Modelling Customer Trajectories with Reinforcement Learning for Practical Retail Insights

Ken Ming Lee
McGill University
Montréal, Canada
ken.m.lee@mail.mcgill.ca

Paul Barde
Mila - Quebec AI Institute
Montréal, Canada
paul.b.barde@gmail.com

Maxime C. Cohen
McGill University
Montréal, Canada
maxime.cohen@mcgill.ca

Derek Nowrouzezahrai
McGill University
Montréal, Canada
derek.nowrouzezahrai@mcgill.ca

ABSTRACT

Understanding customer movement within retail spaces is essential for optimizing store layouts. Real-world trajectory data can provide highly accurate insights, but collecting it is costly and often infeasible for many retailers. Heuristics such as Travelling Salesman Problem (TSP) and Probabilistic Nearest Neighbours (PNN) are commonly used as inexpensive approximations, but actual customer trajectories deviate by an average of 28% from shortest paths, highlighting a tradeoff between accuracy and practicality. We propose an agent-based modelling framework that casts customer trajectory prediction as a maximum entropy reinforcement learning (RL) problem, balancing reward maximization with stochasticity to better reflect customers with bounded rationality. Using real-world trajectory data from a convenience store, we show that RL-generated trajectories align more closely with customer behaviour than TSP and PNN, providing more accurate estimates of impulse purchase rates and shelf traffic densities. Furthermore, only RL-based predictions yield repositioning decisions for impulse products that align with those derived from actual trajectory data, resulting in comparable estimated profit gains. Our work demonstrates that RL provides a practical, behaviourally grounded alternative that bridges the gap between oversimplified heuristics and data-intensive approaches, making accurate layout optimization more accessible. To encourage further research, the source code is available on GitHub.

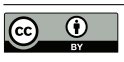
CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

Reinforcement Learning, Application, Retail, IA

ACM Reference Format:
Ken Ming Lee, Paul Barde, Maxime C. Cohen, and Derek Nowrouzezahrai. 2026. Modelling Customer Trajectories with Reinforcement Learning for Practical Retail Insights. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/AJIK9102



Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/AJIK9102

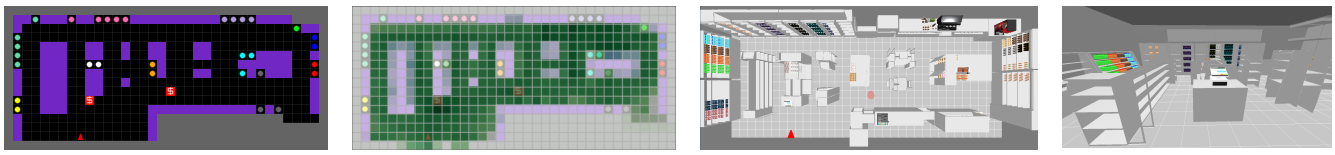
1 INTRODUCTION

Maximizing profit is the central goal of store layout optimization. This is typically achieved by strategically arranging products to maximize impulse purchases, which, unlike planned purchases, are highly influenced by store layout and customer exposure [37]. Therefore, understanding customer movement is paramount in determining the optimal placement for each product.

While customer trajectories can provide these insights, collecting them is costly, making it infeasible for many retailers, particularly smaller ones. As an alternative, heuristics such as the Travelling Salesman Problem (TSP) and Probabilistic Nearest Neighbour (PNN) are commonly used in the literature to approximate customer paths [13, 17, 23, 24, 34]. TSP assumes the global shortest path, while PNN models a stochastic greedy-optimal route. However, prior research shows that customers deviate from shortest paths by an average of 28% [26], highlighting a gap resulting from a tradeoff between these approaches: heuristics provide inexpensive but unrealistic approximations, while actual trajectories offer accuracy but at prohibitive cost.

To address this gap, we propose an agent-based modelling framework in which customers are assumed to follow trajectories derived from maximum entropy reinforcement learning (RL). Maximum entropy RL optimizes for both reward and randomness simultaneously [41], making it a compelling alternative for modelling customers with bounded rationality. Using ground-truth trajectory data collected from a convenience store, we demonstrate that RL-generated paths more closely resemble real-world customer behaviour than those generated by TSP and PNN, as measured by several divergence metrics. These more realistic trajectories translate to more accurate estimates of impulse purchase rates and shelf traffic densities. Furthermore, when using trajectories generated by these methods to inform the repositioning of a single impulse product, only RL-based predictions yield decisions consistent with those derived from actual customer trajectories. Simulating customer trajectories on these revised layouts to estimate profit outcomes further confirms this advantage: only the RL-informed layout achieves a profit increase comparable to that obtained using ground-truth trajectory data.

These findings show that RL-generated trajectories provide a more accurate approximation of customer behaviour than heuristic methods, while serving as a cheaper substitute for actual trajectories. In doing so, RL bridges the gap between oversimplified



(a) RL StoreGrid environment. (b) Overlay of customer trajectories on the RL environment. (c) Top-down view of the digital twin. (d) First-person view of the digital twin.

Figure 1: Various representations of the retail store. (a) Grid-based representation for RL training. (b) Overlay of customer trajectories on the RL environment, illustrating spatial alignment with the physical store. (c-d) Top-down and first-person views of the 3D digital twin constructed from the same discretized representation.

heuristics and data-intensive approaches, making accurate layout optimization more accessible.

The supplementary material, the link to the source code, and a playable demo of the digital twin are publicly available at https://sites.google.com/view/storegrid.

2 BACKGROUND INFORMATION

The facility layout problem studies the optimal arrangement of physical spaces (e.g., workstations, shelves) to achieve operational goals. [2]. In manufacturing, a typical objective is to minimize material handling costs by optimizing the flow of goods between production units [14]. In contrast, the retail setting instead prioritize profitability and customer experience [6], making behavioural modelling essential. Figure 1 illustrates the retail store layout used throughout this paper.

