Learning Deployable Causal Action Geometry under Temporal Non-Stationarity

Changjian Liu[1,2 *†] **Tianyu Wang**[3*] **Yuwei Xu**[4*†] **Xiaoxuan Deng**[3] **Zhilin Zhang**[5]
Yong Gao[1,2‡] **Chuan Yu**[3] **Jian Xu**[3‡] **Bo Zheng**[3]

[1] Institute of Remote Sensing and Geographic Information System,
School of Earth and Space Sciences, Peking University, Beijing, China
[2] Beijing Key Laboratory of Spatio-temporal Perception and Urban Resilience
[3] Taobao & Tmall Group of Alibaba, Beijing, China
[4] The Chinese University of Hong Kong, Shenzhen, China [5] Independent Researcher

Abstract

Learning a response law for continuous decisions from observational time series is central to many real-world decision systems, yet temporal non-stationarity poses a fundamental challenge: as outcomes drift over time, naively pooled models can conflate action effects with shifting baselines and even reverse action rankings. Even when the one-step response is identifiable, unrestricted cross-time generalization remains fundamentally ill-posed. We therefore seek the part of the response that remains stable for decisions even when outcome levels do not. We propose deployable causal action geometry: the action-response shape that remains after anchoring away action-independent level shifts on a selected link scale and profiling out positive amplitude variation. Building on this target, we develop a two-stage orthogonal estimator that first obtains cross-fitted pilot estimates of anchored responses and then performs honest profiled selection over link, representation, and shape classes. The resulting estimator admits oracle guarantees for the restricted target and, under approximate persistence, produces decision rules that remain near optimal in future periods. Experiments on synthetic data and Dominick's support the proposed target and its transfer behavior. In a randomized Taobao A/B test, the deployed system improves GMV and purchase counts by 1.16% and 1.70%.

1 Introduction

Continuous decision making from observational logs is central to many large-scale platforms. A prominent example motivating this work is the budget decision in online shopping platforms. The system repeatedly adjusts spending under evolving traffic, policy, and demand conditions. Similar challenges arise in advertising, pricing, and other resource allocation systems, where actions are continuous and decisions must be made from logged time-series data rather than repeated randomized experimentation [Hua et al., 2021, Liu et al., 2026b]. In such settings, the goal is not merely to predict realized outcomes, but to learn how outcomes would change under counterfactual action levels, so that future actions can be optimized [Athey and Wager, 2021].

This setting raises two distinct challenges that are often conflated. The first is confounding: logged actions are chosen adaptively from histories that also affect outcomes, a problem extensively studied in continuous treatment effect estimation [Kennedy et al., 2017, Nie et al., 2021, Hirano and Imbens,

*Equal contribution.
†Work done during an internship at Taobao, Alibaba Group.
‡Corresponding authors: Yong Gao (gaoyong@pku.edu.cn) and Jian Xu (xiyu.xj@taobao.com).

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

2004] and counterfactual sequence modeling [Lim, 2018, Bica et al., 2020b, Frauen et al., 2025a, Melnychuk et al., 2022]. The second, and the focus of this paper, is temporal non-stationarity. Demand, traffic, policies, and market conditions change over time, causing outcome levels to drift even when the causal effect of an action remains structurally similar. As a result, a pooled predictive model can fit historical outcomes well while absorbing time-varying baselines into the learned action surface. This creates a target mismatch: predictive accuracy on logged data need not imply correct action rankings for future deployment [Crasson et al., 2024, Schultz et al., 2023]. Existing invariant and robust prediction frameworks [Peters et al., 2016, Arjovsky et al., 2019, Rothenhäusler et al., 2021, Krueger et al., 2021] provide important tools for cross-environment stability, but they do not directly define a causal response target for confounded continuous actions under temporal drift.

This distinction leads to two separate questions. The first is identification within an observed period: under standard one-step exchangeability and overlap conditions, the period-specific response is identifiable. The second is transfer across periods: which part of these responses can be used to rank actions in future environments? We show that pooling observations across time can mix action-sensitive variation with drifting baseline structure and can even reverse the action ranking relevant for deployment. More fundamentally, past periods alone do not identify an unrestricted future response law. Thus the main obstacle is not only estimation under distribution shift, but the absence of a well-defined deployable target.

We address this by introducing a restricted target, which we call deployable causal action geometry. Rather than fitting raw outcome levels, we compare actions through anchored contrasts on a family of monotone link scales. Anchoring removes baseline shifts that do not depend on the action. The link scale and a profiled calibration factor isolate the shared action shape that determines intervention rankings. Identity and log links play a canonical role: they remove additive and multiplicative calibration. The resulting target allows period-specific levels and amplitudes to vary while requiring only the action-response shape relevant for decision making to persist. We estimate this target with an honest two-stage procedure. The estimator follows the orthogonal or debiased learning principle [Chernozhukov et al., 2018, Foster and Syrgkanis, 2023]. The first stage constructs cross-fitted orthogonal pilots for anchored responses under standard confounding assumptions. The second stage performs profiled selection over the link function, representation, and shape class, yielding a decision rule over continuous actions. We prove an oracle inequality over a finite candidate library and show that, under approximate persistence of the learned geometry, the induced rule has controlled regret in future periods.

Our contributions are threefold. First, we formalize the target mismatch induced by pooled regression under temporal drift and show that unrestricted future response learning is generally ill posed. Second, we define a deployable causal action geometry based on anchored link-scale contrasts and develop an honest orthogonal estimator with an oracle inequality over a finite candidate library. Third, we characterize deployability under approximate geometry persistence, separating the structural cost of persistence violations from the statistical error of learning the geometry. We validate these claims in synthetic experiments that isolate link selection, nuisance robustness, overlap stress, and persistence violations. We further evaluate the method on the Dominick's dataset and in a large-scale Taobao pricing deployment, including randomized online A/B testing.

2 Problem Setup

2.1 One-Step Response under Temporal Drift

Observed longitudinal data. We consider continuous one-step decision making from temporally non-stationary observational logs. Suppressing the unit index, let H_t denote the pre-treatment history available before period t, $A_t \in \mathcal{A} \subseteq \mathbb{R}$ the continuous action, and Y_{t+1} the next-step outcome. In the pricing application, H_t contains item attributes and recent sales and price histories, A_t is the normalized price intervention, and Y_{t+1} is next-step demand.

Temporal non-stationarity and estimand scope Temporal non-stationarity means that the joint law of (H_t, A_t, Y_{t+1}) changes with t. We decompose this variation into three sources of drift: $p_t(h, a, y) = p_t(y \mid a, h)\, p_t(a \mid h)\, p_t(h)$, corresponding to changes in the outcome mechanism $p_t(y \mid a, h)$, the logging policy $p_t(a \mid h)$, and the history distribution $p_t(h)$. The target mismatch below is driven by the interaction of these components: different action values may be observed

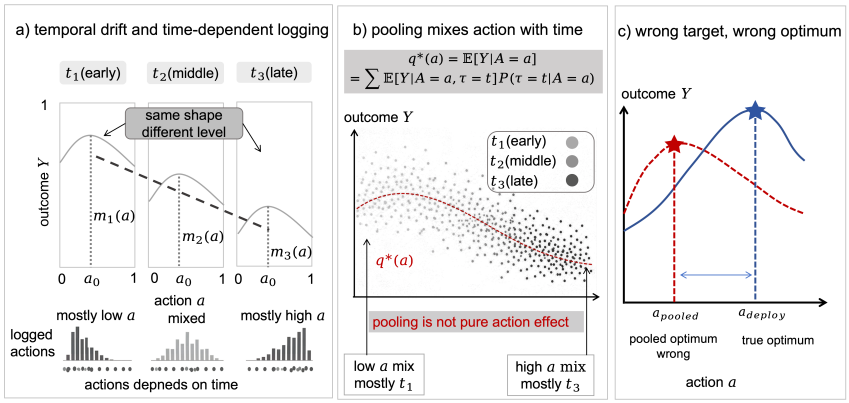


Figure 1: **Target mismatch under drift.** Pooled data under non-stationary environments compares confounded temporal mixtures, leading to action ranking reversal despite good historical fit.

under different mixtures of temporal environments, so $Y \mid A = a$ does not necessarily represent a pure intervention response. Our target is a proximal one-step response, not a long-horizon dynamic treatment effect. The question is therefore not only within-period identification, but whether the induced action ranking remains meaningful under temporal drift.

Within-Time Causal Description For each $a \in \mathcal{A}$, let $Y_{t+1}(a)$ denote the one-step potential outcome, while leaving the realized pre-treatment history H_t. The period-specific causal response is $m_t(h, a) := \mathbb{E}_t[Y_{t+1}(a) \mid H_t = h]$.

Assumption 2.1 (One-Step Identification)**.** *For each t, there exists $\mathcal{T}_t \subseteq \mathcal{H} \times \mathcal{A}$ such that $\forall (h, a) \in \mathcal{T}_t$: (i)* ***Consistency:*** *$A_{i,t} = a \implies Y_{i,t+1} = Y_{i,t+1}(a)$ a.s. (ii)* ***No Interference:*** *$Y_{i,t+1}(a_i, \mathbf{a}_{-i}) = Y_{i,t+1}(a_i)$ $\forall \mathbf{a}_{-i}$. (iii)* ***Overlap:*** *$\exists$ open $\mathcal{N}(a) \ni a$ s.t. $f_{A_t|H_t}(a' \mid h) > 0$ $\forall a' \in \mathcal{N}(a)$. (iv)* ***One-Step Exchangeability:*** *$Y_{i,t+1}(a) \perp A_{i,t} \mid H_{i,t}$.*

Proposition 2.2 (Within-time Identification)**.** *Under Assumption 2.1, $\forall t$ and a.e. $(h, a) \in \mathcal{T}_t$, $m_t(h, a) = \mathbb{E}_t[Y_{i,t+1} \mid H_{i,t} = h, A_{i,t} = a]$*

The proof is standard and deferred to Appendix C.1. Thus, temporal drift does not by itself preclude within-period identification. The remaining issue is deployment: a response learned from pooled historical periods may target the wrong object.

2.2 Target mismatch of pooled regression

To formalize pooled learning, let τ denote the period label of a randomly sampled training observation, with $\mathbb{P}(\tau = t) = \omega_t$. Conditional on $\tau = t$, the observation (H, A, Y) follows the period-t law. We write $\mathbb{E}_{\text{pool}}$ for expectation under this mixture. A standard pooled predictive learner ignores τ and targets $q^\star(x, a) := \mathbb{E}_{\text{pool}}[Y \mid X = x, A = a]$, where $X = \phi(H)$. This is a valid factual prediction target under the pooled training distribution, but it need not be a deployable response surface. The mismatch arises because conditioning on $A = a$ can change the mixture over periods and histories; hence different action values may be evaluated under different temporal environments.

To expose the mismatch, consider an additive response decomposition $m_t(h, a) = b_t(h) + r(h, a)$, where $b_t(h)$ is an action-independent drifting baseline and $r(h, a)$ is the action-sensitive component. Define $s(x, a) := \mathbb{E}_{\text{pool}}[r(H, a) \mid X = x]$, the action-sensitive response averaged within representation cell x. This is the action ranking that pooled prediction would need to preserve for deployment.

Proposition 2.3 (Pooling can reverse the deployable action ranking)**.** *Under Assumption 2.1 and the decomposition above, the pooled predictive target admits $q^\star(x, a) = s(x, a) + \delta(x, a)$, where $\delta(x, a)$ is a drift-induced distortion caused by conditioning on the logged action. If $\delta(x, \cdot)$ varies across actions, the ranking induced by $q^\star(x, \cdot)$ can differ from that induced by $s(x, \cdot)$. In particular, there exist data-generating processes such that*

$$s(x, a) > s(x, a') \qquad \textit{but} \qquad q^\star(x, a) < q^\star(x, a').$$

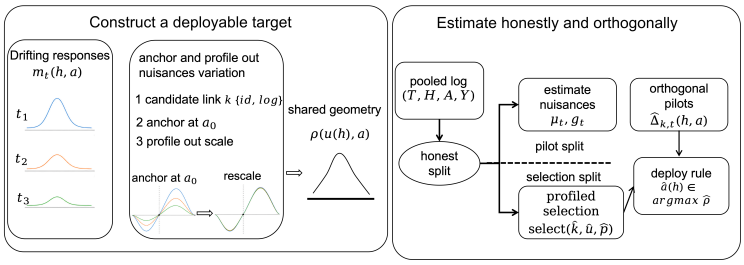


Figure 2: From drifting responses to deployable action geometry.

Appendix C.2 gives the full expression for δ, the proof, and a concrete ranking-reversal construction. Thus, within-period identification alone does not justify optimizing a pooled predictive surface: pooled regression may fit historical associations while selecting the wrong action for deployment.

3 Deployable Action Geometry

Section 2 shows that pooled prediction can learn the wrong action ranking under temporal drift. A natural candidate is an invariant response law. Such a law is both too strong and unnecessary for deployment: outcome levels may drift with demand, traffic, or policy conditions even when the relative ordering of actions is stable, while a one-step decision rule only requires the geometry of action contrasts, not the fully calibrated outcome surface.

We therefore target *deployable causal action geometry*: the action-response shape that remains after removing action-independent offsets and positive amplitude variation on a suitable monotone link scale. Concretely, we anchor the response at a reference action to remove the linked-scale offset and profile out a positive scale factor, leaving the shared shape that determines action rankings. Thus, outcome levels and response amplitudes may vary across periods while the decision-relevant geometry remains transportable. Without a persistence assumption, observed periods cannot identify an unrestricted future response law; Appendix D.1 formalizes this point. We therefore treat this geometry as a restricted, decision-sufficient target for deployment under approximate persistence. See Appendix B for a terminology map.

Let $\mathcal{A}_\star \subseteq \mathcal{A}$ be a compact common-support action region and $a_0 \in \mathcal{A}_\star$ a reference action. Assume $a \mapsto m_t(h,a)$ is continuous on $\mathcal{A}_\star$ for every observed period t and almost every h.

3.1 Anchored link scales

For a strictly increasing link ℓ_k, define the anchored linked response $\Delta_{k,t}(h,a) := \ell_k(m_t(h,a)) - \ell_k(m_t(h,a_0))$. Then $\Delta_{k,t}(h,a_0) = 0$, and anchoring removes an action-independent calibration term while preserving action rankings on the link scale.

We use a finite family of monotone links $\{\ell_k\}_{k\in\mathcal{K}}$. Unless stated otherwise, $\mathcal{K} = \{0,1\}$, with $\ell_0(y) = y$ and $\ell_1(y) = \log y$, where the log link is used only when $m_t(h,a) > 0$ on $\mathcal{A}_\star$. The identity link removes additive calibration in anchored contrasts, while the log link removes multiplicative calibration. The following result formalizes their canonical role.

Proposition 3.1 (Canonical exact nuisance-eliminating links)**.** *Let $I \subseteq \mathbb{R}$ be an interval and let $\ell : I \to \mathbb{R}$ be continuously differentiable and strictly increasing. (i) If $\ell(y+b) - \ell(y_0+b) = \ell(y) - \ell(y_0)$ for all admissible y, y_0, b, then $\ell(y) = \alpha y + \beta$ on I, with $\alpha > 0$. (ii) If $I \subseteq (0,\infty)$ and $\ell(cy) - \ell(cy_0) = \ell(y) - \ell(y_0)$ for all admissible $y, y_0, c > 0$, then $\ell(y) = \alpha \log y + \beta$ on I, with $\alpha > 0$.*

Thus identity and log are not merely preprocessing choices: if anchored contrasts must exactly remove additive or multiplicative calibration, they are canonical up to affine rescaling. Different links define different equivalence classes of response surfaces modulo time-varying calibration.

3.2 Shared action geometry

The deployable object is not a period-specific anchored curve, but the part of that curve that persists after calibration. We seek a link k, representation u, and normalized shape ρ such that

$$\Delta_{k,t}(h,a) \approx s_t(h)\rho(u(h),a), \qquad s_t(h) \geq 0.$$

Where $s_t(h)$ absorbs period-specific calibration, while $\rho(u(h),a)$ carries the action geometry for deployment.

Let $\mathfrak{U}$ be a candidate class of representations $u : \mathcal{H} \to \mathbb{R}^{d_u}$. For each k, u, let $\mathfrak{R}_k(u)$ be a normalized class of shapes $\rho : u(\mathcal{H}) \times \mathcal{A}_\star \to \mathbb{R}$ with $\rho(z, a_0) = 0$; the scale/sign convention is in Appendix D.3.1. A candidate (k, u, ρ) induces the action geometry

$$G_{k,u,\rho}(h) := \arg\max_{a \in \mathcal{A}_\star} \rho(u(h),a).$$

When the maximizer is not unique, fix an arbitrary measurable selector. The latent tuple need not be unique; the deployment target is the induced action correspondence, or its decision-equivalence class (Appendix D.4.3).

