



Learning promotion policies with attention-based deep Q-networks

Yingnan Xu^{1,2} · Xuchun Wu¹ · Zhenjun Li¹ · Congli Liu¹ · Yansheng Zhang¹

Received: 15 July 2025 / Accepted: 11 September 2025 / Published online: 3 October 2025

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2025

Abstract

In financial services, personalized promotion strategies are critical for sustaining customer engagement and driving asset growth. We present FAT-DQN, a deep reinforcement learning framework for off-line environments that models sequential decision-making as a Markov Decision Process (MDP), where promotional actions influence future changes in customer assets under management (AUM). FAT-DQN extends the standard Deep Q-Network (DQN) architecture with a multi-head self-attention mechanism over promotion–reward histories augmented by learnable temporal encodings, and applies Feature-wise Linear Modulation (FiLM) to incorporate customer-segment embeddings. To improve robustness, we employ per-customer reward normalization and evaluate policies with both ranking-based metrics and counterfactual off-policy estimators. Empirical results on real promotion logs show that FAT-DQN consistently outperforms baseline methods, yielding a higher mean NDCG@3 (0.7744) compared to Batch-Constrained deep Q-learning (BCQ, 0.7325) and DQN (0.6852). It further improves alignment between predicted and realized outcomes, achieving a Spearman correlation of 0.2584, compared to 0.1619 for BCQ and 0.1522 for DQN. Counterfactual evaluations further show that FAT-DQN delivers consistently strong off-policy estimates, confirming its robustness across evaluation settings. These findings demonstrate that attention-based architectures with modulation offer a more effective and interpretable alternative to standard reinforcement learning approaches for personalized promotion planning in financial services.

Keywords Offline reinforcement learning · Deep Q-network (DQN) · Customer behavior modeling · Financial promotion strategy · Temporal user embedding · Applied machine learning

1 Introduction

Personalized promotion is a central driver of customer engagement and long-term asset growth in financial services, yet it poses unique challenges. Interactions are sparse, outcomes

such as changes in assets under management (AUM) are delayed, and behavioral responses depend strongly on temporal and contextual factors. Conventional heuristics and segmentation-based scoring adapt poorly to such dynamics, often emphasizing short-term responses while overlooking customer heterogeneity and exposure saturation effects.

Reinforcement learning (RL) offers a principled framework for sequential decision-making under uncertainty [1]. Classical Q-learning [2] estimates the action–value function to guide optimal policies, but its tabular form scales poorly in large state–action spaces. Deep Q-Networks (DQN) address this limitation by approximating Q-values with deep neural networks [3], enabling RL to handle high-dimensional inputs and large discrete action sets. However, their direct application to financial promotion is limited by two key issues: determining which past actions and outcomes remain influential, and accommodating substantial variation in response patterns across customer segments.

Recent advances in representation learning provide building blocks for addressing these issues. Attention mechanisms allow models to focus selectively on the most relevant

✉ Yingnan Xu
yingnanx@mail.smu.edu

Xuchun Wu
wxcwuxuchun@163.com

Zhenjun Li
lzj@qlbchina.com

Congli Liu
lcl@qlbchina.com

Yansheng Zhang
17615850921@163.com

¹ Department of Digital Banking, Qilu Bank, JingShi Rd 10817, 250014 Jinan, Shandong, PR China

² Zhongtai Securities Institute for Financial Studies, Shandong University, Shanda S Rd, 250014 Jinan, Shandong, PR China

past interactions [4], while Feature-wise Linear Modulation (FiLM) offers a parameter-efficient means of adapting representations across heterogeneous populations [5]. Yet these techniques have largely been explored in isolation: DRQN and DARQN [6, 7] apply recurrence or spatial attention to visual control tasks, and sequential recommenders such as SASRec and BERT4Rec [8, 9] employ attention for next-item prediction rather than long-term financial outcomes. Their integration within a reinforcement learning framework for promotion optimization remains underexplored.

To fill this gap, we propose FAT-DQN (FiLM-modulated, Attention-based Temporal DQN), which extends standard DQN with three key components:

- a token-level self-attention encoder with learnable temporal encodings for short, sparse action–reward histories;
- FiLM-based segment conditioning for efficient and interpretable adaptation across customer groups;
- an exposure vector that captures recency and saturation effects.

Together, these components yield a lightweight yet expressive architecture tailored to the offline, sparse, and heterogeneous nature of financial promotion.

The remainder of this paper is organized as follows: Section 2 reviews related works in reinforcement learning for financial services and personalized promotion. Section 3 describes the model architecture. Section 4 details the experimental setup, including environment construction and evaluation metrics. Section 5 presents quantitative and qualitative results, including comparative and ablation studies. Finally, Section 6 concludes the paper with discussions on practical implications and future research directions.

2 Related works

Reinforcement learning (RL) has been extensively applied in finance, particularly in trading and portfolio management. Early studies explored direct RL under noisy and delayed signals [10], while deep RL methods enabled representation learning directly from market data [11]. Risk-sensitive extensions, such as short-selling under volatility, have also been proposed [12]. More recently, attention-enhanced RL frameworks have emerged: transformer-based portfolio strategies leveraging the Black–Litterman model [13], the Financial Transformer RL framework [14], and residual channel–spatial attention networks for stock forecasting [15] demonstrate the utility of attention-based temporal modeling in financial contexts, though these works remain

focused on portfolio optimization and forecasting rather than customer promotion sequencing. Similarly, attention-based architectures for financial time series prediction [16, 17], surveys of transformers in quantitative trading [18], and recent quantitative transformer strategies [19] further validate the growing applicability of transformer-based designs in financial decision systems.

Beyond trading, RL has been explored in pricing and recommendation tasks. Multi-agent RL has been employed for dynamic pricing in supply chains [20], while offline RL methods have been developed for cold-start recommendation and delayed feedback personalization [21]. Large-scale digital platforms have further adapted contextual bandits and RL to optimize user engagement [22–24]. However, these approaches typically emphasize short-term engagement metrics such as clicks, rather than optimizing for long-term financial outcomes.

Attention mechanisms [4] enable selective focus on relevant temporal or spatial signals, while Feature-wise Linear Modulation (FiLM) [5] provides lightweight, context-aware conditioning through affine transformations of latent representations. Together, attention and FiLM have demonstrated versatility across a wide range of domains [25–29]. Their demonstrated generality motivates their integration within reinforcement learning for financial promotion optimization, where temporal adaptability and interpretable segment-specific adaptation are critical.

In sequential modeling, attention has advanced recommendation through architectures such as SASRec, BERT4Rec, and DIN [8, 9, 30], while RL variants including DRQN, DARQN, and MANet [6, 7, 31] employ attention to improve value estimation in partially observable environments. Nonetheless, these methods are largely designed for perceptual control or short-horizon next-item prediction, rather than for the offline, long-horizon optimization of financial promotions.

Building on these threads, our work introduces FAT-DQN, which combines token-level self-attention with FiLM-based segment conditioning in an offline RL framework. This design yields a lightweight and interpretable architecture that jointly models temporal influence, customer heterogeneity, and exposure saturation—challenges that are central to financial promotion optimization but remain insufficiently addressed by existing approaches.

3 Model architecture

3.1 Definitions

We cast personalized promotion recommendation in financial services as a Markov Decision Process (MDP).

Definition 3.1 (State Space $\mathcal{S}$) The state $s_t \in \mathcal{S}$ represents the observable characteristics of a customer at time t , including static demographics (e.g., age, income), dynamic financial variables (e.g., balance, AUM trajectory), and historical interaction features (e.g., past promotions, exposures). This design provides a sufficient summary of the customer's relevant history, enabling the Markov property.

Definition 3.2 (Action Space $\mathcal{A}$) The action space $\mathcal{A}$ is a finite set of promotion types. At each decision epoch t , the agent selects an action $a_t \in \mathcal{A}$ corresponding to which promotion to deliver (e.g., deposit rate offer, fund recommendation, product bundle).

Definition 3.3 (Transition Kernel P) The transition dynamics $P(s_{t+1} \mid s_t, a_t)$ describe the probability of moving to a new customer state s_{t+1} after applying action a_t in state s_t . In practice, P is implicitly defined by historical customer trajectories: the next state arises from observed behavioral and financial responses to promotions.

Definition 3.4 (Reward Function r) The reward $r(s_t, a_t)$ measures the incremental effect of applying promotion a_t in state s_t , quantified as peer-differenced AUM uplift (see Section 3.3). This design isolates the net contribution of the promotion while controlling for market- or cluster-level trends.

Definition 3.5 (Discount Factor γ) The discount factor $\gamma \in [0, 1]$ determines the weight assigned to long-term outcomes. We set $\gamma = 0.99$ to emphasize sustained asset growth and long-horizon customer engagement.