2.1 Retail Layout Design Subproblems

Retail layout design is a hierarchical process spanning aisle configuration (e.g., grid, freeform, racetrack layouts)[31], zoning (i.e., assigning departments to functional zones) [11, 13], shelf-space allocation (i.e., how much shelf space to allocate to each product or category) [16, 18], and product placement (i.e., optimal physical arrangement of items) [9, 10, 15].

This paper is primarily concerned with the product-placement subproblem. Accordingly, all subsequent references to "store layout optimization" refer specifically to this component of the broader retail layout design process.

2.2 Impulse Purchases

In the retail literature, customer purchases are typically categorized into planned and unplanned purchases. Planned purchases refer to items that customers intend to buy before entering the store (e.g., essential products like bread and milk), and are largely unaffected by layout. In contrast, impulse purchases are spontaneous and heavily influenced by in-store stimuli such as product placement, displays, or advertisements[37]. Impulse purchases can account for 30–50% or more of total sales in supermarkets [4, 30, 35, 40], making them the primary source of marginal profit gain from layout optimization.

Hence, much of the literature on product-level layout optimization focuses specifically on maximizing impulse profit — the portion of profit attributable to unplanned purchases influenced by the in-store environment.

2.3 Optimizing Store Layouts to Maximize Impulse Purchases

A central principle in the retail literature is "unseen is unsold" [20, 27], motivating layouts that maximize product exposure. Earlier strategies attempt to increase exposure by distributing essential products throughout the store to lengthen paths [5, 19, 28], but a limitation with this approach is that excessive detours risk reducing shopping convenience and thus overall sales [7, 29].

To address this limitation, subsequent studies have sought to balance between maximizing exposure and customers' convenience. For instance, Ozgormus and Smith [34] alternates between optimizing for revenue and adjacency preferences, while Abdelaziz et al. [1] formulates the problem as a bi-objective constraint optimization problem, where the authors maximizes impulse profit while constraining shopping distance of customers.

2.4 Evaluating Store Layouts

Effectively addressing the product-placement subproblem requires not only a strong optimization algorithm but also an evaluation framework that closely aligns with the ultimate goal of profit maximization. In the literature, evaluation of layouts is typically either integrated directly into the optimization objective — as in constraint optimization approaches (e.g., Flamand et al. [17]) — or performed explicitly as a separate step [12, 13, 24].

Regardless of the approach, most evaluation frameworks rely on predicting customer navigation patterns, where heuristics such as the Travelling Salesman Problem (TSP) and Probabilistic Nearest Neighbour (PNN) dominate [13, 17, 23, 24, 34]. TSP assumes globally shortest routes, while PNN stochastically select the next item based on proximity, assigning probabilities inversely proportional to distance.

Although inexpensive, prior research shows that actual customer trajectories can deviate from shortest-path estimates by an average of 28% [26], highlighting the limitations of such models in capturing realistic customer behaviour. This creates a gap between costly but accurate trajectory data and oversimplified heuristic approximations.

2.5 Maximum Entropy Reinforcement Learning

In Reinforcement Learning (RL), an agent learns a policy, that selects an action given a state, to maximize expected cumulative reward through sequential interaction with an environment [39]. Maximum Entropy (MaxEnt) RL augments this by jointly maximizing

the expected return and the policy entropy [41]. This encourages diverse behaviours and captures multi-modal solutions [21, 22].

In the case of retail, for a given shopping basket, customers may not only purchase products in different orders, but also take different paths for a specific product sequence. Traditional heuristics like TSP and PNN are unable to capture such behavioural variability. Moreover, the space of all possible trajectories through the store grows combinatorially with the number of products to pickup and size of the store, quickly becoming too large for exhaustive enumeration.

In contrast, since maximizing randomness is part of the objective of a MaxEnt RL algorithm, it is naturally incentivized to explore diverse trajectories, including different product-pickup orders or route variations. Additionally, when used in combination with expressive function approximators such as neural networks, MaxEnt RL can scale to explore large trajectory spaces much more effectively than heuristic or exact methods.

3 EXPERIMENTAL SETUP

This section describes the experimental setup, RL formulation of the physical store, preprocessing of real-world trajectory data, and implementation of trajectory modelling methods.

3.1 Modelling a Real-World Convenience Store

We collaborated with a local convenience store outfitted with overhead cameras arrays, and obtained anonymized customer trajectory data collected from September 2023 to February 2024. The dataset contained 3D joint coordinates recorded at 5 Hz and corresponding checkout baskets. From this, we reconstructed customers' 2D in-store positions and associated purchased items. Layout metadata describing boundaries of shelves and checkout locations were obtained from Li [32].

To compare trajectory-prediction methods, we implemented a 2D gridworld environment (Figure 1) based on the Gymnasium Minigrid framework [8]. The environment matched the store's actual dimensions: a 16×36 grid with each cell representing 50×50 cm. This discretization allows the store to be represented as a graph for pathfinding (e.g., Dijkstra's algorithm) and defines a finite action space for RL agents, improving computational tractability.

We focused on the top 61 best-selling products, which account for approximately 51% of sales. Products were grouped into 11 categories: Hot Coffee/Tea, Bakery/Pastries, Hot Food (e.g., Hot Dog, Pizza), Fruits/Yogurt, Energy Drinks, Cold Beverages (e.g., Kombucha, Sparkling), Soft Drinks (e.g., Coke, Pepsi), Snack Bars (e.g., Energy, Granola), Cold Food (e.g., Sandwiches, Wraps, Salad), Cold Coffee/Tea/Shake and Fountain Drinks.

Each shelf holds one category, and both self-checkout and cashier-assisted stations are represented as distinct checkout locations (shown as red cells with dollar signs in Figure 1). For convenience, the term "product" refers to the entire category throughout this paper.

3.2 Preprocessing Trajectory Data

Raw trajectories are preprocessed to align with the gridworld environment as follows. Trajectories without matching basket information are removed, and coordinates outside store boundaries or recorded after checkout are trimmed. Continuous (x, y) positions are then discretized to grid indices, and invalid points (e.g., within shelves or walls) are reassigned to the nearest valid cell. All trajectories are normalized to start at the store entrance and end at one of the two checkout counters. Finally, each trajectory is transformed into a sequence of $(state, action)$ pairs, with pickup actions inferred at the customer's nearest (or latest) approach to a product location. After preprocessing, 3,054 trajectories remained for evaluation purposes.

