Action-Value Geometry Under Hidden Confounding: GeoSCOPE for Reliability-Oriented Off-Policy Evaluation

Jingyang He
Department of Computer Science
Florida State University
Tallahassee, FL 32306
jh24o@fsu.edu

Zhen Lei
Department of Computer Science
Florida State University
Tallahassee, FL, 32306
zl24a@fsu.edu

Ang Li
Department of Computer Science
Florida State University
Tallahassee, FL, 32306
angli@cs.fsu.edu

Abstract

Offline policy evaluation (OPE) is often used to screen policies learned from observational data before deployment. With hidden confounding, an unobserved episode-level variable may affect behavior actions, state transitions, and rewards; observed-state OPE can then overestimate policies whose interventional values are lower. Because this setting is nonidentifiable without additional assumptions, we do not pursue hidden-variable recovery or certified causal value bounds. We instead study a deployable reliability target: reducing optimistic screening errors from one observational dataset. We introduce GeoSCOPE, a proxy-only estimator that adds local, one-sided pessimism to standard FQE. Observation-compatible latent worlds are used only at design time to identify stable, ambiguous, support-fragile, and value-inflated action-value patterns. At evaluation time, GeoSCOPE replaces these oracle roles with observable local-geometry and support proxies, forms a proxy-cautious Bellman reference, and projects only positive gaps. In oracle-evaluable hidden-confounded Sepsis and Two-Corridor experiments, GeoSCOPE reduces optimistic and selected-policy overestimation relative to standard FQE and pessimistic or robust adaptations, exposing a reliability–accuracy tradeoff without using hidden variables, compatible-world indices, true values, or oracle labels.

1 Introduction

Offline policy evaluation (OPE) is often the screening step before policies learned from observational data are considered for deployment [36, 35, 14, 8, 38]. That screening role becomes fragile when the observed state omits a hidden confounder. Such a variable can directly affect the behavior policy's action choices, the state transitions, and the reward. The behavior policy may then select actions using information that the offline evaluator and the target policy will not observe. Even after conditioning on the same observed state and action, the recorded transitions and rewards can be averages over latent groups selected by the behavior policy.

The resulting problem is structural bias, not just finite-sample uncertainty. More data from the same observational process, or reweighting that conditions only on observed states, does not change the latent mixture in the data. A target policy that cannot observe the confounder may face a different mixture after deployment. The screening failure we focus on is optimistic overestimation: the

Preprint.

evaluator gives a policy credit for favorable outcomes selected into the behavior data, and the policy is promoted because its estimated value is too high. In safety-critical domains such as healthcare, this error direction matters more than symmetric accuracy alone.

Without additional assumptions, hidden-confounded OPE is nonidentifiable: the same observed trajectory law can be compatible with different latent mechanisms, and those mechanisms can assign different values to the same observed-state target policy [30, 31, 2, 7]. Because of this nonidentifiability, we do not try to recover the hidden confounder, identify the true latent mechanism, or certify a causal lower bound. Our goal is narrower: given one observational dataset, can an evaluator reduce optimistic screening errors relative to standard observed-state OPE? This is a reliability target, not an identification claim.

Scalar uncertainty alone misses part of this failure mode because fitted OPE methods build a policy score from many local estimates. In FQE, for example, the reported value is aggregated from estimates $Q^\pi(s, a)$ over observed states and actions [10, 40, 1]. Hidden confounding can make a small region look too favorable because the corresponding actions were mostly taken by a favorable latent group. That region need not dominate average error bars, but policies that place probability mass on it can still be over-scored. We therefore study the geometry of learned action values: the local pattern that helps separate estimates that appear stable from those that are ambiguous, support-fragile, or inflated.

We use observation-compatible latent worlds to analyze this failure mode [29, 28, 3, 41]. Compatible worlds are alternative latent factorizations of the same marginalized observational law. They agree on observed data, but can imply different counterfactual values and action-value relations. These worlds are not available to a deployed estimator; they are a design-time lens for studying what stable anchors, ambiguous relations, support-fragile regions, and value-inflated failures look like before replacing those oracle roles with observable proxies.

GeoSCOPE is the deployable version of this idea. It is not a generic pessimism penalty: it constructs local-geometry and support proxies from standard FQE and the observed data, forms a proxy-cautious reference, and projects only positive gaps:

$$Q_{\mathrm{GeoSCOPE}} = \min\{Q_{\mathrm{std}}, Q_{\mathrm{ref}}\}.$$

The projection is local because hidden-confounding distortion is heterogeneous, and one-sided because positive value gaps are the screening failure of interest. Stable local structure moderates the correction, so caution does not collapse into uniform shrinkage [19, 39, 15]. At evaluation time, the estimator uses no hidden variables, compatible-world indices, true values, oracle role labels, policy-family labels, or confounding-strength labels.

We evaluate GeoSCOPE in two oracle-evaluable hidden-confounded environments: a diabetes-confounded Sepsis simulator [32] and the Two-Corridor environment with exact hidden-state dynamic-programming values. These settings are needed because optimistic OPE error requires true interventional values and is not directly observable in ordinary real-world observational data. We compare ordinary OPE, uncertainty-based LCB adaptations, global pessimism, CVaR-FQE, Delphic-LCB diagnostic baselines, and GeoSCOPE [11, 21, 6, 29]. The comparison separates three failure modes: accuracy-oriented estimators can retain optimistic violations despite moderate absolute error, robust lower-tail methods reduce optimism only with a large accuracy cost, and compatible-world LCB baselines are calibration sensitive. In the main settings, GeoSCOPE drives mean optimistic over-gap close to zero and reduces selected-policy overestimation, while making the reliability–accuracy tradeoff explicit.

