MODELING AND DEBIASING TEMPORAL POPULARITY IN RECOMMENDATION WITH COUNTING PROCESSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Implicit-feedback recommender systems are trained on logged interactions in which observed choices conflate user-specific preference with time-varying item popularity. We formulate timestamped user-item interactions as counting processes and decompose the conditional interaction intensity into user activity, dynamic item popularity, and user-specific utility. This formulation reveals an identifiability issue: without additional structure, item-time effects shared across users can be absorbed into the utility function. We show that centering the utility across users resolves this ambiguity up to ranking-equivalent shifts. We further establish that unmodeled popularity can induce a shared direction in user-context representations, leading to geometric homogenization and degraded uniformity. Motivated by this analysis, we propose a general counting process framework that jointly addresses the challenging tasks of debiasing temporal item popularity and identifying user–item-specific effects. Experiments on synthetic and real-world datasets support the theoretical analysis and show consistent improvements across diverse recommendation backbones.

1 INTRODUCTION

Implicit-feedback recommender systems learn user preferences from behavioral signals such as clicks, views, and purchases. These interactions, however, are shaped not only by user preference but also by population-level factors such as item popularity and temporary surges in attention. When popularity evolves over time, recommendation models may attribute popularity-driven interactions to user preference, systematically distorting the learned utility and ranking (Abdollahpouri et al., 2019; Klimashevskaia et al., 2023).

Existing studies mitigate popularity bias through causal disentanglement, representation correction, contrastive learning, and adaptation to temporal popularity shifts (Zheng et al., 2021a; Zhao et al., 2022; Yang et al., 2023; Cai et al., 2024; Yoo et al., 2025; Liu et al., 2025). Regarding the population bias, we focus on a more fundamental issue underlying these approaches: whether user-specific preference and time-varying item popularity can be separated from the same observed interactions. If the two effects are not independently identified, it will be impossible to estimate an individual's intrinsic preference for an item.

This problem extends beyond merely complicating recommendations based on intrinsic user–item preferences once the trend subsides. As our study confirms, recommender systems that fail to disentangle the popularity effect cause the embeddings of users who interacted with the trending item to collapse into homogenization, ultimately degrading overall recommendation performance (Zheng et al., 2021b; Zhang et al., 2023). Consequently, decoupling item popularity bias within a recommendation model is crucial for mitigating recommendation bias across the system.

We formulate timestamped user–item interactions as a counting process (Andersen et al., 2012) and decompose their conditional intensity into user activity, time-varying item popularity, and user-specific utility. The formulation reveals that popularity and utility are not automatically identifiable without an additional constraint. Furthermore, it shows that the identifiability issue essentially causes embedding collapse (Chen et al., 2024; Lin et al., 2025; Yang et al., 2026): when an unmodeled item–time effect is absorbed into utility, the absorbed effect can induce a direction shared across user contexts, concentrating representations and reducing uniformity. Based on this analysis, a general framework that explicitly separates temporal popularity from user-specific utility is proposed.

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.

The proposed framework combines Hawkes-based popularity modeling, importance-corrected estimation, and a centering penalty, while remaining applicable to diverse recommendation backbones.

Experiments on artificial and real-world datasets support both the theoretical analysis and the proposed framework. Controlled experiments show that stronger popularity confounding leads to representation concentration and poorer recovery of the underlying preference ranking when popularity is not modeled separately. Across three real-world datasets and diverse recommendation backbones, the proposed framework substantially reduces representation concentration and improves recommendation performance.

The main contributions of this work are threefold. First, we rigorously formalize timestamped implicit feedback from a counting-process perspective. This formulation exposes the fundamental identifiability ambiguity between temporal popularity and user-specific utility, ensuring ranking-equivalent identifiability via centering. Second, we establish a theoretical connection between this temporal confounding and representation geometry, showing how an absorbed item–time effect induces a shared direction across user contexts and ultimately triggers representation homogenization. Finally, guided by these theoretical insights, we propose a scalable estimation framework that unifies Hawkes-based popularity modeling (Zadeh & Sharda, 2014; Lalor & Swishchuk, 2025; Park et al., 2025), importance-corrected conditional choice estimation, and stochastic centering, demonstrating substantial empirical gains across diverse recommendation architectures and real-world datasets.

2 Related Work

Mitigating Popularity Bias. Popularity bias has been approached through causal factorization, representation correction, and modeling of temporal distribution shifts. Zheng et al. (2021a) separate user interest from conformity using distinct causal representations, while Zhao et al. (2022) further distinguish user–item matching, static item quality, and dynamic conformity in observed interactions. Other studies focus on representation-level correction: Cai et al. (2024) strengthen unpopular-item representations through popularity-aware alignment, and Liu et al. (2025) suppress the popularity-dominant direction induced by BPR optimization. Temporal variation in popularity is considered by Yoo et al. (2025), who address shifts in the item-popularity distribution. Related ideas have also been developed for sequential recommendation. Yang et al. (2023) disentangle preference and conformity signals through contrastive learning, Kim et al. (2023) jointly address long-tailed users and items, and Wu et al. (2023) incorporate dynamic popularity effects through causal reasoning. More recently, Kim & Kang (2025) integrate popularity information into sequential modeling to improve both recommendation accuracy and diversity.