3.3 RL Environment and Agent Design

We model customers as conditional MaxEnt RL agents, trained using Proximal Policy Optimization (PPO) [38] with a convolutional neural network backbone.

Each episode represents a complete shopping trip: the agent starts at the entrance and ends upon performing a checkout action or reaching a time limit. The agent is conditioned on the target basket (i.e., products to purchase), checkout location (two exist in this case), and optionally, a timestep budget (i.e., target trip length). Conditioning on timestep budget enables control over the length of generated trajectories, allowing us to simulate customers with different shopping preferences.

The agent observes a multi-layer 2D tensor that encodes environmental features. Specifically, the observation includes:

(1) an object-type map (indicating walls, shelves, products etc.);
(2) step count in the current episode;
(3) timestep budget (if desired);
(4) a binary mask marking cells yet to be visited;
(5) category identifiers for each product;
(6) conditional basket specifying the customer's intended purchases; and
(7) the agent's current position and orientation.

The discrete action space comprises four actions: move forward, turn left, turn right, and pickup/checkout.

At checkout, the agent receives a reward based on the accuracy of products picked-up, use of the correct checkout point, and adherence to the specified timestep budget. Rewards scale with the proportion of correctly collected items and closeness to the target trip length.

Additional details, including RL network architecture and hyperparameters settings, are provided in Appendix A.

3.4 Training Techniques

To improve learning stability and promote generalizable policies, several practical techniques are applied:

- Conditional baskets, ranging from 0–5 products (which are broader than the real product range), are periodically resampled to improve generalization across basket sizes and product combinations. Training follows a curriculum schedule, progressing from simpler to more complex baskets.
- Parallel environments with independently sampled basket conditions are used to stabilize learning and enhance generalization.
- Observations are normalized per channel to account for differing value ranges.

- A discount factor of $\gamma = 1.0$ ensures that longer trajectories are not penalized, preserving natural path-length variability and supporting time-conditioned control.
- To further promote exploratory behaviour, a bonus reward is granted after all main objectives have been satisfied, based on the number of unique states visited in an episode.

3.5 Trajectory Generation

We generate TSP and PNN trajectories by treating the RL environment as a graph, with adjacent cells as connected nodes.

For the TSP baseline, the shortest possible route is computed, whereas for the PNN baseline, the next product is selected probabilistically, with nearer products more likely to be chosen. In both cases, the checkout is appended as the final waypoint. Detailed description of their implementations are available in Appendix B.

To generate RL trajectories, the trained RL policy is conditioned on a basket, a checkout location, and an optional timestep budget – all of which are encoded into the agent's observation. We then roll out the policy in the environment and only retain trajectories that exceeds a minimum reward threshold, ensuring the specified conditions are satisfied.

4 EXPERIMENTAL RESULTS

We evaluate our agent-based modelling framework through three stages: (1) comparing how closely different trajectory-generation methods reproduce real human movement, (2) assessing their ability to estimate shelf-level visitation and impulse rates, and (3) applying these estimates to reposition an impulse product for profit.

In this section, we describe these stages in detail, and present results alongside a corresponding discussion of them.

4.1 Performance Comparison of Trajectory Generation Methods

We sample 10k trajectories for each method across all baskets combination present in the ground-truth dataset. For those with fewer than 10k available trajectories (such as TSP and human trajectories), we upsample with replacement to ensure fair comparison.

To evaluate each method's ability to replicate real-world customer behaviour, we aggregate their trajectories for each basket and normalize them into 2D probability distributions. They are compared to the real human trajectory distribution using Jensen-Shannon divergence (JSD) and Wasserstein Distance (WD), where smaller values indicate closer alignment with human behaviour.

From Table 1, it can first be observed that PNN's stochastic sampling of waypoints across all possible product locations allows it to better capture the diversity and multimodality present in real-world customer behaviour, allowing it to outperform TSP in terms of JSD and WD.

However, PNN falls short compared to RL. Figure 2 illustrates this qualitatively: TSP fails to capture the mode where the customer navigates through the bottom half of the store to reach the product, because it always selects the shortest global path, which in this case, leads through the top of the store. Although PNN is stochastic in choosing which shelf to visit next, the shortest path to both shelf locations of the Cold Food category happens to go through the upper half of the store, causing PNN to similarly miss the bottom traversal. In contrast, since the RL agent is trained to maximize not only reward but also policy entropy, it is incentivized to explore and recover diverse yet task-valid trajectories, such as the one through the bottom half of the store.

Table 1: Divergence of simulated trajectories compared to human trajectories. Lower (bolded) scores indicate closer alignment with human behaviours.

Divergence (lower is better)	TSP	PNN	RL
Jensen-Shannon divergence (JSD) of average heatmap	0.657	0.580	**0.415**
Wasserstein Distance (WD) of average heatmap	0.0140	0.0120	**0.00800**
Average JSD	0.777	0.676	**0.476**
Average WD	0.0176	0.0142	**0.00920**

In summary, RL trajectories align more closely with real-world customer behaviour than those generated by either PNN or TSP. This trend is consistent across both JSD and WD, whether computed over average occupancy heatmaps (Table 1) or examined on a per-basket basis (Appendix C).

4.2 Use Case 1: Estimating Traffic Density

To maximize impulse purchases, high-impulse items should be placed in areas with high foot traffic. In contrast, essential items — which are purchased regardless of placement — can be used strategically to guide customers through high-value zones [25]. To take advantage of this, it's critical to quantify how likely each shelf is to be visited during an average customer trip.