Contributions.

- We frame hidden-confounded observed-state OPE around optimistic screening error, a deployment-relevant reliability target distinct from symmetric accuracy and from certified causal identification.
- We use observation-compatible latent worlds to diagnose local action-value patterns–stable, ambiguous, support-fragile, and inflated–that explain how hidden confounding can mislead policy screening.
- We introduce GeoSCOPE, a proxy-only positive-gap projection estimator that constructs observable local-geometry and support proxies, forms a proxy-cautious reference, and applies asymmetric correction from single-dataset inputs.

- We evaluate GeoSCOPE in oracle-evaluable hidden-confounded Sepsis and Two-Corridor settings, showing reduced optimistic and selected-policy overestimation and an explicit reliability–accuracy tradeoff.

2 Hidden-confounded OPE and diagnostic compatible worlds

2.1 Observed-state OPE with a hidden trajectory variable

We study episodic off-policy evaluation from observational trajectories [35]. Each episode draws an unobserved variable $u \sim \nu$, which remains fixed throughout the trajectory. The initial-state, transition, and reward mechanisms may depend on u. The behavior policy also observes u, so actions are sampled from $\pi_b(a \mid s, u)$. The offline dataset, however, contains only observed states, actions, rewards, and next states, and the target policies evaluated in this paper are observed-state policies $\pi(a \mid s)$. Figure 1 depicts this structure.

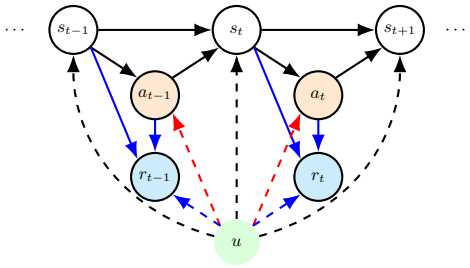

Figure 1: Hidden-confounded observed-state OPE. The behavior policy and environment depend on a latent trajectory variable u, but the offline evaluator observes only (s, a, r, s') trajectories and target policies condition only on s.

For an observed trajectory $\tau = (s_0, a_0, r_0, \ldots, s_T, a_T, r_T, s_{T+1})$, the observational data are sampled from the marginal behavior law

$$P_{\text{obs}}(\tau) = \sum_u \nu(u)\rho_0(s_0 \mid u) \prod_{t=0}^{T} \pi_b(a_t \mid s_t, u) P(r_t, s_{t+1} \mid s_t, a_t, u).$$

The learner observes samples from P_{obs}, but not the latent variable or the conditional factors inside the mixture. The issue is therefore not simply that the state is incomplete. At a fixed observed pair (s, a), the empirical conditional $P_{\text{obs}}(r, s' \mid s, a)$ averages over latent groups selected by the behavior policy. A target-policy intervention changes the action mechanism and may change this latent mixture. Bellman backups computed only from observed states may then use the wrong mixture for the target policy. We use standard observed-state OPE terminology here; related estimators include density-ratio, generalized stationary-value, minimax, and double-RL methods [26, 42, 37, 16].

2.2 Nonidentifiability and the reliability target

The interventional value of an observed-state target policy is defined under the target-policy intervention:

$$V_M^\pi = \mathbb{E}_{M,\pi}\left[\sum_{t=0}^{T} \gamma^t r_t\right],$$

where $\mathbb{E}_{M,\pi}$ samples $u \sim \nu$, $s_0 \sim \rho_0(\cdot \mid u)$, $a_t \sim \pi(\cdot \mid s_t)$, and $(r_t, s_{t+1}) \sim P(\cdot \mid s_t, a_t, u)$ for $t = 0, \ldots, T$. In hidden-variable causal models, this quantity is not generally identified by the marginal observational law P_{obs} [30, 31].

Classical nonidentifiability. The complete identification criterion for recursive semi-Markovian causal models states the relevant limitation: when an intervention is not identifiable, two latent causal models can share the same observational distribution but differ under intervention [34]. For our finite-horizon setting, viewed as an unrolled causal graph, the same observed trajectory law can therefore be compatible with different latent mechanisms that assign different values to the same target policy. This result is used only to set the scope of the paper; it is not a new theoretical contribution.

We therefore do not attempt to recover the hidden variable, identify the true latent mechanism, or produce a certified causal lower bound. Instead, we study a weaker but deployable objective: reducing optimistic screening errors when an evaluator must operate on a single observational dataset. For a target policy π, the dangerous error is the positive estimation gap

$$(\widehat{V}^{\pi} - V_M^{\pi})_+,$$

because it can cause offline screening to promote an overestimated policy. This leads to the metrics in Section 5: violation rate and mean over-gap for reliability, mean absolute error for accuracy, and an appendix under-gap measure that separates conservative underestimation from other accuracy costs.