Representation Geometry and Uniformity. Representation geometry is closely tied to representation quality. Wang & Isola (2020) characterize contrastive learning through alignment and uniformity on the hypersphere, while Bardes et al. (2021) show that collapse yields non-informative embeddings and motivate preserving feature variance. In recommendation, Wang et al. (2022) relate alignment and uniformity to collaborative-filtering performance, and Yu et al. (2021) show that more user and item representations can improve recommendation and alleiviate popularity bias. Lee et al. (2023) extend this view to biased implicit feedback, while Ouyang et al. (2024) and Zhang et al. (2024) improve generalization and reduce redundant solutions through uniformly-based objectives. Guo et al. (2023) further show that collapsed recommendation embeddings occupy a low-dimensional subspace, and Ma et al. (2025) emphasize balancing uniformity and alignment to preserve diverse user sub-interests. These findings indicate that excessive concentration can reduce the discriminative variation required for personalized ranking.

3 Problem Definition

This section formalizes the learning target and the probabilistic structure of timestamped implicit feedback under time-varying popularity. We then characterize the separation of popularity and utility from the observed choice distribution and examine its implications for the geometry of learned user-context representations.

3.1 DATA, CONTEXT, AND TARGET RANKING

Denote the sets of users and items as $\mathcal{U} = \{u_1, \ldots, u_{|\mathcal{U}|}\}$ and $\mathcal{V} = \{v_1, \ldots, v_{|\mathcal{V}|}\}$, respectively. The observed implicit feedback events are denoted by $\mathcal{D}_n = \{(u_i, v_i, t_i)\}_{i=1}^n$ where $u_i \in \mathcal{U}$, $v_i \in \mathcal{V}$ and $t_i \in [0, T]$ indicate that user u_i interacted with item v_i at time t_i. At time t, the context of user u is represented by $x_u(t) = (u, \mathcal{S}_u(t), t)$, where $\mathcal{S}_u(t) = \{(v_i, t_i) : (u_i, v_i, t_i) \in \mathcal{D}_n, u_i = u, t_i < t\}$ collects the interactions of user u observed before time t. Let $\mathcal{F}_{t-}^u = \sigma(\mathcal{S}_u(t))$ denote the information generated by user u's interaction history strictly before time t and $\mathcal{F}_{t-} = \sigma(\cup_{u\in\mathcal{U}}\mathcal{S}_u(t))$. Let $f^*(x_u(t), v)$ be the ground-truth utility of item v under context $x_u(t)$. The goal of recommendation is to rank items according to $f^*(x_u(t), v)$.

Definition 1 (Ranking-equivalence class of utility function) *For a utility function f^*, define its ranking-equivalence class as*

$$[f^*] = \big\{f : f\big(x_u(t), v\big) = f^*\big(x_u(t), v\big) + C\big(x_u(t)\big),\ \forall x_u(t),\ v \in \mathcal{V}\big\}.$$

where $C\big(x_u(t)\big)$ is an arbitrary function that is independent of the item v.

For a fixed context $x_u(t)$, the additive term $C(x_u(t))$ is constant across candidate items and therefore does not affect their relative ordering. Thus, any ranking-equivalent utility $f \in [f^*]$ is sufficient to recover the optimal item ranking induced by f^*.

3.2 PROBABILISTIC MODEL OF LOGGED INTERACTIONS

In this work, the logged user–item interactions are formulated as a counting process. For a user–item pair (u, v), let $N_{uv}(t) = \sum_{i=1}^n \mathbb{I}(u_i = u, v_i = v, t_i \le t)$ count the number of interactions between user u and item v observed by time t. The user–item counting process is characterized by the conditional intensity

$$\lambda_{uv}^0(t|\mathcal{F}_{t^-}^u) = \lim_{\Delta t \downarrow 0} \frac{\Pr\left(N_{uv}(t+\Delta t) - N_{uv}(t) = 1 | \mathcal{F}_{t^-}^u\right)}{\Delta t}.$$

Assuming that two or more interactions occur with probability $o(\Delta t)$ within an interval of length Δt, the conditional probability that item v is selected, given that user u generates an interaction at time t, is $p_0(v|x_u(t)) = \lambda_{uv}^0(t|\mathcal{F}_{t^-}^u) / \sum_{k\in\mathcal{V}} \lambda_{uk}^0(t|\mathcal{F}_{t^-}^u)$. Derivation of this conditional distribution is given in Appendix A.1.

Logged interactions are affected not only by user preference but also by time-varying item attractiveness and popularity (Abdollahpouri et al., 2019; Klimashevskaia et al., 2023). We therefore model the user–item intensity as

$$\lambda_{uv}^0(t|\mathcal{F}_{t-}^u) = r_0\big(x_u(t)\big)\pi_0(v|t)\exp\big\{f^*\big(x_u(t), v\big)\big\},$$

where $r_0(x_u(t)) > 0$ is a context-dependent activity factor, and $\pi_0(v|t) > 0$ is the time-varying item popularity. Substituting this intensity into the conditional distribution $p_0(v|x_u(t))$ gives

$$p_0\big(v|x_u(t)\big) = \frac{\pi_0(v|t)\exp\big(f^*\big(x_u(t), v\big)\big)}{\sum_{k\in\mathcal{V}} \pi_0(k|t)\exp\big(f^*\big(x_u(t), k\big)\big)}. \tag{1}$$

The multiplicative activity factor $r_0(x_u(t))$ scales the overall event rate but cancels when conditioning on the occurrence of an interaction. Thus, equation 1 separates the observed interaction into the user-specific utility $f^*(x_u(t), v)$ and the time-varying item popularity $\pi_0(v|t)$. When $\pi_0(v|t)$ is uniform over items, equation 1 reduces to the preference-only softmax model, which corresponds to the Luce choice model Luce et al. (1959).

