TOWARDS TEMPORAL INTEREST MODELING IN RECOMMENDATION VIA REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement Learning (RL) methods are promising for recommendation systems due to their ability to optimize long-term user-system interactions. However, many RL recommendation methods treat user interactions as homogeneous decision rounds, ignoring irregular temporal gaps and the resulting interest decay over time. To incorporate temporal interest signals into RL, we study the relationship between temporal gaps and user interest through mutual information, finding power-law decay at the population-level and exponential decay at the individual level. Guided by the empirical observations, we formulate a temporal-aware objective for RL-based recommendation, and propose a structured decision-process model: **Temporal Interest Semi-MDP (TIS-MDP)**, which incorporates temporal gaps via a survival function to make RL methods temporal-aware. The effectiveness of TIS-MDP is verified across experiments.

1 INTRODUCTION

The rapid growth of the internet has left users facing overwhelming choices when browsing websites or mobile apps. For e-commerce retailers, content platforms, and the advertising market, this abundance is not only a user-experience challenge but also a direct driver of revenue: the ability to match the right item to the right user at the right moment determines conversions, engagement, and ultimately profit (eMarketer, 2025). As a result, recommendation systems are a core component of modern online services, tasked with inferring user preferences from massive candidate sets and selecting what to display conditioned on context such as position, time, and interaction history. Typically, recommendation models are trained to predict short-term supervised Key Performance Indicators (KPIs) such as click-through rate (CTR) or conversion rate (CVR). However, user behavior is inherently long-term: a good recommendation not only triggers an immediate click but also considers future engagement, retention, and subsequent interactions. This mismatch has motivated a growing line of work on reinforcement learning for recommendation (Afsar et al., 2022), which formulates recommendation as a sequential decision-making problem and learns a policy that optimizes cumulative profit over time.

Reinforcement learning (RL) studies sequential decision making, where an agent interacts with an environment by selecting actions, receiving reward feedback, and updating its policy to maximize the expected cumulative reward over time.

These properties make RL a promising paradigm for improving recommendation systems (Afsar et al., 2022; Yang et al., 2025; Chen et al., 2019b; Ie et al., 2019). Despite this appeal, the recommendation differs from many classical RL settings in that interactions are not homogeneous decision rounds: they are separated by *temporal gaps*, i.e., the elapsed time between two sequential interactions. Moreover, such temporal gaps are not straightforward to incorporate as part of states since the gap is only observed after an action is taken and the next interaction occurs. Temporal information is an important feature in recommendations, as it reflects the correlation between two sequential interactions. However, most RL-based recommendation methods either model interactions as a finite-horizon episodic MDP, capturing sequential dependencies within a window but abstracting away temporal gaps (Cai et al., 2023b; Chen et al., 2019b), or treat the recommendation as a sequence of homogeneous decision rounds (Chen et al., 2018; Liu et al., 2023b), thereby ignoring the temporal information inherent in real user interactions. Even in work that targets long-horizon objectives such as user retention (Cai et al., 2023a), temporal structure is not explicitly modeled or analyzed.

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

Fig.1 illustrates the key issue: should a user's future interaction be attributed to a recommended item equally, whether it happens immediately, after one day, or even after ten days?

In other RL applications with non-negligible action durations, such as driver dispatching (Tang et al., 2019) and robotics (Metelli et al., 2020), a standard method is to adopt a Semi-Markov Decision Process (Semi-MDP) (Sutton et al., 1999). Semi-MDP generalizes the MDP by associating each action with a duration, allowing the model and the return to depend explicitly on the time elapsed between decision rounds. This provides a way to incorporate real-time dynamics when actions consume variable amounts of time.

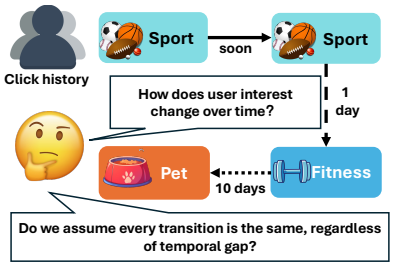


Figure 1: Irregular temporal gaps in recommendation.

In recommendation, we further need to relate elapsed time to interest continuation. First, the temporal gap is determined by the user's behavior under the current state and action and is only observed after the next interaction. Second, our goal is not to model execution time, but to capture how latent user interest decays with elapsed time. This raises two questions:

1. How does user interest decay across the temporal gap? 2. How can we incorporate temporal interest information into RL-based recommendations?

We address these two questions by empirically characterizing how user interest decays with temporal gaps in recommendation and introducing a structured decision process, **Temporal Interest Semi-MDP (TIS-MDP)**. Our main contribution can be summarized as follows,

- We study and quantify the relationship between temporal gaps and interest decay through mutual information, providing an empirical characterization of this connection in recommendation: we find that interest decay is well approximated by a power-law at the population level, but is closer to an exponential trend at the individual level, offering a new macro-to-micro perspective on how interest changes over time.
- Motivated by the observed interest-decay patterns, we introduce a latent interest-continuation view of recommendation, reformulate the RL objective, and propose a structured decision process, Temporal Interest Semi-MDP (TIS-MDP), which incorporates temporal interest signals into RL-based recommendation through a survival function and enables the agent to attribute future outcomes across varying temporal gaps properly. We establish contraction and fixed-point uniqueness of the induced Bellman operator and provide a Law of Large Numbers (LLN) justification for using a population-level power-law survival function. We develop a family of algorithms based on TIS-MDP to construct temporal-aware value estimation and policy updates.
- We integrate TIS-MDP with multiple RL backbones and evaluate it in a synthetic environment, as well as two open-source benchmarks: EasyRL4Rec, which provides an RL-for-recommendation evaluation pipeline, and KuaiSim, which simulates user leave and return dynamics in cross-session settings, thereby verifying the effectiveness of TIS-MDP.

2 Related Works

Applying RL to recommendations has become both feasible and increasingly promising (Shani et al., 2005; Afsar et al., 2022). Most existing approaches formulate the recommendation process as an episodic MDP (Cai et al., 2023b; Zhao et al., 2021; Liu et al., 2023a; Chen et al., 2019a;b; Xin et al., 2020; Liu & Yang, 2019), where each episode corresponds to a short session and interactions within it are treated as immediate. While effective for capturing within-session dynamics, this model largely ignores cross-session signals, which reflect how past states and actions determine long-term user engagement. The recommendation system is modeled as an infinite-horizon MDP in Chen et al. (2018); Ge et al. (2021; 2022); Zhao et al. (2018a; 2020; 2018b); Liu et al. (2023b); Ma et al. (2020), where interactions are treated as homogeneous decision rounds, and the differences induced by irregular temporal gaps are ignored. In Cai et al. (2023a), they model recommendations as an infinite-horizon MDP: interactions within a session are treated as a tightly coupled sequence without decay, whereas cross-session transitions follow the same fixed pattern. This still imposes a coarse temporal structure, treating all cross-session gaps equivalently, whether users return after

hours or weeks. Another related work is structured MDP in Maystre et al. (2023), which decomposes long-term value via an item-specific state and value function. Some works incorporate long- and short-term preferences via state construction (Hou et al., 2023), and hierarchical RL formulations study temporal abstraction by separating decision layers within and out of session (Ji et al., 2024). However, these works still do not distinguish heterogeneous temporal structures, overlooking a key component of recommendation dynamics.

There is also a line of work on temporal-aware recommendation. These methods mainly utilize temporal information to aggregate historical interactions, while our focus is temporal weighting of future credit in RL. General transition-dependent discounting is established in White (2017); emphatic TD primarily addresses off-policy stability (Sutton et al., 2016). Our contribution is the recommendation-specific empirical characterization and use of temporal decay. A detailed discussion is provided in Appendix A.

3 PRELIMINARIES

(Semi-)MDP: A standard Markov decision process (MDP) is modeled by a tuple $(\mathcal{S}, \mathcal{A}, P, r)$, where

- **State $\mathcal{S}$:** each state $s \in \mathcal{S}$ summarizes the user representation.
- **Action $\mathcal{A}$:** each action $a \in \mathcal{A}$ specifies the recommendation decision, typically a recommended item.
- **Transition kernel P:** $P(\cdot \mid s, a)$ describes how the user state changes after taking action a in state s. In the recommendation setting, this transition is often assumed to be deterministic, i.e., $P(s' \mid s, a) = 1$ for a certain s'.
- **Reward function r:** $r(s, a)$ defines the immediate reward of recommending a in state s, such as click, conversion or retention.

At each decision round $t = 1, 2, \ldots$, the agent observes s_t, chooses $a_t \sim \pi(\cdot \mid s_t)$, receives $r_t = r(s_t, a_t)$, and the user state transitions to s_{t+1}. The Markov property means that the conditional distribution of s_{t+1} depends on the past only through (s_t, a_t). In the discounted infinite-horizon setting, the agent maximizes $J(\pi) = \mathbb{E}_\pi \left[\sum_{t \geq 1} \gamma^{t-1} r(s_t, a_t)\right]$, where $\gamma \in (0, 1)$ is a discount factor. The action value function satisfies

$$Q^\pi(s, a) = r(s, a) + \gamma \mathbb{E}_{a' \sim \pi(\cdot|s')}[Q^\pi(s', a')]. \tag{1}$$

The MDP formulation implicitly assumes homogeneous decision rounds. A Semi-Markov Decision Process (Semi-MDP) (Sutton et al., 1999) associates each transition with a (possibly stochastic) duration $\Delta_t \in \mathbb{R}_+$. After taking action a_t in state s_t, the environment produces (s_{t+1}, Δ_t), where Δ_t indicates the elapsed real time until the next decision. Under the discounted criterion, discounting is applied during the time of action execution, yielding

$$Q^\pi(s, a) = r(s, a) + \mathbb{E}_{a' \sim \pi(\cdot|s')} \left[\gamma^{\Delta_t} Q^\pi(s', a')\right], \tag{2}$$

where the expectation is over the joint transition $(s', \Delta) \sim P(\cdot, \cdot \mid s, a)$ and then $a' \sim \pi(\cdot \mid s')$.

Temporal Gap: Similar to Semi-MDP, for recommendation, we define the temporal gap Δ at state s as the elapsed time between two sequential interactions s and s'. Let ts and ts' denote their timestamps, then $\Delta = \mathrm{ts}' - \mathrm{ts}$. The realized gap is available for the learning update only after the next interaction occurs.

Mutual Information: Mutual information (MI) is a measure of statistical dependence between two random variables. For discrete X and Y with joint probability $p(x, y)$ and marginals $p(x)$ and $p(y)$, it is defined as

$$\mathrm{MI}(X, Y) \triangleq \sum_x \sum_y p(x, y) \log \frac{p(x, y)}{p(x)p(y)}. \tag{3}$$

It holds that $\mathrm{MI}(X, Y) \geq 0$, with equality if and only if X and Y are independent.

4 BRIDGING INTEREST DECAY AND TEMPORAL GAPS IN RECOMMENDATION

The main challenge in incorporating interest decay into RL is that we only observe the temporal gaps between interactions, rather than users' latent interest dynamics. This raises a practical question:

How does interest decay with increasing temporal gaps?