Definition 3.6 (Markov Decision Process) A Markov Decision Process is a tuple

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma),$$

where the next state depends only on the current state and action, i.e.,

$$\Pr(s_{t+1} \mid s_0, a_0, \dots, s_t, a_t) = \Pr(s_{t+1} \mid s_t, a_t).$$

This formulation is well suited to the financial promotion recommendation setting, as the state representation s_t is explicitly constructed to capture the essential historical and contextual information relevant to decision-making, including demographics, account balances, and recent promotion exposures. Within this representation, the Markov property can be assumed to hold approximately: conditional on (s_t, a_t) , the distribution of future outcomes depends minimally on the longer trajectory. This abstraction provides a principled basis for applying reinforcement

learning, consistent with established practice in recommender systems and financial decision-making. We acknowledge that certain external influences—such as macroeconomic shocks or unobserved behavioral factors—may not be fully encoded, yet the Markov Decision Process remains a tractable and widely accepted approximation for sequential optimization under uncertainty in real-world financial applications.

3.2 Model design

The proposed **FAT-DQN** (FiLM-modulated, Attention-based Temporal Deep Q-Network) is a value-based reinforcement learning framework specifically designed for financial promotion optimization. FAT-DQN extends the standard Deep Q-Network (DQN) by integrating three key modules: a token-level temporal self-attention encoder that adaptively weighs historical interactions, a FiLM-based conditioning mechanism that enables segment-specific adaptation, and an exposure vector that explicitly models recency and saturation effects. The integration of these components yields an interpretable and parameter-efficient architecture tailored to the characteristics of financial promotion trajectories, which are typically sparse, heterogeneous, and influenced by delayed responses. By jointly capturing temporal dependencies, customer heterogeneity, and diminishing returns, FAT-DQN provides a principled yet lightweight approach for learning effective promotion policies under offline reinforcement learning settings. The overall architecture of FAT-DQN is illustrated in Fig. 1, with the main components and their interactions formally summarized in Algorithm 1, and the notations used throughout the model definition detailed in Table 1.

3.2.1 Tokenized history representation

Each customer trajectory is segmented into monthly windows, where each window records the promotions delivered, the outcomes observed, and their temporal positions. We represent each past interaction in month t as a token

$$z_t = [e^a(a_t), e^r(r_t), e^t(t)], \quad (1)$$

where $e^a(\cdot)$ denotes a promotion embedding, $e^r(\cdot)$ is a reward projection, and $e^t(\cdot)$ is a learnable temporal encoding.

In standard transformer architectures, temporal order is introduced using fixed sinusoidal position encodings that assume smooth, periodic dependencies between positions. However, financial promotions do not necessarily follow such regularities: the effect of recency may decay sharply, persist for irregular intervals, or interact with customer-specific attributes. A fixed encoding cannot capture such irregular, non-stationary dynamics.

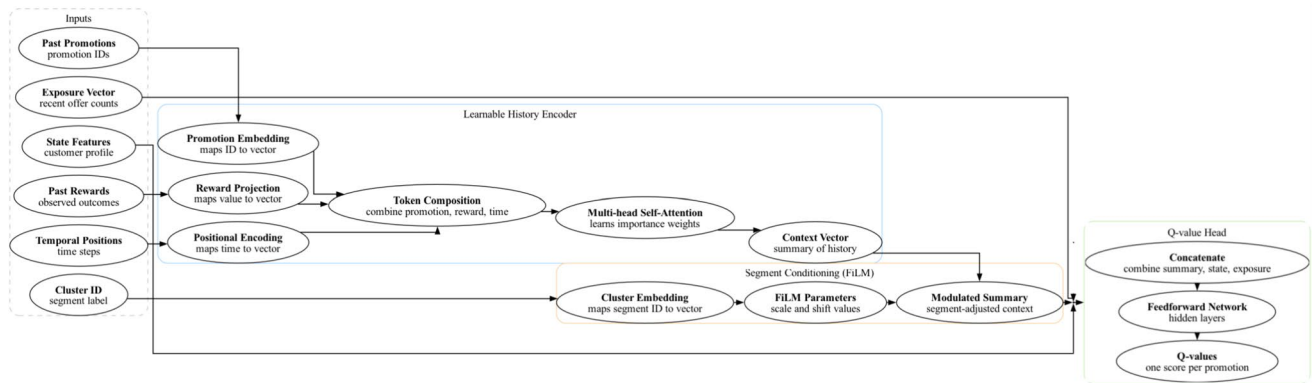


Fig. 1 Overview of the proposed FAT-DQN architecture. Historical promotions and rewards are encoded through a token-level self-attention module, modulated by FiLM-based segment conditioning, and

combined with exposure and state features. The unified representation is passed through a feedforward Q-value head to generate promotion values

Table 1 Summary of notations used in this section

Symbol	Definition
s_t	Customer state at time t (features, demographics, AUM variables)
a_t	Promotion action at time t
r_t	Reward signal (peer-differenced AUM uplift)
H_t	History of past L promotions, rewards, and time indices
x_t^e	Exposure vector (recent counts of promotion categories)
c	Customer cluster / segment ID
$e^a(a_t)$	Promotion embedding
$e^r(r_t)$	Reward projection
$e^t(t)$	Learnable temporal encoding
$e^c(c)$	Cluster embedding for FiLM conditioning
h	Attention-derived context vector from history H_t
$\tilde{h}$	FiLM-modulated context vector
u_t	Unified feature vector $[\tilde{h} x_t^e s_t]$
q_t	Q-values over actions at time t
θ, θ^-	Online and target network parameters

By making $e^t(\cdot)$ learnable, FAT-DQN allows the model to discover temporal patterns directly from data rather than imposing a pre-specified functional form. This design provides flexibility to capture heterogeneous and context-dependent temporal effects—such as short-lived responses to repeated offers or delayed impacts of high-value promotions—while still embedding them in a common latent space with promotion and reward information. As a result, the model can more effectively weigh historical interactions when estimating long-term promotion value.

3.2.2 Temporal self-attention encoder

The tokenized representation described above transforms each past interaction into a vector $z_t = [e^a(a_t), e^r(r_t), e^t(t)]$

that encodes promotion type, outcome magnitude, and temporal position in a learnable manner. A sequence of L such tokens $\{z_{t-L}, \dots, z_{t-1}\}$ is then passed through a multi-head self-attention block [4]:

$$h = \text{SelfAttn}(z_{t-L:t-1}), \quad (2)$$

yielding a context vector h that aggregates the historical sequence.

This design links the learnable temporal encoding with adaptive attention weighting: whereas $e^t(\cdot)$ provides a flexible embedding of temporal position, the self-attention mechanism determines which interactions remain influential for predicting long-term outcomes. Together, they move beyond fixed recency windows or hand-crafted decay features, enabling the model to discover irregular influence patterns—such as short-lived responses, persistent long-tail effects, or context-dependent delays—directly from data. This improves promotion assignment under delayed and sparse financial feedback, ensuring that the encoder captures both the timing and relative importance of past promotions and outcomes.

3.2.3 FiLM-based segment conditioning

To capture systematic differences in promotion responsiveness across customer groups, we incorporate FiLM. Originally proposed in computer vision to adapt convolutional feature maps across tasks, FiLM provides a simple yet effective mechanism for conditioning representations on external signals. In our setting, each customer segment c is represented by an embedding $e^c(c)$, from which FiLM generates affine parameters (γ_c, β_c) applied to the context vector h derived from the temporal self-attention encoder:

$$\tilde{h} = \gamma_c \odot h + \beta_c. \quad (3)$$

This design is particularly necessary in financial promotion because customer heterogeneity is structured but not extreme: segments differ systematically (e.g., by demographics or financial behavior) but still share many commonalities. Training entirely separate models per segment risks statistical inefficiency and overfitting, while a single undifferentiated model fails to capture segment-specific nuances. FiLM strikes a balance by allowing lightweight, segment-aware modulation of the shared representation h , achieving parameter efficiency, robustness in data-scarce clusters, and interpretability. In practice, this means the attention encoder identifies which past actions and outcomes matter, and FiLM adapts this summary to reflect how different segments value similar histories, enabling more personalized and compliant promotion policies.

3.2.4 Promotion-exposure modeling

While the temporal self-attention encoder captures the sequential importance of past promotions and FiLM provides segment-specific adaptation, neither directly accounts for **recency bias**¹ or **saturation effects**² arising from repeated exposures. To address this, we introduce an exposure vector $x^e \in \mathbb{R}^K$ that tracks the recent frequency of offers across K promotion categories. This order-agnostic representation complements the sequence encoder by explicitly encoding how repeated offers diminish marginal impact, providing the model with an additional signal that pure temporal attention may overlook.