3.3 IDENTIFIABILITY

The conditional item-choice model in equation 1 represents the observed choice distribution through the time-varying item popularity $\pi_0(v|t)$ and the user-specific utility $f^*(x_u(t), v)$. To study whether these quantities can be separately identified, the model can be written in log form as $\log p_0(v|x_u(t)) = \log \pi_0(v|t) + f^*(x_u(t), v) + a_u(t)$, where $a_u(t) =$

$-\log \sum_{k\in\mathcal{V}} \pi_0(k|t)\exp\{f^*(x_u(t),k)\}$. This additive structure is analogous to a multi-way ANOVA decomposition, with item popularity shared across users and utility capturing user-specific variation. However, these two quantities are not separately identifiable from the conditional distribution without an additional constraint. For any item- and time- dependent function $b_v(t)$, the transformations $\log \pi_0(v|t) \mapsto \log \pi_0(v|t) + b_v(t)$, $f^*(x_u(t)) \mapsto f^*(x_u(t)) - b_v(t)$ leaves $p_0(v|x_u(t))$ unchanged. We resolve this ambiguity by centering the utility across users.

Assumption 1 (Centered Utility) *For every item v and time t, $\sum_{u\in\mathcal{U}} f(x_u(t),v) = 0,\ \forall v \in \mathcal{V}, t \in [0,T]$.*

Proposition 1 (Identifiability under Centering) *Suppose Assumption 1 holds, and $\pi_0(v|t) > 0$ and $p_0(v|x_u(t)) > 0$ for all $v \in \mathcal{V}$. For any $(\widetilde{\pi}, \widetilde{f})$ satisfying the same conditions,*

$$p_{\widetilde{\pi},\widetilde{f}}\big(v|x_u(t)\big) = p_{\pi_0,f^*}\big(v|x_u(t)\big), \quad \forall u \in \mathcal{U},\ v \in \mathcal{V},\ t \in [0,T],$$

implies

$$\widetilde{\pi}(\cdot|t) = \pi_0(\cdot|t), \qquad \widetilde{f} \in [f^*].$$

The proposition shows that centering rules out item-specific reallocations between popularity and utility. The only remaining freedom is a context-dependent additive shift in the utility, which leaves the induced item ranking unchanged.

3.4 Popularity Bias and Representation Geometry

The non-identifiability in equation 1 also has implications for the geometry of learned user-context representations. Without resolving this non-identifiability through an explicit separation of popularity and utility or an appropriate identification constraint, an item-time component shared across users may be absorbed into the utility function. When this shared component is encoded through a common direction in the context representation space, user contexts become increasingly concentrated around that direction.

Let $\mathcal{S}^{d-1}$ be the unit sphere in d-dimensional space, and $\mathbf{h}_{u,t}, \mathbf{z}_v \in \mathbb{R}^d$ denote the representations of user context $x_u(t)$ and item v, respectively. Let $\mathbf{c}_t = \rho\widetilde{\mathbf{c}}_t$, where $\widetilde{\mathbf{c}}_t \in \mathcal{S}^{d-1}$ denote a direction shared across user contexts at time t, with $\rho > 0$ denoting its magnitude. The following results show how such a shared component can arise from temporal popularity and characterize its effect on representation geometry.

Lemma 1 (Popularity-Induced Shared Component) *Suppose $f^*(x_u(t),v) = \mathbf{h}_{u,t}^\top \mathbf{z}_v$, and, for a fixed t,*

$$(\beta_t, \mathbf{c}_t) \in \operatorname*{arg\,min}_{\beta\in\mathbb{R},\mathbf{c}\in\mathbb{R}^d} \sum_{v\in\mathcal{V}} \big[\log \pi_0(v|t) - \beta - \mathbf{c}^\top \mathbf{z}_v\big]^2.$$

Then the utility learned without explicitly modeling the popularity term has the projected form

$$\widehat{f}\big(x_u(t),v\big) = \big(\mathbf{h}_{u,t} + \mathbf{c}_t\big)^\top \mathbf{z}_v.$$

Theorem 1 (Popularity-Induced Homogenization of User-Context Representations) *Define $B_t = \max_{u\in\mathcal{U}} \parallel \mathbf{h}_{u,t} \parallel_2$. Suppose that an item-time component absorbed into the utility induces the shifted representation*

$$\widetilde{\mathbf{h}}_{u,t} = \frac{\mathbf{h}_{u,t} + \rho\widetilde{\mathbf{c}}_t}{\parallel \mathbf{h}_{u,t} + \rho\widetilde{\mathbf{c}}_t \parallel_2}, \qquad \rho > B_t.$$

Then, for any two users $u_1, u_2 \in \mathcal{U}$,

$$\Big\|\widetilde{\mathbf{h}}_{u_1,t} - \widetilde{\mathbf{h}}_{u_2,t}\Big\|_2^2 \le \frac{4B_t^2}{\rho^2}.$$

Theorem 1 shows that a shared item–time component absorbed into the utility contracts user-context representations at the rate $O(\rho^{2-})$ as its magnitude increases. This contraction also implies a loss of representation uniformity in the sense of Wang & Isola (2020): the corresponding uniformity loss

approaches zero as the shared component becomes dominant. The formal result and its proof are provided in Appendix A.2.

