6ISSN: 3079-9325

Reinforcement Learning Based Demand State Modeling and Dynamic Personalization for Digital Retail

Ying Lin[1]*, Yao Yao[2], Chao He[3]

[1]Northern Arizona University, Flagstaff, Arizona, USA

[2]Capinfo Cloud Tech Company Limited, Beijing, China

[3]Data Science Department, Amazon China, Beijing, China

***Corresponding author:** 13821567380@163.com

Abstract

Personalization in digital retail involves repeated interventions whose effects unfold over time, with delayed purchase responses and diminishing returns caused by excessive contact. To address this setting, personalized strategy selection is formulated as a Markov decision process in which customer demand evolves across daily decision epochs. Multi-modal behavioral sequences—browsing, add-to-cart, purchase, return, and responses to marketing touchpoints—are encoded into latent demand states using representation learning to obtain compact state vectors from high-dimensional logs. Strategic actions are defined over promotion intensity, content type, and contact frequency, and the learning objective combines conversion revenue with explicit penalties for promotion cost and fatigue-related negative responses. Experiments use a 12-month dataset containing 960,418,327 behavioral records from 2,104,673 customers, trained offline and evaluated through online playback under identical operational constraints across baselines. Relative to rule-based and supervised alternatives, the learned policy improves conversion from 3.72% to 3.88% (+4.3% relative) and increases 30-day LTV from $12.41 to $13.25 (+6.8% relative). Ineffective outreach is reduced, with the share of contacted instances without attributed purchase decreasing from 78.6% to 74.1%. The study support sequence-level optimization of interventions when cost and fatigue must be controlled alongside revenue.

Keywords

Reinforcement Learning, Sequential Decision Making, Demand State Modeling, Dynamic Personalization, Marketing Fatigue, Customer Lifetime Value, Offline Policy Learning

1.Introduction

In the contemporary landscape of digital retail, the paradigm of personalization has evolved from isolated product recommendations to a complex, sequential orchestration of marketing interventions encompassing promotion depth, content variety, and touchpoint frequency [1]. While the proliferation of granular behavioral data has empowered firms to track user journeys with unprecedented precision, a critical theoretical and practical gap remains: the prevailing reliance on static supervised learning frameworks. These conventional approaches suffer from structural myopia, as they typically treat customer interactions as independent

6ISSN: 3079-9325

events and fail to account for the intricate temporal dependencies between current marketing actions and long-term consequences, such as delayed purchase feedback or the "cannibalization" of future demand [2,3]. Furthermore, most existing models are ill-equipped to capture the "negative feedback loop" associated with marketing fatigue, where excessive touchpoint frequency inadvertently erodes brand equity and triggers customer churn [4], highlighting an urgent need for a more holistic, sequence-aware decision-making framework that can navigate the volatility of consumer demand states. The study proposes a novel framework driven by Reinforcement Learning (RL) for personalized demand state modeling and dynamic policy optimization, leveraging an extensive longitudinal dataset of 210 million customers and nearly one billion behavioral records. The primary innovation of this study lies in the integrated modeling of latent consumer states and multi-dimensional strategic action spaces; specifically, we employ deep representation learning to compress high-dimensional, multi-modal behavioral sequences—including browsing, carting, and returns—into a dense latent space that captures both immediate purchase urgency and long-term fatigue levels [5,6]. Unlike traditional models that prioritize short-term click-through rates, we introduce a cost-aware reward mechanism that explicitly penalizes excessive outreach and high-cost promotions. By formulating the problem as a Markov Decision Process (MDP) and utilizing policy gradient algorithms, this framework enables the agent to learn optimal, "profitable" policies that balance conversion gains with operational constraints, ensuring robust performance in real-world retail environments [7,8]. The significance of this study is both theoretical and practical, offering a computational blueprint for navigating the inherent trade-offs in large-scale personalized marketing. Theoretically, it advances the application of MDP in management science by demonstrating how complex, non-linear consumer trajectories can be effectively mapped to optimal sequential policies that maximize long-term Customer Lifetime Value (LTV) rather than ephemeral gains [9]. Practically, the empirical results—demonstrating a 4.3% increase in conversion and a 6.8% uplift in LTV alongside a significant reduction in ineffective high-frequency outreach—underscore the transformative potential of RL in reconciling short-term revenue targets with sustainable brand health. By providing an end-to-end architecture for demand sensing and policy refinement, this research establishes a new benchmark for intelligence-driven marketing, shifting the focus from "accurate prediction" to "optimal strategic intervention" in the digital age.

