CAFE-RL: COUNTERFACTUAL AUGMENTED REINFORCEMENT LEARNING FOR MECHANISM-AWARE ONBOARDING FRAUD DETECTION IN E-COMMERCE

Linfeng Cao[1,2*], **Hang Yin**[1], **Yang Zhao**[1], **Xinze Guan**[1], **Qiang Wang**[1], **Ming Ouyang**[1]
[1]eBay Inc.
[2]Department of Computer Science and Engineering, The Ohio State University
cao.1378@osu.edu, {hangyin,yzhao5,xiguan,qiwang,mouyang}@ebay.com

ABSTRACT

Buyer Onboarding Fraud Detection (BOFD), defined as fraud detection at the early stages of user registration and first-transaction screening, is a strategic decision-making problem that defines a platform intervention mechanism and shapes user participation, adversarial behavior, and long-term outcomes such as Gross Merchandise Value (GMV). Existing systems typically optimize these checkpoints independently, leading to suboptimal trade-offs between fraud prevention and legitimate user retention. In this study, we model BOFD as a two-stage Markov Decision Process that captures sequential platform decisions and their long-term effects. Build on this, we propose **CAFE-RL**, a counterfactual augmented offline reinforcement learning framework that learns a unified decision policy from historical logs. To address the strategic and statistical bias induced by deployed mechanisms, CAFE-RL introduces a counterfactual augmentation scheme that constructs complementary transitions, ensuring full state–action coverage. In addition, we propose a hybrid reinforcement–contrastive objective that combines conservative Q-learning with supervised contrastive losses over factual–counterfactual pairs, providing stronger supervision and stabilizing policy convergence. Experiments on large-scale eBay data demonstrate that jointly optimizing early-stage interventions significantly outperforms deployed checkpoint-wise mechanisms, reducing fraud while preserving legitimate transactions.

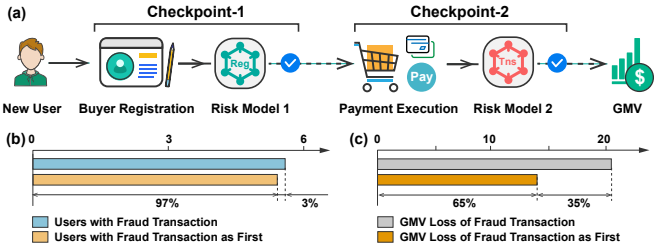


Figure 1: (a) Buyer Onboarding Fraud Detection (BOFD): Sequential fraud detection for new users at registration (Checkpoint-1) and first transaction (Checkpoint-2). (b, c) Fraudulent users and associated GMV loss within 30 days after registration, highlighting the impact of fraudulent first transactions. *Note: Data are anonymized, sampled, and shown only for relative comparison.*

1 INTRODUCTION

Fraud detection is a central challenge for modern e-commerce platforms (e.g., Amazon, eBay, Taobao), where security and Gross Merchandise Value (GMV) must be balanced. As marketplaces expand, malicious registrations Breuer et al. (2020); Li et al. (2022), account takeovers Lin et al. (2022), and fraudulent transactions Lu et al. (2022); Liu et al. (2021) pose serious risks. While fraud

*Work done during an internship at eBay.

causes direct losses and erodes user trust, overly conservative policies that block legitimate users reduce GMV and hinder platform growth, making this trade-off a fundamental strategic challenge.

Fraud-related interventions occur throughout the user lifecycle, but registration approval and first-transaction screening are the earliest and most consequential. As shown in Fig. 1(a), these checkpoints define the platform's initial access and risk-control rules and strongly influence downstream GMV. Empirical analysis over a 30-day window shows that most fraudulent users act at their first transaction, which accounts for the majority of GMV losses (Fig. 1(b,c)). We therefore focus on this two-stage onboarding setting and formalize it as *Buyer Onboarding Fraud Detection* (BOFD).

In practice, BOFD is typically handled by separate risk models for registration Liang et al. (2021); Zhang et al. (2023); Kondeti et al. (2020); Agarwal et al. (2019) and transactions Lu et al. (2022); Liu et al. (2021); Singh et al. (2023); Zhou et al. (2024). While effective locally, this fragmented design ignores the sequential dependence between decisions: early rejections eliminate both fraud risk and legitimate future value, whereas approvals without anticipating downstream risk can incur substantial losses. As a result, *optimizing checkpoints independently often fails to achieve global optimality in terms of secure GMV.* From a mechanism design and strategic decision-making perspective, BOFD concerns how platform intervention rules should be jointly optimized over time. We therefore formulate BOFD as a two-stage Markov Decision Process (MDP) and study it as a sequential decision problem, where early actions shape downstream states and long-term rewards. Reinforcement learning (RL) provides a natural framework for learning such policies.

However, applying RL to BOFD introduces unique challenges: **(1)** *Exploration constraints.* Online exploration is infeasible, as exploratory interventions (e.g., approving fraudulent users) incur unacceptable operational risks, requiring policies to be learned purely from historical logs. **(2)** *Mechanism-induced bias.* Logged data reflect outcomes only for actions taken by previously deployed mechanisms, while the outcomes of unobserved actions (e.g., declined users) are missing, leading to overestimation and poor generalization. **(3)** *Weak decision discrimination.* value-based offline RL often suffers from unstable convergence and limited discriminative power between effective and ineffective actions, as it relies exclusively on bootstrapped targets, making it prone to suboptimal policies. To overcome these challenges, we propose **CAFE-RL** (**C**ounterfactual **A**ugmented **F**ramework for **E**-commerce fraud detection with **R**einforcement **L**earning), a novel offline RL framework for mechanism-level optimization in BOFD. CAFE-RL introduces two key innovations:

- **Counterfactual augmentation:** For each logged transition, we construct a complementary counterfactual counterpart that represents the outcome had the alternative intervention been applied. This augmentation enables full action coverage of platform decisions, thereby mitigates overestimation caused by mechanism-induced selection bias.
- **Hybrid reinforcement–contrastive learning:** Based on the constructed factual–counterfactual pairs, we propose a novel hybrid learning framework that augments the RL loss with a supervised contrastive loss. This additional supervisory signal encourages the policy to distinguish between effective and ineffective actions, complementing value-based learning, stabilizing convergence, and improving overall decision quality.

We evaluate CAFE-RL on real-world data from a large-scale e-commerce platform, covering millions of new users over one month. Results show that CAFE-RL outperforms traditional separate models at both registration and transaction stages, reducing false positives and preserving legitimate GMV while maintaining strong fraud detection accuracy. These findings highlight the effectiveness of treating BOFD as a sequential strategic decision problem and using RL for optimizing platform mechanisms in high-stakes environments.