2.3 Compatible worlds as a diagnostic lens

To see how nonidentifiability appears in learned value functions, we use observation-compatible latent worlds. A compatible world

$$w = (\mathcal{U}_w, \nu_w, \rho_{0,w}, P_w, \pi_{b,w})$$

is a latent factorization that induces the same marginalized observational trajectory law:

$$P_{\mathrm{obs}}(\tau) = \sum_{u \in \mathcal{U}_w} \nu_w(u)\rho_{0,w}(s_0 \mid u) \prod_{t=0}^{T} \pi_{b,w}(a_t \mid s_t, u) P_w(r_t, s_{t+1} \mid s_t, a_t, u).$$

Two compatible worlds agree on all observed trajectory probabilities but may assign different interventional values and action-value functions to the same observed-state target policy.

In practice, we cannot enumerate all compatible worlds. We therefore train a finite, empirically compatible diagnostic family $\widehat{\mathcal{W}}(D) = \{w_1, \ldots, w_K\}$ from the observed trajectories, treating it as an empirical approximation to alternative latent factorizations of the same observational data. The family is for mechanism analysis only; Section 4 replaces compatible-world access with observable proxies. Latent-variable causal representation learning provides related diagnostic background [22]; Appendix A.4.1 gives the construction details and data-access assumptions.

Compatible worlds are a diagnostic device, not a deployment-time estimator. The simulator confounding-strength parameter α is not a compatible-world index. Changing α creates a different stress-test data-generating process; compatible worlds instead refer to alternative latent factorizations of a fixed observed trajectory law.

With this setup, the next section asks which local action-value relations remain stable across compatible explanations and which become distorted.

3 Compatible-world diagnostic of action-value geometry

Section 2 shows that the same observed trajectory law can admit multiple latent explanations. Here we use the finite compatible-world family as a design-time diagnostic rather than a route to the true latent world. The goal is to see where unresolved hidden confounding enters the learned action-value function, and what that implies for an evaluator that must later operate only on observed data.

3.1 Why value relations, not point estimates

Hidden confounding makes the target-policy value nonidentified, but the OPE failure that matters for policy screening is more specific than scalar uncertainty [17, 18, 4, 5, 27]. Offline screening ranks policies through learned value relations. A value function can therefore be misleading even when its average pointwise movement is small. Conversely, pointwise values can move substantially while preserving the relations that determine policy comparison.

Two elementary cases make this concrete. If every compatible latent world shifts all action values by a world-dependent constant, pointwise values vary while all pairwise action-value relations are preserved. If only a small local region changes across worlds, the average pointwise movement can be arbitrarily small even though the affected local relation changes by order one. Pointwise movement is therefore neither necessary nor sufficient for detecting control-relevant distortion. Formal constructions and proofs are given in Appendix A.5. These cases motivate a relation-level

diagnostic: instead of asking only how much individual Q-values move, we ask whether the local action-value geometry used for policy comparison is stable across compatible latent explanations.

3.2 Pairwise action-value geometry under compatible worlds

For a compatible latent world w, let $q_w^\pi(s) = (Q_w^\pi(s,a))_{a\in\mathcal{A}}$ denote the action-value profile of state s. We diagnose relation-level stability through the pairwise Mean Q-gap

$$G_w^\pi(s_i, s_j) = \frac{1}{|\mathcal{A}|} \sum_{a\in\mathcal{A}} \left| Q_w^\pi(s_i, a) - Q_w^\pi(s_j, a) \right| .$$

Intuitively, $G_w^\pi(s_i, s_j)$ measures how far apart two observed states are in their action-value profiles under world w. We use this quantity only to inspect whether local value geometry remains stable across compatible explanations. It is not an OPE estimator, and GeoSCOPE does not use it at deployment time. We use a uniform action average because this section diagnoses the geometry of the learned Q-function itself, rather than the value of a particular policy-induced action distribution.

3.3 Mechanism roles revealed by localized distortion

We summarize the compatible-world analysis with three signals computed only after the fact: cross-world relation distortion, empirical support, and directional value shift (roughly, instability across compatible worlds, observed-data coverage, and signed changes in local value relations; formal definitions are in Appendix A.4.2). In this analysis, the signals separate several local behaviors under hidden confounding [20, 23, 33]. Some local relations remain stable and supported, some are supported but ambiguous across latent explanations, some are dominated by weak observational support, and some show concentrated value inflation. We refer to these cases as mechanism roles: explanatory categories used to motivate the design, not labels available to GeoSCOPE at deployment time.

Table 1: Compatible-world mechanism roles and their design implications. The roles are post-hoc diagnostic prototypes, not deployment-time labels.

Role	Compatible-world signature	Design implication
Stable anchor	Stable relation geometry; adequate support	Offset unnecessary caution
Ambiguous relation	Supported but unstable across worlds	Activate caution
Support-fragile relation	Weak observed support	Guide pair construction and diagnostics
Value-inflated failure	Local distortion or optimistic inflation	Motivate positive-gap projection

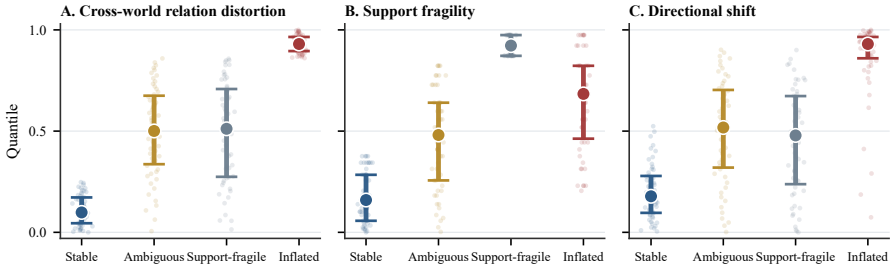

Figure 2: **Compatible-world evidence for localized mechanism roles.** The panels report post-hoc oracle diagnostics used only for analysis: cross-world relation distortion, support fragility, and directional shift. State pairs are grouped by explanatory prototypes. GeoSCOPE does not receive these labels; Section 4 replaces them with observable proxies.

