1.19	10.23	0.987	1.18	8.76	0.988	1.05	5.69	0.989
Contrastive	Unsupervised	1.60	4.99	0.984	1.50	4.45	0.985	1.38	4.09	0.986
Balanced	Task-aware	4.45	9.33	0.595	4.30	8.89	0.579	4.17	9.10	0.574
LDM	Task-aware	2.88	4.92	0.968	2.83	4.76	0.969	2.93	5.01	0.961
SPRL (λ=1)	**Ours**	**1.45**	**2.66**	**0.986**	**1.26**	**2.32**	**0.991**	**1.27**	**2.48**	**0.990**

6.3 Setting B: Amplification Under Overlapping Features

Table 2 reports the median and 95th-percentile per-step action-sensitivity amplification ratio $\Gamma_{\mathcal{Z}}/\Gamma_{\mathcal{S}}$ across K=100 seeds at three confounding strengths, together with U-decodability. All learned representations show fitted amplification ($\Gamma_{\mathcal{Z}}/\Gamma_{\mathcal{S}} > 1$), consistent with the amplification mechanism of Theorem 1 in the multi-step setting.

Unsupervised methods amplify modestly in median (AE $1.37\times$, Contrastive $1.60\times$) but VAE shows a large fitted upper tail (median 1.19, p95 10.23) at $w_u = 0.3$. Task-aware methods are not automatically safe: Balanced amplifies nearly five-fold ($4.45\times$, p95 9.33) despite having the lowest U-AUC of any method (0.595), illustrating that low linear U-decodability does not establish Proposition 2's $U \perp\!\!\!\perp S \mid Z$. Distributional balancing need not prevent within-cell state selection. LDM has median amplification $2.88\times$, consistent with retaining U-correlated transition-predictive features. SPRL reduces fitted amplification (median $1.45\times$, p95 2.66) relative to the task-aware baselines, while preserving high U-decodability, consistent with the mechanism in Proposition 6.

We measure fitted per-step action-sensitivity amplification in multi-step settings because H-step IPW bounds compound conservatively. Setting B's refreshed nuisance coordinates depend on U (Appendix H.5.2), so this experiment does not test the transition-separated action-sequence bound of Corollary 8.

6.4 Ablation: Effect of λ_{prop}

We ablate $\lambda_{\text{prop}} \in \{0, 0.5, 1.0, 2.0, 5.0\}$ at $w_u = 0.3$, $K = 100$ seeds; $\lambda_{\text{prop}} = 0$ recovers LDM. Figure 2(a) shows that the main reduction in median fitted amplification occurs between $\lambda = 0$ and $\lambda = 1$ (2.88 to 1.45; p95 falls from 4.92 to 2.66 in Table 2). Further increases yield diminishing returns while raw latent-transition MSE rises; its magnitude depends on representation scale. We use $\lambda_{\text{prop}} = 1$ as a default, selected using the oracle diagnostic; Appendix I.3 studies logged-data selection. Across confounding strengths, $\lambda = 1$ is the smallest value achieving Med $\Gamma_{\mathcal{Z}}/\Gamma_{\mathcal{S}} < 1.5$ at all three $w_u \in \{0.3, 0.5, 0.8\}$ (Appendix I.3). Figure 2(b,c) shows lower median fitted amplification than LDM across w_u, with U-AUC close to Raw State (≈ 0.99) versus LDM's ≈ 0.96. This supports the design argument of Appendix G: SPRL reduces fitted amplification while preserving U-information in Z, consistent with targeting within-cell state selection rather than scrubbing U. Controlled objective and latent-dimension ablations appear in Appendix I.1–I.2. SPRL retains lower fitted amplification than LDM under the tested Gaussian teacher perturbations (Appendix I.5).

6.5 Semi-Synthetic Validation

We evaluate on two semi-synthetic benchmarks pairing real covariate structure or physics-derived state trajectories with injected confounding: IHDP [Hill, 2011] (25 clinical covariates, one-step, $S_\Delta = \varnothing$) and Minari's `mujoco/halfcheetah/medium-v0` dataset in the D4RL benchmark family [Fu et al., 2020] (17-dimensional states, multi-step, $S_\Delta \neq \varnothing$). Table 3 reports median amplification across the benchmarks; full per-method tables and DGP specifications are given in Appendix J.

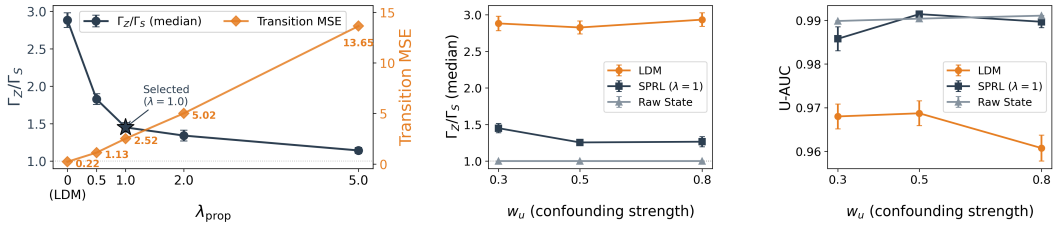


Figure 2: (a) λ_{prop} ablation in Setting B ($w_u = 0.3$, $K = 100$): the mean across seeds of within-seed median amplification (± 1 SE) and raw latent-transition MSE; the latter depends on representation scale. The star marks $\lambda_{\text{prop}} = 1.0$. (b) The same median amplification statistic and (c) mean U-AUC across $w_u \in \{0.3, 0.5, 0.8\}$ (± 1 SE).

Table 3: Mean across $K = 100$ seeds of each seed's median fitted odds-ratio diagnostic divided by nominal $e^{|w_u|}$, at $w_u = 0.5$. IHDP uses an additive readout; Setting B and HalfCheetah use interaction readouts. Setting B fixes Raw State at 1; non-unit IHDP/HalfCheetah Raw-State values reflect fitted-readout error and nominal normalization. SPRL is shown in bold. Paired uncertainty and p95 appear in Appendix J; Setting B p95 is in Table 2.

Method	Setting B	IHDP	HalfCheetah
Raw State	1.00	1.08	0.88
Balanced	4.30	1.49	1.20
LDM	2.83	1.36	1.62
SPRL	1.26	0.73	1.10

SPRL lowers the fitted median diagnostic relative to LDM in all three settings. On IHDP, its paired difference from LDM is -0.63 (95% interval $[-0.80, -0.46]$) at $w_u = 0.5$, and it also improves this diagnostic over Balanced at each strength. Its IHDP intervals are wider and reward errors larger than both comparators. On episode-disjoint HalfCheetah, the paired median difference from LDM is -0.52 $[-0.69, -0.35]$; the difference from Balanced is -0.10 $[-0.21, 0.01]$, which is inconclusive. SPRL also has higher reward-prediction error than both comparators on HalfCheetah. Its injected binary action does not cause the recorded transitions or rewards, so this benchmark tests amplification on physics-derived trajectories rather than binary-action policy value.

7 Related Work

Confounding-robust off-policy evaluation. A growing body of work develops sensitivity analysis for sequential decisions under unobserved confounding [Kallus and Zhou, 2020, Namkoong et al., 2020, Wang et al., 2021, Bruns-Smith, 2021]. Extensions address deep RL [Li et al., 2025], eligibility traces [Zhang and Bareinboim, 2025], and efficient policy learning [Hess et al., 2026]. Their guarantees are formulated in the raw observation space; we study when they transfer through a learned representation. Proximal and latent-confounder OPE seek identification through proxy structure [Bennett and Kallus, 2024, Bennett et al., 2020]; transition-robust OPE bounds transition perturbations [Bennett et al., 2024].

Sensitivity models. Rosenbaum's pairwise odds model [Rosenbaum, 2002] and Tan's marginal propensity-odds model [Tan, 2006] use different anchors; their population relationship is given by Zhao et al. [2019, Section 7.2, Proposition 7.1]. Generalized, continuous-outcome, and sharp outcome-aware variants are studied by Frauen et al. [2024], Melnychuk et al. [2023], Dorn et al. [2022], Zhang et al. [2025]. These works target static settings; we study sequential state aggregation.

State representation learning for RL. State representation learning for RL includes latent dynamics models [Hafner et al., 2023] (our LDM baseline), Bellman-complete OPE representations [Chang et al., 2022], causal representation learning approaches [Schölkopf et al., 2021, Wang et al., 2024, 2025], and balanced representations from treatment effect estimation [Johansson et al., 2016, Shalit et al., 2017] (our Balanced baseline, adapted to RL using MMD); see Appendix K for an extended discussion. These methods optimize for prediction, balance, or structural identifiability. We study

a different objective: whether the learned representation preserves the action-sensitivity class and, under outcome sufficiency, downstream value coverage.

Representation-induced causal bias. Melnychuk et al. [2024] study representation-induced confounding bias (RICB) in static CATE estimation: low-dimensional representations can introduce bias by discarding observed confounders. Their covariate decomposition informs our experimental design. Our setting studies sequential pairwise action sensitivity, where state-selection effects can compound over the horizon (Corollary 8). Our failure mode operates through within-cell state selection by the unobserved confounder, characterized by U-dependence in $P(S \mid Z, U)$ (Section 3). Under outcome sufficiency this has value-coverage implications, rather than being only a point-estimator effect. SPRL's within-cell propensity homogeneity penalty targets this state-selection mechanism directly.

8 Conclusion

Causal sensitivity analysis provides coverage guarantees in the observation space in which bounds are computed, but offline RL pipelines typically apply these bounds in a learned latent space rather than the raw state. We showed that this gap matters: state aggregation under hidden confounding can amplify the effective pairwise action-sensitivity parameter in the latent space, invalidating the nominal class. Under outcome sufficiency and a common target, separate marginal-model constructions show actual value noncoverage. We identified the underlying state-selection mechanism, gave a sufficient condition for action-sensitivity preservation with a matching continuity result, established a binary-confounder propensity-homogeneity condition, and proposed SPRL, which targets the amplification mechanism via a within-cell propensity homogeneity penalty. Across Setting B, IHDP, and HalfCheetah, SPRL lowers fitted amplification relative to LDM; its Setting A and IHDP intervals include the analytic targets, with wider intervals than LDM and reward-prediction costs on IHDP and HalfCheetah.

Limitations and future work. Our action-sensitivity theory uses a pairwise, Rosenbaum-type population constraint; extending the results to other classes, including f-sensitivity, requires separate analysis. SPRL's empirical average penalty does not certify population propensity homogeneity or value coverage. Exact homogeneity preserves action sensitivity for binary U, but can fail for multivalued U (Appendix B.1); the continuity bound of Theorem 3 applies in a restrictive local regime. Our amplification metric uses oracle access to the unobserved confounder for validation, following standard practice in confounding-robust OPE [Kallus and Zhou, 2020, Namkoong et al., 2020, Frauen et al., 2024]. Continuous actions, time-varying confounding, and joint propensity-encoder training are natural next steps. Appendix I.5 tests sensitivity to Gaussian teacher perturbations.

Acknowledgments and Disclosures

Funding. This work received no external funding. **Competing interests.** The authors declare no competing interests.

References

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15619–15629, 2023.

Andrew Bennett and Nathan Kallus. Proximal reinforcement learning: Efficient off-policy evaluation in partially observed markov decision processes. *Oper. Res.*, 72(3):1071–1086, May 2024. ISSN 0030-364X. doi: 10.1287/opre.2021.0781. URL `https://doi.org/10.1287/opre.2021.0781`.

Andrew Bennett, Nathan Kallus, Lihong Li, and Ali Mousavi. Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders. *arXiv preprint arXiv:2007.13893*, 2020.

Andrew Bennett, Nathan Kallus, Miruna Oprescu, Wen Sun, and Kaiwen Wang. Efficient and sharp off-policy evaluation in robust markov decision processes. In *Advances in Neural Information Processing Systems*, 2024.

David A Bruns-Smith. Model-free and model-based policy evaluation when causality is uncertain. In *International Conference on Machine Learning*, pages 1116–1126. PMLR, 2021.

Jonathan D. Chang, Kaiwen Wang, Nathan Kallus, and Wen Sun. Learning bellman complete representations for offline policy evaluation. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

Xiaocong Chen, Siyu Wang, and Lina Yao. Energy-guided diffusion sampling for long-term user behavior prediction in reinforcement learning-based recommendation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 322–331, 2025a. doi: 10.1145/3746252.3761076. URL https://arxiv.org/abs/2510.12815.

Xiaocong Chen, Siyu Wang, Tong Yu, and Lina Yao. Diffusion policies for risk-averse behavior modeling in offline reinforcement learning. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 567–574, 2025b. doi: 10.1109/IROS60139.2025.11245819. URL https://arxiv.org/abs/2403.17646.

Xiaocong Chen, Siyu Wang, and Lina Yao. Continuity-regularized flow matching for offline reinforcement learning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=sV3bCyzAvH.

Jacob Dorn, Kevin Guo, and Nathan Kallus. Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding. *arXiv preprint arXiv:2112.11449*, 2022.

Dennis Frauen, Fergus Imrie, Alicia Curth, Valentyn Melnychuk, Stefan Feuerriegel, and Mihaela van der Schaar. A neural framework for generalized causal sensitivity analysis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ikX6D1oM1c.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Hang Gao, Jiangmeng Li, Wenwen Qiang, Lingyu Si, Bing Xu, Changwen Zheng, and Fuchun Sun. Robust causal graph representation learning against confounding effects. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i6.25925. URL https://doi.org/10.1609/aaai.v37i6.25925.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Konstantin Hess, Dennis Frauen, Valentyn Melnychuk, and Stefan Feuerriegel. Efficient and sharp off-policy learning under unobserved confounding. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=7nTKiJLkWS.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Biwei Huang, Chaochao Lu, Liu Leqi, José Miguel Hernández-Lobato, Clark Glymour, Bernhard Schölkopf, and Kun Zhang. Action-sufficient state representation learning for control with structural constraints. In *International Conference on Machine Learning*, pages 9260–9279. PMLR, 2022.

Ying Jin, Zhimei Ren, and Emmanuel J Candès. Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proceedings of the National Academy of Sciences*, 120(6): e2214889120, 2023.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029. PMLR, 2016.

Nathan Kallus and Angela Zhou. Confounding-robust policy evaluation in infinite-horizon reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22293–22304. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/fd4f21f2556dad0ea8b7a5c04eabebda-Paper.pdf.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Mingxuan Li, Junzhe Zhang, and Elias Bareinboim. Confounding robust deep reinforcement learning: A causal approach. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=9fUr5iFU9j.

Haohong Lin, Wenhao Ding, Jian Chen, Laixi Shi, Jiacheng Zhu, Bo Li, and Ding Zhao. Because: Bilinear causal representation for generalizable offline model-based reinforcement learning. *Advances in Neural Information Processing Systems*, 37:114510–114552, 2024.

Yu-Ren Liu, Biwei Huang, Zhengmao Zhu, Honglong Tian, Mingming Gong, Yang Yu, and Kun Zhang. Learning world models with identifiable factorization. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6449–6459, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Partial counterfactual identification of continuous outcomes with a curvature sensitivity model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Bounds on representation-induced confounding bias for treatment effect estimation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=d3xKPQVjSc.

Hongseok Namkoong, Ramtin Keramati, Steve Yadlowsky, and Emma Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Doina Precup, Richard S. Sutton, and Satinder P. Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, page 759–766, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1558607072.

Paul R Rosenbaum. Observational studies. In *Observational studies*, pages 1–17. Springer, 2002.

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.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pages 3076–3085. PMLR, 2017.

Zhiqiang Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2139–2148, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/thomasa16.html.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. Ieee, 2015.

Lingxiao Wang, Zhuoran Yang, and Zhaoran Wang. Provably efficient causal reinforcement learning with confounded observational data. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=tUeeRzMXJZ.

Siyu Wang, Xiaocong Chen, and Lina Yao. On causally disentangled state representation learning for reinforcement learning based recommender systems. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2390–2399, 2024.

Siyu Wang, Xiaocong Chen, and Lina Yao. Policy-guided causal state representation for offline reinforcement learning recommendation. In *Proceedings of the ACM on Web Conference 2025*, pages 402–412, 2025.

Siyu Wang, Xiaocong Chen, Mingming Gong, Yong Li, Quan Z. Sheng, and Lina Yao. Reward-preserving counterfactual state editing for offline reinforcement learning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/66163.

Hangtong Xu, Yuanbo Xu, Chaozhuo Li, and Fuzhen Zhuang. Causal structure representation learning of unobserved confounders in latent space for recommendation. *ACM Trans. Inf. Syst.*, 43(6), September 2025. ISSN 1046-8188. doi: 10.1145/3731447. URL https://doi.org/10.1145/3731447.