Thus, the non-identifiability described in Section 3.3 can manifest as a measurable concentration of the learned representation space when the shared item–time effect is encoded in user contexts. This provides a geometric motivation for explicitly modeling such shared effects through the popularity component while imposing the centering constraint.

4 Proposed Method

The identifiability analysis suggests a direct estimation strategy: the item-time baseline should be modeled through $\pi_0(v|t)$, while the remaining user-specific deviation should be captured by $f(x_u(t), v)$. We train these two components jointly through the conditional item-choice formulation in equation 1. Scalable estimation consistent with the identifiability convention is achieved by combining three components: a Hawkes intensity for item popularity, a sampled approximation to the conditional choice likelihood, and a centering penalty on the utility component.

4.1 Hawkes Intensity for Dynamic Item Popularity

Motivated by the branching structure of self-exciting process and the temporal propagation of popular item interactions, we use a Hawkes process to model the intensity of item-level point processes (Rizoiu et al., 2017; Laub et al., 2021). Let $\mathcal{H}_t^v = \{t_j : v_j = v,\ t_j < t\}$ be the set of previous event times for item v. The intensity of item v at time t is parameterized as

$$\lambda_v(t; \Phi) = \mu(z_v; \phi) + \sum_{t_j \in \mathcal{H}_t^v} \alpha(z_v; \varphi) \exp\{-\xi(t - t_j)\},$$

where $\Phi = \{\mathbf{z}, \phi, \varphi, \xi\}$, $\mathbf{z} \in \mathbb{R}^{d \times |\mathcal{V}|}$ is an item representation, $\mu(\cdot; \phi) : \mathbb{R}^d \mapsto \mathbb{R}_+$ denotes the background event rate, $\alpha(\cdot; \varphi) : \mathbb{R}^d \mapsto \mathbb{R}_+$ controls the magnitude of the excitation caused by past events, and $\xi \in \mathbb{R}_+$ determines how quickly this excitation decays. The positive constraints can be imposed by applying a positive activation function to the network outputs. The item popularity is obtained by normalizing the item intensities: $\pi_\Phi(v|t) = \lambda_v(t; \Phi) / \sum_{k \in \mathcal{V}} \lambda_k(t; \Phi)$. Thus, $\pi_\Phi(v|t)$ represents the population-level tendency for item v to be selected at time t, before accounting for user-specific preference.

4.2 Importance-Corrected Choice Loss

The utility function $f_\Psi(x_u(t), v)$ is parameterized by the underlying recommendation backbone. Conventional recommenders compute utility from user–item representation matching, whereas sequential recommenders derive it from a history-dependent context representation and the candidate item representation. This formulation allows the proposed framework to be applied to both static and sequential recommendation architectures without altering their underlying utility models. Given the utility score $f_\Psi(x_u(t), v)$ and the Hawkes-based popularity $\pi_\Phi(v|t)$, the conditional item-choice model in equation 1 can be written directly in terms of the item intensities as

$$p_{\Phi,\Psi}\big(v|x_u(t)\big) = \frac{\lambda_v(t; \Phi) \exp\big\{f_\Psi\big(x_u(t), v\big)\big\}}{\sum_{k \in \mathcal{V}} \lambda_k(t; \Phi) \exp\big\{f_\Psi\big(x_u(t), k\big)\big\}}. \tag{2}$$

The normalization of $\pi_\Phi(v|t)$ therefore cancels from the conditional choice probability, allowing the Hawkes intensity to serve directly as the unnormalized item-time popularity component.

The remaining bottleneck is the normalizer $Z_i = \sum_{k \in \mathcal{V}} \lambda_k(t_i; \Phi) \exp\{f_\Psi(x_{u_i}(t_i), k)\}$. Following importance-sampling approaches for large normalization terms (Bengio & Senécal, 2008), an importance-corrected estimator is used to avoid summing over the entire item set. For each observed interaction (u_i, v_i, t_i), M auxiliary items $\bar{v}_{ij} \sim q_i^-(\cdot)$ are sampled independently from a proposal distribution on $\mathcal{V} \backslash \{v_i\}$. The normalizer is then estimated by $\widehat{Z}_i = \lambda_{v_i}(t_i; \Phi) \exp\{f_\Psi(x_{u_i}(t_i), v_i)\} + \frac{1}{M} \sum_{j=1}^{M} \lambda_{\bar{v}_{ij}}(t_i; \Phi) \exp\{f_\Psi(x_{u_i}, \bar{v}_{ij})\} / q_i^-(\bar{v}_{ij})$. This estimator satisfies $\mathbb{E}[\widehat{Z}_i] = Z_i$. The resulting sampled choice loss is

$$\ell_{\text{choice}}(\Phi, \Psi) = -\sum_{i=1}^{n} \Big[\log \lambda_{v_i}(t_i; \Phi) + f_\Psi\big(x_{u_i}(t_i), v_i\big) - \log \widehat{Z}_i\Big].$$