3.3 Oracle target

Let ϖ be a reference measure on $\mathcal{A}_\star$ and write $\|f\|_\varpi^2 := \int_{\mathcal{A}_\star} f(a)^2\, \varpi(da)$. For a candidate (k, u, ρ), define the profiled geometry loss

$$\mathcal{A}_k(u,\rho) := \sum_{t=1}^{T} \pi_t\, \mathbb{E}_t\left[\inf_{s \geq 0} \frac{\|\Delta_{k,t}(H_t,\cdot) - s\rho(u(H_t),\cdot)\|_\varpi^2}{\|\Delta_{k,t}(H_t,\cdot)\|_\varpi^2}\right],$$

with zero contribution when the denominator is zero. The infimum profiles out period-specific calibration, while the normalization removes amplitude, so candidates are compared by geometry rather than outcome level or link-scale magnitude. Appendix D.3.2 gives the closed form of the profiled calibration, and Appendix D.2.1 explains why fixing the anchored scale is necessary.

With regularization, $\mathcal{L}_k(u,\rho) := \mathcal{A}_k(u,\rho) + \lambda_u \Omega_u(u) + \lambda_\rho \Omega_{k,\rho}(\rho)$. Any minimizer $(k^\star, u^\star, \rho^\star) \in \arg\min_{k \in \mathcal{K},\, u \in \mathfrak{U},\, \rho \in \mathfrak{R}_k(u)} \mathcal{L}_k(u,\rho)$ defines an oracle target within the admissible candidate family, with induced oracle geometry

$$G^\star(h) := \arg\max_{a \in \mathcal{A}_\star} \rho^\star(u^\star(h),a).$$

Under exact calibrated-shape specification, $\mathcal{A}_k(u,\rho) = 0$ for a correct candidate; under misspecification, the target is the best regularized approximation within the admissible family. The object of interest is $G^\star$, not a unique representation or a universal raw response surface.

3.4 Deployability under approximate geometry persistence

The oracle geometry is defined from observed periods. For deployment, we do not assume that the full future response surface is identified; we assume only that the anchored geometry persists approximately after calibration.

Assumption 3.2 (Approximate geometry persistence)**.** *For deployment periods $\mathcal{T}_{\mathrm{dep}} \supseteq \{1,\dots,T\}$, suppose that for every $t \in \mathcal{T}_{\mathrm{dep}}$, almost every h, and every $a \in \mathcal{A}_\star$,*

$$\Delta_{k^\star,t}(h,a) = s_t(h)\rho^\star(u^\star(h),a) + \xi_t(h,a), \qquad s_t(h) > 0.$$

The residual ξ_t measures violation of shared-geometry persistence. Small violation is sufficient for near-optimal action selection.

Theorem 3.3 (Near-optimality under geometry persistence)**.** *Suppose Assumption 3.2 holds. Let $a^\rho(h) \in G^\star(h)$ and $a_t^{\mathrm{opt}}(h) \in \arg\max_{a \in \mathcal{A}_\star} m_t(h,a)$. Then for every $t \in \mathcal{T}_{\mathrm{dep}}$ and almost every h,*

$$0 \leq \Delta_{k^\star,t}\big(h, a_t^{\mathrm{opt}}(h)\big) - \Delta_{k^\star,t}(h, a^\rho(h)) \leq 2 \sup_{a \in \mathcal{A}_\star} |\xi_t(h,a)|.$$

If $\xi_t \equiv 0$, every selector from $G^\star$ is optimal in every deployment period; otherwise, linked-response suboptimality is controlled by the persistence residual.

Because $\ell_{k^\star}$ is strictly increasing and anchoring subtracts an action-independent constant, the maximizers of $\Delta_{k^\star,t}(h,\cdot)$, $\ell_{k^\star}(m_t(h,\cdot))$, and $m_t(h,\cdot)$ coincide on the link-valid region. Thus the theorem connects the calibration-quotiented geometry to the original action decision. Section 4 develops an orthogonal procedure for estimating this oracle target.

4 Orthogonal Estimation of Deployable Action Geometry

Section 3 defines the population target through a profiled criterion over link, representation, and shape classes. We estimate this target with an honest two-stage procedure. Stage I constructs cross-fitted orthogonal pilots of the anchored linked responses. Stage II uses an independent selection sample to choose the link, representation, and action geometry by minimizing an empirical profiled loss. The resulting decision rule is induced by the selected geometry rather than by a globally stable raw response surface. The estimator below is the canonical low-dimensional construction analyzed theoretically; Section 6 describes the scalable realization used in the Taobao deployment.

Let $O = (\tau, H, A, Y)$ denote a pooled observation, where τ is the period label, $\mathbb{P}(\tau = t) = \pi_t > 0$, and $Y = Y_{\tau+1}$ conditional on τ. We split the pooled sample into a pilot sample $\mathcal{I}^{\mathrm{pil}}$ and an independent selection sample $\mathcal{I}^{\mathrm{sel}}$, preserving temporal proportions. This separation prevents the criterion used for selection from reusing the data that constructed its response pilots.

4.1 Estimator

Orthogonal anchored-response pilots. For each period t, let $\mu_{0,t}(h,a) = \mathbb{E}[Y \mid H = h, A = a, \tau = t]$ and $g_{0,t}(a \mid h) = f_{A|H,\tau=t}(a \mid h)$. For a candidate representation u, representation kernel L_λ, and action kernel K_b, define the localized orthogonal score

$$\begin{aligned}\phi_{t,\lambda,b}(O; u, h, a, \eta_t) = \frac{\mathbf{1}\{\tau = t\}}{\pi_t} L_\lambda(u(H) - u(h)) \\ \times \left[\frac{K_b(A-a)}{g_t(A \mid H)} \{Y - \mu_t(H,A)\} + \int_{\mathcal{A}_\star} K_b(s-a)\mu_t(H,s)\, ds \right].\end{aligned} \tag{1}$$

where $\eta_t = (\mu_t, g_t)$. The corresponding local mass score is $\psi_{t,\lambda}(O; u, h) = \pi_t^{-1}\mathbf{1}\{\tau = t\}L_\lambda(u(H) - u(h))$. The residualized term in (1) makes the pilot locally robust to first-order nuisance error, while the integral term targets the kernel-smoothed response. Identification and orthogonality are proved in Appendix E.2; regularity conditions are stated in Appendix E.1.

Using cross-fitted nuisance estimates on $\mathcal{I}^{\mathrm{pil}}$, let $\hat{\nu}_{t,\lambda,b}(u,h,a)$ and $\hat{\delta}_{t,\lambda}(u,h)$ be the empirical averages of $\phi_{t,\lambda,b}$ and $\psi_{t,\lambda}$, respectively. With a vanishing truncation threshold $\epsilon_n > 0$, define $\hat{m}_{t,\lambda,b}(u,h,a) = \hat{\nu}_{t,\lambda,b}(u,h,a)/\{\hat{\delta}_{t,\lambda}(u,h) \vee \epsilon_n\}$. For a link ℓ_k, the anchored pilot is $\hat{\Delta}_{k,t}(u,h,a) = \ell_k^{(\epsilon_n)}\{\hat{m}_{t,\lambda,b}(u,h,a)\} - \ell_k^{(\epsilon_n)}\{\hat{m}_{t,\lambda,b}(u,h,a_0)\}$, where $\ell_0^{(\epsilon_n)}(y) = y$ and $\ell_1^{(\epsilon_n)}(y) = \log(y \vee \epsilon_n)$. The truncation stabilizes the estimated log pilot; the population log link remains defined on the link-valid positive region.

Only finite-grid control is needed for the selector. For the Stage-II action grid $\{a^{(m)}\}_{m=1}^M \subset \mathcal{A}_\star$, Appendix E.1 proves that, under the stated conditions,

$$\sup_{k\in\mathcal{K}} \sup_{u\in\mathfrak{U}_n} \sup_{1\le t\le T} \max_{i\in\mathcal{I}^{\mathrm{sel}}} \max_{1\le m\le M} \left| \hat{\Delta}_{k,t}(u, H_i, a^{(m)}) - \Delta_{k,t}(H_i, a^{(m)}) \right| = O_p(\varepsilon_n^{\mathrm{pilot}}). \tag{2}$$

For a fixed evaluation point, the corresponding rate is $(n_t\lambda^{d_u}b)^{-1/2} + \bar{r}_n + \lambda^{\alpha_u} + b^{\alpha_a}$, where $\bar{r}_n$ is the second-order nuisance product error; the uniform sequence $\varepsilon_n^{\mathrm{pilot}}$ additionally absorbs the finite-library and finite-grid union step.

Honest profiled selection. Let $\mathfrak{U}_n = \{u_1, \ldots, u_{M_n}\}$ be a finite representation library. For each $k \in \mathcal{K}$ and $u \in \mathfrak{U}_n$, let $\mathfrak{R}_{k,n}(u)$ be a normalized shape class. Given quadrature points and weights $\{(a^{(m)}, w_m)\}_{m=1}^M$, write $\|f\|_{M,w}^2 = \sum_{m=1}^M w_m f(a^{(m)})^2$. For $i \in \mathcal{I}^{\mathrm{sel}}$, let $\hat{\Delta}_{k,i}^u(\cdot) = \hat{\Delta}_{k,\tau_i}(u, H_i, \cdot)$ and $\rho_i^u(\cdot) = \rho(u(H_i), \cdot)$. The empirical profiled geometry loss is

$$\hat{\mathcal{A}}_k^{\mathrm{sel}}(u,\rho) = \frac{1}{n_{\mathrm{sel}}} \sum_{i\in\mathcal{I}^{\mathrm{sel}}} \inf_{s\ge 0} \frac{\|\hat{\Delta}_{k,i}^u - s\rho_i^u\|_{M,w}^2}{\|\hat{\Delta}_{k,i}^u\|_{M,w}^2}, \tag{3}$$

with the zero-energy convention of Section 3.3. The closed-form profiler, denominator stability, and quadrature approximation are given in Appendix E.4. With penalties Ω_u and $\Omega_{k,\rho}$, define $\hat{\mathcal{L}}_k^{\mathrm{sel}}(u,\rho) = \hat{\mathcal{A}}_k^{\mathrm{sel}}(u,\rho) + \lambda_{u,n}\Omega_u(u) + \lambda_{\rho,n}\Omega_{k,\rho}(\rho)$, and select

$$(\hat{k},\hat{u},\hat{\rho}) \in \arg\min_{k\in\mathcal{K},\ u\in\mathfrak{U}_n,\ \rho\in\mathfrak{R}_{k,n}(u)} \hat{\mathcal{L}}_k^{\mathrm{sel}}(u,\rho).$$

The deployed rule is $\hat{a}(h) \in \arg\max_{a\in\mathcal{A}_\star} \hat{\rho}(\hat{u}(h),a)$. The normalization in (3) is essential: it compares candidates by action geometry rather than by the numerical amplitude of their anchored contrasts.

4.2 Finite-library guarantee

The main guarantee is criterion-level learnability within the admissible finite library. Let p_n be the largest finite-dimensional shape parameter dimension, let q_M denote quadrature error, and define $r_n = \varepsilon_n^{\mathrm{pilot}} + \sqrt{(\log|\mathfrak{U}_n| + p_n \log n)/n_{\mathrm{sel}}} + q_M$.

Theorem 4.1 (Finite-library oracle inequality)**.** *Assume the pilot bound* (2) *holds, together with the finite-library entropy, quadrature, energy-separation, and penalty conditions in Appendix E.5. Then*

$$\mathcal{L}_{\hat{k}}(\hat{u},\hat{\rho}) - \inf_{k\in\mathcal{K},\ u\in\mathfrak{U}_n,\ \rho\in\mathfrak{R}_{k,n}(u)} \mathcal{L}_k(u,\rho) = O_p(r_n). \tag{4}$$

Theorem 4.1 is deliberately criterion-level. It states that, within the chosen link, representation, and shape classes, the selected candidate approaches the best admissible population criterion at rate r_n. It does not by itself imply recovery of a unique latent representation or policy regret in a future period.

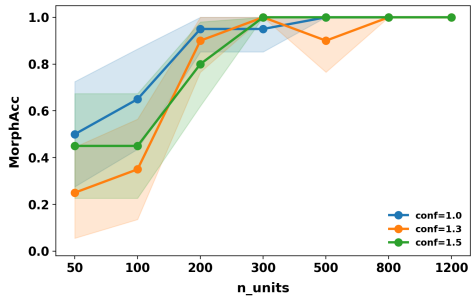
Two conditional consequences connect criterion learnability to link recovery and deployment. First, if $k^\star$ is the unique population link minimizer and the link selection gap $\gamma_n = \min_{k\neq k^\star}\{V_n(k) - V_n(k^\star)\}$, with $V_n(k) = \inf_{u,\rho}\mathcal{L}_k(u,\rho)$, satisfies $r_n = o(\gamma_n)$, then $\mathbb{P}(\hat{k} = k^\star) \to 1$. Second, under the criterion-to-geometry bridge in Appendix E.7, approximate geometry persistence in the deployment period, and a standard margin condition, the induced policy satisfies $\mathrm{Regret}_t(\hat{a}) = O_p(\Psi_n(r_n)^{\kappa+1}) + O(\mathrm{Pers}_t)$, where Pers_t measures violation of shared geometry in period t. Thus statistical error enters through the oracle-criterion rate r_n, while irreducible deployment error is exactly the cost of persistence failure.

5 Synthetic Experiments

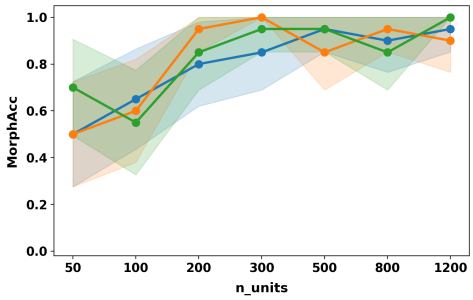
We use synthetic experiments to examine link recovery, future-period decision quality, and the contributions of link selection and orthogonal estimation. We consider four regimes. $R1$ has additive temporal shifts, for which the identity link is correct; $R2$ has multiplicative shifts, for which the log link is correct. $R3$ introduces approximate action-dependent drift, and $R4$ deliberately violates stable geometry through strong time-varying shape drift. Unless otherwise stated, we use $T = 20$ periods, a $12/4/4$ train/validation/test split, a 51-point action grid, and 30 Monte Carlo repetitions. Our primary metrics are future-period stable IMSE (SIMSE) and regret; factual RMSE is reported only as a diagnostic. Full DGP details, implementation, and additional stress tests are in Appendix G.

Main evidence. Figure 3(a,b) examines link recovery in the two exact regimes, with recovery generally improving as sample size increases. Representation recovery is harder, but the deployable object is the induced action geometry rather than the latent representation itself. Figure 3(c,d) and Table 1 show lower future-period SIMSE and regret than pooled prediction and decision performance close to the regime-matched fixed-link reference. The close regret values indicate low adaptation cost when the generating link must be selected. After fixing the selected link-representation pair, the orthogonal refit yields lower regret and action error than direct plug-in in the reported nuisance stress tests. For component attribution, an additional 2×2 ablation varies orthogonal versus controlled plug-in pilots, which omit only the residual correction, at selection and refitting. Under outcome-model misspecification, most regret reduction occurs at refitting, while selection-stage effects on link recovery are regime-dependent. Full ablations are reported in Appendix G.4.

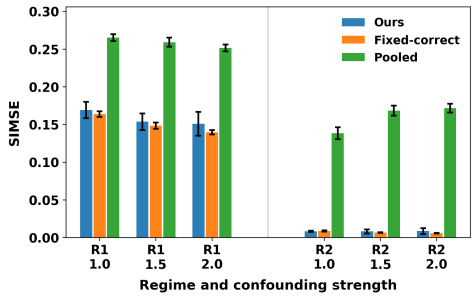
When link selection matters. We additionally shift the history distribution while preserving exact action geometry, so that a wrong fixed link induces curvature outside the shared shape class. Ours



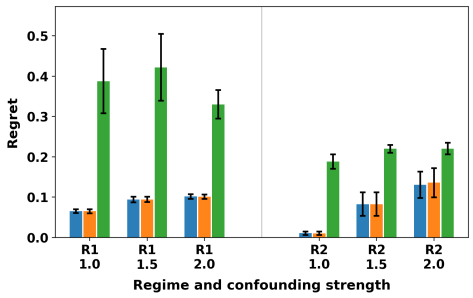
(a) Link recovery in R1.



(b) link recovery in R2.



(c) Future-time SIMSE.



(d) Future-time regret.

Figure 3: **Synthetic evidence.** Top: link recovery in the exact regimes. Bottom: future-period SIMSE and regret.

Table 1: **Future-period deployment.** Each cell reports SIMSE / regret. Fixed-correct denotes the regime-matched fixed link: raw in $R1$ and log in $R2$.