Junzhe Zhang and Elias Bareinboim. Causal eligibility traces for confounding robust off-policy evaluation. In Silvia Chiappa and Sara Magliacane, editors, *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pages 4933–4942. PMLR, 21–25 Jul 2025. URL https://proceedings.mlr.press/v286/zhang25d.html.

Yi Zhang, Wenfu Xu, and Zhiqiang Tan. Enhanced marginal sensitivity model and bounds. *arXiv preprint arXiv:2504.08301*, 2025.

Qingyuan Zhao, Dylan S. Small, and Bhaswar B. Bhattacharya. Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(4):735–761, 2019. doi: 10.1111/rssb.12327.

Broader Impact

Offline reinforcement learning is increasingly applied in high-stakes domains such as healthcare, education, and autonomous systems, where errors in policy evaluation translate directly into harmful decisions. Causal sensitivity analysis is a principled tool for reasoning about policy value under unobserved confounding, but its guarantees are only meaningful if the sensitivity model holds at the level at which bounds are computed. Our results show that a standard engineering step, learning a state representation, can silently break this assumption and produce anti-conservative intervals that

systematically miss the true policy value. Practitioners who trust such intervals may deploy policies that appear validated but are not.

SPRL offers a practical mitigation for representation-induced action-sensitivity amplification in the standard offline RL pipeline. A small SPRL penalty does not certify propensity homogeneity or coverage, so SPRL-based intervals should not be treated as validated on that basis alone. More broadly, our work highlights a general principle: representation learning objectives should be chosen to match the downstream inferential task. Objectives tuned for prediction, reconstruction, or distributional balance need not preserve the causal or statistical properties that subsequent analyses rely on. We hope this motivates further work on representation learning that is accountable to the validity of downstream causal reasoning, particularly in domains where decisions affect individuals.

A Full Proof of Theorem 1

Proof. We construct the counterexample in full detail.

Problem specification. Let $\mathcal{S} = \{s_1, s_2, s_3\}$, $\mathcal{A} = \{0, 1\}$, $\mathcal{U} = \{0, 1\}$. The joint distribution $P(S, U)$:

	$u = 0$	$u = 1$	$P(s)$
s_1	0.40	0.05	0.45
s_2	0.05	0.40	0.45
s_3	0.05	0.05	0.10
$P(u)$	0.50	0.50	1.00

The behavioural policy $\pi_b(a = 1 \mid s, u)$:

	$u = 0$	$u = 1$
s_1	0.80	0.70
s_2	0.30	0.20
s_3	0.50	0.50

True confounding strength. The per-state odds ratios:

$$s_1: \quad \frac{\text{odds}(0.80)}{\text{odds}(0.70)} = \frac{4}{7/3} = \frac{12}{7} \approx 1.714,$$

$$s_2: \quad \frac{\text{odds}(0.30)}{\text{odds}(0.20)} = \frac{3/7}{1/4} = \frac{12}{7} \approx 1.714,$$

$$s_3: \quad \frac{\text{odds}(0.50)}{\text{odds}(0.50)} = 1.$$

Therefore $\Gamma_{\mathcal{S}} = 12/7$.

Compression. $\phi(s_1) = \phi(s_2) = z_1$, $\phi(s_3) = z_3$.

Mixing weights.

$$P(z_1, u{=}0) = 0.40 + 0.05 = 0.45, \qquad P(z_1, u{=}1) = 0.05 + 0.40 = 0.45.$$

$$P(s_1 \mid z_1, u{=}0) = \frac{8}{9}, \qquad P(s_2 \mid z_1, u{=}0) = \frac{1}{9},$$

$$P(s_1 \mid z_1, u{=}1) = \frac{1}{9}, \qquad P(s_2 \mid z_1, u{=}1) = \frac{8}{9}.$$

Induced policy. Using exact fractions:

$$\tilde{\pi}_b(1 \mid z_1, u{=}0) = \frac{8}{9} \cdot \frac{4}{5} + \frac{1}{9} \cdot \frac{3}{10} = \frac{64}{90} + \frac{3}{90} = \frac{67}{90},$$
$$\tilde{\pi}_b(1 \mid z_1, u{=}1) = \frac{1}{9} \cdot \frac{7}{10} + \frac{8}{9} \cdot \frac{1}{5} = \frac{7}{90} + \frac{16}{90} = \frac{23}{90}.$$

Effective confounding.

$$\operatorname{odds}\left(\frac{67}{90}\right) = \frac{67}{23}, \qquad \operatorname{odds}\left(\frac{23}{90}\right) = \frac{23}{67}.$$
$$\Gamma_{\mathcal{Z}} = \frac{67/23}{23/67} = \left(\frac{67}{23}\right)^2 = \frac{4489}{529} \approx 8.49.$$

Since $\Gamma_{\mathcal{Z}} = 4489/529 > 12/7 = \Gamma_{\mathcal{S}}$, the claimed result holds. The amplification factor is $\Gamma_{\mathcal{Z}}/\Gamma_{\mathcal{S}} = (4489 \times 7)/(529 \times 12) \approx 4.95$. □

One-way loss-of-guarantee implication. After compression, the outcome-complete hidden variable is $U^\dagger = (S, U)$ whenever discarded S remains outcome-relevant. Let $\Gamma^\dagger_{\mathcal{Z}}$ be the pairwise odds range of $P(A = 1 \mid Z = z, U^\dagger = (s, u)) = \pi_b(1 \mid s, u)$ over compatible (s, u). For each u, the averaged propensity $\widetilde{\pi}_b(1 \mid z, u)$ lies between its full-$U^\dagger$ extrema; monotonicity of odds therefore gives $\Gamma_{\mathcal{Z}} \leq \Gamma^\dagger_{\mathcal{Z}}$. Hence $\Gamma_{\mathcal{Z}} > \Gamma_{\mathcal{S}}$ implies that the value-preserving compressed DGP lies outside the pairwise class $\mathcal{R}(\Gamma_{\mathcal{S}}, \mathcal{Z})$. This does not by itself imply exclusion from $\mathcal{M}(\Gamma_{\mathcal{S}}, \mathcal{Z})$ or value noncoverage. The reverse preservation implication requires outcome sufficiency or direct control of $\Gamma^\dagger_{\mathcal{Z}}$.

State selection alone. The behavioural policy can be unconfounded at every raw state while compression creates arbitrarily large latent confounding. Let $\pi_b(1 \mid s, u) = \pi_b(1 \mid s)$ equal $(p, 1-p)$ at (s_1, s_2), and let $P(s_1 \mid z, u) = (p, 1-p)$ for $u = (0, 1)$. Then $\Gamma_{\mathcal{S}} = 1$; at $p = 0.9$, the latent propensities are $(0.82, 0.18)$ and $\Gamma_{\mathcal{Z}} \approx 20.75$. Moreover, $\Gamma_{\mathcal{Z}} \to \infty$ as $p \uparrow 1$, while positivity holds for every $p < 1$. Proposition 7 is tight: Factor 1 equals one and $\Gamma_{\mathcal{Z}} = \Gamma_{\text{sel}} = \Gamma_{\mathcal{S}}\Gamma_{\text{sel}}$.

A.1 Pairwise and Marginal Sensitivity Levels

Fix an observed variable X, a hidden variable U, and an observed-compatible population assignment law with $p_x(u) = P(A = 1 \mid X = x, U = u) \in (0, 1)$ and $e_X(x) = P(A = 1 \mid X = x) = \mathbb{E}[p_x(U) \mid X = x]$. Take extrema only over supported (x, u), using essential suprema when needed. Define

$$\Gamma_X = \sup_{x,u,u'} \frac{\operatorname{odds}(p_x(u))}{\operatorname{odds}(p_x(u'))}, \qquad \Lambda_X = \sup_{x,u} \max\left\{ \frac{\operatorname{odds}(p_x(u))}{\operatorname{odds}(e_X(x))}, \frac{\operatorname{odds}(e_X(x))}{\operatorname{odds}(p_x(u))} \right\}.$$

For each x, $e_X(x)$ lies between the smallest and largest supported $p_x(u)$. Since odds is increasing, every odds-to-e_X ratio is at most the largest within-x pairwise ratio; hence $\Lambda_X \leq \Gamma_X$. Conversely, each of the two odds-to-e_X ratios is at most Λ_X, so their product gives $\Gamma_X \leq \Lambda_X^2$. For observed-compatible DGP classes this proves $\mathcal{R}(g, X) \subseteq \mathcal{M}(g, X) \subseteq \mathcal{R}(g^2, X)$ for $g \geq 1$. This is the population comparison in Zhao et al. [2019, Section 7.2, Proposition 7.1], specialized to the same X and U. It cannot compare an averaged-U $\Gamma_{\mathcal{Z}}$ with a marginal level defined using $U^\dagger = (S, U)$, nor does it apply with a fitted propensity substituted for e_X.

Policy extension for mixture proofs. For supported s, fix a version of the behavior kernel and define

$$\pi_b^\circ(1 \mid s, u) := \begin{cases} \pi_b(1 \mid s, u), & (s, u) \text{ supported}, \\ e(s), & \text{otherwise}, \end{cases} \qquad \pi_b^\circ(0 \mid s, u) := 1 - \pi_b^\circ(1 \mid s, u).$$

The extension agrees with the population law almost surely. Because $e(s)$ lies within the supported propensity range at s, it preserves the within-state pairwise odds bound, the marginal level, and any two-sided positivity bound. Appendices C–F use this fixed extension whenever a mixture evaluates a policy at a state–confounder pair outside support. Virtual state-selection factors are defined relative to this extension; no full-support assumption is needed. For continuous variables, the ranges and identities are essential and hold almost surely.

B Full Proof of Proposition 2

Lemma 4 (Mixtures preserve odds-ratio bounds under common weights)**.** *Let $p_i, q_i \in (0,1)$ and weights $w_i \geq 0$ with $\sum_i w_i = 1$. Suppose for some $\Gamma \geq 1$:*

$$\frac{\mathrm{odds}(p_i)}{\mathrm{odds}(q_i)} \leq \Gamma \qquad \textit{for all } i.$$

Then the weighted averages $\bar{p} := \sum_i w_i p_i$ and $\bar{q} := \sum_i w_i q_i$ satisfy $\mathrm{odds}(\bar{p})/\mathrm{odds}(\bar{q}) \leq \Gamma$.

Proof. The assumption is equivalent to $p_i \leq h_\Gamma(q_i)$ where $h_\Gamma(x) := \Gamma x/(1+(\Gamma{-}1)x)$. We verify concavity:

$$h''_\Gamma(x) = \frac{-2\Gamma(\Gamma{-}1)}{(1+(\Gamma{-}1)x)^3} \leq 0 \quad \text{for } \Gamma \geq 1.$$

Applying the pointwise bound and Jensen's inequality:

$$\bar{p} = \sum_i w_i p_i \leq \sum_i w_i h_\Gamma(q_i) \leq h_\Gamma\Big(\sum_i w_i q_i\Big) = h_\Gamma(\bar{q}).$$

Since $\bar{p} \leq h_\Gamma(\bar{q})$ is equivalent to $\mathrm{odds}(\bar{p})/\mathrm{odds}(\bar{q}) \leq \Gamma$, the claim follows. □

Proof of Proposition 2. Fix supported $(z,u), (z,u')$. Under $U \perp\!\!\!\perp S \mid Z$, $P(s \mid z,u) = P(s \mid z)$ for all s, z, u. Therefore:

$$\tilde{\pi}_b(1 \mid z,u) = \sum_{s:\phi(s)=z} w_s\, p_s, \qquad \tilde{\pi}_b(1 \mid z,u') = \sum_{s:\phi(s)=z} w_s\, q_s,$$

where $w_s := P(s \mid z)$, $p_s := \pi_b(1 \mid s,u)$, $q_s := \pi_b(1 \mid s,u')$.

By the observed-space bound, $\mathrm{odds}(p_s)/\mathrm{odds}(q_s) \leq \Gamma_{\mathcal{S}}$ for all s. By positivity, $p_s, q_s \in (0,1)$. Applying Lemma 4:

$$\frac{\mathrm{odds}(\tilde{\pi}_b(1 \mid z,u))}{\mathrm{odds}(\tilde{\pi}_b(1 \mid z,u'))} \;\leq\; \Gamma_{\mathcal{S}}.$$

Since z, u, u' were arbitrary, $\Gamma_{\mathcal{Z}} \leq \Gamma_{\mathcal{S}}$. □

Corollary 5 (Value coverage under outcome sufficiency)**.** *Assume positivity, deterministic ϕ, a ϕ-measurable target policy, and conditional-mean sufficiency $r(s,a,u) = \bar{r}(\phi(s),a,u)$ for every (s,a,u). If $\Gamma_{\mathcal{Z}} \leq \Gamma_{\mathcal{S}}$, then every population mean-value interval valid for the pairwise class $\mathcal{R}(\Gamma_{\mathcal{S}}, \mathcal{Z})$ or the marginal class $\mathcal{M}(\Gamma_{\mathcal{S}}, \mathcal{Z})$ covers the original-DGP value $V(\pi_e)$. Full potential-outcome sufficiency $Y(a) \perp\!\!\!\perp S \mid (Z,U)$ is a stronger sufficient condition.*

Proof of Corollary 5. Marginalize S to obtain the induced assignment mechanism $\widetilde{\pi}_b(a \mid z,u)$ and define the observed-compatible reward kernel

$$P^\dagger(R \mid z,a,u) = \sum_{s:\phi(s)=z} P(R \mid s,a,u) P(s \mid z,a,u).$$

Conditional-mean outcome sufficiency makes its mean $\bar{r}(z,a,u)$ independent of the discarded state, and ϕ-measurability makes its policy value equal to the original $V(\pi_e)$. Since $\Gamma_{\mathcal{Z}} \leq \Gamma_{\mathcal{S}}$, this value-preserving compressed DGP belongs to $\mathcal{R}(\Gamma_{\mathcal{S}}, \mathcal{Z})$; by Appendix A.1, it also belongs to $\mathcal{M}(\Gamma_{\mathcal{S}}, \mathcal{Z})$. The stated population intervals therefore cover. □

B.1 The Binary-Confounder Propensity Bound

Proposition 6 (Binary-confounder propensity bound)**.** *Let $Z = \phi(S)$ be a deterministic compression of a discrete state space, with binary A, U and $\pi_b(a \mid s,u) \in (0,1)$ for supported (s,u) and each a. For a cell z with positive support under both values of U, write $\mathcal{S}_z = \{s : \phi(s) = z,\ P(S=s) > 0\}$, let $e(s) = P(A=1 \mid S=s)$ be the true observed propensity, and define*

$$\Gamma_e(z) := \max_{s,s' \in \mathcal{S}_z} \frac{\mathrm{odds}(e(s))}{\mathrm{odds}(e(s'))}.$$

If $\Gamma_{\mathcal{Z}}(z)$ is the pairwise action-odds level of the induced policy at z, then

$$\Gamma_{\mathcal{Z}}(z) \leq \Gamma_{\mathcal{S}} \Gamma_e(z).$$

Proof of Proposition 6. Fix a cell z as in the proposition and restrict sums to $\mathcal{S}_z$. The posterior odds below are written for interior probabilities; the cross-multiplied inequalities extend to boundary cases by continuity. Write

$$x_s := P(U = 1 \mid S = s, A = 1), \qquad y_s := P(U = 1 \mid S = s, A = 0).$$

For binary U, the 2×2 cross-ratio identity gives

$$\frac{\mathrm{odds}(x_s)}{\mathrm{odds}(y_s)} = \frac{\mathrm{odds}\{\pi_b(1 \mid s, U = 1)\}}{\mathrm{odds}\{\pi_b(1 \mid s, U = 0)\}},$$

so this ratio and its reciprocal are bounded by $\Gamma_{\mathcal{S}}$. Let

$$\alpha_s := P(s \mid z, A = 1), \qquad \beta_s := P(s \mid z, A = 0).$$

Bayes' rule gives

$$\frac{\alpha_s}{\beta_s} = \frac{\mathrm{odds}(e(s))}{\mathrm{odds}\{P(A = 1 \mid Z = z)\}},$$

and hence

$$\frac{\max_s(\alpha_s/\beta_s)}{\min_s(\alpha_s/\beta_s)} = \Gamma_e(z).$$

Define $X := \sum_s \alpha_s x_s$, $X_\beta := \sum_s \beta_s x_s$, and $Y := \sum_s \beta_s y_s$. Lemma 4, applied with the common weights β_s, gives $\mathrm{odds}(X_\beta)/\mathrm{odds}(Y) \le \Gamma_{\mathcal{S}}$. For $\lambda_s := \alpha_s/\beta_s$, note that $\sum_s \beta_s \lambda_s = 1$. Therefore

$$\frac{\mathrm{odds}(X)}{\mathrm{odds}(X_\beta)} = \frac{\sum_s \beta_s \lambda_s x_s}{\sum_s \beta_s x_s} \cdot \frac{\sum_s \beta_s (1 - x_s)}{\sum_s \beta_s \lambda_s (1 - x_s)} \le \frac{\max_s \lambda_s}{\min_s \lambda_s} = \Gamma_e(z).$$

