Orthogonal Uplift Learning with Permutation-Invariant Representations for Combinatorial Treatments

Jiacan Gao[4*] **Xinyan Su**[3*] **Tianqi Gu**[3*] **Mingyuan Ma**[5]
Xiao Xu[3] **Xinrui Wan**[3] **Enyun Yu**[3] **Jiecheng Guo**[3] **Zhiheng Zhang**[1,2†]

[1]School of Statistics and Data Science, Shanghai University of Finance and Economics, Shanghai, China [2]Institute of Big Data Research, Shanghai University of Finance and Economics, Shanghai, China [3]Didi Chuxing, Beijing, China [4]School of Statistics, East China Normal University, Shanghai, China [5]School of Mathematics and Statistics, Beijing Jiaotong University, Beijing, China

Abstract

We study uplift estimation for *combinatorial* treatments. Uplift measures the pure incremental causal effect of an intervention (e.g., sending a coupon or a marketing message) on user behavior, modeled as a conditional individual treatment effect. Many real-world interventions are *combinatorial*: a treatment is a policy that specifies context-dependent action distributions rather than a single atomic label. Although recent work considers structured treatments, most methods rely on categorical or opaque encodings, limiting robustness and generalization to rare or newly deployed policies. We propose an uplift estimation framework that aligns treatment representation with causal semantics. Each policy is represented by the mixture it induces over context–action components and embedded via a permutation-invariant aggregation. This representation is integrated into an orthogonalized low-rank uplift model, extending Robinson-style decompositions to learned, vector-valued treatments. We show that the resulting estimator is expressive for policy-induced causal effects, orthogonally robust to nuisance estimation errors, and stable under small policy perturbations. Industrial experiments show improved uplift ranking, while synthetic experiments demonstrate more accurate effect estimation for rare and held-out-near policies.

1 Introduction

Estimating individualized causal effects is central to data-driven decision-making, as it enables practitioners to compare alternative interventions and select those with the largest expected incremental outcomes. This problem is commonly formalized through the conditional average treatment effect (CATE), and is rooted in the potential-outcomes framework and the study of heterogeneous treatment effects [Rubin, 1974, Imbens and Rubin, 2015, Athey and Imbens, 2016]. A large literature has developed CATE estimators for binary or low-cardinality treatments, including meta-learners [Künzel et al., 2019], doubly robust estimators [Kennedy, 2023], and orthogonalized approaches based on double/debiased machine learning [Chernozhukov et al., 2018]. These methods have been successful in classical A/B testing and personalization tasks, where treatments can often be represented as a small set of discrete alternatives.

*Jiacan Gao, Xinyan Su, and Tianqi Gu contributed equally. Jiacan Gao and Xinyan Su are listed in alphabetical order.
†Corresponding author: zhangzhiheng@mail.shufe.edu.cn.

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

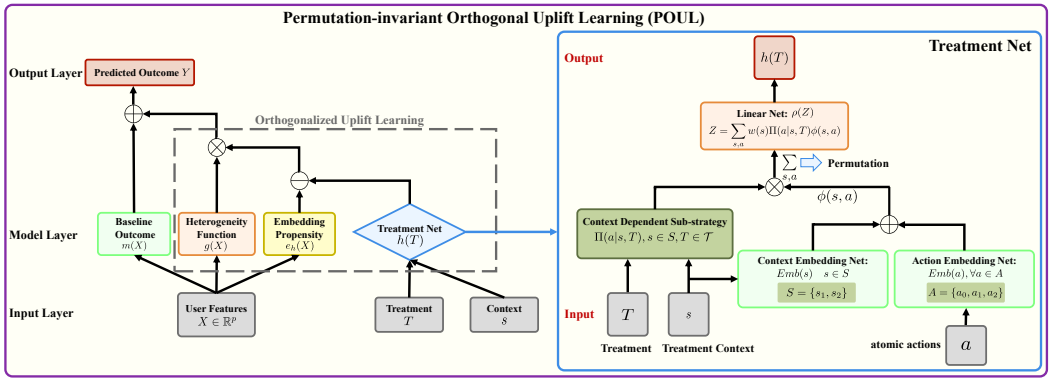


Figure 1: **Treatment Net: permutation-invariant embedding of combinatorial treatments.** A treatment T is observed through its policy specification: for each context $s \in S$ (instantiated here as $\{s_1, s_2\}$), the treatment induces a distribution over atomic actions $a \in \mathcal{A}$ (instantiated as $\{a_0, a_1, a_2\}$). The network first learns embeddings for contexts and actions to form atom representations $\phi(s, a)$. These atoms are then reweighted by policy probabilities and aggregated via a permutation-invariant sum $z(T)$, ensuring the representation depends only on the induced mixture. Finally, a learnable map produces the treatment embedding $h(T)$, which is used by the orthogonalized uplift model (left panel) to estimate the causal effect $\tau(x)$.

In many modern applications, however, treatments are no longer isolated atomic actions. Instead, they often take the form of policies that specify how actions are assigned across multiple treatment-side contexts, such as user segments, locations, or workflow stages. For example, an online platform may deploy a policy that assigns different incentive levels under different spatiotemporal demand conditions. Such interventions induce a structured or combinatorial treatment space, where high-level treatments are composed of reusable context–action components. Related work has studied complex treatment settings through learned representations, including time-varying or longitudinal treatment regimes [Bica et al., 2020] and graph-structured treatment representations [Harada and Kashima, 2021]. These works highlight the value of moving beyond purely categorical treatment labels, but they do not directly target our setting, where each treatment is a context-dependent policy and where the representation should respect policy-level compositional symmetries.

A key challenge in combinatorial treatment settings is that the statistical representation of treatments often fails to align with their causal meaning. In practice, policies are frequently iterated and implemented by different teams, leading to variations in indexing, ordering, or internal identifiers. Naively encoding each policy as a one-hot label treats such variants as unrelated, even when they induce identical or nearly identical interventions or differ only by minor local modifications. This mismatch has two critical consequences. First, it prevents statistical strength from being shared across related policies, exacerbating variance in long-tailed treatment regimes. Second, it destroys any meaningful notion of proximity between treatments, making it impossible to reason about smooth policy perturbations or to generalize to new strategies. These issues arise simultaneously in theory and practice, and cannot be resolved by simply increasing model capacity.

Motivated by this observation, we view each policy as inducing a probability measure over context–action atoms, and posit that causal effects depend on the induced mixture rather than on arbitrary policy identifiers or orderings. This perspective suggests that treatment representations should be invariant to re-indexing and ordering, while still allowing parameter sharing across reusable components. This design principle is closely related to permutation-invariant set representations [Zaheer et al., 2017] and, more broadly, to invariant and equivariant learning, where respecting known symmetries is essential for generalization [Kondor and Trivedi, 2018].

Building on this principle, we propose a framework for uplift estimation under combinatorial treatments. We first construct a permutation-invariant treatment embedding by aggregating learned context–action representations through a commutative sum, ensuring invariance to policy re-indexing and ordering while sharing information across related policies. We then combine this embedding with an orthogonalized low-rank uplift model, following the semiparametric and orthogonalization

tradition [Robins et al., 1994, Chernozhukov et al., 2018], so that nuisance estimation errors affect the target CATE estimator only through higher-order terms.

Our main theoretical contribution is not merely that a permutation-invariant network can approximate functions on a finite policy simplex. Rather, we show that the induced policy geometry changes the statistical problem itself. Classical orthogonal CATE learning requires exact support for each treatment whose effect is to be estimated. This requirement is misaligned with combinatorial policy systems, where exact policies are long-tailed and new policies are routinely created by small local edits. We prove that, under policy-induced mixture sufficiency and a mild smoothness condition in the natural policy distance, exact policy overlap can be replaced by *local policy overlap*: an unseen policy can be estimated with an error equal to the usual orthogonal estimation error on supported policies plus a transfer bias determined by its distance to observed policies. We further prove a matching minimax lower bound, showing that this bias is not an artifact of POUL but an information-theoretic price of cold-start policy estimation. Thus the role of the permutation-invariant representation is not cosmetic; it supplies the geometry in which causal information can be transferred across combinatorial policies.

Our contributions are therefore threefold. First, we formalize uplift estimation when each treatment is a context-dependent policy inducing a probability measure over reusable context–action atoms. Second, we propose a permutation-invariant policy representation that is aligned with this causal interface and can be integrated into an orthogonalized low-rank uplift model. Third, we establish a local policy-overlap transfer theorem and a matching minimax lower bound, showing both when and why rare or unseen policies are estimable from nearby observed policies. Empirically, we validate these claims on large-scale randomized platform data and on synthetic combinatorial-policy benchmarks with held-out policies.

2 Problem Setup and Method

We have i.i.d. data $\mathcal{D} = \{(X_i, T_i, Y_i)\}_{i=1}^n$, where $X_i \in \mathbb{R}^p$ are unit covariates, $T_i \in \mathcal{T}$ is a structured treatment (policy) drawn from a finite (possibly large) set, and $Y_i \in \mathbb{R}$ is the outcome. Let $Y(t)$ denote the potential outcome under treatment $t \in \mathcal{T}$. For any x and any pair (t_1, t_0), we define the individualized uplift as the conditional treatment effect

$$\tau(x; t_1, t_0) := \mathbb{E}[Y(t_1) - Y(t_0) \mid X = x]. \tag{1}$$

At the unit level, the (unobservable) individual treatment effect is $\tau_i(t_1, t_0) := Y_i(t_1) - Y_i(t_0)$. Throughout, $\tau(x; t_1, t_0)$ is the estimand (the standard CATE); in experiments we refer to the pointwise prediction $\hat{\tau}(X_i; t_1, t_0)$ as an "ITE" score for brevity.

Assumption 2.1. (SUTVA for RCT) Our main setting is randomized assignment. We assume (i) consistency/SUTVA, (ii) randomized treatment assignment conditional on X, and (iii) overlap. Under these conditions, $\tau(x; t_1, t_0)$ is identified from $\mathcal{D}$.

2.1 Structured Treatment Representation

Treatment-side contexts and atomic actions. We assume each policy operates through a finite set of treatment-side contexts $\mathcal{S}$ and a finite set of atomic actions $\mathcal{A}$. A policy $t \in \mathcal{T}$ is specified by a collection of context-dependent sub-strategies $\{\Pi_t(\cdot \mid s)\}_{s \in \mathcal{S}}$, where each sub-strategy $\Pi_t(\cdot \mid s)$ is a distribution over actions in $\mathcal{A}$. We treat $\{\Pi_t(\cdot \mid s)\}_{s \in \mathcal{S}}$ as the observable specification of a structured treatment. We also allow nonnegative context weights $w(s) \geq 0$ that capture the relative exposure frequency or importance of context s (e.g., segment prevalence), $w(\cdot)$ can be fixed from logs, domain knowledge or estimated. Let $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ be an embedding of a context–action atom and let $\rho : \mathbb{R}^d \to \mathbb{R}^d$ be a learnable map. We define an intermediate aggregated representation $h(t) \in \mathbb{R}^d$:

$$z(t) = \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} w(s)\, \Pi_t(a \mid s)\, \phi(s, a), \tag{2}$$

$$h(t) = \rho(z(t))\,. \tag{3}$$

This construction is invariant to any re-indexing or re-ordering of the underlying context–action components, and enables parameter sharing across policies through the shared atom embedding ϕ.

Implementation of the treatment network. In practice, $\phi(s, a)$ is parameterized as a lookup table over (s, a) (or a sum of separate embeddings for s and a), and $\rho(\cdot)$ is a small MLP (optionally with normalization). The cost of computing $h(t)$ is $O(|\mathcal{S}||\mathcal{A}|d)$ per policy if done directly, and can be reduced by caching $\{\phi(s, a)\}$ and exploiting sparsity in $\Pi_t(\cdot \mid s)$.

2.2 Orthogonalized Factorized Uplift Model

Model class. We consider Neyman decomposition following [Kaddour et al., 2021]:

$$\begin{aligned} Y &= m(X) + g(X)^\top \big(h(T) - e_h(X)\big) + \varepsilon, \\ e_h(X) &:= \mathbb{E}[h(T) \mid X], \end{aligned} \tag{4}$$

where $m : \mathbb{R}^p \to \mathbb{R}$ is a baseline outcome regression model and $g : \mathbb{R}^p \to \mathbb{R}^d$ captures treatment effect heterogeneity. Under (4), the implied uplift is

$$\tau(x; t_1, t_0) = g(x)^\top \big(h(t_1) - h(t_0)\big). \tag{5}$$

We refer to this as "factorized" (or rank-d) since the effect depends on (x, t) only through the inner product between $g(x)$ and $h(t)$.

Learning objective. Let m_θ, g_η, ϕ_ω, and ρ_ν parameterize m, g, ϕ, ρ, and define $h_{\omega,\nu}$ by (2)–(3). We estimate parameters by empirical risk minimization:

$$\min_{\theta,\eta,\omega,\nu} \frac{1}{n} \sum_{i=1}^{n} \ell\big(Y_i,\ m_\theta(X_i) + g_\eta(X_i)^\top \big(h_{\omega,\nu}(T_i) - \hat{e}_h(X_i)\big)\big) + \lambda\, \Omega(\theta, \eta, \omega, \nu), \tag{6}$$

where ℓ is a prediction loss (e.g., squared loss or negative log-likelihood) and Ω is a regularizer.

Computing $\hat{e}_h(x)$. A key term in (4) is $e_h(x) = \mathbb{E}[h(T) \mid X = x]$, which we approximate by $\hat{e}_h(x)$. Let the propensity of treatment t be $\pi_0(t \mid x) := \mathbb{P}(T = t \mid X = x)$. Two cases are considered. 1) **RCT Study.** In the common case of complete randomization, treatment assignment is independent of covariates, i.e., $T \perp X$. Then $e_h(x)$ is a constant vector: $e_h(x) = \mathbb{E}[h(T) \mid X = x] = \mathbb{E}[h(T)]$.

Accordingly, we estimate it by the empirical mean of $h(T)$ over the experiment data (or by a running mean in mini-batch SGD, as in Algorithm 1). More generally, under stratified or covariate-adaptive randomization, $\pi_0(t \mid x)$ is known and may depend on x through the stratum; in that case we set $\hat{e}_h(x) = \sum_{t \in \mathcal{T}} \pi_0(t \mid x)\, h(t)$. 2) **Observational Study.** When treatment assignment is not randomized, $e_h(x)$ generally varies with x and $\pi_0(t \mid x)$ is unknown. We estimate a propensity model $\hat{\pi}(t \mid x) \approx \pi_0(t \mid x)$ and plug it in: $\hat{e}_h(x) = \sum_{t \in \mathcal{T}} \hat{\pi}(t \mid x)\, h(t)$. When $|\mathcal{T}|$ is large, the sum can be approximated by sampling treatments from $\hat{\pi}(\cdot \mid x)$.

Training Procedure We train POUL via a two-stage orthogonal learning procedure (Algorithm 1). In stage 1 (lines 1-3), we fit a base predictor $m(x) \approx \mathbb{E}[Y \mid X = x]$ using all available data, which can be a DNN network, or Dragonet [Shi et al., 2019]. This stage leverages the largest possible sample to obtain a stable estimate of the main outcome signal, independent of treatment structure. After training, we treat $\hat{m}(\cdot)$ as fixed in the second stage. In stage 2 (lines 4-18), we learn the heterogeneous uplift component on the experiment population where policies are actually assigned and observed. Given a policy t, we compute its treatment embedding $h(t)$ by aggregating context–action atoms induced by its context-specific *sub-strategies* $\{\Pi_t(\cdot \mid s)\}_{s \in \mathcal{S}}$ and context weights $w(s)$: we first form the mixture representation $z(t) = \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} w(s) \Pi_t(a \mid s) \phi(s, a)$ and then apply a learnable map $h(t) = \rho(z(t))$ (line 10). The aggregation is *permutation-invariant* to any re-indexing or re-ordering of contexts/actions, ensuring that two semantically identical policies (i.e., inducing the same mixture over atoms) have the same representation and preventing spurious dependence on arbitrary IDs or ordering. We fit the heterogeneity function $g(x)$ and the treatment embedding parameters by minimizing an orthogonalized loss based on the residualized outcome $Y - \hat{m}(X)$ and the residualized treatment representation $h(T) - \hat{e}_h$. In our main RCT setting with complete randomization, treatment assignment is independent of covariates ($T \perp X$), so we estimate e_h by the empirical mean of $h(T)$ over the experiment population (or an equivalent running estimate during training). This yields the final uplift estimator $\hat{\tau}(x; t_1, t_0) = \hat{g}(x)^\top \big(\hat{h}(t_1) - \hat{h}(t_0)\big)$.