Figure 2 gives a compact empirical check of this taxonomy. The four groups separate in the three-signal space: stable anchors combine low distortion with adequate support, ambiguous relations have support but high distortion, support-fragile relations lack coverage, and value-inflated failures show directional optimistic shift. The point is not to recover these groups at deployment, but to make the design target explicit before replacing them with observable proxies in Section 4.

3.4 From diagnostic roles to design requirements

The compatible-world analysis is not a deployment procedure. Its useful output is a set of design constraints for a proxy-only evaluator that must work without the latent mechanism. Pessimism

should be local: distortion is heterogeneous across state pairs, so a global penalty would also shrink regions that appear stable. Correction should be asymmetric, since the screening failure is optimistic overestimation rather than arbitrary scalar error. This is why uncertainty-only penalties can remain misaligned when the bias is structural rather than sampling-driven. Finally, stable local anchors should offset ambiguity; without them, caution reduces to uniform shrinkage. Section 4 implements these constraints with observable single-dataset proxies.

4 GeoSCOPE: mechanism-guided positive-gap projection

GeoSCOPE implements the three requirements identified in Section 3.4: locality, asymmetry, and role preservation. The estimator receives only the observed dataset and a target policy; it does not observe hidden confounders, compatible-world indices, confounding strength, true values, oracle diagnostic labels, policy-family labels, or evaluation outcomes [25, 13].

The bridge from compatible-world roles to single-dataset proxies is empirical rather than identifying. Appendices C.5 and C.6 report proxy-to-mechanism enrichment, cross-fitting, and sensitivity checks; these support role alignment but do not constitute a causal lower-bound or proxy-identification guarantee [24].

4.1 Observable role proxies

For each nearby observed state pair (s_i, s_j), GeoSCOPE computes an observable pair score $S_{\mathrm{proxy}}(s_i, s_j)$ from local action-value geometry and empirical support:

$$S_{\mathrm{proxy}}(s_i, s_j) = R(\mathrm{sep}(s_i, s_j) + \mathrm{amb}(s_i, s_j) + \mathrm{supp}(s_i, s_j)) \,.$$

Here R rank-normalizes over the current pair table: sep scores learned action-value separation, amb scores disagreement between value-distance, observed-state distance, and support, and supp scores weak empirical support. Low scores indicate stable nearby anchors and high scores indicate ambiguous or support-fragile nearby relations; full formulas, fixed hyperparameters, and sensitivity checks are in Appendices A and C.6.

Let q_{low} and q_{high} be the lower and upper quartiles of S_{proxy} among nearby pairs, where near is a fixed observed-distance neighborhood rule. We define

$$H_{\mathrm{safe}} = \{(s_i, s_j) : \mathrm{near}(s_i, s_j),\ S_{\mathrm{proxy}}(s_i, s_j) \le q_{\mathrm{low}}\},$$

$$H_{\mathrm{mid}} = \{(s_i, s_j) : \mathrm{near}(s_i, s_j),\ S_{\mathrm{proxy}}(s_i, s_j) \ge q_{\mathrm{high}}\}.$$

Low-score nearby pairs serve as stable local anchors, while high-score nearby pairs serve as ambiguity evidence. We aggregate pair evidence into state-level scores,

$$\rho_{\mathrm{safe}}(s) \propto \sum_{(s,j)\in H_{\mathrm{safe}}} (1 - S_{\mathrm{proxy}}(s, j)), \qquad \rho_{\mathrm{mid}}(s) \propto \sum_{(s,j)\in H_{\mathrm{mid}}} S_{\mathrm{proxy}}(s, j).$$

Both state scores are rank-normalized to $[0, 1]$ after aggregation. Support scarcity enters the final projected estimator through the pair score and nearby-pair construction; it is not an additional multiplicative projection gate in the main estimator.

4.2 Proxy-cautious reference

The state-level gate compares ambiguity evidence against stable local evidence:

$$g_{\mathrm{raw}}(s) = \frac{\rho_{\mathrm{mid}}(s) - \rho_{\mathrm{safe}}(s)}{\rho_{\mathrm{mid}}(s) + \rho_{\mathrm{safe}}(s) + \epsilon}, \qquad \eta_{\mathrm{mid}}(s) = \mathcal{N}([g_{\mathrm{raw}}(s)]_+) \,.$$

Here $\mathcal{N}$ is a fixed positive-tail normalization to $[0, 1]$, not tuned using true values or OPE outcomes. Thus stable anchors reduce the ambiguity gate, and caution is activated only when ambiguity dominates nearby stable evidence.