Combining the last two displays yields $\mathrm{odds}(X)/\mathrm{odds}(Y) \le \Gamma_{\mathcal{S}}\Gamma_e(z)$; exchanging the two actions gives the reciprocal bound. Finally, the same binary cross-ratio identity at level $Z = z$ identifies these ratios with the odds ratios of $\widetilde{\pi}_b(1 \mid z, U)$ across the two values of U. Thus $\Gamma_{\mathcal{Z}}(z) \le \Gamma_{\mathcal{S}}\Gamma_e(z)$. □

Why binary U matters. For arbitrary finite U, exact $e(s) = e(z)$ makes the state-mixture weights common across actions but does not reduce the hidden distribution to one scalar. Let

$$q_s^a(u) := P(U = u \mid S = s, A = a), \qquad r_s(u) := \frac{q_s^1(u)}{q_s^0(u)}.$$

Bayes' rule gives

$$r_s(u) = \frac{1 - e(z)}{e(z)} \mathrm{odds}\{\pi_b(1 \mid s, u)\}.$$

The pairwise range of each r_s is at most $\Gamma_{\mathcal{S}}$ and $E_{q_s^0}[r_s] = 1$, so $r_s(u) \in [1/\Gamma_{\mathcal{S}}, \Gamma_{\mathcal{S}}]$. The latent density ratio is a convex combination of these values for each u; taking its ratio across two values of u yields

$$\Gamma_{\mathcal{Z}}(z) \le \Gamma_{\mathcal{S}}^2.$$

The square cannot generally be reduced to $\Gamma_{\mathcal{S}}$. With two equally likely raw states, three-valued U, and $e(s_1) = e(s_2) = 1/2$, take

$$q^0 = \begin{pmatrix} 0.001 & 0.010 & 0.989 \\ 0.010 & 0.001 & 0.989 \end{pmatrix}, \qquad \begin{aligned} r_1 &= (100/109, 1000/109, 100/109), \\ r_2 &= (100/991, 1000/991, 1000/991), \end{aligned}$$

and set $q_s^1(u) = q_s^0(u) r_s(u)$. The implied behavior policies give $\Gamma_{\mathcal{S}} = 10$ but $\Gamma_{\mathcal{Z}} = 100190/2081 \approx 48.145$. A less extreme construction with all $P(U = u \mid S = s) \ge 0.10$ and behavior probabilities in $[0.38, 0.62]$ gives $\Gamma_{\mathcal{S}} \approx 1.967$ and $\Gamma_{\mathcal{Z}} \approx 2.262$.

Trained multivalued-U stress test. We augment the moderate construction with an independent binary task component T and a two-code bottleneck. Rewards follow $R = 2T - 1 + 0.25A + \varepsilon$, so the action is outcome-relevant and the reward-optimal encoder retains T. Because $e(s) = 1/2$ for every raw state, the population $\mathcal{L}_{\mathrm{prop}}$ is identically zero for every encoder. Across $K = 30$ seeds, task-only and task+$\mathcal{L}_{\mathrm{prop}}$ with oracle propensities at $\lambda_{\mathrm{prop}} \in \{1, 10\}$ selected the task-aligned amplifying merge in $30/30$ runs; under matched initialization and minibatch order, these branches were bitwise identical. An independently fitted-propensity branch at $\lambda_{\mathrm{prop}} = 1$ also selected the merge in $30/30$ runs. Every selected merge has $\Gamma_{\mathcal{Z}}/\Gamma_{\mathcal{S}} \approx 1.150$, while a common frozen-code reward probe has mean MSE 1.6×10^{-6} against the noiseless structural reward mean. This controlled two-code hard-bottleneck experiment isolates an objective-level limitation of the marginal-propensity proxy; it does not estimate typical failure frequency, establish value noncoverage, or test the continuous Setting B implementation.

C Proof of Theorem 3

Regime of validity. The bound Equation 6 is meaningful only when Δ is well below the threshold $\eta^3/(1-\eta)$. For $\eta = 0.1$, the threshold is $\Delta < 0.0011$; for $\eta = 0.2$, the threshold relaxes to $\Delta < 0.01$. The bound is informative only in a restrictive regime; outside it, the qualitative content is what connects to the empirical findings of Section 6: representations closer to the action-sensitivity condition incur smaller amplification, with $\Delta = 0$ recovering Proposition 2 exactly.

On the positivity assumption. Theorem 3 requires two-sided positivity $\pi_b(a \mid s, u) \in [\eta, 1-\eta]$. For binary actions this is equivalent to requiring both action probabilities to be at least η on supported (s, u), a uniform strengthening of strict positivity. The extension in Appendix A.1 obeys the same bounds. The proof uses this to bound $\mathrm{odds}'(x) = (1-x)^{-2}$ uniformly on $[\eta, 1-\eta]$; without the upper-tail bound, the Lipschitz constant of $\mathrm{odds}(\cdot)$ blows up as $x \to 1$, and the resulting amplification bound becomes vacuous in regimes where the behavioural policy is near-deterministic on some arm. The assumption is standard in confounding-robust off-policy evaluation, where it appears as bounded log-propensity [Kallus and Zhou, 2020, Namkoong et al., 2020].

Proof. Fix supported $(z, u), (z, u')$. Decompose the induced latent policy into a safe part and a perturbation, using π_b° from Appendix A.1. Define:

$$p_{\mathrm{safe}} := \sum_{s:\phi(s)=z} P(s \mid z)\ \pi_b^\circ(1 \mid s, u),$$
$$q_{\mathrm{safe}} := \sum_{s:\phi(s)=z} P(s \mid z)\ \pi_b^\circ(1 \mid s, u'),$$

and perturbations $\varepsilon_u := \tilde{\pi}_b(1 \mid z, u) - p_{\mathrm{safe}}$ and $\varepsilon_{u'} := \tilde{\pi}_b(1 \mid z, u') - q_{\mathrm{safe}}$.

Bounding the perturbations. Writing $\xi_s := P(s \mid z, u) - P(s \mid z)$, we have $\sum_s |\xi_s| \le \Delta$ and:

$$|\varepsilon_u| \ = \ \Big|\sum_s \xi_s\ \pi_b^\circ(1 \mid s, u)\Big| \ \le \ \sum_s |\xi_s| \cdot 1 \ \le \ \Delta.$$

Similarly $|\varepsilon_{u'}| \le \Delta$.

Safe part. Since p_{safe} and q_{safe} are convex combinations with u-independent weights $P(s \mid z)$, Proposition 2 (via Lemma 4) gives $\mathrm{odds}(p_{\mathrm{safe}})/\mathrm{odds}(q_{\mathrm{safe}}) \le \Gamma_{\mathcal{S}}$.

Lipschitz bound on odds. For $x \in [\eta, 1-\eta]$, the derivative $\mathrm{odds}'(x) = 1/(1-x)^2 \le 1/\eta^2$. Under positivity, $\tilde{\pi}_b(1 \mid z, u) \in [\eta, 1-\eta]$ and $p_{\mathrm{safe}} \in [\eta, 1-\eta]$, so:

$$|\mathrm{odds}(\tilde{\pi}_b(1 \mid z, u)) - \mathrm{odds}(p_{\mathrm{safe}})| \ \le \ \Delta/\eta^2.$$

Bounding the perturbed ratio. Write $A := \mathrm{odds}(p_{\mathrm{safe}})$, $B := \mathrm{odds}(q_{\mathrm{safe}})$, $\varepsilon := \Delta/\eta^2$. Then:

$$\frac{\mathrm{odds}(\tilde{\pi}_b(1 \mid z, u))}{\mathrm{odds}(\tilde{\pi}_b(1 \mid z, u'))} \ \le \ \frac{A+\varepsilon}{B-\varepsilon}.$$

Since $B \ge \mathrm{odds}(\eta) = \eta/(1-\eta)$ and $\varepsilon = \Delta/\eta^2$, the denominator is positive when $\Delta < \eta^3/(1-\eta)$. Using $A \le \Gamma_{\mathcal{S}} B$:

$$\begin{aligned}
\frac{A+\varepsilon}{B-\varepsilon} &= \frac{A}{B-\varepsilon} + \frac{\varepsilon}{B-\varepsilon} \\
&\le \frac{\Gamma_{\mathcal{S}} B}{B-\varepsilon} + \frac{\varepsilon}{B-\varepsilon} \\
&= \Gamma_{\mathcal{S}} + \frac{(1+\Gamma_{\mathcal{S}})\varepsilon}{B-\varepsilon} \\
&\le \Gamma_{\mathcal{S}} + \frac{(1+\Gamma_{\mathcal{S}})\Delta/\eta^2}{\eta/(1-\eta) - \Delta/\eta^2} \\
&= \Gamma_{\mathcal{S}} + \frac{(1+\Gamma_{\mathcal{S}})(1-\eta)\,\Delta}{\eta^3 - (1-\eta)\Delta}.
\end{aligned}$$

Since z, u, u' were arbitrary, the bound holds uniformly, giving Equation 6. □

D Discrete Counterexample and LP Verification

We instantiate the coverage failure of Theorem 1 with a concrete robust importance-weighted estimator applied to a parametric counterexample. Because the compression also discards outcome-relevant state identity, this example exhibits combined coarsening harm; Appendix D.5 gives the clean outcome-sufficient construction.

D.1 Data-Generating Process

Consider a one-step confounded decision problem with $\mathcal{S} = \{s_1, s_2, s_3\}$, $\mathcal{A} = \{0, 1\}$, $\mathcal{U} = \{0, 1\}$, and compression $\phi(s_1) = \phi(s_2) = z_1$, $\phi(s_3) = z_3$.

The joint distribution $P(S, U)$ and behavioural policy $\pi_b(1 \mid s, u)$ are specified in Table 4. The parameter $\alpha = 0.95$ controls the strength of the U–S correlation within the merged code z_1.

Table 4: DGP specification for the non-coverage demonstration. Left: joint distribution. Right: behavioural policy. Entries 0.947 and 0.053 round $18/19$ and $1/19$; all calculations use the exact values.

	u=0	u=1
s_1	0.380	0.020
s_2	0.020	0.380
s_3	0.100	0.100

$\pi_b(1 \mid s, u)$	u=0	u=1
s_1	0.947	0.900
s_2	0.100	0.053
s_3	0.500	0.500

By construction, the per-state pairwise sensitivity parameter is $\Gamma_{\mathcal{S}} = 2.0$. Thus $\Lambda_{\mathcal{S}} \leq 2$ is a conservative population marginal-model input. After compression, the conditional distributions within z_1 are:

$$P(s_1 \mid z_1,\, u{=}0) = 0.95, \quad P(s_1 \mid z_1,\, u{=}1) = 0.05,$$

so the mixing weights nearly reverse with the confounder value. This yields an induced latent policy $\tilde{\pi}_b(1 \mid z_1, u{=}0) = 0.905$ and $\tilde{\pi}_b(1 \mid z_1, u{=}1) = 0.095$, giving $\Gamma_{\mathcal{Z}} = 90.75$ — an amplification factor of approximately $45\times$. The approximate conditional independence violation is $\Delta = 0.90$.

D.2 Statewise HT-IPW Calculation

For a conditioning variable X, let $e_X(x) = P(A = 1 \mid X = x)$ be the true observed propensity and $p_x(u) = P(A = 1 \mid X = x, U = u)$ the confounded propensity. The population marginal class $\mathcal{M}(\Lambda, X)$ implies

$$p_x(u) \;\in\; I_\Lambda(x) := \left[\frac{e_X(x)}{\Lambda + (1-\Lambda)\, e_X(x)},\;\; \frac{\Lambda\, e_X(x)}{1 + (\Lambda - 1)\, e_X(x)}\right]. \tag{10}$$

The reported statewise scalar Horvitz–Thompson (HT-IPW) calculation varies a candidate $e^*(x) \in I_\Lambda(x)$ independently at each state:

$$\begin{aligned} V_{\max}(\pi_e; \Lambda, \mathcal{X}) = \sum_{x \in \mathcal{X}} P(x) \max_{e^*(x) \in I_\Lambda(x)} \Big[& e_X(x) \frac{\pi_e(1 \mid x) r(x, 1)}{e^*(x)} \\ & + (1 - e_X(x)) \frac{\pi_e(0 \mid x) r(x, 0)}{1 - e^*(x)} \Big]. \end{aligned} \tag{11}$$

with $V_{\min}$ obtained by replacing $\max$ with $\min$. In the binary-action discrete case, each state's subproblem is a one-dimensional optimisation over the scalar $e^*(x)$ within the interval Equation 10, which we solve by bounded scalar minimisation.

Remark 2. *Equation 11 is a statewise scalar sensitivity calculation, not a sharp identified interval for the full marginal class. It differs from the stationary occupancy-ratio LP of Kallus and Zhou [2020]. Sharp normalized one-step MSM bounds use observed-data compatibility and outcome-dependent weighting [Dorn et al., 2022]. The sharp recomputation reported below independently confirms exclusion at the same numerical latent input.*

D.3 Reward and Target Policy

We define rewards $r(s, a)$ that create opposing incentives across the merged states:

$$r(s_1, 1) = 1, \quad r(s_1, 0) = 0, \quad r(s_2, 1) = 0, \quad r(s_2, 0) = 1, \quad r(s_3, 0) = r(s_3, 1) = 0.5.$$

The target policy is $\pi_e(1 \mid s) = 1$ for all s (always take action 1). Under the true DGP, the expected policy value is:

$$V(\pi_e) \;=\; \sum_s P(s)\, r(s, 1) \;=\; 0.4 \cdot 1 + 0.4 \cdot 0 + 0.2 \cdot 0.5 \;=\; 0.50.$$

Because the compression maps s_1 (high reward under a=1) and s_2 (zero reward under a=1) to the same latent code, the observed action-conditioned means are $\mathbb{E}[R \mid Z = z_1, A = 0] = \mathbb{E}[R \mid Z = z_1, A = 1] = 0.945$ — the reward structure that distinguishes s_1 from s_2 is obscured.

D.4 Results

Table 5 reports the scalar intervals computed by solving Equation 11 and its minimisation counterpart.

Table 5: Statewise HT-IPW intervals for the target policy $\pi_e(1 \mid \cdot) = 1$ with true value $V(\pi_e) = 0.50$. The latent-space scalar interval at marginal input $\Lambda = \Gamma_{\mathcal{S}}$ fails to cover the true value.

State space	Λ used	$V_{\min}$	$V_{\max}$	Width	Covers $V(\pi_e)$?
Raw $\mathcal{S}$	$\Lambda = \Gamma_{\mathcal{S}} = 2.0$	0.464	0.572	0.108	**Yes**
Latent $\mathcal{Z}$	$\Lambda = \Gamma_{\mathcal{S}} = 2.0$	0.642	1.284	0.642	**No**
Latent $\mathcal{Z}$	$\Lambda = 90.75$	0.433	39.25	38.82	Yes

Sharp normalized one-step MSM bounds. For the always-action-1 target, let F_x be the observed conditional law of R given $(X = x, A = 1)$. The sharp endpoints minimize or maximize

$$\sum_x P(x) \int w_x(r)\, r\, dF_x(r), \qquad \int w_x(r)\, dF_x(r) = 1,$$

subject to $e_X(x) + (1 - e_X(x))/\Lambda \le w_x(r) \le e_X(x) + \Lambda(1 - e_X(x))$ almost surely [Dorn et al., 2022]. Unlike Equation 11, weights depend on the outcome and satisfy normalization within each observed cell. Because the reward kernel is $P(R \mid S, A)$, the raw-state value is identified under overlap. Under the sharp normalized one-step MSM, the raw interval collapses to $[0.500, 0.500]$. In contrast, the sharp latent interval at the same numerical marginal input $\Lambda = 2$ is $[0.834, 0.867]$, which excludes the true value 0.500. The displayed averaged-U pairwise level is 90.751. For the outcome-complete $U^\dagger = (S, U)$, the minimal marginal level containing the true assignment mechanism is 18, while its pairwise range is 324. Thus using $\Lambda = 90.751$ as the MSM input is conservative here and gives $[0.482, 0.878]$, restoring the coverage guarantee, although this relationship does not follow generally from the averaged-U definition.