For instance, Flamand et al. [17] proposed a profit-maximizing store layout formulated as a constrained optimization problem, with the objective:

$$J_{\text{profit}} = \max_{x} \sum_{b \in \mathcal{B}} \sum_{p \in \mathcal{P}} \rho_p \, i_p \, \frac{s_{pb}}{c_b} \, x_{bp} \, \theta_b \tag{1}$$

where

- b indexes shelves, and p indexes product categories,
- ρ_p denotes the per-unit profit of product category p, which can be expressed as $\rho_p = \text{price}_p \times \text{margin}_p$,
- i_p represents impulse rate of product category p,
- $\frac{s_{pb}}{c_b}$ is the ratio between space taken up by product p and total shelf space c_b; conceptually, this ratio measures visibility of p on shelf b, which equals 1 in our case (one category per shelf),
- x_{bp} is a binary placement variable to optimize for, and
- θ_b is the *shelf traffic density* — the probability that shelf b is visited during a typical store visit.

Intuitively, the shelf traffic density θ_b depends not only on the shelf's physical location in the store, but also on the shelf's content. The closer the shelf is to the entrance/exit, the more likely it is to be visited. Likewise, if the product placed on the shelf is popular, or the shelf is close to another popular product, the shelf is more likely to be visited as well. To predict the traffic density of a shelf in

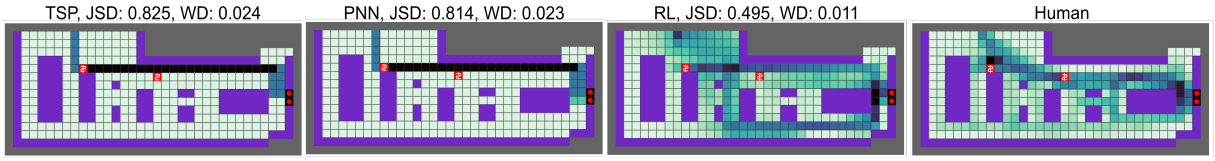
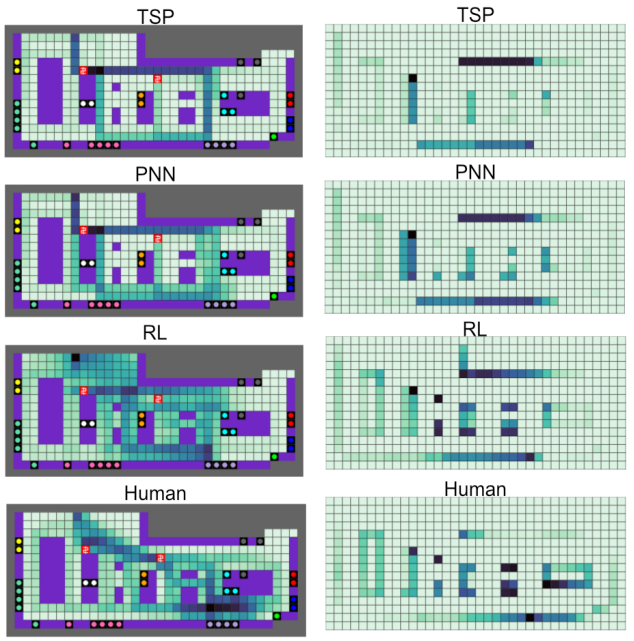


Figure 2: Trajectory heatmap of customers purchasing from the Cold Food category with (9,5) checkout, across all methods. Only RL was able to recover the mode where the customer purchases the product by travelling the longer way from the bottom.



(a) Trajectory heatmap. **(b) Heatmap of shelf-traffic density computed from Figure 3a.**

Figure 3: Heatmap of trajectories and shelf-traffic density generated by TSP, PNN, RL and ground-truth (human) data, for the top 61 most frequently bought baskets.

a new layout, Flamand et al. [17] train a regressor to do so, which takes it's and neighbouring shelves' contents into consideration.

Importantly, since Flamand et al. [17] did not have access to human trajectory data, they instead used simulated TSP trajectories as the ground-truth dataset for training the regressor. Motivated by previous results where RL-generated trajectories more closely resemble real customer movement, we propose deriving shelf traffic density from RL-generated trajectories instead of TSP (and PNN).

To compute shelf traffic density from trajectories, for each trajectory, we iterate through each shelf in the store, and mark a shelf as being visited if the customer gets close to the shelf (where "close" is defined as reaching an adjacent cell). The resulting shelf occupancy map is then normalized by the number of trajectories, yielding a 2D grid where each value represents the average probability that a shelf at that location is visited during a trip (see Figure 3b).

To evaluate the quality of the shelf traffic density prediction by each method, we compute their JSD and WD to that of ground truth human trajectories. From Table 2, it can be observed that RL yields lower JSD and WD than TSP and PNN. This is likely attributed to TSP and PNN's limited coverage over certain shelves, especially for shelves that may not be passed by using shortest path algorithms (see Figure 3b).

Table 2: Divergence between shelf traffic density of different methods and that of ground-truth human trajectories. Lower (bolded) scores indicate closer alignment with human behaviour.

(a) Proportional Sampling: Basket trajectories sampled proportionally to real-world purchase frequencies, potentially biasing results toward popular baskets.

Divergence	TSP	PNN	RL
JSD	0.632	0.549	**0.430**
WD	0.313	0.278	**0.217**

(b) Uniform Sampling: Each basket sampled equally, avoiding overrepresentation of popular baskets. All divergences reported are measured against uniformly-sampled human trajectories. The last column shows divergence between proportionally and uniformly sampled human trajectories to illustrate effect of sampling choice.

Divergence	TSP	PNN	RL	Human
JSD	0.505	0.506	**0.347**	0.224
WD	0.0106	0.0106	**0.00676**	0.00517

4.3 Use Case 2: Estimating Impulse Rates

To optimize store layouts for impulse profit, it is insufficient to only know which shelves receive the most traffic; we must also estimate which products are most likely to be purchased spontaneously — that is, which products exhibit the highest impulse rates.

Prior approaches have estimated impulse rates using domain expertise [34], customer surveys [3], or by assuming a direct correlation with purchase probabilities [13]. However, domain knowledge may be unreliable or unavailable, especially for new or smaller retailers. Surveys are resource-intensive to conduct and may not scale well. Purchase probability is not a reliable proxy for impulse rate, since essential products are frequently purchased but rarely on impulse.