We then solve a proxy-cautious reference

$$Q^{\pi}_{\mathrm{ref}}(s, a) = (1 - \eta_{\mathrm{mid}}(s)) \left[\hat{r}(s, a) + \gamma \sum_{s'} \hat{P}(s'|s, a) \sum_{a'} \pi(a'|s') Q^{\pi}_{\mathrm{ref}}(s', a')\right] .$$

This reference is an operational benchmark for what standard FQE can justify under observed geometry. It is not the final estimator, is not a certified lower bound, and need not be pointwise below standard FQE; the one-sided correction comes from the subsequent minimum projection.

4.3 Positive-gap projection and estimator

GeoSCOPE removes only the positive part of standard FQE that exceeds the proxy-cautious reference:

$$\Delta_Q^\pi(s,a) = [Q_{\mathrm{std}}^\pi(s,a) - Q_{\mathrm{ref}}^\pi(s,a)]_+ ,$$

$$Q_{\mathrm{GeoSCOPE}}^\pi(s,a) = Q_{\mathrm{std}}^\pi(s,a) - \Delta_Q^\pi(s,a) = \min\{Q_{\mathrm{std}}^\pi(s,a), Q_{\mathrm{ref}}^\pi(s,a)\}.$$

This asymmetric correction keeps standard FQE whenever it is already below the reference and corrects only local positive gaps. The final estimate is

$$\widehat{V}_{\mathrm{GeoSCOPE}}^\pi = \sum_s \hat{\rho}_0(s) \sum_a \pi(a|s) Q_{\mathrm{GeoSCOPE}}^\pi(s,a).$$

This projection has two simple operational properties. For any fixed target policy, GeoSCOPE cannot increase the scalar FQE estimate and therefore cannot increase fixed-policy optimistic over-gap relative to standard FQE. Moreover, for a fixed gate $\eta_{\mathrm{mid}} \in [0,1]$, the proxy-cautious reference operator is a γ-contraction. Appendix A.3 gives proofs; these are well-definedness and one-sidedness properties, not causal lower-bound guarantees.

Algorithm 1 GeoSCOPE: proxy-only positive-gap projection

Require: observed dataset $\mathcal{D}$, target policy π, discount γ
1: Fit empirical transition and reward models $(\hat{P}, \hat{r})$ from $\mathcal{D}$.
2: Fit standard FQE on $\mathcal{D}$ to obtain Q_{std}^π.
3: Compute observable pair scores S_{proxy} from local value geometry and support.
4: Select nearby anchor pairs H_{safe} and ambiguous pairs H_{mid}.
5: Aggregate pair evidence into $\rho_{\mathrm{safe}}(s)$ and $\rho_{\mathrm{mid}}(s)$.
6: Construct ambiguity gate $\eta_{\mathrm{mid}}(s)$ and solve the proxy-cautious reference Q_{ref}^π.
7: Project positive gaps: $Q_{\mathrm{GeoSCOPE}}^\pi(s,a) = \min\{Q_{\mathrm{std}}^\pi(s,a), Q_{\mathrm{ref}}^\pi(s,a)\}$.
8: Return $\widehat{V}_{\mathrm{GeoSCOPE}}^\pi = \sum_s \hat{\rho}_0(s) \sum_a \pi(a|s) Q_{\mathrm{GeoSCOPE}}^\pi(s,a)$.

Algorithm 1 summarizes the observed-data computation used to produce the scalar OPE score.

Complexity. GeoSCOPE adds one proxy-construction pass and one additional Bellman fixed-point solve on top of FQE; Appendix D.3 gives the full finite-state complexity.

5 Experiments

Our experiments evaluate OPE as a policy-screening score rather than as a pure absolute-error leaderboard. For each target policy π, let $\widehat{V}^\pi$ be the OPE estimate and V_{true}^π the post-hoc interventional value used only for evaluation. We report

$$\mathrm{Viol} = \mathbb{E}_\pi[\mathbf{1}\{\widehat{V}^\pi > V_{\mathrm{true}}^\pi\}], \quad \mathrm{Over} = \mathbb{E}_\pi[(\widehat{V}^\pi - V_{\mathrm{true}}^\pi)_+],$$
$$\mathrm{Under} = \mathbb{E}_\pi[(V_{\mathrm{true}}^\pi - \widehat{V}^\pi)_+], \quad \mathrm{Abs} = \mathbb{E}_\pi[|\widehat{V}^\pi - V_{\mathrm{true}}^\pi|].$$

Violation and over-gap measure reliability by quantifying optimistic screening risk; absolute error measures symmetric accuracy; under-gap, reported in the appendix, decomposes conservative underestimation. All metrics are lower better. Low absolute error is not sufficient if violations remain frequent, and low violation is not sufficient if values collapse downward.

5.1 Evaluation protocol: oracle-evaluable hidden-confounded OPE

Here we evaluate the screening question directly: when an OPE score ranks candidate policies from a hidden-confounded observational dataset, does it avoid promoting policies whose values are optimistically overestimated? Ordinary observational data cannot answer this question because target-policy interventional values are unobserved, so we use two oracle-evaluable hidden-confounded environments.

Table 2: **Strong-confounding OPE results at $\alpha = 0.75$.** Violation and over-gap measure reliability; absolute error measures accuracy. CVaR-FQE and Delphic-LCB are robust diagnostic baselines shown at fixed operating points; full τ- and λ-sweeps, under-gap decompositions, and standard errors are in Appendix C.