4.3 CENTERING PENALTY

Following the standard centering principle used to identify additive and interaction effects (Lengerich et al., 2019), the population penalty is defined as

$$\mathcal{R}_{\mathrm{cen}}(\Psi) = \int_0^T \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \left[\frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} f_\Psi\big(x_u(t), v\big) \right]^2 d\nu(t),$$

where ν is a reference measure over time. Minimizing this penalty encourages the utility to satisfy Assumption 1, preventing the item-time effect shared across users from being absorbed into f_Ψ. Exact computation requires averaging over all users and items. Instead, motivated by $\mathcal{R}_{\mathrm{cen}}(\Psi)$ being the square of the first moment, we introduce the second-order U-statistics (Hoeffding, 1948) as the estimator of the moment. For each sampled event time $t \in \mathcal{T}$, two independent user sets $\mathcal{U}_t^{(1)}$ and $\mathcal{U}_t^{(2)}$ are drawn together with a candidate item set $\mathcal{C}_t$. The centering loss is approximated by

$$\ell_{\mathrm{cen}}(\Psi) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \frac{1}{|\mathcal{C}_t|} \sum_{v \in \mathcal{C}_t} \left(\frac{1}{|\mathcal{U}_t^{(1)}|} \sum_{u \in \mathcal{U}_t^{(1)}} f_\Psi\big(x_u(t), v\big) \right) \left(\frac{1}{|\mathcal{U}_t^{(2)}|} \sum_{u \in \mathcal{U}_t^{(2)}} f_\Psi\big(x_u(t), v\big) \right).$$

Specifically, the independence between the two user sets ensures that the cross-terms asymptotically vanish, leaving only the desired population variance term without finite-sample bias.

4.4 OPTIMIZATION AND INFERENCE

The popularity model π_Φ and the utility model f_Ψ are optimized jointly. Combining the importance-corrected choice loss with the centering loss gives

$$\mathcal{L}(\Phi, \Psi) = \ell_{\mathrm{choice}}(\Phi, \Psi) + \gamma\, \ell_{\mathrm{cen}}(\Psi),$$

where $\gamma \geq 0$ controls the strength of the centering penalty. To bound the magnitude of the stochastic error in $\ell_{\mathrm{cen}}(\Psi)$, γ is constrained to a carefully chosen bounded range. Through the formulation in equation 2, the Hawkes parameters Φ enter the choice loss directly through the item intensities $\lambda_v(t; \Phi)$. Thus, the same objective jointly updates the popularity and utility components without requiring a separate point-process likelihood. The complete stochastic optimization procedure, including auxiliary-item sampling and the two-sample estimator for the centering penalty, is provided in Appendix A.3.

At test time, the fitted popularity and utility components are combined into a single ranking score. Because the Hawkes intensity depends only on item-side representations and past event times, the popularity distribution $\pi_\Phi(v|t)$ can be evaluated exactly over the full item catalog at inference. The utility term $f_\Psi\big(x_u(t), v\big)$ is likewise evaluated for all candidate items by the recommendation backbone. Items are ranked by

$$s\big(x_u(t), v\big) = \eta \log \pi_\Phi(v|t) + (1 - \eta) f_\Psi\big(x_u(t), v\big), \qquad \eta \in [0, 1], \tag{3}$$

after masking previously consumed items. $\eta = 0$ recovers the fully popularity-debiased ranking induced by $[f^*]$, while $\eta = 1$ reduces to ranking by population popularity alone; η is thus treated as an inference-time parameter, allowing a single trained model to be evaluated across this spectrum without retraining.

5 NUMERICAL EXPERIMENTS

This section empirically validates the theoretical results of temporal popularity confounding in Section 3.4 and evaluates the proposed framework. We first analyze representation geometry on artificial data, and then evaluate recommendation performance, representation uniformity, and parameter sensitivity on three real-world datasets.

5.1 EXPERIMENTS ON ARTIFICIAL DATA

To directly examine the representation-level effect predicted by Theorem 1, controlled experiments are conducted on artificial timestamped interaction data. The strength of temporal popularity confounding is systematically varied while the ground-truth user preference remains fixed. This allows

us to isolate how popularity absorbed into the utility model affects representation uniformity and the recovery of preference-based rankings.

Experimental Setup. Interactions are generated from $p_\rho(v|x_u(t)) \propto \pi_0(v|t)^\rho \exp\{f^*(u,v)\}$, where $f^*(u,v)$ denotes the ground-truth preference and $\rho \in \{0, 0.5, 1, 2, 4\}$ controls the influence of time-varying item popularity. For each ρ, we train a vanilla matrix factorization model and the proposed model on the same interactions. We evaluate (i) the uniformity loss of normalized user representations (Wang & Isola, 2020) and (ii) Recall@10 for recovering the ground-truth preference ranking. Each experiment is repeated over 30 random seeds. Detailed data-generation and training settings are provided in Appendix A.5.