Regime	Conf.	Ours	Fixed-correct	Pooled
$R1$	1.0	0.169±0.011 / 0.065±0.005	0.164±0.004 / 0.065±0.005	0.265±0.005 / 0.388±0.080
	1.5	0.154±0.011 / 0.094±0.007	0.149±0.004 / 0.094±0.007	0.259±0.006 / 0.422±0.083
	2.0	0.151±0.016 / 0.102±0.006	0.140±0.003 / 0.101±0.006	0.252±0.005 / 0.330±0.035
$R2$	1.0	0.008±0.001 / 0.011±0.004	0.009±0.001 / 0.010±0.004	0.139±0.008 / 0.189±0.018
	1.5	0.008±0.003 / 0.083±0.029	0.007±0.001 / 0.083±0.029	0.169±0.007 / 0.220±0.009
	2.0	0.009±0.004 / 0.131±0.033	0.006±0.000 / 0.136±0.036	0.172±0.006 / 0.220±0.014

attains NRegret 0.1385, close to the reference given the generating link (0.1384), and below plug-in selection (0.1412), fixed identity (0.1416), and fixed log (0.1422). This isolates the benefit of selecting the response scale under transport shift;

Boundary and sensitivity diagnostics. Appendix G further probes the assumptions behind the target. In $R3$, increasing action-dependent drift weakens link recovery; an additional controlled sweep shows increasing oracle-geometry regret as the persistence residual grows. In $R4$, no shared geometry exists, so the goal is failure-boundary characterization rather than uniform superiority. We also report overlap stress tests and sensitivity to the action grid and reference action.

6 Real-world Pricing Dataset and Application

We evaluate our framework on a large-scale observational pricing dataset from Taobao, one of China's largest e-commerce platforms. Historical prices reflect adaptive merchant behavior under non-stationary demand driven by traffic, seasonality, and campaigns. We construct an item-day panel where the action is the price normalized near 1. Offline evaluation is conducted via a rolling-origin chronological split on three months of data, containing $\sim 546.3M$ observations across $8.3M$ unique items (averaging $\sim 6.0M$ per day). Full construction details are provided in Appendix H.

Table 2: Offline evaluation.

Method	RMAE ↓	MAPE ↓	Stability ↓	DR Policy Value Index ↑
Logged / behavior	–	–	–	100.00
NPM	0.176	18.27	–	121.72
VCNet	0.385	48.85	0.452	116.09
Causal Transformer	0.174	17.35	0.566	107.18
Pooled predictive	0.337	44.69	0.503	89.43
Time-aware pooled	0.496	57.91	0.433	72.96
Fixed identity	0.223	26.85	0.323	117.41
Fixed log	0.138	16.33	0.282	122.36
Ours	0.103	12.75	0.259	128.32

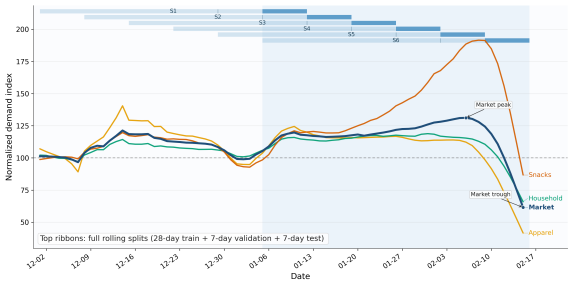
Production implementation. At Taobao scale, we deploy a scalable target aligned neural realization that preserves the same anchored identity/log estimand and honest two-stage separation. A fixed 320-dimensional multimodal encoder maps histories to representations. Separate baseline and treatment nuisance models are trained with item-disjoint cross-fitting, and all out-of-fold predictions are frozen before response fitting. The action-response shape is modeled by an anchored three-layer MLP, trained with an R-loss-style residualized objective. Because the encoder is fixed before link selection, the representation library is a singleton; at each weekly retraining window, identity- and log-link candidates are compared using the profiled criterion on an independent chronological validation period, and the selected checkpoint is exactly the one used for deployment.

6.1 Offline Evaluation

Metrics and baselines We report three primary offline metrics: relative factual MAE $RMAE$ and $MAPE$ on observed outcomes; policy value V_{DR} induced by models evaluated with a localized DR estimator for continuous actions [Kallus and Zhou, 2018]; $\mathrm{StabLoss}$ measures future time transportability. We use baseline like Neural Parametric Model (NPM) which is a strong baseline in the business application, VcNet, Causal Transformer, Pooled MLP, and Time-aware pooled MLP. **Main offline results.** Table 2 reports factual accuracy, future-profile stability, and DR-based offline policy value. The baselines exhibit a clear trade-off: Causal Transformer gives the best factual fit, fixed identity gives the best stability, and NPM gives the strongest policy value. Our method dominates this trade-off, achieving the best accuracy, the lowest stability loss, and the highest DR Policy Value Index. These results support our claim that future-time pricing decisions require stable response learning rather than same-period prediction alone. **Link-scale diagnosis.** Figure 4(b) compares a split level future-profile inconsistency under the raw-scale and log-linked anchored contrasts. The log-linked contrast attains a lower score in four of the six rolling splits and also yields a lower mean score across splits (0.5566 vs. 0.5718). The evidence is strongest in Splits 3 and 5, while Splits 1 and 6 favor the raw-scale contrast. We therefore interpret the evidence as time-varying preference between the two response scales. As intended, this diagnosis is a deployment-oriented structural check based on future-period profile consistency rather than a claim of ground-truth link recovery. **Temporal drift and deployment distance.** We visualize the non-stationary demand environment underlying the real-world pricing task in Figure 4(a). The figure plots normalized market-level and category-level demand indices over time, together with the rolling evaluation windows used in our chronological protocol. The demand trajectories vary substantially across periods, confirming that the deployment environment is far from stationary. This motivates evaluating pricing policies chronologically rather than with random splits.

6.2 Online A/B Result

We finally test whether the offline gains translate into real deployment. We run an online A/B experiment on $\sim$ 280k items over 26 days (2026/03/24–2026/04/18). As shown in Figure 4(c), the treatment policy yields consistent positive lifts on all three core business metrics: GMV (+1.16%, $p = 0.037$), Buy counts (+1.70%, $p = 0.004$), and Commission (+1.51%, $p < 0.001$). These results provide direct online evidence that the learned price-response structure improves decision quality beyond retrospective offline evaluation.



(a) Temporal drift and rolling windows.

(b) Link diagnosis.

Split	Raw	Log	Pref.	Gap
S1	0.6657	0.7311	Raw	8.95
S2	0.5946	0.5848	Log	1.64
S3	0.5183	0.4231	Log	18.36
S4	0.4704	0.4551	Log	3.26
S5	0.5087	0.4297	Log	15.54
S6	0.6733	0.7161	Raw	5.97
Mean	0.5718	0.5566	Log	2.66

(c) Online A/B result.

Metric	Lift (%)	p
GMV	+1.16	0.037
Buy counts	+1.70	0.004
Commission	+1.51	< 0.001

Figure 4: **Real-world diagnostics and deployment.** (a) temporal drift. (b) future profile favors the log link while revealing heterogeneity. (c) Online A/B testing shows positive lift.

6.3 Public Dominick's Evaluation

To complement the proprietary Taobao study, we evaluate our method on the public Dominick's retail dataset [Montgomery, 1997] using chronological splits. We assess factual prediction on the observed panel and counterfactual decision quality in semi-synthetic experiments that retain the real temporal and treatment-assignment design. These experiments cover additive and multiplicative persistence and controlled action-dependent shape drift. Under exact persistence, decision regret is close to that of a reference supplied with the generating link, indicating low adaptation cost. Appendix H.3 provides data preparation, evaluation protocols, outcome generators, implementation details, and full baseline comparisons.

7 Related Work

Correcting confounding and identifying what can transfer are distinct problems. Continuous-treatment estimators and longitudinal counterfactual models address response estimation from observational histories [Kennedy et al., 2017, Nie et al., 2021, Melnychuk et al., 2022], while orthogonal learning reduces sensitivity to nuisance estimation [Chernozhukov et al., 2018, Foster and Syrgkanis, 2023]. Neither causal identification nor nuisance robustness alone specifies which part of a historical response remains useful for future decisions. Causal forecasting and invariant prediction make cross-environment transfer explicit [Schultz et al., 2023, Crasson et al., 2024, Peters et al., 2016, Arjovsky et al., 2019, Liu et al., 2026a]. Our distinction is the object being transported: an anchored conditional action-response shape, rather than its period-specific level or amplitude. Under exact geometry persistence, environments may disagree about outcome magnitudes while agreeing on the preferred actions; approximate persistence quantifies departures from this structure. The proposed target therefore complements confounding adjustment by making the required transfer assumption explicit. Appendix A gives detailed comparisons.

8 Conclusion and Limitations

We studied continuous treatment-response learning under temporal non-stationarity, where pooled prediction can conflate action effects with drifting baselines and target an object that is unsuitable for deployment. We introduced deployable causal action geometry, a decision target that captures the action-response structure preserved after reference-action anchoring and link-scale recalibration. We developed an honest two-stage orthogonal estimator with profiled link and representation selection. Theory and experiments on synthetic data, the Dominick's dataset, and a large-scale pricing application support a central message: under temporal drift, decision learning should prioritize stable action geometry rather than raw prediction. Its validity relies on one-step exchangeability, overlap, a link family, and approximate geometry persistence; future work could relax the structure assumptions and extend the framework to long-horizon sequential decision making.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 42571530) and by Alibaba Group through Alibaba Research Intern Program. We thank the anonymous reviewers and area chairs for their constructive feedback.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. doi: 10.48550/arXiv.1907.02893.

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1):133–161, 2021. doi: 10.3982/ECTA15732.

Ioana Bica, Ahmed Alaa, and Mihaela Van Der Schaar. Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders. In *International conference on machine learning*, volume 119, pages 884–895. PMLR, 2020a.

Ioana Bica, Ahmed M. Alaa, James Jordon, and Mihaela van der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *International Conference on Learning Representations*, 2020b.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10.1111/ectj.12097.

Thomas Crasson, Yacine Nabet, and Mathias Lécuyer. Training and evaluating causal forecasting models for time-series. *arXiv preprint arXiv:2411.00126*, 2024. doi: 10.48550/arXiv.2411.00126.

Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *The Annals of Statistics*, 51(3):879–908, 2023.

Dennis Frauen, Konstantin Hess, and Stefan Feuerriegel. Model-agnostic meta-learners for estimating heterogeneous treatment effects over time. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Mihaela van der Schaar, and Stefan Feuerriegel. Treatment effect estimation for optimal decision-making. *arXiv preprint arXiv:2505.13092*, 2025b.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Keisuke Hirano and Guido W. Imbens. The propensity score with continuous treatments. In *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, Wiley Ser. Probab. Stat., pages 73–84. Wiley, Chichester, 2004. doi: 10.1002/0470090456.ch7.

Junhao Hua, Ling Yan, Huan Xu, and Cheng Yang. Markdowns in e-commerce fresh retail: A counterfactual prediction and multi-period optimization approach. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 3022–3031, 2021. doi: 10.1145/3447548.3467083.

Kosuke Imai and David A Van Dyk. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99(467):854–866, 2004. doi: 10.1198/016214504000001187.

Nathan Kallus and Masatoshi Uehara. Doubly robust off-policy value and gradient estimation for deterministic policies. In *Advances in Neural Information Processing Systems*, volume 33, pages 10420–10430, 2020.

Nathan Kallus and Angela Zhou. Policy evaluation and optimization with continuous treatments. In *International conference on artificial intelligence and statistics*, volume 84, pages 1243–1251. PMLR, 2018.

Edward H Kennedy, Zongming Ma, Matthew D McHugh, and Dylan S Small. Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(4):1229–1245, 2017. doi: 10.1111/rssb.12212.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, pages 5815–5826. PMLR, 2021.

Bryan Lim. Forecasting treatment responses over time using recurrent marginal structural networks. *Advances in neural information processing systems*, 31, 2018.

Changjian Liu, Yong Gao, Yuqing Wang, Leyi Su, Honglei Guo, Zhiyang Wang, Xiaoyu Wang, and Fan Zhang. Opflow: Learning opportunity-conditioned choice potentials for robust od flow prediction, 2026a. URL `https://arxiv.org/abs/2607.03200`.

Changjian Liu, Tianyu Wang, Xiaoxuan Deng, WenTao Zhu, Yuwei Xu, Jungqi Jin, Yong Gao, Chuan Yu, Jian Xu, and Bo Zheng. Multi-channel uplift policy learning, 2026b. URL `https://arxiv.org/abs/2607.28182`.

Sara Magliacane, Thijs Van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. *Advances in neural information processing systems*, 31, 2018.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 15293–15329. PMLR, 2022.

Alan L Montgomery. Creating micro-marketing pricing strategies using supermarket scanner data. *Marketing science*, 16(4):315–337, 1997.

Lizhen Nie, Mao Ye, Qiang Liu, and Dan Nicolae. Vcnet and functional targeted regularization for learning causal effects of continuous treatments. In *International Conference on Learning Representations*, 2021.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016. doi: 10.1111/rssb.12167.

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000.

Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(2): 215–246, 2021. doi: 10.1111/rssb.12398.

Douglas Schultz, Johannes Stephan, Julian Sieber, Trudie Yeh, Manuel Kunz, Patrick Doupe, and Tim Januschowski. Causal forecasting for pricing. *arXiv preprint arXiv:2312.15282*, 2023. doi: 10.48550/arXiv.2312.15282.

Adarsh Subbaswamy and Suchi Saria. From development to deployment: dataset shift, causality, and shift-stable models in health ai. *Biostatistics*, 21(2):345–352, 2020. doi: 10.1093/biostatistics/kxz041.

Shuxi Zeng, Murat Ali Bayir, Joseph J Pfeiffer III, Denis Charles, and Emre Kiciman. Causal transfer random forest: Combining logged data and randomized experiments for robust prediction. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 211–219, 2021. doi: 10.1145/3437963.3441722.

This appendix collects related work, theoretical proofs and supplementary results for Sections 2–4, the implementation algorithm, and additional synthetic-experiment details.

A Related Work

Continuous treatments, policy learning, and orthogonal estimation. Continuous-action causal learning builds on generalized propensity scores, dose-response estimation, and policy learning from observational data [Imai and Van Dyk, 2004, Hirano and Imbens, 2004, Kennedy et al., 2017, Kallus and Zhou, 2018, Athey and Wager, 2021]. Recent neural methods such as VCNet estimate smooth average dose-response curves for continuous treatments [Nie et al., 2021], while continuous-action off-policy evaluation develops kernelized or doubly robust estimators for policy value and gradients [Kallus and Zhou, 2018, Kallus and Uehara, 2020]. Our estimator also uses the orthogonal/debiased learning principle: cross-fitted nuisance estimates enter through scores with second-order sensitivity to nuisance error [Chernozhukov et al., 2018, Foster and Syrgkanis, 2023]. These works provide the estimation primitives for confounded continuous actions. Their target, however, is usually a static dose-response curve, a policy value, or a fixed causal parameter. They do not address the target-level question central to this paper: when outcome levels and logging policies drift over time, which part of a history-conditioned continuous response remains meaningful for future action ranking?

Longitudinal counterfactual modeling. Classical longitudinal causal inference studies time-varying treatments and confounding through marginal structural models and related weighting ideas [Robins et al., 2000]. Modern sequence models extend this perspective with recurrent or transformer architectures: RMSN forecasts responses under planned treatment sequences, CRN balances treatment histories for counterfactual prediction, the Time Series Deconfounder targets multi-cause hidden confounding, and Causal Transformer improves sequential counterfactual modeling with attention-based architectures [Lim, 2018, Bica et al., 2020b,a, Melnychuk et al., 2022]. These methods are highly relevant because they learn from observational time series. Their objective is nevertheless different: they aim to predict counterfactual outcomes over treatment paths or estimate heterogeneous effects over time. We instead study a one-step continuous decision problem under temporal non-stationarity, where the key issue is not trajectory prediction but the construction of an anchored action geometry that can be transported to future periods.

Invariant learning, causal transfer, and deployment shift. A separate literature asks what should remain stable across environments. Domain-adversarial training learns representations that obscure domain identity [Ganin et al., 2016]; invariant causal prediction, IRM, Anchor Regression, and REx seek stable predictors or risk profiles across environments and perturbations [Peters et al., 2016, Arjovsky et al., 2019, Rothenhäusler et al., 2021, Krueger et al., 2021]; causal domain adaptation identifies invariant conditional distributions under distribution shifts [Magliacane et al., 2018]; and deployment-oriented work formalizes dataset shift through a causal lens [Subbaswamy and Saria, 2020]. In industrial prediction, Causal Transfer Random Forest combines logged data with a small randomized experiment to learn predictors robust to feature shifts [Zeng et al., 2021]. These works are conceptually close because they share the goal of stable transfer. The difference is the object being stabilized. Prior work typically stabilizes a predictor, conditional distribution, or robust risk. Our target is a causal response geometry for confounded continuous actions: invariance is imposed only after anchoring, link transformation, and profiling out time-varying calibration.