Non-coverage. In the raw state space, the scalar interval $[0.464,\ 0.572]$ contains $V(\pi_e) = 0.50$. In the latent space at the same sensitivity input $\Lambda = \Gamma_{\mathcal{S}} = 2.0$, the lower bound rises to $0.642 > 0.50$: the true value falls *below* the interval. The sharp MSM interval above also excludes the value at this input. The scalar bounds are anti-conservative: they exclude the truth from below, falsely suggesting that the policy is better than it actually is.

Target-policy measurability check. Table 6 evaluates four ϕ-measurable target action mixtures, including a nonconstant cell-projected greedy policy. By symmetry, every ϕ-measurable policy in this construction has value 0.500; thus the rows compare intervals under different action mixtures, not distinct estimands. All four intervals from Equation 11 exclude this value in the latent space at marginal input $\Lambda = \Gamma_{\mathcal{S}}$.

Let the raw greedy policy break the reward tie at s_3 toward action 0, and define its observed-cell projection by $\bar{\pi}_e(1 \mid z) = \mathbb{E}[\pi_{\text{greedy}}(1 \mid S) \mid Z = z]$. This gives $\bar{\pi}_e(1 \mid z) = (0.5, 0)$ and the lifted policy $\pi_e(1 \mid s) = (0.5, 0.5, 0)$. Hence the raw and latent rows evaluate the same nonconstant policy, rather than changing the target after compression.

Table 6: HT-IPW calculations for four ϕ-measurable target action mixtures. All have true value 0.500 in this symmetric construction; raw coverage is maintained, whereas the latent intervals at $\Lambda = \Gamma_{\mathcal{S}}$ exclude the value.

Target policy π_e	$V(\pi_e)$	Raw ($\Lambda = \Gamma_{\mathcal{S}}$)	$\text{Cvg}_{\mathcal{S}}$	Latent ($\Lambda = \Gamma_{\mathcal{S}}$)	$\text{Cvg}_{\mathcal{Z}}$
Always a=1	0.500	$[0.464,\ 0.572]$	Yes	$[0.642,\ 1.284]$	**No**
Always a=0	0.500	$[0.464,\ 0.572]$	Yes	$[0.642,\ 1.284]$	**No**
Random (0.5)	0.500	$[0.489,\ 0.535]$	Yes	$[0.856,\ 0.963]$	**No**
Projected greedy	0.500	$[0.464,\ 0.572]$	Yes	$[0.831,\ 1.000]$	**No**

D.5 Outcome-Confounded Sequential Construction

To isolate sensitivity transport under hidden outcome confounding, we use an exactly enumerable finite $H = 2$ construction. We reuse the joint $P(S_0, U)$, behavioural policy, and compression above, draw U once per trajectory, and set $S_1 = (S_0, A_0)$. At both steps, the behavioural policy applies the one-step policy table to the S_0 component, so the per-step raw level remains $\Gamma_{\mathcal{S}} = 2$. The transition is deterministic and has no direct U-dependence given (S_0, A_0). The latent state retains the action history: $Z_0 = \phi(S_0)$ and $Z_1 = (\phi(S_0), A_0)$.

The target chooses action 1 at both steps, and the terminal potential reward is

$$Y(a_0, a_1) = \mathbf{1}\{a_0 = a_1 = 1\}(1 - U).$$

Thus the same hidden U affects the logging policy and the outcome, and the target value is $V = 0.500$. Structurally, $Y(a_0, a_1) \perp\!\!\!\perp S_0 \mid (Z_0, U)$: discarded raw-state identity has no direct outcome effect once (Z_0, U) is fixed. Thus U is outcome-complete after compression, cleanly isolating action-sensitivity transport from loss of state-specific structural reward information.

For an observed history h_t, let $e_t = P(A_t = 1 \mid h_t)$ be the true observed propensity. Under a history-conditional marginal constraint at level Λ, the likelihood ratio on a target-matching action satisfies

$$e_t + \frac{1 - e_t}{\Lambda} \ \le\ \frac{e_t}{P(A_t = 1 \mid h_t, U)} \ \le\ e_t + \Lambda(1 - e_t).$$

Multiplying these factors along the target path and aggregating the nonnegative terminal reward gives a conservative sequential HT outer envelope. Exact population enumeration gives:

Table 7: Exact $H = 2$ outcome-confounded extension. The sequential envelopes are conservative outer bounds, not sharp intervals.

State used	Sensitivity level	Interval	Covers $V = 0.500$?
Raw S	$\Lambda = 2$ (conservative from $\Gamma_{\mathcal{S}}$)	$[0.435,\ 0.891]$	Yes
Compressed Z	$\Lambda = 2$	$[0.598,\ 1.000]$	No
Compressed Z	Required $\Lambda = 10.390$	$[0.408,\ 1.000]$	Yes

Because the compressed $\Lambda = 2$ row is a valid outer interval for its history-conditional marginal sensitivity class and already excludes V, the corresponding sharp identified interval contained within it also excludes V. The history-conditional marginal level actually required in Z-space, 10.390, is the maximum confounded-to-observed propensity-odds deviation used by this envelope. Because the construction reuses the step-0 propensities, its distinct pairwise latent-U sensitivity level is again $\Gamma_{\mathcal{Z}} = 90.751$. The two numbers use different sensitivity conventions and should not be conflated. This construction preserves transition-confounding separation and establishes outcome-relevant hidden confounding without introducing transition confounding.

D.6 Transition-Confounded Sequential Construction

We next isolate transport of a transition-kernel sensitivity level in an exactly enumerable $H = 2$ problem. Let

$$P(S_0, U) = \begin{array}{c|cc} & 0 & 1 \\ \hline s_1 & 2/5 & 3/10 \\ s_2 & 1/10 & 1/5 \end{array}, \qquad \pi_b(1 \mid s, u) = \begin{array}{c|cc} & 0 & 1 \\ \hline s_1 & 1/5 & 1/5 \\ s_2 & 3/5 & 4/5 \end{array}.$$

For $A_0 = 1$, the binary transition outcome Y satisfies

$$P(Y = 1 \mid s, A_0 = 1, u) = \begin{array}{c|cc} & 0 & 1 \\ \hline s_1 & 1/10 & 1/10 \\ s_2 & 1/10 & 9/10 \end{array};$$

for $A_0 = 0$ all four entries are $1/2$. Set $S_1 = (S_0, Y)$, draw $A_1 \sim \text{Bernoulli}(1/2)$, and let $R = A_1 Y$. The target takes $A_0 = A_1 = 1$, giving $V = 13/50 = 0.260$. Compression merges s_1, s_2 at time zero and retains Y at time one, so discarded state identity has no direct reward effect.

Let $T^{\text{do}}(\cdot \mid x, a)$ denote the interventional transition row and $T^{\text{obs}}(\cdot \mid x, a)$ its observed, behavior-conditioned counterpart. Define the direct rectangular level

$$\Lambda_T(x, a) = \max_y \max \left\{ \frac{T^{\text{do}}(y \mid x, a)}{T^{\text{obs}}(y \mid x, a)}, \frac{T^{\text{obs}}(y \mid x, a)}{T^{\text{do}}(y \mid x, a)} \right\}.$$

For binary observed row $q = T^{\text{obs}}(Y = 1 \mid x, a)$, the sharp row-rectangular bounds are

$$\max\left\{ \frac{q}{\Lambda}, 1 - \Lambda(1 - q) \right\} \le T^{\text{do}}(Y = 1 \mid x, a) \le \min\left\{ \Lambda q, 1 - \frac{1 - q}{\Lambda} \right\}.$$

Exact enumeration gives:

Table 8: Exact $H = 2$ transition-confounded construction under the direct rectangular transition-row model.

State used	Transition level	Robust interval	Covers $V = 0.260$?
Raw S	$\Lambda_{T,S} = 1.152$	$[0.251, 0.298]$	Yes
Compressed Z	Same 1.152	$[0.395, 0.525]$	No
Compressed Z	$\Lambda_{T,Z} = 1.752$	$[0.260, 0.689]$	Yes

The paper's pairwise action-odds level moves oppositely, $\Gamma_{\mathcal{S}} = 8/3$ to $\Gamma_{\mathcal{Z}} = 99/49$, so an increase in action Γ does not explain the exclusion. Action selection is already absorbed by T^{obs} versus T^{do}; multiplying Γ and Λ_T would require a separate joint sensitivity model. For $Y = 1$, the exact transition decomposition is

$$\frac{T^{\text{obs}}(1 \mid z, A_0 = 1)}{T^{\text{do}}(1 \mid z, do(A_0 = 1))} = \frac{123}{115} \cdot \frac{575}{351} = \frac{205}{117}.$$

This finite result establishes persistence under genuine transition confounding; it is not a full extension of Corollary 8 or a general hidden-U trajectory model.

E Amplification Bound

Theorem 1 shows that compression can amplify the averaged-U action-sensitivity level. This section quantifies the amplification through a multiplicative decomposition into the two action-sensitivity sources identified in Section 3: per-state action confounding, bounded by $\Gamma_{\mathcal{S}}$, and state-selection confounding, a property of ϕ and $P(S \mid Z, U)$.

State-selection confounding factor. For $u^* \in \mathcal{U}$, define the *virtual latent policy*

$$\nu_{u^*}(z, u) \;:=\; \sum_{s:\phi(s)=z} P(s \mid z, u)\, \pi_b^\circ(1 \mid s, u^*),$$

which mixes the within-cell weights $P(s \mid z, u)$ at confounder value u against the per-state policy held fixed at u^*. The *state-selection confounding factor* is

$$\Gamma_{\text{sel}} \;:=\; \max_{z \in \mathcal{Z}} \max_{u, u', u^* \in \mathcal{U}} \frac{\text{odds}\big(\nu_{u^*}(z, u)\big)}{\text{odds}\big(\nu_{u^*}(z, u')\big)}. \tag{12}$$

Γ_{sel} measures how much U can shift the within-cell state composition $P(s \mid z, u)$, evaluated through any per-state action-probability function from the underlying policy family. Here π_b° is the fixed extension in Appendix A.1, and the maximum in Equation 12 uses only supported $(z, u), (z, u')$; u^* may be any confounder value.

Proposition 7 (Multiplicative amplification bound)**.** *Assume positivity ($\pi_b(a \mid s,u) \in (0,1)$ for supported (s,u) and each a). Then*

$$\Gamma_{\mathcal{Z}} \;\le\; \Gamma_{\mathcal{S}} \cdot \Gamma_{\mathrm{sel}}. \tag{13}$$

Moreover, $\Gamma_{\mathrm{sel}} = 1$ whenever $U \perp\!\!\!\perp S \mid Z$, recovering Proposition 2 as a special case.

Proof. Fix supported $(z,u),(z,u')$. Write $w_s^u := P(s \mid z,u)$ and $p_s^u := \pi_b^\circ(1 \mid s,u)$, so that the induced latent policy Equation 1 is $\bar{p}_u := \tilde{\pi}_b(1 \mid z,u) = \sum_s w_s^u p_s^u$.

Introduce the intermediate quantity

$$\bar{q} \;:=\; \sum_{s:\phi(s)=z} w_s^u \, p_s^{u'},$$

which uses the within-cell weights at u together with the per-state policy at u'. We decompose the latent odds ratio as

$$\frac{\mathrm{odds}(\bar{p}_u)}{\mathrm{odds}(\bar{p}_{u'})} \;=\; \underbrace{\frac{\mathrm{odds}(\bar{p}_u)}{\mathrm{odds}(\bar{q})}}_{\text{Factor 1}} \cdot \underbrace{\frac{\mathrm{odds}(\bar{q})}{\mathrm{odds}(\bar{p}_{u'})}}_{\text{Factor 2}} \tag{14}$$

and bound each factor separately.

Factor 1 (per-state action confounding). Both $\bar{p}_u = \sum_s w_s^u p_s^u$ and $\bar{q} = \sum_s w_s^u p_s^{u'}$ are convex combinations with *identical* weights w_s^u, differing only in the per-state policy. By the supported pairwise constraint Equation 4 and the range-preserving extension, $\mathrm{odds}(p_s^u)/\mathrm{odds}(p_s^{u'}) \le \Gamma_{\mathcal{S}}$ for every s. By positivity, $p_s^u, p_s^{u'} \in (0,1)$, so Lemma 4 applies, yielding

$$\frac{\mathrm{odds}(\bar{p}_u)}{\mathrm{odds}(\bar{q})} \;\le\; \Gamma_{\mathcal{S}}.$$

Factor 2 (state-selection confounding). Both $\bar{q} = \sum_s w_s^u p_s^{u'}$ and $\bar{p}_{u'} = \sum_s w_s^{u'} p_s^{u'}$ use the *same* per-state policy $p_s^{u'}$, differing only in the within-cell weights. In the notation of Equation 12, $\bar{q} = \nu_{u'}(z,u)$ and $\bar{p}_{u'} = \nu_{u'}(z,u')$. Therefore

$$\frac{\mathrm{odds}(\bar{q})}{\mathrm{odds}(\bar{p}_{u'})} = \frac{\mathrm{odds}(\nu_{u'}(z,u))}{\mathrm{odds}(\nu_{u'}(z,u'))} \;\le\; \Gamma_{\mathrm{sel}},$$

where the inequality follows because the maximum in Equation 12 is taken over $u^* \in \mathcal{U}$, in particular including $u^* = u'$.

Combining. Substituting both bounds into Equation 14,

$$\frac{\mathrm{odds}(\bar{p}_u)}{\mathrm{odds}(\bar{p}_{u'})} \;\le\; \Gamma_{\mathcal{S}} \cdot \Gamma_{\mathrm{sel}}.$$

Taking the maximum over $z \in \mathcal{Z}$ and $u,u' \in \mathcal{U}$ gives $\Gamma_{\mathcal{Z}} \le \Gamma_{\mathcal{S}} \cdot \Gamma_{\mathrm{sel}}$.

Special case. Under $U \perp\!\!\!\perp S \mid Z$, $P(s \mid z,u) = P(s \mid z)$ for all s,u, so $\nu_{u^*}(z,u) = \nu_{u^*}(z,u')$ for every z,u,u',u^*. Hence $\Gamma_{\mathrm{sel}} = 1$, and Equation 13 reduces to $\Gamma_{\mathcal{Z}} \le \Gamma_{\mathcal{S}}$, recovering Proposition 2. □

Numerical verification. For the counterexample of Theorem 1, the per-state factor is $\Gamma_{\mathcal{S}} = 12/7 \approx 1.71$. Computing Γ_{sel} from Equation 12 requires only the within-cell weights and the per-state policy: at z_1, the weights flip near-completely under U ($P(s_1 \mid z_1,0) = 8/9$ versus $P(s_1 \mid z_1,1) = 1/9$), giving $\Gamma_{\mathrm{sel}} \approx 5.26$ (attained at $u^* \in \{0,1\}$ symmetrically). The bound yields

$$\Gamma_{\mathcal{S}} \cdot \Gamma_{\mathrm{sel}} \;\approx\; 1.71 \times 5.26 \;\approx\; 9.0,$$

compared to the true $\Gamma_{\mathcal{Z}} \approx 8.49$. The bound is loose by roughly 6% on this construction.

History-Conditional State-Selection Factor

The state-selection factor Γ_{sel} defined in Equation 12 uses marginal within-cell weights $P(s \mid z, u)$. In the multi-step setting, the relevant quantity at step t involves *history-conditional* within-cell weights $P(s_t \mid z_{0:t}, a_{0:t-1}, u)$, which condition on the latent trajectory and action sequence through step $t-1$. We define the history-conditional analogue of Γ_{sel} accordingly.