Effect of Popularity Confounding. Figure 1 shows that the two models respond differently as the strength of popularity confounding increases. For Vanilla, the uniformity loss consistently moves toward 0 with increasing ρ, indicating that the learned user representations become progressively more concentrated. This change in representation geometry is accompanied by a substantial overall degradation in ground-truth preference recovery. Although Vanilla performs favorably when popularity confounding is absent, its Recall@10 rapidly decreases once the observed interactions are increasingly influenced by temporal popularity.

In contrast, the proposed method remains stable across the entire range of ρ in both representation uniformity and preference recovery. This behavior is consistent with the representation-contraction mechanism in Theorem 1 and its uniformity implication formalized in Appendix A.2. When popularity is not modeled separately, the item-time component shared across users can be absorbed into the utility representation, thereby homogenizing user representations and distorting the underlying preference ranking. Explicitly separating temporal popularity from utility and imposing the centering constraint substantially reduces this sensitivity to popularity confounding.

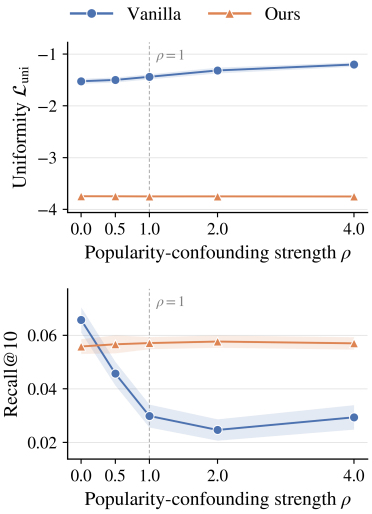


Figure 1: Effect of popularity-confounding strength ρ on representation uniformity and ground-truth preference recovery.

5.2 Experiments on Real Data

To evaluate the proposed framework under realistic temporal dynamics, experiments are conducted on three real-world datasets. The evaluation compares the proposed framework with existing popularity-debiasing methods, and further analyzes its effects on representation geometry and sensitivity to key hyperparameters.

Experimental Setup. Experiments are conducted on three timestamped interaction datasets: MicroVideo, MovieLens-1M, and KuaiRand (Zhao et al., 2023; Harper et al., 2016; Gao et al., 2022). Interactions are chronologically split into training, validation, and test sets in an 8:1:1 ratio. Performance is evaluated under a full-ranking protocol using Recall@10 and NDCG@10, with previously observed items excluded from the candidate set. The proposed framework is evaluated with six recommendation backbones, including both conventional and sequential architectures, and compared with representative popularity-debiasing methods. Each experiment is repeated with four random seeds, and results are reported as mean $\pm$ standard deviation. Detailed experimental settings are provided in Appendix A.6.

Performance across Backbones and Baselines. Table 1 shows that applying the proposed framework generally improves the underlying backbones, with the clearest gains on All and Head items across the three datasets. These improvements are observed across architectures with different mechanisms for modeling user behavior, indicating that the benefit is not specific to a particular backbone. Compared with existing popularity-bias mitigation methods, the proposed variants achieve competitive performance across datasets and evaluation groups, although they do not uniformly outperform the strongest baseline in every setting.

Table 1: Recommendation performance on all, Head, and Tail items, defined by recent three-day interaction frequency. **Bold** compares each backbone with +(Ours); $\star$ and † mark the best and second-best results.