Causal forecasting and decision-oriented learning. Recent work argues that time-series models used for decisions should be evaluated by intervention- or policy-relevant criteria rather than only predictive accuracy. Causal forecasting for pricing combines causal estimation with forecasting architectures for off-policy price-response learning [Schultz et al., 2023], and recent causal-forecasting protocols emphasize training and evaluation under shifted decisions [Crasson et al., 2024]. More broadly, decision-oriented treatment-effect learning shows that estimators optimized for effect accuracy may be suboptimal for downstream policy quality [Frauen et al., 2025b]. Our work shares this deployment-oriented motivation but focuses on an earlier target-definition problem: under temporal drift, the raw response surface itself may not be the right object. We therefore define a restricted action geometry that preserves the intervention rankings needed for deployment while allowing period-specific calibration to vary.

Table 3: Terminology and standard interpretations.

Term	Standard interpretation	Role in our framework
Link function ℓ_k	Monotone response scale	Defines the scale on which action-independent temporal variation can be removed. The main candidates are the identity and log links.
Anchored response $\Delta_{k,t}(h, a)$	Reference-action contrast	$\ell_k\{m_t(h, a)\} - \ell_k\{m_t(h, a_0)\}$. Anchoring removes an action-independent offset while preserving action rankings.
Positive amplitude $s_t(h)$	Nuisance scale factor	Captures period- and history-specific changes in the magnitude of the anchored response. It is profiled out when comparing candidate geometries.
Action geometry $\rho(u(h), a)$	Normalized conditional action-response shape	The decision-relevant shape remaining after anchoring and positive-scale profiling; it determines the action ranking through $\arg\max_a \rho(u(h), a)$.
Deployable causal action geometry	Decision-sufficient response equivalence class	An anchored action-response shape considered modulo linked-scale offsets and positive rescaling, restricted to the chosen link, representation, and shape classes.
Profiled geometry loss	Nuisance-amplitude-profiled discrepancy	Measures how well one shared action geometry explains period-specific anchored responses after optimally rescaling their amplitudes.
Approximate geometry persistence	Cross-period transport condition	Requires future anchored responses to remain close to a positive rescaling of the learned geometry; $\xi_t(h, a)$ measures the violation.
Decision equivalence	Equality of induced action rankings	Different latent representations or response parameterizations are equivalent when they induce the same maximizing action.

Summary. The closest prior ideas are continuous-treatment estimation, longitudinal confounding adjustment, orthogonal learning, and invariant prediction under shift. Our contribution is their combination at the level of the estimand: we define a deployable, restricted continuous-action response on anchored link scales, estimate it with orthogonal pilots and honest profiled selection, and require persistence only for the action-sensitive geometry rather than for the full raw outcome surface.

B Terminology

C Appendix for Section 2: Problem Setup

C.1 Proof of Within-Period Identification

Proof. Fix a period t and suppress the unit index. Let

$$\mu_t(h, a) := \mathbb{E}[Y_{t+1} \mid H_t = h, A_t = a, \text{ period } t]$$

be a version of the observed conditional mean. For any action a in the overlap region, consistency gives $Y_{t+1} = Y_{t+1}(a)$ on the event $\{A_t = a\}$. Therefore,

$$\begin{aligned}\mu_t(h, a) &= \mathbb{E}[Y_{t+1}(a) \mid H_t = h, A_t = a, \text{ period } t] \\ &= \mathbb{E}[Y_{t+1}(a) \mid H_t = h, \text{ period } t] \\ &= m_t(h, a),\end{aligned}$$

where the second equality uses one-step exchangeability and the last equality is the definition of m_t. Hence

$$m_t(h, a) = \mathbb{E}[Y_{t+1} \mid H_t = h, A_t = a, \text{ period } t]$$

for (H_t, A_t)-almost every (h, a) in the relevant support. □

C.2 Proof of Proposition 2.3

Proof. Let τ denote the period label under the pooled distribution. By consistency, one-step exchangeability, and iterated expectation,

$$\begin{aligned} q^\star(x,a) &= \mathbb{E}_{\text{pool}}[Y \mid X=x, A=a] \\ &= \mathbb{E}_{\text{pool}}[Y(a) \mid X=x, A=a] \\ &= \mathbb{E}_{\text{pool}}[\mathbb{E}[Y(a) \mid H, A, \tau] \mid X=x, A=a] \\ &= \mathbb{E}_{\text{pool}}[\mathbb{E}[Y(a) \mid H, \tau] \mid X=x, A=a] \\ &= \mathbb{E}_{\text{pool}}[m_\tau(H,a) \mid X=x, A=a]\,. \end{aligned}$$

Under the decomposition $m_t(h,a) = b_t(h) + r(h,a)$, this becomes

$$q^\star(x,a) = \mathbb{E}_{\text{pool}}[b_\tau(H) \mid X=x, A=a] + \mathbb{E}_{\text{pool}}[r(H,a) \mid X=x, A=a].$$

Adding and subtracting

$$s(x,a) := \mathbb{E}_{\text{pool}}[r(H,a) \mid X=x]$$

yields

$$q^\star(x,a) = s(x,a) + \delta(x,a),$$

where

$$\delta(x,a) := \mathbb{E}_{\text{pool}}[b_\tau(H) \mid X=x, A=a] + \mathbb{E}_{\text{pool}}[r(H,a) \mid X=x, A=a] - \mathbb{E}_{\text{pool}}[r(H,a) \mid X=x].$$

Therefore, for any a, a',

$$q^\star(x,a) - q^\star(x,a') = \{s(x,a) - s(x,a')\} + \{\delta(x,a) - \delta(x,a')\}.$$

Thus pooled prediction preserves all pairwise action rankings at x whenever $\delta(x,\cdot)$ is constant, and can change them when the distortion varies with action. The concrete construction below shows that reversal can occur. □

Corollary C.1 (Ranking reversal under drift-aligned logging)**.** *Let $\mathcal{A} = \{a_1, a_2\}$, and suppose $H \equiv h$ and $X \equiv x$. Consider two periods with equal pooled weights and drifting baselines $b_1(h) = 0$, $b_2(h) = M$. Suppose the action-sensitive component ranks a_1 above a_2:*

$$r(h,a_1) > r(h,a_2).$$

Let the logging policy be aligned with time:

$$\mathbb{P}(A=a_1 \mid \tau=1) = 1-\epsilon, \qquad \mathbb{P}(A=a_1 \mid \tau=2) = \epsilon, \qquad 0 < \epsilon < \frac{1}{2},$$

with $\mathbb{P}(A=a_2 \mid \tau=t) = 1 - \mathbb{P}(A=a_1 \mid \tau=t)$. Then

$$q^\star(x,a_1) - q^\star(x,a_2) = \{r(h,a_1) - r(h,a_2)\} - (1-2\epsilon)M.$$

Consequently, if

$$M > \frac{r(h,a_1) - r(h,a_2)}{1-2\epsilon},$$

the pooled target ranks a_2 above a_1, although the deployable action-sensitive component ranks a_1 above a_2.

Proof. Since the two periods have equal pooled weights, Bayes' rule gives

$$\mathbb{P}(\tau=2 \mid A=a_1) = \epsilon, \qquad \mathbb{P}(\tau=2 \mid A=a_2) = 1-\epsilon.$$

Because $H \equiv h$ and $X \equiv x$,

$$\begin{aligned} q^\star(x,a_1) &= r(h,a_1) + \mathbb{E}[b_\tau(h) \mid A=a_1] = r(h,a_1) + \epsilon M, \\ q^\star(x,a_2) &= r(h,a_2) + \mathbb{E}[b_\tau(h) \mid A=a_2] = r(h,a_2) + (1-\epsilon)M. \end{aligned}$$

Subtracting gives the stated expression, and the reversal condition follows immediately. □

D Additional Results for Section 3

This appendix provides technical details for the deployable action geometry defined in Section 3. Appendix D.1 shows why an unrestricted future response law is not identified from observed periods alone. Appendix D.2 justifies the anchored link family and the need to fix a contrast scale. Appendix D.3.1 and Appendix D.3.2 give the normalization convention and the closed form of the profiled calibration. Appendix D.4 records basic properties of the oracle criterion.

D.1 Unrestricted future deployment is not identified

Proposition D.1 (Observed periods do not identify the future optimum)**.** *Assume the one-step causal conditions hold on periods $t = 1, \dots, T$, so that each observed-period response m_t is identified on $\mathcal{A}_\star$. Then the observed law on periods $1, \dots, T$ does not identify the future response m_{T+1}, and hence does not identify the future-optimal action. More precisely, for any $h \in \mathcal{H}$ and any $a, a' \in \mathcal{A}_\star$, there exist two data-generating processes $\mathbb{P}^+$ and $\mathbb{P}^-$ that induce the same observed law on periods $1, \dots, T$ but satisfy*

$$m^+_{T+1}(h, a) > m^+_{T+1}(h, a'), \qquad m^-_{T+1}(h, a) < m^-_{T+1}(h, a').$$

Proof. Fix any observed-period process $\mathbb{P}$ satisfying the one-step causal assumptions on $t = 1, \dots, T$. We construct two extensions beyond the observed periods.

For $t = 1, \dots, T$, let $\mathbb{P}^+$ and $\mathbb{P}^-$ coincide with $\mathbb{P}$. Hence the observed data law is identical under the two processes. For period $T + 1$, set $H_{T+1} = h$ almost surely under both extensions, and choose any assignment kernel whose support contains $\mathcal{A}_\star$. Define future response functions $\widetilde{m}^+$ and $\widetilde{m}^-$ such that

$$\widetilde{m}^+(h, a) > \widetilde{m}^+(h, a'), \qquad \widetilde{m}^-(h, a) < \widetilde{m}^-(h, a').$$

For example, take $\widetilde{m}^+(h, a) = 1, \widetilde{m}^+(h, a') = 0$ and $\widetilde{m}^-(h, a) = 0, \widetilde{m}^-(h, a') = 1$, with arbitrary admissible values elsewhere.

Finally choose potential outcomes in period $T + 1$ so that

$$\mathbb{E}_{\mathbb{P}^\pm}[Y_{T+2}(b) \mid H_{T+1} = h] = \widetilde{m}^\pm(h, b), \qquad b \in \mathcal{A}_\star,$$

for instance by setting $Y_{T+2}(b) = \widetilde{m}^\pm(h, b) + \varepsilon_b$ with mean-zero noise. The two extensions agree on all observed periods but rank a and a' oppositely in period $T + 1$. Thus the future response and the future optimum are not identified from the observed-period law. □

D.2 Anchored scales and canonical links

D.2.1 Contrast-scale non-canonicity

Proposition D.2 (Anchored factorizations are not canonical without fixing scale)**.** *Suppose an anchored contrast admits*

$$\Delta_t(h, a) = s_t(h)\rho(h, a), \qquad s_t(h) \geq 0, \qquad \rho(h, a_0) = 0.$$

For any $p > 0$, define

$$\Delta^{[p]}_t(h, a) := \operatorname{sgn}\{\Delta_t(h, a)\} \, |\Delta_t(h, a)|^p.$$

Then

$$\Delta^{[p]}_t(h, a) = s_t(h)^p \rho^{[p]}(h, a), \qquad \rho^{[p]}(h, a) := \operatorname{sgn}\{\rho(h, a)\} \, |\rho(h, a)|^p,$$

and $\rho^{[p]}(h, a_0) = 0$.

Proof. By direct substitution,

$$\Delta^{[p]}_t(h, a) = \operatorname{sgn}\{s_t(h)\rho(h, a)\}|s_t(h)\rho(h, a)|^p = s_t(h)^p \operatorname{sgn}\{\rho(h, a)\}|\rho(h, a)|^p,$$

where the sign is unchanged because $s_t(h) \geq 0$. The anchoring condition follows from $\rho(h, a_0) = 0$. □

This construction shows that a stable factorization is not uniquely defined until the anchored contrast scale is fixed. The link family in the main text therefore specifies the admissible link family rather than merely a preprocessing convention.

D.2.2 Canonical nuisance-eliminating links

Proof of Proposition 3.1. We prove the two statements in turn.

For the additive case, fix y, y_0 in the interior of I and define

$$F(b) := \ell(y + b) - \ell(y_0 + b)$$

whenever the arguments lie in I. By assumption F is constant, so

$$0 = F'(b) = \ell'(y + b) - \ell'(y_0 + b).$$

Taking $b = 0$ gives $\ell'(y) = \ell'(y_0)$. Since y, y_0 are arbitrary interior points, ℓ' is constant on the interior of I. Hence $\ell(y) = \alpha y + \beta$ on I, with $\alpha > 0$ because ℓ is strictly increasing.

For the multiplicative case, let $J := \{\log y : y \in I\}$ and define $g(x) := \ell(e^x)$. Then g is continuously differentiable and strictly increasing. The multiplicative invariance condition gives, whenever the arguments are in J,

$$g(x + c) - g(x_0 + c) = \ell(e^{x+c}) - \ell(e^{x_0+c}) = \ell(e^x) - \ell(e^{x_0}) = g(x) - g(x_0).$$

Thus g satisfies the additive invariance condition on J. By the first part, $g(x) = \alpha x + \beta$ with $\alpha > 0$, and therefore

$$\ell(y) = g(\log y) = \alpha \log y + \beta.$$

□

D.3 Normalization and profiled calibration

D.3.1 Normalization convention

The factorization

$$\Delta_{k,t}(h, a) = s_{k,t}(h)\rho(u(h), a)$$

is unchanged under the rescaling

$$\rho \mapsto c\rho, \qquad s_{k,t} \mapsto s_{k,t}/c, \qquad c > 0.$$

We therefore normalize the shape by imposing

$$\rho(z, a_0) = 0, \qquad \|\rho(z, \cdot)\|_\varpi^2 = \int_{\mathcal{A}_\star} \rho(z, a)^2 \, \varpi(da) = 1 \quad \text{for all } z.$$

To fix the remaining sign ambiguity, choose a nonzero loading $b : \mathcal{A}_\star \to \mathbb{R}$ satisfying

$$b(a_0) = 0, \qquad \|b\|_\varpi = 1,$$

and require

$$\int_{\mathcal{A}_\star} \rho(z, a) b(a) \, \varpi(da) \geq 0 \quad \text{for all } z.$$

This convention fixes the within-candidate scale and orientation of ρ. The profiled loss in Section 3.3 additionally normalizes the anchored response energy, so cross-link comparisons are made at the level of geometry rather than contrast amplitude.

D.3.2 Profiled calibration in closed form

Lemma D.3 (Closed form of profiled calibration)**.** *Suppose ρ satisfies the normalization in Appendix D.3.1. For fixed k, t, h, u, ρ, define*

$$\Delta(a) := \Delta_{k,t}(h, a), \qquad R(a) := \rho(u(h), a).$$

If $\|\Delta\|_\varpi > 0$, then the minimizer of

$$\min_{s \geq 0} \|\Delta - sR\|_\varpi^2$$

is

$$s_{k,t}^\star(h; u, \rho) = [\langle \Delta, R \rangle_\varpi]_+ = \left[\int_{\mathcal{A}_\star} \Delta_{k,t}(h, a)\rho(u(h), a) \, \varpi(da) \right]_+ .$$

The same $s^\star$ minimizes the normalized objective

$$\min_{s \geq 0} \frac{\|\Delta - sR\|_\varpi^2}{\|\Delta\|_\varpi^2}.$$

If $\|\Delta\|_\varpi = 0$, the normalized contribution is defined as 0, and $s = 0$ is admissible.

Proof. By the normalization $\|R\|_\varpi = 1$,

$$\|\Delta - sR\|_\varpi^2 = \|\Delta\|_\varpi^2 - 2s\langle \Delta, R\rangle_\varpi + s^2.$$

This is a convex quadratic over $s \geq 0$, whose minimizer is $s^\star = [\langle \Delta, R\rangle_\varpi]_+$. Dividing by the positive constant $\|\Delta\|_\varpi^2$ does not change the minimizer. If $\|\Delta\|_\varpi = 0$, then $\Delta = 0$ in $L_2(\varpi)$, and the stated zero-contribution convention applies. □

D.4 Properties of the oracle criterion

D.4.1 Scale invariance of the profiled loss

Lemma D.4 (Scale invariance)**.** *Let $c_t(h) > 0$ be measurable and define*

$$\widetilde{\Delta}_{k,t}(h, a) := c_t(h)\Delta_{k,t}(h, a).$$

For any candidate (u, ρ), the pointwise normalized profiled loss is unchanged when $\Delta_{k,t}$ is replaced by $\widetilde{\Delta}_{k,t}$.