3.2.5 Q-value estimation

The FiLM-modulated context vector $\tilde{h}$, the exposure vector x^e , and the current customer state s_t are concatenated into a unified representation:

$$u_t = [\tilde{h} \parallel x^e \parallel s_t]. \quad (4)$$

This integration ensures that long-term historical dependencies (captured by attention), segment-specific responsiveness (captured by FiLM), and recency/saturation dynamics (captured by x^e) jointly inform promotion decisions. A feed-forward network f_θ maps this representation to Q-values for all candidate promotions:

$$Q(s_t, a; \theta) = f_\theta(u_t). \quad (5)$$

The parameters θ are optimized using the standard DQN objective:

¹ The tendency of customers to respond more strongly to recent offers while discounting older ones.

² The diminishing marginal impact of repeated exposures to the same type of promotion.

$$\mathcal{L}(\theta) = \mathbb{E} \left[\left(r_t + \gamma \max_{a'} Q(s_{t+1}, a'; \theta^-) - Q(s_t, a_t; \theta) \right)^2 \right], \quad (6)$$

where θ^- denotes the parameters of a slowly updated target network.

This final stage ties together the preceding modules: the attention encoder identifies which historical events matter, FiLM modulates their influence according to customer segment, the exposure vector encodes recency and diminishing returns, and the Q-network integrates these signals into actionable promotion values. In this way, FAT-DQN provides a holistic architecture tailored to the sparse, heterogeneous, and saturation-prone nature of financial promotions.

Algorithm 1 FAT-DQN (FiLM-modulated, Attention-based Temporal DQN).

Require: Logged dataset $\mathcal{D} = \{(s_t, a_t, r_t, s_{t+1}, H_t, H_{t+1}, x_t^e, x_{t+1}^e, c)\}$, history length L , discount γ , target update period K , learning rate η

- 1: Initialize parameters θ (token encoder, self-attention, FiLM, Q-head); set target $\theta^- \leftarrow \theta$
- 2: **function** ENCODEHISTORY(H_t, L) $\triangleright H_t = \{(a_\tau, r_\tau, \tau)\}_{\tau=t-L}^{t-1}$
- 3: Truncate/pad H_t to length L (most recent first)
- 4: **for** $\tau = t - L, \dots, t - 1$ **do**
- 5: $e_\tau^a \leftarrow e^a(a_\tau)$ $\triangleright$ promotion embedding
- 6: $e_\tau^r \leftarrow e^r(r_\tau)$ $\triangleright$ reward projection
- 7: $e_\tau^t \leftarrow e^t(\tau)$ $\triangleright$ learnable temporal encoding
- 8: $z_\tau \leftarrow [e_\tau^a \parallel e_\tau^r \parallel e_\tau^t]$
- 9: **end for**
- 10: **return** $h \leftarrow \text{SelfAttn}(\{z_{t-L}, \dots, z_{t-1}\})$ $\triangleright$ context vector
- 11: **end function**
- 12: **function** FiLMMODULATE(h, c)
- 13: $e^c \leftarrow e^c(c)$ $\triangleright$ cluster embedding
- 14: $(\gamma_c, \beta_c) \leftarrow g(e^c)$ $\triangleright$ FiLM generator (linear layers)
- 15: **return** $\tilde{h} \leftarrow \gamma_c \odot h + \beta_c$
- 16: **end function**
- 17: **function** BUILDFEATURES(s_t, H_t, x_t^e, c, L)
- 18: $h \leftarrow \text{ENCODEHISTORY}(H_t, L)$; $\tilde{h} \leftarrow \text{FiLMMODULATE}(h, c)$
- 19: **return** $u_t \leftarrow [\tilde{h} \parallel x_t^e \parallel s_t]$ $\triangleright$ unified representation
- 20: **end function**

Offline training

- 21: **while** not converged **do**
- 22: Sample minibatch $\mathcal{B} \subset \mathcal{D}$; set $\mathcal{L} \leftarrow 0$
- 23: **for all** $(s_t, a_t, r_t, s_{t+1}, H_t, H_{t+1}, x_t^e, x_{t+1}^e, c) \in \mathcal{B}$ **do**
- 24: $u_t \leftarrow \text{BUILDFEATURES}(s_t, H_t, x_t^e, c, L)$; $\mathbf{q}_t \leftarrow f_\theta(u_t)$
- 25: $u_{t+1} \leftarrow \text{BUILDFEATURES}(s_{t+1}, H_{t+1}, x_{t+1}^e, c, L)$; $\mathbf{q}_{t+1}^+ \leftarrow f_{\theta^-}(u_{t+1})$
- 26: $y_t \leftarrow r_t + \gamma \max_{a'} \mathbf{q}_{t+1}^+[a']$
- 27: $\mathcal{L} \leftarrow \mathcal{L} + (y_t - \mathbf{q}_t[a_t])^2$
- 28: **end for**
- 29: Update $\theta \leftarrow \theta - \eta \nabla_\theta \frac{1}{|\mathcal{B}|} \mathcal{L}$ $\triangleright$ e.g., Adam
- 30: **if** iteration mod $K = 0$ **then**
- 31: $\theta^- \leftarrow \theta$
- 32: **end if**
- 33: **end while**

Inference / action selection

- 34: **function** SELECTACTION(s_t, H_t, x_t^e, c, L)
- 35: $u_t \leftarrow \text{BUILDFEATURES}(s_t, H_t, x_t^e, c, L)$; $\mathbf{q}_t \leftarrow f_\theta(u_t)$
- 36: **return** $\hat{a}_t \leftarrow \arg \max_a \mathbf{q}_t[a]$ $\triangleright$ apply masking if action constraints
- 37: **end function**

3.3 Reward design

We define promotion effectiveness using a peer-differenced pre–post change in assets under management (AUM) within the customer’s cluster. Let k denote the evaluation horizon (e.g., one month). For customer i in cluster c_i , the raw AUM change following a promotion at time t is

$$\Delta \text{AUM}_i(t; k) = \text{AUM}_i(t + k) - \text{AUM}_i(t - 1). \quad (7)$$

To isolate the incremental effect of the promotion, we subtract the average change among untreated peers in the same cluster:

$$\tilde{r}_i(t) = \Delta \text{AUM}_i(t; k) - \mathbb{E}[\Delta \text{AUM}_j(t; k) \mid c_j = c_i, \text{ no promo at } t]. \quad (8)$$

If multiple promotions are sent in month t , the uplift is distributed evenly across them:

$$r_i(t) = \frac{\tilde{r}_i(t)}{N_{\text{promo},i}(t)}, \quad N_{\text{promo},i}(t) \geq 1. \quad (9)$$

For customers without promotions, we assign a small constant reward $r_i(t) = \epsilon$ (with $\epsilon = 0.01$) to encode opportunity cost and discourage indiscriminate over-promotion.

To stabilize learning across actions with heterogeneous value scales, we normalize rewards per cluster–action pair:

$$r_i^{\text{norm}}(t) = \frac{r_i(t) - \mu_{c_i,a}}{\sigma_{c_i,a}}, \quad (10)$$

where $\mu_{c,a}$ and $\sigma_{c,a}$ denote the mean and standard deviation of unnormalized rewards for action a within cluster c .

This design provides several advantages. By differencing against peers within the same cluster, it controls for seasonal and market-wide fluctuations that might otherwise confound promotion effects. The division by the number of promotions in a given month prevents over-attribution when multiple offers are issued simultaneously. Finally, the use of per-action normalization yields scale-adjusted and comparable rewards across actions and clusters, facilitating more stable training and fair optimization.

A limitation of this reward design is its reliance on $\text{AUM}(t + k)$ to evaluate promotions initiated at time t , which introduces a form of **look-ahead bias**³. In our offline setting, however, this formulation is acceptable: the agent is trained only on information available up to t , while future outcomes are used solely to define retrospective reward signals. This enables counterfactual comparisons without leaking future information into the policy.

³ The use of future outcomes that are not observable at decision time.

To reduce noise and stabilize training, we analyze different horizons k to assess sensitivity, apply clipping to extreme outliers in $\tilde{r}_i(t)$, and revert to broader normalization strategies when cluster–action statistics are sparse. The resulting rewards should not be interpreted as unbiased causal estimates of treatment effect, but rather as variance-reduced proxies that are both scale-adjusted and comparable across actions, making them suitable for offline reinforcement learning and ranking under logged financial data.

3.4 Preprocessing and customer clustering

To define the discrete action space $\mathcal{A}$, we categorize promotions by prize type, including Alipay and WeChat coupons, media subscriptions, and car-owner benefits. Each promotion type is assigned a unique integer index, with a dedicated index ($a = 0$) reserved for the no-promotion case. This systematic mapping ensures that the action space is consistently defined and reproducible across the environment and Q-network.