Dataset	Approach	Method	All Items		Head Items		Tail Items	
			Recall@10(↑)	NDCG@10(↑)	Recall@10(↑)	NDCG@10(↑)	Recall@10(↑)	NDCG@10(↑)
Micro Video	Debiased	DICE	$0.0565_{\pm 0.0008}$	$0.0281_{\pm 0.0009}$	$0.0753_{\pm 0.0019}$	$0.0374_{\pm 0.0014}$	$0.0475_{\pm 0.0004}$	$0.0238_{\pm 0.0007}$
		TIDE	$0.0477_{\pm 0.0008}$	$0.0240_{\pm 0.0002}$	$0.0662_{\pm 0.0017}$	$0.0321_{\pm 0.0004}$	$0.0389_{\pm 0.0008}$	$0.0202_{\pm 0.0003}$
		MELT	$0.0005_{\pm 0.0003}$	$0.0002_{\pm 0.0001}$	$0.0004_{\pm 0.0003}$	$0.0001_{\pm 0.0001}$	$0.0005_{\pm 0.0003}$	$0.0002_{\pm 0.0001}$
		DCRec	$^{\star}0.1780_{\pm 0.0017}$	$^{\dagger}0.0860_{\pm 0.0008}$	$^{\star}0.2177_{\pm 0.0043}$	$0.1021_{\pm 0.0017}$	$^{\dagger}0.1592_{\pm 0.0015}$	$^{\dagger}0.0784_{\pm 0.0005}$
		PAUDRec	$0.1366_{\pm 0.0019}$	$^{\star}0.0870_{\pm 0.0006}$	$0.1723_{\pm 0.0044}$	$^{\star}0.1050_{\pm 0.0018}$	$0.1197_{\pm 0.0024}$	$^{\star}0.0785_{\pm 0.0011}$
		PAAC	$0.0622_{\pm 0.0014}$	$0.0302_{\pm 0.0005}$	$0.0832_{\pm 0.0025}$	$0.0393_{\pm 0.0011}$	$0.0523_{\pm 0.0011}$	$0.0259_{\pm 0.0003}$
		SAPID	$0.0011_{\pm 0.0004}$	$0.0005_{\pm 0.0002}$	$0.0016_{\pm 0.0008}$	$0.0007_{\pm 0.0002}$	$0.0008_{\pm 0.0003}$	$0.0005_{\pm 0.0002}$
		TPAB	$0.0842_{\pm 0.0030}$	$0.0598_{\pm 0.0019}$	$0.1177_{\pm 0.0052}$	$0.0742_{\pm 0.0030}$	$0.0609_{\pm 0.0055}$	$0.0394_{\pm 0.0027}$
		DDC	$0.0328_{\pm 0.0009}$	$0.0158_{\pm 0.0002}$	$0.0447_{\pm 0.0025}$	$0.0214_{\pm 0.0007}$	$0.0271_{\pm 0.0005}$	$0.0132_{\pm 0.0002}$
	Proposed	MF	$0.0121_{\pm 0.0009}$	$0.0061_{\pm 0.0006}$	$0.0168_{\pm 0.0017}$	$0.0089_{\pm 0.0009}$	$0.0112_{\pm 0.0005}$	$0.0054_{\pm 0.0005}$
		+(Ours)	$\mathbf{0.0637}_{\pm \mathbf{0.0003}}$	$\mathbf{0.0318}_{\pm \mathbf{0.0002}}$	$\mathbf{0.0928}_{\pm \mathbf{0.0008}}$	$\mathbf{0.0460}_{\pm \mathbf{0.0004}}$	$\mathbf{0.0500}_{\pm \mathbf{0.0003}}$	$\mathbf{0.0250}_{\pm \mathbf{0.0005}}$
		GRU	$0.0633_{\pm 0.0021}$	$0.0417_{\pm 0.0009}$	$0.0909_{\pm 0.0041}$	$0.0577_{\pm 0.0015}$	$0.0503_{\pm 0.0012}$	$0.0341_{\pm 0.0008}$
		+(Ours)	$\mathbf{0.1004}_{\pm \mathbf{0.0013}}$	$\mathbf{0.0609}_{\pm \mathbf{0.0002}}$	$\mathbf{0.1609}_{\pm \mathbf{0.0028}}$	$\mathbf{0.0924}_{\pm \mathbf{0.0030}}$	$\mathbf{0.0718}_{\pm \mathbf{0.0030}}$	$\mathbf{0.0461}_{\pm \mathbf{0.0015}}$
		SASRec	$0.0121_{\pm 0.0009}$	$0.0061_{\pm 0.0006}$	$0.0168_{\pm 0.0017}$	$0.0089_{\pm 0.0009}$	$0.0112_{\pm 0.0005}$	$0.0054_{\pm 0.0005}$
		+(Ours)	$\mathbf{0.1632}_{\pm \mathbf{0.0039}}$	$\mathbf{0.0803}_{\pm \mathbf{0.0022}}$	$\mathbf{0.2017}_{\pm \mathbf{0.0048}}$	$\mathbf{0.0983}_{\pm \mathbf{0.0043}}$	$\mathbf{0.1449}_{\pm \mathbf{0.0040}}$	$\mathbf{0.0718}_{\pm \mathbf{0.0014}}$
		TiSASRec	$0.1477_{\pm 0.0011}$	$0.0704_{\pm 0.0004}$	$0.1851_{\pm 0.0026}$	$0.0854_{\pm 0.0014}$	$0.1286_{\pm 0.0009}$	$0.0625_{\pm 0.0001}$
		+(Ours)	$\mathbf{0.1695}_{\pm \mathbf{0.0014}}$	$\mathbf{0.0845}_{\pm \mathbf{0.0007}}$	$\mathbf{0.1998}_{\pm \mathbf{0.0035}}$	$\mathbf{0.1005}_{\pm \mathbf{0.0015}}$	$\mathbf{0.1552}_{\pm \mathbf{0.0010}}$	$\mathbf{0.0768}_{\pm \mathbf{0.0006}}$