For each $u^* \in \mathcal{U}$, define the history-conditional virtual latent policy using the time-t extension $\pi^{\circ}_{b,t}$: it equals π_b on the support of (S_t, U) and uses $e_t(s) = P(A_t = 1 \mid S_t = s)$ elsewhere

$$\nu^{\mathrm{hist}}_{u^*}(z_{0:t}, a_{0:t-1}, u) := \sum_{s_t : \phi(s_t) = z_t} P(s_t \mid z_{0:t}, a_{0:t-1}, u)\, \pi^{\circ}_{b,t}(1 \mid s_t, u^*).$$

The history-conditional state-selection factor at step t is

$$\widetilde{\Gamma}^{(t)}_{\mathrm{sel}} := \max_{z_{0:t},\, a_{0:t-1}} \max_{u, u', u^* \in \mathcal{U}} \frac{\mathrm{odds}\big(\nu^{\mathrm{hist}}_{u^*}(z_{0:t}, a_{0:t-1}, u)\big)}{\mathrm{odds}\big(\nu^{\mathrm{hist}}_{u^*}(z_{0:t}, a_{0:t-1}, u')\big)}. \tag{15}$$

The maximum ranges over histories supported under both u and u', with arbitrary u^*. $\widetilde{\Gamma}^{(t)}_{\mathrm{sel}}$ generalizes Γ_{sel} of Equation 12 to depend on the latent history rather than only the current latent code. When $P(s_t \mid z_{0:t}, a_{0:t-1}, u) = P(s_t \mid z_t, u)$ (the within-cell state law depends only on z_t and u), the marginal factor bounds every comparison on a commonly supported history, giving $\widetilde{\Gamma}^{(t)}_{\mathrm{sel}} \le \Gamma^{(t)}_{\mathrm{sel}}$. Equality holds when the maximizing marginal comparison admits such a history. In general the two need not coincide or be ordered, because marginalizing over histories changes the U-specific within-cell mixtures.

F Multi-Step Compounding

The one-step amplification bound (Proposition 7) extends to the multi-step setting under dynamics-confounding separation. We bound the likelihood ratio of action sequences between confounder values, the pairwise action-sensitivity component relevant to multi-step off-policy evaluation [Kallus and Zhou, 2020, Namkoong et al., 2020].

Define the sequential action likelihood

$$\widetilde{\pi}_b(a_{0:H-1} \| z_{0:H-1}, u) := \prod_{t=0}^{H-1} \widetilde{\pi}_b(a_t \mid z_{0:t}, a_{0:t-1}, u).$$

The symbol $\|$ distinguishes this policy likelihood from the ordinary conditional probability given the complete future latent trajectory.

Corollary 8 (Multi-step action-likelihood compounding)**.** *Under dynamics-confounding separation and positivity, for any latent trajectory $z_{0:H-1}$, action sequence $a_{0:H-1}$, and confounder values $u, u' \in \mathcal{U}$ for which every conditioning history $(z_{0:t}, a_{0:t-1})$ is supported under both,*

$$\frac{\widetilde{\pi}_b(a_{0:H-1} \| z_{0:H-1}, u)}{\widetilde{\pi}_b(a_{0:H-1} \| z_{0:H-1}, u')} \;\le\; \prod_{t=0}^{H-1} \Gamma^{(t)}_{\mathcal{S}} \cdot \widetilde{\Gamma}^{(t)}_{\mathrm{sel}},$$

where $\Gamma^{(t)}_{\mathcal{S}}$ is the per-state action sensitivity parameter at step t and $\widetilde{\Gamma}^{(t)}_{\mathrm{sel}}$ is the history-conditional state-selection factor defined in Equation 15. If $\Gamma^{(t)}_{\mathcal{S}} \le \overline{\Gamma}_{\mathcal{S}}$ uniformly across t, the action-likelihood bound becomes $\overline{\Gamma}^{H}_{\mathcal{S}} \cdot \prod_{t=0}^{H-1} \widetilde{\Gamma}^{(t)}_{\mathrm{sel}}$: the factor $\overline{\Gamma}^{H}_{\mathcal{S}}$ is the per-step pairwise action-likelihood compounding, and $\prod_t \widetilde{\Gamma}^{(t)}_{\mathrm{sel}}$ is the additional representation-induced amplification, which equals 1 at every step when $U \perp\!\!\!\perp S_t \mid Z_t$ and $P(S_t \mid Z_{0:t}, A_{0:t-1}, U) = P(S_t \mid Z_t, U)$.

This is an action-sequence likelihood-ratio result. A policy-value coverage statement additionally requires consistency, positivity, a history-measurable target, and conditional-mean sufficiency of the trajectory return given the compressed history and U; otherwise the discarded state history must enter the latent variable. It also requires transition separation at the relevant compressed/history level or a separate bound on transition-likelihood factors.

Remark 3 (Transition-confounded dynamics)**.** *The per-step mixture identity remains valid when $P(S_{t+1} \mid S_t, A_t, U)$ depends on U: earlier transitions change the history-conditional weights $P(S_t \mid Z_{0:t}, A_{0:t-1}, U)$ but not the action-policy decomposition. A joint trajectory-sensitivity model must additionally bound the resulting transition likelihood ratios. Multiplicative transition perturbations and their sharp robust value bounds are studied by Bennett et al. [2024]; combining such transition factors with representation-induced action amplification gives a broader robust-OPE model. Appendix D.6 gives an exact finite transition-confounded coverage construction under a direct rectangular transition-row model.*

Proof. By the definition of the sequential action likelihood,

$$\frac{\widetilde{\pi}_b(a_{0:H-1} \| z_{0:H-1}, u)}{\widetilde{\pi}_b(a_{0:H-1} \| z_{0:H-1}, u')} = \prod_{t=0}^{H-1} \frac{\widetilde{\pi}_b(a_t \mid z_{0:t}, a_{0:t-1}, u)}{\widetilde{\pi}_b(a_t \mid z_{0:t}, a_{0:t-1}, u')}.$$

Each per-step ratio is a history-conditional latent action ratio. By total probability, it admits the decomposition

$$\widetilde{\pi}_b(a_t \mid z_{0:t}, a_{0:t-1}, u) = \sum_{s_t : \phi(s_t) = z_t} P(s_t \mid z_{0:t}, a_{0:t-1}, u)\, \pi^{\circ}_{b,t}(a_t \mid s_t, u),$$

where the extension agrees with π_b wherever the weight is positive. This is the same form as the one-step latent policy in Equation 12 with marginal weights replaced by history-conditional weights. Transition dynamics affect these weights but do not enter the policy-likelihood product itself; a joint state-action trajectory likelihood would require separate transition factors. Applying the two-factor argument of Proposition 7 pointwise to each step yields

$$\frac{\widetilde{\pi}_b(a_t \mid z_{0:t}, a_{0:t-1}, u)}{\widetilde{\pi}_b(a_t \mid z_{0:t}, a_{0:t-1}, u')} \leq \Gamma_{\mathcal{S}}^{(t)} \cdot \widetilde{\Gamma}_{\mathrm{sel}}^{(t)},$$

where the per-state action factor is bounded by $\Gamma_{\mathcal{S}}^{(t)}$ exactly as in Proposition 7 (action-confounding bounds are pointwise in s_t and unaffected by history conditioning), and the history-conditional state-selection factor is bounded by $\widetilde{\Gamma}_{\mathrm{sel}}^{(t)}$ by definition Equation 15. Multiplying the per-step bounds yields the action-likelihood bound. The $\overline{\Gamma}_{\mathcal{S}}^{H}$ consequence follows immediately; the stated $U \perp\!\!\!\perp S_t \mid Z_t$ special case follows because $P(s_t \mid z_{0:t}, a_{0:t-1}, u) = P(s_t \mid z_t)$ under these conditions, giving $\widetilde{\Gamma}_{\mathrm{sel}}^{(t)} = 1$. □

Remark 4 (Marginal vs. history-conditional state-selection)**.** *The trajectory bound uses the history-conditional $\widetilde{\Gamma}_{\mathrm{sel}}^{(t)}$ rather than the marginal $\Gamma_{\mathrm{sel}}^{(t)}$ of Proposition 7. If $P(s_t \mid z_{0:t}, a_{0:t-1}, u) = P(s_t \mid z_t, u)$, the history factor is at most the marginal factor; equality additionally requires a commonly supported history for a maximizing marginal comparison. This condition on the within-cell state law is stronger than Markov structure of the latent process alone. Without it, the two factors need not coincide or be ordered. Our empirical evaluation in Section 6 measures per-step latent sensitivity $\Gamma_{\mathcal{Z}}^{(t)}$ directly via logistic regression on the Z-space (Appendix H.4) without relying on the marginal-Γ_{sel} form of the trajectory bound. SPRL acts at each step by penalising within-cell propensity heterogeneity, providing an observable mitigation for the state-selection mechanism represented by $\widetilde{\Gamma}_{\mathrm{sel}}^{(t)}$.*

G Design Argument for the Propensity Penalty

In the small-bandwidth limit $\sigma \to 0$, for fixed finite-valued codes, Proposition 9 identifies the population pair ratio underlying $\mathcal{L}_{\mathrm{prop}}$ with twice a collision-weighted conditional variance of a fixed propensity score. Lemma 10 separately bounds a latent marginal action-probability shift by the within-cell range of the true propensity. The finite-bandwidth empirical penalty encourages local homogeneity of a fitted score; its average does not by itself bound that range or latent action sensitivity.

G.1 Population Interpretation of the Penalty

For a fixed bounded score $q(s)$, define its within-cell variance

$$V_q(z) \;:=\; \mathrm{Var}\big(q(S) \mid Z = z\big). \tag{16}$$

At $q = e$, this is $V_e(z)$; also write $\bar{e}(z) := \mathbb{E}[e(S) \mid Z = z]$.

Proposition 9 (Population pair ratio for finite-valued codes)**.** *Fix a measurable encoder ϕ whose representation $Z = \phi(S)$ takes finitely many values, and fix a bounded measurable score q. Let S, S' be independent population draws, and set $K_\sigma(s, s') := \exp(-\|\phi(s) - \phi(s')\|^2/(2\sigma^2))$ for $\sigma > 0$. The population kernel ratio*

$$L^{\mathrm{pop}}_{q,\sigma}(\phi) := \frac{\mathbb{E}[K_\sigma(S, S')\{q(S) - q(S')\}^2]}{\mathbb{E}[K_\sigma(S, S')]}$$

satisfies

$$\lim_{\sigma \to 0} L^{\mathrm{pop}}_{q,\sigma}(\phi) = 2\,\frac{\sum_z P(Z = z)^2 V_q(z)}{\sum_z P(Z = z)^2}.$$

Proof. As $\sigma \to 0$, $K_\sigma(S, S')$ converges pointwise to $\mathbf{1}[Z = Z']$. Boundedness of q and $0 \le K_\sigma \le 1$ allow dominated convergence in the numerator and denominator. Write $p_z := P(Z = z)$. Independence gives a denominator limit $P(Z = Z') = \sum_z p_z^2 > 0$. Conditional on $Z = Z' = z$, the two scores are independent draws from $q(S) \mid Z = z$, so $\mathbb{E}[\{q(S) - q(S')\}^2 \mid Z = Z' = z] = 2V_q(z)$. The numerator therefore converges to $2\sum_z p_z^2 V_q(z)$, yielding the stated ratio. □

In Equation 8, $q = \hat{e}$ is the fitted teacher, frozen during encoder training. Its finite minibatch ratio excludes self-pairs and, at finite bandwidth, measures a kernel-weighted local discrepancy of that teacher. The proposition concerns independent population draws; taking $\sigma \to 0$ in a fixed minibatch of distinct continuous codes does not establish it. Replacing $V_{\hat{e}}$ by V_e requires an additional accuracy argument for the fitted teacher.

G.2 Bounding the Marginal Latent Shift

We next give a pointwise bound using the range of the true propensity $e(s)$. The latent *marginal* action probability is

$$\tilde{\pi}_b^{\mathrm{marg}}(1 \mid z, u) \;:=\; \sum_{s:\phi(s)=z} P(s \mid z, u)\, e(s),$$

which mixes the marginal per-state propensity $e(s)$ against within-cell weights $P(s \mid z, u)$. The latent confounded policy $\tilde{\pi}_b(1 \mid z, u)$ from Equation 1 mixes the *confounded* per-state policy $\pi_b(1 \mid s, u)$ instead.

Lemma 10 (Marginal latent shift bound)**.** *Define the within-cell propensity range $R(z) := \max_{s:\phi(s)=z} |e(s) - \bar{e}(z)|$, and let $\Delta(z) := \max_{u,u' \in \mathcal{U}} \|P(S \mid Z = z, U = u) - P(S \mid Z = z, U = u')\|_1$. Then for any z, u, u',*

$$\big|\tilde{\pi}_b^{\mathrm{marg}}(1 \mid z, u) - \tilde{\pi}_b^{\mathrm{marg}}(1 \mid z, u')\big| \;\le\; R(z) \cdot \Delta(z). \tag{17}$$

Proof. Let $\Delta P(s) := P(s \mid z, u) - P(s \mid z, u')$, so $\sum_s \Delta P(s) = 0$ and $\sum_s |\Delta P(s)| \le \Delta(z)$. Then

$$\tilde{\pi}_b^{\mathrm{marg}}(1 \mid z, u) - \tilde{\pi}_b^{\mathrm{marg}}(1 \mid z, u') = \sum_s \Delta P(s)\, e(s) = \sum_s \Delta P(s)\,(e(s) - \bar{e}(z))$$

where the last equality subtracts $\bar{e}(z) \cdot \sum_s \Delta P(s) = 0$. Bounding by $L^\infty \cdot L^1$:

$$\Big|\sum_s \Delta P(s)\,(e(s) - \bar{e}(z))\Big| \le \max_s |e(s) - \bar{e}(z)| \cdot \sum_s |\Delta P(s)| \le R(z) \cdot \Delta(z). \qquad \square$$

Lemma 10 gives a small marginal shift when the true within-cell range $R(z)$ is small. An average of conditional variances cannot generally bound $R(z)$: a rare state can have a large propensity difference while contributing little to the average. A small empirical penalty on $\hat{e}$ therefore gives no population range bound without further assumptions.

G.3 Limitations of the Marginal Proxy

The marginal-propensity proxy and the confounded per-state policy differ in two related ways.

(i) Marginal vs. confounded per-state policy. Lemma 10 bounds shifts in $\tilde{\pi}_b^{\mathrm{marg}}(1 \mid z, u)$, not in the confounded $\tilde{\pi}_b(1 \mid z, u)$ from Equation 1. The connection between the two is mediated by per-state confounding: writing $\pi_b(1 \mid s, u) = e(s) \cdot c(s, u)$ for some confounding multiplier $c(s, u)$ with $\mathbb{E}[c(S, U) \mid S = s] = 1$,

$$\widetilde{\pi}_b(1 \mid z, u) \;=\; \widetilde{\pi}_b^{\mathrm{marg}}(1 \mid z, u) \;+\; \underbrace{\sum_s P(s \mid z, u)\, e(s)\, \big(c(s, u) - 1\big)}_{=:\,\tau(z,u)}, \tag{18}$$

where the cross-term $\tau(z, u)$ is what $\mathcal{L}_{\mathrm{prop}}$ does not directly control. Under the pairwise constraint, $|c(s, u) - 1|$ is bounded as a function of $\Gamma_{\mathcal{S}}$, so $|\tau(z, u)|$ contributes at most an additive $O(\Gamma_{\mathcal{S}} - 1)$ term beyond the marginal-shift bound. The full multiplicative decomposition is in Proposition 7.

(ii) Identical propensities, different per-state policies. Two states s_1, s_2 with $e(s_1) = e(s_2)$ can still satisfy $\pi_b(1 \mid s_1, u) \neq \pi_b(1 \mid s_2, u)$ for individual u, provided the discrepancies cancel under $P(u \mid s)$. The fitted-propensity penalty therefore cannot rule out within-cell heterogeneity of the confounded per-state policies. Appendix B.1 makes this boundary precise: exact marginal homogeneity preserves the sensitivity level for binary U, whereas explicit three-valued-U constructions have constant true propensities within cells and $\Gamma_{\mathcal{Z}} > \Gamma_{\mathcal{S}}$.

G.4 Scope of the Proxy

The implemented penalty measures local discrepancies in $\hat{e}$, whereas Lemma 10 uses the range of the true e. For binary U, exact population homogeneity of e gives the separate sensitivity guarantee in Proposition 6; a small empirical penalty does not establish that premise. An additive-logit assignment rule $\pi_b(1 \mid s, u) = \mathrm{logit}^{-1}\{h(s) + \beta(u)\}$ does not generally make $c(s, u) = \pi_b(1 \mid s, u)/e(s)$ constant across states. Thus additive logits do not remove the distinction between the marginal and confounded policies described above.

G.5 The Two-Condition View of Amplification

The state-selection mechanism of Theorem 1 can amplify $\Gamma_{\mathcal{Z}}$ when two conditions hold simultaneously:

(C1) U shifts the within-cell state composition ($\Delta(z) > 0$ in Lemma 10);

(C2) States within the cell have heterogeneous action profiles, so the shift in (C1) translates to a shift in the induced action distribution.

The exact action-sensitivity condition $U \perp\!\!\!\perp S \mid Z$ blocks amplification by negating (C1). This is the cleaner mathematical target but requires oracle access to U, which is unavailable. SPRL targets the observable marginal component of (C2). When the true within-cell propensity range is small, Lemma 10 bounds the resulting marginal shift; the empirical penalty encourages this property without certifying it.