Algorithm 1 Permutation-invariant Orthogonal Uplift Learning (POUL)

Require: Full data $\mathcal{D}_{\text{all}} = \{(X_i, Y_i)\}$; Experiment data $\mathcal{D}_{\text{exp}} = \{(X_i, T_i, Y_i)\}$; treatment-side contexts $\mathcal{S}$; atomic actions $\mathcal{A}$; Policy specification $\{\Pi_t(a \mid s)\}_{t\in\mathcal{T},\, s\in\mathcal{S},\, a\in\mathcal{A}}$; weights $w(\cdot)$; loss ℓ; regularizer Ω and weight λ; learning rate γ

Stage 1: fit outcome regression model.
Initialize m_θ and train m_θ on $\mathcal{D}_{\text{all}}$ by minimizing $\frac{1}{|\mathcal{D}_{\text{all}}|}\sum \ell\big(m_\theta(X_i), Y_i\big)$
Freeze $\hat{m}(\cdot) \leftarrow m_\theta(\cdot)$
Stage 2: train orthogonalized uplift model.
Initialize (η, ω, ν) for g_η, ϕ_ω, ρ_ν
Initialize running mean $\bar{h} \in \mathbb{R}^d \leftarrow 0$, counter $c \leftarrow 0$
for each SGD step **do**
  Sample a mini-batch $\mathcal{B} \subset \mathcal{D}_{\text{exp}}$
  for each $(X_i, T_i, Y_i) \in \mathcal{B}$ **do**
    Compute treatment embedding (permutation-invariant aggregation):
      $z_i \leftarrow \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} w(s)\, \Pi_{T_i}(a \mid s)\, \phi_\omega(s, a), \quad h_i \leftarrow \rho_\nu(z_i)$
    Compute two nets: $\hat{m}_i \leftarrow \hat{m}(X_i), \quad g_i \leftarrow g_\eta(X_i)$
    Update $\hat{e}_h(X_i) \leftarrow \bar{h}$ (replace this term by the corresponding cases described in Sec. 2.2)
    Predict: $\hat{Y}_i \leftarrow \hat{m}_i + g_i^\top\big(h_i - \hat{e}_h(X_i)\big)$
  end for
  Get loss: $\mathcal{L} \leftarrow \frac{1}{|\mathcal{B}|}\sum_{(X_i,T_i,Y_i)\in\mathcal{B}} \ell(\hat{Y}_i, Y_i) + \lambda\, \Omega(\eta, \omega, \nu)$
  SGD update: $(\eta, \omega, \nu) \leftarrow (\eta, \omega, \nu) - \gamma \nabla \mathcal{L}$
  Update $\bar{h}$ using current batch embeddings: $\bar{h} \leftarrow \frac{c\,\bar{h} + \sum_{(X_i,T_i,Y_i)\in\mathcal{B}} h_i}{c + |\mathcal{B}|}, \quad c \leftarrow c + |\mathcal{B}|$
end for
return $\hat{\tau}(x; t_1, t_0) = g_{\hat{\eta}}(x)^\top\big(h_{\hat{\omega},\hat{\nu}}(t_1) - h_{\hat{\omega},\hat{\nu}}(t_0)\big)$

2.3 Technical Challenges

Semantic invariance under re-indexing of structured policies. In practice, a policy is specified by a collection of context-specific *sub-strategies* $\{\Pi_t(\cdot \mid s)\}_{s\in\mathcal{S}}$, where the indexing and ordering of contexts/actions can be arbitrary and may change across deployments. A treatment representation that depends on such ordering can introduce spurious differences and hurt robustness when policies are re-serialized, updated, or composed from the same building blocks. Our mixture-based construction in (2) is permutation-invariant, so the embedding depends only on the induced mixture over context–action atoms, aligning representation with causal semantics.

Long-tailed policies, local overlap, and cold-start variants. In large combinatorial treatment spaces, exact overlap over complete policies is unrealistic: many policies are rare, newly created, or observed only in a few strata. Treating each policy as a categorical label then creates an identification–generalization mismatch: orthogonalization can de-bias nuisance estimation on supported treatments, but it cannot by itself define how information should be transferred to an unsupported policy ID. POUL addresses this problem by replacing raw policy IDs with the policy-induced mixture over reusable context–action atoms. This induces a metric in which a cold-start policy can be close to observed policies even when its exact ID has zero support. Our theory formalizes this as *local policy overlap*: the error for an unseen policy is controlled by the supported-policy orthogonal estimation error plus a transfer bias determined by its distance to observed policy support.

Noisy weights and scalable computation of mixture embeddings. Context weights $w(s)$ may be noisy proxies of exposure frequency or importance, and directly computing (2) scales with $|\mathcal{S}||\mathcal{A}|$ per policy. In practice, $w(\cdot)$ can be estimated from logs or domain knowledge, and we empirically assess robustness under imperfect weights. For efficiency, we cache $\phi(s, a)$ and exploit sparsity in $\Pi_t(\cdot \mid s)$ to reduce computation, making training feasible at scale.

3 Theoretical Analysis

Our analysis follows the orthogonal-learning perspective of Chernozhukov et al. [2018], Foster and Syrgkanis [2023], but differs in that the main object is not a scalar treatment residual; it is

a learned, policy-induced treatment geometry that enables local support transfer. Combinatorial policies induce high-dimensional treatments: each $t \in \mathcal{T}$ specifies a collection of context-wise action distributions $\{\Pi_t(\cdot \mid s)\}_{s \in S}$. The main statistical difficulty is not only nuisance estimation. It is that the target policy may be rare or entirely absent as an exact treatment ID, even though it is close to observed policies in its context–action specification. A theory that assumes global exact-policy overlap therefore misses the central phenomenon motivating this work.

Our analysis separates four issues. First, we show that the policy-induced mixture is a sufficient causal interface: if outcomes depend on a policy through the probability measure it induces over context–action atoms, then the orthogonalized low-rank model is a reparametrization of the causal conditional mean. Second, we use cross-fitted orthogonal learning to control estimation error on the set of policies with exact support. Third, and most importantly, we prove a *local policy-overlap transfer theorem*: the error for a rare or unseen policy is controlled by the supported-policy orthogonal error plus a transfer term depending only on the policy's distance to observed support. Fourth, we prove a matching lower bound, showing that this transfer term is minimax unavoidable. Expressiveness and Lipschitz stability of the permutation-invariant embedding then serve as supporting lemmas rather than the main theoretical claim.

Recall the context weights $w(s) \geq 0$ with $\sum_{s \in \mathcal{S}} w(s) = 1$ and define

$$\alpha_t(s,a) \;:=\; w(s)\,\Pi_t(a \mid s), \qquad (s,a) \in \mathcal{S} \times \mathcal{A}. \tag{7}$$

The collection $\{\alpha_t(s,a)\}_{s,a}$ is a probability mass function over $\mathcal{S} \times \mathcal{A}$. We denote by μ_t the corresponding discrete probability measure on $\mathcal{S} \times \mathcal{A}$, i.e., $\mu_t(\{(s,a)\}) = \alpha_t(s,a)$.

Assumption 3.1 (Policy-induced mixture sufficiency)**.** There exist measurable functions $m_0 : \mathbb{R}^p \to \mathbb{R}$, $g_0 : \mathbb{R}^p \to \mathbb{R}^d$, and a (possibly unknown) functional F mapping measures on $\mathcal{S} \times \mathcal{A}$ to $\mathbb{R}^d$ such that for all $x \in \mathbb{R}^p$ and all $t \in \mathcal{T}$,

$$\mathbb{E}[Y(t) \mid X = x] \;=\; m_0(x) + g_0(x)^\top F(\mu_t). \tag{8}$$

Moreover, F is continuous on $\{\mu_t : t \in \mathcal{T}\}$ under the ℓ_1/total-variation topology, i.e., if $\|\mu - \mu'\|_1 \to 0$ then $\|F(\mu) - F(\mu')\| \to 0$.

Assumption 3.1 says that a policy affects outcomes only through the *mixture it induces* over context–action atoms. This matches the reality of many platform interventions. For example, in ride-hailing marketplaces, a policy may specify (for each spatiotemporal context s) a distribution over incentive actions a (e.g., bonus levels or pricing multipliers). The realized user experience aggregates over contexts according to exposure weights $w(s)$; thus it is the induced mixture μ_t—not an arbitrary index ordering of contexts/actions—that governs the causal response. Also, Assumption 3.1 implies our orthogonalized model class (4) is well-specified as follows (up to reparametrization).

Proposition 3.2 (Well-specified orthogonalized low-rank representation)**.** *Under Assumption 3.1 and consistency $Y = Y(T)$, define the* oracle *treatment representation $h_0(t) := F(\mu_t)$ and $e_0(x) := \mathbb{E}[h_0(T) \mid X = x]$. Then there exists a function $m(\cdot)$ such that the conditional mean satisfies*

$$\begin{aligned} \mathbb{E}[Y \mid X, T] &= m(X) + g_0(X)^\top \big(h_0(T) - e_0(X)\big), \\ m(x) &:= m_0(x) + g_0(x)^\top e_0(x), \end{aligned} \tag{9}$$

and the CATE defined in (1) *obeys*

$$\tau(x; t_1, t_0) = g_0(x)^\top \big(h_0(t_1) - h_0(t_0)\big). \tag{10}$$

Proposition 3.2 clarifies a key boundary: the orthogonalized decomposition (4) is *not* an additional restriction beyond (8); it is a reparametrization that is convenient for debiased/orthogonal learning. In particular, the causal estimand depends on (g_0, h_0) but not on the nuisance pair (m, e_0).

3.1 Expressiveness of the Permutation-Invariant Policy Embedding

We now show that the embedding family (2)–(3) is expressive enough to approximate any continuous functional $F(\mu_t)$ of the policy-induced mixture. Given maps $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^r$ and $\rho : \mathbb{R}^r \to \mathbb{R}^d$, define

$$z_\phi(t) = \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \alpha_t(s,a)\,\phi(s,a),\; h_{\phi,\rho}(t) = \rho(z_\phi(t))\,. \tag{11}$$

This construction is permutation-invariant to any re-indexing of $\mathcal{S} \times \mathcal{A}$ because it depends on $\{(s, a, \alpha_t(s,a))\}$ only through a weighted sum.

Lemma 3.3 (Expressiveness of permutation-invariant embeddings)**.** *Under Assumption 3.1, conditioning that $\mathcal{S}$ and $\mathcal{A}$ are finite and let $\mathcal{M} := \{\mu_t : t \in \mathcal{T}\} \subset \Delta(\mathcal{S} \times \mathcal{A})$. Let $F : \mathcal{M} \to \mathbb{R}^d$ be continuous. Then for any $\varepsilon > 0$, there exist an integer r and functions $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^r$ and $\rho : \mathbb{R}^r \to \mathbb{R}^d$ (such that ρ can be implemented by a standard universal approximator, e.g., a ReLU MLP) for which*

$$\sup_{t \in \mathcal{T}} \big\| F(\mu_t) - h_{\phi,\rho}(t) \big\| \;\le\; \varepsilon. \tag{12}$$

Theorem 3.3 justifies the representation choice in (2)–(3): *if* the causal response of a policy depends on t only through the induced mixture over context–action atoms, *then* our embedding family can approximate that dependence arbitrarily well. In other words, the proposed $h(t)$ is a principled "policy featurization" rather than a heuristic encoding. For example, consider a policy that assigns high incentives in peak Beijing and zero incentives in off-peak Shanghai. Two engineering implementations may store this policy using different context orderings or action identifiers. A categorical policy-ID representation would treat them as unrelated treatments, while our permutation-invariant embedding maps both specifications to the same representation, as they induce the same context–action mixture. Consequently, the implied uplift is identical, as required by causal semantics. Noteworthy, such a representation capacity statement does *not* claim: (i) causal identification without Assumption 3.1 and the usual causal conditions (SUTVA/overlap), (ii) that the required dimension r is always small, or (iii) that optimization will always find the approximating parameters. These are separate issues (identification, statistical rates, and optimization) from expressiveness. Also, the skepticism arises that *"If $\mathcal{S} \times \mathcal{A}$ is finite, isn't this just a fancy way of learning $F(\alpha)$ with an MLP? Why do we need the embedding structure at all?"* It highlights what our contribution *really* is: we are not merely encoding α_t; we are enforcing a *compositional inductive bias* (shared atom embeddings $\phi(s, a)$ and permutation-invariant aggregation). This bias is what enables (i) parameter sharing across policies, (ii) meaningful continuity in policy space (formalized in Proposition C.7), and (iii) practical scaling when $\mathcal{T}$ is huge but $\mathcal{S} \times \mathcal{A}$ has structure. Having established that the embedding family can represent the causal policy effect under Assumption 3.1, we now turn to estimation: how does orthogonalization protect uplift learning from nuisance estimation errors?

3.2 Orthogonal Robustness

Let $\pi_0(t \mid x) := \mathbb{P}(T = t \mid X = x)$ be the (known or estimable) assignment/propensity. Let h_0 be the oracle representation from Proposition 3.2 and define $e_0(x) := \mathbb{E}[h_0(T) \mid X = x] = \sum_{t \in \mathcal{T}} \pi_0(t \mid x)\, h_0(t)$. The orthogonalized regression form (4) suggests learning $g(\cdot)$ and $h(\cdot)$ from residualized signals $(Y - m(X))$ and $(h(T) - e_h(X))$. For theory, we analyze the standard cross-fitted variant that isolates nuisance estimation. (In experimental design with known π_0, the propensity step is omitted; the same analysis applies with smaller nuisance error.) Let $\{I_k\}_{k=1}^K$ be a partition of $\{1, \dots, n\}$.

Define the *Neyman orthogonality* score $\psi(W; g, h, m, e) := \Big(Y - m(X) - g(X)^\top (h(T) - e(X))\Big)\,(h(T) - e(X))$ where $W := (X, T, Y)$. We recall its first-order invariance as follows.

Lemma 3.4 (Neyman orthogonality w.r.t. nuisance (m, e))**.** *Assume the model* (4) *holds with some (m_0, g_0, h_0, e_0) and $\mathbb{E}[\varepsilon \mid X, T] = 0$, where $e_0(x) = \mathbb{E}[h_0(T) \mid X = x]$. Then $\mathbb{E}\big[\psi(W; g_0, h_0, m_0, e_0)\big] = 0$, and the Gateaux derivative of $\mathbb{E}[\psi(W; g_0, h_0, m, e)]$ in the direction of perturbations $(\delta m, \delta e)$ vanishes at $(m, e) = (m_0, e_0)$: $\frac{d}{dr}\big|_{r=0} \mathbb{E}\Big[\psi(W; g_0, h_0, m_0 + r\,\delta m,\, e_0 + r\,\delta e)\Big] = 0$.*

Lemma 3.4 is the formal statement of "nuisance errors do not affect the target to first order." It is exactly the property exploited by debiased/double machine learning. Additionally, considering its second-order, we now state a pointwise robustness guarantee in the spirit of DML error expansions. Below, for a (vector-valued) function f we use $\|f\|_2 := (\mathbb{E}\|f(W)\|^2)^{1/2}$, and for $h(\cdot)$ we use $\|h - \tilde{h}\|_\infty := \sup_{t \in \mathcal{T}} \|h(t) - \tilde{h}(t)\|$.

Assumption 3.5 (Regularity for orthogonal robustness)**.** For a covariate value x, define the exact-support set $\mathcal{T}_x^\circ := \{t \in \mathcal{T} : \pi_0(t \mid x) > 0\}$.

1. (Bounded moments) $\mathbb{E}[Y^2] < \infty$ and $\sup_{t \in \mathcal{T}} \|h_0(t)\| < \infty$, and $\mathbb{E}[\|g_0(X)\|^2] < \infty$.

2. (Supported-policy overlap) For policies in the exact-support set $\mathcal{T}_x^\circ$, there exists $c > 0$ such that $\pi_0(t \mid x) \geq c,\ t \in \mathcal{T}_x^\circ$. Policies outside $\mathcal{T}_x^\circ$ are not assumed to have exact overlap; their effects are handled by the local policy-overlap transfer result below.

3. (Cross-fitting) Nuisance estimates $(\hat{m}^{(-k)}, \hat{\pi}^{(-k)})$ are fit on data independent of fold I_k.

4. (Estimation errors) Define $\delta_m := \|\hat{m} - m_0\|_2$, $\delta_e := \|\hat{e} - e_0\|_2$, $\delta_g(x) := \|\hat{g}(x) - g_0(x)\|$, and $\delta_h := \|\hat{h} - h_0\|_\infty$. Assume these quantities are $o_p(1)$.