Customer heterogeneity is incorporated through clustering. Specifically, K-Means [32] is applied to static demographic attributes and aggregated financial features to obtain $C = 12$ cluster identifiers, $c \in \{1, \dots, 12\}$. These clusters capture distinct customer profiles, such as high-AUM investors, consumption-active clients, or low-balance segments. The resulting cluster identifiers serve a dual role: they support stratified evaluation of policies across heterogeneous customer groups and, at the same time, provide embeddings $e^c(c)$ that act as conditioning signals for FiLM-based modulation within the Q-network. In this way, the value function adapts its estimation to segment-specific characteristics during both training and inference.

4 Experimental design

4.1 Model configurations

We benchmark the proposed FAT-DQN against several reinforcement learning baselines, all trained under identical hyperparameter settings and data splits to ensure comparability. Training is conducted on promotion logs from 2023–2024, while evaluation is performed on held-out 2025 data, enabling an assessment of out-of-sample generalization to future periods. All experiments are implemented in Python 3.10 with PyTorch 2.4.1, and executed on a high-performance computing cluster consisting of a Hygon C86 32-core CPU machine running the Kylin 10 Linux-based operating system, equipped with 512 GB of memory and 8 Hygon DCU GPUs. This environment ensures both reproducibility and computational feasibility.

for large-scale offline reinforcement learning experiments. The models are trained with the Adam optimizer (learning rate 1×10^{-6}), discount factor $\gamma = 0.99$, and ReLU activations, using 50,000 client samples per training run to approximate the logged distribution.

4.2 Baseline models

- **DQN**: A standard feedforward Deep Q-Network that uses only current customer features and cluster embeddings, without temporal history, attention, or FiLM modulation.
- **BCQ**: Batch-Constrained Q-learning [33], which constrains policy updates to remain close to the logged behavior distribution, mitigating extrapolation error in offline reinforcement learning.
- **CQL**: Conservative Q-learning [34], which penalizes overestimation of unseen actions, producing a more pessimistic but robust policy for offline evaluation.

4.3 Ablations and sensitivity analyses

To disentangle the contributions of each architectural component and to probe model sensitivity, we evaluate the following FAT-DQN variants:

- **SA-DQN (Static Attention DQN)**: Applies self-attention over static features (e.g., cluster ID, promotion type) without modeling sequential interaction histories.
- **GRU-DQN**: Replaces the multi-head attention encoder with a gated recurrent unit (GRU) [35], providing a recurrent-memory baseline.
- **AT-DQN (w/o FiLM)**: Removes FiLM-based cluster modulation to assess the role of segment-specific adaptation.
- **FA-DQN (w/o Temporal Encoding)**: Removes learnable temporal encodings to evaluate whether flexible timing improves promotion assignment.
- **History Length Sensitivity**: FAT-DQN is trained with history windows of 3, 6, and 12 months to quantify the value of longer temporal contexts.
- **Attention Head Sensitivity**: The number of attention heads in the multi-head encoder is varied (2, 3, 4, and 6) to study the trade-off between expressiveness and efficiency.

4.4 Evaluation metrics

We assess models along three complementary dimensions: ranking quality, counterfactual estimation, and robustness under noise perturbations.

4.4.1 Ranking quality

We evaluate how well policies rank promotional actions relative to realized outcomes. Normalized Discounted Cumulative Gain (NDCG@ k) [36] measures the relevance of the top- k recommended promotions, using logarithmic discounting to emphasize early ranks:

$$\text{NDCG@}k = \frac{1}{Z_k} \sum_{i=1}^k \frac{2^{y_i} - 1}{\log_2(i + 1)}, \quad (11)$$

where y_i is the realized outcome of the promotion at rank i and Z_k normalizes by the ideal DCG. Higher values indicate better prioritization of effective promotions.

Spearman's rank correlation coefficient [37] measures monotonic agreement between predicted Q-value rankings and realized outcomes:

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}, \quad (12)$$

where d_i is the difference in ranks and n the number of actions. Larger ρ indicates more faithful ranking alignment.

4.4.2 Counterfactual evaluation

Because policies are evaluated offline against logged data from a behavior policy π_b , we employ standard off-policy estimators.

Inverse Propensity Scoring (IPS) [38] reweights rewards by action probabilities under π and π_b :

$$\hat{V}_{\text{IPS}}(\pi) = \frac{1}{N} \sum_{i=1}^N \frac{\pi(a_i | s_i)}{\pi_b(a_i | s_i)} r_i. \quad (13)$$

Self-Normalized IPS (SNIPS) [39] stabilizes variance by renormalization:

$$\hat{V}_{\text{SNIPS}}(\pi) = \frac{\sum_{i=1}^N \frac{\pi(a_i | s_i)}{\pi_b(a_i | s_i)} r_i}{\sum_{i=1}^N \frac{\pi(a_i | s_i)}{\pi_b(a_i | s_i)}}. \quad (14)$$

Doubly Robust (DR) [40] augments IPS with a reward model $\hat{r}(s, a)$, ensuring unbiasedness if either component is correct:

$$\hat{V}_{\text{DR}}(\pi) = \frac{1}{N} \sum_{i=1}^N \left[\hat{r}(s_i, \pi(s_i)) + \frac{\pi(a_i | s_i)}{\pi_b(a_i | s_i)} (r_i - \hat{r}(s_i, a_i)) \right]. \quad (15)$$

Weighted Importance Sampling (WIS) [41] normalizes trajectory weights:

$$\hat{V}_{\text{WIS}}(\pi) = \sum_{i=1}^N \frac{w_i}{\sum_{j=1}^N w_j} R(\tau_i), \quad w_i = \prod_{t=1}^{T_i} \frac{\pi(a_t|s_t)}{\pi_b(a_t|s_t)}. \quad (16)$$

Weighted Per-Decision Importance Sampling (WPDIS) [41] extends this to each step:

$$\hat{V}_{\text{WPDIS}}(\pi) = \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^{T_i} \frac{w_{i,t}}{\sum_{j=1}^N w_{j,t}} \gamma^t r_{i,t}. \quad (17)$$

Finally, Effective Sample Size (ESS) [42] quantifies support:

$$\text{ESS} = \frac{\left(\sum_{i=1}^N w_i\right)^2}{\sum_{i=1}^N w_i^2}, \quad (18)$$

measuring how many independent samples the reweighted dataset effectively represents.

4.4.3 Robustness to reward noise

To test resilience under volatility and measurement error, we evaluate policies on perturbed rewards:

$$\tilde{r}_{i,t} = r_{i,t} + \varepsilon_{i,t}, \quad \varepsilon_{i,t} \sim \mathcal{N}(0, (\alpha\sigma_{\text{ref}})^2), \quad (19)$$

where σ_{ref} is the empirical standard deviation of rewards and $\alpha \in \{0.5, 1.0, 2.0\}$ scales the perturbation.

Choosing $\alpha \leq 2$ is motivated by sub-Gaussian concentration: 2σ covers roughly 95% of natural variability, offering a stringent yet realistic stress test. Values $\alpha = 0.5$ and $\alpha = 1.0$ represent moderate perturbations, while $\alpha = 2.0$ approximates adversarially noisy but still plausible financial volatility. Importantly, noise is injected only **during evaluation**, preserving stability of training dynamics.

5 Results and discussion

In this section, we present ranking quality metrics and counterfactual policy evaluation results, together with robustness tests under additive reward noise and analyses of attention weight distributions and FiLM-based segment conditioning, with all outcomes averaged across five random seeds to ensure statistical reliability and fair comparison across models.

5.1 Promotion ranking performance

Table 2 reports the ranking quality of all models in terms of NDCG@3, NDCG@5, and Spearman's ρ . Several trends emerge. FAT-DQN with four attention heads achieves the

Table 2 Ranking quality across models (NDCG@3, NDCG@5, and Spearman's ρ)

Model	NDCG@3	NDCG@5	Spearman's ρ
BCQ	0.7325	0.7724	0.1619
CQL	0.7684	0.7828	0.2867
DQN	0.6852	0.6988	0.1522
SA-DQN	0.6641	0.7097	0.1506
GRU-DQN	0.7576	0.7986	0.2373
AT-DQN	0.7450	0.7846	0.2112
FA-DQN	0.6464	0.7007	0.1072
FAT-DQN (2 heads)	0.7511	0.7881	0.2123
FAT-DQN (3 heads)	0.7621	0.7949	0.2172
FAT-DQN (4 heads)	0.7744	0.7978	0.2584
FAT-DQN (6 heads)	0.7611	0.7944	0.2437
FAT-DQN (history length 3)	0.7301	0.7677	0.1825
FAT-DQN (history length 6)	0.7446	0.7789	0.1937

highest NDCG@3 (0.7744), while GRU-DQN attains the strongest NDCG@5 (0.7986), highlighting the benefit of sequence modeling for capturing longer-term user interactions. CQL achieves the best rank correlation ($\rho = 0.2867$), reflecting its conservative regularization, which avoids large ranking errors. By contrast, FA-DQN, which removes temporal encodings, shows the weakest ranking performance, confirming that explicit timing features are crucial for promotion effectiveness.

Interestingly, the full FAT-DQN variants generally outperform the baselines, particularly in NDCG, demonstrating the value of combining self-attention with FiLM-based segment conditioning and temporal encoding. Increasing the number of attention heads from two to four improves ranking quality, but gains saturate at six heads, suggesting diminishing returns. Similarly, extending the history window from three to six months yields only minor improvements, indicating that shorter interaction sequences are sufficient for capturing promotional preferences. Overall, these results show that FiLM conditioning and temporal encodings are key for personalization, while multi-head attention improves expressiveness up to a point.

5.2 Counterfactual policy evaluation

Table 3 presents off-policy evaluation metrics across models. Here, the relative ordering of models diverges from the ranking results, reflecting differences between recommendation quality and policy stability. GRU-DQN achieves the highest WIS (3.7094), WPDIS (46.5554), IPS (1.8033), and SNIPS (1.4088), consistent with its ability to closely track logged action sequences. This alignment reduces variance in importance sampling and improves effective sample size (20.2%), but does not translate to the best ranking metrics. In contrast, CQL achieves the strongest doubly robust (DR) estimate (0.5748) and the best Spearman's ρ in Table 2,

Table 3 Off-policy evaluation metrics across models

Model	WIS	WPDIS	IPS	SNIPS	DR	ESS%
BCQ	1.9127	36.9006	0.2945	0.1964	0.2176	13.7
CQL	2.9089	27.5385	0.7649	0.3195	0.5748	16.4
DQN	1.3629	13.2479	0.2967	0.1982	0.2122	12.7
SA-DQN	1.9088	23.6132	0.3060	0.1826	0.4122	12.1
GRU-DQN	3.7094	46.5554	1.8033	1.4088	0.4553	20.2
AT-DQN	3.5063	26.5191	0.9054	0.3691	0.3286	22.8
FA-DQN	1.1137	12.2157	0.2854	0.1162	0.0872	10.4
FAT-DQN (2 heads)	3.5628	28.8670	1.2934	0.9735	0.4122	16.3
FAT-DQN (3 heads)	3.5748	31.3753	1.3698	1.0387	0.4386	16.1
FAT-DQN (4 heads)	3.6951	36.6259	1.4571	1.1552	0.4987	16.5
FAT-DQN (6 heads)	3.5129	33.6404	1.3597	1.0619	0.4736	11.4
FAT-DQN (history length 3)	2.9087	23.2496	0.3978	0.2259	0.2488	12.4
FAT-DQN (history length 6)	2.9556	25.8521	0.6583	0.2574	0.3596	12.7

consistent with its conservative training objective that favors stable value estimation at the expense of exploration.

The attention-based models exhibit complementary trade-offs. AT-DQN (without FiLM) effectively collapses to a more global policy that remains closer to the logging behavior, yielding the largest ESS (22.8%). However, without segment-specific adaptation its ranking quality lags behind FAT-DQN. FA-DQN, which removes learnable temporal encodings, performs poorly across all OPE metrics, underscoring the importance of modeling recency and seasonality effects. By contrast, FAT-DQN variants personalize actions via FiLM and therefore move further from the logging policy; this reduces ESS and can lower IS-based metrics (WIS, WPDIS, IPS, SNIPS), but improves ranking fidelity and maintains competitive model-based value (e.g., FAT-DQN (4 heads) achieves WPDIS (36.6259) and DR (0.4987)). These results reflect a personalization–stability trade-off: greater per-cluster adaptation increases off-policy variance while enhancing recommendation relevance.

Taken together, these results indicate that inconsistencies between ranking and OPE are not contradictions but reflect inherent trade-offs. GRU-DQN emphasizes trajectory mimicry, which benefits importance sampling and weighted policy estimates, while FAT-DQN emphasizes personalization via FiLM modulation and temporal encoding, which benefits ranking quality. CQL provides stability and conservative value estimation, but at the cost of reduced ranking flexibility. These complementary behaviors highlight the need to consider both ranking metrics and OPE when assessing promotion assignment models.

5.3 Robustness to reward noise

We further investigated the robustness of ranking performance under noisy reward signals. Additive Gaussian perturbations were introduced to the reward table, scaled by the empirical standard deviation of each cluster–action pair.

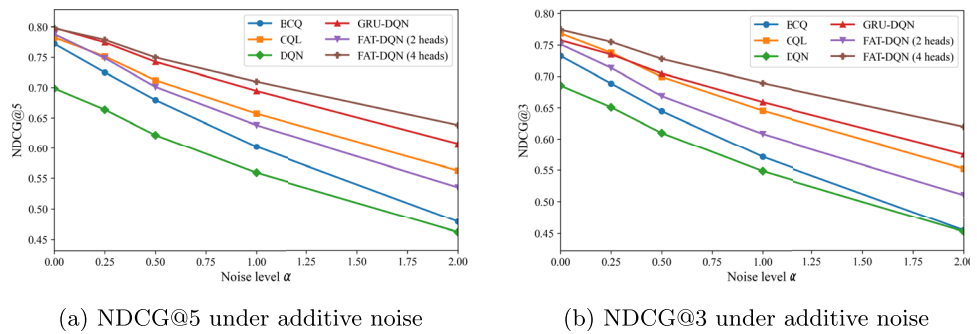
We then evaluated NDCG@3 and NDCG@5 at noise levels $\alpha \in \{0, 0.25, 0.5, 1.0, 2.0\}$. The results are shown in Fig. 2, with panel (a) reporting NDCG@5 and panel (b) reporting NDCG@3.

As expected, all models exhibit monotonic degradation as α increases. The degradation is smooth and approximately linear in α , reflecting the fact that additive noise reduces the effective signal-to-noise ratio proportionally to the inherent variability of each cluster–action pair. Among the baselines, DQN and BCQ are most sensitive, with NDCG@5 decreasing from 0.6988 and 0.7724 at $\alpha = 0$ to 0.4614 and 0.4789 at $\alpha = 2.0$, respectively, corresponding to relative drops of 34% and 38%. The same trend is observed for NDCG@3, where both models fall below 0.46 at the highest noise level. CQL shows intermediate robustness, decreasing from 0.7828 (NDCG@5) at $\alpha = 0$ to 0.5636 at $\alpha = 2.0$ (a 28% drop), and from 0.7684 (NDCG@3) to 0.5532 (a 28% drop).

In contrast, models incorporating temporal modeling and richer attention demonstrate stronger resilience. GRU-DQN maintains an NDCG@5 of 0.607 and an NDCG@3 of 0.576 at $\alpha = 2.0$, corresponding to relative drops of 24% and 24% from baseline, respectively. FAT-DQN with four heads is the most robust across all settings: its NDCG@5 decreases from 0.7978 to 0.6382 (a 20% drop), while its NDCG@3 decreases from 0.7744 to 0.6195 (a 20% drop). FAT-DQN with two heads, while stronger than DQN and BCQ, exhibits larger sensitivity than the four-head variant, falling to 0.5359 (NDCG@5) and 0.5107 (NDCG@3) at $\alpha = 2.0$. This confirms that additional attention heads provide a hedge against noisy or misleading signals by distributing weight more broadly across history.

These results demonstrate that robustness is not solely determined by baseline ranking quality, but also by the ability to integrate noisy historical signals. Models with temporal encoding and multi-head attention, particularly FAT-DQN (4 heads), maintain substantially higher fidelity under severe perturbations, whereas value-based baselines

Fig. 2 Robustness of ranking performance with additive Gaussian noise ($\alpha \in \{0, 0.25, 0.5, 1, 2\}$). Models with richer temporal encoding and multi-head attention are more robust



such as DQN and BCQ degrade sharply. The robustness trends observed here are consistent with the architectural characteristics of the models: those with more diffuse attention distributions tend to hedge across multiple historical signals, making them less brittle under noisy perturbations.

5.4 Attention weights analysis

In Table 4 we report the last-row attention weights for a single representative customer (Cluster 5, Customer A). Here, FAT-DQN with shorter history windows (e.g., 3 months) assigns nearly all weight to the most recent interaction (0.9683), revealing a strong recency bias. Extending the window to six or twelve months introduces secondary peaks, but the relative emphasis on recent history remains dominant. Different head configurations also produce markedly different allocation strategies: with two heads, attention is sharply concentrated on the final steps, whereas four or six heads distribute weight more evenly across earlier positions. AT-DQN shows similar recency dominance without FiLM modulation, while FA-DQN collapses to a uniform distribution, confirming its lack of temporal selectivity.

In contrast, Fig. 3 aggregates mean attention distributions across the full customer population. The population-level trends mirror the single-customer case but with smoother profiles. Models with fewer heads (e.g., two or three) yield sharper peaks around the most recent steps, highlighting strong recency sensitivity. As the number of heads increases, attention spreads more evenly, reflecting a hedging strategy that captures longer-range dependencies. This head-dependent trade-off suggests that few heads encourage highly focused but brittle solutions, while additional heads hedge across horizons, producing more stable but less interpretable allocations.

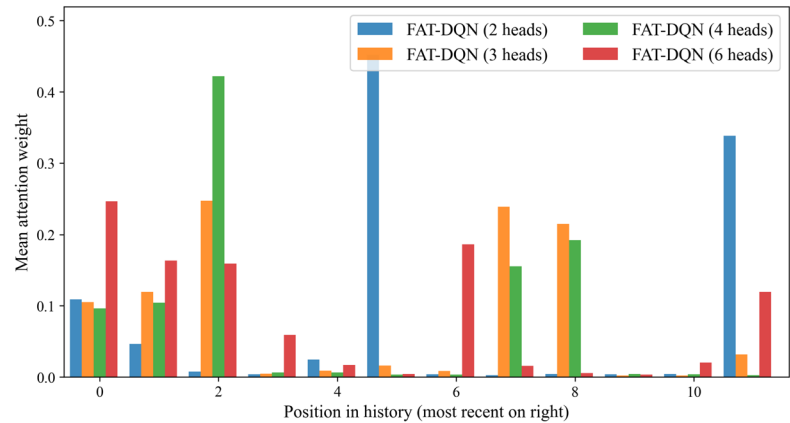
These divergent attention patterns do not result in large performance differences: models with peaked or diffuse profiles achieve comparable NDCG and OPE scores. This convergence indicates that promotion effectiveness signals are redundantly encoded across history, allowing different aggregation strategies to arrive at similar predictions. The poorest performance arises in FA-DQN, whose uniform weighting treats old and recent events equally, thereby diluting recency and saturation signals. This degradation is more severe than in models without history (e.g., DQN or BCQ),

Table 4 Attention weight distributions under different models and settings

Model	History	Attention Weights					
FAT-DQN	3m	0.0114	0.9683	0.0203	—	—	—
FAT-DQN	6m	0.2250	0.1110	0.0054	0.4449	0.1567	0.0570
FAT-DQN (2 heads)	12m	0.4462	0.0034	0.0025	0.0052	0.0009	0.0027
		0.4746	0.0309	0.0011	0.0032	0.0225	0.0067
FAT-DQN (3 heads)	12m	0.0290	0.0015	0.0014	0.2652	0.3000	0.0094
		0.0168	0.0066	0.0026	0.2356	0.1093	0.0226
FAT-DQN (4 heads)	12m	0.0012	0.0031	0.0025	0.2463	0.2065	0.0012
		0.0020	0.0035	0.0026	0.4818	0.0439	0.0054
FAT-DQN (6 heads)	12m	0.1650	0.0264	0.0030	0.0056	0.0183	0.2497
		0.0045	0.0123	0.0632	0.1600	0.1145	0.1775
AT-DQN	12m	0.4850	0.0004	0.0379	0.0004	0.0003	0.0006
		0.0064	0.0112	0.0001	0.0007	0.4559	0.0011
FA-DQN	12m	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833
		0.0833	0.0833	0.0833	0.0833	0.0833	0.0833

Values are normalized last-row attention weights for Customer A in Cluster 5 (oldest $\rightarrow$ most recent)

Fig. 3 Mean last-row attention weights across customers for FAT-DQN with different numbers of heads. More heads distribute attention more evenly across history, while fewer heads yield sharper peaks



since uniform attention forces the model to integrate outdated interactions. Overall, these findings confirm that temporal encodings and non-uniform attention are essential for effective promotion modeling, while raw attention weights should be interpreted with caution.

5.5 FiLM parameter analysis

In this section, we analyze the FiLM parameters learned by FAT-DQN to investigate how the model adapts its latent representations across heterogeneous customer segments. Figure 4 illustrates the average FiLM scaling (γ) and shifting (β) parameters across clusters and feature channels. The left panel (Fig. 4a) shows mean γ values, while the right panel (Fig. 4b) presents mean β values. Both maps reveal pronounced heterogeneity across clusters and channels. For example, clusters 0 and 2 exhibit strong positive scaling in multiple channels, whereas clusters such as 5 and 8 display consistently weaker activations. The β heatmap highlights similarly distinct offsetting patterns, with clusters 0 and 2 again showing large-magnitude adjustments, while smaller clusters (e.g., 8 and 11) remain relatively subdued. These structured modulations confirm that FiLM provides a compact, interpretable mechanism for encoding segment-specific adaptations that cannot be captured by shared network parameters alone.

Table 5 summarizes the ℓ_2 norms of FiLM parameters across clusters, quantifying the overall strength of modulation. Large clusters such as 0 (mean $\|\gamma\|_2 = 16.6$, mean $\|\beta\|_2 = 16.1$) and 2 (mean $\|\gamma\|_2 = 13.9$, mean $\|\beta\|_2 = 13.2$) exhibit the strongest modulation, while clusters 8 and 11 show much weaker strengths (both below 5). These quantitative differences indicate that high-value or behaviorally distinctive customer groups demand stronger representational adjustments, whereas more homogeneous or low-variance groups require only minimal modulation.

Taken together, these results highlight that FiLM modulation substantially enhances interpretability. The heatmaps expose which latent channels are emphasized or suppressed for different customer clusters, effectively providing a visual “signature” of each segment. The norm-based analysis offers a complementary, quantitative perspective, enabling practitioners to rank clusters by their degree of personalization and to identify which populations require highly tailored promotion strategies. Importantly, FiLM achieves this without the need to train separate models per cluster, preserving statistical robustness while offering segment-level transparency.

As a concrete illustration, Cluster 0, characterized by customers with high AUM and active engagement in wealth management products, exhibits the strongest FiLM strengths across all groups. Its heatmaps show both large positive and significantly negative γ/β values across channels, indicating

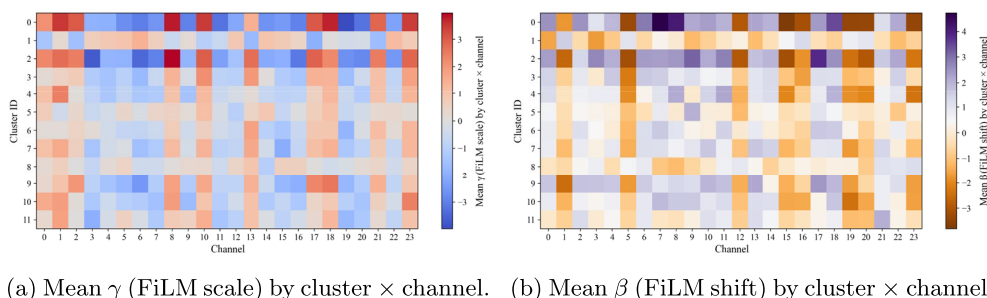


Fig. 4 FiLM modulation patterns across clusters and channels. Both scale (γ) and shift (β) parameters exhibit heterogeneous structures, reflecting customer-segment specific adaptations

Table 5 FiLM strength statistics by cluster (mean $\pm$ std over customers)

Cluster	Mean $\ \gamma\ _2$	Std $\ \gamma\ _2$	Mean $\ \beta\ _2$	Std $\ \beta\ _2$
0	16.5871	4.3347	16.0905	4.2049
1	7.2400	2.3033	6.4168	2.0414
2	13.9410	2.4902	13.2026	2.3583
3	7.6030	2.5013	7.3229	2.4092
4	6.1560	1.8588	8.2076	2.4783
5	3.2215	0.9389	3.8546	1.1234
6	4.7639	1.1375	4.9058	1.1714
7	6.0429	1.2176	5.9112	1.1911
8	2.5138	0.4868	3.5497	0.6874
9	8.4719	2.2831	8.7518	2.3585
10	6.5899	1.6524	6.8078	1.7071
11	5.2073	0.6320	4.1634	0.5053

that the model systematically amplifies informative promotion signals while suppressing less relevant ones. This behavior is consistent with domain expectations: financially affluent customers often display heterogeneous sensitivities to different promotion types, requiring substantial representational adjustment. By contrast, cluster8, characterized by higher liabilities and relatively uniform financial behaviors, displays consistently low modulation strengths and nearly flat channel responses. Such homogeneity reflects lower responsiveness to promotions, where additional adaptation is unnecessary. This contrast confirms that FiLM provides targeted amplification and suppression tailored to segment characteristics, yielding transparent and meaningful segment-level adaptations.

From a practical perspective, FiLM parameters thus serve as interpretable control knobs: γ reflects how strongly particular features are amplified or attenuated, while β encodes systematic offsets in responses. Examining these parameters allows analysts to diagnose which latent features drive promotional effectiveness in specific segments and to infer broader behavioral differences across clusters. In doing so, FiLM bridges the gap between opaque deep reinforcement learning policies and actionable business insights, enabling both compliance auditing and strategy refinement in real-world financial promotion systems.