2 Problem Formulation

We study *Buyer Onboarding Fraud Detection* (BOFD) in e-commerce platforms, which governs how platforms design and optimize intervention mechanisms at the earliest stages of the user lifecycle. BOFD consists of two sequential decision checkpoints, *registration* and the *first transaction*, as shown in Fig. 1(a). Decisions at these checkpoints jointly determine platform security and long-term marketplace outcomes such as Gross Merchandise Value (GMV). Our goal is to design a decision policy that strategically balances fraud prevention and GMV preservation across these checkpoints.

Table 1: Reward function specification for BOFD in the MDP, defined for any user u.

Checkpoint	Registration Stage ($t = 1$)		Transaction Stage ($t = 2$)		
Action	0: "approve"		0: "approve"		1: "decline"
Fraud	✗	✓	✗	✓	✗ / ✓
Reward	0	$-c$	$\alpha \cdot \log(g(u))$	$-\log(g(u))$	$-\epsilon$

2.1 BOFD AS A SEQUENTIAL DECISION PROCESS

We formulate BOFD as a two-stage strategic decision-making problem modeled by a Markov Decision Process (MDP) $(\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \gamma)$. Let $\mathcal{U}$ denote the set of users. Each component of the MDP captures a key element of the platform intervention mechanism:

State space $\mathcal{S}$. A continuous state space representing user information available to the platform at each checkpoint. For a user $u \in \mathcal{U}$ at stage t, the state $s_{u,t}$ encodes profile features, behavioral signals, contextual attributes, and other relevant information used for decision-making.

Action space $\mathcal{A}$. A binary intervention action at each checkpoint,

$$a \in \mathcal{A} = \{0, 1\}, \quad a = 0 : approve,\ a = 1 : decline,$$

which specifies whether the platform grants or denies access at the current stage.

Transition dynamics $\mathbb{P} : \mathcal{S} \times \mathcal{A} \to \Omega(\mathcal{S})$, where $\Omega(\mathcal{S})$ denotes the set of probability distributions over state space $\mathcal{S}$. The transition kernel $\mathbb{P}(s' \mid s, a)$ models how platform interventions shape downstream user states. In our problem, we define declining a user terminates the onboarding process deterministically: $\mathbb{P}(s_{\text{term}} \mid s, a = 1) = 1$, where s_{term} is an absorbing terminal state. Approval decisions allow the process to proceed to the next checkpoint, where further outcomes are realized.

Reward function r: maps each state–action pair (s, a) to an immediate reward, encoding the platform's strategic objective of balancing security risks against economic value. Let $\epsilon > 0$ be a small penalty to discourage unnecessary rejections, $c \gg \epsilon$ quantify the risk of fraudulent registrations[1], $g(u)$ denote the GMV of user u's first transaction, and $\alpha \in (0, 1]$ represent the platform's conversion rate[2]. We summarize the reward function in Table 1.

Discount factor γ: Since BOFD consists of two closely coupled stages, we set $\gamma = 1$ to value registration and transaction outcomes equally.

Under this formulation, the platform seeks a policy $\pi(a \mid s)$ that maximizes expected cumulative reward across both checkpoints, yielding a globally optimized intervention mechanism.

2.2 OFFLINE SETTING

BOFD operates in a high-stakes environment where online exploration is infeasible: exploratory approvals may admit fraudulent users, while unnecessary rejections reduce legitimate participation. Consequently, policy optimization must be conducted in an offline manner using historical logs by previously deployed mechanisms. Formally, we are given a dataset $\mathcal{D} = (s, a, r, s')$ collected under existing checkpoint policies. These logs exhibit mechanism-induced selection bias, as outcomes are observed only for actions historically taken (e.g., approved users), while outcomes of unchosen actions remain unobserved. This presents a central challenge for learning strategic intervention policies, as naively optimizing on $\mathcal{D}$ can lead to *overestimation* on unseen state–action pairs.

3 CAFE-RL: A UNIFIED RL FRAMEWORK FOR BOFD

To address the aforementioned challenges, we propose **CAFE-RL** (**C**ounterfactual **A**ugmented **F**ramework for **E**-commerce fraud detection with **R**einforcement **L**earning), a unified reinforcement learning framework for *mechanism-level* decision optimization in BOFD. As shown in Fig. 2,

[1]The penalty c is calibrated by domain experts to reflect the risk of fraudulent users traffic in the platform.

[2]Most e-commerce platforms (e.g., eBay, Amazon) operate as intermediaries rather than direct sellers. Thus, only a fraction of GMV contributes to actual revenue.

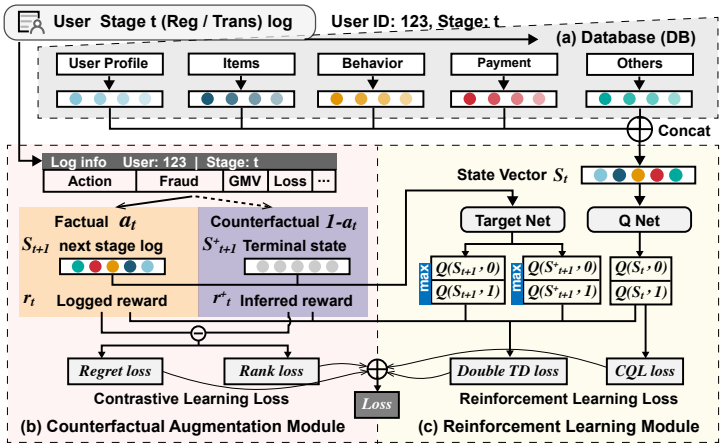


Figure 2: Overview of the CAFE-RL framework for BOFD. (a) **Database (DB) module**: retrieves user profile information and encodes it into state embeddings. (b) **Counterfactual Augmentation module**: constructs factual and counterfactual transition pairs to enable full state–action space coverage and provide supervised contrastive signals under offline data. (c) **Reinforcement Learning module**: learns a unified intervention policy using Double DQN with a conservative Q-learning (CQL) penalty, performing updates over the full action space to mitigate overestimation.

CAFE-RL consists of three components: (a) a *Database (DB) module* that retrieves and encodes user-related information into state representations, (b) a *Counterfactual Augmentation module* that constructs alternative outcomes for unobserved interventions to improve action coverage and provides supervised contrastive signals, and (c) a *Reinforcement Learning module* that learns a unified intervention policy using Double DQN with conservative Q-learning (CQL) and supervised loss penalties. Collectively, these components stabilize offline policy learning from logs under previously deployed mechanisms and enable the joint optimization of registration and transaction decisions as a single sequential strategic decision problem, rather than independent checkpoint-wise classifiers.