This design choice has a counterintuitive empirical signature that we observe in Section 6: SPRL can preserve high U-decodability from Z (U-AUC close to Raw State) while still reducing $\Gamma_{\mathcal{Z}}/\Gamma_{\mathcal{S}}$. The Balanced baseline of Shalit et al. [2017] matches $P(Z \mid A = 0)$ and $P(Z \mid A = 1)$ and exhibits lower linear U-decodability (U-AUC ≈ 0.58 in Setting B) but still amplifies ($\Gamma_{\mathcal{Z}}/\Gamma_{\mathcal{S}} \approx 4.45\times$) because distributional balancing does not prevent the within-cell state-selection mechanism. These findings illustrate why mitigating amplification need not remove U from Z. SPRL encourages homogeneity of the fitted marginal propensity; the binary-U population guarantee concerns exact homogeneity of the true propensity.

H Experimental Details

H.1 Datasets and Reproducibility

The IHDP dataset (25 covariates; 747-row covariate pool resampled with replacement to $n = 5{,}000$ per seed) is obtained from the CEVAE repository [Louizos et al., 2017].[2] We use Minari's `mujoco/halfcheetah/medium-v0` dataset (HalfCheetah-v5, 17-dimensional state, about one million transitions), in the D4RL benchmark family [Fu et al., 2020], rather than the original D4RL `halfcheetah-medium-v0` file.[3] IHDP retains covariates but discards original assignments and outcomes; HalfCheetah retains physics trajectories and rewards while injecting synthetic binary actions and confounding. The implementation, experiment configurations, and seed-level numerical results are available at `https://github.com/sy-siyu/SPRL`. The repository includes an MIT license, environment specifications, training and evaluation commands, table-reconstruction scripts, and instructions for obtaining the external datasets separately.

External asset provenance. The HalfCheetah dataset metadata credits Kallinteris Andreas and records Minari version 0.5.2. The IHDP covariate data, originally introduced by Hill [2011] and distributed via the CEVAE repository [Louizos et al., 2017], is used solely for non-commercial academic research. We could not verify dataset-specific licenses from the hosted HalfCheetah metadata or the CEVAE IHDP distribution. Minari's software license does not establish distribution terms for these dataset files.

H.2 Encoder, Predictor, and Propensity Architectures

All neural components are MLPs with ReLU activations, fully connected, with no dropout, batch normalisation, or layer normalisation.

Encoder φ. A 3-layer MLP $\mathbb{R}^{d_S} \to \mathbb{R}^H \to \mathbb{R}^H \to \mathbb{R}^{dz}$ with hidden width $H = 256$ and ReLU between layers; the output layer is linear. The same encoder class is used for the LDM and SPRL methods; the AE, VAE, contrastive, and balanced baselines use a structurally identical encoder. Per-benchmark input/output dimensions are listed in Table 9.

Table 9: Encoder input/output dimensions per benchmark. d_S: raw state dimension; d_Z: latent dimension; H: hidden width. No dropout or normalisation layers are used in any benchmark.

Benchmark	d_S	d_Z	H	dropout	norm
Setting A	12	2	256	none	none
Setting B	12	5	256	none	none
IHDP	25	2	256	none	none
HalfCheetah	17	4	256	none	none

Reward predictor g_R. A 3-layer MLP $\mathbb{R}^{dz+1} \to \mathbb{R}^{128} \to \mathbb{R}^{128} \to \mathbb{R}$ with ReLU activations and a linear scalar output. The action $A \in \{0, 1\}$ is concatenated with $\varphi(S)$ as a single extra input feature; the output is a scalar $\hat{R}$. The same architecture is used across all four benchmarks.

Transition predictor g_T (multi-step settings only). A 3-layer MLP $\mathbb{R}^{dz+1} \to \mathbb{R}^{128} \to \mathbb{R}^{128} \to \mathbb{R}^{dz}$, ReLU activations, linear output. Following the Joint Embedding Predictive Architecture (JEPA) approach [LeCun et al., 2022, Assran et al., 2023], the prediction target is implemented as $\varphi(S_{t+1}).\mathrm{detach}()$ in PyTorch (the stop-gradient in Equation 7). The transition head is enabled for Setting B and HalfCheetah and disabled for Setting A and IHDP (one-step bandits).

Propensity estimator $\hat{e}$. A 3-layer MLP $\mathbb{R}^{d_S} \to \mathbb{R}^{H_e} \to \mathbb{R}^{H_e} \to \mathbb{R}$ with ReLU activations and a sigmoid output. The width is $H_e = 128$ for Setting A and Setting B (synthetic) and $H_e = 256$ for IHDP and HalfCheetah (matching the encoder hidden width). Trained by binary cross-entropy on the observed (S, A) pairs, optimiser Adam with learning rate 10^{-3} and batch size 256 for 100 epochs;

[2] `https://github.com/AMLab-Amsterdam/CEVAE/tree/master/datasets/IHDP`

[3] `https://minari.farama.org/datasets/mujoco/halfcheetah/medium-v0/`

weights are then *frozen* for the duration of representation training. The frozen estimates $\hat{e}(S_i)$ are passed into $\mathcal{L}_{\text{prop}}$ as a fixed signal, so no gradients flow back through the propensity network.

Kernel bandwidth for $\mathcal{L}_{\text{prop}}$. The Gaussian kernel weight in $\mathcal{L}_{\text{prop}}$ uses the median heuristic: σ is set to the median pairwise Euclidean distance $\text{median}_{i \neq j} \|\varphi(S_i) - \varphi(S_j)\|_2$ within the current minibatch, recomputed every minibatch during training (lower-bounded at 0.1). Diagonal pairs $(i = i)$ are masked out so they do not dilute the denominator.

Optimiser and training schedule. Adam (β_1, β_2 at PyTorch defaults), learning rate 10^{-3}, batch size 256, no weight decay, gradient norm clipped at 1.0. The same optimiser settings are used for all components (encoder, reward head, transition head, propensity model). Total training epochs are 100 for Setting A, IHDP, and HalfCheetah, and 200 for Setting B.

Warm-start schedule for λ_{prop}. SPRL training uses a two-phase schedule. *Phase 1 (warmup).* For the first E_w epochs the encoder and predictors are trained with $\lambda_{\text{prop}} = 0$ (dynamics-only loss); this lets the encoder reach a usable representation before the homogeneity penalty is introduced. *Phase 2 (linear ramp).* Over the next E_r epochs the effective coefficient λ_{eff} is ramped linearly from 0 to the target λ_{prop}: $\lambda_{\text{eff}}(e) = \lambda_{\text{prop}} \cdot \min\big(1, (e - E_w + 1)/E_r\big)$ for $e \geq E_w$, and $\lambda_{\text{eff}} = 0$ otherwise. *Phase 3 (full).* After epoch $E_w + E_r$, $\lambda_{\text{eff}} = \lambda_{\text{prop}}$ for the remainder of training. The schedule values are $(E_w, E_r) = (50, 50)$ for Setting A and Setting B and $(25, 25)$ for IHDP and HalfCheetah. The LDM ablation runs the same architecture with $E_w = E_r = 0$ and $\lambda_{\text{prop}} = 0$ throughout.

Per-benchmark training summary. Training schedules and hyperparameters are summarised in Table 10.

Table 10: Per-benchmark training summary. E_{tot} is total encoder/predictor epochs; (E_w, E_r) are warmup and ramp epochs for λ_{prop}; E_{prop} is the propensity-model pre-training length, run *before* representation training begins and frozen thereafter. The Setting B λ sweep constitutes the ablation in Figure 2; the headline Table 2 reports $\lambda_{\text{prop}} = 1.0$.

Benchmark	E_{tot}	(E_w, E_r)	λ_{prop}	E_{prop}	batch	lr
Setting A	100	$(50, 50)$	1.0	100	256	10^{-3}
Setting B	200	$(50, 50)$	$\{0,\ 0.5,\ 1,\ 2,\ 5\}$	100	256	10^{-3}
IHDP	100	$(25, 25)$	1.0	100	256	10^{-3}
HalfCheetah	100	$(25, 25)$	1.0	100	256	10^{-3}

Reported Γ_Z estimator. Two estimators are used (full description in Appendix H.4): the *additive* model $\text{logit}\ \Pr(A{=}1 \mid Z, U) = \beta_Z^\top Z + \beta_U U$, and the *interaction* model that adds $Z \cdot U$ cross terms and yields per-sample odds ratios. Setting B and HalfCheetah report results from the interaction estimator (per-sample median and 95th percentile across seeds); IHDP uses the additive diagnostic, appropriate for its additive DGP. Setting A's interval evaluation instead uses fitted evaluation propensities.

H.3 Sign-Aware Samplewise Marginal-Sensitivity Evaluation

For a conditioning state X (Raw State S or a learned code Z), the population marginal sensitivity model compares the confounded propensity $P(A = 1 \mid X, U)$ with the *true* observed propensity $e_X(x) = P(A = 1 \mid X = x)$. The numerical evaluation uses $\Lambda = \exp(|w_u|)$ and a per-observation propensity $\hat{e}_i$. For $A_i = 1$, its candidate propensity lies in

$$\ell_i = \frac{\hat{e}_i}{\hat{e}_i + \Lambda(1 - \hat{e}_i)} \ \leq\ e_i^* \ \leq\ h_i = \frac{\Lambda \hat{e}_i}{\Lambda \hat{e}_i + 1 - \hat{e}_i}. \tag{19}$$

The candidate denominator for the observed action lies in $[d_i^-, d_i^+] = [\ell_i, h_i]$ when $A_i = 1$ and $[1 - h_i, 1 - \ell_i]$ when $A_i = 0$. With $t_i = \pi_e(A_i \mid X_i) R_i$, the reported samplewise endpoints are

$$\widehat{V}^- = \frac{1}{n} \sum_{i=1}^{n} \min\left\{ \frac{t_i}{d_i^-}, \frac{t_i}{d_i^+} \right\}, \qquad \widehat{V}^+ = \frac{1}{n} \sum_{i=1}^{n} \max\left\{ \frac{t_i}{d_i^-}, \frac{t_i}{d_i^+} \right\}. \tag{20}$$

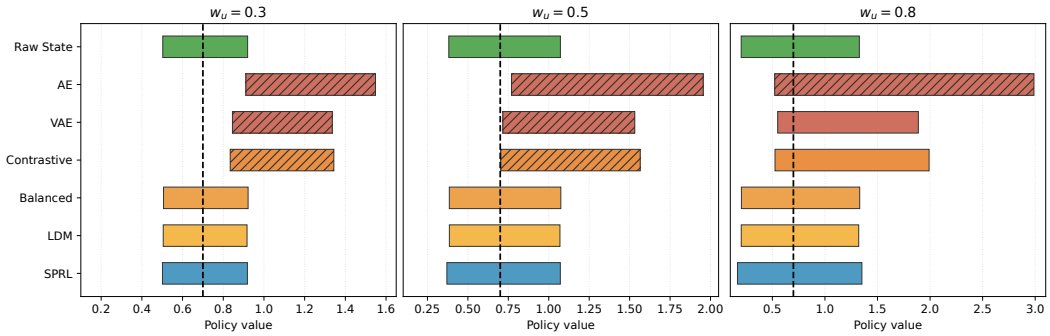


Figure 3: Setting A samplewise intervals. Each bar joins the mean lower and mean upper endpoint over 100 seeds; it is not a confidence interval for either endpoint. The dashed line is the analytic target 0.700; hatching denotes fewer than 95/100 seed intervals including that target.

Table 11: Setting A paired samplewise mean-width differences, SPRL minus comparator, over 100 common seeds per strength. Brackets are pointwise two-sided 95% t intervals; positive values mean wider SPRL intervals.

w_u	Comparator	SPRL minus comparator width [95%]
0.3	LDM	$+0.0055$ $[-0.0013, 0.0123]$
	Balanced	$+0.0017$ $[-0.0037, 0.0072]$
0.5	LDM	$+0.0178$ $[0.0038, 0.0318]$
	Balanced	$+0.0121$ $[0.0008, 0.0233]$
0.8	LDM	$+0.0660$ $[0.0375, 0.0945]$
	Balanced	$+0.0583$ $[0.0360, 0.0806]$

The minimum and maximum respect the sign of each reward contribution; no shared bin propensity is optimized after averaging rewards. Independent pointwise choices give conservative population outer bounds when e_X is known, the marginal sensitivity assumption holds for X, and outcome sufficiency identifies the same target-policy value from that state. They need not be sharp identified bounds. Replacing e_X by a fitted propensity and averaging finite observations adds estimation error: these endpoints are not finite-sample confidence intervals.

In Setting A, $n = 5{,}000$ observations per seed are used for both training and interval evaluation. Raw State uses the DGP's oracle $e_S(s) = \sum_u P(A = 1 \mid s, u)P(u \mid s)$; each learned representation uses its own ℓ_2-regularized logistic fit of A on Z from the same observations. The frozen neural teacher used to train SPRL is separate from this evaluation fit. The always-action-1 target has analytic value $V = 0.5 + 0.2 = 0.700$. The three nominal inputs $\Lambda = \exp(|w_u|)$ are 1.350, 1.649, and 2.226. Table 1 counts inclusion over $K = 100$ seeds. Its width ratios divide the mean method width by the mean samplewise Raw-State width within each strength; the paired width comparisons in Table 11 use the same seed IDs. For learned Z, this nominal input and fitted $\hat{e}_Z$ do not establish marginal-model membership or outcome sufficiency.

The IHDP evaluation uses 5,000 training observations and 10,000 fresh observations per seed. Learned evaluation propensities are fitted on training (Z, A) and then frozen for fresh observations; Raw State uses the exact observed-state propensity from the finite IHDP covariate pool. Its always-action-1 analytic target is 0.300. The same sign-aware calculation and nominal inputs are used, with $K = 100$ paired seeds and four methods. All inclusion counts in this paper describe empirical inclusion of these known targets.

H.4 Fitted Action-Sensitivity Diagnostic

The oracle diagnostic uses U after encoder training to fit an A-readout. For a fixed seed we divide its fitted odds ratios by the *nominal reference* $\exp(|w_u|)$; this is the per-state structural action-odds level in the synthetic assignment laws, not a claim that the fitted Raw-State statistic must equal one. A fitted ratio below one is not a population sensitivity certificate.

For Setting B and HalfCheetah, we fit the interaction model

$$\operatorname{logit} P(A = 1 \mid Z, U) = \beta_Z^\top Z + \beta_U U + \beta_{ZU}^\top (Z \odot U), \tag{21}$$

with standardized Z, giving the observation-level fitted ratio $\exp(|\hat{\beta}_U + \hat{\beta}_{ZU}^\top z_i|)/\exp(|w_u|)$. Each seed contributes its own median and p95 across evaluation observations; tables report the *mean across seeds* of those statistics separately. The p95 is an upper-tail summary, not a worst-case bound. The ℓ_2-regularized fit uses $C = \min(1, 5/d_Z)$ and clips counterfactual probabilities to $[0.005, 0.995]$ when forming odds ratios. Setting B uses 100 seeds. The HalfCheetah readout is fitted on 10 held-out episodes and applied unchanged to 10 further fresh episodes, disjoint from its 50 training episodes; all 100 paired seeds and extreme valid values enter the summaries.

IHDP uses the additive readout

$$\operatorname{logit} P(A = 1 \mid Z, U) = \beta_Z^\top Z + \beta_U U, \tag{22}$$

with $C = 10$ and standardized Z. Its fitted odds ratio $\exp(|\hat{\beta}_U|)/\exp(|w_u|)$ is constant within each seed, so that seed's median and p95 coincide. The readout is fitted on training observations and applied to 10,000 fresh replacement draws from the same finite covariate pool. This additive diagnostic is a modeling choice; additive raw assignment does not imply that every compressed conditional policy is additive. The reported across-seed statistic is the mean of the 100 seed ratios. Setting B's and HalfCheetah's interaction summaries remain distinct from IHDP's additive summaries.

