Multi-channel Uplift Policy Learning

Changjian Liu*
Peking University
Beijing, China
cjliu25@stu.pku.edu.cn

Tianyu Wang
Alibaba Group
Beijing, China
yves.wty@@alibaba-inc.com

Xiaoxuan Deng
Alibaba Group
Beijing, China
dengxiaoxuan.dxx@alibaba-inc.com

Wentao Zhu*
Beihang University
Beijing, China
zy2406229@buaa.edu.cn

Yuwei Xu*
CUHK-shenzhen
Shenzhen, China
yuweixu@link.cuhk.edu.cn

Junqi Jin✉
Alibaba Group
Beijing, China
junqi.jjq@alibaba-inc.com

Yong Gao✉
Peking University
Beijing, China
gaoyong@pku.edu.cn

Chuan Yu
Alibaba Group
Beijing, China
yuchuan.yc@alibaba-inc.com

Jian Xu
Alibaba Group
Beijing, China
xiyu.xj@taobao.com

Bo Zheng
Alibaba Group
Beijing, China
bozheng@alibaba-inc.com

Abstract

E-commerce platforms must allocate fixed marketing budgets across multiple channels to maximize business utility. However, standard predict-then-optimize (PTO) paradigms fail in this compositional space due to observational confounding and severe extrapolation. We formulate this challenge as a simplex-constrained uplift decision problem and propose ReAlloc, a fast-slow causal framework. Specifically, an agile Orthogonal Teacher extracts unbiased local gradients from short-term logs, while an Explanation-Guided Student distills them into a structured marginal field over long-term horizons. This design enables support-aware, conservative decisions that capture cross-channel substitutions. Extensive simulations and large-scale online A/B tests on Taobao platform demonstrate that ReAlloc achieves simultaneous lifts in both pay order and income.

CCS Concepts

• **Applied computing** → **Electronic commerce**; • **Information systems** → **Decision support systems**; *Computational advertising*; • **Computing methodologies** → **Causal reasoning and diagnostics**.

Keywords

uplift, resource allocation, decision making, e-commerce marketing

1 Introduction

In modern e-commerce systems, platforms increasingly act as centralized decision makers. Specifically, they must allocate limited marketing resources across heterogeneous items and diverse intervention channels [2, 5]. Moving beyond mere outcome prediction, the core objective is to learn intervention policies that causally drive business utility, such as sales, income, and Gross Merchandise Volume (GMV). This naturally frames an *uplift policy learning* problem: given a pre-decision state, the platform must allocate resources to maximize the incremental outcome [3, 11, 18].

Most existing uplift and contextual decision methods focus on binary, discrete, or single-channel treatments [29, 30]. However, real-world systems are inherently multi-channel and resource constrained. This transforms the intervention from an independent scalar action into a compositional allocation vector constrained on a budget simplex. Such a constraint fundamentally alters the decision landscape from the absolute individual treatment effects (ITE) to the *relative marginal value* of reallocating resources across channels. A concrete example arises in marketing hosting systems [26], where a fixed total budget is split between traffic-driving channels (e.g., advertising) and conversion enhancing benefits (e.g., coupons and rebates). Although potentially complementary, these channels frequently substitute or cannibalize each other. For instance, shifting funds to advertising increases exposure but dilutes conversion incentives. Consequently, the final outcome is governed by a complex joint response surface.

A common industrial paradigm is predict-then-optimize (PTO): fit a response model, then search to maximize the predicted outcome [9, 27]. However, PTO is fundamentally insufficient for multi-channel uplift learning for three reasons. **First**, channel-wise models ignore cross-channel interactions; their independent curves cannot recover the optimum of the joint response surface. **Second**, even a joint black-box model is unsafe under global optimization; the optimizer may exploit extrapolation errors in unsupported regions, yielding aggressive policies. **Third**, and most critically, prediction accuracy does not imply decision quality. Standard losses (e.g., MSE) optimize average prediction, whereas deployment relies on counterfactual ranking [6]. Minor prediction errors in high-uplift regions are severely amplified by the optimizer, ultimately degrading the result. This optimization vulnerability is further exacerbated by observational confounding. Historical allocations are driven by legacy business policies, systematically entangling treatment assignments with item states. Consequently, models achieve strong factual prediction but yield biased causal gradients, rendering downstream optimization ineffective.

*Work done during an internship at Taobao, Alibaba Group.

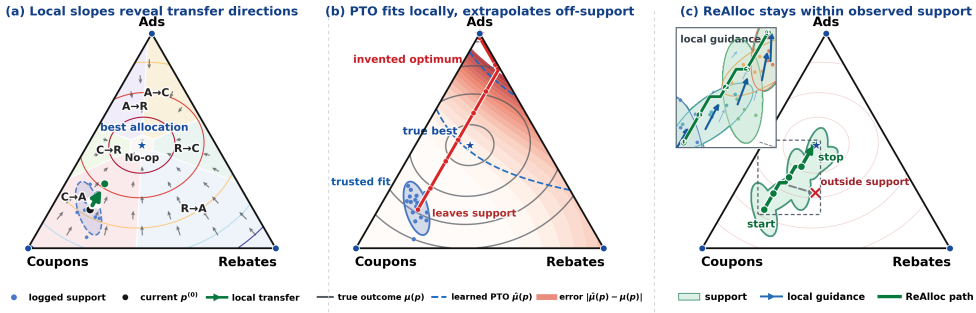

Figure 1: Conceptual illustration of ReAlloc. (a) Local response slopes identify promising reallocation directions. (b) Global PTO extrapolation produces unreliable decisions. (c) ReAlloc composes local slopes within support, leading to better decisions.

To address these challenges, we propose ReAlloc, a framework targeting local reallocation for multi-channel uplift decision problem. It learns the causal marginal gradient of the budget simplex. First, we use an orthogonalized teacher to remove confounding effects and recover unbiased local gradients. These gradients are then distilled into a student marginal field that guides conservative decisions within the observed support. ReAlloc turns unsafe PTO into support-aware causal reallocation, enabling precise and stable optimization under multi-channel budget constraints. Extensive offline experiments and large-scale online A/B tests on the Taobao platform demonstrate its superiority. It significantly outperforms uplift and PTO baselines in counterfactual ranking and decision quality. More importantly, online results reveals substantial improvements in both pay order 3.53 % and income 3.26 pt. Our contributions are:

(1) **Compositional Uplift.** We formulate multi-channel budget allocation as compositional uplift on the simplex, targeting relative marginal effects under zero-sum constraints.

(2) **Support-Aware Causal Reallocation.** We propose ReAlloc, which decouples causal response learning from support-aware decisions through a distilled conservative marginal field.

(3) **Industrial-Scale Deployment.** Deployed in Taobao's system, ReAlloc simultaneously improves pay order and income.

2 Related Work

2.1 Uplift Modeling

Uplift modeling and heterogeneous treatment effect (HTE) estimation aim to quantify the incremental impact of interventions [11, 16, 19, 25]. While classical and modern neural estimators have achieved significant success, they primarily focus on binary or discrete treatments, targeting absolute individual treatment effects (ITE) [17, 21, 23, 31]. Recent extensions consider multiple treatments [18, 30] or continuous doses [12, 13, 28]. However, these formulations typically model treatments as separate arms or a one-dimensional dose, rather than as a coupled composition. At the intersection of causal estimation and resource allocation, existing methods fall short of our setting. Budget-constrained uplift methods [1, 2, 24] allocate scarce resources across users via knapsack optimization, but restrict each user to a scalar or discrete treatment. Multi-channel advertising systems [5, 22] coordinate campaign-level budgets, treating channel responses as aggregate primitives. Multi-touch attribution [10, 15] assigns conversion credit across sequences but lacks explicit simplex-constrained optimization. In contrast, our setting requires learning *item-level* causal substitution under a fixed budget, where increasing one channel's allocation necessarily decreases another's, demanding a fundamentally different compositional formulation.

2.2 Decision-Focused Optimization

Offline policy learning evaluates policies from logged data using inverse-propensity or doubly robust estimators [3, 7, 8]. In industrial pipelines, the dominant paradigm is PTO, which decouples reward prediction from downstream optimization. To bridge the prediction-decision gap, Decision-Focused Learning (DFL) and SPO-style losses directly train predictive models against downstream decision metrics [9, 20, 27]. While PTO and DFL successfully align prediction with decision objectives in deterministic settings, they are fundamentally ill-equipped for causal compositional allocation. First, Factual Objective vs. Counterfactual Gradients. DFL optimizes factual reconstruction end-to-end. However, deployment strictly requires correct counterfactual marginal ranking. Differentiating through the optimizer with factual losses does not guarantee accurate causal gradients, especially in high-uplift regions. Second, Lack of Extrapolation Conservatism. Global optimizers over learned black-box surfaces are notoriously aggressive. Unlike Conservative Q-Learning in offline RL [14], standard DFL lacks mechanisms to penalize out-of-support predictions. On the constrained simplex, DFL actively exploits extrapolation errors, yielding disastrous out-of-distribution policies. Third, Confounding Amplification. Historical allocations are heavily entangled with item states via legacy policies. Because DFL aggressively optimizes these biased factual predictions end-to-end without explicit causal orthogonalization, it amplifies unreliable causal gradients [4, 17], leading to severe policy degradation.

3 Problem Formulation

3.1 Setup

Each decision has a fixed budget across K channels. Let $X = (H, B)$ collect the pre-decision state $H \in \mathcal{H}$ and the available budget $B > 0$.

We observe historical decisions:

$$\mathcal{D} = \{(X_i, P_i, Y_i)\}_{i=1}^{n}, \qquad P_i \in \Delta^{K-1} := \left\{p \in \mathbb{R}_+^K : \mathbf{1}^\top p = 1\right\}, \quad (1)$$

where P_i is the allocation proportion and Y_i is the realized outcome. Uppercase P_i denotes a historical action, whereas lowercase p denotes a candidate decision. Writing $Y(p) \equiv Y(Bp)$ for the potential outcome under allocation p, define $\mu(X, p) := \mathbb{E}[Y(p) \mid X]$ and $V(\pi) := \mathbb{E}_X[\mu(X, \pi(X))]$. Given a policy class Π, our objective is

$$\pi^\star \in \arg\max_{\pi \in \Pi} V(\pi), \qquad \pi : X \mapsto \Delta^{K-1}. \quad (2)$$

3.2 Local Reallocation Is the Decision Primitive

Fixed budget makes every feasible first-order perturbation zero-sum. The tangent space of the simplex and its orthogonal projector are:

$$\mathcal{T} := \{v \in \mathbb{R}^K : \mathbf{1}^\top v = 0\}, \qquad \Pi_{\mathcal{T}} := I_K - \frac{1}{K}\mathbf{1}\mathbf{1}^\top. \quad (3)$$

For channels $k \neq \ell$, an infinitesimal transfer from k to ℓ has local effect

$$D_{k\to\ell}\mu(X, p) = \frac{\mathrm{d}}{\mathrm{d}\delta}\mu(X, p + \delta(e_\ell - e_k))\Big|_{\delta=0^+} = (e_\ell - e_k)^\top \nabla_p \mu(X, p) = (e_\ell - e_k)^\top g^\star(X, p), \quad (4)$$

where $g^\star(X, p) = \Pi_{\mathcal{T}} \nabla_p \mu(X, p)$ is the causal reallocation field. The policy signal is therefore not an absolute channel effect, but the relative marginal return from moving budget between channels.