3.1 Counterfactual World: Learning over the Full State–Action Coverage

Offline logs in BOFD are inherently *sparse and biased*, as historical platform policies predominantly approve users, resulting in limited coverage of the state–action space. As shown in Fig. 3(a), standard offline RL learns solely from logged transitions $\{(s, a, r, s')\}$ and cannot evaluate unexplored interventions, which are often critical for policy improvement Levine et al. (2020); Fujimoto et al. (2019).

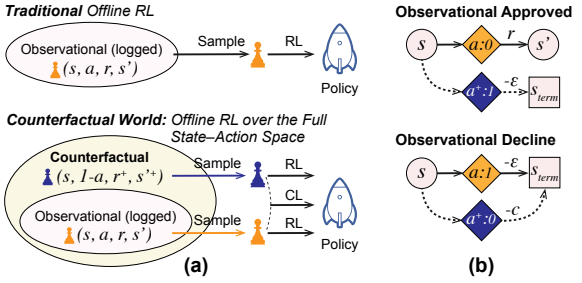


Figure 3: (a) Traditional offline RL only learns from logged transitions, while our counterfactual formulation augments them with complementary actions enabling full state–action coverage. (b) Counterfactual augmentation for BOFD: each logged transition is paired with its complementary action to construct a surrogate MDP. Circles = states, diamonds = actions, solid arrows = logged, dashed = counterfactual.

This mechanism-induced partial feedback motivates a counterfactual perspective: rather than modeling only the observational distribution, we seek to reason over the *counterfactual distribution* that includes outcomes of unobserved actions. Concretely, for each logged transition (s, a, r, s'), we augment its complementary tuple $(s, 1 - a, r^+, s'^+)$, thereby enabling learning over the full state–action coverage and direct comparison between alternative interventions. The augmentation rule is illustrated in Fig. 3(b) and described below.

Case 1: Observational *Approve*. Consider a logged transition (s, a, r, s') with the *approve* action ($a = 0$). The corresponding counterfactual outcome is straightforward to derive. Owing to the *deterministic* transition kernel and reward function associated with the *decline* action (see Transition dynamics and Table 1 in Sec 2.1), the counterfactual next state distribution and reward are given by

$$\mathbb{P}(s'^{+} = s_{\text{term}} \mid s, a^{+} = 1) = 1, \qquad r^{+} = r(s, a^{+}) = -\epsilon.$$

Since both the transition and reward of the *decline* action are fully specified, the constructed surrogate MDP in this case coincides exactly with the underlying true MDP. The corresponding structure for observational *approve* logs is shown in Fig. 3(b, top).

Case 2: Observational *Decline*.

For users declined by the platform's intervention mechanism, a fundamental challenge is *reject inference* (Smith & Elkan, 2004): their outcomes under approval (i.e., whether they would eventually commit fraud or behave legitimately), is inherently unobservable by design. This lack of counterfactual feedback is an inherent consequence of the deployed mechanism and prevents reliable estimation of alternative outcomes from logged data. To address reject inference in practice, we adopt a conservative modeling choice and *treat system-declined users as negative samples (fraud).*

Remark 1 (Conservative modeling via system design)**.** *This conservative modeling aligns with the operational design of industrial risk-control pipelines (e.g., eBay), where intervention mechanisms apply deliberately high decision thresholds and trigger declines only under strong fraud signals or critical information inconsistencies. As a result, declined actions are designed to achieve near-perfect precision, justifying their treatment as negative samples under a risk-averse mechanism.*

Remark 2 (Robustness to imperfect precision)**.** *Importantly, our MDP formulation and learning framework do not require perfect decline precision. Even when the existing intervention mechanism incurs false declines, the proposed approach can still improve long-term outcomes by optimizing downstream decisions for users who pass initial screening and mitigating residual fraud risks. This robustness to imperfect precision is empirically demonstrated in our experiments.*

Under the conservative modeling above, we then consider a logged transition (s, a, r, s') with the *decline* action ($a = 1$). Please note that the *declines* logs are very sparse over the dataset. Since observational *declines* are always labeled as fraudulent users, the counterfactual outcome of approving such a user can be derived directly. In this case, the immediate reward is defined as $r^{+} = -c$, reflecting the penalty of admitting a fraudulent account (see Table 1). For the next state, although approval would nominally lead to the transaction stage, the fraud label implies that no legitimate transaction path exists. We therefore terminate the trajectory: $\mathbb{P}(s'^{+} = s_{\text{term}} \mid s, a^{+} = 0) = 1$. This construction ensures that the surrogate MDP captures the true cost of approving fraudulent users while maintaining a well-defined termination condition, as illustrated in Fig. 3(b, bottom).

Claim 1 (Preservation of Optimality)**.** *The counterfactual surrogate MDP preserves the optimal policy of the underlying true MDP while providing full action coverage in BOFD* [3].

Proof Sketch. For observational *approve* logs, the counterfactual surrogate MDP coincides with the true MDP, as the complementary *decline* branch is deterministic in both transition and reward. For observational *decline*, the surrogate approximates the counterfactual *approve* branch with penalty $-c$ and termination, but the action ranking remains unchanged since $-\epsilon > -c$. Thus, in both cases, the counterfactual surrogate MDP maintains the same optimal policy as the underlying true MDP.

We learn policy under the augmented dataset $\mathcal{D}^{*} = \{(s, a, r, s', a^{+}, r^{+}, s'^{+})\}$, which mitigates extrapolation error and overestimation by exposing the learner to both actions at each state. These factual–counterfactual pairs also serve as the foundation for both *Q-learning* and the *supervised contrastive learning* in CAFE-RL.

3.2 Policy Learning in CAFE-RL

To ensure stable and effective learning, CAFE-RL optimizes a joint objective consisting of two complementary components: (1) a *Q-learning objective* for offline reinforcement learning, and (2) a *supervised contrastive objective* that leverages factual–counterfactual pairs to provide an additional

[3] Full action coverage since for every observed state, both actions (*approve* and *decline*) are represented through either factual or counterfactual transitions.

action-level supervisory signal. Together, these enable CAFE-RL to achieve robust policy convergence while mitigating extrapolation error and overestimation.