Inspired by Zhou et al. (2024), we quantify semantic–temporal correlation via mutual information. For the temporal aspect, we study how the dependency between historical behaviors and target responses varies with the temporal gap. For semantics, the dependency can be instantiated with any discrete semantic feature whose cardinality is moderate enough for reliable estimation. In our empirical study, we use item category as the semantic feature because it alleviates sparsity while preserving meaningful item-level structure. Then we define the following metric, which quantifies the dependency between a user's historical behavior category and the target-item response:

Definition 4.1 (Category-wise Target-aware Temporal-gap Correlation (CTTC))**.** Given a target category c_i and a temporal gap Δ (in days), sample a (target item, history item) pair such that the target category is c_i and the gap is Δ. For any history category c_j, define

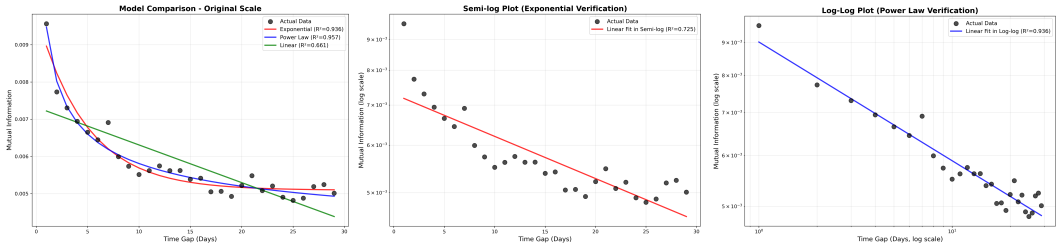
$$X_{c_j,\Delta} \triangleq \mathbb{I}\{\text{the history item has category } c_j\} \in \{0,1\},$$

and let $Y_i \in \{0,1\}$ be the label of the target item. We define

$$\mathrm{CTTC}(c_i, c_j, \Delta) \triangleq \mathrm{MI}\big(X_{c_j,\Delta}, Y_i\big). \tag{4}$$

This category-based score serves as a concrete empirical proxy for semantic–temporal correlation. Other semantic features can be used in the same way as long as their cardinality remains moderate. It quantifies how much observing a history category at gap Δ reduces uncertainty about the label of the target. Accordingly, we can bridge interest dynamics with observable time gaps by examining how the target-category-averaged correlation varies with Δ. In practice, for each target category, we first aggregate the CTTC values over history categories and then average the resulting curves over target categories, where each term is computed from empirical pairs $\big(X_{c_j,\Delta}, Y_i\big)$ aggregated across the dataset.

4.1 POPULATION-LEVEL INTEREST DECAY



(a) Original-scale fit. (b) Semi-log fit (exponential). (c) Log-log fit (power law).

Figure 2: Exponential and power-law fits to population CTTC. Original-scale R^2 is 0.936 and 0.957, respectively; the semi-log and log-log diagnostic fits give 0.725 and 0.936.

We obtain a population-level semantic–temporal correlation curve by aggregating over history categories and averaging over target categories:

$$\mathrm{CTTC}(\Delta) \triangleq \frac{1}{|\mathcal{C}|} \sum_{c_i \in \mathcal{C}} \sum_{c_j \in \mathcal{C}} \mathrm{CTTC}(c_i, c_j, \Delta). \tag{5}$$

where $\mathcal{C}$ denotes the target category set used for averaging, and the inner summation aggregates history-category contributions. The resulting $\mathrm{CTTC}(\Delta)$ summarizes the overall contribution of history items to the target item label given a temporal gap Δ. Fig.2a shows that $\mathrm{CTTC}(\Delta)$ drops sharply at the very beginning, which naturally induces an exponential-decay hypothesis,

$\mathrm{CTTC}(\Delta) \approx Ae^{-\beta\Delta} + C$. However, the empirical curve also suggests a clear heavy-tail pattern, which is inconsistent with a single-rate exponential model. The empirical experiments in Fig.2 verify this observation: a power-law family provides a better match than the exponential fitting. This contrast is particularly evident when plotting the exponential model on a semi-log scale and the power-law model on a log–log scale. We obtain consistent results on additional datasets, as shown in Appendix B.4. In particular, the results show that a Lomax form yields a strong fit:

$$\mathrm{CTTC}(\Delta) \approx A\Big(1 + \frac{\Delta}{\tau}\Big)^{-\alpha} + C. \tag{6}$$

This finding provides a powerful empirical basis for designing temporally aware mechanisms to capture long-range information more precisely in RL.

4.2 Why power-law? A Heterogeneity Mechanism from Micro to Macro

One more step beyond the population-level interest decay is to ask the question: how does user interest decay at the individual level? Our study suggests an answer,

At the individual level, interest decay is better characterized by an exponential trend, whereas at the population level, it yields an approximately power-law decay.

Specifically, for a user u, suppose interest at temporal gap Δ decays exponentially, $g_u(\Delta) \propto \exp(-\lambda_u \Delta)$, where λ_u is a user-specific forgetting rate. If λ_u varies across users and is modeled as a random variable drawn from a Gamma distribution, $\lambda_u \sim \mathrm{Gamma}(\alpha, \tau)$, then the population-level decay follows by the Laplace transform:

$$\mathbb{E}_u\big[\exp(-\lambda_u \Delta)\big] = \mathbb{E}\big[\exp(-\lambda \Delta)\big] = \Big(1 + \frac{\Delta}{\tau}\Big)^{-\alpha}. \tag{7}$$

This identity recovers the power-law form in Eq.(6), providing a concrete mechanism that bridges micro-level decay for individual users and macro-level decay. We further verify this hypothesis empirically in Appendix B.3. However, we focus on the population-level perspective for two reasons: First, individual-level estimation is much noisier due to data sparsity and is more sensitive to sampling bias. Second, incorporating per-user decay models into RL would require user-specific parameterization, which is often impractical at scale. We therefore design our modeling and algorithms based on the population-level power-law observation.

5 TIS-MDP: Temporal Interest Continuation for Recommendation

Given the empirical relationship between temporal gaps and interest continuation, the next question is

How should we incorporate temporal interest decay into the RL objective for recommendation?

To address this question, we reformulate the RL objective through introducing the latent user interest mechanism: in recommendation, future rewards should be credited to the current state–action pair only when they still belong to the same underlying user-interest chain. When an interest break occurs, subsequent interactions may still be valuable for later decisions, but they should no longer be assigned as long-range credit to decisions before the break.

Latent interest-continuation objective. Let $\Delta_t := \mathrm{ts}_{t+1} - \mathrm{ts}_t$ denote the temporal gap between two consecutive interactions. We introduce a latent continuation variable $C_t \in \{0, 1\}$, where $C_t = 1$ indicates that the current interest chain continues from interaction t to $t + 1$, and $C_t = 0$ indicates an interest break. Under this view, the chain-local return from step t is defined as

$$G_t^{\mathrm{TIS}} = \sum_{\ell=0}^{\infty} \gamma_{\mathrm{base}}^{\ell} \left(\prod_{j=0}^{\ell-1} C_{t+j} \right) r_{t+\ell}, \tag{8}$$

where $\gamma_{\mathrm{base}} \in (0, 1)$ is a fixed base discount factor, and the empty product is defined as one. The return in (8) means that the reward $r_{t+\ell}$ is credited to the current state–action pair only if the same latent interest chain survives through every intermediate transition from t to $t + \ell$.

We define the action value by $Q^\pi(s,a) = \mathbb{E}_\pi[G_t^{\mathrm{TIS}} \mid s_t = s, a_t = a]$; future interactions after an interest break remain valuable as new decisions.

Temporal Interest Semi-MDP. The above objective naturally induces a structured decision process, Temporal Interest Semi-MDP (TIS-MDP). Concretely, let $\gamma_{\mathrm{base}} \in (0,1)$ be a fixed base discount factor. For each transition $(s_t, a_t, \mathrm{ts}_t) \to (s_{t+1}, \mathrm{ts}_{t+1})$, we observe the time gap $\Delta_t := \mathrm{ts}_{t+1} - \mathrm{ts}_t$ and apply a temporal-aware effective discount $\gamma_{\mathrm{eff}}(\Delta_t) \triangleq \gamma_{\mathrm{base}} \cdot s(\Delta_t)$, where $s(\Delta_t)$ denotes the survival (interest-continuation) probability at gap Δ_t.

For the gap-only model, we assume $\Pr(C_t = 1 \mid s_t, a_t, s_{t+1}, \Delta_t) = s(\Delta_t)$ and that the next state is sufficient for the subsequent return conditional on continuation. This is a population-level approximation; residual dependence on user context, state, or action may bias credit assignment. The Bellman equation in TIS-MDP under a fixed policy π then becomes

$$Q^\pi(s_t, a_t) = \mathbb{E}\left[r(s_t, a_t) + \gamma_{\mathrm{eff}} Q^\pi(s_{t+1}, a_{t+1}) | s_t, a_t\right],$$

where expectation is taken over the transition $s_{t+1} \sim P(\cdot \mid s_t, a_t)$, the induced time gap Δ_t and the action $a_{t+1} \sim \pi(\cdot|s_{t+1})$. Intuitively, the temporal-aware discounting truncates credit assignment across long gaps where user interest is likely to have shifted, while preserving standard bootstrapping for short gaps. We instantiate a family of algorithms in Algorithm 1 by plugging the above temporal-aware backup into RL backbones that use bootstrapped value or critic targets.

Algorithm 1 TIS-MDP Learning

1: **Input:** fixed $\gamma_{\mathrm{base}} \in (0,1)$; survival $s(\cdot)$; data $\mathcal{D}$ with timestamps; value function Q_ϕ; policy π_θ; a backbone RL policy update $\mathcal{U}$.
2: **repeat**
3: Sample a mini-batch $\{(s, a, r, s', \mathrm{ts}, \mathrm{ts}')\}$ from $\mathcal{D}$.
4: $\Delta \leftarrow \mathrm{ts}' - \mathrm{ts}$, $\gamma_{\mathrm{eff}} \leftarrow \gamma_{\mathrm{base}} s(\Delta)$.
5: Compute bootstrap value $V_{\mathrm{boot}}(s')$ according to the RL backbone (e.g., $\max_{a'} Q_\phi(s', a')$, or $Q_\phi(s', \pi_\theta(s'))$).
6: Temporal-aware TD target: $y \leftarrow r + \gamma_{\mathrm{eff}} \cdot V_{\mathrm{boot}}(s')$.
7: Critic update: $\phi \leftarrow \arg\min_\phi \ \mathbb{E}\left[(Q_\phi(s,a) - y)^2\right]$.
8: Policy update: $\theta \leftarrow \mathcal{U}\left(\theta; Q_\phi, \{(s, a, r, s')\}\right)$.
9: **until** converged

This plug-in modification applies to RL backbones whose learning rule contains a bootstrapped value or critic target, including value-based methods and actor–critic or offline RL algorithms with TD-style targets. For these methods, TIS-MDP makes the backup temporal-aware by replacing the constant discount γ in the target with the transition-dependent discount $\gamma_{\mathrm{eff}} = \gamma_{\mathrm{base}} s(\Delta)$. Here, CTTC is not treated as a calibrated continuation probability. Only the normalized decay shape is used; the fitted amplitude and offset are not continuation probabilities. Instead, it serves as a target-aware empirical proxy for identifying the decay shape of user interest, from which we instantiate a monotone survival function $s(\Delta)$ with $s(0) = 1$ and $0 \le s(\Delta) \le 1$. This perspective places TIS-MDP between two common credit-assignment choices. Compared with episodic RL, which treats the end of each session as a hard terminal state and discards subsequent credit assignment, TIS-MDP provides a soft, gap-dependent termination: long gaps lead to small $s(\Delta)$ and thus effectively truncate bootstrapping, while short gaps preserve temporal continuity. Compared with the standard infinite-horizon framework that uses a fixed discount, TIS-MDP modulates bootstrapping by the estimated interest-continuation probability, reducing wrong long-range credit assignment when user interest is likely to have shifted. Guided by the empirical CTTC decay pattern and scalability considerations, we use the population-level power-law form $s(\Delta) = [(\tau + \Delta)/\tau]^{-\alpha}$ as the survival function. This choice matches the observed heavy-tailed population trend, whereas an individual-level exponential model would require estimating and maintaining sparse per-user parameters, which is impractical at scale.