Method	Sepsis			Two-Corridor		
	Viol.	Over	Abs.	Viol.	Over	Abs.
PDWIS	0.650	0.035	0.044	1.000	2.054	2.054
Standard FQE	0.683	0.047	0.062	1.000	0.494	0.494
Bootstrap LCB q10	0.667	0.042	0.059	1.000	0.489	0.489
Ensemble LCB	0.650	0.042	0.058	1.000	2.226	2.226
Global coarse	0.667	0.045	0.061	1.000	0.381	0.381
Target-Q global	0.683	0.047	0.062	1.000	0.237	0.237
CVaR-FQE ($\tau = 0.1$)	0.000	0.000	1.248	0.000	0.000	0.820
Delphic-LCB ($\lambda = 5$)	0.133	0.005	0.109	0.333	0.144	2.931
GeoSCOPE	0.117	0.005	0.106	**0.250**	**0.001**	0.037

Sepsis. In the diabetes-confounded Sepsis simulator, diabetic status is observed by the behavior policy and affects downstream outcomes, but it is omitted from the offline dataset, target policies, and all OPE estimators. The main strong-confounding setting uses $\alpha = 0.75$ and evaluates 60 held-out target policies from BC, BCQ, and CQL families [9, 19]. True target-policy values are estimated only after OPE by simulator rollouts.

Two-Corridor. Two-Corridor is a finite navigation task with a hidden route-risk variable. The behavior policy observes this variable and tends to choose the safer corridor, whereas the offline dataset and target policies omit it. Because the hidden-state MDP is known for evaluation, true target-policy values are computed by exact dynamic programming. We use five independently generated offline datasets, each with 60 target policies, and report mean $\pm$ standard error over dataset seeds.

In both environments, hidden variables, oracle values, and confounding-strength labels are reserved for post-hoc evaluation and diagnostics; every OPE method receives only the observed dataset and target policy. We compare GeoSCOPE with standard FQE, PDWIS, bootstrap and ensemble LCB-FQE adaptations [11], global pessimism, Target-Q global pessimism [21], CVaR-FQE, and Delphic-LCB diagnostic baselines. CVaR-FQE serves as a lower-tail, sensitivity-style robust Bellman baseline [6], while Delphic-LCB directly uses compatible-world value dispersion at evaluation time [29]. These rows are diagnostic operating points, fixed without true values, not claims of dominance over guarantee-driven MSM or confounded-POMDP OPE methods. Those methods address complementary regimes with explicit uncertainty sets or latent-state assumptions; GeoSCOPE is a proxy-only OPE-stage correction. Full baseline details and sweeps are in Appendices B.4–C.

5.2 Main results: reliability–accuracy frontier

Table 2 reports strong-confounding OPE results at $\alpha = 0.75$, comparing accuracy-oriented estimators, global and robust pessimism, compatible-world pessimism, and GeoSCOPE under the same held-out policy banks.

Interpretation. The table is a reliability–accuracy frontier, not an absolute-error leaderboard. Standard FQE and related accuracy-oriented baselines retain frequent optimistic violations despite moderate absolute error, which is exactly the screening failure considered here. CVaR-FQE removes optimism at this operating point by moving far below the true values. Delphic-LCB is competitive in Sepsis for an aggressive λ, but it uses compatible-world values at evaluation time and is sensitive to that coefficient. GeoSCOPE stays in the proxy-only setting: compatible worlds are used only for diagnosis, and the deployed correction is a local positive-gap projection from observed data.

Selected-policy deployment diagnostic. The practical use of OPE is often ranking, not estimating all policies equally. For each estimator, we select the top ten policies according to its own OPE scores and then measure selected optimistic over-gap and absolute error. Table 3 reports the selected-policy diagnostic for representative ordinary, robust, and compatible-world diagnostic baselines.

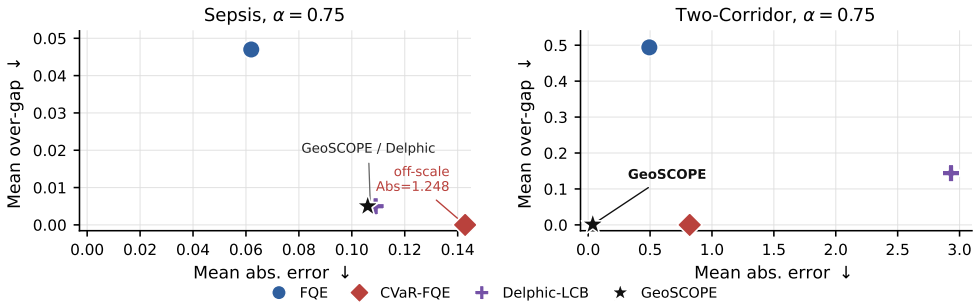

Figure 3: **Reliability–accuracy frontier at strong confounding.** The x-axis is mean absolute error and the y-axis is mean over-gap; lower-left is better. Representative estimators are shown, with full results in Table 2.