In our case, access to simulated customer trajectories enables a data-driven estimation of impulse rates using the following formulation adapted from Azar and Daou [3]:

$$P_{\text{purchase}} = P_{\text{visit\_shelf}} \times P_{\text{prod\_visibility}} \times P_{\text{impulse rate}} \quad (2)$$

Table 3: Purchase probabilities (P_{purchase}) for each product category across three customer clusters (Cluster 1-3).

Product Category	Cluster 1	Cluster 2	Cluster 3
Hot Coffee/Tea	0	0.642	1
Bakery/Pastries	0	1	0
Hot Food	0.145	0	0
Fruits/Yogurt	0.0427	0.0363	0
Energy Drinks	0.287	0	0
Cold Beverages	0.274	0	0.00809
Soft Drinks	0.0668	0.0130	0
Snack Bars	0.0438	0	0
Cold Food	0.0296	0	0
Cold Coffee/Tea/Shake	0.0482	0	0
Fountain Drinks	0.0635	0	0

where for a given product,

- P_{purchase} is the probability that the product is purchased,
- $P_{\text{visit\_shelf}}$ represents the probability that the product's corresponding shelves are visited,
- $P_{\text{prod\_visibility}}$ is the probability that the product is visible to the customer upon visiting the shelf, and
- $P_{\text{impulse rate}}$ is the probability that the product is spontaneously purchased upon being seen.

In our case, each shelf contains only a single product category, so $P_{\text{prod\_visibility}} = 1$, simplifying Equation 2 to:

$$P_{\text{purchase}} = P_{\text{visit\_shelf}} \times P_{\text{impulse rate}} \tag{3}$$

This equation expresses the intuition that a product can only be purchased if the shelf is visited (and the product seen), and then selected impulsively.

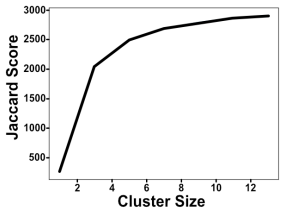


Figure 4: Within-Cluster Sum of Squares score against the number of clusters.

To compute P_{purchase}, we follow the method used by Dorismond et al. [13] and clustered all 61 baskets using the elbow method, resulting in three clusters (Figure 4). Computing the average basket per cluster then provides us with P_{purchase} values for each product category (Table 3). Following their definition, within each cluster, products with $P_{\text{purchase}} < 20\%$ are classified as impulse products.

To estimate $P_{\text{visit\_shelf}}$, we simulate customers purchasing solely essential (i.e., non-impulse) products. This reflects the nature of impulse products: they should be picked up spontaneously, not as part of a preplanned list. For products stored across multiple shelves, we compute $P_{\text{visit\_shelf}}$ as the sum of visit probabilities across all relevant shelves (capped to a maximum of 1). $P_{\text{impulse rate}}$ is then computed by dividing P_{purchase} by $P_{\text{visit\_shelf}}$ for each impulse product, per Equation 3.

As a concrete example, consider Cluster 2 in Table 3, where Hot Coffee/Tea and Bakery/Pastries are essential items, and Soft Drinks and Fruits/Yogurt are impulse products. 10,000 trajectories are generated, such that customers purchase Hot Coffee/Tea with a probability of 0.642, and Bakery/Pastries with a probability of 1. Checkout destinations are sampled to match the empirical checkout distribution observed in the original dataset.

Table 4: Estimated impulse purchase rates across all methods for Cluster 2's impulse products . `Inf` indicates a product was purchased despite no recorded shelf visits.

Product Category	PNN	TSP	RL	Human
Fruits/Yogurt	`Inf`	`Inf`	0.0577	0.115
Soft Drinks	`Inf`	`Inf`	7.20	3.20

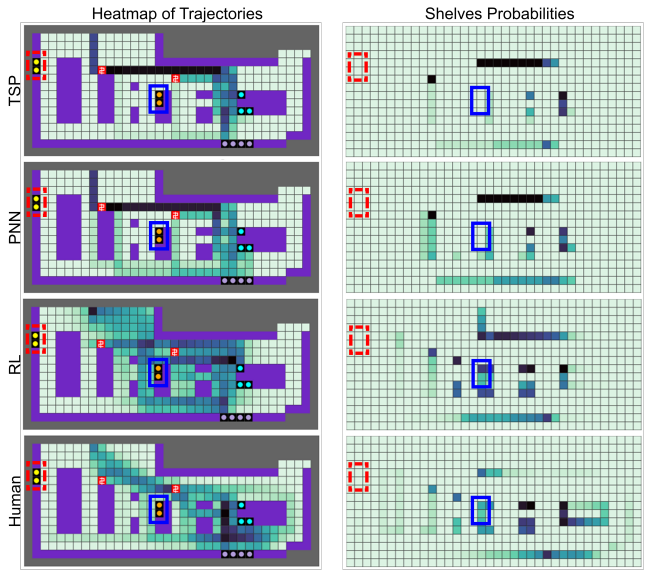


Figure 5: Left column shows trajectory heatmaps for Cluster 2; right column shows the corresponding shelf traffic density. Red (dotted) and blue (solid) boxes denote shelf regions for Soft Drinks and Fruits/Yogurt, respectively. It can be observed that generated trajectories for TSP and PNN have not visited these shelves.

From Table 4, it can be observed that the generated trajectories of TSP and PNN miss the shelves housing both impulse items entirely, resulting in a division by zero when computing impulse rates via Equation 3. This is likely because the shelves of Cluster 2's impulse products (i.e., Soft Drinks and Fruits/Yogurt) lie outside the direct routes connecting its essential products to checkout counters (illustrated at Figure 5), hence posing a challenge for shortest-path based methods.

In contrast, RL-generated trajectories exhibit more diverse routing behaviour and successfully reach the shelves containing Cluster 2's impulse items. As observed in Table 4, while the estimated absolute values deviate somewhat from the ground truth, the relative ordering is preserved — Soft Drinks appear more impulsive than Fruits/Yogurt. This insight is particularly valuable for downstream product placement /promotional tasks.