H.5 Data-Generating Process Specifications

H.5.1 Synthetic Setting A (One-Step, $S_\Delta = \varnothing$)

Setting A uses a one-step confounded bandit with disjoint feature groups. The state $S \in \mathbb{R}^{12}$ is partitioned into noise ($S_\varnothing \in \mathbb{R}^2$), instruments ($S_a \in \mathbb{R}^5$), and clean dynamics features ($S_r \in \mathbb{R}^5$), with $S_\Delta = \varnothing$. The DGP is:

$$\begin{cases} U \sim \mathrm{Bern}(0.5), \\ S_\varnothing \sim \mathcal{N}(\mathbf{0},\, I_2), \\ S_a \sim \mathcal{N}\big((2U{-}1) \cdot c_s/\sqrt{5},\; I_5\big), \\ S_r \sim \mathcal{N}(\mathbf{0},\, I_5), \\ A \sim \mathrm{Bern}\big(\sigma(w_s^\top S + w_u \cdot U + b)\big), \\ R = w_r^\top S + 0.2 \cdot A + 0.5 + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,\, 0.01), \end{cases} \tag{23}$$

where $c_s = 2.0$ is the confounding shift magnitude, σ is the logistic function, $b = 0.0$, and w_r is nonzero only on S_r with entries drawn as $\mathcal{N}(0, 0.25)$. The policy weights are $w_s[S_a] = 0.5 \cdot \hat{c}$ (aligned with the confounding direction) and $w_s[S_r] = 0.45 \cdot \hat{w}_r$ (aligned with the reward direction), where $\hat{c}$ and $\hat{w}_r$ denote unit-normalised directions. The sensitivity parameter is $\Gamma_\mathcal{S} = \exp(|w_u|)$; we evaluate $w_u \in \{0.3, 0.5, 0.8\}$, giving $\Gamma_\mathcal{S} \in \{1.35, 1.65, 2.23\}$.

We use $n = 5{,}000$ samples per seed, $K = 100$ seeds, latent dimension $d_Z = 2$, hidden dimension 256, and 100 training epochs. SPRL uses $\lambda_{\text{prop}} = 1.0$, warmup 50, ramp 50. The target policy is $\pi_e(1 \mid s) = 1$ with true value $V_{\text{true}} = 0.700$.

H.5.2 Synthetic Setting B (Multi-Step, $S_\Delta \neq \varnothing$)

Setting B extends to a multi-step confounded MDP ($H = 5$) with overlapping feature groups. The state $S \in \mathbb{R}^{12}$ is partitioned into noise ($S_\varnothing \in \mathbb{R}^2$), instruments ($S_a \in \mathbb{R}^2$), confounded dynamics features ($S_\Delta \in \mathbb{R}^3$), and clean dynamics features ($S_r \in \mathbb{R}^5$). Let $S_{\text{dyn}} = S_\Delta \cup S_r \in \mathbb{R}^8$. The DGP is:

$$\begin{cases} U \sim \mathrm{Bern}(0.5), \\ S_0 \sim \mathcal{N}\big((2U{-}1) \cdot w_{\text{conf}},\; I_{12}\big), \\ A_t \sim \mathrm{Bern}\big(\sigma(w_s^\top S_t + w_u \cdot U - 0.3)\big), \\ S_{t+1,\text{dyn}} = W \cdot S_{t,\text{dyn}} + w_{\text{act}} \cdot A_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(\mathbf{0},\, 0.01 \cdot I_8), \\ R_t = w_r^\top S_t + 0.2 \cdot A_t + 0.5 + \varepsilon_r, \quad \varepsilon_r \sim \mathcal{N}(0,\, 0.01), \end{cases} \tag{24}$$

where $w_{\text{conf}}[S_a \cup S_\Delta] = c_s/\sqrt{|S_a \cup S_\Delta|}$ with $c_s = 2.0$ (zero elsewhere), $w_s \sim \mathcal{N}(0, 0.09 \cdot I_{12})$ (drawn once per seed), $W \in \mathbb{R}^{8\times 8}$ is a stable random matrix with spectral radius scaled to 0.8, $w_{\text{act}} \sim \mathcal{N}(0, 0.09 \cdot I_8)$, and w_r is nonzero on S_{dyn} with entries drawn as $\mathcal{N}(0, 0.25)$. Noise and instrument dims are re-sampled at each step (not dynamic), with instruments re-shifted by $(2U-1) \cdot w_{\text{conf}}[S_a]$. Thus the dynamical coordinates are conditionally U-separated, while the full-state transition kernel depends on U through the refreshed instrument coordinates. Setting B evaluates per-step action amplification under this broader DGP rather than the transition-separated action-likelihood bound of Corollary 8.

We use $n = 5{,}000$ trajectories (25,000 transitions) per seed, $K = 100$ seeds, $d_Z = 5$, hidden dimension 256, 200 training epochs. SPRL is tested at $\lambda_{\text{prop}} \in \{0, 0.5, 1.0, 2.0, 5.0\}$ with warmup 50 epochs and ramp 50 epochs.

H.5.3 IHDP Semi-Synthetic (One-Step)

We use the IHDP dataset [Hill, 2011] (747 observations, 25 covariates: 6 continuous, 19 binary) as a source of covariate structure. The 25 actual covariates exclude the original treatment and factual/counterfactual outcome columns. A scaler is fitted once on the 747 covariate rows. Each seed draws 5,000 training and 10,000 fresh replacement observations from this same subject pool; the fresh draws are not unseen subjects. The synthetic DGP is:

$$\begin{cases} U \sim \text{Bern}(0.5), \\ S \sim \text{resample from IHDP covariates}, \\ \tilde{S}[S_a] = S[S_a] + U \cdot c_s, \quad \tilde{S}[S_r] = S[S_r], \\ A \sim \text{Bern}\big(\sigma(w_s^\top \tilde{S} + w_u \cdot U)\big), \\ R = w_r^\top \tilde{S} + 0.3 \cdot A + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,\, 0.01), \end{cases} \tag{25}$$

where S_a consists of the first 10 dimensions (shifted by U, not reward-relevant), S_r consists of the remaining 15 dimensions (reward-relevant, not shifted), $c_s = 2.0$, $w_s[S_a] \sim \mathcal{N}(0, 0.01)$, $w_s[S_r] = 0.6 \cdot \hat{w}_r$ (aligned with the reward direction), and w_r is nonzero only on S_r with a normalised random direction scaled by 0.5. The nominal per-state action-odds level is $\exp(|w_u|)$ for $w_u \in \{0.3, 0.5, 0.8\}$. Raw-State evaluation uses the exact finite-support posterior $P(U \mid \tilde{S})$, obtained from counts of the two shifted covariate supports, to form $e_S(\tilde{s}) = \sum_u P(A = 1 \mid \tilde{s}, u) P(u \mid \tilde{s})$.

We use $K = 100$ paired seeds, $d_Z = 2$, hidden dimension 256, and 100 training epochs. SPRL uses $\lambda_{\text{prop}} = 1.0$, warmup 25, ramp 25. The target policy is $\pi_e(1 \mid s) = 1$ with analytic value 0.300; its reward coefficient vanishes on the shifted coordinates.

H.5.4 HalfCheetah Semi-Synthetic (Multi-Step)

We use Minari's `mujoco/halfcheetah/medium-v0` (HalfCheetah-v5) dataset, in the D4RL benchmark family [Fu et al., 2020]; it is not the original D4RL `halfcheetah-medium-v0` file. The environment has a 17-dimensional state space. We retain the original transitions $S_{t+1} = f_{\text{MuJoCo}}(S_t, A_t^{\text{orig}})$ and inject synthetic confounding per episode. The DGP is:

$$\begin{cases} U \sim \text{Bern}(0.5), \\ \tilde{S}_{t,d} = S_{t,d} + U \cdot c_s \quad \forall\, d \in \text{conf_dims},\ \forall\, t, \\ A_t \sim \text{Bern}\big(\sigma(w_s^\top \tilde{S}_t + 0.5 \cdot U)\big), \\ R_t = R_t^{\text{Minari}} \quad \text{(original physics reward)}, \end{cases} \tag{26}$$

where $c_s = 0.5$, conf_dims $= \{0, 1, 8, 11, 12\}$ (height, torso angle, forward velocity, back-thigh angular velocity, and back-shin angular velocity, using the default zero-based 17-coordinate HalfCheetah-v5 ordering), and the same shift is applied to both S_t and S_{t+1} targets. The injected binary A_t is recorded for the diagnostic; the preserved reward and next state were generated under the original continuous action. We therefore report no binary-action policy-value coverage for this benchmark. The retained state sequence follows the original MuJoCo physics, but the injected binary action does not cause the recorded S_{t+1}. The transition head therefore evaluates forecasting on physics-derived

trajectories rather than action-dependent control. Policy weights are placed on selected state coordinates: $w_s[16] = 0.3$ (front-foot angular velocity), $w_s[8] = -0.2$ (forward velocity), $w_s[1] = 0.15$ (torso angle), $w_s[10] = 0.1$ (torso angular velocity), with $b = 0.0$. The nominal structural action-odds reference is $\exp(0.5) \approx 1.65$; it need not equal the minimal supported sensitivity level of the empirical state distribution.

Per seed, 50 whole episodes train the encoder, 10 disjoint episodes fit the U-assisted readout, and 10 further episodes evaluate it (50,000/10,000/10,000 transitions). A state scaler is fitted on training episodes alone and frozen for the other splits. Representations use $d_Z = 4$, hidden dimension 256, 100 training epochs. SPRL uses $\lambda_{\text{prop}} = 1.0$, warmup 25, ramp 25. $K = 100$ paired seeds; no episode crosses splits.

H.6 Compute Resources

Setting B and supporting ablations used a workstation with an NVIDIA RTX 5080 (16 GB); its recorded total was approximately 35–40 GPU-hours. The Setting A evaluation used a local CPU and took about two wall-clock hours. The paired IHDP and HalfCheetah evaluations used a local CPU and took about 5.5 wall-clock hours combined (1.11 IHDP and 4.49 HalfCheetah summed training hours).

The reported configurations contain seven methods at three strengths and 100 seeds for Setting A and Setting B; IHDP contains four methods at three strengths and 100 paired seeds, and HalfCheetah contains four methods at one strength and 100 paired seeds. The controlled objective comparison uses ten paired Setting B seeds (Appendix I.1). The Minari `mujoco/halfcheetah/medium-v0` dataset is the largest external asset, about 1 GB on disk.

I Ablation Studies

This appendix isolates the contribution of $\mathcal{L}_{\text{prop}}$, studies sensitivity to latent dimension, examines the reported λ_{prop} choice across confounding strengths, and tests sensitivity to Gaussian teacher perturbations.

I.1 Controlled Objective Ablations

We run paired $K = 10$ Setting B experiments at $w_u = 0.3$, using common initialization within each pair to isolate the objective change from initialization variance. Adding $\mathcal{L}_{\text{prop}}$ lowers p95 amplification under both matched objectives:

Table 12: Paired objective ablations. Negative differences favour the objective with $\mathcal{L}_{\text{prop}}$.

Comparison	Mean paired Δ p95	Lower in seeds
Reward+$\mathcal{L}_{\text{prop}}$ vs. reward-only	-4.641	$10/10$
Full SPRL vs. LDM	-3.807	$9/10$

Thus the reduction occurs when $\mathcal{L}_{\text{prop}}$ is added either to reward-only or to LDM. The full-SPRL minus LDM p95 difference has a bootstrap 95% interval $[-6.97, -1.70]$; reward-probe MSE rises from 0.015 to 0.035. Propensity-only attains p95 2.240, but its reward and standardized dynamics probe MSEs are $0.417/0.725$, versus $0.035/0.265$ for full SPRL, and its latent scale is poorly conditioned. Unstabilized reward-only+$\mathcal{L}_{\text{prop}}$ variants share this conditioning issue. In a unit-sphere reward-only comparison, adding $\mathcal{L}_{\text{prop}}$ lowers p95 from 6.715 to 1.947 in $10/10$ seeds, while held-out reward MSE changes from 0.0174 to 0.0194. The direction therefore survives explicit scale control with a small reward-prediction cost.

I.2 Sensitivity to Latent Dimension

We repeat the paired Setting B comparison at $d_Z \in \{2, 5, 10\}$:

SPRL's benefit is capacity-dependent. At $d_Z = 2$ it does not improve Setting B amplification: relative to LDM, effective rank falls from 1.6 to 1.2, reward-probe MSE rises from 0.196 to 0.300, and

Table 13: Latent-dimension sensitivity at $w_u = 0.3$ ($K = 10$, mean $\pm$ SD). Negative differences favour SPRL.

d_Z	LDM p95	SPRL p95	SPRL–LDM	Lower in seeds
2	4.11 ± 1.16	4.53 ± 0.45	$+0.41$	$3/10$
5	6.06 ± 4.56	2.25 ± 0.89	-3.81	$9/10$
10	8.78 ± 3.81	2.02 ± 0.66	-6.76	$10/10$

dynamics-probe MSE rises from 0.491 to 0.563, consistent with a severe bottleneck. For context, the one-step Setting A and IHDP experiments use no transition head; their SPRL intervals include the analytic target in 100/100 seeds at $d_Z = 2$. Latent dimension must therefore provide enough capacity for the task objective alongside the propensity penalty.

I.3 Pareto Stability of λ_{prop} Across Confounding Strengths

The reported experimental default $\lambda_{\text{prop}} = 1.0$ was chosen from the oracle median-amplification/task-fidelity frontier at $w_u = 0.3$ (Figure 2a); it is not a deployable non-oracle automatic tuner. We examine whether this fixed experimental choice is consistent across confounding strengths by extracting the full λ sweep at all three w_u values from the same Setting B $K = 100$ run.

Table 14 reports Med $\Gamma_{\mathcal{Z}}/\Gamma_{\mathcal{S}}$ across the full grid; Table 15 reports transition MSE; Table 16 reports U-AUC.

Table 14: Med $\Gamma_{\mathcal{Z}}/\Gamma_{\mathcal{S}}$ across confounding strengths and λ_{prop} values (Setting B, K=100 each cell, interaction estimator). Bold marks the smallest λ at which Med $\Gamma_{\mathcal{Z}}/\Gamma_{\mathcal{S}} < 1.5$.

w_u ($\Gamma_{\mathcal{S}}$)	λ=0	λ=0.5	λ=1.0	λ=2.0	λ=5.0
0.3 (1.35)	2.88 ± 0.99	1.83 ± 0.72	$\mathbf{1.45 \pm 0.63}$	1.34 ± 0.72	1.14 ± 0.41
0.5 (1.65)	2.83 ± 0.88	1.68 ± 0.74	$\mathbf{1.26 \pm 0.41}$	1.17 ± 0.51	1.06 ± 0.46
0.8 (2.23)	2.93 ± 0.93	1.58 ± 0.77	$\mathbf{1.27 \pm 0.69}$	1.12 ± 0.54	0.89 ± 0.35

Table 15: Transition MSE across the (w_u, λ) grid.

w_u	λ=0	λ=0.5	λ=1.0	λ=2.0	λ=5.0
0.3	0.22	1.13	2.52	5.02	13.65
0.5	0.16	1.35	2.73	5.78	16.10
0.8	0.43	1.49	3.09	6.40	18.48

Across all three strengths, $\lambda = 1.0$ is the smallest tested value with median $\Gamma_{\mathcal{Z}}/\Gamma_{\mathcal{S}} < 1.5$. This consistency explains the reported fixed experimental default; it does not provide automatic selection of the operating point, which depends on the user's task-fidelity budget.

Logged-data selection analysis. We additionally run a $K = 10$ leave-one-seed-out analysis over $\lambda_{\text{prop}} \in \{0, 0.5, 1, 2, 5\}$. Selection uses only held-out logged-data quantities: reward-head NMSE and standardized ridge dynamics-probe MSE, each expressed relative to $\lambda = 0$, together with a robust neighborhood propensity-odds risk score. The latter is an observable diagnostic, not a sensitivity certificate. For each held-out seed, candidate statistics are computed from the other nine seeds; hidden U is revealed only after the choice is fixed.

A constrained rule minimizes observable risk subject to both task-error ratios being at most $1 + \tau$. At $\lambda = 1$, the leave-one-out median reward and dynamics ratios are 1.472 and 1.149. With $\tau = 0.25$, the rule selects $\lambda = 0$ in $10/10$ folds; with $\tau = 0.50$, it selects $\lambda = 1$ in $10/10$. Post-selection oracle evaluation gives mean p95 amplification 5.058 and 2.670, respectively, with the latter improving over $\lambda = 0$ in $8/10$ seeds. A threshold-free Pareto-knee rule selects $\lambda = 1$ in $8/10$ folds and 0.5 in $2/10$, also improving in $8/10$. Thus logged data expose a propensity-risk/task-fidelity trade-off, but the operating point changes with the rule or user-specified task budget; these results do not validate an automatic non-oracle tuner.