Table 3: **Selected-policy diagnostic at strong confounding.**

Method	Sepsis			Two-Corridor		
	Over@10	Viol.@10	Abs@10	Over@10	Viol.@10	Abs@10
Standard FQE	0.138	1.000	0.138	0.497	1.000	0.497
Bootstrap LCB q10	0.128	1.000	0.128	0.492	1.000	0.492
Ensemble LCB	0.128	1.000	0.128	2.257	1.000	2.257
CVaR-FQE ($\tau = 0.1$)	0.000	0.000	1.235	0.000	0.000	0.821
Delphic-LCB ($\lambda = 5$)	0.019	0.600	0.031	0.465	1.000	0.465
GeoSCOPE	0.013	0.500	0.029	0.004	0.600	0.010

Table 4: **Mechanism role ablation on raw-state Sepsis at $\alpha = 0.75$.**

Variant	Viol.	Over	Under	Abs.
Standard FQE	0.683	0.0466	0.0155	0.0620
Amb-only proxy-cautious reference	0.000	0.0000	0.3220	0.3220
Anchor-def. proxy-cautious reference	0.017	0.0006	0.1894	0.1900
Amb-masked proj.	0.233	0.0157	0.0546	0.0703
Anchor-def. mask	0.467	0.0291	0.0262	0.0553
GeoSCOPE	**0.117**	**0.0046**	0.1017	0.1064

The selected-policy diagnostic looks at the policies an evaluator would actually promote. Standard FQE and uncertainty-penalty baselines still select overestimated policies. CVaR-FQE avoids selected optimism by large underestimation. Delphic-LCB reduces Sepsis top-ten optimism only with compatible-world values at evaluation time, and remains optimistic in Two-Corridor. GeoSCOPE keeps selected over-gap low in both domains with a smaller accuracy cost and no compatible-world access at deployment. Bootstrap and Ensemble LCB show the same mismatch as in the full-policy table: uncertainty around observational FQE does not remove structural optimistic bias.

5.3 Ablations and stress tests

Mechanism ablation. Table 4 separates the roles of ambiguity, stable anchors, and projection. Reference-only variants remove optimism but pay large under-gap, while masked projection lowers under-gap yet leaves more violations. GeoSCOPE combines ambiguity and anchor evidence through positive-gap projection, keeping over-gap low with bounded accuracy cost.

Stress tests and cross-fitting. Appendices C.6–C.8 report cross-fitting, projection-strength, weak-confounding, threshold/neighborhood, policy-class-shift, global-calibration, and controlled hidden-risk treatment checks. The policy-class-shift check sweeps route-preference policies in Two-Corridor from behavior-aligned to more aggressive route choices under full action coverage, separating hidden-U selection bias from ordinary missing-action support. Projection-strength and threshold/neighborhood sweeps show the expected reliability–accuracy frontier: stronger projection improves optimism control while increasing conservative underestimation. Cross-fit GeoSCOPE preserves the reliability pattern when proxy construction and projected evaluation are separated, and the hidden-risk treatment MDP is used only as a third mechanistic stress test rather than a benchmark.

6 Conclusion

Hidden-confounded OPE is nonidentifiable, but the deployment failure studied here is more specific: an evaluator can promote policies whose values are optimistically biased. GeoSCOPE starts from compatible-world diagnostics of stable, ambiguous, support-fragile, and value-inflated local relations, and turns these roles into a proxy-only estimator that constructs a proxy-cautious reference and clips positive FQE gaps. Across Sepsis and Two-Corridor, the results support a reliability–accuracy view of OPE: low absolute error can still be unsafe when high scores are optimistic, while broad pessimism can avoid optimism only by sacrificing substantial accuracy. GeoSCOPE should therefore be read as a reliability-oriented evaluator under unresolved hidden confounding, not as an identification method or a certified causal lower bound.

References

[1] Rishabh Agarwal, Marlos C. Machado, Pablo Samuel Castro, and Marc G Bellemare. Contrastive behavioral similarity embeddings for generalization in reinforcement learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=qda7-sVg84.

[2] Alexander Balke and Judea Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997. doi: 10.1080/01621459.1997.10474074. URL https://doi.org/10.1080/01621459.1997.10474074.

[3] Elias Bareinboim, Andrew Forney, and Judea Pearl. Bandits with unobserved confounders: A causal approach. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/795c7a7a5ec6b460ec00c5841019b9e9-Paper.pdf.

[4] Andrew Bennett and Nathan Kallus. Policy evaluation with latent confounders via optimal balance. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/7c4bf50b715509a963ce81b168ca674b-Paper.pdf.

[5] Andrew Bennett, Nathan Kallus, Lihong Li, and Ali Mousavi. Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1999–2007. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/bennett21a.html.

[6] Yinlam Chow and Mohammad Ghavamzadeh. Algorithms for cvar optimization in mdps. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/35f1050a4381d2d216bf56ad46b0277d-Paper.pdf.

[7] Alexander D'Amour. On multi-cause approaches to causal inference with unobserved counfounding: Two cautionary failure cases and a promising alternative. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 3478–3486. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/d-amour19a.html.

[8] Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1447–1456. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/farajtabar18a.html.

[9] Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2052–2062. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/fujimoto19a.html.

[10] Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G. Bellemare. DeepMDP: Learning continuous latent space models for representation learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2170–2179. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/gelada19a.html.

[11] Kamyar Ghasemipour, Shixiang Shane Gu, and Ofir Nachum. Why so pessimistic? estimating uncertainties for offline rl through ensembles, and why their independence matters. In *Advances in Neural Information Processing Systems*, volume 35, pages 18267–18281, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/7423902b5534e2b267438c85444a54b1-Paper-Conference.pdf.