Table 5: Estimated impulse rates (i_p) and per-unit impulse profits (π_p) for Cluster 2 impulse products across different methods. π_p is computed using Equation 4, with the product price of $3.79 (Fruits/Yogurt) and $3.59 (Soft Drinks), and a uniform margin of 5%. Only RL identifies Soft Drinks as the more profitable impulse product, which is consistent with the ground-truth trajectories (shown in bold).

Metric	Method	Fruits/Yogurt	Soft Drinks
i_p	PNN	0.0363	0.013
	TSP	0.0363	0.013
	RL	0.0577	7.20
	Human	0.115	3.20
π_p	PNN	0.00688	0.00233
	TSP	0.00688	0.00233
	RL	0.0109	**1.29**
	Human	0.0218	**0.574**

4.4 Use Case 3: Informing Layout Changes

Having computed both impulse rates and shelf traffic densities, we now demonstrate how these metrics can inform concrete layout decisions to improve store profitability. For instance, these values can be directly used in optimization formulations such as those proposed by Abdelaziz et al. [1] and Flamand et al. [17]. In our case, we utilize these findings to reposition a single impulse product to maximize store profit, demonstrating tangible improvements that these insights can bring to the store.

Step 1: Identifying the Most Profitable Impulse Product. We begin by selecting the impulse product with the highest potential for profit. This is determined using the following heuristic, which estimates the expected impulse profit for product p. It is derived from the profit maximization objective in Equation 1, under the assumption that all shelves (and thus all products) are always visited:

$$\begin{aligned} \pi_p &= i_p \times \rho_p \\ &= i_p \times \text{price}_p \times \text{margin}_p \end{aligned} \tag{4}$$

Here, π_p denotes the expected impulse profit of product p, i_p is the impulse rate, and ρ_p is the per-unit profit (price times margin). Due to the lack of product-level profit margin data, we assume a uniform margin of 5% based on industry estimates [36].

Consistent with Use Case 2, we focus on Cluster 2, where Soft Drinks and Fruits/Yogurt are impulse products. Since TSP and PNN trajectories do not intersect with the shelves of both of these products, their impulse rates i_p are approximated using purchase probabilities (from Table 3) instead, with the assumption that higher purchase frequency correlates with higher impulsivity.

Table 5 compares the estimated impulse rates and per-unit profits across all methods. While TSP and PNN rank Fruits/Yogurt as more profitable, RL identifies Soft Drinks as the more profitable impulse product, which is consistent with the ground-truth trajectories, once again demonstrating RL's stronger behavioural alignment with customers movement patterns.

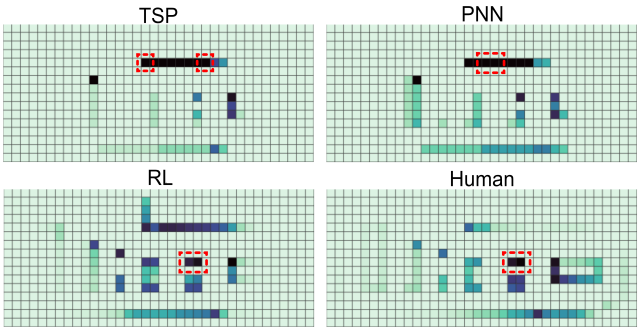


Figure 6: Shelf traffic density heatmaps for Cluster 2 trajectories. Red dotted boxes highlight the top two *empty* shelves with the highest visitation probability for each method. Notably, RL identifies the same high-traffic shelves as the human data.

Step 2: Selecting the Best Shelf Location. Intuitively, to maximize profit, the most profitable impulse product should ideally be placed on the store's most frequently visited *unoccupied* shelves. Alternatively, one could also swap this product with highly-visited shelves that are currently *occupied* by low-profit products. While both strategies similarly increase visibility and thus purchase likelihood for the impulse product, highly-visited, low-profit products are often essential products that customers deliberately seek for. Retailers usually place these essential products strategically to draw customers to specific areas, therefore rearranging them risks disrupting the overall traffic flow, potentially reducing visits to certain parts of the store, leading to reduced sales. From this perspective, the first strategy—using unoccupied shelves—is less risky and more promising, and is therefore adopted in this work.

To implement this, we rank all unoccupied shelves by their estimated visitation probability and select the top two (since both candidate impulse products occupy two shelves). Shelf traffic density heatmaps with top-ranked shelves highlighted are shown in Figure 6.

Step 3: Repositioning the Product. Then, for each method, we place its chosen product on the top two unoccupied shelves. These layout decisions can be visualized in Figure 7.

Step 4: Evaluating the New Layouts. Since only the location of impulse product changes, human trajectories that purchase essential products in the original layout remain valid, allowing them to be used for evaluation.

Actual customer trajectories from Cluster 2 are rolled-out in the modified environments, where ground-truth impulse rates (rightmost column of Table 4) are applied to determine whether a customer makes an impulse purchase upon encountering the repositioned impulse product.

Then, using the product prices and assumed profit margin, we compute the average impulse profit per customer under each method. As shown in Table 6, the RL-informed layout yields the highest profit gain per customer, closely matching the optimal result derived from ground-truth trajectories. In contrast, layouts based on TSP and PNN achieve substantially lower returns. This is due to two key limitations: first, their predictions of most visited shelves

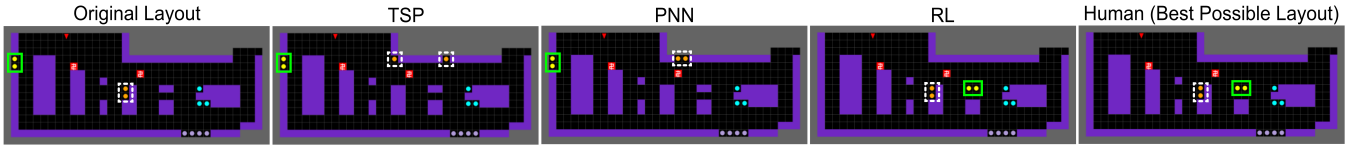


Figure 7: Visualization of shelf layout recommendations by different methods. Suggested shelf placements for Soft Drinks and Fruits/Yogurt are marked with green (solid) and white (dotted) boxes, respectively. Only RL replicates the ground-truth layout and achieves the highest profit gain (Table 6).