3.2.1 DOUBLE CONSERVATIVE Q-LEARNING.

Our counterfactual augmentation provides full *action* coverage at the *logged states*. However, offline RL can still suffer from overestimation and instability because: (i) coverage is limited to the empirical state distribution (next-state distributions and long-horizon compounding errors remain), (ii) counterfactual branches (e.g., Case 2) introduce modeling simplifications, and (iii) bootstrapping can amplify small value errors. Conservative Q-Learning (CQL) Kumar et al. (2020) mitigates these issues by penalizing large Q-values on weakly supported actions/states, biasing learning toward conservative, in-distribution estimates.

Given our augmented dataset $\mathcal{D}^*$ containing both observational and counterfactual tuples, we obtain Q-values $Q_\theta(s, a)$ for *both actions* at each state. This allows us to update over the full action space simultaneously, rather than only the logged action.

For an augmented transition $(s, a, r, s', a^+, r^+, s'^+)$, the Double DQN target for $a' \in \{a, a^+\}$ is

$$y_{a'} = r_{a'} + \gamma\, Q_{\bar{\theta}}\big(s'_{a'}, \arg\max_{a''} Q_\theta(s'_{a''}, a'')\big),$$

where $(r_{a'}, s'_{a'})$ denotes the reward and next state associated with action a', and $Q_{\bar{\theta}}$ is the target network. The double temporal-difference (TD) loss then averages over both actions:

$$\mathcal{L}_{\text{TD}}(\theta) = \mathbb{E}_{(s,a,\cdot)\sim\mathcal{D}^*}\big[\tfrac{1}{2}\big((Q_\theta(s,a) - y_a)^2 + (Q_\theta(s,a^+) - y_{a^+})^2\big)\big].$$

To further curb overestimation, we additionally introduce the CQL penalty (Kumar et al., 2020):

$$\mathcal{L}_{\text{CQL}}(\theta) = \mathbb{E}_{s\sim\mathcal{D}^*}\big[\log \textstyle\sum_{a\in\{0,1\}} \exp(Q_\theta(s,a)) - \tfrac{1}{2}\sum_{a\in\{0,1\}} Q_\theta(s,a)\big].$$

These two losses enable updates over both actions simultaneously while stabilizing the learning of Q-values for weakly supported state–action pairs.

3.2.2 SUPERVISED CONTRASTIVE LEARNING

Beyond Q-learning, CAFE-RL incorporates *supervised contrastive objectives* that exploit the paired *factual–counterfactual* transitions. These objectives provide auxiliary supervision that aligns Q-values with observed outcomes and improves stability. We design two contrastive losses as follows:

Regret Loss. The regret loss captures the intuition that if the policy π_θ assigns high probability to an action whose expected return is significantly worse than that of its counterfactual, the model should reduce its confidence in that action. Let $\pi_\theta(a|s)$ denote the softmax distribution over Q-values at state s, and let $R(a|s)$ denote the expected future return associated with action a: $R(a|s) = r(s,a) + \gamma \max_{a'} r(s', a')$. Let $a^* = \arg\max_a Q_\theta(s,a)$ be the action selected by the policy, and let a^+ denote its counterfactual alternative. We define the regret at state s as:

$$\text{Regret}(s) = R(a^*|s) - R(a^+|s).$$

The regret loss is then given by

$$\mathcal{L}_{\text{regret}} = -\,\mathbb{E}_{s\sim\mathcal{D}^*}\Big[\text{Regret}(s)\cdot \log \pi_\theta(a^*|s)\Big].$$

This penalizes the policy when it is overconfident in actions that exhibit high regret compared to their counterfactuals.

Rank Loss. The rank loss enforces consistency between the relative ordering of Q-values and the empirical ordering of expected returns between factual–counterfactual pairs. Let $\Delta R = R(a|s) - R(a^+|s)$ be the difference in expected returns, and $\Delta Q = Q_\theta(s,a) - Q_\theta(s,a^+)$ be the predicted Q-value difference. The rank loss is defined as

$$\mathcal{L}_{\text{rank}} = \mathbb{E}_{s\sim\mathcal{D}^*}\big[\max\big(0, -\operatorname{sign}(\Delta R)\cdot \Delta Q + m\big)\big],$$

where $m > 0$ is a margin hyperparameter. This ensures that Q-value ranking follows the observed return ranking.

3.2.3 FINAL OBJECTIVE.

Bringing everything together, the overall training objective of CAFE-RL combines the temporal-difference loss, conservative regularization, and supervised contrastive losses. The final loss is

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{TD}} + \lambda_1 \cdot \mathcal{L}_{\text{CQL}} + \lambda_2 \cdot \mathcal{L}_{\text{regret}} + \lambda_3 \cdot \mathcal{L}_{\text{rank}},$$

where $\lambda_1, \lambda_2, \lambda_3 > 0$ are trade-off hyperparameters. This formulation allows CAFE-RL to simultaneously benefit from conservative value estimation, action-level supervision, and ranking consistency, yielding more stable and effective policy learning.

4 EXPERIMENTS

We evaluate the proposed **CAFE-RL** framework on with the sampled data from real e-commercial platform of eBay (All reported data are anonymized and do not reflect the actual distribution of the underlying platform), with the goal of answering the following questions:

- **RQ1:** Does jointly optimizing registration and transaction decisions improve long-term platform outcomes compared to checkpoint-wise risk control mechanisms?
- **RQ2:** Can counterfactual augmentation and contrastive learning mitigate the limitations of mechanism-induced partial feedback and yield improved policy?
- **RQ3:** Does the proposed framework remains robust when learning from conservative and imperfect intervention mechanisms, where reject inference is unavoidable by design?

Dataset and Features. We evaluate CAFE-RL on anonymized real-world buyer registration and first-transaction logs from a large-scale e-commerce marketplace (eBay), covering millions of new users over one month, with first transactions observed within a 30-day window after registration. Data are split chronologically into training, validation, and an out-of-time test set to assess temporal generalization. User states are represented by tabular features selected using standard IV and GBDT importance criteria, resulting in a 750-dimensional input representation. Additional dataset description and preprocessing details are provided in Appendix B.1 and B.2 respectively.

Baselines. We compare our method with the current eBay platform-deployed separate risk score models (denoted as **Ckpt Model**) as well as representative imitation learning and offline RL approaches including Behavior Cloning **(BC)**, Conservative Q-Learning **(CQL)** (Kumar et al., 2020), and Model-Based Policy Optimization **(MBPO)** (Janner et al., 2019). Details of these baselines are presented in Appendix B.3.

Experimental Setup. We adopt Double Dueling DQN Van Hasselt et al. (2016); Wang et al. (2016) as the deep RL backbone for our proposed CAFE-RL framework. The network consists of three hidden layers with a hidden size of 256. We train the model using the Adam optimizer with a learning rate of 0.001. The hyperparameters $\lambda_1, \lambda_2, \lambda_3$ are set to 0.3, 0.3, 0.2, respectively. Training is conducted with a mini-batch strategy, where the batch size is 600,000 and the total number of training episodes is 400. For the reward function in Table 1, we define the conversion rate $\alpha = 0.1$, the penalty constants of $\epsilon = 1$ and $c = 5$ respectively.

Evaluation Metrics. We evaluate performance from both risk-control and economic perspectives. For fraud detection accuracy, we report `Precision` and `Recall` at both checkpoints.

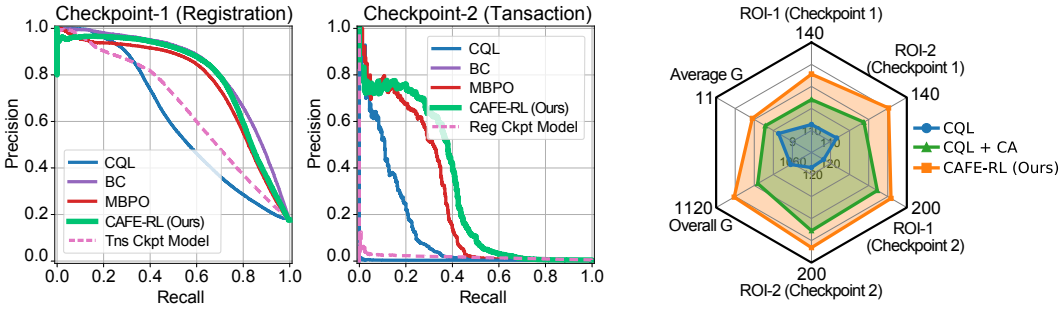
To assess platform-level outcomes, we further introduce profit-oriented metrics that capture the trade-off between fraud prevention and GMV. Specifically, we report two variants of Return on Investment (ROI). `ROI-1` measures the gain from approving legitimate users relative to the losses incurred by fraudulent approvals and false rejections, reflecting a risk-sensitive objective. `ROI-2` additionally rewards successful fraud rejections, providing a more holistic evaluation of intervention effectiveness. In addition, we report absolute profit gains, including `Overall Gain` (G) and `Average Gain` ($\overline{G}$), to quantify the economic impact of different policies. Formal definitions of all metrics are provided in Appendix B.4.

4.1 RQ1: EFFECTIVENESS OF UNIFIED SEQUENTIAL OPTIMIZATION

We first analyze whether jointly optimizing registration and transaction decisions as a unified sequential mechanism leads to better platform-level outcomes than checkpoint-wise risk control.

Table 2: Profit-related results including `ROI` and `Gain` at both registration (Checkpoint-1) and transaction (Checkpoint-2). Decision thresholds for all models are tuned to prioritize similar decline rates. (All `Gain` values are anonymized and reported only to preserve the relative ordering.)

Model	Checkpoint-1 (Registration)			Checkpoint-2 (Transaction)			Gain	
	ROI-1 ↑	ROI-2 ↑	Decline Rate	ROI-1 ↑	ROI-2 ↑	Decline Rate	Overall G ↑	Average $\overline{G}$ ↑
CQL	117.63	117.92	0.105	130.21	130.96	0.0006	1073.11	9.69
BC	121.97	123.71	0.099	141.26	141.26	0.00001	1079.83	9.73
MBPO	124.46	127.47	0.105	173.31	174.13	0.0012	1091.65	9.89
Ckpt Model	109.13	109.82	0.102	133.16	133.44	0.0012	1043.08	9.45
CAFE-RL (Ours)	**131.34**	**134.26**	0.110	**186.73**	**189.13**	0.0012	**1108.65**	**10.24**



(a) Precision-Recall Curves of two checkpoints.

(b) Ablations on CA

Figure 4: (a) Precision-Recall Curves of BOFD. **Left:** Buyer registration stage (Checkpoint-1). **Right:** Buyer transaction stage (Checkpoint-2). (b) Analysis of Counterfactual Augmentation (CA).

Figure 4a compares precision–recall curves at both checkpoints. While the deployed checkpoint models achieve competitive precision locally, their performance varies substantially across stages, reflecting the lack of coordination between registration and transaction decisions. In contrast, CAFE-RL consistently maintains strong precision while extending recall at both checkpoints, indicating improved global decision quality rather than isolated local gains.

Beyond classification metrics, Table 2 reports ROI and profit-related outcomes. For Ckpt Model baseline, after we obtain the corresponding risk score, we tune the decision threshold to control the decline rate. Under comparable decline rates, we observe nearlly all other sequential optimization methods outperforms checkpoint-wise baselines in terms of ROI and overall GMV gains, where CAFE-RL outperforms others significantly. These results demonstrate that modeling buyer onboarding fraud detection as a unified sequential decision problem yields strictly better global outcomes than locally optimized interventions. From a mechanism design perspective, CAFE-RL effectively learns a globally optimized intervention mechanism that internalizes long-term trade-offs across stages, rather than relying on myopic checkpoint-wise rules.

Insight-1: *Sequential optimization in BOFD internalizes long-term trade-offs between early rejection risk and downstream GMV that isolated local rules fail to capture.*

4.2 RQ2: Effect of Counterfactual Augmentation and Hybrid Learning

We evaluate whether counterfactual augmentation and hybrid learning objectives improve offline policy learning under mechanism-induced partial feedback.

We first isolate the effect of counterfactual augmentation by comparing CQL, CQL augmented with counterfactual transitions (CQL+CA), and the full CAFE-RL framework. As shown in Fig. 4b, introducing counterfactual augmentation leads to a substantial and consistent improvement over vanilla CQL across all ROI and gain metrics, with particularly large gains at the transaction checkpoint and in overall profit. This highlights a fundamental limitation of pure CQL: when learning from conservative logs, value estimation is confined to the support of historically executed actions. Coun-

Table 3: Profit-related results including `ROI` and `Gain` of ablations study on hybrid learning. Decision thresholds for all models are tuned to prioritize similar decline rates. (All `Gain` values are anonymized separately and only preserve the relative ordering among methods.)

Model	Checkpoint-1 (Registration)			Checkpoint-2 (Transaction)			Gain	
	ROI-1	ROI-2	Decline Rate	ROI-1	ROI-2	Decline Rate	G	$\overline{G}$
CAFE-RL	**131.34**	**134.26**	0.110	**186.73**	**189.13**	0.0012	**1108.65**	**10.24**
CAFE-RL w/o $\mathcal{L}_{\text{rank}}$	130.40	133.22	0.109	183.24	187.71	0.0011	1100.62	10.03
CAFE-RL w/o $\mathcal{L}_{\text{regret}}$	125.31	127.83	0.110	170.71	172.64	0.0013	1100.45	9.98
CAFE-RL w/o $\mathcal{L}_{\text{CQL}}$	131.20	133.87	0.109	185.13	187.26	0.0011	1106.88	10.03

terfactual augmentation expands effective state–action coverage, enabling the policy to reason about alternative interventions that are unobserved but essential for improvement.

In addition, Table 3 reports an ablation study where key components of the hybrid learning objective in CAFE-RL are removed while keeping decline rates comparable across methods. Across all ablated variants, removing any component leads to consistent degradation in ROI and profit-related metrics, with the most pronounced impact observed at the transaction checkpoint and in overall gains. This pattern indicates that each component contributes to effective policy improvement, and that combining counterfactual augmentation with auxiliary contrastive supervision yields more reliable optimization than value-based learning alone.

Insight-2: *Counterfactual augmentation is necessary to overcome mechanism-induced support limitations in offline BOFD, while hybrid reinforcement–contrastive objectives are critical for stabilizing value estimation and refining action selection.*

4.3 RQ3: Improvement over Imperfect Intervention Mechanisms

We study whether CAFE-RL can improve decision quality when learning from imperfect platform intervention mechanisms, where reject inference and noisy declines are unavoidable by design.

Fig. 4a reports the precision–recall curves across both onboarding checkpoints. The deployed platform mechanisms at registration and transaction do not exhibit perfect precision, implying that the logged data are generated by conservative yet imperfect intervention rules. Despite being trained entirely on such biased logs, CAFE-RL consistently dominates all baselines across the precision–recall spectrum at both checkpoints. This result demonstrates that CAFE-RL does not rely on an idealized or precision-perfect logging policy. Instead, it learns *strictly improved intervention rules on top of the existing imperfect mechanism*, simultaneously at registration and transaction. In other words, even when the historical policy incurs false declines and partial feedback, the learned RL policy achieves uniformly better trade-offs between fraud control and GMV preservation.

Crucially, these improvements are not isolated to a single decision point. By optimizing the onboarding process as a unified sequential decision problem, CAFE-RL improves intervention quality at both stages under the same imperfect logging mechanism, yielding consistent gains throughout the onboarding pipeline.

Insight-3: *Even when historical intervention mechanisms are imperfect, CAFE-RL can reliably improve both early-stage and downstream decisions, demonstrating robustness to reject inference and mechanism-induced bias.*

5 Conclusion and Future Work

We study Buyer Onboarding Fraud Detection (BOFD) from a mechanism design and strategic decision-making perspective and propose **CAFE-RL**, a counterfactual-augmented offline reinforcement learning framework for optimizing early-stage platform interventions. By modeling registration and first-transaction screening as a sequential decision process, CAFE-RL improves upon

existing mechanisms through counterfactual reasoning and a hybrid reinforcement–contrastive objective. Experiments on large-scale real-world data show that CAFE-RL consistently learns strictly improved intervention policies on top of imperfect and conservative logging mechanisms, achieving better trade-offs between fraud control and GMV preservation across stages. These results highlight the value of offline RL as a principled tool for mechanism optimization under partial feedback.

Future work includes extending the framework to longer user lifecycles and studying safe online deployment and adaptive mechanism updates in dynamic marketplaces.

REFERENCES

Nancy Agarwal, Suraiya Jabin, Syed Zeeshan Hussain, et al. Analyzing real and fake users in facebook network based on emotions. In *2019 11th International Conference on Communication Systems & Networks (COMSNETS)*, pp. 110–117. IEEE, 2019.

Adam Breuer, Roee Eilat, and Udi Weinsberg. Friend or faux: Graph-based early detection of fake accounts on social networks. In *Proceedings of the web conference 2020*, pp. 1287–1297, 2020.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Dawei Cheng, Sheng Xiang, Chencheng Shang, Yiyi Zhang, Fangzhou Yang, and Liqing Zhang. Spatio-temporal attention-based neural network for credit card fraud detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 362–369, 2020.

R Renuga Devi, Joseph Emerson Raja, and Yeo Boon Chin. Reinforcement learning with graph neural network (rl-gnn) fusion for real-time financial fraud detection: a context-aware community mining approach. *Scientific Reports*, 2025.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*, pp. 2052–2062. PMLR, 2019.

Jia Guo, Guannan Liu, Yuan Zuo, and Junjie Wu. Learning sequential behavior representations for fraud detection. In *2018 IEEE international conference on data mining (ICDM)*, pp. 127–136. IEEE, 2018.

Zehong Hu, Zhen Wang, Zhao Li, Shichang Hu, Shasha Ruan, and Jie Zhang. Fraud regulating policy for e-commerce via constrained contextual bandits. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 1377–1385, 2019.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. *Advances in neural information processing systems*, 32, 2019.

Jiaxin Jiang, Yuan Li, Bingsheng He, Bryan Hooi, Jia Chen, and Johan Kok Zhi Kang. Spade: A real-time fraud detection framework on evolving graphs. *Proceedings of the VLDB Endowment*, 16(3):461–469, 2022.

Samira Khodabandehlou and Alireza Hashemi Golpayegani. Fifraud: unsupervised financial fraud detection in dynamic graph streams. *ACM Transactions on Knowledge Discovery from Data*, 18 (5):1–29, 2024.

Priyanka Kondeti, Lakshmi Pranathi Yerramreddy, Anita Pradhan, and Gandharba Swain. Fake account detection using machine learning. In *Evolutionary computing and mobile sustainable networks: Proceedings of ICECMSN 2020*, pp. 791–802. Springer, 2020.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in neural information processing systems*, 33:1179–1191, 2020.

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

Siyu Li, Jin Yang, Gang Liang, Tianrui Li, and Kui Zhao. Sybilflyover: Heterogeneous graph-based fake account detection model on social networks. *Knowledge-Based Systems*, 258:110038, 2022.

Xiao Liang, Zheng Yang, Binghui Wang, Shaofeng Hu, Zijie Yang, Dong Yuan, Neil Zhenqiang Gong, Qi Li, and Fang He. Unveiling fake accounts at the time of registration: An unsupervised approach. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 3240–3250, 2021.

Xu Lin, Panagiotis Ilia, Saumya Solanki, and Jason Polakis. Phish in sheep's clothing: Exploring the authentication pitfalls of browser fingerprinting. In *31st USENIX Security Symposium (USENIX Security 22)*, pp. 1651–1668, 2022.

Can Liu, Li Sun, Xiang Ao, Jinghua Feng, Qing He, and Hao Yang. Intention-aware heterogeneous graph attention networks for fraud transactions detection. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 3280–3288, 2021.

Mingxuan Lu, Zhichao Han, Susie Xi Rao, Zitao Zhang, Yang Zhao, Yinan Shan, Ramesh Raghunathan, Ce Zhang, and Jiawei Jiang. Bright-graph neural networks in real-time fraud detection. In *Proceedings of the 31st ACM international conference on information & knowledge management*, pp. 3342–3351, 2022.

Adrian Mead, Tyler Lewris, Sai Prasanth, Stephen Adams, Peter Alonzi, and Peter Beling. Detecting fraud in adversarial environments: A reinforcement learning approach. In *2018 Systems and Information Engineering Design Symposium (SIEDS)*, pp. 118–122. IEEE, 2018.

Mao V Ngo, Tie Luo, and Tony QS Quek. Adaptive anomaly detection for internet of things in hierarchical edge computing: A contextual-bandit approach. *ACM Transactions on Internet of Things*, 3(1):1–23, 2021.

Karandeep Singh, Yu-Che Tsai, Cheng-Te Li, Meeyoung Cha, and Shou-De Lin. Graphfc: Customs fraud detection with label scarcity. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 4829–4835, 2023.

Andrew Smith and Charles Elkan. A bayesian network framework for reject inference. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 286–295, 2004.

Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Siddharth Vimal, Kanishka Kayathwal, Hardik Wadhwa, and Gaurav Dhama. Application of deep reinforcement learning to payment fraud. *arXiv preprint arXiv:2112.04236*, 2021.

Ziming Wang, Qianru Wu, Baolin Zheng, Junjie Wang, Kaiyu Huang, and Yanjie Shi. Sequence as genes: An user behavior modeling framework for fraud transaction detection in e-commerce. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5194–5203, 2023.

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas. Dueling network architectures for deep reinforcement learning. In *International conference on machine learning*, pp. 1995–2003. PMLR, 2016.

Wei Zhang, Yufei Zhang, Yilin Huang, Fangshu Chen, Jiahui Wang, and Xiaoming Hu. Gufad: a graph-based unsupervised fraud account detection framework. In *Proceedings of the 2023 4th International Conference on Machine Learning and Computer Application*, pp. 401–406, 2023.

Wen Zheng, Bingbing Xu, Emiao Lu, Yang Li, Qi Cao, Xuan Zong, and Huawei Shen. Midlg: Mutual information based dual level gnn for transaction fraud complaint verification. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5685–5694, 2023.

Andy Zhou, Xiaojun Xu, Ramesh Raghunathan, Alok Lal, Xinze Guan, Bin Yu, and Bo Li. Knowgraph: Knowledge-enabled anomaly detection via logical reasoning on graph data. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 168–182, 2024.

Xiaozhen Zhou, Shanping Li, Cheng Chang, Jianfeng Wu, and Kai Liu. Information-value-based feature selection algorithm for anomaly detection over data streams. *Technical Gazette/Tehnički Vjesnik*, 21(2), 2014.

A Related Works

Fraud Detection as Platform Intervention. Fraud detection has been extensively studied in e-commerce platforms, with methods targeting malicious registrations, account takeovers, and fraudulent transactions using supervised learning, graph-based models, and representation learning Breuer et al. (2020); Li et al. (2022); Liang et al. (2021); Zhang et al. (2023); Lu et al. (2022); Singh et al. (2023). In practice, these models are deployed as *intervention rules* at different checkpoints of the user lifecycle, most commonly at registration Liang et al. (2021) and transaction stages Liu et al. (2021); Lu et al. (2022); Zheng et al. (2023), often through thresholded risk scores or rule-based pipelines.

From a mechanism design perspective, such checkpoint-wise deployments implicitly define access and screening mechanisms that regulate user participation and risk exposure. However, existing systems typically optimize each checkpoint in isolation, focusing on local prediction accuracy or loss minimization. This fragmented design overlooks how early intervention decisions reshape the population that reaches downstream stages, and thus fails to optimize system-level objectives such as long-term GMV or overall marketplace efficiency.

Sequential Decision-Making and RL for Fraud Control. To capture temporal structure in fraud behavior, prior work has modeled fraud as a sequential or evolving process using behavior sequences Guo et al. (2018); Wang et al. (2023), temporal graphs Khodabandehlou & Golpayegani (2024); Jiang et al. (2022), and dynamic representations Cheng et al. (2020). While effective for prediction, these approaches remain largely *passive*: they characterize fraud risk but do not model the strategic impact of intervention actions on future states.

Several studies move toward active decision-making by formulating fraud control as a bandit or reinforcement learning problem Hu et al. (2019); Ngo et al. (2021); Vimal et al. (2021); Mead et al. (2018); Devi et al. (2025). These methods demonstrate the potential of adaptive policies to balance fraud loss and revenue. However, most existing approaches treat fraud decisions as per-transaction approvals or apply RL as a meta-controller, without explicitly modeling how early interventions constrain downstream state distributions or how intervention rules jointly determine long-term system outcomes.

In contrast to prior work, our approach frames early-stage fraud detection as a *mechanism optimization problem*, where platform intervention rules across stages must be designed jointly under partial feedback and strategic constraints.

B Detailed Experimental Settings

B.1 Dataset

We evaluate the proposed framework using real-world buyer registration and transaction records from the e-commercial platform marketplace. We sample millions of newly registered users over a one-month period, along with their first transactions within a 30-day window after registration. For data partitioning, we split the dataset according to user registration time. Specifically, the first 20 days of sampled registrations and their corresponding first-transaction logs are used for training, while the subsequent sampled users within the same month are reserved for validation. To further assess generalization across time, we adopt an out-of-time validation (OTV) setting: three additional consecutive days of sampled registrations and transactions (without down-sampling) are used as the test set.

B.2 Feature Construction and Selection

To construct the feature set, we apply both Information Value (IV) Zhou et al. (2014) and Gradient Boosting Decision Tree (GBDT) Chen & Guestrin (2016) importance weights to perform feature selection, ensuring that only informative predictors are retained. For categorical (tabular) features, we employ one-hot encoding to capture discrete attribute information in a machine-readable form. After feature selection and transformation, the final feature representation contains 750 dimensions, which serves as the input to our proposed deep-RL framework.

B.3 BASELINES

We compare our method with the current platform-deployed risk models as well as representative offline RL approaches:

- **Checkpoint Risk Model (Ckpt Model):** The production-grade risk control baseline deployed on the *eBay* platform. At each checkpoint, this baseline consists of an ensemble-style scoring pipeline that aggregates multiple heterogeneous expert models, including tree-based models (e.g., GBDT), rule-based systems, and deep learning models, to produce a unified risk score for each user or transaction.

  Intervention decisions are implemented by applying a configurable threshold to this score: instances above the threshold are declined, while others are approved. In our experiments, we use this aggregated score as a surrogate decision variable and simulate platform interventions by adjusting the decision threshold. By varying the threshold, we precisely control the decline rate, enabling fair and meaningful comparisons across methods under matched operational constraints.

  This setup reflects how intervention mechanisms are deployed and tuned in real-world e-commerce systems, while allowing controlled evaluation of alternative decision policies.
- **Behavior Cloning (BC):** A supervised learning baseline that directly imitates the historical logging policy by minimizing the discrepancy between the model's predicted actions and those recorded in the dataset. It provides a useful benchmark for understanding the performance gap between pure imitation and reinforcement learning–based optimization.
- **Conservative Q-Learning (CQL) Kumar et al. (2020):** A state-of-the-art offline RL algorithm designed to address overestimation issues when learning from static logged data. CQL penalizes unseen state–action pairs by learning a conservative Q-function, thereby reducing the risk of selecting actions that were rarely observed in the dataset. Moreover, we combine CQL with our counterfactual augmentation (CA) for comparison, denoted as **CQL+CA**.
- **Model-Based Policy Optimization (MBPO) Janner et al. (2019):** A model-based offline RL method that learns a dynamics model of the environment to generate synthetic trajectories for policy training. By augmenting limited logged data with model-generated rollouts, MBPO aims to improve sample efficiency and generalization. In our setting, MBPO leverages the registration and transaction logs for dynamic model training.

B.4 EVALUATION METRICS

We evaluate the proposed performance based on several metrics. For fraud detection, we choose the widely adopted precision and recall for evaluation. Additionally, we also aim to evaluate the actual profit the platform can earn. To this end, we introduce the following evaluation metrics:

Return on Investment (ROI). It considers the actual profit from GMV increasing and the loss the platform will have with fraud transactions. Specifically, for any user $u \in \mathcal{U}$, we define

- $A(u)$: the event that user u is approved, i.e., $a(u) = \texttt{Approve}$;
- $D(u)$: the event that user u is declined, i.e., $a(u) = \texttt{Decline}$;
- $F(u)$: the event that user u is fraudulent, i.e., $u \in \mathcal{F}$;
- $g(u)$: the GMV associated with user u;
- α: a fixed conversion rate (we set $\alpha = 0.1$ in our evaluation).

Then we define `ROI-1` as the ratio on the gain from approving non-fraud users versus the risk of approving frauds or rejecting legitimate users:

$$\text{ROI-1} = \frac{\alpha \sum_{u \in \mathcal{U}} \mathbb{1}_{A(u) \wedge \neg F(u)} \cdot g(u)}{\sum_{u \in \mathcal{U}} \left[\mathbb{1}_{A(u) \wedge F(u)} \cdot g(u) + \alpha \cdot \mathbb{1}_{D(u) \wedge \neg F(u)} \cdot g(u) \right]}.$$

`ROI-2` additionally rewards the system for successfully declining fraud users, leading to a more holistic metric:

$$\text{ROI-2} = \frac{\sum_{u \in \mathcal{U}} [\alpha \mathbb{1}_{A(u) \wedge \neg F(u)} \cdot g(u) + \mathbb{1}_{D(u) \wedge F(u)} \cdot g(u)]}{\sum_{u \in \mathcal{U}} \left[\mathbb{1}_{A(u) \wedge F(u)} \cdot g(u) + \alpha \cdot \mathbb{1}_{D(u) \wedge \neg F(u)} \cdot g(u) \right]}.$$

Profit Gain. It approximates the actual profit gain that the platform can obtain given the conversion rate. Specifically, we define the `Overall-Gain` and `Averaged-Gain` as follows:

$$\text{Overall-Gain: } G = \sum_{u \in \mathcal{U}} \left[\alpha \cdot g(u) - \mathbb{1}_{A(u) \wedge F(u)} \cdot g(u)\right],$$

$$\text{Average-Gain: } \overline{G} = \frac{\text{Overall-Gain}}{|\mathcal{U}|}.$$

C Additional Experimental Results

Table 4: Reward function (with uniform transaction reward) specification for BOFD in the MDP, defined for any user u.

Checkpoint	Registration Stage ($t = 1$)		Transaction Stage ($t = 2$)		
Action	0: "approve"		0: "approve"		1: "decline"
Fraud	✗	✓	✗	✓	✗ / ✓
Reward	0	$-c$	$\alpha \cdot m$	$-m$	$-\epsilon$

Table 5: Profit-related results including `ROI` and `Gain` of CAFE-RL with uniform transaction reward function. Decision thresholds for all models are tuned to prioritize similar decline rates. (All `Gain` values are anonymized separately and only preserve the relative ordering among methods.)

Model	Checkpoint-1 (Registration)			Checkpoint-2 (Transaction)			Gain	
	ROI-1	ROI-2	Decline Rate	ROI-1	ROI-2	Decline Rate	G	$\overline{G}$
CAFE-RL	**131.34**	**134.26**	0.110	**186.73**	**189.13**	0.0012	**1108.65**	**10.24**
CAFE-RL (r def in Tab. 4)	126.10	128.73	0.109	167.83	169.31	0.0009	1099.93	9.94

C.1 CAFE-RL with Uniform Reward Function

We additionally investigate the effect of GMV-related transaction reward function definition on the profit-related performance. For comparison, we replace the GMV-related transaction reward with a uniform reward across all users, as summarized in Table 4. Concretely, we fix the conversion rate to $\alpha = 0.1$, set penalty constants to $\epsilon = 1$ and $c = 5$, and assign a uniform transaction reward constant $m = 10$.

The results, reported in Table 5, show that CAFE-RL with the uniform reward function performs consistently worse than the GMV-related design. By contrast, the GMV-based reward function provides more fine-grained supervision aligned with profit value, enabling the policy to better balance fraud prevention against GMV preservation.

These findings highlight the importance of incorporating GMV into reward design: beyond fraud detection accuracy, profit-aware rewards guide the RL agent toward decisions that optimize both platform security and long-term business performance.