5.1 Contraction Analysis of TIS-MDP

To analyze the contraction of the Bellman equations under interest-continuation marginalization, we define the policy evaluation operator acting on bounded action value functions $(T^\pi Q)(s,a)$ as

$$(T^\pi Q)(s,a) := r(s,a) + \mathbb{E}\left[\gamma_{\text{base}} s(\Delta) \mathbb{E}_{a' \sim \pi(\cdot|s')} Q(s',a') \,\middle|\, s,a\right]. \tag{9}$$

where the outer expectation is over the joint next-state and temporal-gap transition (s', Δ). Since $0 < s(\Delta) \leq 1$ for all $\Delta \geq 0$, we have $\max \gamma_{\text{base}} s(\Delta) \leq \gamma_{\text{base}} < 1$. Assume the boundedness of reward that $|r(s,a)| \leq R_{\max}$ for all (s,a), define sup norm $\|Q\|_\infty \triangleq \sup_{s,a} |Q(s,a)|$, then we can obtain:

Proposition 5.1 (Contraction of T^π)**.** *For any policy π and any valid survival function $s(\cdot)$, the associated action-value function satisfies $\|Q^\pi\|_\infty \leq \frac{R_{\max}}{1-\gamma_{\text{base}}}$. Moreover, for any bounded $Q_1,\ Q_2$,*

$$\|T^\pi Q_1 - T^\pi Q_2\|_\infty \leq \gamma_{\text{base}} \|Q_1 - Q_2\|_\infty.$$

Hence, T^π is a γ_{base}-contraction in the sup norm.

By Proposition 5.1, T^π is a contraction on the complete metric space of bounded functions under $\|\cdot\|_\infty$. The Banach fixed-point theorem then yields the following Corollary.

Corollary 5.2 (Existence and uniqueness of Q^π)**.** *For any policy π, there exists a unique fixed point Q^π satisfying $Q^\pi = T^\pi Q^\pi$.*

Similarly, define the optimality operator as

$$(T^\star Q)(s,a) := r(s,a) + \mathbb{E}\left[\gamma_{\text{base}} s(\Delta) \max_{a'} Q(s',a') \,\middle|\, s,a\right].$$

The same argument implies that $T^\star$ is also a γ_{base}-contraction, hence the optimal action value function $Q^\star$ exists and is unique.

5.2 A Law of Large Numbers for Temporal-aware RL-based Recommendation

Section 4.2 shows how heterogeneous exponential decay curves can yield a population-level power law. TIS-MDP uses the population model to avoid fitting noisy, sparse user-specific curves. We now relate the population Bellman operator to the average of individual operators and establish a Law of Large Numbers (LLN) for the resulting value functions.

For each user u, define the policy evaluation operator T_u^π by

$$(T_u^\pi Q)(s,a) := r(s,a) + \mathbb{E}\left[\gamma_{\text{base}} s_u(\Delta) \mathbb{E}_{a' \sim \pi(\cdot|s')} Q(s',a') \,\middle|\, s,a\right]. \tag{10}$$

where $s_u(\Delta) = e^{-\lambda_u \Delta}$ and $\lambda_u \sim \text{Gamma}(\alpha, \tau)$. We assume rewards, transitions, and the temporal gap are population-shared and do not depend on u. Then recall the definition in (9), we can directly obtain that

$$(T^\pi Q)(s,a) = \mathbb{E}_u\big[(T_u^\pi Q)(s,a)\big], \tag{11}$$

Given N i.i.d. users $u_1, \ldots, u_N$ from the population, we further define the empirical operator

$$(T_N^\pi Q)(s,a) := \frac{1}{N} \sum_{i=1}^{N} (T_{u_i}^\pi Q)(s,a). \tag{12}$$

Similar to the proofs of Proposition 5.1 and Corollary 5.2, we have that T_u^π and T_N^π are also γ_{base}-contraction and corresponding fixed points are denoted by Q_u^π and Q_N^π. Then we provide the following theorem:

Theorem 5.3. *For finite $\mathcal{S}, \mathcal{A}$, $N \geq 1$, and $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\|Q_N^\pi - Q^\pi\|_\infty \leq \frac{2\gamma_{\text{base}} R_{\max}}{(1-\gamma_{\text{base}})^2} \sqrt{\frac{\log\big(2|\mathcal{S}||\mathcal{A}|/\delta\big)}{2N}}. \tag{13}$$

The proof is deferred to Appendix E. Under the shared-transition assumptions, Theorem 5.3 bounds the difference between the empirical and population fixed points as the number of users grows. It justifies population aggregation under the specified model, rather than identifying the decay family from data.

6 EXPERIMENTS

In this section, we conduct experiments in three settings: a synthetic environment, an offline RL benchmark for recommendation, and an online simulator that models user retention. Unless otherwise specified, we fix the decay exponent at $\alpha = 0.185$ (estimated from the Amazon dataset) and select the tuning parameter τ from a fixed set for TIS-MDP. Additional results and implementation details are provided in Appendix D. The Amazon exponent is a fixed default rather than a universal constant; the KuaiRand estimate near 0.25 is covered by the sensitivity analysis.

6.1 SYNTHETIC EXPERIMENT

To evaluate temporal-aware reinforcement learning for recommendation, we first design a synthetic experiment with known user interaction and interest-decay dynamics. This is a matched-model verification, rather than a test of robustness to misspecified decay. The detailed setup is described below.

States. The environment maintains a discrete *focus* state $s_t \in \{0, 1, 2\}$, corresponding to high, medium, and low focus, respectively. Within an interest chain, focus is monotone non-increasing: if the chain continues, the next focus level is restricted to staying the same or degrading by one level, i.e., $0 \to 1 \to 2$, and never improves unless an interest break occurs.

Actions and rewards. Action is a binary set $a_t \in \{0, 1\}$, which denotes conservative and aggressive recommendation respectively. Intuitively, an aggressive recommendation yields higher gains under high focus but may perform worse under low focus.

Temporal gap. The environment samples a temporal gap $\Delta_t \in \{1, 5, 15\}$ from a conditional distribution $P(\Delta_t \mid s_t, a_t)$ after the action a_t. By construction, lower focus and more aggressive actions stochastically lead to larger Δ_t, reflecting large temporal gap when recommendations are less aligned with the current interest.

Interest break and reset. The interest chain continues with probability $\Pr(\text{continue} \mid \Delta_t) = s(\Delta_t), s(\Delta) = [(\tau + \Delta)/\tau]^{-\alpha}$, where $s(\Delta)$ is a power-law survival function with parameters $\alpha, \tau > 0$. Upon an interest break, the user enters a new chain and the focus state is re-initialized as $s_{t+1} \sim \text{Unif}\{0, 1, 2\}$. Otherwise, if the chain continues, the focus transition follows the monotone degradation rule, i.e., when the chain continues, focus either stays unchanged or degrades by one level; it does not improve within a chain.

Under the above setting, the objective is to maximize the cumulative reward over an episode with a day-level horizon $T = 500$. To evaluate the benefit of exploiting temporal information via TIS-MDP, we consider three standard RL methods: (i) **DQN**, which learns an action-value function $Q(s, a)$ and selects actions via ϵ-greedy exploration; (ii) **A2C**, an actor–critic method that jointly learns a stochastic policy (actor) and a value baseline (critic) to reduce gradient variance; and (iii) **PPO**, which performs stable policy updates by optimizing a clipped surrogate objective to constrain the change of the policy between iterations. For each backbone, we additionally instantiate a temporally aware version by replacing the standard TD backup with the TIS-MDP update.

Method	Base	TIS	Gain (%)
DQN	47.29	53.27	+12.65%
A2C	52.81	**55.77**	+5.60%
PPO	51.86	<u>55.51</u>	+7.04%

Table 1: Synthetic rewards. TIS uses the temporal-aware TD backup.

Table 1 reports the results of the synthetic experiment. DQN gains the most ($+12.65\%$), while A2C and PPO show smaller gains ($+5.60\%$ and $+7.04\%$). These results verify the benefit of the temporal backup in the matched environment; they do not isolate the reasons for the different gains across backbones.

6.2 RECOMMENDATION VIA OFFLINE REINFORCEMENT LEARNING

To further validate the effectiveness of TIS-MDP, we conduct experiments on EasyRL4Rec (Yu et al., 2024), a benchmark for RL-based recommendation. Since the benchmark does not support online simulation of temporal gaps, we focus on offline RL methods in this setting. Specifically, we consider the following methods: (i) **BCQ**, which constrains policy updates to high-confidence,

in-distribution actions by sampling candidate actions from a behavior-conditioned generative model; (ii) **CQL**, which alleviates Q-value overestimation through a conservative regularization term that penalizes unseen actions; and (iii) **CRR**, which learns policies via advantage-weighted behavior cloning, thereby naturally avoiding out-of-distribution actions. Experiments are conducted on two popular real-world datasets, KuaiRand and MovieLens. Due to space constraints, we report results only on KuaiRand, as it provides more informative temporal-gap signals. We focus on three key metrics: average reward per step $\mathrm{R}_{\mathrm{avg}}$, episode length during testing Len, and cumulative reward per test episode $\mathrm{R}_{\mathrm{cumu}}$. Results on the MovieLens dataset, along with additional evaluation metrics, are provided in the Appendix D.2.

From Table 2, we observe that TIS-MDP improves all three backbones on cumulative reward. For BCQ, the policy is constrained by a generative model to remain close to the behavior data, so TIS-MDP mainly changes the critic backup and reduces over-propagation across large gaps, which tends to stabilize continuation decisions, and is reflected as longer Len. For CQL, the critic is explicitly regularized to be conservative on unseen actions, so the main benefit of a temporal-aware backup is better value calibration on in-distribution actions, which shows up more in $\mathrm{R}_{\mathrm{avg}}$ than in Len. For CRR, the actor update is largely a weighted regression on behavior actions and is strongly anchored to the dataset, so TIS-MDP has limited influence on the learned policy, leading to the smallest overall improvement.

Method	R_{cumu}	Len	R_{avg}
BCQ	<u>1.93</u>	4.80	<u>0.40</u>
BCQ (TIS)	**2.08**	5.14	**0.41**
CQL	0.85	3.13	0.27
CQL (TIS)	0.98	3.09	0.32
CRR	1.26	<u>5.83</u>	0.22
CRR (TIS)	1.27	**5.90**	0.22

Table 2: KuaiRand results; full metrics are in Appendix D.2.

The power-law versus exponential comparison is reported in Table 8; its mean results show a trade-off across reward metrics.

6.3 Cross-session User Retention Optimization

In this section, we adopt the cross-session setting in KuaiSim (Zhao et al., 2023), which simulates long-term user retention on short-video platforms. KuaiSim includes a *user-leave* module that predicts whether a user exits a session, and a user-return module that predicts when the user returns to the platform. We treat the predicted return interval as the temporal gap. Under the KuaiSim environment, we evaluate three RL methods: (i) **TD3**, a strong off-policy actor–critic baseline for continuous control that mitigates overestimation via clipped double Q-learning and delayed policy updates, (ii) **RLUR** (Cai et al., 2023a), which models recommendation as an infinite-horizon MDP for retention: interactions within a session are treated as undiscounted ($\gamma_{\mathrm{eff}} = 1$), while cross-session transitions are discounted by a fixed factor ($\gamma_{\mathrm{eff}} = \gamma_{\mathrm{base}}$). (iii) **TD3 (TIS)**, which uses TD3 as the backbone but replaces the standard TD backup with the TIS-MDP update rule to explicitly incorporate temporal gaps in value estimation.