Table 6: Average impulse profit per customer (in $) under the original and suggested layouts (visualized in Figure 7), when evaluated with each method's own trajectories or with human trajectories. As shown in bold, the RL-informed layout yields the highest profit when evaluated with human data.

Layout	TSP	PNN	RL	Human
Original (Evaluated with own trajectories)	0	0	0.0140	0.00760
Suggested (Evaluated with own trajectories)	0.180	0.171	0.163	0.162
Suggested (Evaluated with human trajectories)	0.0900	0.0501	**0.162**	**0.162**

are suboptimal. Second, their reliance on shortest paths prevent them from exploring other routes that customers may take when pursuing the same basket, leading to their failure in computing impulse rates for products located away from shortest paths. By contrast, RL — under the maximum entropy framework — recovers multiple plausible customer routes, allowing it to identify the same most visited shelves and most impulsive products as those derived from ground truth human trajectories.

5 DISCUSSION AND FUTURE WORK

In this section, we discuss key implications of our study and outline several directions for future research, organized by theme.

Modelling Customers with RL. Our RL-based approach to trajectory modelling demonstrates clear behavioural advantages over heuristics, but also introduces challenges. Training robust policies is computationally expensive, whereas TSP and PNN require no training and can be computed on-the-fly. Moreover, our current RL policies must be retrained for each new layout, limiting their practicality for iterative optimization. A promising direction is to perform domain randomization on store layouts, enabling the policy to generalize across store configurations without retraining from scratch. Additionally, while the computational cost of TSP and PNN paths scale with the basket size (exponentially and quadratically, respectively), inference cost of the RL policy remains constant. Hence, for stores with large assortments, the upfront training cost of RL is amortized over repeated use.

Evaluation Methods. Our evaluation showed that RL-driven estimates of shelf traffic and impulse rates can inform product repositioning decisions, yielding profit gains comparable to those obtained from ground-truth trajectories. However, this analysis was restricted to a single customer type and one impulse product, limiting generalizability. Future work could extend evaluation to multiple customer segments and product categories, and ultimately validate the approach through real-world deployments to assess its practical impact in retail environments.

End-to-end Layout Optimization with RL. An intriguing direction is to extend RL beyond simulating customer trajectories to directly optimizing store layouts. Inspired by RL's success in design tasks such as chip placement [33], one could envision a hierarchical framework where an inner agent models customer behaviour while an outer agent iteratively adjusts product placements, enabling end-to-end layout optimization.

Increasing Accessibility. Research in this area has traditionally relied on proprietary data and solving complex optimization tasks with specialized software, making replication difficult. Aside from reducing the need for costly real-world data through RL-generated trajectories, our work also aims to make store layout optimization more accessible. By releasing our code, including implementations of all methods and the environment, alongside installation/running instructions, we hope to provide a first step toward open, reproducible approaches, with the goal of enabling retailers of all sizes to benefit from these layout optimization methods.

6 CONCLUSION

Understanding customer behaviour through their movements play a critical role in store layout optimization. While real trajectory data offers accurate insights, it is costly to collect and often infeasible for many retailers. Heuristics such as the Travelling Salesman Problem (TSP) and Probabilistic Nearest Neighbour (PNN) are widely used as inexpensive substitutes, but real customer paths deviate by an average of 28%, highlighting their limitations.

To address this gap, we framed customers as maximum entropy reinforcement learning (RL) agents, which explicitly models reward-seeking and stochasticity, better capturing bounded rationality. Our results show that RL-generated trajectories align more closely with customers movement than TSP or PNN, yielding more accurate predictions of shelf traffic densities and impulse purchase rates.

Beyond trajectory similarity, when used to guide a product repositioning task, RL produced layout improvements that matched those derived from ground-truth data — a result unattainable with heuristics.

Overall, RL emerges as a practical and behaviourally grounded alternative, bridging the gap between oversimplified heuristics and costly trajectory datasets, making accurate layout optimization more broadly accessible.

REFERENCES

[1] Fouad Ben Abdelaziz, Bacel Maddah, Tülay Flamand, and Jimmy Azar. 2024. Store-wide space planning balancing impulse and convenience. *European Journal of Operational Research* 312, 1 (2024), 211–226.

[2] Salam Qaddoori Dawood Al-Zubaidi, Gualtiero Fantoni, and Franco Failli. 2021. Analysis of drivers for solving facility layout problems: A Literature review. *Journal of industrial information integration* 21 (2021), 100187.

[3] Jimmy Azar and Hoda Daou. 2023. In-Store Traffic Density Estimation. In *Retail Space Analytics*. Springer, 35–50.

[4] Danny N Bellenger, Dan H Robertson, and Elizabeth C Hirschman. 1978. Impulse buying varies by product. *Journal of advertising research* 18, 6 (1978), 15–18.

[5] Joyendu Bhadury, Rajan Batta, Jessica Dorismond, Chien-Chih Peng, and Shrideep Sadhale. 2016. *Store layout using location modelling to increase purchases.* Technical Report. University of Buffalo working paper. http://www. acsu. buffalo. edu/˜ batta ....

[6] Ahmet Reha Botsali. 2007. *Retail facility layout design.* Ph.D. Dissertation. Texas A & M University.

[7] A Reha Botsalı, Georgia-Ann Klutke, and Brett A Peters. 2023. Effect of Customer Travel Behavior on Grid Layout and Shelf Space Allocation in Retail Facilities. In *Retail Space Analytics*. Springer, 1–20.

[8] Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo Perez-Vicente, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan Terry. 2023. Minigrid & Miniworld: Modular & Customizable Reinforcement Learning Environments for Goal-Oriented Tasks. In *Advances in Neural Information Processing Systems 36, New Orleans, LA, USA*.

[9] Marcel Corstjens and Peter Doyle. 1981. A model for optimizing retail space allocations. *Management Science* 27, 7 (1981), 822–833.