3.3 Production Failure Modes of PTO

A standard industrial solution to (2) is PTO:

$$\widehat{\theta} \in \arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, \widehat{\mu}_\theta(X_i, P_i)), \quad (5)$$

$$\widehat{\pi}_{\mathrm{PTO}}(X) \in \arg\max_{p \in \Delta^{K-1}} \widehat{\mu}_{\widehat{\theta}}(X, p). \quad (6)$$

Despite strong factual accuracy, this pipeline faces three failures. **Observational assignment:** historical allocations are selected by legacy policies, so causal interpretation requires adjustment. **Objective mismatch:** factual prediction loss does not control the marginal slopes or candidate rankings consumed by the optimizer. **Support mismatch:** global optimization may exploit response estimates outside the logged action region; These failures motivate the three stages of ReAlloc: orthogonal local estimation, marginal gradient distillation, and support-aware decision.

4 Method

We propose ReAlloc, a casual teacher-student framework for fixed-budget multi-channel uplift learning; see Fig. 2. To handle the inherent non-stationarity of marketing environments and the vulnerabilities of standard PTO, ReAlloc operates as a **dual system**. At each update round, a **fast teacher** is trained on recent logs to extract unbiased local causal geometry, while a **slow student** accumulates this through a replay buffer to produce stable decisions. As detailed below, the three stages of ReAlloc systematically address the challenges of observational confounding, prediction-decision mismatch, and extrapolation vulnerability; see Algorithm 1.

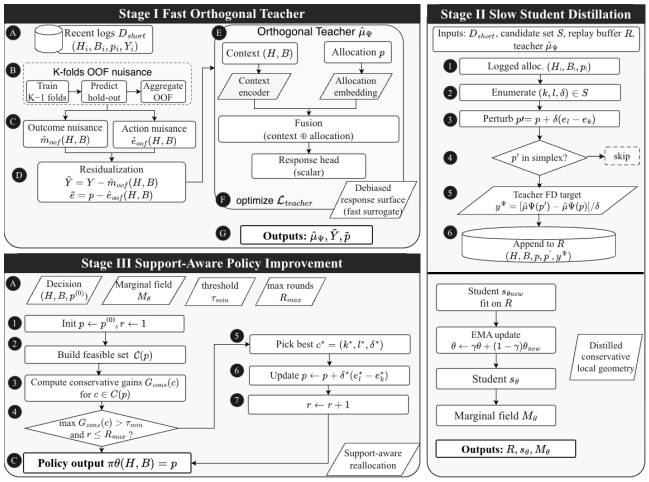

Figure 2: Overview of ReAlloc

4.1 Stage I: Orthogonal Response Teacher

To address *observational confounding* (Challenge 1), we use orthogonalization, akin to Robinson's transformation in double machine learning (DML). For each recent training window $\mathcal{D}_{\mathrm{short}}$, we first estimate two nuisance functions via cross-fitting:

$$\hat{m}(H, B) \approx \mathbb{E}[Y \mid H, B], \qquad \hat{e}(H, B) \approx \mathbb{E}[p \mid H, B], \quad (7)$$

where $\hat{m}$ captures the baseline demand and $\hat{e} \in \Delta^{K-1}$ captures the conditional mean action, i.e., the legacy policy tendency. We then construct the residualized outcome and treatment:

$$\tilde{Y}_i = Y_i - \hat{m}(H_i, B_i), \qquad \tilde{p}_i = p_i - \hat{e}(H_i, B_i). \quad (8)$$

Constrained by simplex the simplex, $\tilde{p}_i$ is a valid tangent-space deviation from the historical allocation, with the predictable component of treatment assignment removed. The teacher's response surface is parameterized as:

$$\hat{\mu}_\psi(H, B, p) = \hat{m}(H, B) + r_\psi\big(H, B, p - \hat{e}(H, B)\big), \quad (9)$$

where r_ψ is a residual response branch anchored at $r_\psi(H, B, 0) = 0$, ensuring that it models only the outcome variation driven by allocation deviations. Let $\mathcal{P}_0 = I - \frac{1}{K}\mathbf{1}\mathbf{1}^\top$ be the projection onto the tangent space. The teacher's local marginal field is

$$g_\psi(H, B) = \mathcal{P}_0 \nabla_z r_\psi(H, B, z)\big|_{z=0}. \quad (10)$$

We train the teacher by minimizing the composite objective

$$\mathcal{L}_{\mathrm{teacher}} = \frac{1}{|\mathcal{D}_{\mathrm{short}}|}\sum_{i \in \mathcal{D}_{\mathrm{short}}} \Big[\ell\big(Y_i, \hat{\mu}_\psi(H_i, B_i, p_i)\big) + \lambda_{\mathrm{res}}\ell\big(\tilde{Y}_i, r_\psi(H_i, B_i, \tilde{p}_i)\big) + \lambda_{\mathrm{grad}}\ell\big(\tilde{Y}_i, \langle g_\psi(H_i, B_i), \tilde{p}_i\rangle\big)\Big]. \quad (11)$$

The first term maintains factual predictive accuracy, the second forces the residual branch to explain residual outcome variation, and the third explicitly regularizes the local directional derivative, yielding reliable causal gradients in supported neighborhoods.

4.2 Stage II: Student Marginal Distillation

While the fast teacher adapts to recent dynamics, directly optimizing its response surface remains prone to the PTO failure mode (Challenge 2). We therefore train a slow student to distill the teacher's local geometry into a globally consistent marginal field. To ensure that the learned marginal field is integrable, i.e., path-consistent, we define the student through a scalar potential function $s_\theta(H, B, p) \in \mathbb{R}$. Its projected marginal utility is

$$u_\theta(H, B, p) = \mathcal{P}_0 \nabla_p s_\theta(H, B, p). \tag{12}$$

The predicted local marginal gain for reallocating budget from channel k to channel l is

$$M_{\theta,k\to l}(H, B, p) = u_{\theta,l}(H, B, p) - u_{\theta,k}(H, B, p). \tag{13}$$

This parameterization guarantees that finite allocation changes can be scored consistently through potential differences, avoiding cyclic or contradictory gradients.

At each round, the teacher generates supervision targets. For a feasible perturbation $p' = p + \delta(e_l - e_k)$, the finite-difference target is

$$y^{\psi}_{k\to l,\delta} = \frac{\hat{\mu}_\psi(H, B, p') - \hat{\mu}_\psi(H, B, p)}{\delta}. \tag{14}$$

We also extract the teacher's projected Jacobian $g^\psi = \mathcal{P}_0 \nabla_p \hat{\mu}_\psi(H, B, p)$. The student is optimized over a replay buffer $\mathcal{R}$ of these targets:

$$\mathcal{L}_{\text{student}} = \mathbb{E}_{\mathcal{R}}\Big[\lambda_{\text{pair}}\big(M_{\theta,k\to l} - y^{\psi}_{k\to l,\delta}\big)^2 + \lambda_{\text{jac}}\left\|u_\theta - g^\psi\right\|_2^2\Big]. \tag{15}$$

By accumulating targets in $\mathcal{R}$ and updating the deployed student via exponential moving average, $\theta^{(t)} = \gamma\theta^{(t-1)} + (1-\gamma)\theta_{\text{new}}$, the student learns a stable, decision-focused marginal geometry.

4.3 Stage III: Support-Aware Decision

To prevent *extrapolation vulnerability* (Challenge 3), ReAlloc abandons unconstrained global maximization and instead performs conservative local reallocations. Given the current allocation p, we enumerate feasible local steps

$$C(p) = \left\{(k, l, \delta) : p + \delta(e_l - e_k) \in \Delta^{K-1}\right\}. \tag{16}$$

For a candidate $c = (k, l, \delta)$, let $p_c = p + \delta(e_l - e_k)$. The expected gain is computed from the student's potential difference:

$$\widehat{G}_c = s_\theta(H, B, p_c) - s_\theta(H, B, p). \tag{17}$$

To penalize unreliable updates, we define the conservative gain

$$\widehat{G}^{\text{cons}}_c = \widehat{G}_c - \beta\widehat{\sigma}_c - \lambda_s\varphi(\omega_c), \qquad \varphi(\omega_c) = -\log(\omega_c + \epsilon). \tag{18}$$

Here, $\widehat{\sigma}_c$ is the predictive uncertainty, estimated using ensemble variance or MC Dropout, and $\omega_c \in [0, 1]$ is a directional support score, computed using kernel density or inverse KNN distance to the replay buffer $\mathcal{R}$. The latter measures whether similar reallocations are empirically supported by historical data. The policy selects the optimal conservative candidate:

$$c^\star = \arg\max_{c\in C(p)} \widehat{G}^{\text{cons}}_c. \tag{19}$$

The allocation is updated only if the conservative gain exceeds a safety threshold $\tau_{\min}$; otherwise, a no-op is executed. This localized, support-regularized search avoids unsupported, high-risk regions of the simplex while enabling conservative policy improvement.

Algorithm 1: Implementation of ReAlloc

Input: Recent logs $\mathcal{D}_{\text{short}}$, replay buffer $\mathcal{R}$, candidate set $\mathcal{S}$, threshold $\tau_{\min}$, EMA rate γ
Output: Potential student s_θ, marginal field M_θ, policy π_θ

Stage I: Fast Orthogonal Teacher;
Estimate nuisances $\hat{m}(H, B)$ and $\hat{e}(H, B)$ on $\mathcal{D}_{\text{short}}$;
Compute residuals $\tilde{Y}_i$ and $\tilde{p}_i$;
Train teacher $\hat{\mu}_\psi$ by minimizing $\mathcal{L}_{\text{teacher}}$;
Stage II: Slow Student Distillation;
foreach $(H_i, B_i, p_i) \in \mathcal{D}_{\text{short}}$ **do**
  foreach $(k, l, \delta) \in \mathcal{S}$ **do**
    $p_i' \leftarrow p_i + \delta(e_l - e_k)$;
    if $p_i' \in \Delta^{K-1}$ **then**
      $y^{\psi}_{i,k\to l,\delta} \leftarrow \dfrac{\hat{\mu}_\psi(H_i, B_i, p_i') - \hat{\mu}_\psi(H_i, B_i, p_i)}{\delta}$;
      Add $(H_i, B_i, p_i, k, l, \delta, y^{\psi}_{i,k\to l,\delta})$ to $\mathcal{R}$;
    end
  end
end
Update potential student $s_{\theta_{\text{new}}}$ on $\mathcal{R}$ by minimizing $\mathcal{L}_{\text{student}}$;
$\theta \leftarrow \gamma\theta + (1-\gamma)\theta_{\text{new}}$;
Stage III: Support-Aware Policy Improvement;
foreach *decision instance* $(H, B, p^{(0)})$ **do**
  $p \leftarrow p^{(0)}$; $r \leftarrow 1$;
  while $r \le R_{\max}$ *and* $\max_c \widehat{G}^{\text{cons}}_c(H, B, p) > \tau_{\min}$ **do**
    $c^\star \leftarrow \arg\max_{c\in C(p)} \widehat{G}^{\text{cons}}_c(H, B, p)$; `// Best conservative candidate`
    $p \leftarrow p + \delta^\star(e_{l^\star} - e_{k^\star})$; $r \leftarrow r + 1$;
  end
  $\pi_\theta(H, B) \leftarrow p$;
end
return s_θ, M_θ, π_θ;

5 Theory

Let $p_0(X)$ denote the incumbent allocation and let $C(X) \subseteq \Delta^{K-1}$ denote the supported region. For an absolutely continuous path $\gamma : [0, 1] \to \Delta^{K-1}$, define the path length:

$$L(\gamma) := \int_0^1 \|\dot{\gamma}(t)\|_2 \, \mathrm{d}t. \tag{20}$$

The path is support-admissible if $\gamma(t) \in C(X)$ for all $t \in [0, 1]$. Define the **Pointwise Uplift** Δ and **Uplift** Γ:

$$\Delta(X, p) := \mu(X, p) - \mu(X, p_0(X)), \tag{21}$$

$$\Gamma(\pi) := \mathbb{E}_X[\Delta(X, \pi(X))], \qquad \mathcal{R} := \{(X, p) : p \in C(X)\}. \tag{22}$$

Formal assumptions and proofs are deferred to Appendix A.

5.1 Global Uplift Accumulates Local Return

Theorem 5.1 (Simplex uplift along a feasible path). *Let γ be an absolutely continuous path with $\gamma(0) = p_0(X)$ and $\gamma(1) = p$. If*

$\mu(X,\cdot)$ is continuously differentiable along γ, then

$$\Delta(X,p) = \int_0^1 g^\star(X,\gamma(t))^\top \dot\gamma(t)\,\mathrm{d}t. \tag{23}$$

Finite business value is the accumulated marginal return along the allocation moves that the system actually executes. ReAlloc therefore concentrates estimation capacity on supported local traces rather than reconstructing a globally response surface. Crucially, this integral equivalence motivates us to parameterize the student as a scalar potential function, ensuring that its utility difference between allocations exactly equals the path integral of local returns, thus serving as a direct and reliable proxy for the true uplift. The proof is given in Appendix A.2.

5.2 Factual Fit Does Not Certify a Policy

Proposition 1 (Factual risk does not control decision quality). *There exist smooth response surfaces and PTO predictors with zero factual risk under the logging distribution but constant decision regret. Moreover, there exists a sequence $\widehat{\mu}_n$ such that $\|\widehat{\mu}_n - \mu\|_{L_2} \to 0$ while $\left\|\Pi_{\mathcal{T}}\nabla_p\widehat{\mu}_n - \Pi_{\mathcal{T}}\nabla_p\mu\right\|_\infty \to \infty$.*

The constructions are given in Appendix A.3. Let $m(X,p) := \mathbb{E}[Y \mid X, P=p]$. A factual response model targets the observational field

$$g_{\mathrm{obs}}(X,p) := \Pi_{\mathcal{T}}\nabla_p m(X,p) = g^\star(X,p) + b_{\mathrm{sc}}(X,p), \tag{24}$$

$$b_{\mathrm{sc}}(X,p) := \Pi_{\mathcal{T}}\nabla_p\{m(X,p) - \mu(X,p)\}. \tag{25}$$

The exact shortcut bias decomposition is given in Appendix A.4. It vanishes under conditional ignorability; orthogonalization controls nuisance-estimation sensitivity after identification, but does not recover omitted confounders. Offline RMSE and calibration are not launch certificates for an allocation policy. An optimizer can amplify small slope errors, and observational selection can make a historically favored channel appear incrementally valuable. Predictive fit, local directional accuracy, support coverage, and policy value must therefore be validated separately.

5.3 Orthogonal Estimation of Supported Local Effects

Let $d = K-1$ and let $Q \in \mathbb{R}^{K\times d}$ be an orthonormal basis of $\mathcal{T}$. With $z = Q^\top p$, $Z = Q^\top P$, and $\widetilde{Z} = Z - z$, let $\mathbb{E}_{h,z}[\cdot \mid X]$ denote kernel localization around z. Define

$$e_{h,z}(X) := \mathbb{E}_{h,z}[\widetilde{Z} \mid X], \qquad m_{h,z}(X) := \mathbb{E}_{h,z}[Y \mid X]. \tag{26}$$

The orthogonal component of the teacher is characterized by

$$\Psi_{h,z}(\beta,m,e;X) := \mathbb{E}_{h,z}\Big[\{\widetilde{Z} - e(X)\}\{Y - m(X) - \beta(X,z)^\top(\widetilde{Z} - e(X))\}\,\Big|\,X\Big]. \tag{27}$$

Theorem 5.2 (Orthogonal local-field estimation). *Under Assumptions 1–4, the population root $\beta_h^\dagger$ of (27) satisfies*

$$\sup_{(X,p)\in\mathcal{R}} \left\|Q\beta_h^\dagger(X,Q^\top p) - g^\star(X,p)\right\|_2 \le \epsilon_{\mathrm{loc}}(h), \tag{28}$$

and the score is Neyman orthogonal with respect to (m,e). Let $\widehat{g}_T^{\mathrm{orth}} := Q\widehat{\beta}$ be the cross-fitted empirical root and let $\widehat{g}_T$ be the field returned by the implemented composite teacher. If

$$\|\widehat{g}_T - \widehat{g}_T^{\mathrm{orth}}\|_{\infty,\mathcal{R}} \le \epsilon_{\mathrm{comp}}, \tag{29}$$

then, with probability at least $1-\delta$,

$$\|\widehat{g}_T - g^\star\|_{\infty,\mathcal{R}} \le \epsilon_T(n,h,\delta) + \epsilon_{\mathrm{comp}}, \tag{30}$$

$$\epsilon_T(n,h,\delta) := \epsilon_{\mathrm{loc}}(h) + C\sqrt{\frac{\mathfrak{C}_{\mathcal{G}} + \log(1/\delta)}{nh^{K+1}}} + Cr_{\mathrm{orth}}. \tag{31}$$

The normalized second-order remainder r_{orth} and the proof are given in Appendix A.5. The rate exposes the production trade-off. A smaller neighborhood reduces local approximation bias but also reduces effective sample size; the h^{K+1} term makes channel dimensionality an explicit data requirement. Orthogonality allows the outcome and logging models to be improved modularly, while ϵ_{comp} records approximation and optimization error from the full neural objective.

5.4 Potential Accuracy Controls Regret

The deployed student is parameterized by a scalar potential: $\widehat{g}(X,p) := \Pi_{\mathcal{T}}\nabla_p s_\theta(X,p)$, $\widehat{\Delta}_\theta(X,p) := s_\theta(X,p) - s_\theta(X,p_0(X))$. For $p \in C(X)$, let $\mathfrak{P}_X(p)$ denote the set of support-admissible paths from $p_0(X)$ to p, and define the reachable action set:

$$\mathcal{A}_{L_{\max}}(X) := \{p : \exists\gamma \in \mathfrak{P}_X(p),\ L(\gamma) \le L_{\max}\}. \tag{32}$$

Let $\pi^\star_{L_{\max}}(X) \in \arg\max_{p\in\mathcal{A}_{L_{\max}}(X)} \mu(X,p)$ and define

$$\mathrm{Reg}_{L_{\max}}(\widehat{\pi}) := \mathbb{E}_X\left[\mu(X,\pi^\star_{L_{\max}}(X)) - \mu(X,\widehat{\pi}(X))\right]. \tag{33}$$

Assume the local search returns $\widehat{\pi}(X) \in \mathcal{A}_{L_{\max}}(X)$ and satisfies

$$\mathbb{E}_X\left[\sup_{p\in\mathcal{A}_{L_{\max}}(X)} \widehat{\Delta}_\theta(X,p) - \widehat{\Delta}_\theta(X,\widehat{\pi}(X))\right] \le \epsilon_{\mathrm{search}}. \tag{34}$$

Theorem 5.3 (Trace-wise field error controls value and regret). *Suppose $\|\widehat{g} - g^\star\|_{\infty,\mathcal{R}} \le \epsilon_g$. For any $p \in \mathcal{A}_{L_{\max}}(X)$ and any $\gamma \in \mathfrak{P}_X(p)$,*

$$\widehat{\Delta}_\theta(X,p) = \int_0^1 \widehat{g}(X,\gamma(t))^\top\dot\gamma(t)\,\mathrm{d}t, \tag{35}$$

$$\left|\widehat{\Delta}_\theta(X,p) - \Delta(X,p)\right| \le L(\gamma)\epsilon_g. \tag{36}$$

Consequently,

$$\mathrm{Reg}_{L_{\max}}(\widehat{\pi}) \le 2L_{\max}\epsilon_g + \epsilon_{\mathrm{search}}. \tag{37}$$

Corollary 5.4 (End-to-end teacher–student regret). *Under Assumption 5 and the event in Theorem 5.2,*

$$\mathrm{Reg}_{L_{\max}}(\widehat{\pi}) \le 2L_{\max}\{\epsilon_T(n,h,\delta) + \epsilon_{\mathrm{comp}} + \epsilon_S\} + \epsilon_{\mathrm{search}}. \tag{38}$$

The proofs are given in Appendix A.6. $L_{\max}$ is a deployment blast-radius control: larger cumulative budget movement creates more upside but also amplifies field error. The term $\epsilon_{\mathrm{search}}$ captures the value lost to finite candidate sets, greedy search, and latency constraints. Because the student is a scalar potential, gains telescope across the actual accepted greedy trace and remain path-consistent.

6 Synthetic Experiment

6.1 Data Generation Process (DGP)

We construct a synthetic environment with known counterfactual outcomes to evaluate *ReAlloc*. The DGP is designed to preserve three core challenges: state-dependent assignment, locally identifiable but globally under-supported actions, and complex channel interactions. Each observation is an item-period tuple $(H_{it}, B_{it}, P_{it}, Y_{it})$, where the allocation $P_{it} \in \Delta^{K-1}$ lies on a simplex with $K = 3$ channels.

State-dependent logging. We decompose the state as $H = (C, E, S)$, where C acts as the primary confounder, E modifies the response, and S captures channel preference. To handle compositional allocations P, we map the simplex to a vector space using the isometric log-ratio (ILR) transform. The logging policy is specifically designed to induce three critical properties: (1) **State-dependent confounding**, controlled by a parameter α_{cf}, which couples the primary confounder C with the allocation; (2) **Boundary skewness**, controlled by α_{bd}, which pushes the propensity mean toward simplex boundaries to mimic extreme budget skews; and (3) **Strict overlap**, achieved via a logistic-normal mixture noise that balances local exploitation (low variance) with sparse global exploration. Crucially, since all confounders are fully observed in H, the DGP strictly satisfies conditional ignorability. *(See Appendix B.)*

Joint response surface. Outcomes follow $Y = \mu_\star(H, B, P) + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma_Y^2)$. The causal response is anchored at a state-independent baseline allocation $c(H, B)$. Let z denote the allocation deviation from the baseline in the ILR space. The response surface:

$$\begin{aligned}\mu_\star(H, B, P) &= m_0(H, B) \\ &+ \rho(H, B)\Big[\omega_{\text{loc}} g_{\text{loc}}(H, z) + \omega_{\text{int}} g_{\text{int}}(H, z) + \omega_{\text{far}} g_{\text{far}}(H, z)\Big].\end{aligned} \tag{39}$$

This formulation explicitly disentangles the ROI curve into three regimes: (1) The **local term** $g_{\text{loc}}(H, z)$ captures linear marginal returns, providing an identifiable first-order reallocation signal. (2) The **interaction term** g_{int} models cross-channel substitution and complementarity. (3) The **far-field term** g_{far} is flat near the anchor but introduces non-linear saturation and cannibalization for out-of-support (OOS) allocations. The component scales (ω) are calibrated to ensure local gradients remain learnable from logs, while distant curvature is weakly identified, rigorously testing the model's extrapolation capability.

Temporal variations. For the temporal experiment, we keep $\mu_\star$ and all response parameters fixed, and apply a smooth window-specific bias to the logging logits. This rotates the observed action support across channels without introducing sequential treatment effects. Hence, we evaluate the retention of local geometric knowledge under changing support, rather than mechanism drift.

6.2 Setup

We design synthetic experiments to address three questions: **RQ1 (Safe Policy Improvement):** Under increasingly severe confounding and low overlap, can ReAlloc achieve higher deployable uplift than baselines? **RQ2 (Mechanism of Improvement):** Which components drive the performance gains? **RQ3 (Fast-Slow Temporal Memory):** Can the slow student effectively accumulate and retain local geometric knowledge from successive teachers?

Baselines. We compare ReAlloc against a comprehensive suite of reference, industry-style, and causal baselines: (1) *Logging* (maintains historical allocations) and *Uniform* (equal budget split) serve as reference policies. (2) *Additive ROI* models channel responses independently, reflecting the decoupled paradigm in industry. (3) *Joint S-Learner PTO* fits a global response surface using a standard S-learner and applies global or local optimization. (4) *R-Learner Local* orthogonally residualizes outcomes and treatments to estimate heterogeneous effects, selecting the optimal locally.

Evaluation Metrics. Standard offline policy evaluation often ignores the risk of OOD recommendations. We introduce a fallback mechanism: if a policy $\hat{\pi}$ recommends an allocation $\hat{p}_i$ that falls outside the historically supported region $\mathcal{S}_i$, the system safely defaults to the factual logging allocation P_i. Let $\Delta_i(\hat{\pi}) = \mu_\star(H_i, B_i, \hat{p}_i) - \mu_\star(H_i, B_i, P_i)$ be the pointwise oracle uplift. We define the **Deployable Uplift** as our primary metric:

$$\text{Uplift}_{\text{dep}}(\hat{\pi}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{\hat{p}_i \in \mathcal{S}_i\}\Delta_i(\hat{\pi}), \tag{40}$$

To provide a comprehensive evaluation, we additionally report the OOS rate (the fraction of rejected recommendations), the Safe Local Recovery ratio (comparing U_{dep} against an oracle), and geometric ranking metrics (EdgeNDCG, TopEdgeAcc, TopEdgeRegret, PairwiseCorr) that evaluate the local directional accuracy independent of policy visitation.

6.3 Results and Analysis

RQ1 result. Table 1 reports policy quality under decreasing logging overlap. Unconstrained PTO attains high raw uplift but converts little of it into deployable value, whereas support-constrained S-learners recover most of this loss, showing that unsafe global search is a major source of PTO failure. ReAlloc remains the strongest learned policy across all regimes, achieving deployable uplifts of .97/.79/.71/.57 and recovering .88/.86/.84/.83 of the local oracle opportunity; its margin over the strongest S-NN variant is .04/.07/.09/.06. Additive ROI and the orthogonal R-Learner remain feasible but obtain substantially lower value. Notably, the R-Learner slightly outperforms ReAlloc on hard-regime logged-anchor NDCG and regret, yet reaches only .21 deployable uplift versus .71, indicating that single-step edge quality alone does not determine effective multi-step reallocation.

Figure 3 complements the aggregate results. For a fixed context and oracle response surface, increasing assignment severity concentrates the conditional logged actions, isolating support deterioration from changes in the underlying outcome function. The best-edge map further shows that the preferred local transfer depends on the current simplex position. The trajectory comparison illustrates how global PTO moves toward a distant unsupported action, whereas the shared local-support policies remain feasible and follow distinct reallocation paths.

RQ2 result. Table 2 shows that ReAlloc requires both accurate local geometry and support-aware deployment. The results separate geometry errors from safety errors. Teacher-only local greedy remains support-safe, but it achieves weak edge correlation and recovers only a small fraction of the local oracle gain, showing that directly deploying noisy teacher targets is insufficient for stable

Table 1: Static policy quality under decreasing overlap. Cells show Deployable uplift / Safe Recovery (D/SR). For unconstrained PTO ($_G$), Raw uplift (R) is shown as superscript.

	Uplift (D/SR)				Hard Regime Diagnostics		
Method	Benign	Medium	Hard	Extreme	Bias	NDCG	Regret
References							
Logging	.00/–	.00/–	.00/–	.00/–	–	–	–
Uniform	−.01/–	−.03/–	−.03/–	−.01/–	–	–	–
Add. ROI	.22/.20	.21/.22	.18/.22	.16/.23	−.02	.89	.78
Unconstrained global PTO							
S-GBDT$_G$	**.08**$^{1.0}$/.07	**.03**$^{.8}$/.03	**.01**$^{.6}$/.01	**.02**$^{.7}$/.02	+.12	.70	2.62
S-NN$_G$	**.02**$^{1.2}$/.02	**.01**$^{1.1}$/.01	**.01**$^{1.0}$/.01	**.00**$^{1.0}$/.00	−.15	.89	.70
Support-constrained PTO							
S-GBDT$_C$	.75/.68	.46/.50	.32/.39	.36/.52	+.03	.70	2.62
S-NN$_C$	.93/.84	.72/.78	.62/.74	.50/.73	−.12	.89	.70
Shared local-support search							
S-GBDT$_L$	.01/.01	.01/.01	.00/.00	.00/.01	+.00	.70	2.62
S-NN$_L$	.91/.82	.71/.76	.61/.73	.51/.74	−.11	.89	.70
R-Learner$_L$	.24/.21	.22/.24	.21/.25	.18/.27	−.03	**.92**	**.49**
ReAlloc	**.97**/.88	**.79**/.86	**.71**/.84	**.57**/.83	+.05	.92	.50
Oracle Local	1.10/1.00	.92/1.00	.83/1.00	.68/1.00	–	–	–

Subscripts: G=global, C=constrained, L=local search.

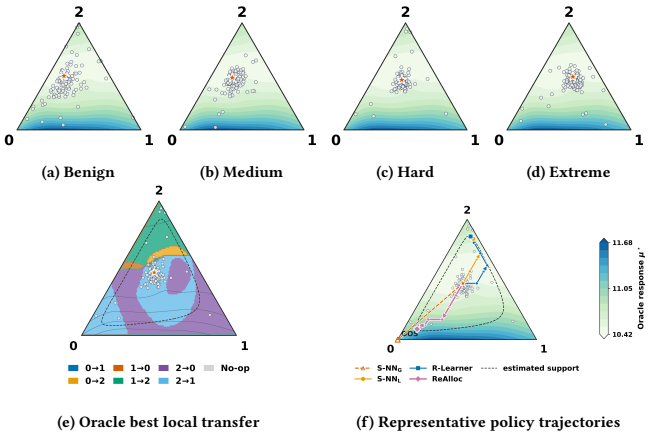

Figure 3: Orcle response surface and policy behavior.

Table 2: Mechanism ablation under hard interaction stress.

Variant	Local geometry				Decision / safety			
	Edge ↑	Top ↑	Corr ↑	Err ↓	Dep. ↑	OOS ↓	Gap ↓	Rec. ↑
Teacher-only	.816	.343	.429	1.173	.144	.000	.000	.129
w/o Orthogonalization	.790	.294	.407	.977	.412	.000	.000	.357
w/o Support	.784	.291	.407	.922	.001	.984	.589	.001
ReAlloc	**.828**	**.370**	**.492**	1.011	**.522**	**.000**	**.000**	**.455**

policy improvement. The most severe safety failure occurs when the support check is removed: despite using a learned local field, the policy frequently leaves the empirical support, leading to high OOS and a non-negligible raw-safe gap. Overall, ReAlloc achieves the best combination of edge-ranking accuracy, deployable uplift, and support safety among learned variants, recovering about 60% of the support-aware local oracle gain without accessing oracle counterfactuals.

Table 3: Final rolling-window summary.

Method	Curr. ↑	Past ↑	Union ↑	Memory ↓	Time ↓
Fresh Teacher	.144	.045	.051	0.50×	0.34×
Reset Student	.152	.047	.053	0.50×	0.43×
Warm-start Teacher	.146	.048	.054	0.50×	0.32×
Equal-memory Raw Replay	.208	.068	.077	1.00×	0.95×
Fast-Slow ReAlloc	.216	.071	.080	1.00×	1.00×
Pooled Teacher	**.224**	**.075**	**.084**	8.00×	2.71×

RQ3 result. We finally evaluate periodic support rotation under a fixed response surface. With N windows and n rows per window, a *Pooled Teacher* stores $\Theta(Nn)$ rows and incurs $\Theta(N^2n)$ cumulative fitting work by repeatedly retraining on the growing history. In contrast, Fast-Slow ReAlloc updates a current-window teacher alongside a fixed-memory student, requiring only $\Theta(n+B)$ storage and $\Theta(N(n+B))$ cumulative work, where B is the student's buffer size. In our configuration ($B \approx n$, $N = 16$), Fast-Slow uses only 12.5% of the pooled storage and 36.9% of its cumulative fitting time, while retaining 94.9% of its final union uplift (0.080 vs. 0.084). This bounded update cost becomes increasingly important when full-history retraining grows progressively more expensive.

7 Real-World Evaluation on Taobao

We evaluate ReAlloc on a 60-day Taobao production dataset comprising 500K items, each allocating a fixed budget across paid advertising and promotional benefits. Evaluations are conducted via matched prospective replay and online A/B testing.

7.1 Offline Evaluation

Offline evaluation combines two complementary data because routine production logs suffer from low overlap and unstable IPS estimates, as risk controls concentrate reallocations near incumbent allocations with small incremental effects. Thus, we use routine traffic solely for an out-of-time matched replay to evaluate directional ranking. For each moved event $e = (i, t)$, we identify no-move controls ($\Delta\mathbf{p} = 0$) from the same date and first-level category using caliper-constrained KNN matching on pre-decision covariates. An evaluation model m_{eval}, trained strictly on dates preceding the replay window, first removes predictable demand variation: $u_{it} = y_{it} - m_{\text{eval}}(\mathbf{x}_{it})$. The matched residual response is then defined as $R_e = u_e - \frac{1}{|C(e)|}\sum_{j\in C(e)} u_j$, where $C(e)$ denotes the matched no-move controls. For method m, let $G_m(e)$ denote its predicted gain for the realized reallocation and let $S_m(e) = \frac{G_m(e)}{\|\Delta\mathbf{p}_e\|_1+\epsilon}$ be the corresponding score. Matching removes predictable demand fluctuations and imbalance in observed covariates, but cannot eliminate unobserved confounding. We therefore use R_e only for evaluating directional ranking.

Causal policy value is evaluated on an exploration set containing 10% of the items. In this set, reallocation directions and magnitudes are randomized with controlled, estimable behavior propensities, which permits DR OPE. We additionally evaluate deployment support using historical reallocations. These evaluations provide three complementary views of a method: score validity, empirical action support, and policy value. We report ρ (Spearman correlation between G_m and R_e), within-item concordance (pairwise S_m–R_e

Table 4: Offline evaluation on routine production traffic and randomized exploration traffic.

	A. Routine traffic				B. Randomized exploration traffic			
Method	$\rho(G,R)\uparrow$	Within-item conc.↑	A–R gap↑	Supp. viol. @ upd.↓	DR lift↑	Action rate	Lift/act.↑	Agreement lift↑
ReAlloc	$.028_{[.021,.036]}$	.512	.023	**.000**@.31	$\mathbf{.025}_{[.012,.043]}$	.33	**.069**	$.098_{[.067,.153]}$
ReAlloc w/o Supp.	$.028_{[.021,.036]}$	.512	.023	.678@.32	$-.021_{[-.028,-.012]}$	.96	−.021	$-.040_{[-.061,-.022]}$
PTO-Local	$.010_{[.000,.019]}$	.507	.012	.675@.31	$-.019_{[-.026,-.011]}$	.97	−.020	$-.025_{[-.044,-.004]}$
Teacher-Only	$.017_{[.009,.024]}$	.509	.011	.663@.27	$-.019_{[-.028,-.011]}$	.92	−.021	$-.024_{[-.045,-.002]}$
Context-Only	$.019_{[.007,.025]}$	.506	.007	.696@.33	$-.015_{[-.023,-.006]}$	.97	−.015	$-.022_{[-.043,-.001]}$
VCNet	$.029_{[.021,.039]}$	.514	.026	.660@.34	$.022_{[.015,.032]}$	.98	.023	$.073_{[.055,.093]}$
GIKS	$.026_{[.018,.035]}$	.513	.023	.657@.33	$.022_{[.014,.031]}$	.98	.022	$.070_{[.051,.092]}$
AdditiveROI	$\mathbf{.031}_{[.022,.040]}$	**.516**	**.029**	.575@.34	$.023_{[.015,.032]}$	.97	.024	$.078_{[.059,.098]}$

agreement within items), aligned–reverse gap (response separation between top and bottom score quantiles), support violation at update rate (fraction of executed recommendations outside support, reported with the update rate), DR lift (DR improvement over the logging policy), per-acted lift (DR lift per non-noop action), and agreement lift (response difference when logged actions agree with versus oppose the recommendation). Full definitions and inference details are in Appendix C.2.

Table 4 presents the results. On routine traffic, the matched residual $R = \tau + \epsilon$ exhibits a noise floor: the residual standard deviation $sd(R) \approx 0.34$ vastly overshadows the true per-move causal effect $sd(\tau) \approx 0.018$, yielding a SNR of merely ≈ 0.05. Even an oracle ranker with perfect knowledge of τ attains a maximum Spearman $\rho^\star \approx sd(\tau)/sd(R) \approx 0.05$ (verified via Monte-Carlo: 0.050 ± 0.004). Within this highly noisy regime, ReAlloc achieves a competitive ρ of 0.028 (recovering over 50% of the oracle bound), performing on par with strong baselines like AdditiveROI. More importantly, ReAlloc's explicit support layer completely eliminates historical support violations (0.000 vs. 0.678 for the ablated version, and 57.5%–69.6% for baselines) without sacrificing the update rate. This highlights a critical flaw in existing methods: they frequently recommend actions in regions lacking empirical evidence. The catastrophic failure of ReAlloc w/o Supp. (negative DR lift of −0.021) empirically proves that constraining decisions within the data support is indispensable for preventing disastrous extrapolation errors. On the randomized exploration set, ReAlloc achieves the highest DR lift (0.025) while intervening on only 33% of the cases, contrasting sharply with the near-ubiquitous interventions (97%) triggered by AdditiveROI. This translates to a significantly higher per-action lift (0.069 vs. 0.024). In real-world e-commerce operations, frequent budget or price adjustments incur implicit friction costs, destabilize item pricing, and can degrade user trust. Therefore, a "less but more accurate" intervention strategy is highly preferred. By delivering superior aggregate causal lift through fewer, highly targeted, and empirically supported interventions, ReAlloc demonstrates exceptional operational efficiency and deployment readiness.

7.2 Randomized Online A/B Test

We conducted a 14-day online A/B test on 300K eligible items, assigning 10% to the treatment arm (ReAlloc) and the remainder to the control (AdditiveROI). Strict item-level randomization prevents budget interference, with both arms sharing identical eligibility and risk constraints. As shown in Table 5, ReAlloc increases pay orders by 3.53%. This improvement is achieved without additional marketing expenditure: total realized spend decreases by 2.47%, while overall marketing ROI remains statistically unchanged. At the same time, profit margin improves by 3.26 percentage points, and the corresponding platform income improves by 3.26 percentage points. The higher transaction volume does not translate into higher aggregate GMV during the current observation, GMV decreases by 2.64%. Thus, ReAlloc generates more transactions with lower total expenditure and improved profitability, but exhibits a trade-off against average order value and GMV.

Table 5: Online A/B-test results.

Metric	Effect	95% CI	(p)-value
Pay orders	**+3.5**%	[**+2.8**%, **+5.1**%]	.001
GMV	−2.6%	[−3.1%, −1.1%]	.003
Total cost	−2.5%	[−2.0%, −0.8%]	.006
Marketing ROI	−0.1%	[−1.8%, +1.5%]	.820
Profit margin	**+1.42**pt	[**+0.8**pt, **+2.0**pt]	.001
Platform income	+3.2pt	[**+2.1**pt, **+4.4**pt]	.001

8 Conclusion

We formulate fixed-budget multi-channel uplift as policy learning on the simplex and identify supported local causal reallocation as the relevant decision primitive. ReAlloc combines orthogonal local-effect estimation, marginal-field distillation, and support-aware updates. Synthetic and offline studies show improved deployable policy value under confounding and limited overlap. An A/B test on TaoBao increases pay orders and profitability while revealing a GMV trade-off.

References

[1] Meng Ai, Biao Li, Heyang Gong, Qingwei Yu, Shengjie Xue, Yuan Zhang, Yunzhou Zhang, and Peng Jiang. 2022. LBCF: A Large-Scale Budget-Constrained Causal Forest Algorithm. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 2310–2319. doi:10.1145/3485447.3512103

[2] Javier Albert and Dmitri Goldenberg. 2022. E-Commerce Promotions Personalization via Online Multiple-Choice Knapsack with Uplift Modeling. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (Atlanta, GA, USA) *(CIKM '22)*. Association for Computing Machinery, New York, NY, USA, 2863–2872. doi:10.1145/3511808.3557100

[3] Susan Athey and Stefan Wager. 2021. Policy Learning with Observational Data. *Econometrica* 89, 1 (2021), 133–161. doi:10.3982/ECTA15732

[4] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. Double/Debiased Machine Learning for Treatment and Structural Parameters. *The Econometrics Journal* 21, 1 (2018), C1–C68. doi:10.1111/ectj.12097

[5] Yuan Deng, Negin Golrezaei, Patrick Jaillet, Jason Cheuk Nam Liang, and Vahab Mirrokni. 2023. Multi-channel Autobidding with Budget and ROI Constraints. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 7617–7644. https://proceedings.mlr.press/v202/deng23c.html

[6] Floris Devriendt, Jente Van Belle, Tias Guns, and Wouter Verbeke. 2022. Learning to Rank for Uplift Modeling. *IEEE Transactions on Knowledge and Data Engineering* 34, 10 (2022), 4888–4904. doi:10.1109/TKDE.2020.3048510

[7] Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. 2014. Doubly Robust Policy Evaluation and Optimization. *Statist. Sci.* 29, 4 (2014), 485–511. doi:10.1214/14-STS500

[8] Miroslav Dudík, John Langford, and Lihong Li. 2011. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning* (Bellevue, Washington, USA) *(ICML'11)*. Omnipress, Madison, WI, USA, 1097–1104.

[9] Adam N. Elmachtoub and Paul Grigas. 2022. Smart "Predict, then Optimize". *Management Science* 68, 1 (2022), 9–26. doi:10.1287/mnsc.2020.3922

[10] Sahin Cem Geyik, Abhishek Saxena, and Ali Dasdan. 2015. Multi-Touch Attribution Based Budget Allocation in Online Advertising. *arXiv preprint arXiv:1502.06657* (2015).

[11] Pierre Gutierrez and Jean-Yves Gérardy. 2017. Causal Inference and Uplift Modelling: A Review of the Literature. In *Proceedings of The 3rd International Conference on Predictive Applications and APIs (Proceedings of Machine Learning Research, Vol. 67)*, Claire Hardgrove, Louis Dorard, Keiran Thompson, and Florian Douetteau (Eds.). PMLR, 1–13.

[12] Keisuke Hirano and Guido W. Imbens. 2004. *The Propensity Score with Continuous Treatments*. John Wiley & Sons, Ltd, Chapter 7, 73–84. doi:10.1002/0470090456.ch7

[13] Edward H. Kennedy, Zongming Ma, Matthew D. McHugh, and Dylan S. Small. 2017. Non-parametric Methods for Doubly Robust Estimation of Continuous Treatment Effects. *Journal of the Royal Statistical Society: Series B* 79, 4 (2017), 1229–1245. doi:10.1111/rssb.12212

[14] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. 2020. Conservative Q-Learning for Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1179–1191. https://proceedings.neurips.cc/paper_files/paper/2020/file/0d2b2061826a5df3221116a5085a6052-Paper.pdf

[15] Sachin Kumar, Garima Gupta, Ranjitha Prasad, Arnab Chatterjee, Lovekesh Vig, and Gautam Shroff. 2020. CAMTA: Causal Attention Model for Multi-touch Attribution. *arXiv preprint arXiv:2012.11403* (2020).

[16] Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. 2019. Metalearners for Estimating Heterogeneous Treatment Effects Using Machine Learning. *Proceedings of the National Academy of Sciences* 116, 10 (2019), 4156–4165. doi:10.1073/pnas.1804597116

[17] Xinkun Nie and Stefan Wager. 2021. Quasi-Oracle Estimation of Heterogeneous Treatment Effects. *Biometrika* 108, 2 (2021), 299–319. doi:10.1093/biomet/asaa076

[18] Diego Olaya, Kristof Coussement, and Wouter Verbeke. 2020. A Survey and Benchmarking Study of Multitreatment Uplift Modeling. *Data Mining and Knowledge Discovery* 34, 2 (2020), 273–308. doi:10.1007/s10618-019-00670-y

[19] Nicholas Radcliffe. 2007. Using Control Groups to Target on Predicted Lift: Building and Assessing Uplift Model. *Direct Marketing Analytics Journal* (2007), 14–21.

[20] Utsav Sadana, Abhilash Chenreddy, Erick Delage, Alexandre Forel, Emma Frejinger, and Thibaut Vidal. 2025. A Survey of Contextual Optimization Methods for Decision-Making under Uncertainty. *European Journal of Operational Research* 320, 2 (2025), 271–289. doi:10.1016/j.ejor.2024.03.020

[21] Uri Shalit, Fredrik D. Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 3076–3085. https://proceedings.mlr.press/v70/shalit17a.html

[22] Guangyuan Shen, Shenjie Sun, Dehong Gao, Shaolei Li, Libin Yang, Yongping Shi, and Wei Ning. 2023. Cross-channel Budget Coordination for Online Advertising System. *arXiv preprint arXiv:2305.06883* (2023).

[23] Claudia Shi, David Blei, and Victor Veitch. 2019. Adapting Neural Networks for the Estimation of Treatment Effects. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/8fb5f8be2aa9d6c64a04e3ab9f63feee-Paper.pdf

[24] Zexu Sun, Hao Yang, Dugang Liu, Yunpeng Weng, Xing Tang, and Xiuqiang He. 2024. End-to-End Cost-Effective Incentive Recommendation under Budget Constraint with Uplift Modeling. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 560–569. doi:10.1145/3640457.3688147

[25] Stefan Wager and Susan Athey. 2018. Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests. *J. Amer. Statist. Assoc.* 113, 523 (2018), 1228–1242. doi:10.1080/01621459.2017.1319839

[26] Bingzhe Wang, Tianyu Wang, Qi Qi, Xiaoxuan Deng, Zhilin Zhang, and Chuan Yu. 2026. Marketing Hosting: From Fixed to Endogenous Budgets. In *Proceedings of the ACM Web Conference 2026* (United Arab Emirates) *(WWW '26)*. Association for Computing Machinery, New York, NY, USA, 327–338. doi:10.1145/3774904.3792519

[27] Bryan Wilder, Bistra Dilkina, and Milind Tambe. 2019. Melding the Data-Decisions Pipeline: Decision-Focused Learning for Combinatorial Optimization. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (Jul. 2019), 1658–1665. doi:10.1609/aaai.v33i01.33011658

[28] Justin R. Williams and Catherine M. Crespi. 2020. Causal Inference for Multiple Continuous Exposures via the Multivariate Generalized Propensity Score. *arXiv preprint arXiv:2008.13767* (2020). doi:10.48550/arXiv.2008.13767

[29] Yan Zhao, Xiao Fang, and David Simchi-Levi. 2017. Uplift Modeling with Multiple Treatments and General Response Types. In *Proceedings of the 2017 SIAM International Conference on Data Mining*. SIAM, 588–596. doi:10.1137/1.9781611974973.66