Method	Return day ↓	Retention ↑
TD3 (TIS)	**2.1454**	**0.8772**
RLUR	<u>2.2415</u>	0.8032
TD3	2.2842	<u>0.8264</u>

Table 3: KuaiSim results. Retention means returning within three days.

The results in Table 3 show the highest retention rate and the lowest average return day for TD3 with TIS-MDP. A lower return day means an earlier user return. These simulator results support a benefit from temporal credit assignment, but do not measure training stability.

7 Conclusion

In this paper, we highlight the role of temporal gaps in long-term reinforcement learning optimization for recommendations. We study the relationship between temporal gaps and user interest, reformulate the reinforcement learning objective in recommendation, and propose the Temporal Interest Semi-MDP (TIS-MDP), which incorporates temporal information through a survival function. We establish a contraction analysis of TIS-MDP together with an LLN justification for the population-level survival function choice. The effectiveness of TIS-MDP is verified across multiple experiments.

AI USE STATEMENT

AI tools were used only for language polishing and LaTeX formatting.

REFERENCES

M Mehdi Afsar, Trafford Crump, and Behrouz Far. Reinforcement learning based recommender systems: A survey. *ACM Computing Surveys*, 55(7):1–38, 2022.

Qingpeng Cai, Shuchang Liu, Xueliang Wang, Tianyou Zuo, Wentao Xie, Bin Yang, Dong Zheng, Peng Jiang, and Kun Gai. Reinforcing user retention in a billion scale short video recommender system. In *Companion Proceedings of the ACM Web Conference 2023*, pp. 421–426, 2023a.

Qingpeng Cai, Zhenghai Xue, Chi Zhang, Wanqi Xue, Shuchang Liu, Ruohan Zhan, Xueliang Wang, Tianyou Zuo, Wentao Xie, Dong Zheng, et al. Two-stage constrained actor-critic for short video recommendation. In *Proceedings of the ACM web conference 2023*, pp. 865–875, 2023b.

Haokun Chen, Xinyi Dai, Han Cai, Weinan Zhang, Xuejian Wang, Ruiming Tang, Yuzhou Zhang, and Yong Yu. Large-scale interactive recommendation with tree-structured policy gradient. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 3312–3320, 2019a.

Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H Chi. Top-k off-policy correction for a reinforce recommender system. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pp. 456–464, 2019b.

Shi-Yong Chen, Yang Yu, Qing Da, Jun Tan, Hai-Kuan Huang, and Hai-Hong Tang. Stabilizing reinforcement learning in dynamic environment with application to online recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1187–1196, 2018.

eMarketer. Worldwide ad spending forecast 2025, 2025. URL https://www.emarketer.com/content/worldwide-ad-spending-forecast-2025. Accessed 2025-09-04.

Chongming Gao, Shijun Li, Yuan Zhang, Jiawei Chen, Biao Li, Wenqiang Lei, Peng Jiang, and Xiangnan He. Kuairand: An unbiased sequential recommendation dataset with randomly exposed videos. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, CIKM '22, pp. 3953–3957, 2022. doi: 10.1145/3511808.3557624. URL https://doi.org/10.1145/3511808.3557624.

Yingqiang Ge, Shuchang Liu, Ruoyuan Gao, Yikun Xian, Yunqi Li, Xiangyu Zhao, Changhua Pei, Fei Sun, Junfeng Ge, Wenwu Ou, et al. Towards long-term fairness in recommendation. In *Proceedings of the 14th ACM international conference on web search and data mining*, pp. 445–453, 2021.

Yingqiang Ge, Xiaoting Zhao, Lucia Yu, Saurabh Paul, Diane Hu, Chu-Cheng Hsieh, and Yongfeng Zhang. Toward pareto efficient fairness-utility trade-off in recommendation through reinforcement learning. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pp. 316–324, 2022.

Ashraf Ghiye, Baptiste Barreau, Laurent Carlier, and Michalis Vazirgiannis. Adaptive collaborative filtering with personalized time decay functions for financial product recommendation. In *Proceedings of the 17th ACM conference on recommender systems*, pp. 798–804, 2023.

Yan-e Hou, Wenbo Gu, WeiChuan Dong, and Lanxue Dang. A deep reinforcement learning real-time recommendation model based on long and short-term preference. *International Journal of Computational Intelligence Systems*, 16(1):4, 2023.

Eugene Ie, Vihan Jain, Jing Wang, Sanmit Narvekar, Ritesh Agarwal, Rui Wu, Heng-Tze Cheng, Morgane Lustman, Vince Gatto, Paul Covington, et al. Reinforcement learning for slate-based recommender systems: A tractable decomposition and practical methodology. *arXiv preprint arXiv:1905.12767*, 2019.

Luo Ji, Gao Liu, Mingyang Yin, Hongxia Yang, and Jingren Zhou. Hierarchical reinforcement learning for temporal abstraction of listwise recommendation. *arXiv preprint arXiv:2409.07416*, 2024.

Shantao Jiang, Yiping Wen, Jun Shen, Gaoxian Peng, Guosheng Kang, and Jianxun Liu. Personalized learning path recommendation with time-aware attention-based reinforcement learning. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–24, 2025.

Dong Liu and Chenyang Yang. A deep reinforcement learning approach to proactive content pushing and recommendation for mobile users. *IEEE Access*, 7:83120–83136, 2019.

Shuchang Liu, Qingpeng Cai, Bowen Sun, Yuhao Wang, Ji Jiang, Dong Zheng, Peng Jiang, Kun Gai, Xiangyu Zhao, and Yongfeng Zhang. Exploration and regularization of the latent action space in recommendation. In *Proceedings of the ACM Web Conference 2023*, pp. 833–844, 2023a.

Ziru Liu, Jiejie Tian, Qingpeng Cai, Xiangyu Zhao, Jingtong Gao, Shuchang Liu, Dayou Chen, Tonghao He, Dong Zheng, Peng Jiang, et al. Multi-task recommendations with reinforcement learning. In *Proceedings of the ACM web conference 2023*, pp. 1273–1282, 2023b.

Jiaqi Ma, Zhe Zhao, Xinyang Yi, Ji Yang, Minmin Chen, Jiaxi Tang, Lichan Hong, and Ed H Chi. Off-policy learning in two-stage recommender systems. In *Proceedings of The Web Conference 2020*, pp. 463–473, 2020.

Lucas Maystre, Daniel Russo, and Yu Zhao. Optimizing audio recommendations for the long-term: A reinforcement learning perspective. *arXiv preprint arXiv:2302.03561*, 2023.

Alberto Maria Metelli, Flavio Mazzolini, Lorenzo Bisi, Luca Sabbioni, and Marcello Restelli. Control frequency adaptation via action persistence in batch reinforcement learning. In *International Conference on Machine Learning*, pp. 6862–6873. PMLR, 2020.

Guy Shani, David Heckerman, and Ronen I Brafman. An mdp-based recommender system. *Journal of machine Learning research*, 6(Sep):1265–1295, 2005.

Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

Richard S. Sutton, A. Rupam Mahmood, and Martha White. An emphatic approach to the problem of off-policy temporal-difference learning. *Journal of Machine Learning Research*, 17(73):1–29, 2016. URL https://jmlr.org/papers/v17/14-488.html.

Xiaocheng Tang, Zhiwei Qin, Fan Zhang, Zhaodong Wang, Zhe Xu, Yintai Ma, Hongtu Zhu, and Jieping Ye. A deep value-network based approach for multi-driver order dispatching. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1780–1790, 2019.

Martha White. Unifying task specification in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 3742–3750. PMLR, 2017. URL https://proceedings.mlr.press/v70/white17a.html.

Pei Wu, Chi Ho Yeung, Weiping Liu, Cihang Jin, and Yi-Cheng Zhang. Time-aware collaborative filtering with the piecewise decay function. *arXiv preprint arXiv:1010.3988*, 2010.

Xin Xin, Alexandros Karatzoglou, Ioannis Arapakis, and Joemon M Jose. Self-supervised reinforcement learning for recommender systems. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pp. 931–940, 2020.

Dingqi Yang, Daqing Zhang, Vincent W. Zheng, and Zhiyong Yu. Modeling user activity preference by leveraging user spatial temporal characteristics in lbsns. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(1):129–142, 2015.

Xiao Yang, Mehdi Ayed, Longyu Zhao, Fan Zhou, Yuchen Shen, Abe Engle, Jinfeng Zhuang, Ling Leng, Jiajing Xu, Charles Rosenberg, et al. Deep reinforcement learning for ranking utility tuning in the ad recommender system at pinterest. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, pp. 945–948, 2025.

Yuanqing Yu, Chongming Gao, Jiawei Chen, Heng Tang, Yuefeng Sun, Qian Chen, Weizhi Ma, and Min Zhang. Easyrl4rec: An easy-to-use library for reinforcement learning based recommender systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 977–987, 2024.

Kesen Zhao, Shuchang Liu, Qingpeng Cai, Xiangyu Zhao, Ziru Liu, Dong Zheng, Peng Jiang, and Kun Gai. Kuaisim: A comprehensive simulator for recommender systems. *Advances in Neural Information Processing Systems*, 36:44880–44897, 2023.

Xiangyu Zhao, Long Xia, Liang Zhang, Zhuoye Ding, Dawei Yin, and Jiliang Tang. Deep reinforcement learning for page-wise recommendations. In *Proceedings of the 12th ACM conference on recommender systems*, pp. 95–103, 2018a.

Xiangyu Zhao, Liang Zhang, Zhuoye Ding, Long Xia, Jiliang Tang, and Dawei Yin. Recommendations with negative feedback via pairwise deep reinforcement learning. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1040–1048, 2018b.

Xiangyu Zhao, Long Xia, Lixin Zou, Hui Liu, Dawei Yin, and Jiliang Tang. Whole-chain recommendations. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pp. 1883–1891, 2020.

Xiangyu Zhao, Changsheng Gu, Haoshenglun Zhang, Xiwang Yang, Xiaobing Liu, Jiliang Tang, and Hui Liu. Dear: Deep reinforcement learning for online advertising impression in recommender systems. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 750–758, 2021.

Haolin Zhou, Junwei Pan, Xinyi Zhou, Xihua Chen, Jie Jiang, Xiaofeng Gao, and Guihai Chen. Temporal interest network for user response prediction. In *Companion Proceedings of the ACM Web Conference 2024*, pp. 413–422, 2024.

A Related Works

RL-based recommendation. Applying RL to recommendations has become both feasible and increasingly promising (Shani et al., 2005; Afsar et al., 2022). Most existing approaches formulate the recommendation process as an episodic MDP (Cai et al., 2023b; Zhao et al., 2021; Liu et al., 2023a; Chen et al., 2019a;b; Xin et al., 2020; Liu & Yang, 2019), where each episode corresponds to a short session and interactions within it are treated as immediate. While effective for capturing within-session dynamics, this model largely ignores cross-session signals, which reflect how past states and actions determine long-term user engagement. The recommendation system is modeled as an infinite-horizon MDP in Chen et al. (2018); Ge et al. (2021; 2022); Zhao et al. (2018a; 2020; 2018b); Liu et al. (2023b); Ma et al. (2020), where interactions are treated as homogeneous decision rounds, and the differences induced by irregular temporal gaps are ignored. In Cai et al. (2023a), they model recommendations as an infinite-horizon MDP: interactions within a session are treated as a tightly coupled sequence without decay, whereas cross-session transitions follow the same fixed pattern. This still imposes a coarse temporal structure, treating all cross-session gaps equivalently, whether users return after hours or weeks. Another related work is structured MDP in Maystre et al. (2023), which decomposes long-term value via an item-specific state and value function. Some works incorporate long- and short-term preferences via state construction (Hou et al., 2023), and hierarchical RL formulations study temporal abstraction by separating decision layers within and out of session (Ji et al., 2024). However, these works still do not distinguish heterogeneous temporal structures, overlooking a key component of recommendation dynamics.