[12] Lan-Zhe Guo and Yu-Feng Li. Class-imbalanced semi-supervised learning with adaptive thresholding. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8082–8094. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/guo22e.html.

[13] Mao Hong, Zhengling Qi, and Yanxun Xu. Model-based reinforcement learning for confounded POMDPs. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 18668–18710. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/hong24d.html.

[14] Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 652–661, 2016. URL https://proceedings.mlr.press/v48/jiang16.html.

[15] Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5084–5096. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/jin21e.html.

[16] Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167):1–63, 2020. URL http://jmlr.org/papers/v21/19-827.html.

[17] Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/3a09a524440d44d7f19870070a5ad42f-Paper.pdf.

[18] Nathan Kallus and Angela Zhou. Confounding-robust policy evaluation in infinite-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 22293–22304, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/fd4f21f2556dad0ea8b7a5c04eabebda-Paper.pdf.

[19] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 1179–1191, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/0d2b2061826a5df3221116a5085a6052-Paper.pdf.

[20] Daniel Kumor, Junzhe Zhang, and Elias Bareinboim. Sequential causal imitation learning with unobserved confounders. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 14669–14680. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/7b670d553471ad0fd7491c75bad587ff-Paper.pdf.

[21] Jie Liu, Yinmin Zhang, Chuming Li, Yaodong Yang, Yu Liu, and Wanli Ouyang. Adaptive pessimism via target q-value for offline reinforcement learning. *Neural Networks*, 180: 106588, 2024. doi: 10.1016/j.neunet.2024.106588. URL https://www.sciencedirect.com/science/article/pii/S0893608024005124.

[22] Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/94b5bde6de888ddf9cde6748ad2523d1-Paper.pdf.

[23] Miao Lu, Yifei Min, Zhaoran Wang, and Zhuoran Yang. Pessimism in the face of confounders: Provably efficient offline reinforcement learning in partially observable markov decision processes. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PbkBDQ5_UbV.

[24] Afsaneh Mastouri, Yuchen Zhu, Limor Gultchin, Anna Korba, Ricardo Silva, Matt Kusner, Arthur Gretton, and Krikamol Muandet. Proximal causal learning with kernels: Two-stage estimation and moment restriction. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7512–7523. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/mastouri21a.html.

[25] Rui Miao, Zhengling Qi, and Xiaoke Zhang. Off-policy evaluation for episodic partially observable markov decision processes under non-parametric models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 593–606. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/03dfa2a7755635f756b160e9f4c6b789-Paper-Conference.pdf.

[26] Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/cf9a242b70f45317ffd281241fa66502-Paper.pdf.

[27] Hongseok Namkoong, Ramtin Keramati, Steve Yadlowsky, and Emma Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18819–18831. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/da21bae82c02d1e2b8168d57cd3fbab7-Paper.pdf.

[28] Michael Oberst and David Sontag. Counterfactual off-policy evaluation with gumbel-max structural causal models. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4881–4890, 2019. URL https://proceedings.mlr.press/v97/oberst19a.html.

[29] Alizée Pace, Hugo Yèche, David Ha, Gunnar Ratsch, and Guy Tennenholtz. Delphic offline reinforcement learning under nonidentifiable hidden confounding. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/0b9ff9baa226af39f86045ecdc173672-Paper-Conference.pdf.

[30] JUDEA PEARL. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 12 1995. ISSN 0006-3444. doi: 10.1093/biomet/82.4.669. URL https://doi.org/10.1093/biomet/82.4.669.

[31] Judea Pearl. *Causality*. Cambridge university press, 2009.

[32] Aniruddh Raghu, Matthieu Komorowski, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach. In Finale Doshi-Velez, Jim Fackler, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 147–163. PMLR, 18–19 Aug 2017. URL https://proceedings.mlr.press/v68/raghu17a.html.

[33] Chengchun Shi, Masatoshi Uehara, Jiawei Huang, and Nan Jiang. A minimax learning approach to off-policy evaluation in confounded partially observable Markov decision processes. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 20057–20094. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/shi22f.html.

[34] Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence*, pages 1219–1226, 2006. URL https://cdn.aaai.org/AAAI/2006/AAAI06-191.pdf.

[35] 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.

[36] Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2380–2388, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/thomas15.html.

[37] Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9659–9668. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/uehara20a.html.

[38] Cameron Voloshin, Hoang Minh Le, Nan Jiang, and Yisong Yue. Empirical study of off-policy policy evaluation for reinforcement learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL https://openreview.net/forum?id=IsK8iKbL-I.

[39] Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 6683–6694. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/34f98c7c5d7063181da890ea8d25265a-Paper.pdf.

[40] Amy Zhang, Rowan Thomas McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=-2FCwDKRREu.

[41] Junzhe Zhang, Daniel Kumor, and Elias Bareinboim. Causal imitation learning with unobserved confounders. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12263–12274. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/8fdd149fcaa7058caccc9c4ad5b0d89a-Paper.pdf.

[42] Ruiyi Zhang, Bo Dai, Lihong Li, and Dale Schuurmans. Gendice: Generalized offline estimation of stationary values. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HkxlcnVFwB.