FROM GLOBAL BUDGETS TO PER-USER LTV HEALTH: OFFLINE REINFORCEMENT LEARNING FOR PERSONALIZED COUPON ALLOCATION UNDER EXTREME POPULATION SKEW

Anonymous authors
Paper under double-blind review

ABSTRACT

Coupon incentives are widely used to promote purchases and sustain user engagement. However, existing systems typically apply uniform rules within user tiers or optimize aggregate returns under a global budget, overlooking differences in users' lifetime value (LTV). This limitation is particularly severe in e-commerce referral platforms dominated by low-activity users, where sparse histories and highly skewed populations further complicate LTV estimation and offline policy learning. To address these challenges, we formulate daily joint coupon allocation as an offline reinforcement learning problem for personalized fee-rate control with user-specific, LTV-aware penalties. We propose a group-relative offline RL algorithm built on implicit Q-learning, termed LIGER, that represents evolving LTV health through time-varying state features and a structured reward balancing immediate incentive efficiency with long-term value changes. To mitigate majority-group dominance, LIGER combines group-balanced sampling with within-group reward and advantage normalization. It further replaces a separately fitted state-value network with a temperature-controlled softmax Bellman target derived from twin Q-critics, avoiding an additional value-function fit on imbalanced data. Offline evaluations and online A/B tests in a real-world e-commerce setting demonstrate the effectiveness of LIGER.

1 INTRODUCTION

Coupon-based incentives are widely used by e-commerce platforms to stimulate user engagement and promote purchase conversion (Zhao et al., 2019; Xiao et al., 2019). A fundamental question is how much subsidy should be allocated to each user: insufficient incentives may fail to activate purchase intent, whereas excessive subsidies reduce platform margins and may eventually result in negative lifetime value (LTV). Existing learning-based approaches typically optimize aggregate outcomes under a global budget (Zhao et al., 2019; Zhang et al., 2021). Although effective for controlling total expenditure, this formulation treats subsidy cost as fungible across users: excessive investment in one user can be offset by reduced investment in another. Consequently, aggregate budget efficiency does not necessarily ensure the long-term profitability of each individual user.

This limitation is particularly important in e-commerce referral domains, where engagement channels embedded in a broader ecosystem—such as daily check-in programs—use coupons to redirect users toward product purchases. We study such a platform serving more than ten million daily active users. Its production policy determines coupon thresholds and face values solely from users' monthly purchase-frequency tiers (M-levels), assigning a uniform fee rate within each tier. However, users within the same tier can differ substantially in purchase intent and LTV: users with persistently negative LTV may continue to receive subsidies, while profitable users with strong purchase potential may remain under-invested. Moreover, an earlier RL system that jointly optimized coupon thresholds and face values was costly to deploy because threshold changes required product-pool reconfiguration and frontend modifications.

Main challenges. The central difficulty is the extreme heterogeneity of the user population, which manifests as two coupled challenges. First, extreme user-tier imbalance causes a global critic to be

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

dominated by the low-activity majority, leading to miscalibrated advantages and a near-degenerate policy, which we term *value-space collapse*. Second, safety caps estimated from sparse histories are systematically underestimated for low-activity users. The resulting penalties can overwhelm behavioral signals and even reverse meaningful feature responses, a failure we call the *safety–feature paradox*, while standard training diagnostics remain apparently normal. Together, these challenges require an offline RL method that accommodates group-specific value scales and uncertain safety constraints under a highly heterogeneous population.

To address these challenges, we formulate personalized coupon allocation as an offline reinforcement learning problem with per-user lifetime-value (LTV) safety constraints and propose an **L**TV-aware **I**mplicit **G**roup-r**E**lative Q-lea**R**ning algorithm, termed **LIGER**. Our formulation incorporates an LTV-aware structured reward that extends beyond binary redemption outcomes to capture investment returns, fee-rate efficiency, LTV health-state transitions, and investment incentives aligned with user value. To mitigate errors in the safety caps estimated from sparse behavioral histories, we introduce *KAPPA-FLOOR*, a calibration mechanism that lower-bounds the penalty reference and prevents excessive safety penalties from overwhelming meaningful behavioral signals. At the algorithmic level, LIGER removes the separate value network used in standard IQL (Kostrikov et al., 2022) and introduces group-relative normalization inspired by Group Relative Policy Optimization (GRPO) (Shao et al., 2024). Specifically, it combines group-wise reward normalization, group-balanced sampling, and within-group advantage standardization, allowing each user group to learn within an appropriate value reference frame and mitigating cross-group value dominance. Together, these designs make policy learning robust to the strong heterogeneity of the user population. **Our main contributions are summarized as follows:**

- **LTV-aware coupon allocation formulation.** We formulate personalized coupon allocation as an offline RL problem with per-user LTV safety constraints, replacing the conventional global-budget objective (Zhang et al., 2021; Zhao et al., 2019). We further design a structured reward that jointly captures investment return, fee-rate efficiency, LTV health transitions, and user-value-aligned incentives.
- **Personalized Q-learning algorithm.** We propose a new implicit Q-learning algorithm, LIGER, which removes the separate value network through a temperature-controlled softmax-Q Bellman target and performs group-relative learning via reward normalization, balanced sampling, and advantage standardization to address user distribution heterogeneity. A sparse-history-aware safety-cap calibration mechanism further prevents unreliable LTV estimates from dominating policy learning.
- **Offline and online empirical evaluation.** We analyze state-conditioned LIGER actions, ablate its group-relative modules, and compare it with BC and standard IQL using model-based value and action-distribution measures. We further assess deployment outcomes through an online A/B experiment.

2 Related Work

Coupon and subsidy allocation. Coupon and marketing-incentive allocation is commonly formulated as constrained optimization under an aggregate budget. Zhao et al. (2019) combine treatment-effect estimation with global marketing-budget optimization, while Xiao et al. (2019) model sequential incentive marketing as a model-based constrained Markov decision process (CMDP). Cai et al. (2023) study marketing-budget allocation with offline constrained deep reinforcement learning. Most closely related to our application, BCORLE(λ) (Zhang et al., 2021) addresses coupon allocation by combining regularized batch-constrained Q-learning with off-policy evaluation and amortizing Lagrangian optimization through λ-generalization. Despite their different learning paradigms, these methods optimize an exogenously specified population-level budget, $\sum_u c_u \leq B$, under which subsidy cost is fungible across users. They do not jointly model a user-specific LTV cap, uncertainty in estimating that cap, and explicit group-level population heterogeneity. LIGER changes the constraint modality by deriving a separate LTV safety cap from each user's history and calibrating policy learning within user groups. Table 1 summarizes this problem-level distinction. Uplift-based pipelines (Rzepakowski & Jaroszewicz, 2010; Rudaś & Jaroszewicz, 2018) offer an interpretable alternative but typically optimize interventions independently and therefore do not capture sequential LTV health transitions.

Table 1: Comparison with representative marketing-incentive and coupon-allocation methods. "Group handling" denotes explicit mechanisms for group-level population heterogeneity; a dash indicates that the corresponding property is not explicitly modeled.

Method	Application	Learning paradigm	Constraint	Cap source	LTV-aware	Group handling
Zhao et al. (2019)	Marketing allocation	Uplift optimization	Global budget	Given	–	–
Xiao et al. (2019)	Incentive marketing	Model-based CMDP	Global budget	Given	–	–
Cai et al. (2023)	Marketing allocation	Offline constrained RL	Global budget	Given	–	–
BCORLE (2021)	Coupon allocation	Offline RL	Global budget	Given	–	–
LIGER (Ours)	**Coupon allocation**	**Offline RL**	**Per-user LTV**	**Estimated**	✓	✓

Offline RL and group-relative learning. Offline RL learns policies from fixed datasets while mitigating behavior-policy distribution shift (Levine et al., 2020). Representative methods include conservative value estimation in CQL (Kumar et al., 2020), behavior-support constraints in BCQ (Fujimoto et al., 2019), and expectile-based policy extraction in IQL (Kostrikov et al., 2022). LIGER follows advantage-weighted regression (Peng et al., 2019) and shares AWAC's value-network-free design (Nair et al., 2020), but computes a soft expectation over the complete discrete action space, connecting its Bellman target to soft Q-learning (Haarnoja et al., 2017; 2018). These methods, however, do not explicitly account for group-dependent value scales. Group-relative estimation has been explored in language-model alignment through GRPO (Shao et al., 2024) and RLOO (Ahmadian et al., 2024), both of which construct relative baselines among responses to the same prompt. We extend this principle from online policy optimization to value-based offline RL, using persistent user attributes to define groups and calibrating rewards, sampling, and advantages within those groups.

Safe RL with estimated constraints. Our per-user LTV constraint places the problem within the broader CMDP and safe-RL literature (Altman, 1999; García & Fernández, 2015). Lagrangian and primal-dual methods incorporate constraint violations through learned multipliers (Achiam et al., 2017; Ding et al., 2020; Paternain et al., 2019); reward-shaping methods use fixed penalties (Tessler et al., 2019); and shielding or safety-layer methods restrict unsafe actions at execution time (Alshiekh et al., 2018; Dalal et al., 2018; Chow et al., 2018). Offline approaches such as CPQ (Xu et al., 2022) penalize constraint-violating actions during value learning, whereas COptiDICE (Lee et al., 2022) enforces constraints through stationary-distribution correction. These methods primarily address how a specified constraint should be enforced. Unlike prior work, our constraint threshold is estimated from sparse user histories and can therefore be systematically underestimated. LIGER addresses this estimation error by combining reward-based enforcement with KAPPA-FLOOR, which lower-bounds the penalty reference without changing the platform guardrail.

LTV-aware decision-making. Customer lifetime value has long been studied in marketing science (Fader et al., 2005) and has been incorporated into downstream decisions such as LTV-constrained advertising and bidding (Cai et al., 2017). Much of this literature treats LTV as a static estimate supplied to a subsequent optimizer. In contrast, our formulation uses LTV in two complementary roles: as a user-specific safety constraint on subsidy intensity and as a dynamic learning signal that rewards improvements and penalizes deterioration in LTV health.

3 Problem Formulation

In this section, we formulate personalized coupon allocation as an offline reinforcement learning problem with the platform's operational constraints. In particular, the formulation specifies the feasible decision space, the offline sequential decision process, the estimated user-level LTV constraint, and the heterogeneous group structure that motivates group-relative learning.

3.1 Decision Setting and Offline Formulation

In our formulation, users are stratified by their monthly purchase frequency. Let $M_u \in \mathcal{M} = \{M_0, M_1, M_2, M_{3+}\}$ denote the tier of user u, where the subscript records the number of purchases in the preceding month and M_{3+} aggregates users with at least three purchases. For coupon $q \in \{1, 2\}$, tier M_u determines a threshold $T_q(M_u)$ and a base face value $B_q(M_u)$. These quantities may differ

across tiers. The thresholds remain fixed during policy execution because modifying them requires product-pool reconfiguration and interface changes.

The policy therefore personalizes only the face values, which is equivalent to setting each coupon's fee rate, i.e., the decision a our method ultimately makes for each user. In particularm, at each daily decision point, the policy selects an action $a = (k_1, k_2)$, where each k_q is the fee-rate adjustment level applied to coupon q and is drawn from the ordered finite set $\mathcal{K} = \{-K_-, \ldots, 0, \ldots, K_+\}$. A negative k_q shrinks and a positive k_q enlarges the issued face value relative to the tier baseline $B_q(M_u)$, thereby lowering or raising the fee rate. Here $K_-, K_+ \in \mathbb{N}$ specify the maximum downward and upward adjustments. The individual endpoints are operational parameters rather than elements of the general formulation, and their concrete range is specified in Appendix B. For a relative step size $\Delta > 0$, chosen such that $1 - \Delta K_- \geq 0$, the action a maps to the issued face value D_q and, crucially, the coupon fee rate ρ_q that is also the ratio of the issued face value to the fixed threshold. Mathmatically, the per-coupon subsidy intensity we personalize can be given by as follows:

$$D_q(u, a) = B_q(M_u)(1 + \Delta k_q), \qquad \rho_q(u, a) = \frac{D_q(u, a)}{T_q(M_u)}. \tag{1}$$

The action space is $\mathcal{A} = \mathcal{K} \times \mathcal{K}$. The interaction is modeled as a discounted Markov decision process $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$. The state $s \in \mathbb{R}^p$ summarizes the user's LTV health, purchasing profile, coupon-response history, temporal context, and intent signals; the complete feature specification is provided in the appendix. An action determines the face values for the current decision period, and the subsequent transition incorporates redemption outcomes and the resulting behavioral and LTV updates. The reward R captures immediate investment efficiency and changes in long-term LTV health.

Given above definitions, our policy learning uses a fixed dataset of logged transitions over T days, $\mathcal{D}_{\mathrm{tr}} = \{(s_i, a_i, r_i, s_i^+, \delta_i)\}_{i=1}^T$, where s_i^+ is the successor state and $\delta_i \in \{0, 1\}$ indicates termination. User identifiers and derived quantities used only to construct s_i or r_i are omitted from the tuple. In particular, the purchase and health tiers are recovered as $M_i = M(s_i)$ and $H_i = H(s_i)$, while the safety reference is computed during reward preprocessing. Let μ_{de} denote the joint distribution of users and initial states in the deployment population. The business objective is to maximize the expected discounted return subject to the per-user coupon fee-rate constraints:

$$\begin{aligned} &\pi^\star \in \arg\max_\pi J_{\mathrm{de}}(\pi) = \mathbb{E}_{\tau \sim (\mu_{\mathrm{de}}, \pi, P)} \left[\sum_{t=0}^{T-1} \gamma^t R(s_t, a_t, s_{t+1}) \right], \\ &\text{s.t.} \underbrace{\rho_q(T) = \frac{1}{T} \sum_{t=0}^{T-1} \rho_q(u, a_t) \leq \kappa_u^{\mathrm{est}} + \epsilon}_{\text{user-level LTV target}}, \ \underbrace{\rho_q(u, a_t) \leq \kappa_{\mathrm{plat}}}_{\text{platform risk ceiling}}, \ q \in \{1, 2\}, \end{aligned} \tag{2}$$

where the two caps operationalize a single normative requirement: every issued fee rate should respect a user-level profitability target and a platform-wide ceiling. In particular, κ_u^{est} denotes the largest fee rate consistent with serving user u profitably over the relevant LTV horizon with a tolerance margin $\varepsilon_{\mathrm{b}} \geq 0$. The user-level cap κ_u^{est} is latent and must be estimated from a behavioral history of effective length n_u. Over the valid windows $\mathcal{W}_u = \{w \in \{7, 30, 90\} : \mathrm{GMV}_{u,w} > 0\}$, the plug-in estimate is

$$\kappa_u^{\mathrm{est}} = \begin{cases} \mathrm{clip}_{[0, \kappa_{\max}]} \left(\max_{w \in \mathcal{W}_u} \dfrac{\mathrm{LTV}_{u,w}}{\mathrm{GMV}_{u,w}} - \varepsilon_{\mathrm{m}} \right), & \mathcal{W}_u \neq \varnothing, \\ 0, & \mathcal{W}_u = \varnothing. \end{cases} \tag{3}$$

where w is in days, $\mathrm{LTV}_{u,w} = \mathrm{Income}_{u,w} - \mathrm{Cost}_{u,w}$, $\varepsilon_{\mathrm{m}} \geq 0$ is a profitability margin, and $\kappa_{\max} > 0$ a platform-specified upper bound; the zero fallback encodes the absence of evidence for a positive cap. Since only κ_u^{est} is available at training time and its reliability depends on n_u, LIGER enforces the user-level target as a soft reward penalty on κ_u^{est}, while retaining the platform ceiling κ_{plat} as a hard operational constraint.

Remark 1. The per-user constraint in (2) departs fundamentally from the global-budget paradigm considered in exsting works (Zhang et al., 2021; Zhao et al., 2019). A global budget treats subsidy cost as fungible across users: an aggregate cap lets overspending on one user be offset by underspending on another, so individual profitability is never enforced. Our constraint is instead imposed *per user*

and is *dynamic*, i.e., the cap κ_u^{est} is re-estimated from each user's evolving LTV history and thus tracks changes in individual value over time. This distinction carries a direct economic benefit: high-value users, whose caps remain high, keep receiving escalating investment, whereas persistently unprofitable users are cut off decisively as their caps collapse. The policy therefore steers subsidy toward users where it is profitable, realizing personalized investment rather than mere aggregate budget efficiency.

3.2 Extreme User Heterogeneity

We have specified the decision process, the objective, and the per-user LTV constraint. Note that a defining feature of our setting is that these must be learned over a *highly heterogeneous* user population, varying widely in purchase activity, reward scale, prevalence, and LTV health. This heterogeneity is also extremely imbalanced: in our referral platform the low-activity M_0 tier alone accounts for about 90% of the users, whereas the profitable higher-frequency tiers (M_1–M_2 and M_{3+}) together form a small minority. A globally trained critic is thus dominated by the M_0 majority, and learning that ignores this skew tends to overfit the low-activity tier while under-serving the profitable minority.

This heterogeneity motivates LIGER's group-relative design, which represents the population structure at three resolutions: reward statistics are indexed by the pair (M, H) of purchase and health tiers; advantage statistics are indexed by M alone; and balanced sampling uses the coarse stratum

$$C = \chi(M) \in \mathcal{C} = \{M_0,\ M_{1\text{–}2},\ M_{3+}\},$$

where $M_{1\text{–}2}$ merges the two intermediate purchase-frequency tiers. This hierarchy preserves LTV-sensitive reward information while avoiding unstable moment estimates in sparsely populated (M, H) cells.

The same heterogeneity that shapes this group structure also drives two coupled sources of miscalibration. First, value learning under a highly non-uniform group prior must reconcile group-dependent reward scales, which can suppress within-tier action-value differences and, in the extreme, collapse the policy onto a handful of dominant actions, i.e., a failure we term *value-space collapse*. We diagnose it at the population level within a stratum C, by a small number of distinct greedy actions together with a dominant-action probability close to one; a policy may thus be nearly deterministic at each state while still assigning distinct actions across users. Second, the estimation error $\kappa_u^{\text{est}} - \kappa_u^{\star}$ is group dependent, since low-activity users provide shorter purchase histories, and the resulting systematic underestimation can let the safety penalty dominate informative behavioral variation.

Addressing these two coupled failures calls for a learning procedure that (i) calibrates value estimates within each group so that no single tier's reward scale dominates, (ii) keeps the empirical majority-group prior from steering updates toward the low-activity tier, and (iii) treats each user's LTV cap as an uncertain, estimated constraint rather than a known bound. Therefore, we design a LIGER algorithm to meet these requirements in the following.

4 Offline Policy Learning: LIGER

In this section, we present LIGER, a new algorithm for learning a policy from logged transitions in a heterogeneous user population. In particular, LIGER addresses group-dependent value scales and history-dependent constraint errors. Its overall idea follows the *implicit* Q-learning paradigm: the policy is trained entirely from logged transitions without ever querying actions outside the data support, thereby avoiding the value overestimation that out-of-distribution (OOD) actions induce in offline RL. Building on this foundation, it combines group-calibrated Q-learning and policy extraction with an LTV-aware reward and a calibrated safety reference. Figure 1 summarizes the architecture, and Algorithm 1 provides the training procedure.

4.1 Group-Calibrated Bellman Learning

For a scalar sample x_i belonging to group G_i, define the empirical standardization operator

$$\mathcal{Z}_{G_i}(x_i) = \frac{x_i - \mu_{G_i}(x)}{\max\{\sigma_{G_i}(x), \varepsilon_{\text{z}}\}}, \tag{4}$$

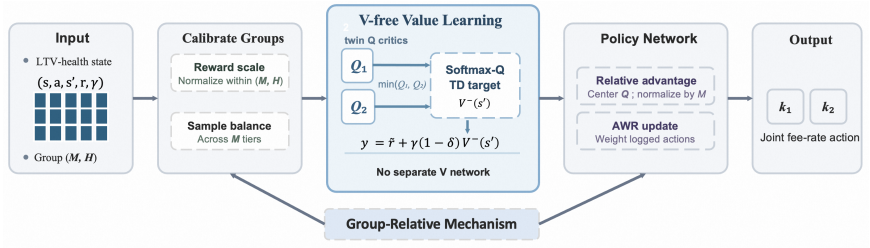


Figure 1: LIGER architecture. Fine-grained (M, H) statistics calibrate rewards, coarse strata determine balanced sampling, and M-level statistics calibrate policy advantages. A tier-conditioned encoder feeds twin Q-heads and a policy head. The critics use a value-network-free softmax-Q target, and the policy is fitted to logged actions by group-relative advantage-weighted regression.

where $\mu_G(x)$ and $\sigma_G(x)$ are the empirical mean and standard deviation within G, and $\varepsilon_z > 0$ ensures numerical stability. Reward calibration uses $G_i^r = (M_i, H_i)$: $\widetilde{r}_i = \mathcal{Z}_{G_i^r}(r_i)$. Cells with fewer than $n_{\min}$ transitions use statistics from their purchase tier M_i. Mini-batches are sampled uniformly over $C_i = \chi(M_i) \in \mathcal{C}$, with replacement when necessary. These operations limit reward-scale and sample-count dominance before value learning.

LIGER uses twin critics Q_{θ_1} and Q_{θ_2} with exponential-moving-average target parameters $\bar{\theta}_1$ and $\bar{\theta}_2$, but no separately parameterized state-value network. Define the clipped target critic as $Q^-_{\min}(s,a) = \min_{\ell\in\{1,2\}} Q_{\bar{\theta}_\ell}(s,a)$. Because the action space is finite, the continuation value is computed directly from this critic:

$$\omega^-(b \mid s) = \frac{\exp\big(Q^-_{\min}(s,b)/\tau_Q\big)}{\sum_{c\in\mathcal{A}} \exp\big(Q^-_{\min}(s,c)/\tau_Q\big)}, V^-(s) = \sum_{b\in\mathcal{A}} \omega^-(b \mid s) Q^-_{\min}(s,b), \\ y_i = \widetilde{r}_i + \gamma(1-\delta_i) V^-(s_i^+). \tag{5}$$

Here b and c are dummy action variables, and $\tau_Q > 0$ is the softmax temperature. For a balanced mini-batch $\mathcal{B}$, the critic loss is

$$\mathcal{L}_Q = \frac{1}{|\mathcal{B}|} \sum_{i\in\mathcal{B}} \left[\sum_{\ell=1}^{2} \big(Q_{\theta_\ell}(s_i,a_i) - y_i\big)^2 \right]. \tag{6}$$

As $\tau_Q \to 0$, V^- approaches the maximum target Q-value; as $\tau_Q \to \infty$, it approaches their arithmetic mean. For intermediate temperatures, V^- is a smooth, parameter-free functional of the target critics and replaces the separately fitted value network used by IQL (Kostrikov et al., 2022).

The minimum over twin critics mitigates overestimation but does not guarantee action support. The policy is therefore initialized by behavior cloning, and the AWR objective is evaluated only on logged actions. The encoder receives $s \oplus \mathrm{onehot}(M)$ before branching into the two Q-heads and the policy head.

4.2 Group-Relative Policy Extraction

Policy extraction uses the state-wise, action-centered score

$$A_Q(s,a) = Q_{\min}(s,a) - \frac{1}{|\mathcal{A}|} \sum_{b\in\mathcal{A}} Q_{\min}(s,b), \; Q_{\min}(s,a) = \min_{\ell\in\{1,2\}} Q_{\theta_\ell}(s,a). \tag{7}$$

The centering term removes state-dependent offsets without changing action rankings. Logged-action scores are then standardized within purchase tier M_i:

$$\widetilde{A}_i = \mathcal{Z}_{M_i}(A_Q(s_i,a_i)). \tag{8}$$

The policy is updated by entropy-regularized advantage-weighted regression:

$$\mathcal{L}_\pi = -\frac{1}{|\mathcal{B}|} \sum_{i\in\mathcal{B}} \Big[w_i \log \pi_\phi(a_i \mid s_i) + \alpha\, \mathcal{H}(\pi_\phi(\cdot \mid s_i)) \Big]. \tag{9}$$

where $w_i = \exp(\widetilde{A}_i/\beta)$, $\beta > 0$ is the advantage temperature and $\alpha \geq 0$ is the entropy coefficient. This construction adapts the group-relative principle of GRPO (Shao et al., 2024) to value-based offline learning. GRPO groups responses sampled for one prompt; LIGER groups persistent user strata and must calibrate the reward, sampling distribution, and learned advantage.

What normalization removes. For any group-specific $\xi_G > 0$ and ζ_G, standardization without an active numerical floor satisfies

$$\mathcal{Z}_G(\xi_G A + \zeta_G) = \mathcal{Z}_G(A). \tag{10}$$

The equality holds when $\sigma_G(A) > \varepsilon_{\mathrm{z}}$ and $\xi_G \sigma_G(A) > \varepsilon_{\mathrm{z}}$; otherwise, the floor intentionally trades exact invariance for numerical stability. Subject to this condition, the policy weights are invariant to group-wise affine transformations of the advantage scale. This property does not remove action-dependent Q-errors, which can still change action rankings. Relative to an IQL advantage $Q_\theta - V_\psi$, Eq. (7) introduces no error from a separately fitted value network. Estimation error in the group moments remains, motivating purchase-tier normalization and the small-cell fallback.

4.3 LTV-Aware Structured Reward

Binary engagement rewards do not distinguish profitable redemption from costly activation. We instead organize the reward around investment return and change in LTV health:

$$R = R_{\mathrm{invest}} + R_{\mathrm{heal}} + R_{\mathrm{aux}} - P_{\mathrm{breach}}, \tag{11}$$

where R_{aux} collects fee-rate saving, exploration and recovery, dense matching, value-conditioned investment, and non-redemption terms. Crucially, every component is regulated by the user's LTV signal: the continuous LTV margin $\bar{\ell}_u$ scales the investment return and the change in LTV-health score drives R_{heal}, so that long-term LTV, rather than raw redemption, dominates the reward. The full component definitions and all coefficients, including the two terms R_{invest} and R_{heal} that define the sequential LTV objective—are provided in Appendix E.

4.4 Calibrating an Uncertain Safety Reference

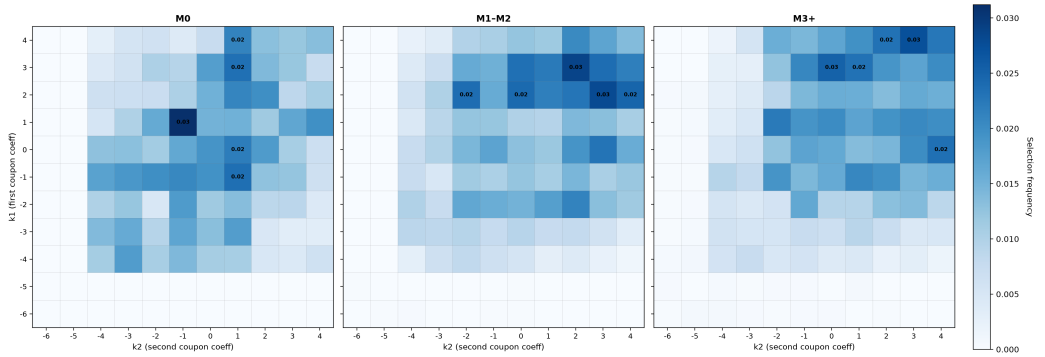
The estimate κ_u^{est} enters learning through the breach penalty P_{breach}. For the low-activity majority it can be systematically underestimated, so the penalty stays active over most of the action range and suppresses state-dependent behavior largely independently of the reward signal, producing the safety–feature paradox that motivates a calibrated penalty reference.

The cap is under-estimated, not tight. The correction is warranted only if low estimates for the majority tier arise from data scarcity rather than genuine risk. To isolate this effect, the user is held fixed while the available history is varied. For users with long records, let κ_u^{full} be the estimate computed from the complete history and $\kappa_{u,n}^{\mathrm{est}}$ the estimate computed from its first n days. The truncation bias is $\mathbb{E}[\kappa_{u,n}^{\mathrm{est}} - \kappa_u^{\mathrm{full}}]$. A systematic negative bias for small n would motivate calibration of the penalty reference.

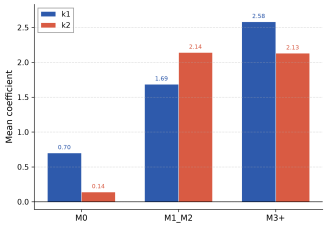
The correction therefore acts on the *penalty reference*, not on the operational guardrail:$\kappa_u^{\mathrm{ref}} = \max\{\kappa_u^{\mathrm{est}}, \kappa_{\mathrm{floor}}\}$. We set κ_{floor} close to the structural rate of the lowest-threshold tier, and use $P_{\mathrm{breach}} = \lambda_{\mathrm{breach}} \sum_{q=1}^{2} \big[\rho_q(T) - \kappa_u^{\mathrm{ref}} - \varepsilon_{\mathrm{b}}\big]_+$, where $[x]_+ = \max\{x, 0\}$. Together with a reduced breach weight $\lambda_{\mathrm{breach}}$, this calibration brings the majority-tier penalty onto the scale of the behavioral terms, so that the safety term no longer overwhelms them. The cart-intent response in the mean adjustment $\bar{k}(a) = \frac{1}{2}(k_1 + k_2)$ correspondingly reverses sign, turning a suppressed adjustment into a positive one. KAPPA-FLOOR deliberately relaxes a *soft penalty induced by an uncertain estimate* and it does not alter κ_{plat}.

5 Experiments

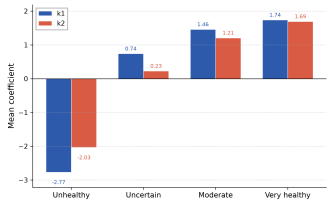
We evaluate LIGER along two complementary axes: an offline study on logged data and an online A/B test on the production platform. For the offline case, we analyze how LIGER conditions coupon actions on user state, isolate the contribution of its group-relative modules through ablation, and compare it with behavioral cloning (BC) and standard IQL using model-based fitted-Q evaluation



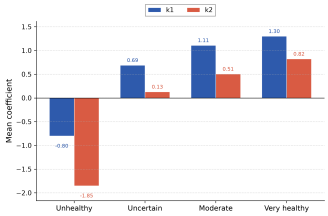
(a) Policy-evaluation action-selection frequency by M-level.



(b) LIGER mean coefficients by M-level.



(c) LIGER mean coefficients by health tier.



(d) LIGER health tiers within $M0$.

Figure 2: Policy-evaluation action frequencies and state-conditioned LIGER policy actions on the held-out test split. Panel (a) reports policy-evaluation frequencies over the 11-by-11 (k_1, k_2) grid for each M-level. Panels (b)–(d) report mean coefficients of greedy LIGER actions obtained during policy evaluation.

(FQE) and action-distribution measures under a shared protocol. For the online case, we deploy LIGER against the incumbent tier-based rule in a month-long A/B test to assess realized business outcomes such as GMV, purchase conversion, and coupon expenditure.

Experimental setup. We evaluate on user-day transitions from a continuous 27-day production stream, split by user into training, validation, and test sets (2.49M/0.64M/0.77M transitions). Each state has 33 features, and the action is a pair of coupon fee-rate coefficients (k_1, k_2) with 121 discrete combinations; results are reported over three purchase-frequency strata ($M0$, $M1$–$M2$, $M3+$). We compare LIGER against behavioral cloning (BC) and standard IQL under matched data, network, reward, and training budget. Offline, we use model-based fitted-Q evaluation (FQE) for counterfactual value differences (GMV, cost, and 7-/30-day LTV) relative to a fixed reference, together with action-distribution descriptors (policy entropy, top-action share, and the number of distinct greedy actions). Full data construction, training protocol, hyperparameters, and evaluation details are provided in Appendix B.

5.1 Offline Analysis

Figure 2 summarizes state-conditioned action selection on the held-out test split. The evaluated policy covers over 100 of the 121 actions in every stratum with no single dominant action (top-action share below 3.1%), yet concentrates its probability mass on a deliberate, stratum-specific region of the (k_1, k_2) grid rather than spreading uniformly. This mass migrates toward stronger coupon combinations as user value increases: mean greedy coefficients rise from $(0.70, 0.14)$ for $M0$ to $(1.69, 2.14)$ for $M1$–$M2$ and $(2.58, 2.13)$ for $M3+$, and LTV health forms a second axis of differentiation that remains discriminative even within $M0$. This joint, monotone, state-conditioned allocation is the opposite of value-space collapse and realizes the per-user LTV reallocation that the global-budget formulation of Section 1 cannot express. Full per-stratum and per-health statistics are reported in Appendix C.

Table 2: Model-based FQE counterfactual value differences on held-out test initial states, relative to a fixed reference policy.

Method	GMV difference (Yuan) ↑	Cost difference (Yuan) ↓	LTV7 difference ↑	LTV30 difference ↑
BC	$+0.23$	-3.93	$+2.27$	$+3.08$
Standard IQL	$+0.42$	-4.14	$+4.24$	$+4.06$
LIGER	$\mathbf{+0.69}$	$\mathbf{-4.54}$	$\mathbf{+5.16}$	$\mathbf{+5.58}$

Table 3: Relative changes in online metrics for LIGER versus the deployed rule in the online experiment. Subgroup comparisons are descriptive.

Population	Metric	Relative change
Overall	GMV (Yuan)	$+3.90\%$
	Visit-to-purchase rate	$+1.40\%$
	Coupon expenditure	-0.12%
Moderately healthy	GMV	$+5.51\%$
	Visit-to-purchase rate	$+2.07\%$
Unhealthy, Q1+	GMV	$+53.51\%$
	Visit-to-purchase rate	$+18.11\%$
	Redeemed-coupon fee rate	-13.28%

5.1.1 Offline Value Comparison

Table 2 reports model-based FQE value differences on held-out test initial states relative to a fixed reference policy. LIGER attains the largest predicted GMV gain ($+0.69$), the largest 7- and 30-day LTV improvements ($+5.16$ and $+5.58$), and the lowest predicted cost (-4.54). These gains compound across method classes: replacing BC with standard IQL enlarges the GMV gain by roughly 80%, and replacing IQL with LIGER adds a further 60%, so LIGER is estimated to generate more value while spending less, matching the per-user LTV objective of Section 1. Complementary behavioral diagnostics confirm that this improvement is not driven by reliance on a single action: ablating any one group-relative module improves one action-distribution metric while degrading another, and only the full method jointly attains the highest entropy, the lowest top-action share, and the broadest action coverage. The complete component-ablation and cross-method action-distribution tables are provided in Appendix C.

5.2 Online A/B Evaluation

We ran a month-long online A/B test against the deployed tier-based rule, with approximately $300{,}000$ users per arm. Relative to the rule, LIGER increased GMV by 3.9% and the visit-to-purchase rate by 1.4% while aggregate coupon expenditure stayed essentially flat (-0.12%), so the additional purchasing activity was not driven by higher overall coupon spending. The effects were strongest for less-healthy users: in the unhealthy Q1+ segment (users who purchased more than one item in a quarter), GMV and the visit-to-purchase rate rose by 53.51% and 18.11% while the redeemed-coupon fee rate fell by 13.28% (Table 3; subgroup rows are descriptive and should be interpreted alongside their baseline levels and sample sizes). These realized business outcomes corroborate the offline FQE analysis of the structured reward.

6 Conclusion

We presented LIGER algorithm for personalized coupon allocation under user heterogeneity and estimated per-user LTV constraints. Our LIGER replaces the separate state-value network with a softmax estimate over twin Q-critics and calibrates rewards, sampling, and advantages within user groups, complemented by an LTV-aware reward and sparse-history-aware safety calibration. The resulting policy yields state-conditioned coupon actions that intensify with user value and LTV health, serving the per-user LTV objective, compared to global-budget formulations. Offline evaluation and an online A/B test confirm improved value and cost outcomes.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning (ICML)*, pp. 22–31, 2017.

Arash Ahmadian, Chris Cremer, Matthias Gallé, et al. Back to basics: Revisiting REINFORCE style optimization for learning from human feedback in LLMs. *arXiv preprint arXiv:2402.14740*, 2024.

Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Eitan Altman. *Constrained Markov Decision Processes*. Chapman and Hall/CRC, 1999.

Han Cai, Kan Ren, Weinan Zhang, Kleanthis Malber, Jun Wang, Yong Yu, and Tie-Yan Liu. Real-time bidding by reinforcement learning in display advertising. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 661–670, 2017.

Tianchi Cai, Jiyan Bao, Jiawei Xu, Jinjie Zhou, Xiaohu Jiang, Defu Lian, Enhong Chen, Weiqiang Guo, and Longfei Wang. Marketing budget allocation with offline constrained deep reinforcement learning. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 432–440, 2023.

Yinlam Chow, Ofir Nachum, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. A Lyapunov-based approach to safe reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Gal Dalal, Krishnamurthy Dvijotham, Matej Vecerik, Todd Hester, Cosmin Paduraru, and Yuval Tassa. Safe exploration in continuous action spaces. *arXiv preprint arXiv:1801.08757*, 2018.

Dongsheng Ding, Kaiqing Zhang, Tamer Basar, and Mihailo Jovanovic. Natural policy gradient primal-dual method for constrained Markov decision processes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Peter S. Fader, Bruce G. S. Hardie, and Ka Lok Lee. "Counting Your Customers" the easy way: An alternative to the Pareto/NBD model. *Marketing Science*, 24(2):275–284, 2005.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning (ICML)*, 2019.

Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning (ICML)*, 2017.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft Actor-Critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning (ICML)*, 2018.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit Q-Learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-Learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Jongmin Lee, Cosmin Paduraru, Daniel J Mankowitz, Nicolas Heess, Doina Precup, Kee-Eung Kim, and Arthur Guez. COptiDICE: Offline constrained reinforcement learning via stationary distribution correction estimation. In *International Conference on Learning Representations (ICLR)*, 2022.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Ashvin Nair, Abhishek Gupta, Murtaza Dalal, and Sergey Levine. AWAC: Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359*, 2020.

Santiago Paternain, Luiz F O Chamon, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained reinforcement learning has zero duality gap. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

Krzysztof Rudaś and Szymon Jaroszewicz. Linear regression for uplift modeling. *Data Mining and Knowledge Discovery*, 32(5):1275–1305, 2018.

Piotr Rzepakowski and Szymon Jaroszewicz. Decision trees for uplift modeling. In *IEEE International Conference on Data Mining (ICDM)*, pp. 441–450, 2010.

Zhihong Shao, Peiyi Wang, Qihao Zhu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Chen Tessler, Daniel J Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *International Conference on Learning Representations (ICLR)*, 2019.

Shuai Xiao, Le Guo, Zaifan Jiang, Lei Lv, Yuanbo Chen, Jun Zhu, and Shuang Yang. Model-based constrained MDP for budget allocation in sequential incentive marketing. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 971–980, 2019.

Haoran Xu, Xianyuan Zhan, and Xiangyu Zhu. Constraints penalized Q-Learning for safe offline reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

Yang Zhang, Bo Tang, Qingyu Yang, Dou An, Hongyin Tang, Chenyang Xi, Xueying Li, and Feiyu Xiong. BCORLE(λ): An offline reinforcement learning and evaluation framework for coupons allocation in e-commerce market. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 20410–20422, 2021.

Kui Zhao, Junhao Hua, Ling Yan, Qi Zhang, Huan Xu, and Cheng Yang. A unified framework for marketing budget allocation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1820–1830, 2019.