[30] Zhenyu Zhao and Totte Harinen. 2019. Uplift Modeling for Multiple Treatments with Cost Optimization. In *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 422–431. doi:10.1109/DSAA.2019.00057

[31] Kailiang Zhong, Fengtong Xiao, Yan Ren, Yaorong Liang, Wenqing Yao, Xiaofeng Yang, and Ling Cen. 2022. 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* (Washington DC, USA) *(KDD '22)*. Association for Computing Machinery, New York, NY, USA, 4612–4620. doi:10.1145/3534678.3539198

A Formal Assumptions and Proofs

A.1 Notation and Assumptions

Let $d := K - 1$, $\bar{p} := K^{-1}\mathbf{1}$, and choose $Q \in \mathbb{R}^{K\times d}$ such that

$$Q^\top Q = I_d, \qquad QQ^\top = \Pi_{\mathcal{T}}, \qquad Q^\top \mathbf{1} = 0. \tag{41}$$

Write

$$p = \bar{p} + Qz, \qquad Z := Q^\top P, \qquad \widetilde{Z} := Z - z,$$
$$v(X, z) := \mu(X, \bar{p} + Qz), \quad \beta^\star(X, z) := \nabla_z v(X, z). \tag{42}$$

Then

$$g^\star(X, p) = Q\beta^\star(X, Q^\top p). \tag{43}$$

For a bounded, compactly supported kernel K and bandwidth $h > 0$, define

$$\mathbb{E}_{h,z}[A \mid X] := \frac{\mathbb{E}[\mathsf{K}((Z - z)/h)A \mid X]}{\mathbb{E}[\mathsf{K}((Z - z)/h) \mid X]}, \tag{44}$$
$$e_{h,z}(X) := \mathbb{E}_{h,z}[\widetilde{Z} \mid X], \qquad m_{h,z}(X) := \mathbb{E}_{h,z}[Y \mid X]. \tag{45}$$

Let $\Psi_{h,z}$ be the score in (27) and $\|f\|_{\infty,\mathcal{R}} := \sup_{(X,p)\in\mathcal{R}} \|f(X, p)\|_2$.

Assumption 1 (Identification). *For every $p \in C(X)$,*

$$Y = Y(P), \qquad \{Y(p) : p \in C(X)\} \perp P \mid X, \tag{46}$$

and all conditional moments below exist.

Assumption 2 (Supported paths and local overlap). *For almost every X, $C(X) \subseteq \Delta^{K-1}$ is compact, contains $p_0(X)$, and contains every path invoked below. Uniformly over $(X, p) \in \mathcal{R}$, with $z = Q^\top p$,*

$$c_0 h^d \le \mathbb{E}[\mathsf{K}((Z - z)/h) \mid X] \le C_0 h^d \tag{47}$$

for all sufficiently small h and constants $0 < c_0 \le C_0 < \infty$.

Assumption 3 (Local smoothness and curvature). *The response is continuously differentiable on $C(X)$. Define*

$$R_{X,z}(u) := v(X, z + u) - v(X, z) - \beta^\star(X, z)^\top u, \tag{48}$$
$$a_{h,z} := \widetilde{Z} - e_{h,z}(X),$$
$$\Sigma_{h,z}(X) := \mathbb{E}_{h,z}[a_{h,z}a_{h,z}^\top \mid X], \tag{49}$$
$$C_{h,z}(X) := \mathbb{E}_{h,z}\left[a_{h,z}\{R_{X,z}(\widetilde{Z}) - \mathbb{E}_{h,z}[R_{X,z}(\widetilde{Z}) \mid X]\} \mid X\right]. \tag{50}$$

Uniformly on $\mathcal{R}$,

$$\kappa h^2 I_d \preceq \Sigma_{h,z}(X) \preceq \bar{\kappa} h^2 I_d, \qquad \|\Sigma_{h,z}(X)^{-1} C_{h,z}(X)\|_2 \le \epsilon_{\mathrm{loc}}(h), \tag{51}$$

where $\epsilon_{\mathrm{loc}}(h) \to 0$.

Assumption 4 (Cross-fitting and empirical complexity). *The nuisance estimators are cross-fitted. For the population root $\beta_h^\dagger$ of $\Psi_{h,z}(\beta, m_{h,z}, e_{h,z}; X) = 0$, let*

$$\delta_m(X, z) := \widehat{m}_{h,z}(X) - m_{h,z}(X), \quad \delta_e(X, z) := \widehat{e}_{h,z}(X) - e_{h,z}(X). \tag{52}$$

Suppressing (X, z), set

$$u_{h,z} := \Sigma_{h,z}^{-1}\delta_e, \qquad q_{h,z} := \delta_m - \beta_h^{\dagger\top}\delta_e,$$
$$r_{\mathrm{orth}} := \sup_{(X,p)\in\mathcal{R}} \|u_{h,z}q_{h,z}\|_2. \tag{53}$$

With probability at least $1 - \delta$,

$$\sup_{(X,p)\in\mathcal{R}} \left\|\widehat{\Psi}_{h,z}(\beta_h^\dagger, \widehat{m}_{h,z}, \widehat{e}_{h,z}; X) - \Psi_{h,z}(\beta_h^\dagger, \widehat{m}_{h,z}, \widehat{e}_{h,z}; X)\right\|_2 \le Ch\sqrt{\frac{\mathfrak{C}_{\mathcal{G}} + \log(1/\delta)}{nh^d}}. \tag{54}$$

The empirical local curvature is at least $\kappa h^2/2$, and root-solving error is of smaller order than the right-hand side of (54).

Assumption 5 (Composite teacher and student approximation). *The implemented teacher and deployed student satisfy*

$$\|\widehat{g}_T - \widehat{g}_T^{\mathrm{orth}}\|_{\infty,\mathcal{R}} \le \epsilon_{\mathrm{comp}}, \qquad \|\widehat{g} - \widehat{g}_T\|_{\infty,\mathcal{R}} \le \epsilon_S. \tag{55}$$

A.2 Proof of Theorem 5.1

Proof. Since $\mathbf{1}^\top\gamma(t) = 1$, $\dot{\gamma}(t) \in \mathcal{T}$ almost everywhere. Hence

$$\begin{aligned}\Delta(X, p) &= \int_0^1 \nabla_p\mu(X, \gamma(t))^\top\dot{\gamma}(t)\,\mathrm{d}t \\ &= \int_0^1 \{\Pi_{\mathcal{T}}\nabla_p\mu(X, \gamma(t))\}^\top\dot{\gamma}(t)\,\mathrm{d}t \\ &= \int_0^1 g^\star(X, \gamma(t))^\top\dot{\gamma}(t)\,\mathrm{d}t.\end{aligned} \tag{56}$$

□

A.3 Proof of Proposition 1

Proof. Identify the two-channel simplex with $[0, 1]$. Let $P = 0$ almost surely, $\mu(p) = -cp$, and $\widehat{\mu}(p) = Mp$, where $c, M > 0$. Then

$$\mathbb{E}_{\log}[\{\widehat{\mu}(P) - \mu(P)\}^2] = 0, \qquad p^\star = 0,$$
$$\widehat{p}_{\mathrm{PTO}} = 1, \qquad \mu(p^\star) - \mu(\widehat{p}_{\mathrm{PTO}}) = c. \tag{57}$$

For $\mu \equiv 0$ and $\widehat{\mu}_n(p) = n^{-1/2}\sin(np)$,

$$\|\widehat{\mu}_n - \mu\|_{L_2([0,1])}^2 \le n^{-1} \to 0, \qquad \|\widehat{\mu}_n' - \mu'\|_\infty = \sqrt{n} \to \infty. \tag{58}$$

□

A.4 Shortcut-Gradient Decomposition

Assume

$$Y = \phi(X, P, U) + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid X, P, U] = 0, \tag{59}$$

and let $\pi_b(p \mid X, U)$ be the logging density. With $F_p := F(U \mid X, P = p)$ and $s_b(U) := \nabla_p \log \pi_b(p \mid X, U)$, Bayes' rule yields

$$\nabla_p \log f(U \mid X, P = p) = s_b(U) - \mathbb{E}_{F_p}[s_b(U)]. \tag{60}$$

Under dominated differentiation,

$$\nabla_p m(X, p) = \mathbb{E}_{F_p}[\nabla_p\phi(X, p, U)] + \mathrm{Cov}_{F_p}(\phi(X, p, U), s_b(U)), \tag{61}$$
$$\nabla_p\mu(X, p) = \mathbb{E}_{F(U|X)}[\nabla_p\phi(X, p, U)]. \tag{62}$$

Therefore

$$b_{\mathrm{sc}}(X, p) = \Pi_{\mathcal{T}}\int \nabla_p\phi(X, p, u)\{\mathrm{d}F_p(u) - \mathrm{d}F(u \mid X)\} + \Pi_{\mathcal{T}}\,\mathrm{Cov}_{F_p}(\phi(X, p, U), s_b(U)). \tag{63}$$

Assumption 1 implies $m(X, p) = \mu(X, p)$ on $C(X)$; orthogonality alone does not.

A.5 Proof of Theorem 5.2

Proof. Fix (X, z) and abbreviate $e = e_{h,z}(X)$, $m = m_{h,z}(X)$, $a = \widetilde{Z} - e$, $R = R_{X,z}(\widetilde{Z})$, and $\bar{R} = \mathbb{E}_{h,z}[R \mid X]$. Under Assumption 1,

$$Y - m = \beta^\star(X, z)^\top a + (R - \bar{R}) + \varepsilon, \qquad \mathbb{E}_{h,z}[\varepsilon \mid X, Z] = 0. \quad (64)$$

Substitution into (27) gives

$$\Psi_{h,z}(\beta, m, e; X) = \Sigma_{h,z}(X)\{\beta^\star(X, z) - \beta(X, z)\} + C_{h,z}(X), \quad (65)$$

$$\beta_h^\dagger(X, z) - \beta^\star(X, z) = \Sigma_{h,z}(X)^{-1} C_{h,z}(X). \quad (66)$$

Assumption 3 and $\|Qv\|_2 = \|v\|_2$ imply (28).

Let $r := Y - m - \beta_h^{\dagger\top}(\widetilde{Z} - e)$. Since $\mathbb{E}_{h,z}[\widetilde{Z} - e \mid X] = \mathbb{E}_{h,z}[r \mid X] = 0$, for arbitrary directions η_m, η_e,

$$D_m \Psi_{h,z}[\eta_m] = -\eta_m \mathbb{E}_{h,z}[\widetilde{Z} - e \mid X] = 0, \quad (67)$$

$$D_e \Psi_{h,z}[\eta_e] = -\eta_e \mathbb{E}_{h,z}[r \mid X] + \mathbb{E}_{h,z}[\widetilde{Z} - e \mid X]\beta_h^{\dagger\top}\eta_e = 0. \quad (68)$$

Direct expansion yields

$$\Psi_{h,z}(\beta_h^\dagger, \widehat{m}_{h,z}, \widehat{e}_{h,z}; X) - \Psi_{h,z}(\beta_h^\dagger, m_{h,z}, e_{h,z}; X) = \delta_e\{\delta_m - \beta_h^{\dagger\top}\delta_e\}. \quad (69)$$

After multiplication by $\Sigma_{h,z}^{-1}$, its norm is bounded by r_{orth}. In raw nuisance norms, if $|\delta_m| \le \bar{r}_m$, $\|\delta_e\|_2 \le \bar{r}_e$, and $\|\beta_h^\dagger\|_2 \le M_\beta$, then

$$r_{\text{orth}} \le \frac{\bar{r}_e(\bar{r}_m + M_\beta \bar{r}_e)}{\kappa h^2}. \quad (70)$$

Assumption 4 and empirical curvature give

$$\|\widehat{\beta} - \beta_h^\dagger\|_{\infty,\mathcal{R}} \le C\sqrt{\frac{\mathfrak{C}_\mathcal{G} + \log(1/\delta)}{nh^{d+2}}} + C r_{\text{orth}}. \quad (71)$$

Combining (28) and (71), using $d + 2 = K + 1$, yields

$$\|\widehat{g}_T^{\text{orth}} - g^\star\|_{\infty,\mathcal{R}} \le \epsilon_T(n, h, \delta). \quad (72)$$

Finally, (29) and the triangle inequality imply (30). □

A.6 Proof of Theorem 5.3

Proof. For any $\gamma \in \mathfrak{P}_X(p)$, $\dot{\gamma}(t) \in \mathcal{T}$ almost everywhere; hence

$$\begin{aligned}\widehat{\Delta}_\theta(X, p) &= s_\theta(X, p) - s_\theta(X, p_0(X)) \\ &= \int_0^1 \nabla_p s_\theta(X, \gamma(t))^\top \dot{\gamma}(t)\, \mathrm{d}t \\ &= \int_0^1 \widehat{g}(X, \gamma(t))^\top \dot{\gamma}(t)\, \mathrm{d}t. \end{aligned} \quad (73)$$

Theorem 5.1 and Cauchy–Schwarz give

$$\begin{aligned}|\widehat{\Delta}_\theta(X, p) - \Delta(X, p)| &\le \int_0^1 \|\widehat{g}(X, \gamma(t)) - g^\star(X, \gamma(t))\|_2 \|\dot{\gamma}(t)\|_2\, \mathrm{d}t \\ &\le L(\gamma)\epsilon_g, \end{aligned} \quad (74)$$

which proves (36).

Let $\pi^\star = \pi^\star_{L_{\max}}$. Adding and subtracting estimated uplift gives

$$\begin{aligned}\mathrm{Reg}_{L_{\max}}(\widehat{\pi}) &= \mathbb{E}_X\left[\Delta(X, \pi^\star(X)) - \Delta(X, \widehat{\pi}(X))\right] \\ &\le 2L_{\max}\epsilon_g + \mathbb{E}_X\left[\widehat{\Delta}_\theta(X, \pi^\star(X)) - \widehat{\Delta}_\theta(X, \widehat{\pi}(X))\right] \\ &\le 2L_{\max}\epsilon_g + \epsilon_{\text{search}}, \end{aligned} \quad (75)$$

proving (37).

Under Assumption 5 and Theorem 5.2,

$$\|\widehat{g} - g^\star\|_{\infty,\mathcal{R}} \le \epsilon_S + \epsilon_{\text{comp}} + \epsilon_T(n, h, \delta), \quad (76)$$

which yields (38). □

Implemented greedy trace. Let γ_{gr} be the piecewise-linear interpolation of

$$p^{(0)} = p_0(X), \quad p^{(1)}, \ldots, p^{(R)}.$$

If every accepted segment lies in $C(X)$, then

$$L(\gamma_{\text{gr}}) = \sum_{r=0}^{R-1} \|p^{(r+1)} - p^{(r)}\|_2, \quad (77)$$

$$\sum_{r=0}^{R-1}\{s_\theta(X, p^{(r+1)}) - s_\theta(X, p^{(r)})\} = s_\theta(X, p^{(R)}) - s_\theta(X, p^{(0)}). \quad (78)$$

A.7 Conditional Outcome Improvement

Corollary A.1 (Conditional outcome improvement). *Suppose $U(X, p)$ satisfies, on an event $\mathcal{E}_U$,*

$$|\widehat{\Delta}_\theta(X, p) - \Delta(X, p)| \le U(X, p), \qquad \forall X,\ p \in \mathcal{A}_{L_{\max}}(X). \quad (79)$$

Define

$$\pi_U(X) := \begin{cases} \widehat{\pi}(X), & \widehat{\Delta}_\theta(X, \widehat{\pi}(X)) - U(X, \widehat{\pi}(X)) \ge 0, \\ p_0(X), & \textit{otherwise.} \end{cases} \quad (80)$$

Then $\Delta(X, \pi_U(X)) \ge 0$ for every X on $\mathcal{E}_U$.

Proof. The fallback case has zero uplift. Otherwise,

$$\Delta(X, \pi_U(X)) \ge \widehat{\Delta}_\theta(X, \widehat{\pi}(X)) - U(X, \widehat{\pi}(X)) \ge 0. \quad (81)$$

□

Without separate calibration, ensemble or MC-dropout dispersion is not assumed to satisfy (79).

B Additional Details of the Synthetic Experiments

B.1 State and Budget Generation

For each item i, we independently sample persistent state blocks:

$$C_i^{\text{item}} \sim \mathcal{N}(0, I_{d_C}), \quad E_i^{\text{item}} \sim \mathcal{N}(0, I_{d_E}), \quad S_i^{\text{item}} \sim \mathcal{N}(0, I_{d_S}).$$

For each period t, we sample lower-variance temporal shocks:

$$C_t^{\text{time}}, E_t^{\text{time}}, S_t^{\text{time}} \sim \mathcal{N}(0, \sigma_t^2 I).$$

The observed state at time t is constructed as:

$$C_{it} = C_i^{\text{item}} + C_t^{\text{time}}, \quad E_{it} = E_i^{\text{item}} + E_t^{\text{time}}, \quad S_{it} = S_i^{\text{item}} + S_t^{\text{time}}, \quad H_{it} = [C_{it}, E_{it}, S_{it}].$$

We set the dimensions $d_C = d_E = d_S = 6$ and the temporal variance $\sigma_t = 0.35$. The total budget is generated via:

$$B_{it} = \text{clip}\left(\exp\{a_B^\top C_{it} + \zeta_{it}\}, B_{\min}, B_{\max}\right), \qquad \zeta_{it} \sim \mathcal{N}(0, 0.2^2), \quad (82)$$

with boundary constraints $B_{\min} = 0.2$ and $B_{\max} = 5$. The train, validation, and test splits consist of mutually disjoint items to prevent data leakage.

Table 6: Static assignment regimes. The underlying true response surface is identical across all regimes.

Setting	α_{cf}	σ_{ov}	α_{bd}	ε_{exp}
Benign	0.25	0.40	0.0	0.10
Medium	0.75	0.25	0.0	0.05
Hard	1.25	0.15	0.0	0.02
Extreme	1.75	0.10	0.5	0.01

B.2 Conditional Logging Policy

To handle compositional allocations on the simplex, we utilize the isometric log-ratio (ILR) transform. Let $Q^\top Q = I_{K-1}$ and $Q^\top \mathbf{1} = 0$. For an interior simplex point p, we define $u(p) = Q^\top \log p$.

The logging policy is governed by a state-dependent propensity mean, which was introduced conceptually in the main text and is defined mathematically here as:

$$q(H,B) = \text{softmax}\big((1+\alpha_{\text{bd}})\lambda_{\log}\left[W_S S + \alpha_{\text{cf}} W_C C + w_B \log(1+B)\right]\big). \tag{83}$$

The context-dependent covariance matrix is formulated as:

$$\Sigma_H = \sigma_{\text{ov}}^2 \left\{1 + 0.2\tanh(E_1)\right\}^2 \text{diag}(v_1, \ldots, v_{K-1}), \tag{84}$$

where σ_{ov} dictates the overlap scale and $v_j > 0$ ensures every tangent direction remains locally non-degenerate.

Given the logging mean $q(H,B)$ from Eq. (83), the final allocation is sampled via a logistic-normal mixture:

$$u(P) = u(q(H,B)) + \xi, \;\; \xi \sim (1-\varepsilon_{\text{exp}})\mathcal{N}(0,\Sigma_H) + \varepsilon_{\text{exp}}\mathcal{N}(0, c_{\text{exp}}^2 \Sigma_H), \tag{85}$$

and mapped back to the simplex through $P = \text{softmax}\{Q[u(q)+\xi]\}$. The broad exploration component uses a scale multiplier $c_{\text{exp}} > 1$. Because the policy is logistic-normal, allocations strictly remain in the simplex interior, and no point mass is placed on any vertex.

The assignment parameters vary independently to test different regimes: α_{cf} controls dependence on the baseline-driving block C; σ_{ov} controls conditional action dispersion; α_{bd} smoothly sharpens the logging mean toward boundaries; and ε_{exp} controls the rate of broad exploration. Crucially, the true response surface and all response parameters are held fixed across these assignment regimes.

B.3 Static Assignment Regimes

The primary severity curve modulates confounding and overlap, utilizing the *Hard* regime to represent conditional low-overlap settings. The *Extreme* regime additionally stresses boundary coverage and serves as an assumption-stress test when local overlap assumptions fail.

We use a shared logging-logit scale $\lambda_{\log} = 0.45$.

B.4 Response Surface

The state-dependent baseline response is:

$$m_0(H,B) = 10 + 2C_1 + C_2^2 + 1.25\sin(C_3) + 0.6C_1C_4 + 2\log(1+B). \tag{86}$$

The treatment scale modifier is:

$$\rho(H,B) = \alpha_\rho \log(1+B)\left\{1 + 0.2\tanh(E_1) + 0.1\tanh(C_1)\right\}. \tag{87}$$

The context-dependent response anchor $c(H,B) \in \Delta^{K-1}$ is generated from fixed context coefficients, and the deviation in the ILR space is defined as $z = Q^\top\{p - c(H,B)\}$.

The **local component** captures linear marginal returns:

$$g_{\text{loc}}(H,z) = a(H)^\top z, \qquad a(H) = \tanh(W_E E + 0.35 W_C C + b_a). \tag{88}$$

For the **interaction component**, let

$$A(H) = \gamma_{\text{int}} \sum_{r=1}^{R} \tanh(v_r^\top E + b_r) A_r,$$

where each A_r is a normalized symmetric matrix. We employ a shifted quadratic representation:

$$g_{\text{int}}(H,z) = \frac{1}{2} z^\top A(H) z + b_{\text{int}}(H)^\top z, \tag{89}$$

where $b_{\text{int}}(H)$ is induced by a small center shift $s_{\text{int}} = 0.4$. This preserves smoothness while allowing interaction gradients to remain active around realistic logged allocations.

For the **far-field component**, let $r = \|z\|_2$. We define:

$$g_{\text{far}}(H,z) = r^2 \sigma\left(\frac{r - d_0}{s_0}\right) \times \left[\sum_{m=1}^{M} w_m(H) \exp\left(-\frac{\|z - v_m\|_2^2}{2\ell_m^2}\right) - \kappa_{\text{can}}\right]. \tag{90}$$

The r^2 gate ensures that both the value and the first derivative vanish at the anchor point. Meanwhile, the low-frequency radial basis functions model distant saturation and cannibalization without introducing discontinuities or adversarial high-frequency oscillations.

Random coefficient matrices, quadratic bases, and RBF centers are sampled once per seed and subsequently frozen. On an independent calibration sample, we compute the expected gradient energy for each component:

$$\mathcal{E}_j = \mathbb{E}\left[\left\|\mathcal{P}_0 \nabla_p g_j(H,p)\right\|_2^2\right],$$

and rescale the three components to target normalized gradient-energy shares of $\mathcal{E}_{\text{loc}} : \mathcal{E}_{\text{int}} : \mathcal{E}_{\text{far}} \approx 0.45 : 0.35 : 0.20$. We use a treatment-response scale of 4.0 across all regimes.