Table 16: Linear-probe U-AUC across the same grid.

w_u	λ=0	λ=0.5	λ=1.0	λ=2.0	λ=5.0
0.3	0.968	0.985	0.986	0.989	0.992
0.5	0.969	0.987	0.991	0.991	0.992
0.8	0.961	0.988	0.990	0.992	0.993

I.4 Post-hoc Diagnostic for Frozen Representations

Proposition 6 also suggests an observable screen for an already learned encoder. Given a frozen ϕ, we fit $\hat{e}(s) = P(A = 1 \mid S = s)$ from logged (S, A) data. For each evaluation point z_i, let $\mathcal{N}_k(i)$ be its k nearest neighbours in latent space and define the robust local odds range

$$\widehat{\rho}_k(i) := \frac{Q_{0.95}\{\text{odds}\hat{e}(S_j) : j \in \mathcal{N}_k(i)\}}{Q_{0.05}\{\text{odds}\hat{e}(S_j) : j \in \mathcal{N}_k(i)\}}.$$

We use the 95th percentile of $\widehat{\rho}_k(i)$ over evaluation points as a risk score. This is a robust continuous-space analogue of Γ_e, not a formal bound: neighbourhoods approximate rather than equal ϕ-preimages, and percentiles replace the population extrema.

We evaluate this diagnostic in a fresh $K = 10$ Setting B analysis at $w_u = 0.3$, using independent evaluation samples and architecture-matched LDM and SPRL encoders. Hidden U is used only after the observable score is fixed to measure oracle amplification. Across the ten paired seeds, SPRL ($\lambda_{\text{prop}} = 1$) has lower oracle p95 amplification in $8/10$ seeds, with means 2.670 versus 5.058 for LDM. Table 17 shows that the observable score ranks this pair in the same direction at all three neighbourhood resolutions and tracks oracle amplification across the five candidate objectives.

Table 17: Logged-data diagnostic on Setting B ($w_u = 0.3$, $K = 10$). "SPRL lower" is the number of paired seeds for which SPRL has lower risk than LDM. Correlations and oracle-best matches use LDM and four positive λ_{prop} settings.

k	LDM risk	SPRL risk	SPRL lower	Spearman ρ	Oracle-best matched
25	778.95	444.01	8/10	0.611	7/10
50	921.26	490.20	8/10	0.620	7/10
100	994.68	522.72	9/10	0.635	5/10

For an empirical check, we compose each encoder with a ten-cell K-means quantizer fitted on training representations. On an independent evaluation sample, we approximate Γ_e using sample extrema of the analytic propensity, and $\Gamma_{\mathcal{Z}}$ using Monte Carlo averages of the analytic behavior policy within each hidden group. Only cells containing at least 50 sampled observations from each group are eligible. Among 10 seeds and 6 representations, 58 partitions have eligible cells; all 58 satisfy the numerical comparison $\widehat{\Gamma}_{\mathcal{Z}} \leq \Gamma_{\mathcal{S}}\widehat{\Gamma}_e$, with a median right-to-left ratio of about 757. The other two partitions have no eligible cells. These sample calculations neither cover excluded cells nor compute the population worst-case range over entire cells. On the exact discrete cells of Theorem 1 and Appendix D, the corresponding ratios are 2.8 and 6.5. Proposition 7 is tighter when its state-selection factor is available, but that factor depends on hidden U; Γ_e trades tightness for observability.

The neighbourhood score can therefore flag and compare frozen representations using logged data, but it is a risk diagnostic rather than a coverage certificate or an automatic selector of λ_{prop}.

I.5 Sensitivity to Gaussian Teacher Perturbations

SPRL's penalty $\mathcal{L}_{prop}$ uses a pretrained, frozen propensity estimator $\hat{e}(s)$ (Section 5). We test sensitivity to Gaussian teacher perturbations by corrupting $\hat{e}$: $\hat{e}_{\text{corrupt}}(s) = \text{clip}(\hat{e}(s)+\varepsilon,\, 0.01,\, 0.99)$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, with ε freshly resampled at each minibatch.

Table 18 reports SPRL trained under this corruption on Setting B at $w_u = 0.5$, K=100 seeds across $\sigma \in \{0, 0.02, 0.05, 0.10, 0.20\}$. At $\sigma = 0.20$, mean fitted Med $\Gamma_{\mathcal{Z}}/\Gamma_{\mathcal{S}}$ rises from 1.26 at $\sigma = 0$ to 1.43, a 13.5% degradation that remains substantially below LDM's amplification of 2.83 at the

same w_u. The linear-probe U-AUC remains essentially flat across all σ, confirming that the encoder continues to retain U-information even when the propensity signal is degraded.

Table 18: SPRL under Gaussian teacher perturbations on Setting B (w_u=0.5, λ_{prop}=1, $\Gamma_{\mathcal{S}}$=1.65, K=100 seeds; mean ± SD). At each minibatch we resample $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and feed $\text{clip}(\hat{e} + \varepsilon, 0.01, 0.99)$ into $\mathcal{L}_{prop}$.

σ	Med $\Gamma_{\mathcal{Z}}/\Gamma_{\mathcal{S}}$	p95 $\Gamma_{\mathcal{Z}}/\Gamma_{\mathcal{S}}$	U-AUC	Trans MSE
0.00	1.26 ± 0.41	2.32 ± 1.40	0.991 ± 0.003	2.73
0.02	1.34 ± 0.68	2.26 ± 1.19	0.988 ± 0.017	2.71 ± 0.70
0.05	1.46 ± 0.83	2.47 ± 1.31	0.988 ± 0.011	2.66 ± 0.74
0.10	1.40 ± 0.65	2.33 ± 0.99	0.990 ± 0.007	2.50 ± 0.65
0.20	1.43 ± 0.69	2.45 ± 1.03	0.989 ± 0.011	2.03 ± 0.46

Across-seed standard deviation of fitted Med increases from 0.41 without noise to 0.65–0.83 at the tested noisy levels. Mean raw latent-transition MSE decreases from 2.73 to 2.03 as σ increases to 0.20; this scale-dependent statistic does not identify the mechanism of that change. Across these Gaussian perturbations, SPRL retains lower mean fitted median amplification than LDM. The experiment addresses the specified additive-noise protocol and does not establish robustness to systematic teacher bias or general propensity-model misspecification.

J Full Semi-Synthetic Results

SPRL uses a frozen neural propensity teacher fitted on raw logged (S, A) pairs. IHDP and HalfCheetah evaluate the same four methods within every paired seed: Raw State, Balanced, LDM, and SPRL, with 100 seeds per condition. Each method's fitted amplification is divided by nominal $\exp(|w_u|)$; the reported median is a within-seed median followed by a mean over seeds. Brackets on means and paired differences are pointwise two-sided 95% t intervals over paired seeds. Raw-relative width-ratio brackets are paired-bootstrap 95% intervals; inclusion counts have Wilson intervals.

J.1 IHDP

The one-step evaluation uses the 25 IHDP covariates and the sign-aware samplewise calculation of Appendix H.3. All four methods include the analytic action-1 target 0.300 in 100/100 intervals at each strength (Wilson 95% interval $[96.3\%, 100\%]$). SPRL's mean samplewise width exceeds both LDM's and Balanced's at each strength, while its fitted additive amplification is lower. The additive readout gives a constant ratio within a seed, so its within-seed p95 equals its within-seed median; the table reports their common mean.

The mean IHDP widths at $w_u = 0.3/0.5/0.8$ are $0.1160/0.1877/0.3005$ for Raw State, $0.1178/0.1913/0.3067$ for Balanced, $0.1159/0.1884/0.3013$ for LDM, and $0.1196/0.1953/0.3154$ for SPRL. Width ratios in Table 19 are ratios of these unrounded means. The SPRL-to-Raw ratio has paired-bootstrap intervals $[1.020, 1.041]$, $[1.031, 1.051]$, and $[1.039, 1.060]$. Fresh action-1 prediction MSE is $0.0146/0.0188/0.0145$ for SPRL versus $0.0039/0.0038/0.0038$ for LDM and 0.0033 for Balanced at each strength. The fitted evaluation propensity is distinct from the frozen neural teacher used in SPRL training.

J.2 HalfCheetah

The Minari evaluation applies the frozen interaction readout on episodes disjoint from training and readout fitting. Table 21 separates mean within-seed median and p95 ratios. The upper tail is skewed: Balanced has a valid seed with p95 1177.009 and SPRL one with p95 95.727; both are retained. The median diagnostic favors SPRL over LDM. Its paired difference against Balanced includes zero. Raw latent-transition MSEs are omitted as a method ranking because the latent target variances differ substantially (LDM 1.465, SPRL 102.532).

SPRL's fresh observed-reward MSE exceeds LDM's by 0.0563 $[0.0463, 0.0663]$ and Balanced's by 0.0557 $[0.0477, 0.0636]$ in paired means. Its fresh action-readout log loss is 0.408, versus 0.435

Table 19: IHDP, $K = 100$ paired seeds per w_u. Incl is empirical target inclusion; W/R is mean interval width divided by samplewise Raw-State mean width. Amp is mean seed-level fitted median (also p95 for this additive readout), normalized by nominal $e^{|w_u|}$. Reward MSE uses fresh observed rewards.

w_u	Method	Incl	W/R	Amp	Reward MSE
0.3	Raw State	100/100	1.000	1.136	—
	Balanced	100/100	1.015	1.616	0.0133
	LDM	100/100	0.999	1.459	0.0138
	SPRL	100/100	1.031	0.892	0.0244
0.5	Raw State	100/100	1.000	1.081	—
	Balanced	100/100	1.019	1.487	0.0133
	LDM	100/100	1.004	1.359	0.0139
	SPRL	100/100	1.041	0.729	0.0287
0.8	Raw State	100/100	1.000	1.035	—
	Balanced	100/100	1.021	1.419	0.0133
	LDM	100/100	1.003	1.309	0.0139
	SPRL	100/100	1.050	0.559	0.0244

Table 20: Paired IHDP SPRL-minus-comparator mean differences and pointwise 95% intervals. Negative amplification favors SPRL; positive width and reward-MSE differences favor the comparator. Amp is the additive readout's common within-seed median and p95.

w_u	Comparator	Amp difference	Width difference	Reward-MSE difference
0.3	LDM	-0.567 $[-0.734, -0.401]$	$+0.0037$ $[0.0025, 0.0048]$	$+0.0106$ $[0.0060, 0.0152]$
	Balanced	-0.724 $[-0.896, -0.551]$	$+0.0018$ $[0.0007, 0.0029]$	$+0.0111$ $[0.0065, 0.0158]$
0.5	LDM	-0.629 $[-0.798, -0.461]$	$+0.0070$ $[0.0050, 0.0090]$	$+0.0148$ $[0.0107, 0.0190]$
	Balanced	-0.757 $[-0.934, -0.581]$	$+0.0040$ $[0.0023, 0.0058]$	$+0.0154$ $[0.0112, 0.0196]$
0.8	LDM	-0.750 $[-0.923, -0.577]$	$+0.0141$ $[0.0108, 0.0173]$	$+0.0106$ $[0.0080, 0.0131]$
	Balanced	-0.861 $[-1.036, -0.685]$	$+0.0086$ $[0.0061, 0.0111]$	$+0.0112$ $[0.0086, 0.0137]$

for LDM and 0.512 for Balanced. These metrics quantify a prediction trade-off; the retained Minari reward was generated under the original continuous action, so no binary-action policy-value conclusion is made for this benchmark.

Table 21: Episode-disjoint HalfCheetah at $w_u = 0.5$ ($K = 100$ paired seeds). Med and p95 are separate means of within-seed interaction-readout statistics normalized by nominal $e^{0.5}$; median brackets are 95% t intervals. P95-seed med is the median across seed-level p95 statistics. Reward MSE is on fresh observed rewards.

Method	Med [95%]	Mean p95	P95-seed med	Reward MSE
Raw State	$0.884\ [0.857, 0.911]$	2.694	1.610	—
Balanced	$1.204\ [1.113, 1.294]$	15.988	2.366	0.0622
LDM	$1.621\ [1.467, 1.774]$	18.224	9.668	0.0616
SPRL	$1.103\ [1.043, 1.164]$	7.627	3.711	0.1179

Table 22: HalfCheetah paired SPRL-minus-comparator differences. Mean differences and brackets use 95% t intervals; the paired p95 median difference uses a 10,000-resample paired-bootstrap interval. All 100 seeds, including extremes, are retained.

Comparator	Median mean difference	P95 mean difference	Paired p95 median difference
LDM	$-0.517\ [-0.687, -0.347]$	$-10.597\ [-16.476, -4.718]$	$-3.602\ [-7.978, -0.768]$
Balanced	$-0.100\ [-0.211, 0.010]$	$-8.361\ [-31.652, 14.930]$	$+0.908\ [-0.043, 1.634]$

K Extended Related Work

This appendix expands on the state representation learning discussion in Section 7.

State representation learning for RL. Learned representations in RL include latent dynamics models [Hafner et al., 2023] and causally structured encoders that learn bilinear causal factorisations [Lin et al., 2024], action-sufficient representations with structural constraints [Huang et al., 2022], identifiable latent factorisation [Liu et al., 2023], latent confounder models for recommendation [Xu et al., 2025], and confounding-robust features [Gao et al., 2023]. Reward-preserving counterfactual state editing constrains augmented observations by reward consistency under the recorded action [Wang et al., 2026]. In the treatment effect literature, balanced representations matching treated and control distributions via Integral Probability Metric (IPM) penalties [Johansson et al., 2016, Shalit et al., 2017] predate similar ideas in offline RL; our Balanced baseline adapts this to RL using MMD. Learned states also serve as inputs to generative offline policies, including continuity-regularized flow models [Chen et al., 2026] and energy-guided diffusion policies for recommendation [Chen et al., 2025a]. These methods connect representation learning with prediction, policy quality, balance, and structural identifiability.

Representations and confounded OPE. Bellman-complete representations preserve structure needed for offline policy evaluation [Chang et al., 2022], while latent-confounder and proximal methods use proxy structure to identify value [Bennett et al., 2020, Bennett and Kallus, 2024]. Robust-MDP analysis instead bounds transition-kernel perturbations [Bennett et al., 2024], and sharp one-step MSM work characterizes outcome-aware partial-identification bounds [Dorn et al., 2022]. Distributional offline RL models return distributions for risk-averse policy learning [Chen et al., 2025b].

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 introduction distinguish the pairwise action-sensitivity results and conditional value-coverage theory from fitted empirical diagnostics. Setting A and IHDP target inclusion, amplification comparisons, and width/prediction costs appear in Section 6 and the experimental appendices.

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 8 discusses the pairwise theory's scope, the fitted marginal-propensity proxy, oracle-U diagnostics, and empirical interval and prediction costs.

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 results state their assumptions explicitly. Full proofs and constructions appear in Appendices A–G, including the binary-confounder proof in Appendix B.1.

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: Appendix H describes the architectures, training schedules, DGPs, sign-aware interval calculation, and evaluation protocols. The public repository linked in Appendix H.1 provides executable code, environment specifications, configurations, and seed-level results for table reconstruction.

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, configurations, numerical results, and reproduction instructions are publicly available at `https://github.com/sy-siyu/SPRL`. Synthetic data are generated locally; instructions identify the separate IHDP and versioned Minari HalfCheetah sources (Appendix H.1).

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: Appendix H gives training details, DGPs, and the Setting A, IHDP, and HalfCheetah evaluation protocols.

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: Section 6 reports 100-seed inclusion counts and seed-level summary means. Appendix tables give paired 95% intervals for width and fitted amplification differences, Wilson intervals for inclusion, and separate median and p95 summaries; the matched objective ablation has ten paired seeds.

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: Appendix H.6 reports the GPU and CPU resources and approximate runtimes for the experiments.

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 research conforms with the NeurIPS Code of Ethics. The experiments use simulated data and existing benchmark data; potential deployment harms are discussed in the Broader Impact section.

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: Broader Impact section discusses positive impacts and negative impacts.

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: This work releases neither pre-trained models nor scraped datasets; there are no high-misuse-risk artifacts.

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: [No]

    Justification: IHDP and the versioned Minari HalfCheetah source are credited in Appendix H.1, but we could not verify dataset-specific licenses from the distribution sources. Minari's software license does not resolve these dataset distribution terms.

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 code release includes a README, an MIT license, pinned dependencies, experiment configurations, data-access instructions, and documentation of the evaluation protocols and saved numerical results. External datasets and trained checkpoints are not redistributed.

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: No participants were recruited and no crowdsourcing was conducted. The IHDP experiments reuse existing benchmark covariates (Appendix H.1).

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: No new participant study was conducted. The IHDP analysis reuses publicly distributed benchmark covariates with synthetic actions and outcomes (Appendix H.5.3).

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 were not used as a non-standard component of the methodology.