5.6 Training complexity and efficiency

In this section, we analyze the convergence behavior and computational efficiency of all evaluated models. Figure 5 depicts training loss trajectories over 50,000 episodes. All models eventually reach stable regimes, but their convergence patterns differ. Standard DQN and GRU-DQN exhibit relatively smooth and monotonic loss reduction, consistent with their simpler architectures. In contrast, attention-based models (e.g., SA-DQN, AT-DQN, FAT-DQN) show higher

variance during early training, attributable to the additional complexity of learning temporal dependencies and attention weights. Within the FAT-DQN family, increasing the number of attention heads accelerates early loss reduction, indicating improved representational capacity, but also leads to higher variance and increased computational overhead. Conservative offline algorithms such as CQL demonstrate slower yet steadier declines in training loss, consistent with their pessimistic regularization objective. Overall, these results confirm that FAT-DQN achieves stable convergence while maintaining sufficient expressiveness to capture historical influence.

Training efficiency is a critical consideration for production deployment in banking environments, where both turnaround time and resource allocation are constrained. Table 6 reports wall-clock training times and peak GPU memory consumption for all DQN-based variants. The results highlight clear trade-offs: while feedforward and recurrent baselines remain efficient, attention-based variants impose additional costs, scaling with the number of heads. Nevertheless, the computational overhead remains tractable on modern GPUs, supporting the practicality of FAT-DQN for real-world financial applications.

The results illustrate three trends. First, multi-head self-attention incurs higher computational cost than feedforward or recurrent encoders, with training time and memory scaling approximately linearly in the number of heads. Second, GRU-DQN is moderately more expensive than feedforward baselines due to sequential recurrence, but remains more efficient than attention with large head counts. Third, the static-attention variant (SA-DQN) requires additional computation and memory relative to DQN, even without temporal modeling, reflecting the overhead of self-attention layers.

In terms of time complexity, DQN and SA-DQN scale with $\mathcal{O}(d^2)$ per layer where d is hidden size, GRU-DQN scales with $\mathcal{O}(Ld^2)$ where L is history length, and FAT-DQN scales with $\mathcal{O}(H L d^2)$ where H is the number of attention heads. Memory usage follows a similar pattern, dominated by attention key–query–value projections and intermediate activations. These differences are consistent with the empirical measurements in Table 6.

Overall, FAT-DQN remains tractable on modern GPUs, with the 4-head configuration requiring approximately $2.3\times$ the training time and $1.8\times$ the memory of a standard DQN, while providing substantially improved recommendation accuracy.

6 Conclusion

In this work we proposed FAT-DQN, a FiLM-modulated, attention-based temporal deep Q-network for optimizing financial promotion strategies. By jointly modeling

Fig. 5 Training loss convergence across baselines and FAT-DQN variants. Logarithmic scale is used for clarity across several orders of magnitude

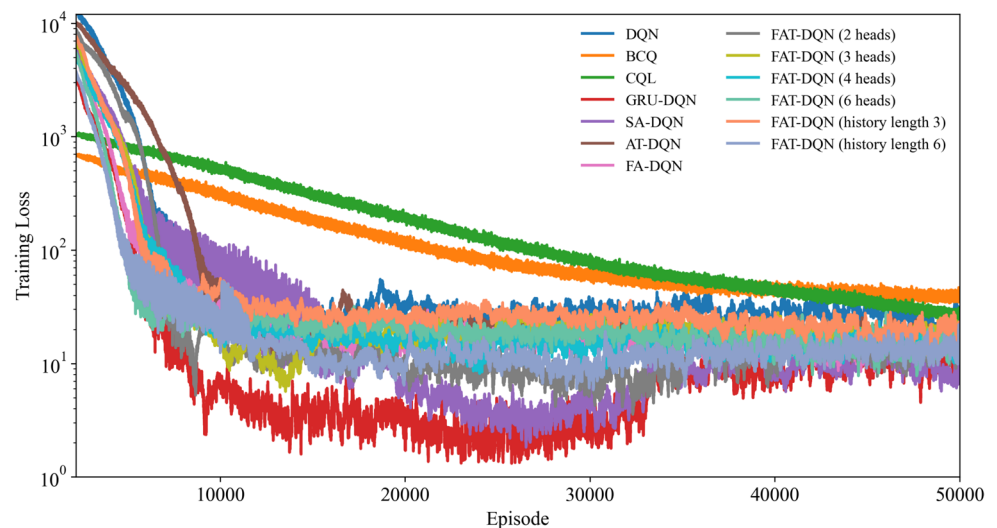


Table 6 Training time and GPU memory usage across DQN-based models (50,000 samples)

Model	Total Training Time	Peak GPU Memory
DQN	3h 14m	424 MB
SA-DQN	3h 44m	551 MB
GRU-DQN	4h 02m	594 MB
FAT-DQN (2 heads)	3h 52m	636 MB
FAT-DQN (3 heads)	4h 12m	678 MB
FAT-DQN (4 heads)	4h 36m	763 MB
FAT-DQN (6 heads)	5h 30m	933 MB

Times are reported per run and for the two-run protocol used in experiments

temporal purchase histories, customer heterogeneity, and promotion exposure effects, the framework is designed to address challenges directly relevant to real banking applications, where institutions must balance personalized marketing effectiveness with operational constraints and regulatory requirements.

Empirical results demonstrate that FAT-DQN improves promotion ranking quality and achieves balanced performance in counterfactual policy evaluation compared with standard offline reinforcement learning baselines. In practice, this implies more precise targeting of promotional offers, higher likelihood of customer engagement, and better resource allocation across marketing campaigns. The robustness experiments further highlight that multi-head attention and temporal encoding make the model resilient to noise in customer response data, a common reality in banking environments where transaction records are often incomplete or noisy.

At the same time, several limitations must be acknowledged. While FAT-DQN delivers stronger personalization, its reliance on offline logs and counterfactual evaluation cannot fully substitute for real-world evidence. Interpretability

remains a challenge: attention weights provide insights into temporal focus, but their variability across configurations limits their use as definitive explanations for business stakeholders. Moreover, conservative baselines such as CQL and recurrent baselines such as GRU-DQN offer greater stability in certain policy evaluation metrics, underscoring that no single method dominates across all dimensions.

For deployment in banks, future work should focus on rigorous validation through online A/B testing, which provides a gold-standard measure of uplift beyond logged-policy alignment. Integrating FAT-DQN into live promotion systems will require monitoring for distributional drift, fairness constraints, and compliance with marketing regulations. Beyond A/B testing, hybrid approaches that combine FAT-DQN's segment conditioning with causal inference or uplift modeling may further strengthen its ability to attribute promotion effects. Another important direction is designing adaptive mechanisms to trade off interpretability, robustness, and effectiveness depending on campaign requirements.

In summary, FAT-DQN offers a promising path toward intelligent, robust, and personalized promotion in banking. Its architectural innovations enable banks to better capture customer dynamics, adapt across segments, and maintain stability under noisy conditions. When combined with ongoing evaluation through A/B testing and fairness-aware monitoring, such models have the potential to transform marketing strategies into more data-driven, efficient, and customer-centric systems.

Acknowledgements We gratefully acknowledge Qilu Bank and its Mathematical Finance Laboratory for providing access to confidential promotion data and high-performance computing resources. Their support in enabling both data availability and computational infrastructure was essential for conducting this research.

Compliance and Ethics

The data used in this study consist of masked customer transaction and asset records from Qilu Bank, a commercial bank headquartered in Jinan, Shandong, China. Due to contractual confidentiality obligations and privacy regulations, the dataset is not publicly available. All data were anonymized and analyzed under appropriate institutional agreements and ethical guidelines. No personal identifiers were used, and no individuals were contacted or affected by the study.

References

- Sutton RS, Barto AG (2018) Reinforcement Learning: An Introduction. A Bradford Book, Cambridge, MA, USA
- Watkins CJ, Dayan P (1992) Q-learning. *Mach Learn* 8(3–4):279–292. <https://doi.org/10.1007/BF00992698>
- Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, Petersen S, Beattie C, Sadik A, Antonoglou I, King H, Kumaran D, Wierstra D, Legg S, Hassabis D (2015) Human-level control through deep reinforcement learning. *Nature* 518(7540):529–533. <https://doi.org/10.1038/nature14236>
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. *CoRR abs/1706.03762*. [arXiv:1706.03762](https://arxiv.org/abs/1706.03762)
- Perez E, Strub F, Vries H, Dumoulin V, Courville AC (2017) Film: Visual reasoning with a general conditioning layer. *CoRR abs/1709.07871*. [arXiv:1709.07871](https://arxiv.org/abs/1709.07871)
- Hausknecht MJ, Stone P (2015) Deep recurrent q-learning for partially observable mdps. *CoRR abs/1507.06527*. [arXiv:1507.06527](https://arxiv.org/abs/1507.06527)
- Sorokin I, Seleznev A, Pavlov M, Fedorov A, Ignateva A (2015) Deep attention recurrent q-network. *CoRR abs/1512.01693*. [arXiv:1512.01693](https://arxiv.org/abs/1512.01693)
- Kang W, McAuley JJ (2018) Self-attentive sequential recommendation. *CoRR abs/1808.09781*. [arXiv:1808.09781](https://arxiv.org/abs/1808.09781)
- Sun F, Liu J, Wu J, Pei C, Lin X, Ou W, Jiang P (2019) Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. *CoRR abs/1904.06690*. [arXiv:1904.06690](https://arxiv.org/abs/1904.06690)
- Moody J, Saffell M (2001) Learning to trade via direct reinforcement. *IEEE Trans Neural Netw* 12(4):875–889. <https://doi.org/10.1109/72.935097>
- Deng Y, Bao F, Kong Y, Ren Z, Dai Q (2017) Deep direct reinforcement learning for financial signal representation and trading. *IEEE Trans Neural Netw Learn Syst* 28(3):653–664. <https://doi.org/10.1109/TNNLS.2016.2522401>
- Gu F, Jiang Z, García-Fernández F, Stefanidis A, Su J, Li H (2025) MTS: A Deep Reinforcement Learning Portfolio Management Framework with Time-Awareness and Short-Selling. [arXiv:2503.04143](https://arxiv.org/abs/2503.04143)
- Sun R, Stefanidis A, Jiang Z, Su J (2024) Combining transformer based deep reinforcement learning with black-litterman model for portfolio optimization. [arXiv:2402.16609](https://arxiv.org/abs/2402.16609) [q-fin.PM]
- Ren X, Sun R, Jiang Z, Stefanidis A, Liu H, Su J (2025) Time series is not enough: financial transformer reinforcement learning for portfolio management. *Neurocomputing* 647:130451. <https://doi.org/10.1016/j.neucom.2025.130451>
- Sun W, Liu Z, Yuan C, Zhou X, Pei Y, Wei C (2025) Rcsan: residual-enhanced channel-spatial attention network for stock price forecasting. *Sci Rep* 15:21800. <https://doi.org/10.1038/s41598-025-06885-y>
- Hadizadeh A, Tarokh MJ, Ghazani MM (2025) A novel transformer-based dual attention architecture for the prediction of financial time series. *J King Saud Univ—Comput Inform Sci* 37:(72). <https://doi.org/10.1007/s44443-025-00045-y>
- Zhang Q, Qin C, Zhang Y, Bao F, Zhang C, Liu P (2022) Transformer-based attention network for stock movement prediction. *Expert Syst Appl* 202:117239. <https://doi.org/10.1016/j.eswa.2022.117239>
- Silva L, Freitas Fonseca G, Castro PAL (2024) Transformers and attention-based networks in quantitative trading: a comprehensive survey. In: Proceedings of the 5th ACM international conference on AI in Finance (ICAIF '24), pp 822–830. Association for computing machinery, New York, NY, USA. <https://doi.org/10.1145/3677052.3698684>
- Zhang Z, Chen B, Zhu S, Langrené N (2025) Quantformer: from attention to profit with a quantitative transformer trading strategy. [arXiv:2404.00424](https://arxiv.org/abs/2404.00424)
- Hazenbert T, Ma Y, Ziahari SSM, Rijswijk M (2025) Multi-Agent Reinforcement Learning for Dynamic Pricing in Supply Chains: Benchmarking Strategic Agent Behaviours under Realistically Simulated Market Conditions. [arXiv:2507.02698](https://arxiv.org/abs/2507.02698)
- Wang Y, Ge Y, Li L, Chen R, Xu T (2020) Offline Meta-level Model-based Reinforcement Learning Approach for Cold-Start Recommendation. [arXiv:2012.02476](https://arxiv.org/abs/2012.02476)
- Li L, Chu W, Langford J, Schapire RE (2010) A contextual-bandit approach to personalized news article recommendation. *CoRR abs/1003.0146*. [arXiv:1003.0146](https://arxiv.org/abs/1003.0146)
- Chen H, Dai X, Cai H, Zhang W, Wang X, Tang R, Zhang Y, Yu Y (2018) Large-scale Interactive Recommendation with Tree-structured Policy Gradient. [arXiv:1811.05869](https://arxiv.org/abs/1811.05869)
- Zhao X, Gu C, Zhang H, Yang X, Liu X, Tang J, Liu H (2021) DEAR: Deep Reinforcement Learning for Online Advertising Impression in Recommender Systems. [arXiv:1909.03602](https://arxiv.org/abs/1909.03602)
- Oroojlooy A, Nazari M, Hajinezhad D, Silva J (2020) Attend-light: Universal attention-based reinforcement learning model for traffic signal control. *CoRR abs/2010.05772*. [arXiv:2010.05772](https://arxiv.org/abs/2010.05772)
- Zhang D, Hao X, Liang L, Liu W, Qin C (2022) A novel deep convolutional neural network algorithm for surface defect detection. *J Comput Des Eng* 9(5):1616–1632. <https://doi.org/10.1093/jcde/qwac071>. <https://arxiv.org/abs/academic.oup.com/jcde/article-pdf/9/5/1616/46878068/qwac071.pdf>
- Qin C, Ran X, Zhang D (2025) Unsupervised image stitching based on generative adversarial networks and feature frequency awareness algorithm. *Appl Soft Comput* 183:113466. <https://doi.org/10.1016/j.asoc.2025.113466>
- Zhang D, Wang Y, Meng L, Yan J, Qin C (2024) Adaptive critic design for safety-optimal ftc of unknown nonlinear systems with asymmetric constrained-input. *ISA Trans* 155:309–318. <https://doi.org/10.1016/j.isatra.2024.09.018>
- Qin C, Jiang K, Wang Y, Zhu T, Wu Y, Zhang D (2025) Event-triggered h ∞ control for unknown constrained nonlinear systems with application to robot arm. *Appl Math Model* 144:116089. <https://doi.org/10.1016/j.apm.2025.116089>
- Zhou G, Zhu X, Song C, Fan Y, Zhu H, Ma X, Yan Y, Jin J, Li H, Gai K (2018) Deep interest network for click-through rate prediction. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. KDD '18, pp 1059–1068. Association for computing machinery, New York, NY, USA. <https://doi.org/10.1145/3219819.3219823>
- Choi J, Lee B, Zhang B (2017) Multi-focus attention network for efficient deep reinforcement learning. *CoRR abs/1712.04603*. [arXiv:1712.04603](https://arxiv.org/abs/1712.04603)
- MacQueen J (1967) Multivariate observations. In: Proceedings of the 5th Berkeley symposium on mathematical statistics and probability, vol 1, pp 281–297
- Fujimoto S, Meger D, Precup D (2019) Off-Policy Deep Reinforcement Learning without Exploration. [arXiv:1812.02900](https://arxiv.org/abs/1812.02900)

34. Kumar A, Zhou A, Tucker G, Levine S (2020) Conservative Q-Learning for Offline Reinforcement Learning. [arXiv:2006.04779](https://arxiv.org/abs/2006.04779)
35. Chung J, Gülçehre Ç, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR* **abs/1412.3555**. [arXiv:1412.3555](https://arxiv.org/abs/1412.3555)
36. Järvelin K, Kekäläinen J (2002) Cumulated gain-based evaluation of ir techniques. *ACM Trans Inf Syst* 20(4):422–446. <https://doi.org/10.1145/582415.582418>
37. Spearman C (2010) The proof and measurement of association between two things. *Int J Epidemiol* 39(5):1137–1150. <https://doi.org/10.1093/ije/dyq191>. <https://arxiv.org/abs/academic.oup.com/ije/article-pdf/39/5/1137/18481215/dyq191.pdf>
38. Strehl AL, Langford J, Kakade SM (2010) Learning from logged implicit exploration data. *CoRR* **abs/1003.0120**. [arXiv:1003.0120](https://arxiv.org/abs/1003.0120)
39. Swaminathan A, Joachims T (2015) Counterfactual risk minimization: Learning from logged bandit feedback, vol. abs/1502.02362. [arXiv:1502.02362](https://arxiv.org/abs/1502.02362)
40. Dudík M, Langford J, Li L (2011) Doubly robust policy evaluation and learning, vol. abs/1103.4601. [arXiv:1103.4601](https://arxiv.org/abs/1103.4601)
41. Precup D, Sutton RS, Singh SP (2000) Eligibility traces for off-policy policy evaluation. In: Proceedings of the seventeenth international conference on machine learning. ICML '00, pp 759–766. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA
42. Kish L (1965) Survey Sampling. Wiley. <https://books.google.com/books?id=94gDzwEACAAJ>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.