B.5 Conditional Path Support

Learned policies utilize an estimated conditional support model, whereas synthetic deployability metrics rely on the oracle logging density. For the estimated model, we predict the conditional log-ratio mean $\hat{m}_u(H,B)$ and estimate a regularized local covariance $\hat{\Sigma}_u(H,B)$ from nearby training contexts, conditioning on both the predicted allocation anchor and the heteroskedastic effect coordinate. The pointwise nonconformity score is:

$$d_{\mathcal{S}}(H,B,p) = \frac{1}{2}\Big[(u(p) - \hat{m}_u)^\top \hat{\Sigma}_u^{-1}(u(p) - \hat{m}_u) + \log\det\hat{\Sigma}_u + (K-1)\log(2\pi)\Big]. \tag{91}$$

The threshold $\epsilon_{\mathcal{S}}$ is calibrated from held-out factual actions to obtain a fixed conditional high-density region.

Since evaluating endpoint support alone might inadvertently connect disjoint action regions, a recommendation p is evaluated

along the linear segment from the factual action P:

$$d_{\text{path}}(H, B, P, p) = \max_{s \in \mathcal{G}_J} d_S(H, B, (1-s)P + sp), \tag{92}$$

where $\mathcal{G}_J = \{0, 1/J, \ldots, 1\}$. We use at least ten interpolation intervals and verify robustness to finer discretizations. The oracle evaluator replaces $(\hat{m}_u, \hat{\Sigma}_u)$ with the known conditional logistic-normal mixture density.

B.6 Oracle Geometry and Evaluation Metrics

The oracle directed effect of reallocating budget from channel k to l is defined as:

$$D^{\star}_{k \to l}(H, B, p) = \frac{\partial \mu_{\star}(H, B, p)}{\partial p_l} - \frac{\partial \mu_{\star}(H, B, p)}{\partial p_k}. \tag{93}$$

Let $\mathcal{E} = \{(k, l) : k \neq l\}$. On a common held-out anchor set:

- PairwiseCorr is the Pearson correlation between learned and oracle scores over $\mathcal{E}$.
- TopEdgeAcc is the fraction of anchors identifying the correct highest-gain edge.
- EdgeNDCG evaluates the full ranking of directed edges using nonnegative oracle gains as relevance.

The support-aware oracle local-greedy policy uses the identical candidate step set, path-support rule, and total movement budget as the learned local policies, but scores candidates using the true $\mu_{\star}$. It defines the denominator of SafeLocalRecovery. The global support-constrained oracle searches all supported grid candidates and is reported strictly as a diagnostic upper bound; it is not directly comparable to the local improvement objective of ReAlloc. We additionally report the mean L_1 movement and the 90th percentile of path nonconformity.

B.7 Temporal Support Rotation

For temporal window w, we perturb only the logging logits:

$$q^{(w)}_{it} = \text{softmax}\left\{\log q(H_{it}, B_{it}) + \lambda_{\text{temp}} b_w\right\}, \tag{94}$$

where b_w follows a smooth cyclic schedule across channels and $\lambda_{\text{temp}} = 1.2$. All response parameters and the true surface $\mu_{\star}$ remain fixed. Consequently, the temporal experiment isolates whether the slow student can retain local geometric gradients learned from different support fragments; it intentionally does not model delayed or sequential causal effects.

B.8 Baseline Implementations

To ensure a fair comparison, all neural baselines share the same backbone architecture (a 3-layer MLP with ReLU activations and Layer Normalization) and are trained with the same optimizer and learning rate schedule.

- **Additive ROI:** Models the response as $\mu(H, B, P) = m_0(H, B) + \sum_{k=1}^{K} f_k(H, B, p_k)$. It completely ignores cross-channel interactions and optimizes each channel marginally subject to the budget constraint.
- **Joint S-Learner PTO:** Predicts the factual outcome $\hat{Y} = f(H, B, P)$ using Mean Squared Error (MSE). During inference, it employs a gradient-based global optimizer (L-BFGS) over the learned surface $\hat{Y}$ to find the optimal allocation on the simplex.
- **R-Learner Local:** Estimates the Conditional Average Treatment Effect (CATE) by minimizing the orthogonalized loss: $\mathbb{E}[((Y - \hat{m}(H, B)) - \sum_k \hat{\tau}_k(H, B)(p_k - \hat{e}_k(H, B)))^2]$. It then selects the local reallocation direction that maximizes the estimated marginal gain, restricted to the estimated support region.

B.9 Detailed Evaluation Metrics

In addition to the Deployable Uplift (U_{dep}) defined in the main text, we utilize the following metrics for comprehensive evaluation:

Raw Oracle Uplift. The unconstrained theoretical uplift, ignoring support boundaries:

$$U_{\text{raw}}(\hat{\pi}) = \frac{1}{N} \sum_{i=1}^{N} \Delta_i(\hat{\pi}). \tag{95}$$

Out-of-Support (OOS) Metrics. We measure the aggressiveness of the policy via the OOS rate and the corresponding gain lost due to fallback:

$$\text{OOS} = 1 - \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{\hat{p}_i \in \mathcal{S}_i\}, \qquad \text{OOSGain} = U_{\text{raw}}(\hat{\pi}) - U_{\text{dep}}(\hat{\pi}). \tag{96}$$

A high OOSGain indicates that the model is hallucinating high rewards in unsupported regions, which are subsequently clipped by the safety mechanism.

Safe Local Recovery. To quantify how close the learned policy is to the theoretical best local policy, we define:

$$\text{SafeLocalRecovery} = \frac{U_{\text{dep}}(\hat{\pi})}{U_{\text{dep}}(\pi^{\star}_{\text{local}}) + \varepsilon}, \tag{97}$$

where $\pi^{\star}_{\text{local}}$ is the support-aware oracle local-greedy policy operating under the identical movement budget constraint.

Geometric Ranking Metrics. To evaluate the local geometry independently of the final policy deployment, we compute the oracle directed effect of reallocating budget from channel k to l:

$$D^{\star}_{k \to l}(H, B, p) = \frac{\partial \mu_{\star}(H, B, p)}{\partial p_l} - \frac{\partial \mu_{\star}(H, B, p)}{\partial p_k}. \tag{98}$$

Based on $D^{\star}$, we report PairwiseCorr (Pearson correlation of edge scores), TopEdgeAcc (accuracy of identifying the highest-gain edge), and EdgeNDCG (ranking quality of all directed edges).

C Additional Details of the Taobao Dataset

C.1 Matching Protocol and Balance Diagnostics

Standard IPS is unsuitable for our continuous-action setting due to explosive variance caused by concentrated logging policies. Instead, we employ a retrospective matched replay to construct a reliable counterfactual baseline. For every item that underwent a budget reallocation (the *treated move*), we seek control items that experienced *no budget change* ($\|\Delta p\|_1 \approx 0$) on the same date and within the same product category. To ensure the treated and control items are highly comparable before the move, we perform nearest-neighbor matching based on key pre-treatment business features, primarily: **baseline GMV**, **total assigned budget**, and **historical traffic trends**. This ensures that any observed post-move difference in outcomes can be reasonably attributed to the budget reallocation

itself, rather than pre-existing differences in item popularity or scale.

To verify the quality of our matching, we measure the Standardized Mean Difference (SMD) for all matching covariates. An SMD below 0.1 is widely accepted as indicating excellent balance between the treatment and control groups, and our matching-relaxation selector enforces this exact threshold as a guardrail. As shown in Table 7, before matching, the reallocated (treated) items differed substantially from the general pool of no-move items on their pre-move covariates—most notably in their starting paid-ad share and assigned budget / price level. After applying our category × date stratified, caliper nearest-neighbour matching protocol, the maximum absolute SMD across all five covariates drops well below the 0.1 threshold (specifically to 0.018). This confirms that, on all measured pre-move covariates, our matched control group is well balanced and serves as an unbiased counterfactual baseline for evaluating the directional accuracy of our uplift models.

Table 7: Covariate balance before and after matching.

Business Feature	SMD Before	SMD After
Ad Share Prev	0.375	0.018
Log Price	0.257	0.005
Log budget	0.199	0.05
Recent Sales Level	0.047	0.008
m pred	0.082	0.007
Maximum Absolute SMD	**0.375**	**0.018**

C.2 Evaluation Metrics

Let $e = (i, t)$ denote an observed local reallocation event, with starting allocation p_e, realized movement Δp_e, and post-move outcome Y_e^+. For each event, the predicted directional score is

$$S_e^{(b)} = \begin{cases} \dfrac{\widehat{M}_b(H_e, p_e)^\top \Delta p_e}{\|\Delta p_e\|_1}, & \text{for marginal-field methods,} \\ \dfrac{\widehat{\mu}_b(H_e, p_e + \Delta p_e) - \widehat{\mu}_b(H_e, p_e)}{\|\Delta p_e\|_1}, & \text{for response-surface methods,} \end{cases}$$

where b indexes the evaluated method. This formulation evaluates all methods on the same realized movement and normalizes out the movement magnitude.

To remove predictable demand variation, we train an independent evaluation model $\widehat{m}_{\text{eval}}$ using only data preceding the evaluation period. The residualized outcome of event e is

$$\widetilde{Y}_e = Y_e^+ - \widehat{m}_{\text{eval}}(H_e).$$

Let $C(e)$ be its matched no-move controls and w_{ec} their normalized matching weights, where $\sum_{c \in C(e)} w_{ec} = 1$. We define the matched residual response as

$$R_e = \widetilde{Y}_e - \sum_{c \in C(e)} w_{ec} \widetilde{Y}_c.$$

We use R_e only as an observational proxy for evaluating directional ranking.

Marginal Rank Correlation. We measure the global alignment between predicted directions and empirical responses using Spearman's rank correlation:

$$\rho_{\text{MR}}^{(b)} = \text{Spearman}\left(\left\{S_e^{(b)}\right\}_{e \in \mathcal{E}}, \{R_e\}_{e \in \mathcal{E}}\right),$$

where $\mathcal{E}$ denotes all successfully matched reallocation events. A larger value indicates that the model more accurately ranks the relative value of observed local movements.

Within-Item Concordance. To test whether a model adapts to the current allocation rather than assigning a fixed channel preference to each item, we compare repeated events from the same item. Let $\mathcal{P}$ contain pairs (e, e') from the same item whose pre-decision contexts satisfy the prescribed similarity caliper. We compute

$$C_{\text{WI}}^{(b)} = \frac{1}{|\mathcal{P}|} \sum_{(e,e') \in \mathcal{P}} \left[\mathbb{I}\left(\left(S_e^{(b)} - S_{e'}^{(b)}\right)\left(R_e - R_{e'}\right) > 0 \right) + \frac{1}{2} \mathbb{I}\left(\left(S_e^{(b)} - S_{e'}^{(b)}\right)\left(R_e - R_{e'}\right) \right.\right.$$

The metric equals 0.5 under random pairwise ordering and approaches 1 when score differences consistently agree with response differences.

Aligned–Reverse Response Gap. Let $q_\alpha^{(b)}$ and $q_{1-\alpha}^{(b)}$ denote the lower and upper α-quantiles of the score distribution, with $\alpha = 0.2$. We define

$$\mathcal{E}_b^+ = \left\{ e : S_e^{(b)} \geq q_{1-\alpha}^{(b)} \right\}, \qquad \mathcal{E}_b^- = \left\{ e : S_e^{(b)} \leq q_\alpha^{(b)} \right\}.$$

The response gap is

$$G_{\text{AR}}^{(b)} = \frac{1}{|\mathcal{E}_b^+|} \sum_{e \in \mathcal{E}_b^+} R_e - \frac{1}{|\mathcal{E}_b^-|} \sum_{e \in \mathcal{E}_b^-} R_e.$$

This metric directly compares the empirical responses of movements most aligned with the model against those ranked least favorably by the model.