Proof. If $\|\Delta_{k,t}(h, \cdot)\|_\varpi = 0$, then also $\|\widetilde{\Delta}_{k,t}(h, \cdot)\|_\varpi = 0$, and both contributions are defined as 0. Otherwise,

$$\begin{aligned}\inf_{s\geq 0} \frac{\|\widetilde{\Delta}_{k,t}(h, \cdot) - s\rho(u(h), \cdot)\|_\varpi^2}{\|\widetilde{\Delta}_{k,t}(h, \cdot)\|_\varpi^2} &= \inf_{s\geq 0} \frac{\|c_t(h)\Delta_{k,t}(h, \cdot) - s\rho(u(h), \cdot)\|_\varpi^2}{c_t(h)^2\|\Delta_{k,t}(h, \cdot)\|_\varpi^2} \\ &= \inf_{r\geq 0} \frac{\|\Delta_{k,t}(h, \cdot) - r\rho(u(h), \cdot)\|_\varpi^2}{\|\Delta_{k,t}(h, \cdot)\|_\varpi^2},\end{aligned}$$

where $r = s/c_t(h)$. This proves the claim. □

D.4.2 Exact specification implies zero approximation loss

Lemma D.5 (Exact calibrated shape gives zero loss)**.** *Suppose that for some $k \in \mathcal{K}$, $u \in \mathfrak{U}$, and $\rho \in \mathfrak{R}_k(u)$,*

$$\Delta_{k,t}(h, a) = s_{k,t}(h)\rho(u(h), a)$$

for every $t = 1, \ldots, T$, almost every h, and every $a \in \mathcal{A}_\star$, with $s_{k,t}(h) \geq 0$. Then

$$\mathcal{A}_k(u, \rho) = 0.$$

Proof. Fix t, h. If $\|\Delta_{k,t}(h, \cdot)\|_\varpi = 0$, the pointwise contribution to $\mathcal{A}_k(u, \rho)$ is 0 by convention. Otherwise, choosing $s = s_{k,t}(h)$ makes the numerator of the profiled loss equal to 0. Hence the contribution is 0 pointwise, and the expectation over t, h is also 0. □

D.4.3 Non-uniqueness of unrestricted representations

Proposition D.6 (Unrestricted representation search is non-unique)**.** *Suppose (u, ρ) satisfies*

$$\Delta_{k,t}(h, a) = s_{k,t}(h)\rho(u(h), a)$$

for all t, h, a. Let $w : \mathcal{H} \to \mathbb{R}^r$ be any measurable map and define

$$\widetilde{u}(h) := (u(h), w(h)), \qquad \widetilde{\rho}((z, r), a) := \rho(z, a).$$

Then the same factorization holds with $(\widetilde{u}, \widetilde{\rho})$, and the induced action geometry is unchanged:

$$\arg\max_{a\in\mathcal{A}_\star} \widetilde{\rho}(\widetilde{u}(h), a) = \arg\max_{a\in\mathcal{A}_\star} \rho(u(h), a).$$

Proof. By construction,

$$\widetilde{\rho}(\widetilde{u}(h), a) = \widetilde{\rho}((u(h), w(h)), a) = \rho(u(h), a).$$

Thus the factorization and the induced action correspondence are unchanged. This shows that the representation itself is not identifiable without controlled classes or additional restrictions; the decision-relevant object is the induced geometry. □

D.5 Proof of Theorem 3.3

Proof. Fix a deployment period $t \in \mathcal{T}_{\text{dep}}$ and a history h for which Assumption 3.2 holds. Let $a^\rho(h) \in G^\star(h)$, so that

$$\rho^\star(u^\star(h), a^\rho(h)) \geq \rho^\star(u^\star(h), a) \qquad \forall a \in \mathcal{A}_\star.$$

Let $a_t^{\text{opt}}(h) \in \arg\max_{a \in \mathcal{A}_\star} m_t(h, a)$. Because $\ell_{k^\star}$ is strictly increasing and anchoring subtracts the action-independent constant $\ell_{k^\star}(m_t(h, a_0))$, the maximizers of $m_t(h, \cdot)$ and $\Delta_{k^\star,t}(h, \cdot)$ coincide. Hence

$$\Delta_{k^\star,t}(h, a_t^{\text{opt}}(h)) \geq \Delta_{k^\star,t}(h, a^\rho(h)),$$

which gives the lower bound.

For the upper bound, write

$$\Delta_{k^\star,t}(h, a) = s_t(h)\rho^\star(u^\star(h), a) + \xi_t(h, a), \qquad s_t(h) > 0.$$

Then

$$\begin{aligned}
&\Delta_{k^\star,t}(h, a_t^{\text{opt}}(h)) - \Delta_{k^\star,t}(h, a^\rho(h)) \\
&= s_t(h)\Big[\rho^\star(u^\star(h), a_t^{\text{opt}}(h)) - \rho^\star(u^\star(h), a^\rho(h))\Big] \\
&\qquad + \xi_t(h, a_t^{\text{opt}}(h)) - \xi_t(h, a^\rho(h)).
\end{aligned}$$

The bracketed term is non-positive because $a^\rho(h)$ maximizes $\rho^\star(u^\star(h), \cdot)$ and $s_t(h) > 0$. Therefore,

$$\begin{aligned}
&\Delta_{k^\star,t}(h, a_t^{\text{opt}}(h)) - \Delta_{k^\star,t}(h, a^\rho(h)) \\
&\leq \xi_t(h, a_t^{\text{opt}}(h)) - \xi_t(h, a^\rho(h)) \leq 2 \sup_{a \in \mathcal{A}_\star} |\xi_t(h, a)|.
\end{aligned}$$

This proves the stated bound. If $\xi_t \equiv 0$, the upper bound is zero, so every selector from $G^\star$ is optimal in period t. Since t and h were arbitrary on the set where the persistence condition holds, the result follows. □

E Additional Results for Section 4

This appendix gives the technical details behind the estimator in Section 4. Appendix E.1 states the Stage-I conditions. Appendix E.2 proves identification and Neyman orthogonality of the localized score. Appendix E.3 derives the finite-grid pilot bound used in (2). Appendix E.4 records stability of the profiled finite-grid criterion. Appendix E.5 proves the finite-library oracle inequality. Appendix E.6 clarifies the scope of the guarantee, and Appendix E.7 gives the criterion-to-policy bridge used in the decision consequence.

E.1 Stage-I regularity

Assumption E.1 (Stage-I regularity)**.** *For each observed period $t \in \{1, \ldots, T\}$, the following conditions hold on the target region.*

1. ***One-step identification.*** *Consistency, no interference, overlap, and one-step exchangeability hold as in Assumption 2.1. Hence $\mu_{0,t}(h, a) := \mathbb{E}[Y \mid H = h, A = a, \tau = t] = m_t(h, a)$ on the target support.*

2. ***Continuous-action overlap.*** *The generalized propensity density $g_{0,t}(a \mid h) := f_{A|H,\tau=t}(a \mid h)$ exists and satisfies $0 < \underline{g} \leq g_{0,t}(a \mid h) \leq \overline{g} < \infty$ on the target region.*

3. ***Smoothness and boundedness.*** *The response $m_t(h, a)$ is uniformly bounded and α_a-Hölder in a. For each $u \in \mathfrak{U}_n$, the law of $U = u(H) \mid \tau = t$ admits a density bounded and bounded away from zero on its target region, and the representation-indexed response is α_u-Hölder in u.*

4. ***Representation localization.*** *The finite-grid pilot statement is written for $m_t(h, a)$. This holds if, on the Stage-II evaluation region, $m_t(h, a)$ is identified by localization in $u(h)$ for each candidate $u \in \mathfrak{U}_n$. If a candidate representation is not sufficient, the same analysis applies to the projected response $m_{t,u}(z, a) := \mathbb{E}[m_t(H, a) \mid u(H) = z, \tau = t]$, and the projection error is absorbed into the restricted oracle approximation term.*

5. ***Kernels and bandwidths.*** *The kernels K and L are bounded, compactly supported, integrate to one, and have finite second moments. The bandwidths satisfy $\lambda \to 0$, $b \to 0$, $n_t \lambda^{d_u} b \to \infty$, and $n_t \lambda^{d_u} \to \infty$.*

6. ***Cross-fitted nuisance rates.*** *The nuisance estimators are fitted on folds independent of the observations on which their scores are evaluated and satisfy $\|\hat{\mu}_t - \mu_{0,t}\|_{L_2(P_t)} = O_p(r_{\mu,t,n})$ and $\|\hat{g}_t - g_{0,t}\|_{L_2(P_t)} = O_p(r_{g,t,n})$, with $\bar{r}_n := \max_t r_{\mu,t,n} r_{g,t,n} = o(1)$.*

7. ***Log-link positivity.*** *For the log link, $m_t(h,a)$ is uniformly bounded away from zero on the link-valid target region.*

E.2 Localized score: identification and orthogonality

Fix t, u, h, a, and write $U = u(H)$. Define

$$\nu_{t,\lambda,b}(u,h,a) := \mathbb{E}\left[\frac{\mathbf{1}\{\tau = t\}}{\pi_t} L_\lambda(U - u(h)) \int_{\mathcal{A}_\star} K_b(s-a) m_t(H,s)\, ds\right],$$

and

$$\delta_{t,\lambda}(u,h) := \mathbb{E}\left[\frac{\mathbf{1}\{\tau = t\}}{\pi_t} L_\lambda(U - u(h))\right].$$

Let $\bar{m}_{t,\lambda,b}(u,h,a) := \nu_{t,\lambda,b}(u,h,a)/\delta_{t,\lambda}(u,h)$.

Proposition E.2 (Score identification and Neyman orthogonality)**.** *Under Assumption E.1,*

$$\mathbb{E}[\phi_{t,\lambda,b}(O;u,h,a,\eta_{0,t})] = \nu_{t,\lambda,b}(u,h,a), \qquad \mathbb{E}[\psi_{t,\lambda}(O;u,h)] = \delta_{t,\lambda}(u,h).$$

Moreover, the map $\eta_t \mapsto \mathbb{E}[\phi_{t,\lambda,b}(O;u,h,a,\eta_t)]$ is Gateaux-orthogonal at $\eta_{0,t} = (\mu_{0,t}, g_{0,t})$.

Proof. At the true nuisances, $\mathbb{E}[Y - \mu_{0,t}(H,A) \mid H, A, \tau = t] = 0$, so the residualized term in ϕ has mean zero. The remaining term equals

$$\mathbb{E}\left[\frac{\mathbf{1}\{\tau = t\}}{\pi_t} L_\lambda(U - u(h)) \int_{\mathcal{A}_\star} K_b(s-a) \mu_{0,t}(H,s)\, ds\right],$$

which is $\nu_{t,\lambda,b}(u,h,a)$ because $\mu_{0,t}(H,s) = m_t(H,s)$ under one-step identification. The identity for ψ is immediate.

For orthogonality, let $\eta_{t,r} = (\mu_{t,r}, g_{t,r})$ be a regular path through $\eta_{0,t}$ with derivatives $(\dot{\mu}_t, \dot{g}_t)$. Differentiating the score mean at $r = 0$, the derivative in the μ-direction is

$$\mathbb{E}\left[\frac{\mathbf{1}\{\tau = t\}}{\pi_t} L_\lambda(U - u(h)) \left\{-\frac{K_b(A-a)}{g_{0,t}(A \mid H)} \dot{\mu}_t(H,A) + \int K_b(s-a) \dot{\mu}_t(H,s)\, ds\right\}\right],$$

which is zero after conditioning on $(H, \tau = t)$. The derivative in the g-direction is proportional to

$$\mathbb{E}\left[\frac{\mathbf{1}\{\tau = t\}}{\pi_t} L_\lambda(U - u(h)) \frac{K_b(A-a)\dot{g}_t(A \mid H)}{g_{0,t}(A \mid H)^2} \{Y - \mu_{0,t}(H,A)\}\right],$$

which is zero after conditioning on $(H, A, \tau = t)$. Hence the first derivative vanishes in every regular nuisance direction. □

E.3 Pilot expansion and finite-grid control

Lemma E.3 (Ratio and log-ratio stability)**.** *If $x_n - x = O_p(r_n)$, $y_n - y = O_p(r_n)$, and y is bounded away from zero, then*

$$\frac{x_n}{y_n} - \frac{x}{y} = \frac{x_n - x}{y} - \frac{x}{y^2}(y_n - y) + O_p(r_n^2).$$

If also x is bounded away from zero, then

$$(\log x_n - \log y_n) - (\log x - \log y) = \frac{x_n - x}{x} - \frac{y_n - y}{y} + O_p(r_n^2).$$

Proof. Both claims are first-order Taylor expansions on compact subsets bounded away from zero. □

Let $\bar{\Delta}_{k,t,\lambda,b}$ denote the anchored contrast constructed from $\bar{m}_{t,\lambda,b}$ with the same link ℓ_k.

Lemma E.4 (Pointwise anchored-pilot expansion)**.** *Under Assumption E.1, for every fixed* (k, t, u, h, a)*,*

$$\hat{\Delta}_{k,t}(u,h,a) - \Delta_{k,t}(h,a) = O_p\Big((n_t\lambda^{d_u}b)^{-1/2} + \bar{r}_n + \lambda^{\alpha_u} + b^{\alpha_a}\Big).$$

Proof. By Proposition E.2 and cross-fitting,

$$\hat{\nu}_{t,\lambda,b}(u,h,a) - \nu_{t,\lambda,b}(u,h,a) = (\mathbb{P}_n - P)\phi_{t,\lambda,b}(O; u, h, a, \eta_{0,t}) + O_p(\bar{r}_n),$$

where the remainder is second order in the nuisance errors. The empirical-process term has order $O_p((n_t\lambda^{d_u}b)^{-1/2})$. Similarly, $\hat{\delta}_{t,\lambda}(u,h) - \delta_{t,\lambda}(u,h) = O_p((n_t\lambda^{d_u})^{-1/2})$, which is no larger than the displayed variance order. Lemma E.3 transfers these rates to the ratio estimator and, under log-link positivity, to the log-ratio estimator. The bias $\bar{m}_{t,\lambda,b}(u,h,a) - m_t(h,a)$ is the standard kernel smoothing bias in the representation and action coordinates, namely $O(\lambda^{\alpha_u} + b^{\alpha_a})$. Anchoring at a_0 preserves the same order. □

Proposition E.5 (Finite-grid pilot control)**.** *Under Assumption E.1, there exists a deterministic sequence* $\varepsilon_n^{\mathrm{pilot}} \downarrow 0$ *such that the finite-grid pilot bound in* (2) *holds.*

Proof. Condition on the independent selection sample. Lemma E.4 gives the stated pointwise rate at each evaluation tuple $(k, u, t, H_i, a^{(m)})$. Since $\mathcal{K}$, T, $\mathfrak{U}_n$, and the Stage-II action grid are finite, a union bound over these indices and over $i \in \mathcal{I}^{\mathrm{sel}}$ yields

$$\sup_{k,u,t,i,m} \left|\hat{\Delta}_{k,t}(u, H_i, a^{(m)}) - \Delta_{k,t}(H_i, a^{(m)})\right| = O_p(\varepsilon_n^{\mathrm{pilot}}),$$

where $\varepsilon_n^{\mathrm{pilot}}$ absorbs the pointwise rate and the finite-library/grid union factor. This is exactly (2). □

E.4 Finite-grid profiled criterion

Let

$$\Theta_n := \{(k, u, \rho) : k \in \mathcal{K},\ u \in \mathfrak{U}_n,\ \rho \in \mathfrak{R}_{k,n}(u)\}.$$

For $\theta = (k, u, \rho)$ and $i \in \mathcal{I}^{\mathrm{sel}}$, define the population grid energy $V_i(\theta) = \sum_m w_m \Delta_{k,\tau_i}(H_i, a^{(m)})^2$.

Assumption E.6 (Finite-grid stability)**.** *On the nonzero-energy selection region, there exists* $\underline{v} > 0$ *such that* $\inf_{\theta\in\Theta_n} \min_i V_i(\theta) \geq \underline{v}$*. The shape classes are uniformly bounded and normalized, and the quadrature rule satisfies*

$$\sup_{\theta\in\Theta_n, i} \left|\sum_m w_m \rho(u(H_i), a^{(m)})^2 - 1\right| \leq q_M, \qquad q_M \downarrow 0.$$

Population zero-energy terms are handled by the zero-energy convention in the main text and may be removed from the normalized criterion.