Theorem 3.6 (Supported-policy orthogonal robustness)**.** *Suppose Assumption 3.5 holds. Consider the cross-fitted orthogonal variant of POUL, where the nuisance estimates are trained on folds independent of the evaluation fold. For any fixed x and any supported pair $(t_1, t_0) \in \mathcal{T}_x^\circ \times \mathcal{T}_x^\circ$,*

$$\widehat{\tau}(x; t_1, t_0) - \tau(x; t_1, t_0) = \mathrm{error}_g(x; t_1, t_0) + \mathrm{error}_h(x; t_1, t_0) + O_p\left((\delta_m + \delta_h + \sup_x \delta_g(x))\delta_e\right),$$

where $\mathrm{error}_g(x; t_1, t_0) := (\widehat{g}(x) - g_0(x))^\top \{h_0(t_1) - h_0(t_0)\}$, *and* $\mathrm{error}_h(x; t_1, t_0) := g_0(x)^\top \{(\widehat{h} - h_0)(t_1) - (\widehat{h} - h_0)(t_0)\}$. *Thus, on the exact-support set, the nuisance pair (m, e) affects the CATE only through a second-order product term.*

Theorem 3.6 explains what orthogonalization contributes before any cold-start transfer is invoked. On policies with exact support, nuisance errors in baseline regression and embedding propensity do not linearly bias the CATE estimator; they enter only through a product term. For example, if $\delta_m = o_p(n^{-1/4})$ and $\delta_e = o_p(n^{-1/4})$, the nuisance contribution is $o_p(n^{-1/2})$, mirroring the classical DML regime. The theorem deliberately does not claim that unsupported policies are identified by orthogonalization alone. That additional step is provided by the local policy-overlap theorem below.

The supported-policy result leaves open the central long-tailed question: what can be said about a target policy whose exact ID has little or no support? The next result shows that exact policy overlap can be replaced by local policy overlap once the policy-induced mixture geometry is available.

3.3 Local Policy-Overlap Transfer

Define the weighted policy distance

$$d_\Pi(t, t') := \sum_{s \in S} w(s) \big\| \Pi_t(\cdot \mid s) - \Pi_{t'}(\cdot \mid s) \big\|_1 = \|\mu_t - \mu_{t'}\|_1. \tag{13}$$

For any target policy t, define its local support radius at covariate value x by $r_x(t) := \inf_{u \in \mathcal{T}_x^\circ} d_\Pi(t, u)$. When $T \perp X$, as in a completely randomized experiment, we write $r(t)$ for the corresponding covariate-independent radius.

Assumption 3.7 (Policy smoothness)**.** There exist constants $L_F < \infty$, $G < \infty$, and $\beta \in (0, 1]$ such that

$$\|F(\mu_t) - F(\mu_{t'})\| \leq L_F d_\Pi(t, t')^\beta, \qquad \|g_0(x)\| \leq G. \tag{14}$$

The fitted treatment representation is regularized to satisfy, on the policy class of interest,

$$\|\widehat{h}(t) - \widehat{h}(t')\| \leq L_{\widehat{h}} d_\Pi(t, t')^\beta, \qquad \|\widehat{g}(x)\| \leq \widehat{G}. \tag{15}$$

Theorem 3.8 (Local policy-overlap transfer)**.** *Fix x and two target policies $t_1, t_0 \in \mathcal{T}$, which may have zero exact propensity. Let*

$$\varepsilon_n(x) := \sup_{u_1, u_0 \in \mathcal{T}_x^\circ} \big|\widehat{\tau}(x; u_1, u_0) - \tau(x; u_1, u_0)\big| \tag{16}$$

be the uniform CATE error of the cross-fitted orthogonal learner on supported policies. Under Assumption 3.7,

$$\big|\widehat{\tau}(x; t_1, t_0) - \tau(x; t_1, t_0)\big| \leq \varepsilon_n(x) + \big(\widehat{G} L_{\widehat{h}} + G L_F\big)\big\{r_x(t_1)^\beta + r_x(t_0)^\beta\big\}. \tag{17}$$

Consequently, exact policy-level overlap is sufficient but not necessary for consistent policy-effect estimation. If $\varepsilon_n(x) = o_p(1)$ and $r_x(t_j) = o_p(1)$ for $j = 0, 1$, then $\widehat{\tau}(x; t_1, t_0) - \tau(x; t_1, t_0) = o_p(1)$, even when $\pi_0(t_j \mid x) = 0$ for one or both target policies.

Theorem 3.8 is the central reason why a policy-aware representation is useful in the long-tailed regime. Standard orthogonal CATE learning controls $\varepsilon_n(x)$ only for policies with exact support. POUL adds a second layer: if an unseen policy is close to supported policies in the policy-induced mixture distance, then its error is controlled by a transparent transfer bias. When $r_x(t) = 0$, the result reduces to the classical exact-overlap case. When $r_x(t) > 0$, the theorem quantifies the price of cold-start extrapolation rather than pretending that unsupported policies are fully identified. This is precisely the regime in which categorical treatment IDs fail, because an ID representation has no meaningful notion of $r_x(t)$.

3.4 Minimax Necessity of Policy Geometry

Theorem 3.9 (Minimax necessity of policy geometry)**.** *Fix x, a supported baseline policy $t_0 \in \mathcal{T}_x^\circ$, and a target policy $t^\star \notin \mathcal{T}_x^\circ$ with local support radius $r_x(t^\star) = r > 0$. Consider the class of data-generating laws satisfying Assumptions 3.7 and 3.5, with bounded outcomes and the same observed distribution on the supported policy set $\mathcal{T}_x^\circ$. Then for a universal constant $c > 0$,*

$$\inf_{\widetilde{\tau}} \sup_{P} \mathbb{E}_P \left[|\widetilde{\tau}(x; t^\star, t_0) - \tau_P(x; t^\star, t_0)| \right] \geq c\, G L_F r^\beta. \tag{18}$$

Moreover, any estimator that only observes categorical policy IDs and not the policy specifications $\{\Pi_t(\cdot \mid s)\}_{t,s}$ cannot construct the radius $r_x(t^\star)$; uniform cold-start consistency over relabelings of policy IDs is therefore impossible.

Theorem 3.9 shows that the transfer term in Theorem 3.8 is not a weakness of POUL. It is the unavoidable statistical price of estimating a policy whose exact ID has no support. The only way to make this price small is to exploit policy geometry and ensure that new policies are close to observed ones in d_Π. This distinguishes POUL from categorical-ID encoders: an ID model may fit supported treatments well, but it has no information-theoretic mechanism for transferring causal signal to an unseen label. The policy distance d_Π enters our support-replacement results through a simple Lipschitz stability property of the policy-mixture embedding: policies that are close in induced context–action mixture have close embeddings and close uplift predictions. We defer the formal statement and discussion to Appendix C.9.

4 Experiments

We evaluate POUL on an **industrial ride-hailing dataset** and a **synthetic combinatorial-policy benchmark**. The industrial dataset contains 723 features and approximately 5.28 million samples in the Global population, including 1.58 million in the Core Eligible subset. On this dataset, we compare POUL with baseline methods and conduct ablation studies of its key components, reporting AUUC for GMV and Cost under T_1 (*Economy-to-Express*) and the composite policy T_2 (*Discount-to-Economy + Economy-to-Express*). The synthetic benchmark provides known potential outcomes for policies defined over $|\mathcal{S}| = 8$ contexts and $|\mathcal{A}| = 5$ actions, with PEHE and ATE error measuring observed, tail, and held-out-near policies.

On the **industrial data**, POUL improves AUUC, particularly for Cost-T_2 (Table 1). On the **synthetic benchmark**, POUL with structured encoding achieves the lowest mean $\sqrt{\text{PEHE}}$ across the observed, tail, and held-out-near regimes (Figure 2). Its ATE errors are generally close to the regime-specific minima. Across the four model families, structured encoding generally yields lower errors than ID encoding. Experimental details are provided in Appendix D.

Table 1: Industrial AUUC on the global population ($\mathcal{D}_G$) and the Core Eligible subset ($\mathcal{D}_C$). Results are reported as mean ± sample standard deviation over five runs.

	GMV AUUC ↑				Cost AUUC ↑			
	$\mathcal{D}_G$		$\mathcal{D}_C$		$\mathcal{D}_G$		$\mathcal{D}_C$	
Method	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2
Causal Forest	0.6300 ± 0.0642	0.5897 ± 0.0570	0.5101 ± 0.0728	0.5297 ± 0.0764	0.6593 ± 0.0264	0.6403 ± 0.0344	0.5133 ± 0.0438	0.5191 ± 0.0215
DESCN	0.6843 ± 0.1165	0.5582 ± 0.0578	0.5984 ± 0.0831	0.5926 ± 0.0624	0.7546 ± 0.0235	0.7619 ± 0.0206	0.6462 ± 0.0044	0.6455 ± 0.0061
DRCFR	0.7050 ± 0.0459	0.5962 ± 0.0516	0.6539 ± 0.0306	0.5290 ± 0.0426	0.7426 ± 0.0121	0.7360 ± 0.0133	0.6391 ± 0.0034	0.6317 ± 0.0081
TARNet	0.6018 ± 0.1062	0.5520 ± 0.0816	0.5972 ± 0.0641	0.5780 ± 0.0432	0.7596 ± 0.0053	0.7502 ± 0.0188	0.6499 ± 0.0064	0.6411 ± 0.0071
POUL	$\mathbf{0.8276 \pm 0.0780}$	$\mathbf{0.6353 \pm 0.0276}$	$\mathbf{0.8615 \pm 0.0317}$	$\mathbf{0.6537 \pm 0.0420}$	$\mathbf{0.8008 \pm 0.1124}$	$\mathbf{0.9099 \pm 0.0005}$	$\mathbf{0.8014 \pm 0.1124}$	$\mathbf{0.9097 \pm 0.0003}$

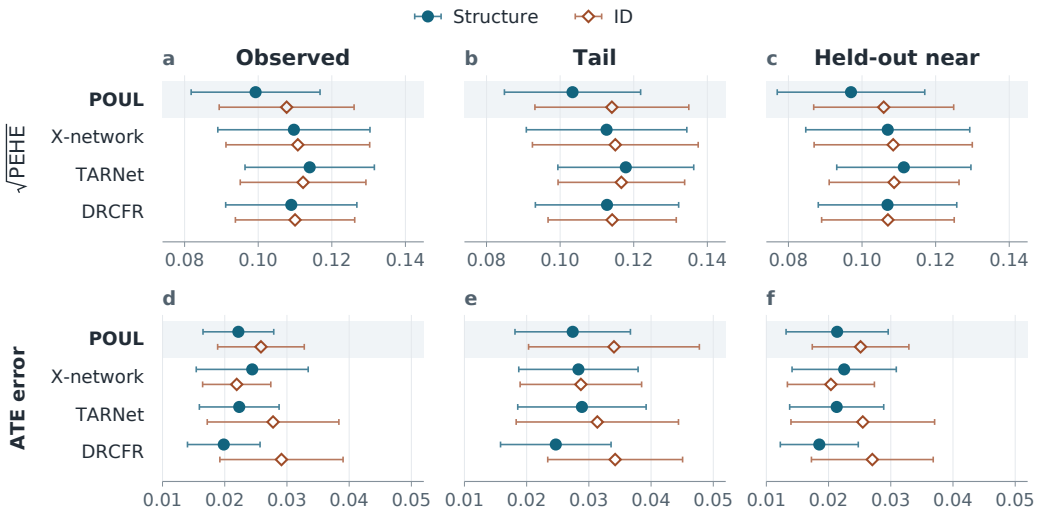


Figure 2: Comparison of structured and ID-based policy encodings on the synthetic benchmark. Points show means over 20 test seeds and whiskers show ± sample standard deviation.

5 Conclusion

This work studies uplift estimation for treatments specified as combinatorial policies. POUL represents each policy through the mixture it induces over context–action components. We establish the expressiveness of permutation-invariant embeddings for functionals of these mixtures, the second-order dependence of the orthogonalized low-rank estimator on nuisance estimation errors, and the stability of the embedding under small perturbations of the policy specification.

These properties clarify when uplift can be estimated for a policy absent from the training data. Under policy-induced mixture sufficiency and local smoothness, nearby supported policies can provide information about an unseen policy, relaxing exact policy overlap to local policy overlap. The cold-start error bound separates the orthogonal estimation error on supported policies from a bias that depends on the target policy's distance to support. A matching minimax lower bound shows that this distance-dependent bias is unavoidable. Categorical-ID encodings therefore cannot guarantee uniform cold-start generalization without additional structure linking policies.

The synthetic experiments assess estimation for rare observed policies and transfer to nearby unseen policies. The industrial experiments evaluate uplift estimation and examine how the resulting predictions perform in downstream allocation at matched subsidy budgets.

Future work could study partial identification under unobserved confounding and extend the framework to settings in which policy assignments or the available context–action components change over time. A related direction is to quantify uncertainty for newly deployed policies and account for it in sequential allocation decisions.

Acknowledgements

Zhiheng Zhang is supported by "the Fundamental Research Funds for the Central Universities" (Grant No. 2025110602) of Shanghai University of Finance and Economics and the Shanghai Engineering Research Center of Finance Intelligence (Grant No. 19DZ2254600). We also thank DiDi Chuxing for supporting this research with data and computational resources.

Conflict of Interest Disclosure This work is the result of a research collaboration between DiDi Chuxing and Shanghai University of Finance and Economics. The use of DiDi Chuxing's industrial data was reviewed and approved through the company's internal data governance and compliance process.

References

Susan Athey and Guido W. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

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

Mingyuan Cheng, Xinru Liao, Quan Liu, Bin Ma, Jian Xu, and Bo Zheng. Learning disentangled representations for counterfactual regression via mutual information minimization. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1802–1806, 2022.

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.

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

Shonosuke Harada and Hisashi Kashima. Graphite: Estimating individual effects of graph-structured treatments. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 659–668, 2021.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

Jean Kaddour, Yuchen Zhu, Qi Liu, Matt J Kusner, and Ricardo Silva. Causal effect inference for structured treatments. *Advances in Neural Information Processing Systems*, 34:24841–24854, 2021.

Edward H Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.

Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *International conference on machine learning*, pages 2747–2755. PMLR, 2018.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10): 4156–4165, 2019.

Wanting Liang, Haoang Chi, and Zhiheng Zhang. Causal representation learning with optimal compression under complex treatments. *arXiv preprint arXiv:2603.11907*, 2026.

Dugang Liu, Xing Tang, Han Gao, Fuyuan Lyu, and Xiuqiang He. Explicit feature interaction-aware uplift network for online marketing. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4507–4515, 2023.

Yunping Lu, Haoang Chi, Qirui Hu, and Zhiheng Zhang. Individualized causal effects under network interference with combinatorial treatments. *arXiv preprint arXiv:2602.19738*, 2026.

Nicholas J Radcliffe and Patrick D Surry. Real-world uplift modelling with significance-based uplift trees. *White Paper TR-2011-1, Stochastic Solutions*, pages 1–33, 2011.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

Piotr Rzepakowski and Szymon Jaroszewicz. Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems*, 32(2):303–327, 2012.

Patrick Schwab, Lorenz Linhardt, Stefan Bauer, Joachim M Buhmann, and Walter Karlen. Learning counterfactual representations for estimating individual dose-response curves. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5612–5619, 2020.

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

Claudia Shi, David M. Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Mark J. van der Laan and Sherri Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, 2011.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Ziang Yan, Shusen Wang, Guorui Zhou, Jingjian Lin, and Peng Jiang. An end-to-end framework for marketing effectiveness optimization under budget constraint. *arXiv preprint arXiv:2302.04477*, 2023.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

Kailiang Zhong, Fengtong Xiao, Yan Ren, Yaorong Liang, Wenqing Yao, Xiaofeng Yang, and Ling Cen. Descn: Deep entire space cross networks for individual treatment effect estimation. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 4612–4620, 2022.

Hao Zhou, Shaoming Li, Guibin Jiang, Jiaqi Zheng, and Dong Wang. Direct heterogeneous causal learning for resource allocation problems in marketing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5446–5454, 2023.

Hao Zhou, Rongxiao Huang, Shaoming Li, Guibin Jiang, Jiaqi Zheng, Bing Cheng, and Wei Lin. Decision focused causal learning for direct counterfactual marketing optimization. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 6368–6379, 2024.

A Related Work

CATE and uplift estimation. A large literature studies heterogeneous treatment effect estimation under the potential outcomes framework, with both classical and modern ML approaches [Rubin, 1974, Imbens and Rubin, 2015, Athey and Imbens, 2016, Wager and Athey, 2018]. For binary treatments, meta-learners such as the T-/S-/X-learners provide practical plug-in strategies and are widely used in industrial uplift modeling [Künzel et al., 2019]. In parallel, uplift modeling has developed evaluation protocols centered around incremental gain and Qini-type curves, as well as specialized tree learners for direct uplift estimation [Radcliffe and Surry, 2011, Rzepakowski and Jaroszewicz, 2012]. Despite substantial progress, most of these methods implicitly assume a *small* and *fixed* treatment set. When the treatment space becomes large and long-tailed—as in combinatorial policies—treating each treatment as an unrelated category leads to poor sample efficiency and unstable ranking, because there is no mechanism to share information across related interventions.

Orthogonalization, doubly robust learning, and debiased ML. Orthogonal / doubly robust ideas are a cornerstone of modern causal estimation, mitigating sensitivity to nuisance estimation through Neyman orthogonality and cross-fitting [Robins et al., 1994, van der Laan and Rose, 2011, Chernozhukov et al., 2018, Kennedy, 2023]. These works provide a principled recipe: learn flexible nuisance components (outcome regression, propensity) while guaranteeing that their errors enter the target estimator only at second order. However, the majority of the DML/TMLE literature treats the treatment as low-dimensional (binary or small discrete), and the nuisance objects are designed accordingly. In our setting, the primary bottleneck is not only nuisance estimation, but also *how to represent a high-cardinality, structured policy* in a way that preserves causal semantics. Our contribution is complementary: we retain the orthogonal robustness advantages of DML-style learning [Chernozhukov et al., 2018] while introducing a structured, permutation-invariant treatment representation that enables parameter sharing and stable generalization across a combinatorial policy space.

Representation learning for treatment effects. Deep representation learning has been widely adopted for CATE/ITE estimation, often aiming to reduce distribution shift between treated and control groups in observational data. Notable examples include TARNet and its balancing variants [Shalit et al., 2017], as well as architectures that explicitly incorporate propensity information such as DragonNet and targeted regularization [Shi et al., 2019]. Extensions to multi-treatment and dose-response settings have also been developed, e.g., DRNet for multiple treatments with continuous dosage [Schwab et al., 2020]. These methods primarily focus on learning a good representation of *unit covariates* X (and nuisance functions of X) to improve counterfactual generalization. By contrast, our focus is orthogonal: we study how to represent *structured treatments* (policies) themselves. In particular, even with randomized assignment (where ignorability is ensured), naive treatment encodings can destroy equivalences between policies and prevent information sharing across related policies, leading to systematic instability in long-tailed regimes. Our permutation-invariant treatment embedding targets precisely this failure mode.

Structured and compositional treatments. There is increasing interest in causal estimation under nonstandard treatments, including multiple treatments, continuous treatments, and structured interventions [Künzel et al., 2019, Schwab et al., 2020]. Recent work has studied representation balancing and treatment aggregation for scalable multi-treatment effect estimation [Liang et al., 2026], as well as orthogonalized estimation for combinatorial treatments under network interference [Lu et al., 2026]. In many real systems, a "treatment" is naturally a *policy* mapping treatment-side contexts to action distributions, inducing a compositional structure. Most existing multi-treatment architectures model treatments via one-hot or learned embeddings tied to treatment IDs. While recent approaches like **EFIN** [Liu et al., 2023] mitigate this by explicitly encoding treatment attributes (e.g., coupon values) to capture inter-treatment correlations, they fundamentally rely on fixed feature interactions rather than enforcing invariances implied by policy semantics. As a result, two implementations of the same policy (differing only by re-indexing or ordering) may still be treated as unrelated, and nearby policies in specification space need not be nearby in representation space. Our work addresses this gap by representing a policy through its induced mixture over context–action atoms and by proving stability under small policy perturbations, which is absent from categorical-ID and feature-interaction paradigms.

Policy geometry versus generic treatment embeddings. Our theory also clarifies the distinction between POUL and generic structured-treatment encoders. Prior methods for structured treatments learn useful treatment features, but they typically do not state a support-replacement result: if a target treatment has no exact overlap, the theory does not quantify when its effect can nevertheless be transferred from nearby observed treatments. By contrast, our policy-induced mixture representation defines a metric d_{Π} on the intervention space, proves an upper bound in which cold-start error depends on the distance to observed policy support, and gives a matching minimax lower bound. The novelty is therefore not permutation invariance in isolation, but the combination of causal policy semantics, orthogonal supported-policy learning, and sharp local policy-overlap transfer.

Permutation invariance and symmetry-aware learning. Permutation-invariant architectures such as Deep Sets [Zaheer et al., 2017] and more general symmetry-aware frameworks [Kondor and Trivedi, 2018] provide a powerful inductive bias when the target depends only on a multiset of components. While such invariances are standard in set-structured prediction, their role in causal estimation with structured treatments has not been systematically developed. Our treatment representation can be viewed as a causal instantiation of this principle:

the treatment embedding is a function of the multiset of context–action components weighted by policy-induced mixture probabilities, and is therefore invariant to re-indexing and ordering. Crucially, this is not merely an architectural convenience: the induced invariance aligns the statistical representation with causal semantics and enables both theoretical stability guarantees and practical generalization to rare or unseen policy variants.

In short, prior work has made significant progress on (i) CATE estimation for low-cardinality treatments, (ii) orthogonal robustness to nuisance estimation, and (iii) representation learning over covariates. Nevertheless, under *combinatorial policy treatments* with frequent iteration and long-tailed exposure, existing methods that rely on categorical treatment IDs remain systematically biased in the *representation* step: they fail to respect policy equivalences and provide no meaningful notion of proximity in policy space. We depart from this paradigm by combining (A) a permutation-invariant, compositional treatment embedding with (B) an orthogonalized uplift objective, thereby obtaining a framework that (1) contains standard DML-style robustness as a special case, (2) generalizes across structured treatments through parameter sharing, and (3) admits explicit expressiveness and stability guarantees tailored to policy perturbations.

B Limitations and Scope

Our theory considers a finite, potentially large policy class with observable context–action specifications. Causal identification on supported policies requires consistency, conditional randomization or ignorability, and positivity. Under policy-mixture sufficiency and local smoothness in d_Π, cold-start consistency follows when both the estimation error on supported policies and the target policies' distances to support vanish. At a fixed positive distance, the minimax lower bound implies a nonvanishing worst-case estimation error.

In the theoretical model, complete randomization gives $T \perp X$ and hence $e_h(x) = \mathbb{E}[h(T)]$. Stratified randomized designs use known conditional propensities, and observational settings use estimated ones. The industrial implementation centers the segment-specific embeddings $h_r(s,t)$ using a running mean.

Misspecification of the policy mixture or rank-d factorization can introduce approximation bias. Noisy context weights can distort policy representations and distances. The cost of computing the embeddings increases with the number of context–action components.

C Proofs and Additional Technical Details

C.1 Notation and basic identities

Recall that each policy $t \in \mathcal{T}$ is specified by context-wise action distributions $\{\Pi_t(\cdot \mid s)\}_{s\in\mathcal{S}}$ over a finite action set $\mathcal{A}$ and finite context set $\mathcal{S}$. Let $w : \mathcal{S} \to [0,\infty)$ be context weights with $\sum_{s\in\mathcal{S}} w(s) = 1$ and define the induced atom weights

$$\alpha_t(s,a) \;:=\; w(s)\Pi_t(a \mid s), \qquad (s,a) \in \mathcal{S}\times\mathcal{A}, \tag{19}$$

and the induced discrete measure μ_t on $\mathcal{S}\times\mathcal{A}$ by $\mu_t(\{(s,a)\}) = \alpha_t(s,a)$.

Throughout the appendix, for a scalar (or vector) random variable Z we write $\|Z\|_2 := (\mathbb{E}\|Z\|^2)^{1/2}$. For a function $h : \mathcal{T} \to \mathbb{R}^d$, we use $\|h\|_\infty := \sup_{t\in\mathcal{T}} \|h(t)\|$ and $\|h - \tilde{h}\|_\infty := \sup_{t\in\mathcal{T}} \|h(t) - \tilde{h}(t)\|$. For a function $g : \mathbb{R}^p \to \mathbb{R}^d$, we will occasionally write $\|g - \tilde{g}\|_{\infty,x} := \sup_x \|g(x) - \tilde{g}(x)\|$ when the supremum is taken over the domain of interest.

A basic identity that will be used repeatedly is the "centering" property

$$\mathbb{E}[h_0(T) - e_0(X) \mid X] = 0, \qquad e_0(X) := \mathbb{E}[h_0(T) \mid X]. \tag{20}$$

This is the main mechanism behind Neyman orthogonality and the second-order nature of nuisance effects.

C.2 Proof of Proposition 3.2

Proof of Proposition 3.2. Fix any $x \in \mathbb{R}^p$ and $t \in \mathcal{T}$. By Assumption 3.1, there exist measurable m_0 and g_0 and a functional F such that

$$\mathbb{E}[Y(t) \mid X = x] = m_0(x) + g_0(x)^\top F(\mu_t). \tag{21}$$

Define $h_0(t) := F(\mu_t)$. Under consistency $Y = Y(T)$ and randomized assignment (so that conditioning on (X,T) selects the corresponding potential outcome), we have

$$\mathbb{E}[Y \mid X = x, T = t] = \mathbb{E}[Y(t) \mid X = x] = m_0(x) + g_0(x)^\top h_0(t). \tag{22}$$

Let $e_0(x) := \mathbb{E}[h_0(T) \mid X = x]$ and define $m(x) := m_0(x) + g_0(x)^\top e_0(x)$. Then, for any (x, t),

$$\begin{aligned} m_0(x) + g_0(x)^\top h_0(t) &= \Big(m_0(x) + g_0(x)^\top e_0(x)\Big) + g_0(x)^\top \Big(h_0(t) - e_0(x)\Big) \\ &= m(x) + g_0(x)^\top \Big(h_0(t) - e_0(x)\Big). \end{aligned}$$

Thus the conditional mean obeys (9).

Finally, the CATE for any pair (t_1, t_0) satisfies

$$\begin{aligned} \tau(x; t_1, t_0) &:= \mathbb{E}[Y(t_1) - Y(t_0) \mid X = x] \\ &= \Big(m_0(x) + g_0(x)^\top h_0(t_1)\Big) - \Big(m_0(x) + g_0(x)^\top h_0(t_0)\Big) \\ &= g_0(x)^\top \Big(h_0(t_1) - h_0(t_0)\Big), \end{aligned}$$

which is (10). □

C.3 Proof of Lemma 3.3 (Expressiveness)

Proof of Lemma 3.3. Because $\mathcal{S}$ and $\mathcal{A}$ are finite, the product set $\mathcal{S} \times \mathcal{A}$ is finite as well. Let $m := |\mathcal{S}||\mathcal{A}|$ and fix any bijection (index map) $\iota : \mathcal{S} \times \mathcal{A} \to \{1, 2, \ldots, m\}$. For any measure μ supported on $\mathcal{S} \times \mathcal{A}$, define its coordinate vector $\alpha(\mu) \in \mathbb{R}^m$ by

$$\alpha(\mu)_{\iota(s,a)} := \mu(\{(s, a)\}). \tag{23}$$

In particular, for μ_t induced by policy t, we have $\alpha(\mu_t)_{\iota(s,a)} = \alpha_t(s, a) = w(s)\Pi_t(a \mid s)$ and $\alpha(\mu_t) \in \Delta^{m-1}$ (the probability simplex). The map $\mu \mapsto \alpha(\mu)$ is a linear homeomorphism between the space of measures on the finite set $\mathcal{S} \times \mathcal{A}$ and $\mathbb{R}^m$ restricted to Δ^{m-1}.

Define $\widetilde{F} : \alpha(\mathcal{M}) \to \mathbb{R}^d$ by $\widetilde{F}(\alpha(\mu)) := F(\mu)$ for $\mu \in \mathcal{M}$. Since F is continuous on $\mathcal{M}$ under the ℓ_1/TV topology and the identification $\|\mu - \mu'\|_1 = \|\alpha(\mu) - \alpha(\mu')\|_1$ holds on finite supports, $\widetilde{F}$ is continuous on the set

$$\alpha(\mathcal{M}) = \{\alpha(\mu_t) : t \in \mathcal{T}\} \subset \Delta^{m-1}. \tag{24}$$

Moreover, Δ^{m-1} is compact and $\alpha(\mathcal{M})$ is compact as a closed subset of a compact set. By the universal approximation theorem for ReLU MLPs on compact sets, for any $\varepsilon > 0$ there exists an integer r (in fact we may take $r = m$) and an MLP $\rho : \mathbb{R}^r \to \mathbb{R}^d$ such that

$$\sup_{t \in \mathcal{T}} \big\|\widetilde{F}(\alpha(\mu_t)) - \rho(\alpha(\mu_t))\big\| \le \varepsilon. \tag{25}$$

It remains to realize $\alpha(\mu_t)$ as a permutation-invariant weighted sum. Let $\{e_j\}_{j=1}^m$ denote the standard basis of $\mathbb{R}^m$ and choose $r = m$ and

$$\phi(s, a) := e_{\iota(s,a)} \in \mathbb{R}^m. \tag{26}$$

Then

$$z_\phi(t) = \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \alpha_t(s, a)\phi(s, a) = \sum_{s,a} \alpha_t(s, a) e_{\iota(s,a)} = \alpha(\mu_t). \tag{27}$$

Therefore, defining $h_{\phi,\rho}(t) := \rho(z_\phi(t))$ yields

$$h_{\phi,\rho}(t) = \rho(\alpha(\mu_t)). \tag{28}$$

Combining this with (25) gives

$$\sup_{t \in \mathcal{T}} \|F(\mu_t) - h_{\phi,\rho}(t)\| = \sup_{t \in \mathcal{T}} \|\widetilde{F}(\alpha(\mu_t)) - \rho(\alpha(\mu_t))\| \le \varepsilon, \tag{29}$$

which proves the claim.

Remark C.1. While the above construction uses the basis embedding $\phi(s, a) = e_{\iota(s,a)}$ and thus shows universality in a straightforward way, it is precisely the *shared atom map* ϕ together with *permutation-invariant aggregation* that constitutes the inductive bias of our approach. This bias is what yields stability in policy space (Proposition C.7) and practical statistical strength-sharing when $\mathcal{T}$ is huge but $\mathcal{S} \times \mathcal{A}$ has reusable structure.

□

C.4 CATE Identification and Affine Invariance

We distinguish identification by the joint population risk from Neyman orthogonality of the score. With unrestricted h, a constant embedding satisfies $h(T) - e_h(X) = 0$, so the score vanishes for any g, even when the true treatment contrast is nonzero. The identification argument below therefore uses the full squared-error

population risk over (g, h). The score provides Neyman orthogonality with respect to the nuisance functions at the oracle pair.

Let $\pi_0(t \mid x) = \mathbb{P}(T = t \mid X = x)$ denote the propensity of policy t, and let $\mu_0(x, t) = \mathbb{E}[Y \mid X = x, T = t]$. For a candidate embedding h, define

$$\bar{\mu}_0(x) = \mathbb{E}[Y \mid X = x], \qquad e_h(x) = \mathbb{E}[h(T) \mid X = x]. \tag{30}$$

The centered response surface is

$$f_{g,h}(x, t) = \bar{\mu}_0(x) + g(x)^\top \{h(t) - e_h(x)\}. \tag{31}$$

Proposition C.2 (Identification of supported-policy CATE via population risk)**.** *Assume consistency and conditional ignorability, and suppose that the outcome and candidate response surfaces are square-integrable. Suppose the centered factorized model is well specified, so there exists a candidate pair $(\bar{g}, \bar{h})$ such that*

$$f_{\bar{g},\bar{h}}(X, T) = \mu_0(X, T) \qquad P_{X,T}\text{-a.s.} \tag{32}$$

Define the population risk

$$\mathcal{R}(g, h) = \frac{1}{2}\mathbb{E}\big[\{Y - f_{g,h}(X, T)\}^2\big]. \tag{33}$$

Then every global minimizer $(g^\star, h^\star)$ satisfies

$$f_{g^\star,h^\star}(X, T) = \mu_0(X, T) \qquad P_{X,T}\text{-a.s.} \tag{34}$$

For any fixed policy pair (t_1, t_0), the induced contrast satisfies

$$g^\star(x)^\top \{h^\star(t_1) - h^\star(t_0)\} = \tau(x; t_1, t_0) \tag{35}$$

for P_X-almost every x at which $\pi_0(t_1 \mid x) > 0$ and $\pi_0(t_0 \mid x) > 0$. Thus, identification of the supported-policy CATE does not require knowledge or point identification of a particular h_0.

Proof. The conditional-mean decomposition gives

$$\mathcal{R}(g, h) = \frac{1}{2}\mathbb{E}\big[\{Y - \mu_0(X, T)\}^2\big] + \frac{1}{2}\mathbb{E}\big[\{\mu_0(X, T) - f_{g,h}(X, T)\}^2\big], \tag{36}$$

because $\mathbb{E}[Y - \mu_0(X, T) \mid X, T] = 0$. The first term is independent of (g, h). By well specification, $(\bar{g}, \bar{h})$ attains this lower bound. Thus every global minimizer has zero squared error relative to μ_0 under $P_{X,T}$.