[10] Marcel Corstjens and Peter Doyle. 1983. A dynamic model for strategically allocating retail space. *Journal of the Operational Research Society* 34, 10 (1983), 943–951.

[11] Elif Danisman and Alice E Smith. 2023. Data-Driven Analytical Grocery Store Design. In *Retail Space Analytics*. Springer, 75–101.

[12] Jessica Dorismond. 2016. Supermarket optimization: Simulation modeling and analysis of a grocery store layout. In *2016 Winter Simulation Conference (WSC)*. 3656–3657. https://doi.org/10.1109/WSC.2016.7822385

[13] Jessica Dorismond, Jose L Walteros, and Rajan Batta. 2023. A Simulation Based Tool to Guide Periodic Changes in a Supermarket Layout. In *Retail Space Analytics*. Springer, 51–74.

[14] Amine Drira, Henri Pierreval, and Sonia Hajri-Gabouj. 2007. Facility layout problems: A survey. *Annual reviews in control* 31, 2 (2007), 255–267.

[15] Gihan S Edirisinghe and Charles L Munson. 2023. Strategic rearrangement of retail shelf space allocations: Using data insights to encourage impulse buying. *Expert Systems with Applications* 216 (2023), 119442.

[16] Tulay Flamand, Ahmed Ghoniem, and Bacel Maddah. 2016. Promoting impulse buying by allocating retail shelf space to grouped product categories. *Journal of the Operational Research Society* 67, 7 (2016), 953–969.

[17] Tülay Flamand, Ahmed Ghoniem, and Bacel Maddah. 2023. Store-Wide Shelf-Space Allocation with Ripple Effects Driving Traffic. *Operations Research* 71, 4 (2023), 1073–1092. https://doi.org/10.1287/opre.2023.2437 arXiv:https://doi.org/10.1287/opre.2023.2437

[18] Ahmed Ghoniem, Tulay Flamand, and Mohamed Haouari. 2016. Optimization-based very large-scale neighborhood search for generalized assignment problems with location/allocation considerations. *INFORMS Journal on Computing* 28, 3 (2016), 575–588.

[19] Donald H Granbois. 1968. Improving the study of customer in-store behavior. *Journal of Marketing* 32, 4_part_1 (1968), 28–33.

[20] Evren Gul, Alvin Lim, and Jiefeng Xu. 2023. Retail store layout optimization for maximum product visibility. *Journal of the Operational Research Society* 74, 4 (2023), 1079–1091.

[21] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. 2017. Reinforcement learning with deep energy-based policies. In *International conference on machine learning*. PMLR, 1352–1361.

[22] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*. Pmlr, 1861–1870.

[23] Sagarkumar Hirpara and Pratik J Parikh. 2021. Retail facility layout considering shopper path. *Computers & Industrial Engineering* 154 (2021), 106919.

[24] Kimberly Holmgren. 2021. Customer path generation simulation for selection from proposed grocery store layouts. In *2021 Winter Simulation Conference (WSC)*. IEEE, 1–11.

[25] Jan Holmström. 1997. Product range management: a case study of supply chain operations in the European grocery industry. *Supply Chain Management: An International Journal* 2, 3 (1997), 107–115.

[26] Sam K Hui, Peter S Fader, and Eric T Bradlow. 2009. Research note—the traveling salesman goes shopping: The systematic deviations of grocery paths from TSP optimality. *Marketing science* 28, 3 (2009), 566–572.

[27] Sam K Hui, J Jeffrey Inman, Yanliu Huang, and Jacob Suher. 2013. The effect of in-store travel distance on unplanned spending: Applications to mobile promotion strategies. *Journal of Marketing* 77, 2 (2013), 1–16.

[28] Easwar S Iyer. 1989. Unplanned Purchasing: Knowledge of shopping environment and. *Journal of retailing* 65, 1 (1989), 40.

[29] Lene Granzau Juel-Jacobsen. 2015. Aisles of life: outline of a customer-centric approach to retail space management. *The International Review of Retail, Distribution and Consumer Research* 25, 2 (2015), 162–180.

[30] David T Kollat and Ronald P Willett. 1967. Customer impulse purchasing behavior. *Journal of marketing research* 4, 1 (1967), 21–31.

[31] Chen Li. 2011. *A facility layout design methodology for retail environments.* Ph.D. Dissertation. University of Pittsburgh.

[32] Xiangyu Li. 2023. *Dynamic Digital Twins for On-Shelf Availability in the Retail Store.* McGill University (Canada).

[33] Azalia Mirhoseini, Anna Goldie, Mustafa Yazgan, Joe Wenjie Jiang, Ebrahim Songhori, Shen Wang, Young-Joon Lee, Eric Johnson, Omkar Pathak, Azade Nova, et al. 2021. A graph placement methodology for fast chip design. *Nature* 594, 7862 (2021), 207–212.

[34] Elif Ozgormus and Alice E Smith. 2020. A data-driven approach to grocery store block layout. *Computers & Industrial Engineering* 139 (2020), 105562.

[35] POPAI. 2014. The 2014 POPAI Mass Merchant Shopper Engagement Study: Media Report.

[36] Remi. 2024. How Profitable is a Convenience Store? Revenue & Profits Analysis — sharpsheets.io. https://sharpsheets.io/blog/how-profitable-is-a-convenience-store/. [Accessed 14-05-2025].

[37] Dennis W. Rook and Robert J. Fisher. 1995. Normative Influences on Impulsive Buying Behavior. *Journal of Consumer Research* 22, 3 (12 1995), 305–313. https://doi.org/10.1086/209452 arXiv:https://academic.oup.com/jcr/article-pdf/22/3/305/5069267/22-3-305.pdf

[38] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[39] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction.* MIT press.

[40] Paco Underhill. 2009. *Why we buy: The science of shopping–updated and revised for the Internet, the global consumer, and beyond.* Simon and Schuster.

[41] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. 2008. Maximum entropy inverse reinforcement learning.. In *Aaai*, Vol. 8. Chicago, IL, USA, 1433–1438.