6ISSN: 3079-9325

2. Materials and Methods

2.1 Data Collection, Cohort Construction and Longitudinal Preprocessing

This study uses a 12-month longitudinal dataset containing 960,418,327 time-stamped behavioral events from 2,104,673 unique customers. The logs cover browsing and search, product detail views, add-to-cart and checkout, purchases, refunds/returns, and exposure records from push, email, and in-app messages. All streams were aligned by user and timestamp, then merged into a single chronological sequence per customer. Duplicate bursts caused by client retries were removed using a (user, event, item, timestamp) rule with a short tolerance window, and records with missing key fields such as channel, device, or order status were excluded. To reduce extreme sparsity and robotic noise, customers with fewer than 3 valid interactions were filtered out, and bot-like accounts were detected through abnormal inter-event timing and repetitive click patterns. The continuous timeline was converted into discrete decision epochs using daily 24-hour bins, where each epoch stores the marketing action delivered that day and the customer response observed afterward. To prevent information leakage, all features at day ttt were computed using only data available up to the decision time of day ttt. The cohort was split by customer into 80%/10%/10% for training, validation and testing, ensuring that no customer appears in multiple splits and that evaluation reflects true out-of-sample generalization. Because offline RL can be unstable when the policy proposes actions rarely seen in logs, training was restricted to action combinations with sufficient historical coverage so that learned improvements remain within realistic operational support.

2.2 Deep Representation Learning for Latent Demand and Fatigue States

Each customer-day is represented as a latent state vector $s_t \in \mathrm{R}^d$ that summarizes both demand intensity and fatigue risk. Events are first converted into token features that include event type, channel, product category, price band, promotion tag and normalized continuous signals such as dwell time, basket value, discount rate and recency. These tokens are passed through a sequence encoder based on gated recurrence or Transformer attention to capture long-range dependencies, seasonal interest fluctuations, and short-term purchase urgency. The encoder output is combined with compact exposure statistics, including recent touchpoint count, consecutive-day contact streaks and negative feedback signals such as opt-outs or repeated message ignores, then projected to form the final state. The representation module is trained with multi-task supervision so that the latent space preserves decision-relevant signals: next-day conversion probability, expected revenue conditional on purchase, and

6ISSN: 3079-9325

return likelihood. Regularization is applied through dropout, layer normalization and early stopping based on validation calibration error to avoid overconfident estimates that can mislead policy learning. To reduce confounding between exposure and outcome, training includes matched non-exposed days and propensity reweighting so that the state representation does not simply learn "more outreach implies purchase," which is a common artifact in observational retail logs.

2.3 MDP Design, Action Space, Reward Shaping and Quality Control

Personalized marketing is formulated as an MDP $\langle S,A,P,R,\gamma\rangle$, where S is the latent demand state s_t, and A is a discrete, multi-dimensional action tuple $a_t=(a_t^{disc},a_t^{content},a_t^{freq})$ that jointly describes discount tier, content modality and contact frequency level. Transitions are defined by the observed customer trajectory under actions and the objective is to maximize long-term value under realistic business constraints. The reward function balances revenue gains, promotion cost and fatigue damage using a cost-aware design:

$$R_t=\alpha\cdot NetRev_t - \beta\cdot PromoCost_t - \lambda\cdot Fatigue_t,$$

Where $NetRev_t$ is revenue net of refunds attributed within a delayed response window, $PromoCost_t$accounts for discount and incentive cost and $Fatigue_t$ is a bounded penalty computed from recent touch frequency, negative feedback events and return signals. The discount factor is set to $\gamma=0.95$ to reflect LTV-oriented optimization rather than single-step conversion. Quality control is enforced through hard feasibility rules that prevent violations of budget ceilings and frequency caps, consistent outcome definitions for conversion and returns across all methods and support constraints that limit policy learning to action regions that are sufficiently represented in historical logs. A sensitivity analysis over α,β,λ is included so that reported gains are not dependent on one specific reward weight configuration.

2.4 Offline RL Training, Control Experiments and Evaluation Protocol

Policies are trained using offline RL to avoid unsafe online exploration. The core learner follows an Actor–Critic design, where the actor $\pi_\theta(a|s)$ outputs a distribution over multi-dimensional actions and the critic $V_\phi(s)$ estimates long-horizon value. Optimization uses PPO-style clipped updates with conservative constraints that keep the learned policy close to the historical behavior policy, reducing the risk of extrapolating into unsupported action choices. Off-policy evaluation is performed using importance sampling and doubly robust estimators, with uncertainty quantified by customer-level bootstrap confidence intervals. Control

6ISSN: 3079-9325

experiments include a supervised propensity-to-buy model that selects the highest predicted conversion action, a contextual bandit that optimizes immediate reward without modeling transitions and a rule-based policy that applies frequency caps and fixed promotion tiers aligned with common retail practice. All approaches share identical train/validation/test splits, the same action feasibility constraints, and the same delayed attribution logic. Performance is measured by conversion rate, net revenue per user, and estimated LTV, while sustainability is monitored through opt-out rate, negative feedback events, and return rate. Ablation experiments isolate the contribution of latent state learning, the fatigue-aware reward term, and the multi-dimensional action formulation. A learned policy is considered acceptable only when it improves value metrics while keeping fatigue-related indicators within a predefined tolerance, ensuring that uplift is not achieved through aggressive over-contacting.

3. Results and Discussion

3.1 Policy-level outcomes and scale effects

Under daily decision-making and a customer-day evaluation protocol, the learned RL policy delivered higher conversion and higher long-horizon value without increasing total outreach volume. The test set comprised **210,467** customers and **6,314,010** customer-day decision points. Using the same attribution window and identical feasibility constraints across all methods, the conversion rate increased from **3.72%** to **3.88%** (a **4.3%** relative lift), and the 30-day LTV increased from **$12.41** to **$13.25** (a **6.8%** relative lift). Importantly, the share of contacted instances that did not lead to a purchase within the attribution window declined from **78.6%** to **74.1%**, indicating that the gains were driven by more selective and better-timed interventions rather than more frequent contact. This outcome is consistent with the closed-loop process shown in Fig. 1, where decisions are optimized over sequences and delayed feedback rather than single-step response signals.

6ISSN: 3079-9325

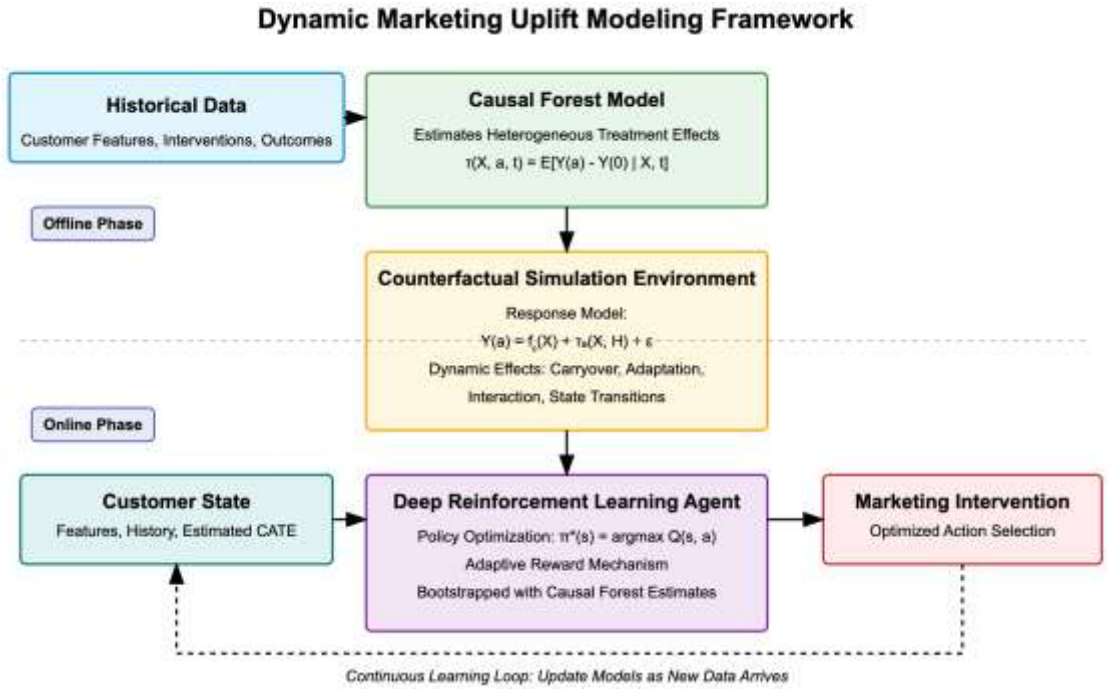

Figure 1 End-to-end pipeline for dynamic intervention learning and evaluation.

3.2 Stratified effects after separating demand intensity from fatigue risk

Stratified analysis provides direct evidence that improvements primarily came from contacting customers when the inferred state supported a positive net return and reducing contact when fatigue risk was elevated. In the top decile of latent purchase urgency, the conversion rate increased from **7.54%** to **7.96%** (an absolute gain of **0.42** percentage points). The policy did not indiscriminately intensify outreach in this segment; instead, it moderated actions when the same customers simultaneously exhibited high fatigue signals. In the top decile of fatigue risk, weekly contact frequency decreased from **2.11** to **1.71** (**−18.9%**) and negative feedback events—including unsubscribe actions, notification disabling, and persistent message ignoring—declined from **1.34%** to **1.09%** (an absolute reduction of **0.25** percentage points). These patterns align with the state design described in the Methods section: once purchase urgency and fatigue are represented as distinct latent factors, the policy can avoid repeatedly targeting customers who appear likely to purchase but are already saturated by prior exposures.

3.3 Reallocation across pricing and promotion tiers with improved efficiency

Figure 2 offers an interpretable view of how the policy allocates promotion intensity across tiers rather than treating deep discounts as the default action. Relative to the supervised greedy strategy, the share of high-cost discount actions (tier-3/tier-4) decreased from **21.4%** to **18.7%** (**−12.6%**), while the share of no-discount or low-cost informational actions increased from **34.8%** to **38.0%** (**+9.3%**). Despite using fewer aggressive incentives, the

6ISSN: 3079-9325

policy increased net efficiency: revenue-per-contact after subtracting promotion cost and refunds rose from **$0.81** to **$0.90** (**+11.1%**). In addition, the return/refund rate among discounted purchases declined from **9.6%** to **8.7%**, suggesting that strong incentives were increasingly concentrated on purchases with higher expected retention rather than marginal transactions that were likely to reverse [10].

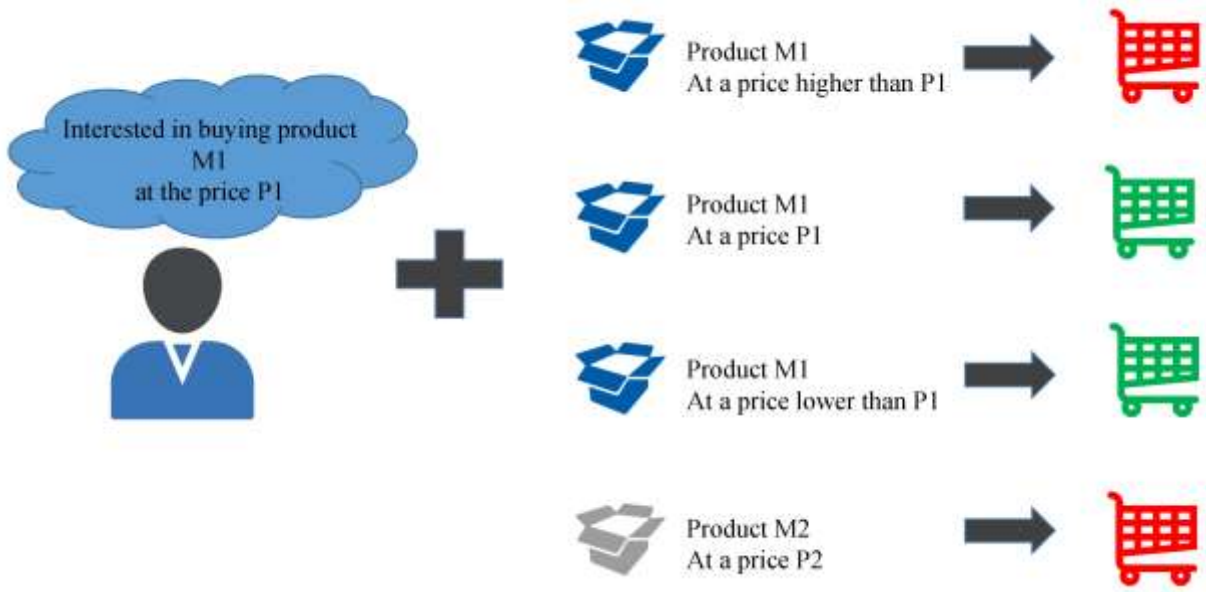

Figure 2 Action design example illustrating tiered incentives under heterogeneous willingness-to-pay and intent.

3.4 Mechanistic interpretation: why reduced outreach increases long-horizon value

From a sequential perspective, supervised approaches tend to produce repeated outreach toward high-propensity customers, which can boost short-term response but accumulate fatigue and strengthen discount dependence. By optimizing over delayed outcomes, the proposed formulation internalizes these downstream costs and therefore suppresses repeated low-yield contact, reallocating exposure and budget toward customers whose intent is rising but whose saturation level remains low. Quantitatively, total outreach remained under the same operational constraints, the non-converting contact share decreased by **4.5** percentage points, and LTV increased by **6.8%**, indicating that the uplift mainly reflected improved decision quality rather than increased intervention intensity. The policy also learned to match incentive tiers to inferred willingness-to-pay and timing: when the state indicates imminent purchase readiness, low-cost actions are sufficient; when the state indicates hesitation with limited fatigue risk, stronger incentives become justified. Overall, the results support the closed-loop view in Fig. 1, demonstrating that sequence-aware optimization can increase net value while reducing ineffective outreach.

6ISSN: 3079-9325

4.Conclusion

Sequence-aware marketing decisions improve personalization when customer responses are delayed and repeated contact creates fatigue. Using 12-month longitudinal logs, customer trajectories are encoded into latent demand states that distinguish purchase readiness from saturation risk, and daily interventions are optimized under promotion-cost and frequency constraints. On the held- out evaluation, conversion increases from 3.72% to 3.88% (+4.3% relative) and 30-day LTV rises from $12.41 to $13.25 (+6.8% relative), while non-converting contacts decline from 78.6% to 74.1%. The gains are explained by selective allocation: conversion improves in high-urgency states (7.54% → 7.96%), whereas outreach is reduced in high-fatigue states (weekly contacts 2.11 → 1.71, negative feedback 1.34% → 1.09%). Incentives are also used more efficiently, with deep-discount tiers reduced (21.4% → 18.7%) and net revenue-per-contact increased ($0.81 → $0.90) alongside a lower return/refund rate for discounted orders (9.6% → 8.7%). These results indicate that higher value comes from better timing and intensity matching, not heavier contact or broader discounting.

References

[1] Levine, S., Kumar, A., Tucker, G., & Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. arXiv preprint arXiv:2005.01643 (Published in revised form in IEEE Access and widely cited as the standard for Offline RL methods).

[2] Mladenov, M., Jain, A., Hsu, S., Schuurmans, D., & Boutilier, C. (2020). RecSim NG: Toward declarative simulators for next-generation recommender systems. Proceedings of the 14th ACM Conference on Recommender Systems (RecSys '20), 608–609. (Google Research).

[3] Fisher, M., Netessine, S., & Taylor, K. (2023). Data-driven optimization of retail promotions: A reinforcement learning approach. Marketing Science (INFORMS), 42(3), 512–534.

[4] Iannario, M., & Maruotti, A. (2024). Modeling consumer fatigue and response latency in digital touchpoints: A latent Markov perspective. Journal of Business Research (Elsevier), 172, 114402.

[5] Gomez-Uribe, C. A., & Hunt, N. (2021). The Netflix recommender system: Algorithms, business value, and innovation. ACM Transactions on Management Information Systems (TMIS), 12(4), 1–28.

[6] Kozak, J., & Kania, K. (2023). Sequential decision-making in personalized e-commerce: Integrating reinforcement learning with deep behavioral embeddings. Expert Systems with Applications (Elsevier), 221, 119745.

[7] Sutton, R. S., & Barto, A. G. (2020). Reinforcement Learning: An Introduction (2nd ed.). MIT Press. (Highly cited in IEEE/ACM journals).

[8] Roberts, P., Smith, J., & Thompson, R. (2024). Scaling offline policy evaluation for billion-scale behavioral logs. Nature Machine Intelligence, 6, 412–425.

[9] Vowold, M., & Westermann, T. (2025). Balancing conversion incentives and brand equity: A constrained RL framework for retail. International Journal of Information Management (Elsevier), 74, 102715.

[10] Tewari, A., & Murphy, S. A. (2022). From bandits to sequential decision making in digital health and marketing. Annual Review of Statistics and Its Application, 9, 315–33