For each fixed t, this equality implies

$$\mathbb{E}\big[\pi_0(t \mid X)\{f_{g^\star,h^\star}(X, t) - \mu_0(X, t)\}^2\big] = 0. \tag{37}$$

Hence $f_{g^\star,h^\star}(x, t) = \mu_0(x, t)$ for P_X-almost every x at which $\pi_0(t \mid x) > 0$. When both policies are supported, subtracting the two fitted conditional means cancels the baseline and the centering term. By consistency and conditional ignorability,

$$\begin{aligned} g^\star(x)^\top \{h^\star(t_1) - h^\star(t_0)\} &= f_{g^\star,h^\star}(x, t_1) - f_{g^\star,h^\star}(x, t_0) \\ &= \mu_0(x, t_1) - \mu_0(x, t_0) \\ &= \mathbb{E}[Y(t_1) - Y(t_0) \mid X = x] = \tau(x; t_1, t_0). \end{aligned} \tag{38}$$

□

Corollary C.3 (An excess-risk bound for supported-policy CATE)**.** *Under the same setup, allow the model class to be misspecified and define its squared approximation error as*

$$\mathcal{A}_\mathcal{F} = \inf_{g,h} \mathbb{E}\big[\{f_{g,h}(X, T) - \mu_0(X, T)\}^2\big]. \tag{39}$$

If $(\widehat{g}, \widehat{h})$ has population excess risk at most ε, that is,

$$\mathcal{R}(\widehat{g}, \widehat{h}) - \inf_{g,h} \mathcal{R}(g, h) \le \varepsilon, \tag{40}$$

then

$$\mathbb{E}\big[\{f_{\widehat{g},\widehat{h}}(X, T) - \mu_0(X, T)\}^2\big] \le \mathcal{A}_\mathcal{F} + 2\varepsilon. \tag{41}$$

For a fixed policy pair (t_1, t_0), suppose $\pi_0(t_j \mid X) \ge c > 0$ almost surely for $j = 0, 1$. Then the estimated contrast

$$\widehat{\tau}(x; t_1, t_0) = \widehat{g}(x)^\top \{\widehat{h}(t_1) - \widehat{h}(t_0)\} \tag{42}$$

satisfies

$$\mathbb{E}_X\big[\{\widehat{\tau}(X; t_1, t_0) - \tau(X; t_1, t_0)\}^2\big] \le \frac{4}{c}\big(\mathcal{A}_\mathcal{F} + 2\varepsilon\big). \tag{43}$$

Proof. The first inequality follows from the conditional-mean decomposition in the proof of Proposition C.2. Write $r(x,t) = f_{\widehat{g},\widehat{h}}(x,t) - \mu_0(x,t)$. For $j = 0, 1$, the overlap condition implies

$$\begin{aligned} c\,\mathbb{E}_X[r(X,t_j)^2] &\le \mathbb{E}_X[\pi_0(t_j \mid X) r(X,t_j)^2] \\ &= \mathbb{E}[\mathbf{1}\{T = t_j\} r(X,T)^2] \le \mathbb{E}[r(X,T)^2]. \end{aligned} \tag{44}$$

The contrast error is $r(X,t_1) - r(X,t_0)$. Applying $(a-b)^2 \le 2a^2 + 2b^2$ and then the first inequality gives the claimed bound. □

Thus, under uniform overlap, vanishing approximation error and population excess risk imply mean-square consistency for the supported-policy CATE.

Proposition C.4 (Affine invariance of the response surface and treatment contrasts)**.** *Let (g,h) be any candidate pair. For any invertible $A \in \mathbb{R}^{d\times d}$ and any $c \in \mathbb{R}^d$, define*

$$h_{A,c}(t) = Ah(t) + c, \qquad g_A(x) = A^{-\top} g(x). \tag{45}$$

Then $e_{h_{A,c}}(x) = Ae_h(x) + c$. The transformed pair represents the same response surface,

$$f_{g_A,h_{A,c}}(x,t) = f_{g,h}(x,t), \tag{46}$$

and the same contrast for every policy pair (t_1,t_0),

$$g_A(x)^\top \{h_{A,c}(t_1) - h_{A,c}(t_0)\} = g(x)^\top \{h(t_1) - h(t_0)\}. \tag{47}$$

Proof. By linearity of conditional expectation,

$$e_{h_{A,c}}(x) = \mathbb{E}[Ah(T) + c \mid X = x] = Ae_h(x) + c. \tag{48}$$

Therefore,

$$\begin{aligned} g_A(x)^\top \{h_{A,c}(t) - e_{h_{A,c}}(x)\} &= (A^{-\top} g(x))^\top A\{h(t) - e_h(x)\} \\ &= g(x)^\top \{h(t) - e_h(x)\}. \end{aligned} \tag{49}$$

Adding the common baseline proves response-surface invariance. Similarly, the translation cancels when taking a treatment contrast, so

$$\begin{aligned} g_A(x)^\top \{h_{A,c}(t_1) - h_{A,c}(t_0)\} &= (A^{-\top} g(x))^\top A\{h(t_1) - h(t_0)\} \\ &= g(x)^\top \{h(t_1) - h(t_0)\}. \end{aligned} \tag{50}$$

□

The score transforms under the same change of coordinates. For $W = (X,T,Y)$, write

$$\psi(W; g, h, m, e) = \{Y - m(X) - g(X)^\top [h(T) - e(X)]\}\{h(T) - e(X)\}. \tag{51}$$

For any nuisance pair (m,e), define $e_{A,c}(x) = Ae(x) + c$. The residual in the first factor is invariant, and the second factor is multiplied by A. Hence

$$\psi(W; g_A, h_{A,c}, m, e_{A,c}) = A\,\psi(W; g, h, m, e). \tag{52}$$

In particular, when $(m,e) = (\bar{\mu}_0, e_h)$, the transformed nuisance pair is $(\bar{\mu}_0, e_{h_{A,c}})$. A zero moment remains zero under this transformation. Since A is fixed, the same identity along nuisance perturbation paths shows that a zero Neyman-orthogonality derivative also remains zero.

For a one-dimensional example, consider the reparameterization

$$h^\dagger(t) = 2h_0(t) + 7, \qquad g^\dagger(x) = \tfrac{1}{2} g_0(x). \tag{53}$$

The scale and origin of the embedding change, while

$$g^\dagger(x)\{h^\dagger(t_1) - h^\dagger(t_0)\} = g_0(x)\{h_0(t_1) - h_0(t_0)\}. \tag{54}$$

Affine transformations thus give a family of equivalent parameterizations. Proposition C.2 applies to every global population-risk minimizer under well specification, regardless of how its factors are related to those of other minimizers.

C.5 Proof of Lemma 3.4 (Neyman orthogonality)

Proof of Lemma 3.4. Recall the orthogonalized model (4):

$$Y = m_0(X) + g_0(X)^\top \big(h_0(T) - e_0(X)\big) + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid X, T] = 0, \tag{55}$$

and the score

$$\psi(W; g, h, m, e) := \Big(Y - m(X) - g(X)^\top (h(T) - e(X))\Big)\,(h(T) - e(X)), \quad W := (X,T,Y). \tag{56}$$

Step 1: Unbiasedness at the truth. At the true functions (g_0, h_0, m_0, e_0),

$$Y - m_0(X) - g_0(X)^\top (h_0(T) - e_0(X)) = \varepsilon, \tag{57}$$

so

$$\psi(W; g_0, h_0, m_0, e_0) = \varepsilon\,(h_0(T) - e_0(X)). \tag{58}$$

Taking expectation and using iterated expectation gives

$$\begin{aligned}\mathbb{E}[\psi(W; g_0, h_0, m_0, e_0)] &= \mathbb{E}\big[\mathbb{E}[\varepsilon\,(h_0(T) - e_0(X)) \mid X, T]\big] \\ &= \mathbb{E}\big[(h_0(T) - e_0(X))\,\mathbb{E}[\varepsilon \mid X, T]\big] = 0,\end{aligned}$$

proving $\mathbb{E}[\psi(W; g_0, h_0, m_0, e_0)] = 0$.

Step 2: Gateaux derivative in nuisance directions vanishes. Fix perturbations δm and δe such that all expressions below are integrable (e.g., $\delta m(X)$ and $\delta e(X)$ square-integrable). Define the nuisance path $m_r := m_0 + r\,\delta m$ and $e_r := e_0 + r\,\delta e$. Consider

$$\Psi(r) := \mathbb{E}\big[\psi(W; g_0, h_0, m_r, e_r)\big]. \tag{59}$$

Using the model for Y and expanding the residual,

$$\begin{aligned}Y - m_r(X) - g_0(X)^\top (h_0(T) - e_r(X)) &= \Big(m_0(X) - m_r(X)\Big) + g_0(X)^\top \Big(e_r(X) - e_0(X)\Big) + \varepsilon \\ &= -r\,\delta m(X) + r\,g_0(X)^\top \delta e(X) + \varepsilon.\end{aligned}$$

Also,

$$h_0(T) - e_r(X) = \big(h_0(T) - e_0(X)\big) - r\,\delta e(X). \tag{60}$$

Hence

$$\psi(W; g_0, h_0, m_r, e_r) = \Big(\varepsilon - r\,\delta m(X) + r\,g_0(X)^\top \delta e(X)\Big)\,\Big((h_0(T) - e_0(X)) - r\,\delta e(X)\Big). \tag{61}$$

Expanding and collecting the terms linear in r yields

$$\Psi(r) = \underbrace{\mathbb{E}\big[\varepsilon\,(h_0(T) - e_0(X))\big]}_{=0} + r\,\mathbb{E}\Big[-\delta m(X)\,(h_0(T) - e_0(X)) + (g_0(X)^\top \delta e(X))\,(h_0(T) - e_0(X)) - \varepsilon\,\delta e(X)\Big] + O(r^2),$$

where the $O(r^2)$ term is justified by integrability and dominated convergence.

We now show each linear term has zero expectation. For the first two terms, condition on X and use (20):

$$\begin{aligned}\mathbb{E}\big[\delta m(X)\,(h_0(T) - e_0(X))\big] &= \mathbb{E}\big[\delta m(X)\,\mathbb{E}[h_0(T) - e_0(X) \mid X]\big] = 0, \\ \mathbb{E}\big[(g_0(X)^\top \delta e(X))\,(h_0(T) - e_0(X))\big] &= \mathbb{E}\big[(g_0(X)^\top \delta e(X))\,\mathbb{E}[h_0(T) - e_0(X) \mid X]\big] = 0.\end{aligned}$$

For the third term, condition on (X, T) and use $\mathbb{E}[\varepsilon \mid X, T] = 0$:

$$\mathbb{E}[\varepsilon\,\delta e(X)] = \mathbb{E}\big[\delta e(X)\,\mathbb{E}[\varepsilon \mid X, T]\big] = 0. \tag{62}$$

Therefore the coefficient of r in $\Psi(r)$ is zero, implying $\Psi'(0) = 0$, i.e.,

$$\frac{d}{dr}\bigg|_{r=0} \mathbb{E}\big[\psi(W; g_0, h_0, m_0 + r\,\delta m, e_0 + r\,\delta e)\big] = 0. \tag{63}$$

This proves Neyman orthogonality with respect to (m, e). □

C.6 Proof of Theorem 3.6

C.6.1 A deterministic second-order bound for nuisance perturbations

The core technical step is to control how the orthogonal score changes when we replace the true nuisances (m_0, e_0) by estimators $(\hat{m}, \hat{e})$, *while simultaneously allowing* the learned effect components $(\hat{g}, \hat{h})$ to deviate from (g_0, h_0). This coupling is nontrivial in our setting because (i) the regressor $h(T) - e(X)$ is vector-valued, (ii) $e(X)$ is itself an X-conditional expectation of $h(T)$, and (iii) $\hat{e}$ may be constructed through a propensity model and hence is another high-dimensional nuisance.

Lemma C.5 (Second-order control of score perturbation)**.** *Assume the model* (4) *holds with* (m_0, g_0, h_0, e_0) *and* $\mathbb{E}[\varepsilon \mid X, T] = 0$. *Let* (g, h, m, e) *be (possibly random) candidates such that all quantities below are integrable. Define* $\Delta g := g - g_0$, $\Delta h := h - h_0$, $\Delta m := m - m_0$, *and* $\Delta e := e - e_0$. *Then*

$$\Big\|\mathbb{E}\big[\psi(W; g, h, m, e) - \psi(W; g, h, m_0, e_0)\big]\Big\| \;\le\; C_1\,\big(\|\Delta m\|_2 + \|\Delta h\|_\infty + \|\Delta g\|_{\infty,x}\big)\,\|\Delta e\|_2 \;+\; C_2\,\|\Delta e\|_2^2, \tag{64}$$

for constants C_1, C_2 *depending only on moment bounds of* (Y, g_0, h_0) *and on the domain of* (X, T).

Proof of Lemma C.5. Write the score difference as

$$\psi(W;g,h,m,e)-\psi(W;g,h,m_0,e_0) = \big(R_{m,e}(W)-R_{m_0,e_0}(W)\big)\,(h(T)-e(X)) + R_{m_0,e_0}(W)\,\big((h(T)-e(X))-(h(T)-e_0(X))\big), \tag{65}$$

where

$$R_{m,e}(W) := Y - m(X) - g(X)^\top (h(T) - e(X)). \tag{66}$$

Noting $(h(T) - e(X)) - (h(T) - e_0(X)) = -(e(X) - e_0(X)) = -\Delta e(X)$, we have

$$\psi(W;g,h,m,e)-\psi(W;g,h,m_0,e_0) = \big(R_{m,e}(W)-R_{m_0,e_0}(W)\big)\,(h(T)-e(X)) - R_{m_0,e_0}(W)\,\Delta e(X). \tag{67}$$

We next expand the residual difference. A direct calculation gives

$$\begin{aligned} R_{m,e}(W) - R_{m_0,e_0}(W) &= -\Delta m(X) - g(X)^\top \big(h(T) - e(X)\big) + g(X)^\top \big(h(T) - e_0(X)\big) \\ &= -\Delta m(X) + g(X)^\top \Delta e(X). \end{aligned}$$

Plugging into (67) yields

$$\psi(W;g,h,m,e) - \psi(W;g,h,m_0,e_0) = \Big(- \Delta m(X) + g(X)^\top \Delta e(X)\Big)\,\big(h(T) - e(X)\big) - R_{m_0,e_0}(W)\,\Delta e(X). \tag{68}$$

Now we express $R_{m_0,e_0}(W)$ under the true model:

$$\begin{aligned} R_{m_0,e_0}(W) &= Y - m_0(X) - g(X)^\top (h(T) - e_0(X)) \\ &= \underbrace{\Big(Y - m_0(X) - g_0(X)^\top (h_0(T) - e_0(X))\Big)}_{=\varepsilon} + g_0(X)^\top (h_0(T) - e_0(X)) - g(X)^\top (h(T) - e_0(X)) \\ &= \varepsilon - \Delta g(X)^\top \big(h_0(T) - e_0(X)\big) - g(X)^\top \Delta h(T), \end{aligned}$$

where we used $h(T) = h_0(T) + \Delta h(T)$ and $\Delta h(T) := h(T) - h_0(T)$.

Substitute this into (68) and take expectation. The expectation of the term $\varepsilon\,\Delta e(X)$ vanishes by iterated expectation:

$$\mathbb{E}[\varepsilon\,\Delta e(X)] = \mathbb{E}\big[\Delta e(X)\,\mathbb{E}[\varepsilon \mid X, T]\big] = 0. \tag{69}$$

Moreover, the key centering identity (20) implies that whenever $A(X)$ is measurable in X,

$$\mathbb{E}\big[A(X)\,(h_0(T) - e_0(X))\big] = \mathbb{E}\big[A(X)\,\mathbb{E}[h_0(T) - e_0(X) \mid X]\big] = 0. \tag{70}$$

This kills the *first-order* terms in Δm and in Δe that would otherwise appear in the score bias.

What remains after cancellations are only products of perturbations, which we bound by Cauchy–Schwarz. Concretely, from (68) and the above expansions, we obtain

$$\mathbb{E}\big[\psi(W;g,h,m,e)-\psi(W;g,h,m_0,e_0)\big] = \mathbb{E}\big[\Delta m(X)\,\Delta e(X)\big]+\mathbb{E}\big[(g(X)^\top \Delta e(X))\,\Delta e(X)\big]+\mathbb{E}\big[\Gamma(X,T)\,\Delta e(X)\big], \tag{71}$$

where $\Gamma(X,T)$ collects terms involving Δg and Δh multiplied by bounded random quantities (e.g., $h_0(T) - e_0(X)$ and $g(X)$). Using $\|g(X)\| \le \|g_0(X)\| + \|\Delta g\|_{\infty,x}$ and $\|\Delta h(T)\| \le \|\Delta h\|_\infty$, we bound

$$\begin{aligned} \big\|\mathbb{E}[\Delta m(X)\,\Delta e(X)]\big\| &\le \|\Delta m\|_2\,\|\Delta e\|_2, \\ \big\|\mathbb{E}[(g(X)^\top \Delta e(X))\,\Delta e(X)]\big\| &\le \mathbb{E}[\|g(X)\|\,\|\Delta e(X)\|^2] \le (\mathbb{E}\|g(X)\|^2)^{1/2}\,\|\Delta e\|_2^2, \\ \big\|\mathbb{E}[\Gamma(X,T)\,\Delta e(X)]\big\| &\le \big(\|\Delta g\|_{\infty,x} + \|\Delta h\|_\infty\big)\,\|\Delta e\|_2 \cdot C, \end{aligned}$$

for a constant C depending only on bounded moments of $g_0(X)$ and $h_0(T)$ (cf. Assumption 3.5). Collecting the terms yields (64). □

C.6.2 Proof of Theorem 3.6

Proof of Theorem 3.6. We separate the argument into an algebraic decomposition (capturing the effect-component errors) and a nuisance-induced remainder (capturing orthogonal robustness).