Lemma E.7 (Profiled calibration on the finite grid)**.** *For fixed* (θ, i)*, write* $\Delta_{i,m} = \Delta_{k,\tau_i}(H_i, a^{(m)})$ *and* $\rho_{i,m} = \rho(u(H_i), a^{(m)})$*. The gridwise minimizer of* $\min_{s\geq 0} \sum_m w_m(\Delta_{i,m} - s\rho_{i,m})^2$ *is*

$$s_i^\star = \left[\frac{\sum_m w_m \Delta_{i,m}\rho_{i,m}}{\sum_m w_m \rho_{i,m}^2}\right]_+.$$

The same $s_i^\star$ *minimizes the normalized pointwise loss. Under Assumption E.6, replacing population contrasts by pilots changes the normalized profiled loss by* $O_p(\varepsilon_n^{\mathrm{pilot}})$ *uniformly over* Θ_n*.*

Proof. The unnormalized grid objective is the quadratic

$$Q_i(s) = \sum_m w_m \Delta_{i,m}^2 - 2s \sum_m w_m \Delta_{i,m} \rho_{i,m} + s^2 \sum_m w_m \rho_{i,m}^2,$$

so the constrained minimizer is the displayed positive part. Dividing by the energy $\sum_m w_m \Delta_{i,m}^2$ does not change the minimizer. Uniform pilot stability follows because the numerator, denominator, and profiled quadratic are locally Lipschitz in the finite vector $(\Delta_{i,1}, \ldots, \Delta_{i,M})$ on the event where the energy is bounded below by $\underline{v}/2$. Proposition E.5 and Assumption E.6 give this event with probability tending to one. □

E.5 Finite-library oracle inequality

Lemma E.8 (Uniform empirical approximation)**.** *Assume Proposition E.5, Assumption E.6, and that each $\mathfrak{R}_{k,n}(u)$ admits a bounded finite-dimensional Lipschitz parameterization of dimension at most p_n. Then*

$$\sup_{\theta \in \Theta_n} \left| \hat{\mathcal{L}}^{\mathrm{sel}}(\theta) - \mathcal{L}(\theta) \right| = O_p\left(\varepsilon_n^{\mathrm{pilot}} + \sqrt{\frac{\log |\mathfrak{U}_n| + p_n \log n}{n_{\mathrm{sel}}}} + q_M \right).$$

Proof. The difference decomposes into pilot error, sampling error, and quadrature error. By Lemma E.7, replacing population contrasts by pilot contrasts changes the profiled normalized loss by $O_p(\varepsilon_n^{\mathrm{pilot}})$ uniformly. Conditional on the pilots being replaced by population contrasts, the criterion is a bounded empirical process over a finite representation library and p_n-dimensional Lipschitz shape sieves, giving

$$O_p\left(\sqrt{\frac{\log |\mathfrak{U}_n| + p_n \log n}{n_{\mathrm{sel}}}} \right).$$

The finite-grid approximation contributes q_M. The deterministic penalties are identical in empirical and population criteria, so they do not affect the supremum difference. □

Proof of Theorem 4.1. Let $\hat{\theta} = (\hat{k}, \hat{u}, \hat{\rho})$ and let $\theta_n^\star \in \arg\min_{\theta \in \Theta_n} \mathcal{L}(\theta)$. Empirical optimality gives $\hat{\mathcal{L}}^{\mathrm{sel}}(\hat{\theta}) \le \hat{\mathcal{L}}^{\mathrm{sel}}(\theta_n^\star)$. Hence

$$\mathcal{L}(\hat{\theta}) - \mathcal{L}(\theta_n^\star) \le 2 \sup_{\theta \in \Theta_n} |\hat{\mathcal{L}}^{\mathrm{sel}}(\theta) - \mathcal{L}(\theta)|.$$

Lemma E.8 yields the claimed $O_p(r_n)$ bound. □

Corollary E.9 (Link selection consistency)**.** *Let $V_n(k) := \inf_{u,\rho} \mathcal{L}_k(u, \rho)$, where $u \in \mathfrak{U}_n$ and $\rho \in \mathfrak{R}_{k,n}(u)$. Suppose $k^\star$ is the unique minimizer and $\gamma_n := \min_{k \neq k^\star} \{V_n(k) - V_n(k^\star)\}$. If $r_n = o(\gamma_n)$, then $\mathbb{P}(\hat{k} = k^\star) \to 1$.*

Proof. If $\hat{k} \neq k^\star$, then $\mathcal{L}(\hat{k}, \hat{u}, \hat{\rho}) - \inf_{\theta \in \Theta_n} \mathcal{L}(\theta) \ge \gamma_n$. Theorem 4.1 bounds the left-hand side by $O_p(r_n)$. Therefore $r_n = o(\gamma_n)$ implies $\mathbb{P}(\hat{k} \neq k^\star) \to 0$. □

E.6 Scope of the finite-library guarantee

Theorem 4.1 is a finite-library, criterion-level result. It should be read as follows. The population object is the restricted oracle criterion from Section 3. The statistical statement controls the excess of this criterion over the admissible library Θ_n. It does not claim identification of a unique latent representation, a globally stable identity response surface, or convergence of an unrestricted nonconvex neural representation learner. In implementations with flexible encoders, the theorem applies to the selected finite candidate library or to a sieve approximation of that library; additional optimization error can be added to r_n in the usual way.

E.7 Decision consequences

The oracle inequality controls a geometry criterion. We next state sufficient conditions under which this criterion control yields policy regret. The result is stated for the finite decision grid used by the estimator. Continuous-action extensions require the corresponding continuous-action margin or curvature condition.

Let $\mathcal{A}_{\mathrm{dec}} \subset \mathcal{A}_\star$ be the finite decision grid. Let $\mathcal{L}_n^\star := \inf_{\theta\in\Theta_n} \mathcal{L}(\theta)$ and $\Theta_n^\star := \arg\min_{\theta\in\Theta_n} \mathcal{L}(\theta)$. For $\theta = (k, u, \rho)$ and $\theta' = (k', u', \rho')$, define

$$d_\infty(\theta, \theta') := \operatorname*{ess\,sup}_h \max_{a\in\mathcal{A}_{\mathrm{dec}}} |\rho(u(h), a) - \rho'(u'(h), a)|, \quad d_\infty(\theta, \Theta_n^\star) := \inf_{\theta'\in\Theta_n^\star} d_\infty(\theta, \theta').$$

Assumption E.10 (Criterion-to-geometry modulus)**.** *There exists a nondecreasing deterministic modulus Ψ_n, with $\Psi_n(\delta) \to 0$ as $\delta \downarrow 0$, such that*

$$d_\infty(\theta, \Theta_n^\star) \le \Psi_n\{\mathcal{L}(\theta) - \mathcal{L}_n^\star\} \qquad \forall \theta \in \Theta_n.$$

Proposition E.11 (Geometry recovery from criterion learnability)**.** *Under Theorem 4.1 and Assumption E.10,*

$$d_\infty\{(\hat{k}, \hat{u}, \hat{\rho}), \Theta_n^\star\} = O_p\{\Psi_n(r_n)\}.$$

Proof. Let $\hat\theta = (\hat{k}, \hat{u}, \hat{\rho})$. By Theorem 4.1, $\mathcal{L}(\hat\theta) - \mathcal{L}_n^\star = O_p(r_n)$. Applying Assumption E.10 gives the result. □

Fix an oracle representative $\theta_n^\star = (k_n^\star, u_n^\star, \rho_n^\star) \in \Theta_n^\star$, and let $a_n^\star(h) \in \arg\max_{a\in\mathcal{A}_{\mathrm{dec}}} \rho_n^\star(u_n^\star(h), a)$.

Assumption E.12 (Deployment bridge and margin)**.** *For every deployment period $t \in \mathcal{T}_{\mathrm{dep}}$, assume*

$$\Delta_{k_n^\star,t}(h, a) = s_t(h)\rho_n^\star(u_n^\star(h), a) + \xi_t(h, a), \qquad 0 < s_t(h) \le \overline{s},$$

for all $a \in \mathcal{A}_{\mathrm{dec}}$. Also assume a linked-to-raw comparison: for some $C_{\mathrm{lr}} < \infty$,

$$m_t(h, a_t^{\mathrm{opt}}(h)) - m_t(h, a) \le C_{\mathrm{lr}}\{\Delta_{k_n^\star,t}(h, a_t^{\mathrm{opt}}(h)) - \Delta_{k_n^\star,t}(h, a)\}.$$

Finally, define the grid margin

$$\Gamma_n(h) := \rho_n^\star(u_n^\star(h), a_n^\star(h)) - \max_{a\in\mathcal{A}_{\mathrm{dec}},\, a\neq a_n^\star(h)} \rho_n^\star(u_n^\star(h), a),$$

and assume $\sup_{t\in\mathcal{T}_{\mathrm{dep}}} \mathbb{P}\{\Gamma_n(H_t) \le \epsilon \mid \tau = t\} \le C\epsilon^\kappa$.

For the identity link, one may take $C_{\mathrm{lr}} = 1$. For the log link, the linked-to-raw comparison follows from a uniform upper bound on m_t over the deployment region.

Corollary E.13 (Deployment regret from finite-library geometry error)**.** *Suppose Proposition E.11 and Assumption E.12 hold. Define*

$$\mathrm{Pers}_t := \mathbb{E}_t\left[\max_{a\in\mathcal{A}_{\mathrm{dec}}} |\xi_t(H_t, a)|\right].$$

Then, for each $t \in \mathcal{T}_{\mathrm{dep}}$,

$$\mathrm{Regret}_t(\hat{a}) := \mathbb{E}_t[m_t(H_t, a_t^{\mathrm{opt}}(H_t)) - m_t(H_t, \hat{a}(H_t))] = O_p\big(\Psi_n(r_n)^{\kappa+1}\big) + O(\mathrm{Pers}_t).$$

Proof. Let

$$\delta_n := \operatorname*{ess\,sup}_h \max_{a\in\mathcal{A}_{\mathrm{dec}}} |\hat\rho(\hat{u}(h), a) - \rho_n^\star(u_n^\star(h), a)|.$$

By Proposition E.11, $\delta_n = O_p\{\Psi_n(r_n)\}$. On the event $\{\delta_n \le \delta\}$, if $\hat{a}(h) \neq a_n^\star(h)$, then $\Gamma_n(h) \le 2\delta$; otherwise the uniform perturbation could not change the maximizer. Hence

$$0 \le \rho_n^\star(u_n^\star(h), a_n^\star(h)) - \rho_n^\star(u_n^\star(h), \hat{a}(h)) \le 2\delta\, \mathbf{1}\{\Gamma_n(h) \le 2\delta\}.$$

Using the deployment bridge and $s_t(h) \le \overline{s}$,

$$\Delta_{k_n^\star,t}(h, a_t^{\mathrm{opt}}(h)) - \Delta_{k_n^\star,t}(h, \hat{a}(h)) \le 2\overline{s}\,\delta\, \mathbf{1}\{\Gamma_n(h) \le 2\delta\} + 2 \max_{a\in\mathcal{A}_{\mathrm{dec}}} |\xi_t(h, a)|.$$

Taking conditional expectation and applying the margin condition gives $O(\delta^{\kappa+1}) + O(\mathrm{Pers}_t)$ for linked-response regret. The linked-to-raw comparison transfers the bound to raw regret. Substituting $\delta = \delta_n$ completes the proof. □

F Implementation Algorithm

Algorithm 1 Selection and optional honest refit for deployable action geometry

Require: Data $\mathcal{D} = \{(T_i, H_i, A_i, Y_i)\}_{i=1}^n$, link set $\mathcal{K}$, representation library $\mathfrak{U}_n$, normalized shape classes $\{\mathfrak{R}_{k,n}(u)\}$, kernels (K_b, L_λ), grid $\{(a^{(m)}, w_m)\}_{m=1}^M$, penalties, truncation τ_n.

Ensure: $(\hat{k}, \hat{u}, \hat{\rho})$ and rule $\hat{a}(h) \in \arg\max_{a \in \mathcal{A}_\star} \hat{\rho}(\hat{u}(h), a)$.

Split $\mathcal{D}$ into $\mathcal{I}^{\mathrm{pil}}$, $\mathcal{I}^{\mathrm{sel}}$ (and optionally $\mathcal{I}^{\mathrm{ref}}$), preserving temporal proportions.

Fit cross-fitted nuisance estimates $\hat{\mu}_t, \hat{g}_t$ on $\mathcal{I}^{\mathrm{pil}}$.

for each $u \in \mathfrak{U}_n$ and $i \in \mathcal{I}^{\mathrm{sel}}$ **do**

Compute $\hat{m}_{t,\lambda,b}(u, H_i, a) = \hat{\nu}_{t,\lambda,b}(u, H_i, a)/(\hat{\delta}_{t,\lambda}(u, H_i) \vee \tau_n)$ on the grid.

Define anchored pilots:

$$\hat{\Delta}^{(m)}_{0,T_i,i}(u) = \hat{m}(u, H_i, a^{(m)}) - \hat{m}(u, H_i, a_0),$$

$$\hat{\Delta}^{(m)}_{1,T_i,i}(u) = \log(\hat{m}(u, H_i, a^{(m)}) \vee \tau_n) - \log(\hat{m}(u, H_i, a_0) \vee \tau_n).$$

For each $k \in \mathcal{K}$, compute

$$\hat{V}_{k,i}(u) = \sum_{m=1}^M w_m \big(\hat{\Delta}^{(m)}_{k,T_i,i}(u)\big)^2.$$

end for

for each (k, u, ρ) with $k \in \mathcal{K}$, $u \in \mathfrak{U}_n$, $\rho \in \mathfrak{R}_{k,n}(u)$ **do**

for each $i \in \mathcal{I}^{\mathrm{sel}}$ **do**

Compute

$$\hat{s}_{k,i}(u, \rho) = \Big[\sum_{m=1}^M w_m \hat{\Delta}^{(m)}_{k,T_i,i}(u) \rho(u(H_i), a^{(m)}) \Big]_+.$$

Define

$$\hat{R}_{k,i}(u, \rho) = \begin{cases} \dfrac{\sum_{m=1}^M w_m \big(\hat{\Delta}^{(m)}_{k,T_i,i}(u) - \hat{s}_{k,i}(u, \rho)\rho(u(H_i), a^{(m)})\big)^2}{\hat{V}_{k,i}(u)}, & \hat{V}_{k,i}(u) > 0, \\ 0, & \hat{V}_{k,i}(u) = 0. \end{cases}$$

end for

Form

$$\hat{\mathcal{L}}^{\mathrm{sel}}_k(u, \rho) = |\mathcal{I}^{\mathrm{sel}}|^{-1} \sum_{i \in \mathcal{I}^{\mathrm{sel}}} \hat{R}_{k,i}(u, \rho) + \lambda_{u,n}\Omega_u(u) + \lambda_{\rho,n}\Omega_{k,\rho}(\rho).$$

end for

Select

$$(\hat{k}, \hat{u}, \hat{\rho}^{\mathrm{sel}}) \in \arg\min_{k,u,\rho} \hat{\mathcal{L}}^{\mathrm{sel}}_k(u, \rho).$$

if no honest refit is used **then**

Set $\hat{\rho} \leftarrow \hat{\rho}^{\mathrm{sel}}$.

else

Rebuild pilots on $\mathcal{I}^{\mathrm{ref}}$ only for $(\hat{k}, \hat{u})$ and minimize the corresponding refit criterion over $\rho \in \mathfrak{R}_{\hat{k},n}(\hat{u})$.

end if

Return $\hat{a}(h) \in \arg\max_{a \in \mathcal{A}_\star} \hat{\rho}(\hat{u}(h), a)$.

G Additional Synthetic Experiments

This appendix provides the simulation design and additional diagnostics supporting Section 5. Table 4 summarizes the role of each synthetic experiment. We use $R1$ and $R2$ as exact regimes, and $R3$ and $R4$ as boundary regimes.

Table 4: Synthetic experiment roadmap.

Experiment	Regimes	Purpose	Reported evidence
Link recovery	$R1, R2$	recover identity/log links	Figure 3(a,b), Tab. 5
Future deployment	$R1, R2$	improve future decisions	Figure 3(c,d), Tab. 1
Orthogonal ablation	$R1, R2$	nuisance robustness	Tab. 6
Approximate drift	$R3$	persistence violation	Fig. 6
No stable geometry	$R4$	structural failure	Fig. 6
Overlap stress	$R1, R2$	support degradation	Fig. 7
Grid sensitivity	$R1, R2$	action discretization robustness	Fig. 8
Anchor sensitivity	$R1, R2$	reference-action robustness	Fig. 8
End-to-end ablation	$R1, R2$	selection versus refitting	Tab. 7
Persistence magnitude	$R3$	structural deployment loss	Tab. 8
Link-selection stress	Transport shift	response-scale selection	Tab. 9

G.1 Data-generating process

We simulate n units over T periods. Each unit has latent state $S_{i,t} = (U_{i,t}, Z_{i,t})$, where $U_{i,t}$ determines the stable action-sensitive component and $Z_{i,t}$ drives temporal drift and treatment assignment. The observed history is a nonlinear, time-varying transformation

$$H_{i,t} = \phi_t(U_{i,t}, Z_{i,t}, Y_{i,t-1}, A_{i,t-1}, \varepsilon^H_{i,t}),$$

so the deployable representation is not directly observed. Latent states follow first-order dynamics

$$U_{i,t+1} = 0.82U_{i,t} + \xi^U_{i,t+1}, \qquad Z_{i,t+1} = 0.72Z_{i,t} + \Lambda_A A_{i,t} + \xi^Z_{i,t+1}.$$

Treatment assignment is observational and time-varying. Unless stated otherwise,

$$A_{i,t} \mid H_{i,t} \sim \mathrm{Beta}(\kappa\mu_{i,t}, \kappa(1-\mu_{i,t}))\,,$$

where $\mu_{i,t}$ depends on latent coordinates, a compressed lagged outcome, and period index. The global confounding-strength parameter multiplies the assignment signal. In $R1$, we use a bridge-style assignment rule that makes future action ranking more challenging while preserving the raw-scale stable geometry.