Temporal-aware recommendation. Beyond the RL formulations discussed above, temporal signals are already widely applied in recommendation. TIN explicitly models semantic–temporal correlations between historical behaviors and a target item through target-aware temporal encoding and interaction for user response prediction (Zhou et al., 2024). Wu et al. (2010) measure how similarity signals vary over time and incorporate a piecewise decay function into item-based collaborative filtering; Ghiye et al. (2023) further learn personalized time-decay functions for financial product recommendation. Jiang et al. (2025) combine a time-aware attention mechanism with policy-gradient reinforcement learning for learning-path recommendation, where temporal information is mainly used to trace evolving preferences for path generation. These studies support the importance of temporal information, but they primarily use time to reweight historical interactions, model target-aware responses, or construct preference representations. Applying temporal information to future-value estimation remains unexplored.

B Extensive Study on Interest Decay.

B.1 Semantic–Temporal Correlation Heatmaps

We first visualize $\mathrm{CTTC}(c_i, c_j, \Delta)$ for a fixed target category $c_i = 351$ across the top-5 history categories c_j following Zhou et al. (2024), whereas they report position-relative correlations, we report *temporal*-relative correlations. Since the Amazon dataset only provides reviews for purchased items (i.e., positive feedback), we construct negative responses by replacing the target item with a randomly sampled item. Note that CTTC is computed at the population level, where we pool interactions from all users and aggregate co-occurrences for each (c_i, c_j, Δ) so that $\mathrm{CTTC}(c_i, c_j, \Delta)$ reflects an average semantic–temporal dependence. Fig.3 shows the resulting normalized heatmap where we exclude $\Delta = 0$ for visualization, since same-day interactions can overwhelm the scale.

The representative heatmap provides a semantic-level view of interest decay. To verify that the same trends hold beyond a single example, we provide additional heatmaps for more target categories. Moreover, we report two versions of the visualization: one excluding day 0 (Fig.4) and one including day 0 (Fig.5).

In practice, the mutual information at day 0 is typically the largest due to same-day interactions. When the heatmap is normalized with day 0 included, this dominant mass can compress the scale, making correlations at $\Delta \geq 1$ appear less evident. By contrasting these two cases, we obtain a clearer picture of the temporal decay: CTTC drops sharply from $\Delta = 0$ to small positive gaps, and then decays much more slowly over longer horizons, consistent with a long-tail persistence of historical influence. For most target items, the strongest correlation occurs when the history item

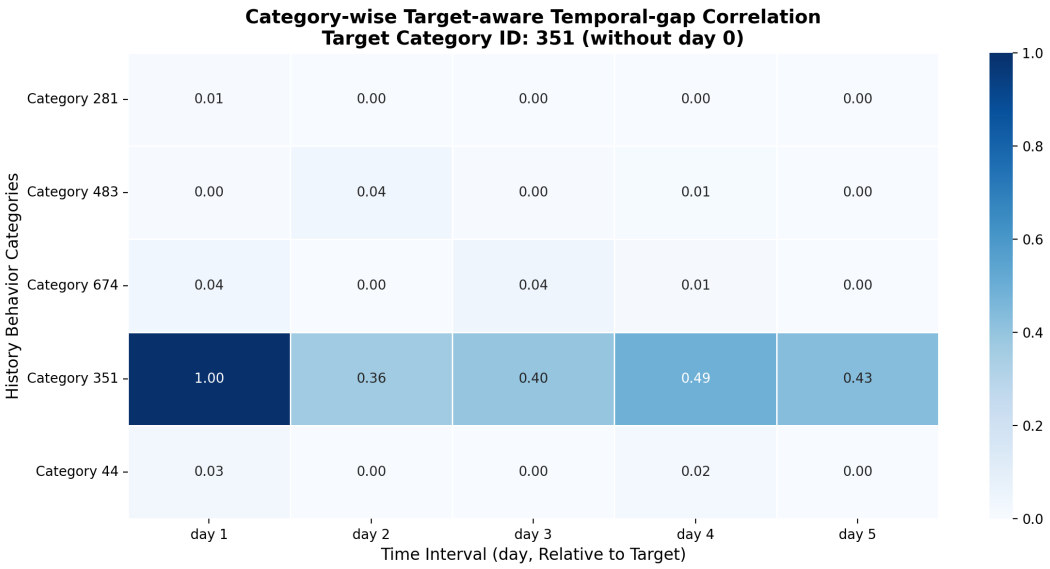
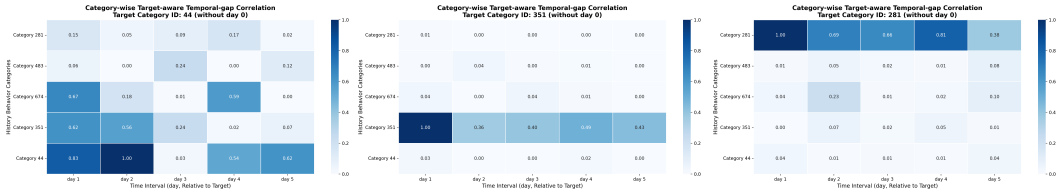


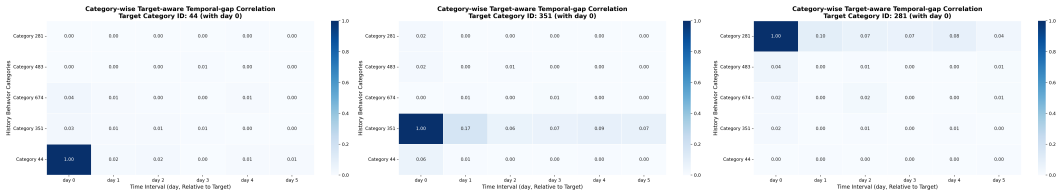
Figure 3: The temporal correlation between the history items of 5 frequent categories (y-axis) at various target-relative gaps from 1 to 5 Days, and the target category of 351.

shares the same category, while meaningful (but weaker) correlations also appear across related categories.



(a) Target Item: 44. (b) Target Item: 351. (c) Target Item: 281.

Figure 4: The temporal correlation between the history items of 5 frequent categories (without day 0).

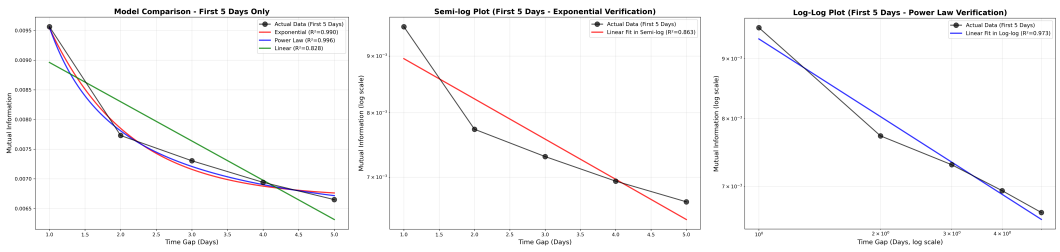


(a) Target Item: 44. (b) Target Item: 351. (c) Target Item: 281.

Figure 5: The temporal correlation between the history items of 5 frequent categories (with day 0).

B.2 Population-level Interest Decay

While Section 4.1 reports interest-decay fitting over a broad range of temporal gaps on 30-days data, we additionally evaluate the fit restricted to the first five days. This short-horizon regime contains denser observations and is therefore less noisy; moreover, fitting over fewer points is typically more stable. The conclusions remain unchanged: the power-law model achieves a better fit than the exponential baseline ($R^2 = 0.996$ vs. $R^2 = 0.990$). This difference becomes even more obvious under diagnostic plots: the exponential fit degrades on a semi-log scale ($R^2 = 0.863$), whereas the power-law remains nearly linear on a log–log scale ($R^2 = 0.973$).



(a) Original-scale fit. (b) Semi-log fit (exponential). (c) Log-log fit (power law).

Figure 6: Goodness-of-fit comparison (5 days) between exponential and power-law models across original, semi-log, and log-log environments.

B.3 Micro to Macro: Exploring The Mechanism of Interest Decay

Recall that we provide a hypothesis that:

At the individual level, interest decay is better characterized by an exponential trend, whereas at the population level, it yields an approximately power-law decay.

We next provide an empirical justification on it through the following steps:

Individual Fitting: Table 4 reports per-user fitting results for the top-10 most active users. We observe that despite the population-level decay curve being better characterized by a power-law model, the individual-level fits are more often better explained by exponential decay. Concretely, exponential fitting achieves higher R^2 for 9 out of 10 users, suggesting that user-specific interest decay may be closer to an exponential forgetting process with a personalized rate parameter. Meanwhile, the absolute R^2 values are relatively small for most users, indicating that the per-user CTTC curves remain noisy. This is expected because, compared with population-level aggregation, the individual contingency tables for (c_i, c_j, Δ) are substantially sparser, especially at larger gaps, which amplifies variance in mutual-information estimation. Overall, these results motivate modeling heterogeneous exponential decay at the user level, and attributing the emergent population-level power-law behavior to cross-user heterogeneity rather than a universal power-law mechanism for each user.

For the top-50 users, exponential fitting achieves higher R^2 for 41 out of 50 users, further supporting the same trend.

rank	user_id	interactions	num_categories	R^2_Exponential	R^2_Power Law	better_model
0	1	160874	858	0.019091	0.011839	Exponential
1	2	136885	810	0.065749	0.057767	Exponential
2	3	150917	730	0.002781	0.006005	Power Law
3	4	150637	700	0.099060	0.051829	Exponential
4	5	149640	698	0.040042	0.036551	Exponential
5	6	34695	662	0.039545	0.002443	Exponential
6	7	110981	558	0.610281	0.585068	Exponential
7	8	180378	530	0.119046	0.095460	Exponential
8	9	58729	518	0.496268	0.495567	Exponential
9	10	187873	516	0.196228	0.158496	Exponential

Table 4: Per-user fitting comparison on the top-10 users (original scale).

Distribution Test: Motivated by the individual-fitting results, we next examine whether the heterogeneity across users can be captured by a simple parametric distribution. Specifically, we extract a per-user exponential decay rate $\hat{\lambda}_u$ from the exponential fit of $\mathrm{CTTC}_u(\Delta)$, and fit a Gamma distribution to $\{\hat{\lambda}_u\}$. Across both user sets, the Kolmogorov–Smirnov (K–S) goodness-of-fit tests indicate that the Gamma hypothesis cannot be rejected at standard significance levels. For the top-10 users, the K–S statistic is 0.2540 with p-value 0.4644; for the top-50 users, the K–S statistic is 0.1572 with p-value 0.1516. These results suggest that a Gamma prior provides a plausible and parsimonious model for the cross-user variability of forgetting rates.

We further compare alternative families and observe that a Weibull distribution yields an even higher K–S p-value (e.g., 0.6183), indicating a potentially better empirical fit under this particular test. However, we primarily adopt the Gamma model in our subsequent analysis since it admits a closed-form Laplace transform and directly leads to the Lomax/power-law aggregation in Eq. (7), which enables a micro-to-macro explanation of the observed population-level decay.