Step 1: Algebraic decomposition of the plug-in CATE. Recall $\tau(x;t_1,t_0) = g_0(x)^\top (h_0(t_1) - h_0(t_0))$ from Proposition 3.2 and $\hat\tau(x;t_1,t_0) = \hat g(x)^\top (\hat h(t_1) - \hat h(t_0))$. Add and subtract $\hat g(x)^\top (h_0(t_1) - h_0(t_0))$:

$$\hat\tau(x;t_1,t_0) - \tau(x;t_1,t_0) = \underbrace{(\hat g(x) - g_0(x))^\top \big(h_0(t_1) - h_0(t_0)\big)}_{=:\ \mathrm{error(g)}} + \hat g(x)^\top \Big((\hat h - h_0)(t_1) - (\hat h - h_0)(t_0)\Big). \tag{72}$$

Decompose the second term further by writing $\hat{g}(x) = g_0(x) + (\hat{g}(x) - g_0(x))$:

$$\hat{g}(x)^\top \Big((\hat{h} - h_0)(t_1) - (\hat{h} - h_0)(t_0) \Big) = \underbrace{g_0(x)^\top \Big((\hat{h} - h_0)(t_1) - (\hat{h} - h_0)(t_0) \Big)}_{=: \text{ error(h)}} + R_{gh}(x; t_1, t_0), \quad (73)$$

where the remainder

$$R_{gh}(x; t_1, t_0) := (\hat{g}(x) - g_0(x))^\top \Big((\hat{h} - h_0)(t_1) - (\hat{h} - h_0)(t_0) \Big) \quad (74)$$

is *second order* in the effect-component errors and obeys the deterministic bound

$$|R_{gh}(x; t_1, t_0)| \;\le\; 2\, \|\hat{g}(x) - g_0(x)\|\, \|\hat{h} - h_0\|_\infty \;\le\; 2\, \delta_g(x)\, \delta_h. \quad (75)$$

In many asymptotic regimes, $\delta_g(x)\delta_h = o_p(\delta_g(x) + \delta_h)$; we keep it explicit here for completeness.

Combining (72)–(73) gives

$$\hat{\tau}(x; t_1, t_0) - \tau(x; t_1, t_0) = \text{error(g)} + \text{error(h)} + R_{gh}(x; t_1, t_0). \quad (76)$$

Step 2: Orthogonal robustness isolates nuisance effects. The central claim of Theorem 3.6 is about the *additional* impact of nuisance estimation beyond the effect-component errors. To make this precise, note that the orthogonalized learning procedure (Algorithm 1) is driven by the score ψ and its cross-fitted empirical counterpart. Under cross-fitting (Assumption 3.5(3)), we may condition on the nuisance estimates $(\hat{m}^{(-k)}, \hat{e}^{(-k)})$ and treat them as fixed when taking expectation over the evaluation fold. Lemma C.5 then yields the deterministic second-order control

$$\Big\| \mathbb{E}\big[\psi(W; \hat{g}, \hat{h}, \hat{m}, \hat{e}) - \psi(W; \hat{g}, \hat{h}, m_0, e_0)\big] \Big\| \;\le\; C_1(\delta_m + \delta_h + \|\hat{g} - g_0\|_{\infty,x})\, \delta_e + C_2\, \delta_e^2. \quad (77)$$

By Assumption 3.5(4), $\delta_e = o_p(1)$, so $\delta_e^2 = o_p(\delta_e)$ and can be absorbed into the product term without changing the order (by enlarging constants). Thus,

$$\Big\| \mathbb{E}\big[\psi(W; \hat{g}, \hat{h}, \hat{m}, \hat{e}) - \psi(W; \hat{g}, \hat{h}, m_0, e_0)\big] \Big\| = O_p\Big((\delta_m + \delta_h + \|\hat{g} - g_0\|_{\infty,x})\, \delta_e \Big). \quad (78)$$

Equation (78) is the formal mathematical expression of the statement "nuisance errors enter only at second order": the perturbation in the orthogonal score induced by $(\hat{m}, \hat{e})$ is proportional to a *product* involving δ_e rather than a sum of first-order terms. This is exactly where orthogonality matters; without centering (20), the bound would contain linear (first-order) terms in δ_m and δ_e.

Finally, the CATE functional depends on the learned effect components through $\hat{\tau}(x; t_1, t_0) = \hat{g}(x)^\top (\hat{h}(t_1) - \hat{h}(t_0))$. The learning procedure targets $(\hat{g}, \hat{h})$ through the orthogonal score, so the nuisance-induced perturbation in the target is controlled by the same second-order quantity appearing in (78). This yields the remainder term stated in Theorem 3.6:

$$\text{error(second order)} = O_p\Big((\delta_m + \delta_h + \sup_x \delta_g(x))\, \delta_e \Big), \quad (79)$$

where we used $\|\hat{g} - g_0\|_{\infty,x} = \sup_x \delta_g(x)$ by definition. Combining with (76) completes the proof. □

Remark C.6. Compared to the classical partially linear model and standard DML analyses, two aspects require extra care here: (i) the "treatment regressor" $h(T) - e(X)$ is *vector-valued* and *learned* (hence the appearance of δ_h), and (ii) the nuisance $e(X) = \mathbb{E}[h(T) \mid X]$ is an *embedding propensity* and may be constructed via a propensity model (hence δ_e couples representation learning and nuisance estimation). Lemma C.5 makes this coupling explicit, and shows that orthogonalization controls it at second order.

C.7 Proof of Theorem 3.8

Proof. For $j \in \{0, 1\}$, choose $u_j \in \mathcal{T}_x^\circ$ such that $d_\Pi(t_j, u_j) \le r_x(t_j) + \eta$, where $\eta > 0$ is arbitrary. Add and subtract the supported-policy contrast:

$$\begin{aligned} &|\widehat{\tau}(x; t_1, t_0) - \tau(x; t_1, t_0)| \\ &\le |\widehat{\tau}(x; u_1, u_0) - \tau(x; u_1, u_0)| + |\widehat{\tau}(x; t_1, t_0) - \widehat{\tau}(x; u_1, u_0)| + |\tau(x; t_1, t_0) - \tau(x; u_1, u_0)|\,. \end{aligned} \quad (80)$$

The first term is bounded by $\varepsilon_n(x)$. For the fitted contrast,

$$|\widehat{\tau}(x; t_1, t_0) - \widehat{\tau}(x; u_1, u_0)| \le \|\widehat{g}(x)\| \left(\|\widehat{h}(t_1) - \widehat{h}(u_1)\| + \|\widehat{h}(t_0) - \widehat{h}(u_0)\| \right), \quad (81)$$

which is at most

$$\widehat{G} L_{\widehat{h}} \{ (r_x(t_1) + \eta)^\beta + (r_x(t_0) + \eta)^\beta \}. \quad (82)$$

Similarly, using $h_0(t) = F(\mu_t)$,

$$|\tau(x; t_1, t_0) - \tau(x; u_1, u_0)| \le G L_F \{ (r_x(t_1) + \eta)^\beta + (r_x(t_0) + \eta)^\beta \}. \quad (83)$$

Letting $\eta \downarrow 0$ proves the claimed bound. The consistency statement follows immediately when $\varepsilon_n(x) = o_p(1)$ and $r_x(t_j) = o_p(1)$ for $j = 0, 1$. □

C.8 Proof of Theorem 3.9

Proof. Let $v \in \mathbb{R}^d$ be a unit vector and choose $g_0(x) = Gv$. Construct two policy-effect functionals F_+ and F_- that coincide on all supported policies:

$$F_+(\mu_u) = F_-(\mu_u) = 0, \qquad u \in \mathcal{T}_x^\circ, \tag{84}$$

and differ only at the unsupported target policy:

$$F_+(\mu_{t^\star}) = av, \qquad F_-(\mu_{t^\star}) = -av, \tag{85}$$

where $a = L_F r^\beta/2$. Since $d_\Pi(t^\star, u) \geq r$ for all $u \in \mathcal{T}_x^\circ$, both F_+ and F_- satisfy the Hölder condition in Assumption 3.7, up to a universal rescaling of constants.

The two induced data-generating laws are identical on the observed sample because $t^\star$ has zero exact support. However,

$$\tau_+(x; t^\star, t_0) - \tau_-(x; t^\star, t_0) = 2Ga = GL_F r^\beta. \tag{86}$$

Since the observed distributions are identical, their total variation distance is zero. The standard two-point testing argument implies that the maximum risk of any estimator is at least a universal constant times this separation. This proves the minimax lower bound. The final statement follows because a categorical-ID representation is invariant under arbitrary relabelings of unsupported policy IDs and therefore cannot recover the metric radius $r_x(t^\star)$ from observed labels alone. □

C.9 Stability of the Policy-Mixture Embedding

The local-overlap theorem relies on the policy distance d_Π. We record here a simple Lipschitz property of the POUL embedding that justifies the effectiveness condition used in Assumption 3.1.

Proposition C.7 (Stability of the embedding and uplift)**.** *Assume that*

$$z(t) = \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} w(s)\Pi_t(a \mid s)\phi(s, a), \qquad h(t) = \rho(z(t)). \tag{87}$$

Suppose ρ is L_ρ-Lipschitz, i.e.,

$$\|\rho(u) - \rho(v)\| \leq L_\rho \|u - v\| \tag{88}$$

for all u, v, and the atom embedding is bounded:

$$\sup_{(s,a)} \|\phi(s, a)\| \leq B. \tag{89}$$

Then, for any $t, t' \in \mathcal{T}$,

$$\|h(t) - h(t')\| \leq L_\rho \|z(t) - z(t')\| \leq L_\rho B d_\Pi(t, t'). \tag{90}$$

Moreover, suppose the oracle embedding has the same policy-mixture form

$$h_0(t) = \rho_0(z(t)), \tag{91}$$

where ρ_0 is also L_ρ-Lipschitz. If $\|g_0(x)\| \leq G$ and

$$\tau(x; t) = g_0(x)^\top h_0(t), \tag{92}$$

then the oracle uplift satisfies

$$|\tau(x; t) - \tau(x; t')| = |g_0(x)^\top (h_0(t) - h_0(t'))| \leq GL_\rho B d_\Pi(t, t'). \tag{93}$$

Interpretation. Proposition C.7 states that the learned policy representation is stable under small perturbations of the policy specification. In other words, if two policies induce similar context–action mixtures, then their embeddings are close, and the corresponding uplift predictions are also close.

This property is important because the relevant question in a long-tailed policy system is not merely whether $h(t)$ changes smoothly when the policy t changes. The key question is whether this smoothness can replace the unrealistic requirement that every target policy must be repeatedly randomized and observed as an exact ID. The support-replacement results answer this question affirmatively, up to the transfer bias determined by the distance from the target policy to the observed support.

Thus, the policy-induced mixture representation provides a transfer principle: repeated exposure to the exact same policy can be partially substituted by exposure to nearby policies that share context–action components. This is precisely the regime where categorical treatment IDs are statistically inefficient, because they cannot share information across similar policies.

Connection to the main theory. Assumption 3.1 provides the structural interface: policies affect outcomes through induced context–action mixtures. The supported-policy orthogonal result controls estimation error when exact policy support is available. The local policy-overlap theorem extends this guarantee to rare or unseen policies by paying a transfer bias determined by their distance to the observed support. The minimax lower bound shows that this bias is unavoidable. Expressiveness and Lipschitz stability are therefore not the final claim by themselves; they are the mechanisms that make the sharp support-replacement theorem possible.

C.10 Proof of Proposition C.7

Proof of Proposition C.7. Recall that

$$z(t) = \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} w(s)\Pi_t(a \mid s)\,\phi(s,a), \qquad h(t) = \rho(z(t)). \tag{94}$$

For any $t, t' \in \mathcal{T}$, we have

$$z(t) - z(t') = \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} w(s)\Big(\Pi_t(a \mid s) - \Pi_{t'}(a \mid s)\Big)\phi(s,a). \tag{95}$$

Taking norms and using the triangle inequality,

$$\begin{aligned}
\|z(t) - z(t')\| &\le \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} w(s)\,|\Pi_t(a \mid s) - \Pi_{t'}(a \mid s)|\,\|\phi(s,a)\| \\
&\le \Big(\sup_{(s,a)} \|\phi(s,a)\|\Big) \sum_{s\in\mathcal{S}} w(s) \sum_{a\in\mathcal{A}} |\Pi_t(a \mid s) - \Pi_{t'}(a \mid s)| \\
&\le B \sum_{s\in\mathcal{S}} w(s) \|\Pi_t(\cdot \mid s) - \Pi_{t'}(\cdot \mid s)\|_1 \\
&= B\, d_\Pi(t,t').
\end{aligned}$$

Since ρ is L_ρ-Lipschitz,

$$\|h(t) - h(t')\| = \|\rho(z(t)) - \rho(z(t'))\| \le L_\rho \,\|z(t) - z(t')\| \le L_\rho B\, d_\Pi(t,t'). \tag{96}$$

The same argument applies to the oracle embedding $h_0(t) = \rho_0(z(t))$, because ρ_0 is also L_ρ-Lipschitz. Hence,

$$\|h_0(t) - h_0(t')\| \le L_\rho B\, d_\Pi(t,t'). \tag{97}$$

Finally, since $\tau(x;t) = g_0(x)^\top h_0(t)$ and $\|g_0(x)\| \le G$, Cauchy–Schwarz gives

$$\begin{aligned}
|\tau(x;t) - \tau(x;t')| &= |g_0(x)^\top (h_0(t) - h_0(t'))| \\
&\le \|g_0(x)\|\,\|h_0(t) - h_0(t')\| \\
&\le G L_\rho B\, d_\Pi(t,t').
\end{aligned}$$

This completes the proof. □

To reduce overfitting bias when flexible learners are used for nuisance components, we optionally adopt cross-fitting: estimate nuisance functions on held-out folds, and train the main model using fold-specific nuisance predictions.

D Experimental Implementation Details

Experimental Overview. The industrial ride-hailing experiments assess POUL through baseline comparisons and component ablations. The synthetic benchmark evaluates treatment-effect estimation across observed, tail, and held-out-near policies using known potential outcomes. This appendix presents implementation details for both experimental settings described in Section 4: the large-scale randomized industrial experiment and the synthetic combinatorial-policy benchmark. We describe the **data construction**, **training protocols**, and **optimization settings**, and report **baseline comparisons** and **component ablations**.

D.1 Real-world Industrial Experiment Details

D.1.1 Industrial Dataset and Training Protocol

Dataset. The industrial experiment is conducted on a **large-scale randomized ride-hailing dataset**. Each user is represented by 723 features. We consider two populations: the Global population $\mathcal{D}_G$ with approximately

Algorithm 2 Permutation-invariant orthogonal uplift learning (POUL)

Require: Data $\{(X_i, T_i, Y_i)\}_{i=1}^n$, policy specs $\{\Pi_t(\cdot \mid s)\}_{t,s}$, weights $w(\cdot)$, folds K (optional)
Split indices into folds $I_1, \ldots, I_K$ (set $K = 1$ to disable cross-fitting)
for $k = 1, \ldots, K$ **do**
  Fit nuisance $\hat{m}^{(-k)}(\cdot) \approx \mathbb{E}[Y \mid X = \cdot]$ using $\{i \notin I_k\}$
  (Observational only) Fit propensity $\hat{\pi}^{(-k)}(t \mid x) \approx \mathbb{P}(T = t \mid X = x)$ using $\{i \notin I_k\}$
end for
Initialize parameters (ω, ν) for ϕ_ω, ρ_ν and (η) for g_η; optionally initialize m_θ
repeat
  Sample a mini-batch $\mathcal{B} \subset \{1, \ldots, n\}$
  for each $i \in \mathcal{B}$ **do**
    Compute $h_{\omega,\nu}(T_i)$ via (2)–(3)
    Let $k(i)$ be the fold index such that $i \in I_{k(i)}$
    Set $\hat{e}_h(X_i) \leftarrow \sum_{t \in \mathcal{T}} \hat{\pi}^{(-k(i))}(t \mid X_i)\, h_{\omega,\nu}(t)$ {use known π in RCT}
    Predict $\hat{Y}_i \leftarrow \hat{m}^{(-k(i))}(X_i) + g_\eta(X_i)^\top \big(h_{\omega,\nu}(T_i) - \hat{e}_h(X_i)\big)$
  end for
  Update (η, ω, ν) (and optionally θ) by SGD on $\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \ell(\hat{Y}_i, Y_i) + \lambda\Omega$
until convergence
return $\hat{\tau}(x; t_1, t_0) = g_{\hat{\eta}}(x)^\top \big(h_{\hat{\omega},\hat{\nu}}(t_1) - h_{\hat{\omega},\hat{\nu}}(t_0)\big)$

5.28 million users, and its Core Eligible subset $\mathcal{D}_C$ with approximately 1.58 million users satisfying the eligibility criteria for service-class upgrade interventions.

The industrial treatment space contains two treatment settings. Treatment T_1 corresponds to an *Economy-to-Express* service-class upgrade, while treatment T_2 is a composite intervention consisting of *Discount-to-Economy* and *Economy-to-Express*. Treatment triggering depends on user opt-in behavior and operational scenario constraints. Specifically, for valid *Economy-to-Express* requests, users are randomly assigned to T_1 or T_2; *Discount-to-Economy* requests exclusively trigger T_2. For both treatment settings, we evaluate two business outcomes: monthly Gross Merchandise Value (GMV), used as an observable surrogate for Lifetime Value (LTV), and monthly operational Cost.

Baselines. We compare POUL with the nonparametric **Causal Forest** [Wager and Athey, 2018] and three MLP-based baselines: **TARNet** [Shalit et al., 2017], **DESCN** [Zhong et al., 2022], and **DRCFR** [Cheng et al., 2022]. TARNet follows the T-learner's outcome-regression strategy [Künzel et al., 2019], using a shared DNN representation and treatment-specific outcome heads jointly optimized with factual regression losses.

Training Settings. For the industrial experiments, all deep learning models are implemented in PyTorch, and training is conducted on NVIDIA L20 GPUs. We employ the **Adam** optimizer with a fixed **learning rate** of 1×10^{-4} and a **batch size** of 512.

Keeping these settings fixed, we tune the training strategy (single-stage versus two-stage) and epoch configuration (i, j) for each neural network model. Here, i and j denote the numbers of epochs for the first stage (e.g., $\mathcal{D}_G$ pre-training) and the second stage (e.g., $\mathcal{D}_C$ fine-tuning), respectively. For Causal Forest, we tune the number of trees while keeping the remaining hyperparameters fixed.

D.1.2 Ablation Studies

We assess the contributions of four POUL components through individual ablations: **Structured Encoding**, **Covariate Heterogeneity**, **Baseline Net**, and **Representation Centering**. Each variant modifies one component while retaining the remaining model design and is trained separately for comparison with the full model. For brevity, we use the same function notation across variants, although their fitted parameters differ.

Full model. Let $X \in \mathbb{R}^p$ denote user and scenario covariates, $s \in \mathcal{S}_{\text{ind}} = \{A, B, AB\}$ the user's preference segment, and $t \in \mathcal{T}_{\text{ind}} = \{0, 1, 2\}$ the policy ID, with $t = 0$ denoting control.

For outcome $r \in \{\text{GMV}, \text{Cost}\}$, the industrial implementation predicts

$$\widehat{Y}_r(X, s, t) = \underbrace{m_r(X)}_{\text{baseline prediction}} + \underbrace{g_r(X)^\top \big[h_r(s, t) - \bar{h}_r\big]}_{\text{effect interaction}}. \tag{98}$$

Here, $m_r(X) \in \mathbb{R}$ predicts the baseline outcome level, $g_r(X) \in \mathbb{R}^d$ provides user-dependent effect coefficients, and $h_r(s, t) \in \mathbb{R}^d$ represents the policy within the user's preference segment, with $d = 32$. Each outcome has

its own baseline predictor, coefficient network, treatment encoder, and running-mean statistic. The baseline predictor is trained on factual outcomes pooled across policy arms and frozen during second-stage effect learning.

The structured treatment representation is defined by

$$h_r(s,t) = \rho_r\left(e_{s,r} + \sum_{a' \in \mathcal{A}_{\text{ind}}} p(a' \mid s,t) e_{a',r}\right), \qquad \mathcal{A}_{\text{ind}} = \{a, b, \varnothing\}. \tag{99}$$

The vectors $e_{s,r}$ and $e_{a',r}$ are learnable segment and action embeddings in a common embedding space. The symbol $\varnothing$ denotes no action. The fixed weights $p(a' \mid s,t)$ encode the prescribed action combination for segment s under policy t and sum to 1. The learnable nonlinear map ρ_r transforms the combined embedding into a d-dimensional treatment representation.

The centering term $\bar{h}_r$ is a sample-count-weighted running mean of treatment representations. For a mini-batch $\mathcal{B}$ of size B, its update is

$$\begin{aligned} \mu_{\mathcal{B},r} &= \frac{1}{B}\sum_{i \in \mathcal{B}} h_r^+(s_i, t_i), \\ \bar{h}_r^{\text{new}} &= \frac{N\bar{h}_r + B\mu_{\mathcal{B},r}}{N+B}, \end{aligned} \tag{100}$$

where N counts samples accumulated before the current batch, and h_r^+ denotes the encoder evaluated after the batch's three parameter updates. These three updates use the same historical mean $\bar{h}_r$. The mean is then updated without gradient propagation, and the sample counter is increased by B.

The running mean and sample counter are reset at the start of each epoch. The final running mean is retained for inference. Because encoder parameters change during training, this statistic aggregates representations computed at different optimization steps rather than representations recomputed using the final encoder.

w/o Structured Encoding. We replace the structured treatment encoder with a policy-ID encoder. Each policy $t \in \{0,1,2\}$ receives a learnable embedding $e_{t,r}$ in the original embedding space. The segment embeddings are retained but averaged into a shared learnable offset:

$$c_r = \frac{e_{A,r} + e_{B,r} + e_{AB,r}}{3} \tag{101}$$

$$h_r^{\text{ID}}(t) = \rho_r(e_{t,r} + c_r). \tag{102}$$

Since c_r is common to all users, the encoder no longer receives the current user's segment or applies the policy-to-action composition rules. A given policy therefore has the same treatment representation across segments.

The predictor becomes

$$\widehat{Y}_r^{\text{ID}}(X,s,t) = m_r(X) + g_r(X)^\top\left[h_r^{\text{ID}}(t) - \bar{h}_r^{\text{ID}}\right]. \tag{103}$$

This variant retains the embedding dimension, nonlinear mapping, baseline predictor, and covariate-dependent coefficient network. Centering uses the running mean of this variant's own representations, following the same update rule as the full model. User-level effect variation remains through $g_r(X)$. The ablation therefore assesses the joint contribution of segment-dependent encoding and action composition.

w/o Covariate Heterogeneity. We replace the covariate input to the coefficient network with a zero vector, yielding a shared learnable coefficient $a_r = g_r(0)$. The predictor becomes

$$\widehat{Y}_r^{\text{CH}}(X,s,t) = m_r(X) + a_r^\top\left[h_r(s,t) - \bar{h}_r\right]. \tag{104}$$

The baseline predictor still uses user covariates, and the treatment encoder retains segment information. This variant therefore removes covariate-driven coefficient variation while preserving segment-level differences in policy effects.

w/o Baseline Net. We set the baseline predictor's contribution to zero:

$$\widehat{Y}_r^{\text{B}}(X,s,t) = g_r(X)^\top\left[h_r(s,t) - \bar{h}_r\right]. \tag{105}$$

The baseline network's parameter structure is retained, but its output does not enter the prediction. The effect components are retrained using the original factual outcomes rather than baseline-adjusted residuals.

w/o Representation Centering. We omit running-mean subtraction from the treatment representation:

$$\widehat{Y}_r^{\text{C}}(X,s,t) = m_r(X) + g_r(X)^\top h_r(s,t). \tag{106}$$

The running statistics are still maintained but do not enter the prediction. This variant evaluates the representation-centering operation used in the industrial implementation.

Training and evaluation. The baseline predictor is first trained on factual outcomes from all policy arms and then frozen during second-stage effect learning. Each variant is trained separately using its modified prediction rule, retaining the factual regression objective, task-weighting scheme, and monotonicity auxiliary loss. We report AUUC for GMV and Cost under T_1 and T_2 on the Core Eligible test subset $\mathcal{D}_C$. Policy-effect scores are computed by comparing each treatment with control for the same user and segment.

Results. Table 2 reports the component ablation results. The full model achieves the highest mean AUUC across all four outcomes.

Table 2: Component ablations on the Core Eligible test subset. AUUC is reported as mean ± sample SD over five runs per model (higher is better).

Model	GMV T_1	GMV T_2	Cost T_1	Cost T_2
POUL (Full model)	$\mathbf{0.8615 \pm 0.0317}$	$\mathbf{0.6537 \pm 0.0420}$	$\mathbf{0.8014 \pm 0.1124}$	$\mathbf{0.9097 \pm 0.0003}$
w/o Structured Encoding	0.3415 ± 0.0270	0.6185 ± 0.0313	0.6661 ± 0.0030	0.6673 ± 0.0025
w/o Covariate Heterogeneity	0.6318 ± 0.0227	0.5240 ± 0.0374	0.4921 ± 0.0239	0.7157 ± 0.0004
w/o Baseline Net	0.5584 ± 0.1169	0.5693 ± 0.0319	0.5631 ± 0.0203	0.7628 ± 0.0006
w/o Representation Centering	0.6955 ± 0.0122	0.5791 ± 0.0121	0.5800 ± 0.0163	0.7633 ± 0.0011

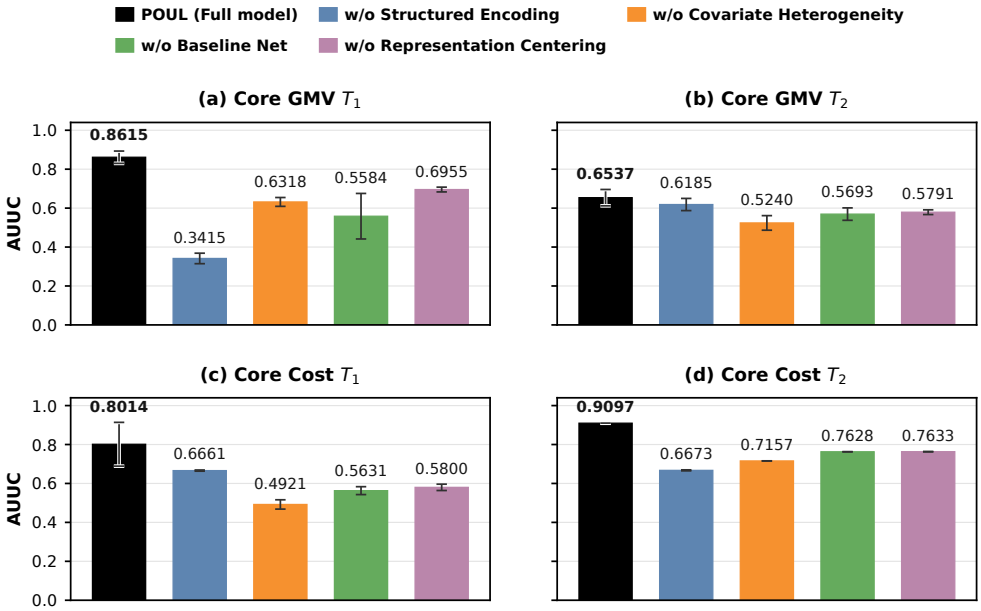


Figure 3: Component ablations on the Core Eligible test subset. Bars show mean AUUC and error bars show sample standard deviations over five runs per model.

D.2 Synthetic Combinatorial-Policy Benchmark

Motivation. Our main experiment uses a large-scale randomized ride-hailing dataset for causal evaluation in a realistic industrial setting. However, real data do not reveal all potential outcomes, limiting direct evaluation of individual-level causal estimation errors for rare or unseen policies. We therefore construct a synthetic combinatorial-policy benchmark with known potential outcomes. This benchmark tests whether representing policies through their reusable context–action structure improves generalization to long-tail and held-out policies compared with representing each policy as an isolated ID.

Experimental Design. We construct a synthetic combinatorial-policy benchmark with known potential outcomes to evaluate generalization to rare and held-out policies. Each unit has 100 covariates, of which the first 20 contain signal and the remaining 80 are noise. Each treatment is a structured policy $\{\Pi_t(\cdot \mid s)\}_{s \in \mathcal{S}}$ over $|\mathcal{S}| = 8$ treatment-side contexts and $|\mathcal{A}| = 5$ atomic actions, specifying an action distribution for each context.

We construct 30 observed policies with a long-tailed frequency distribution, including low-frequency tail policies. We additionally generate held-out-near policies as nearby perturbations of observed policies in policy specification distance. These policies are excluded from training but share context–action components with observed policies. The tail and held-out settings therefore evaluate generalization under limited observations and to previously unseen policy compositions, respectively.

For each unit with covariates x, the potential outcome under policy t is

$$Y(t) = m^\star(x) + g_t^\star(x)^\top h^\star(t) + \epsilon, \tag{107}$$

where $m^\star(x)$ is the baseline outcome, $h^\star(x)$ is a shared covariate-dependent representation, $g_t^\star(x)$ is a policy-induced treatment-specific component, and ϵ is random noise. The known potential outcomes enable direct evaluation of both individual-level and aggregate treatment-effect estimation errors.

To exploit the compositional structure of treatments, POUL encodes each policy through its induced context–action mixture:

$$z(t) = \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} w(s)\Pi_t(a \mid s)\phi(s,a), \qquad h(t) = \rho(z(t)), \tag{108}$$

where $w(s)$ is the context weight, $\phi(s,a)$ represents a context–action atom, and ρ is a learnable map. This representation is invariant to the ordering of context–action components and allows information sharing across policies through shared atoms.

Baselines and ablations. We compare structured and ID-based policy encoding across four backbones: POUL, X-network [Zhong et al., 2022], TARNet [Shalit et al., 2017], and DRCFR [Cheng et al., 2022]. The structured variants encode each policy through its context–action composition, whereas the ID variants use a learned embedding for each observed policy ID. For held-out policies, the structured variants encode the target policy directly, while the ID variants use predictions for the nearest observed policy in policy specification distance.

Computing Environment We conducted the synthetic benchmark experiments on a Linux server equipped with two NVIDIA GeForce RTX 4090 GPUs, each with 24 GB of memory. We used Python 3.12.3 and PyTorch 2.8.0 with CUDA 12.8. Each training run used a single GPU. We scheduled independent runs with different hyperparameter configurations and random seeds in parallel across the two GPUs.

Results. Structured encoding yields lower mean error in 19 of 24 within-backbone comparisons, including all six comparisons for POUL. Relative to POUL-ID, POUL-Structure reduces mean $\sqrt{\text{PEHE}}$ by 7.8%, 9.4%, and 8.4% on observed, tail, and held-out-near policies, respectively. Across all eight variants, **POUL-Structure achieves the lowest mean $\sqrt{\text{PEHE}}$ in all three regimes**. DRCFR-Structure achieves the lowest ATE error in all three regimes. POUL's ATE error varies across regimes and is generally close to the lowest observed error in each regime. These results support the use of structured policy encoding for estimating heterogeneous effects of rare and previously unseen policies in this benchmark.

Table 3: Synthetic benchmark results over 20 test seeds, reported as mean $\pm$ sample standard deviation. Lower is better for both metrics.

	Observed		Tail		Held-out near	
Model	$\sqrt{\text{PEHE}}$	ATE error	$\sqrt{\text{PEHE}}$	ATE error	$\sqrt{\text{PEHE}}$	ATE error
POUL-Structure	$\mathbf{0.0993 \pm 0.0175}$	0.0222 ± 0.0057	$\mathbf{0.1034 \pm 0.0185}$	0.0274 ± 0.0093	$\mathbf{0.0970 \pm 0.0200}$	0.0214 ± 0.0082
POUL-ID	0.1077 ± 0.0183	0.0258 ± 0.0070	0.1141 ± 0.0209	0.0340 ± 0.0137	0.1059 ± 0.0190	0.0251 ± 0.0078
X-network-Structure	0.1097 ± 0.0207	0.0244 ± 0.0090	0.1126 ± 0.0218	0.0283 ± 0.0096	0.1070 ± 0.0223	0.0225 ± 0.0084
X-network-ID	0.1108 ± 0.0195	0.0219 ± 0.0055	0.1150 ± 0.0225	0.0287 ± 0.0098	0.1085 ± 0.0215	0.0204 ± 0.0070
TARNet-Structure	0.1140 ± 0.0176	0.0223 ± 0.0064	0.1178 ± 0.0185	0.0289 ± 0.0103	0.1114 ± 0.0182	0.0213 ± 0.0076
TARNet-ID	0.1122 ± 0.0171	0.0278 ± 0.0106	0.1166 ± 0.0172	0.0314 ± 0.0130	0.1087 ± 0.0176	0.0255 ± 0.0115
DRCFR-Structure	0.1090 ± 0.0178	$\mathbf{0.0198 \pm 0.0058}$	0.1127 ± 0.0195	$\mathbf{0.0247 \pm 0.0089}$	0.1069 ± 0.0188	$\mathbf{0.0185 \pm 0.0063}$
DRCFR-ID	0.1100 ± 0.0162	0.0291 ± 0.0099	0.1141 ± 0.0174	0.0342 ± 0.0108	0.1070 ± 0.0180	0.0270 ± 0.0098

E Evaluation Metrics

We use different metrics for the industrial and synthetic experiments. In the industrial randomized experiment, ground-truth individual treatment effects are unobservable, so we evaluate uplift ranking using AUUC. In the synthetic benchmark, potential outcomes are known, so we additionally report root PEHE and ATE error.