Let $a_0 \in (0,1)$ be the anchor and $d = a - a_0$. The shared action shape is

$$\rho^\star(u,a) = (0.35{+}0.25\tanh u_1)d{+}(0.18{+}0.10\tanh u_2)\sin(2\pi a){+}(-1.10{-}0.20\tanh(u_1{-}u_2))d^2,$$

with $\rho^\star(u, a_0) = 0$. The four regimes are

$$\begin{aligned}
R1: &\quad m_t(h,a) = B^{(1)}_t(h) + \rho^\star(u^\star(h),a),\\
R2: &\quad m_t(h,a) = D_t(h)\exp\{\rho^\star(u^\star(h),a)\},\\
R3: &\quad m_t(h,a) = B^{\mathrm{add}}_t(h) + \rho^\star(u^\star(h),a) + \epsilon_{\mathrm{drift}} q_t(u^\star(h),a),\\
R4: &\quad m_t(h,a) = 3.0 + B^{\mathrm{add}}_t(h) + \rho^\star(u^\star(h),a) + \Delta_t \widetilde{q}_t(u^\star(h),a).
\end{aligned}$$

Thus $R1$ has an exact additive reduction, $R2$ has an exact multiplicative reduction, $R3$ introduces controlled action-dependent drift, and $R4$ breaks stable geometry through strong time-varying shape drift. For $R1, R3, R4$, outcomes are generated as clipped additive Gaussian outcomes. For $R2$, noise is added on the log scale, so the observation model is log-normal. All oracle actions and regrets are evaluated on a dense grid.

G.2 Candidate library and evaluation metrics

The learner selects over a finite representation library $\mathcal{U}_n = \{u^{(1)}, u^{(2)}, u^{(3)}, u^{(4)}\}$:

$$
\begin{aligned}
u^{(1)}(H) &= U, && \text{oracle representation,} \\
u^{(2)}(H) &= U + 0.4\eta(H), && \text{deterministically perturbed oracle,} \\
u^{(3)}(H) &= U_1, && \text{under-specified representation,} \\
u^{(4)}(H) &= f_{\text{rand}}(H), && \text{fixed nonlinear random summary.}
\end{aligned}
$$

This library tests whether the profiled criterion can recover a deployable low-dimensional geometry rather than a time-entangled summary. The primary metrics are future-period stable IMSE (SIMSE) and regret. SIMSE measures recovery of the stable target on the future split; regret compares the oracle action with the learned action on the evaluation grid. Factual RMSE is reported only as a diagnostic and is not used for model selection.

G.3 Exact regimes: recovery and deployment

Table 5 summarizes the exact-regime sweep. Link recovery is substantially easier than full representation recovery, which is consistent with the paper's target: deployment requires the induced action geometry, not a unique latent representation. Figure 5 reports link selection accuracy and regret across sample size and confounding strength. Fixed-correct is supplied with the generating link, but still estimates the response shape and profiles its amplitude. Close decision performance relative to this reference indicates low adaptation cost.

Table 5: **Exact-regime recovery summary.** Fixed-correct denotes raw in $R1$ and log in $R2$; Fixed-wrong uses the other link.

Regime	LinkAcc	ReprAcc	Ours SIMSE	Fixed-correct SIMSE	Fixed-wrong SIMSE
$R1$	0.786 ± 0.178	0.352 ± 0.200	0.181 ± 0.019	0.160 ± 0.006	0.259 ± 0.003
$R2$	0.833 ± 0.148	0.195 ± 0.162	0.011 ± 0.002	0.012 ± 0.002	0.0140 ± 0.004

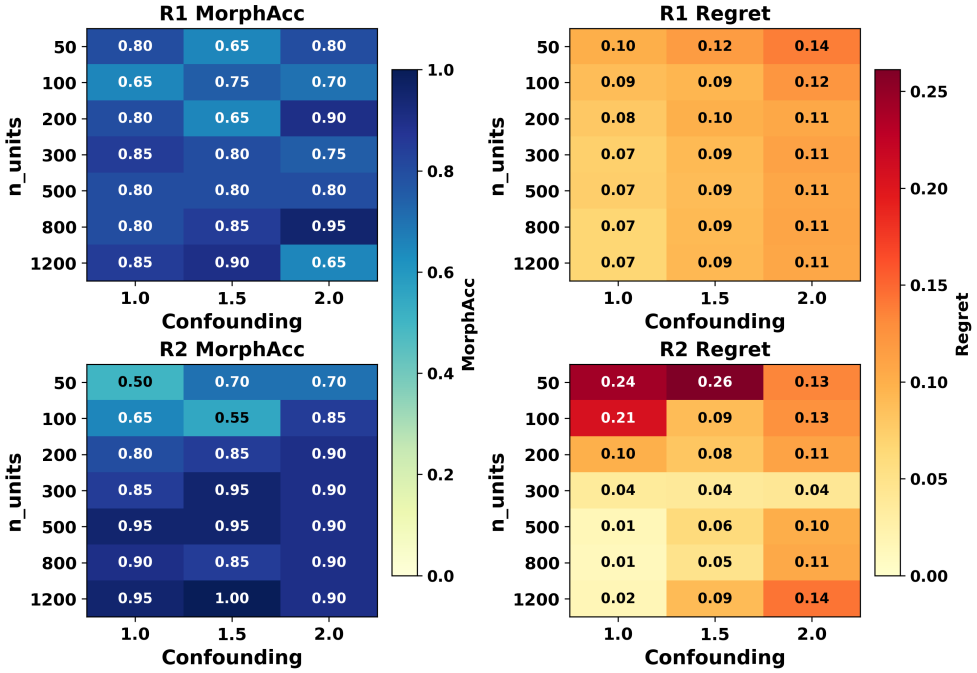


Figure 5: **Full exact-regime sweep.** Heatmaps over sample size and confounding strength for $R1$ and $R2$. Panels show link selection accuracy and future regret.

G.4 Post-selection and end-to-end orthogonalization

We freeze the selected link-representation pair and compare the orthogonal refit with direct plug-in under nuisance perturbations. This is a post-selection estimator ablation rather than a second target-selection experiment. The orthogonal estimator consistently yields lower regret and smaller action error than the plug-in estimator. The effect is strongest in the log-linked regime $R2$: under outcome-side misspecification, the plug-in estimator collapses, whereas the orthogonal estimator remains usable despite increased action error. In $R1$, SIMSE is similar across estimators, but the orthogonal estimator gives substantially lower regret, showing that predictive curve error and decision quality can diverge.

Table 6: Post-selection estimator ablation under nuisance misspecification.

Regime / Misspec.	Method	SIMSE ↓	Regret ↓	Action MAE ↓
R1				
none	OURS	0.1732 ± 0.0173	$\mathbf{0.0669 \pm 0.0088}$	$\mathbf{0.1205 \pm 0.0072}$
	PLUG-IN	$\mathbf{0.1697 \pm 0.0163}$	0.2438 ± 0.0492	0.2446 ± 0.0249
outcome	OURS	0.1737 ± 0.0173	$\mathbf{0.0683 \pm 0.0095}$	$\mathbf{0.1231 \pm 0.0075}$
	PLUG-IN	$\mathbf{0.1689 \pm 0.0174}$	0.1583 ± 0.0196	0.1920 ± 0.0148
treatment	OURS	0.1779 ± 0.0165	$\mathbf{0.0710 \pm 0.0050}$	$\mathbf{0.1250 \pm 0.0046}$
	PLUG-IN	$\mathbf{0.1697 \pm 0.0163}$	0.2438 ± 0.0492	0.2446 ± 0.0249
R2				
none	OURS	$\mathbf{0.0081 \pm 0.0011}$	$\mathbf{0.0154 \pm 0.0091}$	$\mathbf{0.0367 \pm 0.0168}$
	PLUG-IN	0.0145 ± 0.0124	0.3834 ± 0.0361	0.4732 ± 0.0427
outcome	OURS	$\mathbf{0.0066 \pm 0.0024}$	$\mathbf{0.1326 \pm 0.0560}$	$\mathbf{0.2076 \pm 0.0779}$
	PLUG-IN	4.8947 ± 2.3391	1.2555 ± 0.0052	0.4151 ± 0.0035
treatment	OURS	$\mathbf{0.0082 \pm 0.0011}$	$\mathbf{0.0163 \pm 0.0101}$	$\mathbf{0.0386 \pm 0.0183}$
	PLUG-IN	0.0145 ± 0.0124	0.3834 ± 0.0361	0.4732 ± 0.0427

Controlled end-to-end ablation. For component attribution, we additionally use a controlled plug-in pilot that omits only the residual-correction term. Localization, smoothing, shape fitting, penalties, and data splits are otherwise held fixed. We independently vary the pilot type at selection and honest refitting: O denotes orthogonal and P controlled plug-in, with the first letter indicating selection and the second refitting. Link and representation selection are repeated for each selection-stage pilot type.

The close regrets of OO and PO, and of OP and PP, locate most of the average improvement at orthogonal refitting in these settings. Selection-stage effects are regime-dependent: under moderate $R2$ outcome-model misspecification, link recovery is 0.57 with orthogonal selection and 0.33 with controlled plug-in selection. Both pilot types select the under-specified representation in the full-library ablation; this comparison does not establish recovery of the generating representation.

G.5 Boundary regimes: approximate and failed geometry

Figure 6 reports the two boundary regimes. In $R3$, increasing ϵ_{drift} adds action-dependent temporal drift to an otherwise stable geometry. Link recovery weakens as the persistence residual grows, which is the expected cost of misspecification. In $R4$, the action-sensitive shape itself changes over time. This is a structural failure regime; the method should not be expected to uniformly dominate. The purpose of $R4$ is to show where the stable-geometry target ceases to be appropriate.

Quantifying persistence violations. We additionally distinguish the structural cost of persistence failure from estimation error. For the identity-link $R3$ sweep, let $s_t^\dagger(h)$ denote the best positive rescaling of the true shared shape and set $\xi_t^\dagger(h,a) = \Delta_{0,t}(h,a) - s_t^\dagger(h)\rho^\star(u^\star(h),a)$. The relative violation is

$$\operatorname{Rel}\xi_\infty = \operatorname{median}_{t,h} \frac{\|\xi_t^\dagger(h,\cdot)\|_\infty}{\|s_t^\dagger(h)\rho^\star(u^\star(h),\cdot)\|_\infty + \epsilon},$$

where $\epsilon > 0$ stabilizes the denominator. The oracle-geometry policy maximizes the true shared shape.

Table 7: **End-to-end orthogonalization ablation.** Mean dense-grid regret under outcome-model misspecification over 30 paired runs using the full representation library.

Selection → refit	$R1$ regret ↓	$R2$ regret ↓
$O \to O$	0.0754	0.1220
$P \to O$	0.0754	0.1220
$O \to P$	0.1292	0.1444
$P \to P$	0.1292	0.1445
$PP - OO$	0.0538 [0.0505, 0.0569]	0.0225 [0.0092, 0.0422]

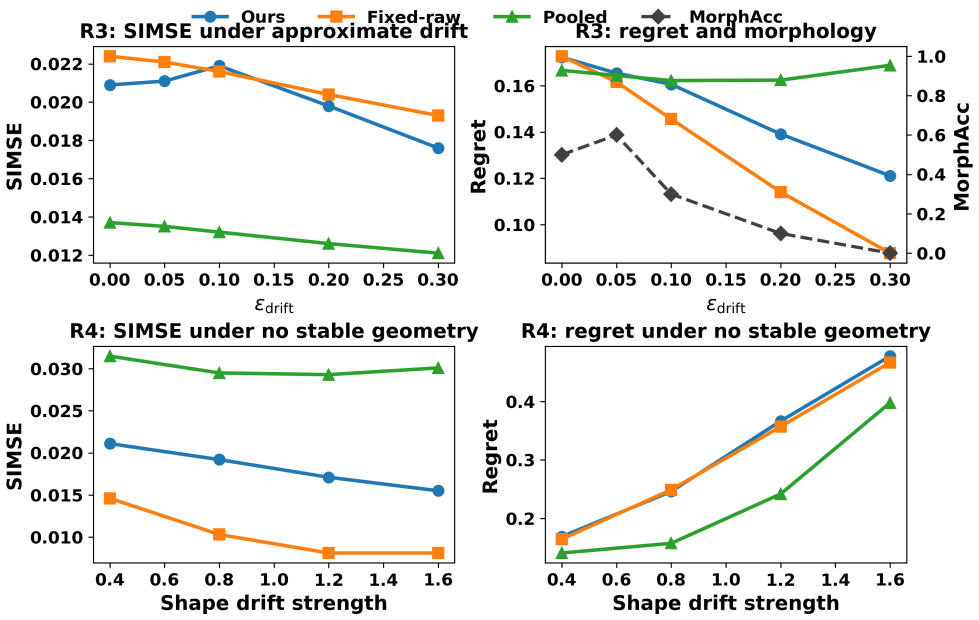


Figure 6: **Boundary diagnostics.** Left: approximate persistence violation in $R3$ as ϵ_{drift} increases. Right: no-stable-geometry stress in $R4$ as shape drift increases.

Table 8 shows increasing oracle-geometry regret as the persistence residual grows in this controlled sweep. It measures the cost of transporting the true shared geometry.

G.6 Overlap stress

We stress overlap by increasing assignment severity. This decreases empirical support fraction and conditional treatment variance. Figure 7 reports support diagnostics together with SIMSE and regret. Severe support loss harms all estimators; the goal is not to claim that the method solves positivity violation, but to show predictable degradation and to contrast it with pooled prediction.

G.7 Grid and anchor sensitivity

We further examine whether the conclusions depend on two implementation choices: the action grid used in Stage-II selection and the reference action used for anchoring. For grid sensitivity, the learner selects actions on grids of size $G \in \{21, 51, 101, 201\}$, while all metrics are evaluated against a fixed 501-point dense grid oracle. This separates optimization discretization from evaluation. For anchor sensitivity, we keep the data-generating process and true response surface fixed, and vary only the reference action used to construct anchored contrasts, $a_0 \in \{0.4, 0.5, 0.6\}$. Thus raw regret is always computed relative to the same underlying response oracle, while anchored SIMSE may vary with the reference point. Figure 8 reports the results. Grid sensitivity is benign: in both $R1$ and $R2$, regret is nearly unchanged once the grid is moderately dense, and the oracle discretization gap is already negligible at the 51-point grid used in the main experiments. Anchor sensitivity is more diagnostic. In $R1$, decision quality remains stable across the tested anchors. In $R2$, $a_0 = 0.4$ and the default

Table 8: **Controlled persistence-violation sweep in $R3$.** Relative persistence violation and the regret of the oracle-geometry policy as action-dependent drift increases.

ϵ_{drift}	$\text{Rel}\,\xi_\infty$	Oracle-geometry regret
0.00	0.000	0.0000
0.05	0.027	0.0004
0.10	0.055	0.0018
0.20	0.114	0.0095
0.30	0.179	0.0242

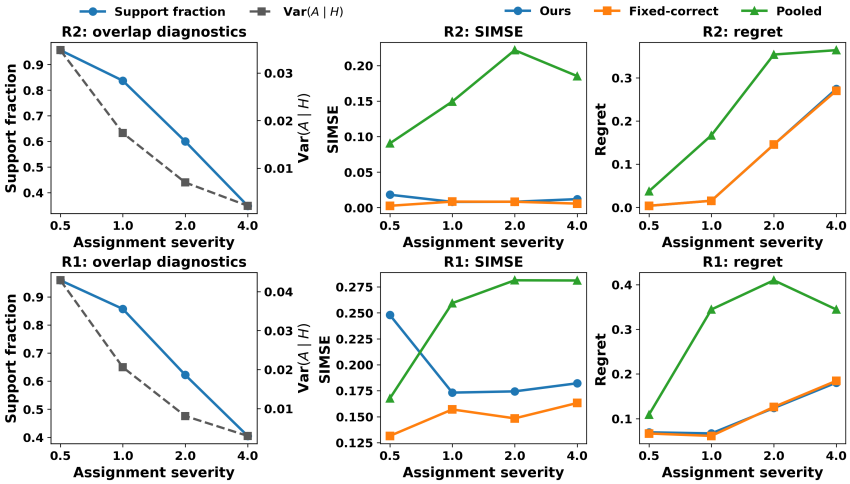


Figure 7: Overlap stress.

$a_0 = 0.5$ both outperform pooled prediction, whereas $a_0 = 0.6$ leads to substantially worse decisions even for the fixed-correct log link. This suggests that the anchor should be chosen as a semantically meaningful, high-support reference action rather than treated as an arbitrary tuning parameter.

G.8 Link selection under transport shift

To examine when selecting the response scale benefits deployment, we consider a supplementary transport-shift setting. The history distribution changes across periods while the generating link admits exact shared geometry, $\xi_t(h, a) = 0$. Under a wrong fixed link, the induced history-dependent curvature lies outside the shared shape class. The given-link reference is supplied with the generating link, not the true response surface or the optimal policy; its amplitude remains profiled as for Ours.