We emphasize that "cannot reject" does not imply the chosen family is uniquely correct, especially under limited sample sizes where goodness-of-fit tests have limited power. Nevertheless, the fact that Gamma remains statistically plausible while also offering analytical tractability provides supportive evidence for modeling user heterogeneity via $\lambda_u \sim \mathrm{Gamma}(\alpha, \tau)$.

From Hypothesis to Verification: Having established that (i) individual users exhibit exponential decay patterns, and (ii) the cross-user decay rates $\{\hat{\lambda}_u\}$ can be plausibly modeled by a Gamma distribution, we now verify whether this micro-level characterization mathematically reproduces the observed population-level power-law behavior.

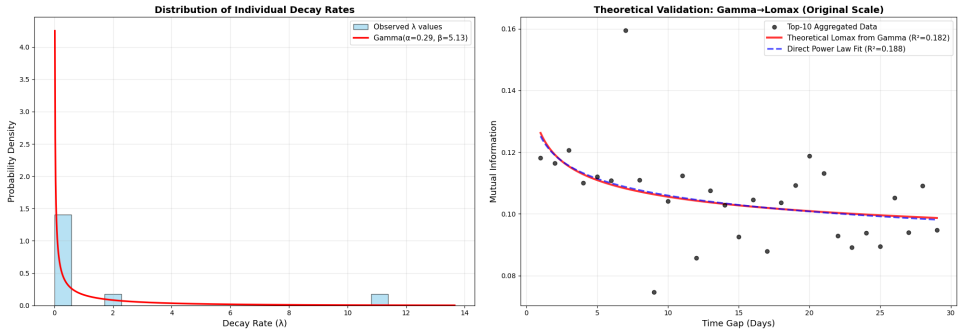
According to the theoretical derivation in Eq. (7), if each user u follows an exponential decay $\exp(-\lambda_u \Delta)$ with $\lambda_u \sim \mathrm{Gamma}(\alpha, \tau)$, then the expected aggregated decay across users should follow a Lomax form:

$$\mathbb{E}[\exp(-\lambda \Delta)] = \left(1 + \frac{\Delta}{\tau}\right)^{-\alpha},$$

which follows a power-law decay in the population-level CTTC curve.

To test this hypothesis, we compare two quantities:

1. **Theoretical Lomax**: Using the Gamma parameters (α, τ) estimated from $\{\hat{\lambda}_u\}$ via maximum likelihood, we construct the theoretical curve $(1 + \Delta/\tau)^{-\alpha}$ (with fitted amplitude and offset).
2. **Direct Power-Law Fit**: We directly fit a power-law model to the aggregated CTTC curve from the same user set.



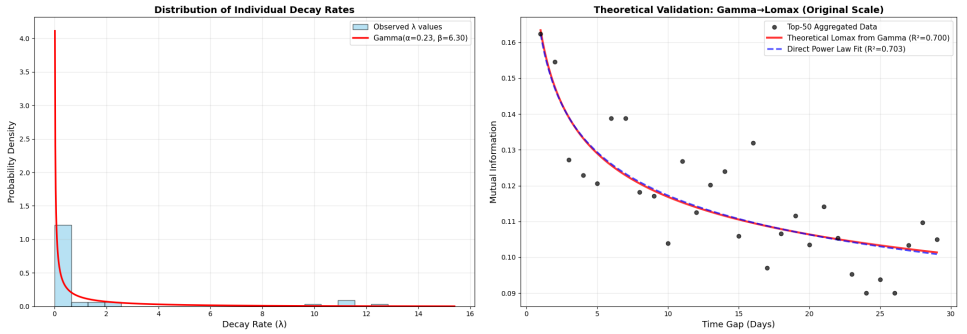
(a) Individual decay rates.

(b) Theoretical Lomax vs. Direct Power-Law Fitting.

Figure 7: Top-10 users Gamma distribution fitting.

Top-10 Results. Fig.7(a) displays the scatter of individual decay rates $\hat{\lambda}_u$ across the top-10 users, along with the fitted Gamma density. The Gamma model captures the right-skewed distribution, with most users exhibiting moderate decay rates and a few showing rapid forgetting. Fig.7(b) shows the theoretical validation on the aggregated decay curve. While the R^2 is low due to the noisy data, these two curves are highly consistent.

Top-50 Results. Fig.8(a) presents the decay-rate distribution for the top-50 users. With a larger sample, the histogram more clearly reveals the shape of the distribution. Fig.8(b) demonstrates a substantially improved theoretical match: the theoretical Lomax achieves $R^2 = 0.700$, nearly identical to the direct power-law fit's $R^2 = 0.703$ (difference $\Delta R^2 = 0.003$). This near-perfect agreement provides strong empirical support for the hypothesis that population-level power-law decay emerges from aggregating heterogeneous exponential processes.



(a) Individual decay rates.

(b) Theoretical Lomax vs. Direct Power-Law Fitting.

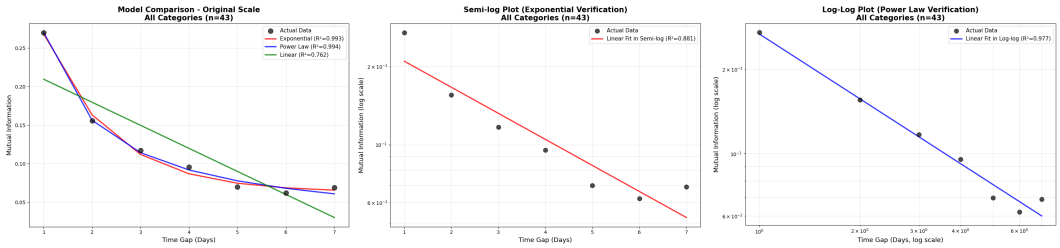
Figure 8: Top-50 users Gamma distribution fitting.

These results validate our hypothesis: individual users forget exponentially at personalized rates λ_u, and the cross-user heterogeneity in forgetting speeds follows the power-law pattern observed at the population level. An additional observation is that users with the most interactions (top-10 users) tend to exhibit slower interest decay, suggesting that highly active users retain longer-lasting preferences. The close match between theory and data indicates that the Gamma-Exponential mixture model is not merely a statistical convenience, but captures a genuine regularity in how user interest decays over time. While alternative distributions (Weibull) may fit the $\{\lambda_u\}$ distribution equally well or better statistically, the Gamma's theoretical tractability and its natural connection to the Lomax form make it a principled choice for both descriptive and predictive modeling.

Although TIS-MDP adopts a population-level power-law survival function, this observation remains valuable: it provides a micro-level explanation for the emergence of power-law decay after aggregation, and suggests how per-user temporal models could be incorporated when sufficient data and personalization capacity are available.

B.4 Validation on Additional Datasets

To further validate our conclusion on the relationship between temporal gaps and interest decay, we repeat the CTTC analysis on the KuaiRand-1K dataset (Gao et al., 2022). Figure 9 compares exponential and power-law models. While both models can provide a reasonable fit on the original scale, the distinction becomes clearer in log-space: the exponential hypothesis exhibits noticeably weaker linearity on a semi-log plot ($R^2 = 0.881$), whereas the power-law model is close to linear on a log–log plot ($R^2 = 0.977$). We further apply the same CTTC analysis to the Foursquare check-in dataset (Yang et al., 2015), which provides long-term user check-ins with venue categories, and observe the same pattern. These results further verify our earlier findings on Amazon.

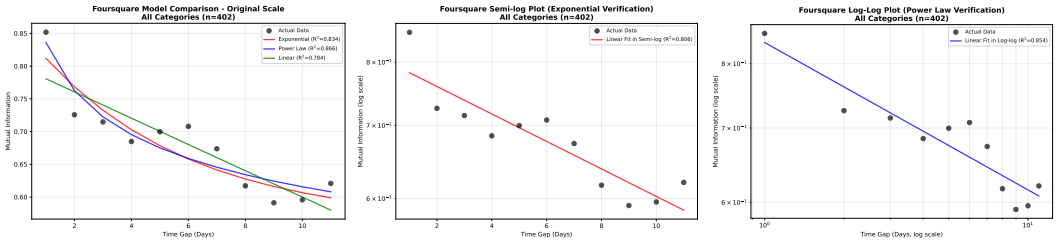


(a) Original-scale fit.

(b) Semi-log fit (exponential).

(c) Log-log fit (power law).

Figure 9: Goodness-of-fit comparison between exponential and power-law models across original, semi-log, and log-log environments on KuaiRand-1K dataset.



(a) Original-scale fit. (b) Semi-log fit (exponential). (c) Log-log fit (power law).

Figure 10: Goodness-of-fit comparison between exponential and power-law models across original, semi-log, and log-log environments on the Foursquare dataset.

C Sensitivity Analysis of TIS-MDP

Since we use an empirical model as a proxy for the true survival probability, we next quantify how estimation errors in the survival probability propagate to the resulting value functions. Let $s^\star(\Delta)$ be the true survival probability and $\hat{s}(\Delta)$ be an estimate, and define the error bound as

$$\varepsilon := \sup_{\Delta \geq 0} \left| \hat{s}(\Delta) - s^\star(\Delta) \right|. \tag{14}$$

Let $T^\pi_{\hat{s}}$ and $T^\pi_{s^\star}$ denote the evaluation operators corresponding $\hat{s}$ and $s^\star$, respectively, and let $Q^\pi_{\hat{s}}$ and $Q^\pi_{s^\star}$ be the action value function.

Theorem C.1. *Under the above assumptions,*

$$\|Q^\pi_{\hat{s}} - Q^\pi_{s^\star}\|_\infty \leq \frac{\gamma_{\text{base}} \varepsilon R_{\max}}{(1-\gamma_{\text{base}})^2}. \tag{15}$$

Proof. Since $Q^\pi_{\hat{s}}$ and $Q^\pi_{s^\star}$ are the fixed action value function, we have

$$\begin{aligned} Q^\pi_{\hat{s}} - Q^\pi_{s^\star} &= T^\pi_{\hat{s}} Q^\pi_{\hat{s}} - T^\pi_{s^\star} Q^\pi_{s^\star} \\ &= \left(T^\pi_{\hat{s}} Q^\pi_{\hat{s}} - T^\pi_{\hat{s}} Q^\pi_{s^\star}\right) + \left(T^\pi_{\hat{s}} Q^\pi_{s^\star} - T^\pi_{s^\star} Q^\pi_{s^\star}\right). \end{aligned}$$

Taking sup norms and applying Proposition 5.1 to the first term yields $\|T^\pi_{\hat{s}} Q^\pi_{\hat{s}} - T^\pi_{\hat{s}} Q^\pi_{s^\star}\|_\infty \leq \gamma_{\text{base}} \|Q^\pi_{\hat{s}} - Q^\pi_{s^\star}\|_\infty$. For the second term, for any (s,a) we have

$$\begin{aligned} &\left|(T^\pi_{\hat{s}} Q^\pi_{s^\star})(s,a) - (T^\pi_{s^\star} Q^\pi_{s^\star})(s,a)\right| \\ =& \left|\mathbb{E}\left[\gamma_{\text{base}}\big(\hat{s}(\Delta) - s^\star(\Delta)\big) \mathbb{E}_{a' \sim \pi(\cdot|s')} Q^\pi_{s^\star}(s',a') \;\middle|\; s,a\right]\right| \\ \leq& \gamma_{\text{base}} \varepsilon \left\|V^\pi_{s^\star}\right\|_\infty \leq \gamma_{\text{base}} \varepsilon \left\|Q^\pi_{s^\star}\right\|_\infty. \end{aligned}$$

Recall $\|Q^\pi_{s^\star}\|_\infty \leq \frac{R_{\max}}{1-\gamma_{\text{base}}}$, which gives

$$\|Q^\pi_{\hat{s}} - Q^\pi_{s^\star}\|_\infty \leq \gamma_{\text{base}} \|Q^\pi_{\hat{s}} - Q^\pi_{s^\star}\|_\infty + \gamma_{\text{base}} \varepsilon \frac{R_{\max}}{1-\gamma_{\text{base}}}.$$

Then we complete the proof by rearranging the inequality. □

Theorem C.1 shows that improving the estimation accuracy of $s(\Delta)$ directly reduces the induced value-function error.