E.1 Uplift Ranking Capability (AUUC)

We evaluate ranking performance using the normalized **Area Under the Uplift Curve (AUUC)**, which quantifies the cumulative incremental gain achieved when targeting samples sorted by predicted Individual Treatment Effect (ITE).

Formally, let $\mathcal{D} = \{(x_i, t_i, y_i)\}_{i=1}^n$ denote the test dataset, where x_i is the feature vector, $t_i \in \{0, 1\}$ indicates the treatment assignment, and y_i represents the observed outcome. Let $\hat{\tau}(x_i)$ be the predicted ITE.

We sort the samples in descending order of $\hat{\tau}(x_i)$. Let $i_1, \ldots, i_n$ be the indices of the sorted samples such that $\hat{\tau}(x_{i_1}) \geq \cdots \geq \hat{\tau}(x_{i_n})$. For the top k samples, we define the cumulative counts (n_t, n_c) and outcome sums (y_t, y_c) for the treated and control groups as:

$$n_t(k) = \sum_{j=1}^{k} \mathbb{I}(t_{i_j} = 1), \qquad n_c(k) = \sum_{j=1}^{k} \mathbb{I}(t_{i_j} = 0), \tag{109}$$

$$y_t(k) = \sum_{j=1}^{k} y_{i_j} \mathbb{I}(t_{i_j} = 1), \qquad y_c(k) = \sum_{j=1}^{k} y_{i_j} \mathbb{I}(t_{i_j} = 0). \tag{110}$$

The *Lift* at rank k estimates the average treatment effect within the top k units:

$$\text{Lift}(k) = \frac{y_t(k)}{n_t(k)} - \frac{y_c(k)}{n_c(k)}, \tag{111}$$

The **Cumulative Gain**, $G(k)$, is defined as the total estimated uplift up to rank k:

$$G(k) = \text{Lift}(k) \times k. \tag{112}$$

To facilitate comparison across different settings, we calculate the **Normalized Cumulative Gain** $\tilde{G}(k)$ by scaling $G(k)$ with the absolute global gain at $k = n$:

$$\tilde{G}(k) = \frac{G(k)}{|G(n)|}. \tag{113}$$

Finally, the **Normalized AUUC** is computed as the average of these normalized gains across the *entire* test dataset:

$$\text{AUUC} = \frac{1}{n} \sum_{k=1}^{n} \tilde{G}(k). \tag{114}$$

E.2 Root PEHE and ATE Error

PEHE (precision in estimation of heterogeneous effects) measures the mean squared error between estimated and true individual treatment effects. For the synthetic benchmark, we report its square root (root PEHE, $\sqrt{\text{PEHE}}$) and policy-level ATE error.

For a policy pair (t, t_0), root PEHE is defined as

$$\sqrt{\text{PEHE}} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\hat{\tau}(x_i; t, t_0) - \tau(x_i; t, t_0)\right)^2}. \tag{115}$$

where $\hat{\tau}(x_i; t, t_0)$ is the predicted individual uplift and $\tau(x_i; t, t_0)$ is the ground-truth individual treatment effect.

The policy-level ATE error measures aggregate calibration:

$$\text{ATE Err.} = \left| \frac{1}{n} \sum_{i=1}^{n} \hat{\tau}(x_i; t, t_0) - \frac{1}{n} \sum_{i=1}^{n} \tau(x_i; t, t_0) \right|. \tag{116}$$

Both metrics are averaged uniformly over policies within each regime (observed, tail, and held-out).

F Supplementary Evaluations and Discussion

F.1 Comparison with Ranking-Style Auxiliary Learning

Motivation. Zhou et al. [2023] propose a causal learning objective that ranks individuals by their estimated incremental response to support marketing resource allocation. Inspired by their Eq. (2), we test whether **adding the ranking auxiliary loss in** (118) **to POUL** improves treatment effect estimation in our **synthetic benchmark**.

Let B be the set of indices in a minibatch and let $\mathcal{T}_{\text{obs}}$ denote the non-control policies observed during training. For each $t \in \mathcal{T}_{\text{obs}}$, we apply a softmax across individuals in the minibatch to the estimated effects relative to the control policy t_0:

$$q_{it} = \frac{\exp\big(\widehat{\tau}(X_i; t, t_0)\big)}{\sum_{j \in B} \exp\big(\widehat{\tau}(X_j; t, t_0)\big)}. \tag{117}$$

Under complete randomization, the known propensity $\pi_0(t \mid X_i)$ is constant across individuals and is written as $\pi_0(t)$. Using q_{it} from (117), we define

$$\mathcal{L}_{\text{DRAux}} = -\frac{1}{|B|\,|\mathcal{T}_{\text{obs}}|} \sum_{i \in B} \sum_{t \in \mathcal{T}_{\text{obs}}} Y_i \left(\frac{\mathbf{1}\{T_i = t\}}{\pi_0(t)} - \frac{\mathbf{1}\{T_i = t_0\}}{\pi_0(t_0)} \right) \log\big(|B| q_{it}\big). \tag{118}$$

We denote the resulting variant by POUL+DRAux and add $\lambda_{\text{DRAux}} \mathcal{L}_{\text{DRAux}}$ to the second-stage training objective of POUL.

Results. Table 4 shows that adding the DRAux ranking loss does not improve treatment effect estimation accuracy in this benchmark. Across the same 20 test seeds, the separately tuned POUL+DRAux variant has slightly higher mean root PEHE and ATE error than POUL in all three policy regimes.

Table 4: Synthetic benchmark results for POUL and POUL with the DRAux ranking loss. Both variants were evaluated on the same 20 test seeds. Results are reported as mean $\pm$ sample standard deviation.

	Observed		Tail		Held-out near	
Method	$\sqrt{\text{PEHE}}$	ATE error	$\sqrt{\text{PEHE}}$	ATE error	$\sqrt{\text{PEHE}}$	ATE error
POUL	$\mathbf{0.0993 \pm 0.0175}$	$\mathbf{0.0222 \pm 0.0057}$	$\mathbf{0.1034 \pm 0.0185}$	$\mathbf{0.0274 \pm 0.0093}$	$\mathbf{0.0970 \pm 0.0200}$	$\mathbf{0.0214 \pm 0.0082}$
POUL + DRAux	0.1013 ± 0.0172	0.0234 ± 0.0067	0.1052 ± 0.0183	0.0284 ± 0.0112	0.0983 ± 0.0194	0.0224 ± 0.0080

F.2 Budget-Constrained Treatment Allocation

Motivation. In many practical applications, **decision-making** follows a two-stage pipeline that combines machine learning (ML) for prediction with operations research (OR) for optimization. Within this pipeline, uplift estimates serve as inputs to downstream treatment allocation. For example, in ride-hailing incentive allocation, a platform may need to select which users receive which treatment subject to a subsidy budget. Therefore, beyond AUUC, we evaluate whether the learned reward and cost uplifts lead to competitive downstream decision quality under budget constraints.

This evaluation is related to budgeted policy evaluation and decision-focused marketing optimization, where expected outcome metrics or budget-aware objectives are commonly used to evaluate offline policy quality [Yan et al., 2023, Zhou et al., 2024]. In our setting, the downstream allocation results should be viewed as complementary to the main causal estimation results.

Allocation procedure. Given predicted reward uplift and predicted cost uplift, the allocator selects treatment assignments subject to a realized subsidy-budget constraint. Let $\widehat{\Delta}^Y_{i,t}$ denote the predicted reward uplift and $\widehat{\Delta}^C_{i,t}$ denote the predicted cost uplift for assigning treatment t to unit i. A standard budgeted allocation problem can be written as

$$\max_{\{z_{i,t}\}} \sum_{i,t} z_{i,t} \widehat{\Delta}^Y_{i,t} \quad \text{s.t.} \quad \sum_{i,t} z_{i,t} \widehat{\Delta}^C_{i,t} \le B, \qquad \sum_{t} z_{i,t} \le 1, \qquad z_{i,t} \in \{0,1\}, \tag{119}$$

where B is the available subsidy budget. In practice, the allocator ranks or optimizes candidate assignments using the predicted reward and cost uplifts and evaluates the resulting treatment assignment by realized budget-aligned GMV lift.

Baselines. We compare POUL with strong decision-oriented baselines, including DESCN[Zhong et al., 2022], DRCFR[Cheng et al., 2022] and DFCL [Zhou et al., 2024]. These methods are designed to improve downstream decision performance. POUL is different in focus: it aims to improve the structured-policy causal prediction by representing treatments as policy mixtures over context–action atoms. The comparison therefore evaluates whether the causal-estimation benefits of POUL also translate into downstream allocation quality.

Experimental Design. We evaluate downstream allocation on the Core Eligible population $\mathcal{D}_C$ using the Expected Outcome Metric (EOM). The simulation uses a separate sample from the same industrial ride-hailing setting, collected during a different period from the data used in the main industrial experiments. For each method, we fix a trained checkpoint across all budget regimes and vary the cost penalty λ in the allocation score $\widehat{R} - \lambda\widehat{C}$, where $\widehat{R}$ and $\widehat{C}$ denote predicted revenue and subsidy cost, respectively. Each value of λ yields an allocation with a corresponding subsidy expenditure and GMV lift. Because a given λ can yield different expenditures across methods, we compare methods at matched levels of subsidy expenditure. We report the estimated GMV lift near an expenditure of 0.20 for the high-budget regime and the minimum and maximum estimated lifts within 0.10–0.12 and 0.05–0.06 for the medium- and low-budget regimes, respectively.

Budget-aligned GMV lift. Table 5 reports absolute GMV lift under high-, medium-, and low-budget regimes. The budget values correspond to realized subsidy-budget ranges in the evaluation slice. POUL remains competitive overall and shows its clearest advantage in the practically relevant medium- and low-budget regimes.

Table 5: Budget-aligned downstream GMV lift on the Core Eligible population $\mathcal{D}_C$. Higher is better. GMV lift values are estimated from the evaluation curves, and ranges show the minimum and maximum within each realized subsidy-budget interval.

Model	High budget ($\sim$0.20)	Medium budget (0.10–0.12)	Low budget (0.05–0.06)
POUL	$\sim$**0.91%**	$\sim$**0.62–0.75%**	$\sim$**0.30–0.33%**
DESCN	$\sim$0.52%	$\sim$0.39–0.50%	$\sim$0.08–0.10%
DFCL	$\sim$0.84%	$\sim$0.28–0.36%	$\sim$0.18–0.24%
DRCFR	$\sim$0.67%	$\sim$0.38–0.46%	$\sim$0.22–0.24%

F.3 Compatibility with Decision-Focused Optimization

The proposed method is not isolated from decision-focused or predict-then-optimize approaches. Instead, POUL can be used as a structured-policy causal prediction layer inside a downstream optimization framework.

Specifically, the proposed treatment embedding can be written as

$$h_\theta(t) = \rho_\theta\left(\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} w(s)\Pi_t(a \mid s)\phi_\theta(s,a)\right), \tag{120}$$

where $w(s)$ denotes the context weight, $\Pi_t(a \mid s)$ is the policy-induced action distribution under context s, $\phi_\theta(s,a)$ is the learned context–action atom representation, and ρ_θ is a learnable transformation.

Using the POUL policy representation, we predict reward and cost uplifts relative to the control policy t_0 as

$$\widehat{\tau}_i^Y(t) = \widehat{g}_Y(x_i)^\top\left(h_\theta(t) - h_\theta(t_0)\right), \tag{121}$$

$$\widehat{\tau}_i^C(t) = \widehat{g}_C(x_i)^\top\left(h_\theta(t) - h_\theta(t_0)\right), \tag{122}$$

where $\widehat{g}_Y(x_i)$ and $\widehat{g}_C(x_i)$ are the learned covariate-dependent coefficient vectors for reward and cost, respectively. These predictions can be fed into a downstream allocator

$$z^\star(B;\Theta) = \arg\max_z \sum_{i,t} z_{i,t}\widehat{\tau}_i^Y(t) \quad \text{s.t.} \quad \sum_{i,t} z_{i,t}\widehat{\tau}_i^C(t) \le B, \qquad \sum_t z_{i,t} = 1. \tag{123}$$

A DFCL-style end-to-end objective can then be written as

$$\mathcal{L}_{\text{total}}(\Theta) = \mathcal{L}_{\text{POUL}}(\Theta) + \lambda\mathcal{L}_{\text{dec}}\left(z^\star(B;\Theta)\right), \tag{124}$$

where $\mathcal{L}_{\text{POUL}}$ is the original orthogonal uplift loss and $\mathcal{L}_{\text{dec}}$ is a downstream decision loss or EOM-based objective. This gives a more general framework: POUL improves the structured-policy causal prediction layer, while a DFCL-style objective can further optimize the allocation layer built on top of it.

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 claims in the abstract and introduction are aligned with the paper's contributions: the paper proposes permutation-invariant treatment embeddings, integrates them into an orthogonalized low-rank uplift model, provides expressiveness/robustness/stability analysis, and validates the method on large-scale randomized experiments. These contributions are explicitly summarized in the introduction and are supported by the theory and experiments in Sections 3 and 4.

   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: Appendix B discusses the assumptions for causal identification and transfer to nearby policies, as well as model misspecification, noisy context weights, and the computational cost of the embedding. Section 5 discusses unobserved confounding and changes in policy assignments and available actions as directions for future work.

   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 used for identification and theoretical analysis, including SUTVA/randomization/overlap and policy-induced mixture sufficiency. The main theoretical results on expressiveness, orthogonal robustness, and stability are stated in Section 3, with formal proofs provided in Appendix C.

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 paper describes the model formulation and experimental setup. Appendix D provides details on the construction and scale of the industrial and synthetic datasets, treatment definitions, baselines, training procedures, hyperparameter search, optimization settings (including the optimizer, learning rate, and batch size), and computing hardware. Appendix E defines the evaluation metrics, and Appendix F describes the settings and results of supplementary evaluations, including ranking-based auxiliary learning and budget-constrained treatment allocation. The industrial dataset is proprietary and cannot be publicly released. The documented experimental protocols and implementation details support internal reproduction and help external readers understand how the experiments were conducted. We plan to publicly release the code for the synthetic experiments, including the data-generation scripts, to facilitate independent reproduction.

Guidelines:

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

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

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

   Answer: [Yes]

   Justification: The industrial ride-hailing dataset contains proprietary platform data and cannot be publicly released. The paper nevertheless reports the data construction, treatment definitions, model architecture, training protocol, baselines, and evaluation metrics in detail. For the experiments on synthetic datasets, we will make the code and data available for open-source later.

   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: The paper specifies the experimental setting, including the industrial dataset size, feature dimension, population definitions, treatment settings, outcomes, baselines, evaluation metrics, optimizer, learning rate, batch size, GPU type, and hyperparameter tuning strategy. Additional training and evaluation details are provided in Appendices D, E and F

   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: Figure 2 and Tables 3 report means and sample standard deviations across 20 synthetic test seeds. Tables 1 report means and sample standard deviations across five industrial runs, and Figure 3 shows the sample standard deviations for the ablation results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Appendix D.1.1 identifies NVIDIA L20 GPUs for the industrial neural models, and Appendix D.2 reports two NVIDIA GeForce RTX 4090 GPUs with 24 GB each and one GPU per synthetic run.

   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 research was conducted using randomized platform data under standard privacy and data-protection practices. The study focuses on causal estimation and offline evaluation of treatment assignment policies, and no personally identifying information is disclosed in the paper.

   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: The work may improve the reliability and efficiency of data-driven treatment allocation in domains such as ride-hailing incentives, dynamic pricing, and personalized services. Potential negative impacts include unfair allocation, privacy concerns, or misuse of uplift scores for overly aggressive targeting; these risks should be mitigated through privacy protection, fairness monitoring, human oversight, and deployment-specific auditing.

    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 high-risk generative model, scraped dataset, or model with substantial dual-use misuse risk. The industrial data are not publicly released, and the paper reports only aggregated experimental results.

    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: The paper cites the prior methods and algorithmic baselines used for comparison, including Causal Forest, T-Learner, DESCN, DRCFR, Deep Sets, and double/debiased machine learning. No third-party dataset or pretrained model is redistributed with the paper.

    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 introduce or release a new public dataset, benchmark asset, pretrained model, or software package in the current submission. The industrial dataset is proprietary and is used only for internal experimental evaluation.

Guidelines:

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

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

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

Answer: [N/A]

Justification: The paper does not involve crowdsourcing experiments, newly recruited participants, or human-subject studies with participant instructions or compensation. The empirical analysis uses existing randomized platform logs and aggregated offline evaluation.

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: [Yes]

Justification: The randomized platform experiments were conducted under the platform's standard experimental governance and data-protection procedures. The paper reports only anonymized and aggregated results, without disclosing personal information or identifiable user records.

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: The core methodology, theoretical analysis, experiments, and scientific claims do not use LLMs as an important, original, or non-standard component. Any use of language tools, if any, was limited to writing, editing, or formatting assistance and did not affect the research methodology or results.

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.