Ours has lower mean NRegret than the plug-in selector and both fixed-link alternatives, while the paired interval for the given-link reference includes zero. This complements the exact-regime comparison by examining selection benefits under transport shift, rather than only the cost of adapting to an unknown generating link.

H Additional details of the real-world experiment

H.1 Dataset Construction and Evaluation Protocol

Data and Action Definition. The panel spans December 1, 2025 to March 1, 2026, containing $\sim$ 546.3M observations across 8.3M unique items ($\sim$ 6.0M daily). The continuous action is the realized effective price normalized by the 14-day median of positive prices, yielding an item-specific anchor $a_0 \approx 1$. The full-sample action distribution is centered near 1 with a 5th–95th percentile range of $[0.922, 1.095]$.

Support Region and Cohort Filtering. Analysis is restricted to the empirical common-support region $A^\star = [0.93, 1.08]$ ($[0.90, 1.10]$ for sensitivity), requiring $\geq$ 10 positive-price days in the

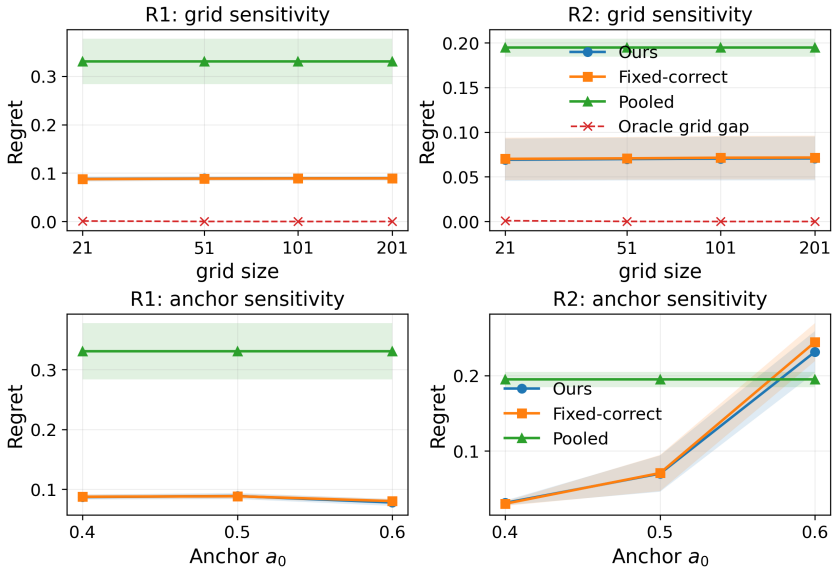


Figure 8: **Grid and anchor sensitivity.** Top: sensitivity to the Stage-II action grid. Bottom: sensitivity to the anchor a_0 with the data-generating process and raw-response oracle fixed.

Table 9: **Link selection under transport shift.** Lower NRegret is better. Paired differences are method minus Ours; a positive interval excluding zero favors Ours.

Method	NRegret	Difference	Paired 95% CI
Given-link reference	0.1384	−0.00008	[−0.00025, 0.00008]
Ours	0.1385	—	—
Plug-in selector	0.1412	0.00269	[0.00160, 0.00378]
Fixed identity	0.1416	0.00309	[0.00205, 0.00413]
Fixed log	0.1422	0.00368	[0.00291, 0.00445]

historical window. To control cross-item heterogeneity and ensure sufficient variation, we dynamically retain items satisfying: (i) ≥ 28 observed days, (ii) ≥ 5 distinct rounded action bins, (iii) action span $\mathrm{P90}(A) - \mathrm{P10}(A) \geq 0.04$, and (iv) zero-sales rate ≤ 0.10.

Evaluation Protocol. We use a rolling-origin chronological setup with six main splits (28-day train, 7-day validation, 7-day test, advanced weekly). An additional split after February 16, 2026, evaluates robustness under a shifted demand regime. After filtering, each split contains $\sim$ 56–58M training observations ($\sim$ 2.0M items) and 12–14M validation/test observations.

H.2 Metric definition

We evaluate each method along three complementary axes: factual prediction, future-time deployability, and decision quality.

Factual prediction error. For observed item-day samples (H_i, A_i, Y_i) in the evaluation period, we report the RMAE

$$\mathrm{RMAE} = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{\sum_{i=1}^{n} y_i}, \tag{5}$$

where $\hat{m}(H, a)$ denotes the learned response surface. RMAE is included as a diagnostic predictive metric, but it is not our primary deployment criterion.

Deployment stability loss. To measure whether the learned response geometry remains stable when transported to a future period, we compare the predicted anchored response curve with a future-period anchored target curve on a common action grid $\mathcal{A}^\star = \{a_1, \ldots, a_K\}$. For a history H_i, define the

anchored predicted curve

$$\hat{\rho}_i(a) = \frac{\hat{m}(H_i, a)}{\hat{m}(H_i, a_0)}, \qquad a_0 = 1. \tag{6}$$

Let $\tilde{\rho}_i^{\mathrm{fut}}(a)$ denote the future-period anchored target curve (estimated from held-out future data or its pilot surrogate). We allow a nonnegative scalar recalibration $\alpha_i \geq 0$ and define the per-sample stability loss as

$$\mathcal{L}_{\mathrm{stab},i} = \min_{\alpha_i \geq 0} \frac{1}{K} \sum_{k=1}^{K} \Big(\tilde{\rho}_i^{\mathrm{fut}}(a_k) - \alpha_i \hat{\rho}_i(a_k) \Big)^2. \tag{7}$$

The reported deployment stability loss is the average over the evaluation set,

$$\mathrm{StabLoss} = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_{\mathrm{stab},i}. \tag{8}$$

A lower value indicates that the learned shape is more transportable to future environments.

Offline policy value. Given a learned surface $\hat{m}(H, a)$, the induced pricing policy recommends

$$\hat{\pi}(H) = \arg\max_{a \in \mathcal{A}^\star} \hat{m}(H, a), \tag{9}$$

or the corresponding revenue/profit objective if the reward is not sales. We evaluate the policy on future data using a localized doubly robust estimator. Let R_i denote the observed reward, $\hat{q}(H_i, a)$ the nuisance outcome model, $\hat{g}(a \mid H_i)$ the estimated generalized propensity score, and $K_h(\cdot)$ a kernel centered at the recommended action. The value estimator is

$$\widehat{V}_{\mathrm{DR}}(\hat{\pi}) = \frac{1}{n} \sum_{i=1}^{n} \left[\hat{q}(H_i, \hat{\pi}(H_i)) + \frac{K_h(A_i - \hat{\pi}(H_i))}{\hat{g}(A_i \mid H_i)} \Big(R_i - \hat{q}(H_i, A_i) \Big) \right]. \tag{10}$$

Higher value indicates a better induced decision rule.

In the real-world study, RMSE measures factual fit, $\mathrm{StabLoss}$ measures future-time transportability of the learned response geometry, and $\widehat{V}_{\mathrm{DR}}$ measures the downstream quality of the induced pricing policy. Our primary emphasis is on the latter two metrics rather than predictive fit alone.

H.3 Public-data Evaluation on Dominick's

Data and protocol. To provide a reproducible complement to the proprietary Taobao study, we use the 11-brand orange-juice panel distributed with `bayesm` 3.1-7 [Montgomery, 1997], retaining store–brand–week records with `deal=0` and `feat=0`. The action is price per ounce divided by its median over the preceding 13 weeks, requiring at least eight observed weeks, with anchor $a_0 = 1$. Observed histories, factual prices, temporal environments, and missingness are retained; simulated outcomes are not fed back into subsequent histories. The semi-synthetic evaluation uses four expanding-window folds. The first uses weeks 53–91 for training, 92–104 for validation, and 105–117 for testing; subsequent folds advance validation and testing by 13 weeks. The 83 stores are partitioned into disjoint pilot, selection, and refit sets of 42, 21, and 20 stores, respectively.

The estimator uses brand-based and history-clustered representations with 11, 22, or 33 cells and a Gaussian action kernel. A 31-point grid, including a_0, is constructed from the intersection of training-period 5th–95th percentile action ranges and $[0.85, 1.15]$. Bandwidth $b \in \{0.04, 0.06, 0.08\}$ is selected on the chronological validation split. We use a representation penalty of 0.002 per log-cell and select the shape-ridge coefficient from $\{0, 10^{-3}, 10^{-2}\}$ on the same validation split. Nuisances are cross-fitted within pilot stores; selection uses validation data from selection stores, and the selected geometry is refitted using training and validation data from refit stores.

Semi-synthetic outcome construction. We generate only the outcome law, so counterfactual response curves are known while the observed temporal and treatment-assignment design is preserved. Let $u_1(H)$ and $u_2(H)$ be the standardized lagged 13-week mean sales and one-week-lagged mean competitor price, respectively. Let z_t be the standardized weekly market mean log-sales index from the original panel, where t denotes calendar week. Standardization constants and median imputation

Table 10: **Semi-synthetic evaluation on Dominick's.** Each entry reports mean true future NRegret (reported uncertainty in parentheses); lower is better. A and M denote exact additive and log-linked persistence. V averages over four levels of action-dependent shape drift.

Method	A	M	V
Ours	$\mathbf{0.123}$ (0.007)	$\mathbf{0.063}$ (0.003)	$\mathbf{0.075}$ (0.003)
Pooled	0.259 (0.008)	0.072 (0.005)	0.089 (0.004)
Time-aware pooled	0.248 (0.008)	0.067 (0.004)	0.091 (0.005)
VCNet	0.761 (0.015)	0.730 (0.012)	0.697 (0.009)
Causal Transformer	0.772 (0.013)	0.731 (0.012)	0.726 (0.009)

for missing competitor prices are fixed using weeks 53–91. With $x = x(a) = (a - 0.85)/0.30$, $x_0 = 0.5$, and $d = x - x_0$, the shared anchored shape is

$$\begin{aligned}\rho^\star(u, a) = &(0.35 + 0.25 \tanh u_1)d \\ &+ (0.18 + 0.10 \tanh u_2)[\sin(2\pi x) - \sin(2\pi x_0)] \\ &- (1.10 + 0.20 \tanh(u_1 - u_2))d^2.\end{aligned}$$

Define $\eta_t(H) = 0.35u_1 + 0.15u_2 + 0.60z_t$ and $s_t(H) = 0.60 + 1.40\,\sigma(0.80z_t + 0.25u_1)$, where $\sigma(v) = (1 + e^{-v})^{-1}$. The noiseless response surfaces are

$$\begin{aligned}m_t^A(H, a) &= 4.0 + 1.2 \tanh \eta_t(H) + s_t(H)\rho^\star(u(H), a), \\ m_t^M(H, a) &= \exp\{2.5 + 0.35 \tanh \eta_t(H) + s_t(H)\rho^\star(u(H), a)\}, \\ m_t^V(H, a) &= m_t^M(H, a) \exp\{\epsilon_{\text{drift}} q_t(u(H), a)\},\end{aligned}$$

where $q_t(u, a) = [\sin(2\pi t/13) + 0.5 \tanh u_1]d^2$ and $\epsilon_{\text{drift}} \in \{0, 0.25, 0.5, 1\}$. The noiseless surfaces in A and M satisfy exact persistence on the identity and log scales, respectively; V adds time-varying, action-dependent curvature on the log scale.

Observed outcomes are generated as

$$\begin{aligned}Y^A &= \max\{m_t^A(H, A) + e_A, 10^{-3}\}, \quad e_A \sim \mathcal{N}(0, \sigma_A^2), \\ Y^r &= m_t^r(H, A) \exp(e_L - 0.15^2/2), \quad e_L \sim \mathcal{N}(0, 0.15^2), \quad r \in \{M, V\},\end{aligned}$$

with $\sigma_A = 0.20\,\mathrm{SD}\{m_t^A(H, A)\}$ computed once on the earliest training block. Evaluation uses the noiseless response surfaces.

Counterfactual decision quality. For each test row i, let $m_i(a)$ denote its noiseless response and $\widehat{a}_i$ the recommended action on the fold-specific evaluation grid $\mathcal{A}_{\text{eval}}$. We compute

$$\text{NRegret} = \frac{1}{n} \sum_{i=1}^{n} \frac{\max_{a \in \mathcal{A}_{\text{eval}}} m_i(a) - m_i(\widehat{a}_i)}{\max_{a \in \mathcal{A}_{\text{eval}}} m_i(a) - \min_{a \in \mathcal{A}_{\text{eval}}} m_i(a) + 10^{-8}}.$$

A and M use 10 seeds across four folds; V uses five seeds across the same folds and four drift levels. Table 10 reports true future NRegret, with V averaged across drift levels. Ours achieves 0.123, 0.063, and 0.075 in A, M, and V, respectively, and has lower reported mean regret than pooled prediction, time-aware pooled prediction, VCNet, and Causal Transformer in all three settings.

Link-selection and persistence diagnostics. The exact regimes also allow us to assess the cost of selecting the response scale. In A, Ours has NRegret 0.1227, compared with 0.1223 for the fixed-correct identity link and 0.1258 for the fixed-wrong log link. In M, the corresponding values are 0.0633, 0.0629, and 0.0654, with the log link now fixed-correct. These comparisons indicate low adaptation cost relative to a learner supplied with the generating link.

On the original observed panel, we additionally evaluate next-week factual prediction and rolling persistence diagnostics. Across 78 weekly origins, a larger held-out profiled residual is associated with worse next-week geometry loss (Spearman correlation 0.37). We use this residual only as a prospective risk-ranking and falsification diagnostic, rather than as a certificate that persistence will hold in an unseen future period.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and Section 1 state the paper's scope as one-step continuous decision learning under temporal non-stationarity. The claims are matched to the identification result in Section 2, the anchored geometry target in Section 3, the finite-library estimator in Section 4, and the empirical evidence in Sections 5 and 6.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 8 discusses the main limitations, including reliance on one-step exchangeability and overlap, dependence on a prescribed morphology family, approximate geometry persistence, and the fact that the analysis does not cover long-horizon dynamic regimes or repeated-intervention feedback.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: The paper states the assumptions for within-period identification, anchored geometry persistence, and finite-library estimation in Sections 2–4. The main results are Proposition 2.2, Proposition 2.3, Proposition 3.1, Theorem 3.3, and Theorem 4.1; proofs and supporting lemmas are provided in Appendices C.1–E.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The synthetic experiments are specified in Section 5 and Appendices G.1–G, including regimes, splits, action grids, repetitions, metrics, and ablations. The proprietary real-world study cannot be exactly reproduced without platform access, but Section 6 and Appendix H describe the data construction, support filtering, chronological protocol, metrics, and aggregate results.

   Guidelines:

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

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [No]

   Justification: We provide the synthetic data generator and experimental implementation in the supplementary material or anonymized repository. However, the large-scale pricing logs and online A/B experiment data are proprietary platform data and cannot be publicly released. We therefore cannot provide open access to all data needed to reproduce every real-world result.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Section 5 specifies the synthetic regimes, temporal split, action grid, repetitions, metrics, and baselines, with full details in Appendix G. Section 6 and Appendix H specify the real-world data period, rolling-origin split, action definition, support filtering, baselines, and evaluation metrics.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: Synthetic results are reported with uncertainty over Monte Carlo repetitions, as shown in Table 1, Figure 3, and Appendix G. The online A/B experiment reports p-values for GMV, buy counts, and commission in Figure 4.

   Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [No]

   Justification: The synthetic experiments are lightweight and reproducible from the released implementation. However, the real-world experiments were executed on an internal company compute platform with abstracted GPU/CPU scheduling. Because exact hardware type, memory, runtime per job, and total compute are not reliably exposed to the authors, we do not provide a complete compute-resource report and avoid unverifiable estimates.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

Justification: The authors have reviewed the NeurIPS Code of Ethics and believe the submission conforms to it. The real-world results are reported only in aggregate form, and the paper does not release personally identifiable information or proprietary user-level logs.

Guidelines:

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

10. **Broader impacts**

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

Answer: [Yes]

Justification: Section 8 discusses the deployment scope and limitations of the method. The application domain includes pricing and advertising decisions, where unsupported or unstable deployment may affect consumers, merchants, and platform allocation; the paper therefore emphasizes support, overlap, and future-period validation.

Guidelines:

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

11. **Safeguards**

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

Answer: [N/A]

Justification: The paper does not release a pretrained model, generative system, scraped dataset, or other artifact with high misuse or dual-use risk.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: Existing methods and baselines are cited in the main text and related work. The paper does not redistribute third-party datasets or pretrained models. Proprietary platform data are used only for aggregate evaluation and are not released.

    Guidelines:

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

13. **New assets**

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

    Answer: [N/A]

    Justification: The paper does not release a new public dataset or pretrained model.

    Guidelines:

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

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

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

    Answer: [Yes]

    Justification: The online experiment was conducted within the platform's standard experimentation and privacy-review process. The submission reports only aggregate, de-identified metrics and does not disclose personally identifiable information or user-level records.

Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The submission reports only aggregate, de-identified results and does not disclose personally identifiable information.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: LLMs are not used as an important, original, or non-standard component of the proposed method, theory, experiments, or analysis.

    Guidelines:

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