D Additional Experimental Results and Setup

D.1 Synthetic Experiment

This appendix provides the full details of the synthetic environment.

State space. The environment maintains a discrete focus state $s_t \in \{0, 1, 2\}$, where 0, 1, and 2 correspond to *high*, *medium*, and *low* focus, respectively.

Action space. The agent chooses a binary action $a_t \in \{0, 1\}$, where $a_t = 0$ denotes a **conservative** recommendation and $a_t = 1$ denotes an **aggressive** recommendation.

Immediate reward. Given (s_t, a_t), the environment samples an immediate binary reward

$$r_t \sim \text{Bernoulli}\big(p(s_t, a_t)\big).$$

We set the click probabilities $p(s, a)$ as follows: aggressive recommendations yield higher immediate reward under high focus but can harm under low focus. Concretely, we use

$p(s,a)$	$a=0$ (cons.)	$a=1$ (agg.)
$s=0$ (high)	0.60	0.80
$s=1$ (med)	0.50	0.55
$s=2$ (low)	0.45	0.25

Temporal gap. After taking action a_t, the environment samples a day-level temporal gap

$$\Delta_t \in \{1, 5, 15\} \quad \text{from} \quad \Delta_t \sim P(\cdot \mid s_t, a_t).$$

The temporal gap depends on the state and action pair. We choose $P(\Delta \mid s, a)$ such that (i) lower focus leads to larger gaps, and (ii) aggressive actions lead to larger gaps, reflecting that misaligned recommendations slow down user return. We use the following conditional distribution:

$P(\Delta=1,5,15 \mid s,a)$	$a=0$ (cons.)	$a=1$ (agg.)
$s=0$	$(0.80,\ 0.18,\ 0.02)$	$(0.60,\ 0.30,\ 0.10)$
$s=1$	$(0.60,\ 0.35,\ 0.05)$	$(0.30,\ 0.50,\ 0.20)$
$s=2$	$(0.20,\ 0.50,\ 0.30)$	$(0.10,\ 0.40,\ 0.50)$

Interest break and reset. Given the realized gap Δ_t, the interest chain *continues* with probability

$$\Pr(\text{continue} \mid \Delta_t) = s(\Delta_t), \qquad s(\Delta) = \left(\tfrac{\tau}{\tau+\Delta}\right)^{\alpha},$$

where $\alpha, \tau > 0$. With the complementary probability $1 - s(\Delta_t)$, an interest break happens and the user enters a new chain; the next focus state is re-initialized as

$$s_{t+1} \sim \text{Unif}\{0, 1, 2\}. \tag{16}$$

Within-chain focus dynamics (only when the chain continues). If the chain continues (i.e., no reset), then focus transitions *within the same chain* according to a monotone stochastic degradation rule:

$$s_{t+1} = \begin{cases} \min(s_t + 1, 2), & \text{with probability } d(s_t, a_t), \\ s_t, & \text{with probability } 1 - d(s_t, a_t). \end{cases}$$

That is, conditioned on continuation, focus either stays unchanged or degrades by one level; it never improves within a chain. We set $d(s, a)$ so that aggressive actions induce faster fatigue:

$d(s,a)$	$a=0$ (cons.)	$a=1$ (agg.)
$s=0$	0.10	0.20
$s=1$	0.15	0.30
$s=2$	0.00	0.00

(17)

Backbone algorithms. We consider three standard RL backbones: **DQN**, **A2C**, and **PPO**, following their standard implementations. All methods operate on the same observation s_t and do not have access to Δ_t at action selection time.

Baseline TD backup. In baseline implementations, value bootstrapping uses a constant discount factor γ_{base} and ignores the realized gap:

$$y_t^{\text{base}} = r_t + \gamma_{\text{base}}\, V(s_{t+1}) \quad \text{or} \quad y_t^{\text{base}} = r_t + \gamma_{\text{base}} \max_{a'} Q(s_{t+1}, a'),$$

depending on whether the method is actor–critic or Q-learning.

TIS-MDP TD backup. For the temporally-aware versions, we replace the standard bootstrap weight by a temporal-aware factor induced by the realized gap. In our main experiments, we use

$$y_t^{\text{tis}} = r_t + \gamma_{\text{base}}\, s(\Delta_t)\, V(s_{t+1}) \quad \text{or} \quad y_t^{\text{tis}} = r_t + \gamma_{\text{base}}\, s(\Delta_t)\, \max_{a'} Q(s_{t+1}, a').$$

Fig. 11 reports the training curves of different algorithms in the synthetic experiment. Overall, the TIS-MDP variants converge substantially faster in the early stage of training, suggesting that incorporating temporal gaps into the TD backup provides a more informative and stable learning signal under irregular user return intervals.

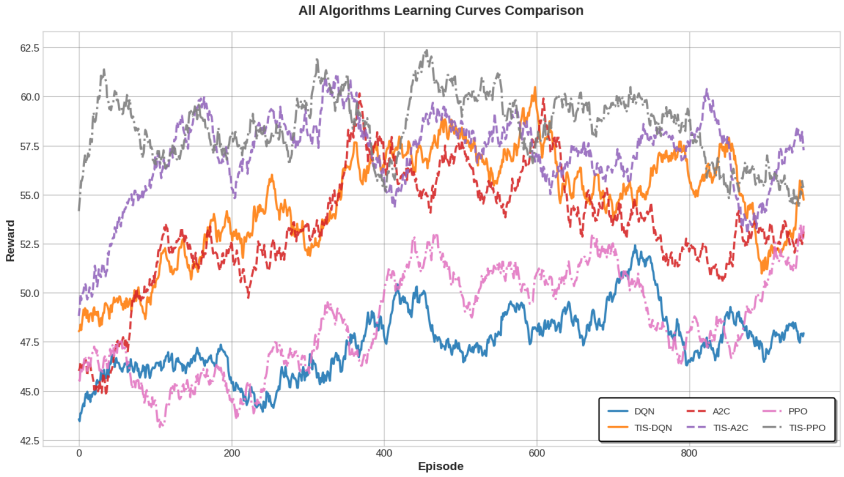


Figure 11: Training curve on synthetic experiment.

D.2 Recommendation via Offline Reinforcement Learning

In this section, we present a comprehensive evaluation of the results on the EasyRL4Rec benchmark. We conduct experiments across two real-world datasets: KuaiRand and MovieLens[1], whose detailed statistics are summarized in Table 5.

Datasets	Domain	Usage	#Users	#Items	#Interactions	# Categories
MovieLens	Movie	Train	6,040	3,952	800.4k	31
		Test	6,040	3,952	800.4k	
KuaiRand	Video	Train	26,210	7,551	1,141.6k	46
		Test	27,285	7,583	1,186.1k	

Table 5: Statistics of the datasets.

MovieLens. MovieLens is a classic recommender systems dataset, providing dense user-item interaction records.

Following the EasyRL4Rec protocol, all metrics except R_{best} are computed by averaging over the final 20% of training to reduce variance, while R_{best} records the maximum evaluation reward observed during training (evaluated every 1000 steps). As shown in Table 6, TIS-MDP yields modest but consistent gains for BCQ and CQL, with BCQ (TIS) improving R_{cumu} and Len by $+2.75\%$ and $+2.17\%$, respectively. CQL (TIS) achieves small improvements in reward and length while notably increasing diversity ($+7.79\%$) and novelty ($+9.52\%$). In contrast, CRR shows no change under the TIS formulation. This behavior is expected on MovieLens, where sequential interactions typically have very small time gaps; the gap-dependent scaling in TIS-MDP therefore becomes nearly constant (close to 1), causing the update to effectively reduce to the standard MDP backup, especially for CRR, whose advantage-weighted regression is already strongly anchored to the data.

Notably, 98% of interactions in MovieLens occur within the same day, leaving limited temporal variation for modeling interest decay. As a result, the improvement is relatively small. Never-

[1] https://grouplens.org/datasets/movielens/1m

Algorithm	R_{cumu}	Impv (%)	Len	Impv (%)	R_{avg}	Impv (%)	R_{best}	Impv (%)	Diversity	Impv (%)	Novelty	Impv (%)
BCQ	10.26	/	2.90	/	3.54	/	11.29	/	0.88	/	5.25	/
BCQ (TIS)	10.54	+2.75%	2.96	+2.17%	3.56	+0.56%	11.41	+1.03%	0.89	+0.56%	4.73	−9.88%
CQL	<u>10.68</u>	/	2.92	/	**3.66**	/	**12.27**	/	0.67	/	4.33	/
CQL (TIS)	**10.76**	+0.68%	2.99	+2.54%	3.59	−1.81%	**12.27**	+0.00%	0.73	+7.79%	4.74	+9.52%
CRR	9.97	/	3.42	/	2.91	/	11.06	/	**0.99**	/	**5.36**	/
CRR (TIS)	9.97	+0.00%	**3.42**	+0.00%	2.91	+0.00%	11.06	+0.00%	**0.99**	+0.00%	**5.36**	+0.00%

Table 6: Performance comparison on MovieLens. We evaluate the models on cumulative reward ($\mathbf{R_{cumu}}$), interaction length (**Len**), average reward per step ($\mathbf{R_{avg}}$), and maximum cumulative reward during the whole training epoch ($\mathbf{R_{best}}$), as well as recommendation diversity (**Diversity**) and novelty (**Novelty**). (TIS) indicates the use of the TIS-MDP formulation, and Impv (%) denotes the relative improvement over the baseline without TIS-MDP.

theless, the results still show a positive trend, providing a meaningful case study of TIS-MDP under limited temporal information.

KuaiRand. KuaiRand is a large-scale dataset collected from the video-sharing platform Kuaishou, distinguished by its inclusion of both unbiased random exposure logs and standard recommendation logs, which facilitates robust policy evaluation. The performance of various algorithms across different metrics on this dataset is presented in Table 7.

Algorithm	R_{cumu}	Impv (%)	Len	Impv (%)	R_{avg}	Impv (%)	R_{best}	Impv (%)	Diversity	Impv (%)	Novelty	Impv (%)
BCQ	<u>1.93</u>	/	4.80	/	<u>0.40</u>	/	<u>2.22</u>	/	0.85	/	3.13	/
BCQ (TIS)	**2.08**	+7.83%	5.14	+7.26%	**0.41**	+0.53%	**2.64**	+18.80%	0.88	+3.43%	3.18	+1.50%
CQL	0.85	/	3.13	/	0.27	/	1.57	/	0.48	/	4.82	/
CQL (TIS)	0.98	+15.12%	3.09	−1.07%	0.32	+16.36%	1.56	−0.75%	0.49	+2.29%	4.94	+2.45%
CRR	1.26	/	<u>5.83</u>	/	0.22	/	1.59	/	<u>0.93</u>	/	**6.34**	/
CRR (TIS)	1.27	+0.95%	**5.90**	+1.09%	0.22	−0.14%	1.56	−1.97%	**0.93**	+0.14%	<u>6.31</u>	−0.49%

Table 7: Performance comparison on KuaiRand. The metrics follow the same definitions as in Table 6.

As shown in Table 7, TIS-MDP improves all three offline backbones on KuaiRand, with gains manifesting differently across algorithms and metrics. For BCQ, which restricts actions to a behavior-supported set via a learned generative model, the main effect of TIS-MDP is to correct the critic backup under irregular gaps, reducing value over-propagation and yielding more reliable continuation decisions. This is reflected by consistent improvements in Len ($+7.26\%$) and $\mathrm{R_{cumu}}$ ($+7.83\%$), together with a boost in peak performance $\mathrm{R_{best}}$ ($+18.80\%$), and modest gains in diversity and novelty. For CQL, the critic is trained to be conservative on unseen actions. Accordingly, TIS-MDP yields the largest gains in $\mathrm{R_{avg}}$ ($+16.36\%$) and $\mathrm{R_{cumu}}$ ($+15.12\%$), while Len remains essentially unchanged. Diversity and novelty also increase slightly, suggesting improved generalization without sacrificing conservatism. For CRR, the actor is updated via advantage-weighted regression on behavior actions and is therefore strongly anchored to the dataset. As a result, modifying the TD backup has a limited influence on the learned policy.

D.3 Ablation and Sensitivity Studies

Power-law versus exponential survival. Table 8 reports the additional KuaiRand comparison. Power law has higher mean cumulative reward and interaction length, whereas exponential survival has higher mean average reward. These three-seed summaries do not establish a universal or statistically significant advantage.

Survival	R_{cumu}	Len	R_{avg}
Exponential	1.3809 ± 0.0736	4.5812 ± 0.1847	0.3051 ± 0.0092
Power law	1.3976 ± 0.0761	4.6820 ± 0.1905	0.3006 ± 0.0088

Table 8: KuaiRand survival-family comparison: mean $\pm$ standard deviation over three independently seeded runs.

To further investigate the TIS-MDP formulation, we conduct an ablation study on α, which controls the decay rate of the survival function. We evaluate the performance on the EasyRL4Rec benchmark, aggregate results across all baseline backbones and multiple τ ($\tau \in [5, 10, 30]$), and report the average relative improvement over the standard MDP formulation.

Hyperparam	R_{cumu}	Impv (%)	Len	Impv (%)	R_{avg}	Impv (%)	R_{best}	Impv (%)	Diversity	Impv (%)	Novelty	Impv (%)
$\alpha = 0.1$	**1.40**	**+3.94%**	4.49	-2.14%	**0.31**	**+4.88%**	**1.89**	**+5.33%**	0.74	-1.59%	4.57	-4.12%
$\alpha = 0.185$	1.38	$+2.49\%$	4.65	$+1.40\%$	0.30	$+0.44\%$	1.87	+4.15%	0.76	+1.10%	4.78	$+0.25\%$
$\alpha = 0.25$	1.38	$+2.38\%$	4.62	$+0.74\%$	0.30	$+1.41\%$	1.86	$+3.52\%$	**0.77**	**+1.13%**	4.75	-0.36%
$\alpha = 0.5$	1.38	$+2.38\%$	4.62	$+0.65\%$	0.30	$+1.51\%$	1.86	$+3.81\%$	0.77	$+0.76\%$	4.76	-0.07%
$\alpha = 1.0$	1.38	$+2.20\%$	**4.65**	**+1.46%**	0.30	$+0.20\%$	1.83	$+1.83\%$	0.76	$+1.03\%$	4.77	$+0.10\%$
$\alpha = 5.0$	1.34	-0.46%	4.49	-2.00%	0.31	$+2.79\%$	1.78	-0.84%	0.77	-1.09%	**4.80**	**+0.61%**

Table 9: Impact of hyperparameter α on model performance. We report the average values and relative improvements (Impv %) across all baselines on the KuaiRand dataset. The best values and improvements in each column are highlighted in bold.

As shown in Table 9, TIS-MDP yields positive improvements for a wide range of α, indicating that the gains are not overly sensitive to precise decay calibration. Performance degrades only when α is very large (e.g., $\alpha = 5$), which corresponds to overly fast decay and can over-truncate temporal credit assignment. Interestingly, the best overall R_{cumu} and R_{avg} are achieved at $\alpha = 0.1$, while values around $\alpha \in [0.185, 1.0]$ provide stable and competitive improvements and better peak performance (R_{best}). Notably, $\alpha = 0.25$ is close to the decay ratio estimated from KuaiRand, yet it is not universally optimal, suggesting that the best algorithmic discounting strength depends not only on the empirical decay trend.

We further conduct sensitivity studies on the scaling parameter τ and the base discount γ_{base}, and report the aggregated results on KuaiRand in Tables 10 and 11.

Hyperparam	R_{cumu}	Impv (%)	Len	Impv (%)	R_{avg}	Impv (%)	R_{best}	Impv (%)	Diversity	Impv (%)	Novelty	Impv (%)
$\tau = 5$	**1.3962**	**+3.5%**	**4.6597**	**+1.6%**	0.3016	+1.3%	1.8101	+0.8%	**0.7622**	**+1.1%**	4.7801	+0.3%
$\tau = 10$	1.3732	+1.8%	4.5817	-0.1%	0.3026	+1.7%	**1.8960**	**+5.6%**	0.7560	+0.3%	4.7533	-0.3%
$\tau = 30$	1.3738	+1.9%	4.6038	+0.4%	0.3018	+1.4%	1.8846	+5.0%	0.7608	+0.9%	4.7546	-0.2%
$\tau = 50$	1.3449	-0.3%	4.5297	-1.2%	0.3010	+1.1%	1.7983	+0.2%	0.7534	-0.1%	**4.7912**	**+0.5%**
$\tau = 100$	1.3682	+1.4%	4.5805	-0.1%	**0.3027**	**+1.7%**	1.8143	+1.1%	0.7606	+0.9%	4.7719	+0.1%

Table 10: Impact of hyperparameter τ on model performance. We report the average values and relative improvements (Impv %) over the no-gap baseline on the KuaiRand dataset. The best values and improvements in each column are highlighted in bold.

Hyperparam	R_{cumu}	Impv (%)	Len	Impv (%)	R_{avg}	Impv (%)	R_{best}	Impv (%)	Diversity	Impv (%)	Novelty	Impv (%)
$\gamma_{\text{base}} = 0.5$	1.3534	+0.3%	4.4775	-2.3%	0.3100	+4.2%	1.7622	-1.8%	0.7343	-2.6%	4.7506	-0.3%
$\gamma_{\text{base}} = 0.7$	1.3330	-1.2%	4.4000	-4.0%	**0.3123**	**+4.9%**	1.8276	+1.8%	0.7340	-2.7%	**4.8222**	**+1.2%**
$\gamma_{\text{base}} = 0.8$	1.3497	+0.1%	4.5145	-1.5%	0.3053	+2.6%	1.7918	-0.2%	0.7532	-0.1%	4.7859	+0.4%
$\gamma_{\text{base}} = 0.9$	**1.3942**	**+3.4%**	**4.6845**	**+2.2%**	0.3009	+1.1%	**1.8609**	**+3.7%**	**0.7649**	**+1.4%**	4.8008	+0.7%

Table 11: Impact of base discount γ_{base} on model performance. We report the average values and relative improvements (Impv %) over the no-gap baseline on the KuaiRand dataset. The best values and improvements in each column are highlighted in bold.

The sensitivity results show that TIS-MDP remains beneficial under a range of temporal scaling and discounting choices. Smaller τ values give stronger gains in cumulative reward and interaction length, while moderate-to-large τ values can still improve peak reward or novelty. For γ_{base}, a larger value such as 0.9 better preserves long-term credit assignment and achieves the strongest cumulative reward and length, whereas smaller values emphasize immediate per-step reward more strongly. Overall, these results indicate that TIS-MDP is not dependent on a single narrow hyperparameter setting.

E Proofs for Section 5

Proof of Proposition 5.1. The bound $\|Q^\pi\|_\infty \le \frac{R_{\max}}{1-\gamma_{\text{base}}}$ follows from $|r| \le R_{\max}$ and the fact that $\gamma_{\text{base}} \in (0,1)$. For any (s,a), the following inequality holds that

$$\begin{aligned}
&|(T^\pi Q_1)(s,a) - (T^\pi Q_2)(s,a)| \\
=&\left|\mathbb{E}\left[\gamma_{\text{base}} s(\Delta) \mathbb{E}_{a'}\big(Q_1 - Q_2\big)(s',a') \mid s,a\right]\right| \\
\le&\mathbb{E}\left[\gamma_{\text{base}} s(\Delta) \left|\mathbb{E}_{a'}\big(Q_1 - Q_2\big)(s',a')\right| \mid s,a\right] \\
\le&\mathbb{E}\left[\gamma_{\text{base}} s(\Delta) \mid s,a\right] \|Q_1 - Q_2\|_\infty \le \gamma_{\text{base}} \|Q_1 - Q_2\|_\infty,
\end{aligned}$$

where the first inequality holds due to Jensen's inequality. □

Proof of Theorem 5.3. For the difference, we first obtain that:

$$\begin{aligned}
&\|Q_N^\pi - Q^\pi\|_\infty = \|T_N^\pi Q_N^\pi - T^\pi Q^\pi\|_\infty \\
\le&\|T_N^\pi Q_N^\pi - T_N^\pi Q^\pi\|_\infty + \|T_N^\pi Q^\pi - T^\pi Q^\pi\|_\infty \\
\le&\gamma_{\text{base}} \|Q_N^\pi - Q^\pi\|_\infty + \|T_N^\pi Q^\pi - T^\pi Q^\pi\|_\infty,
\end{aligned}$$

where the last inequality uses the γ_{base}-contraction of T_N^π. Rearranging the inequality, we have

$$\|Q_N^\pi - Q^\pi\|_\infty \le \frac{1}{1-\gamma_{\text{base}}} \|T_N^\pi Q^\pi - T^\pi Q^\pi\|_\infty. \tag{18}$$

Recall the definition of T_u^π and the fact that $s_u(\Delta) \in (0,1]$, we have that

$$(T_u^\pi Q^\pi)(s,a) \in r(s,a) + [-\gamma_{\text{base}} \|Q^\pi\|_\infty,\ \gamma_{\text{base}} \|Q^\pi\|_\infty].$$

which implies the range bound

$$\text{range}\big((T_u^\pi Q^\pi)(s,a)\big) \le 2\gamma_{\text{base}} \|Q^\pi\|_\infty.$$

Then applying Hoeffding's inequality yields that for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\|T_N^\pi Q^\pi - T^\pi Q^\pi\|_\infty \le 2\gamma_{\text{base}} \|Q^\pi\|_\infty \sqrt{\frac{\log\big(2|\mathcal{S}||\mathcal{A}|/\delta\big)}{2N}}.$$

Combining with (18) and the boundedness of Q-functions completes the proof. □

F Limitations

CTTC is descriptive rather than causal and may reflect exposure policies, activity schedules, seasonality, and negative sampling. The gap-only continuation model is a population-level approximation; its contraction property does not validate that approximation. Comparisons with hard temporal cutoffs, historical-gap state features, and misspecified survival dynamics remain open.

TIS-MDP introduces temporal-aware value updates, which require a recommender system to expose, log, and update temporal signals during both policy learning and serving. This requirement imposes additional constraints on the system architecture. Therefore, our current evaluation focuses on offline datasets and simulators, and does not yet include online A/B experiments in a live recommender system.

In addition, the benefit of TIS-MDP depends on the amount of temporal information carried by user interaction gaps. When user interactions are very dense, the realized gaps contain limited temporal information and the survival factor $s(\Delta)$ becomes close to a constant across transitions. In such cases, the temporal-aware Bellman update degenerates toward a standard RL update.