		FEARec	$0.1433_{\pm 0.0013}$	$0.0695_{\pm 0.0003}$	$0.1758_{\pm 0.0048}$	$0.0833_{\pm 0.0015}$	$0.1279_{\pm 0.0008}$	$0.0629_{\pm 0.0005}$
		+(Ours)	$^{\dagger}\mathbf{0.1742}_{\pm \mathbf{0.0007}}$	$\mathbf{0.0855}_{\pm \mathbf{0.0004}}$	$^{\dagger}\mathbf{0.2124}_{\pm \mathbf{0.0010}}$	$^{\dagger}\mathbf{0.1046}_{\pm \mathbf{0.0013}}$	$^{\dagger}\mathbf{0.1561}_{\pm \mathbf{0.0007}}$	$\mathbf{0.0765}_{\pm \mathbf{0.0003}}$
		BSARec	$0.0902_{\pm 0.0178}$	$0.0504_{\pm 0.0104}$	$0.1193_{\pm 0.0182}$	$0.0636_{\pm 0.0097}$	$0.0764_{\pm 0.0178}$	$0.0441_{\pm 0.0108}$
		+(Ours)	$\mathbf{0.1715}_{\pm \mathbf{0.0009}}$	$\mathbf{0.0848}_{\pm \mathbf{0.0003}}$	$\mathbf{0.2105}_{\pm \mathbf{0.0039}}$	$\mathbf{0.1045}_{\pm \mathbf{0.0014}}$	$\mathbf{0.1530}_{\pm \mathbf{0.0012}}$	$\mathbf{0.0755}_{\pm \mathbf{0.0004}}$
MovieLens	Debiased	DICE	$0.0322_{\pm 0.0003}$	$0.0150_{\pm 0.0001}$	$0.0473_{\pm 0.0005}$	$0.0217_{\pm 0.0002}$	$0.0258_{\pm 0.0002}$	$0.0121_{\pm 0.0001}$
		TIDE	$0.0283_{\pm 0.0004}$	$0.0129_{\pm 0.0002}$	$0.0384_{\pm 0.0006}$	$0.0172_{\pm 0.0002}$	$0.0239_{\pm 0.0005}$	$0.0111_{\pm 0.0002}$
		MELT	$0.0383_{\pm 0.0002}$	$0.0176_{\pm 0.0001}$	$0.0788_{\pm 0.0006}$	$0.0352_{\pm 0.0003}$	$0.0211_{\pm 0.0003}$	$0.0101_{\pm 0.0001}$
		DCRec	$^{\star}0.1092_{\pm 0.0004}$	$^{\star}0.0524_{\pm 0.0002}$	$^{\dagger}0.1317_{\pm 0.0008}$	$0.0646_{\pm 0.0008}$	$^{\star}0.0997_{\pm 0.0008}$	$^{\star}0.0472_{\pm 0.0004}$
		PAUDRec	$0.0925_{\pm 0.0006}$	$0.0450_{\pm 0.0003}$	$0.1136_{\pm 0.0020}$	$0.0555_{\pm 0.0010}$	$0.0835_{\pm 0.0010}$	$0.0406_{\pm 0.0003}$
		PAAC	$0.0349_{\pm 0.0002}$	$0.0166_{\pm 0.0001}$	$0.0571_{\pm 0.0007}$	$0.0270_{\pm 0.0004}$	$0.0255_{\pm 0.0003}$	$0.0121_{\pm 0.0001}$
		SAPID	$0.0458_{\pm 0.0002}$	$0.0213_{\pm 0.0001}$	$0.0876_{\pm 0.0003}$	$0.0407_{\pm 0.0003}$	$0.0281_{\pm 0.0003}$	$0.0131_{\pm 0.0002}$
		TPAB	$0.0055_{\pm 0.0016}$	$0.0200_{\pm 0.0015}$	$0.0056_{\pm 0.0013}$	$0.0071_{\pm 0.0011}$	$0.0060_{\pm 0.0005}$	$0.0190_{\pm 0.0008}$
		DDC	$0.0197_{\pm 0.0005}$	$0.0091_{\pm 0.0003}$	$0.0234_{\pm 0.0005}$	$0.0109_{\pm 0.0005}$	$0.0181_{\pm 0.0006}$	$0.0084_{\pm 0.0002}$
	Proposed	MF	$0.0308_{\pm 0.0006}$	$0.0145_{\pm 0.0004}$	$0.0488_{\pm 0.0012}$	$0.0223_{\pm 0.0005}$	$\mathbf{0.0235}_{\pm \mathbf{0.0003}}$	$\mathbf{0.0112}_{\pm \mathbf{0.0004}}$
		+(Ours)	$\mathbf{0.0375}_{\pm \mathbf{0.0004}}$	$\mathbf{0.0177}_{\pm \mathbf{0.0001}}$	$\mathbf{0.0728}_{\pm \mathbf{0.0003}}$	$\mathbf{0.0341}_{\pm \mathbf{0.0002}}$	$0.0225_{\pm 0.0005}$	$0.0108_{\pm 0.0002}$
		GRU	$0.0463_{\pm 0.0008}$	$0.0225_{\pm 0.0004}$	$0.0556_{\pm 0.0022}$	$0.0270_{\pm 0.0014}$	$0.0423_{\pm 0.0005}$	$0.0206_{\pm 0.0002}$
		+(Ours)	$\mathbf{0.0803}_{\pm \mathbf{0.0009}}$	$\mathbf{0.0389}_{\pm \mathbf{0.0003}}$	$\mathbf{0.1061}_{\pm \mathbf{0.0005}}$	$\mathbf{0.0533}_{\pm \mathbf{0.0004}}$	$\mathbf{0.0694}_{\pm \mathbf{0.0012}}$	$\mathbf{0.0328}_{\pm \mathbf{0.0006}}$
		SASRec	$0.0916_{\pm 0.0006}$	$0.0389_{\pm 0.0004}$	$0.1055_{\pm 0.0016}$	$0.0455_{\pm 0.0009}$	$0.0867_{\pm 0.0017}$	$0.0365_{\pm 0.0009}$
		+(Ours)	$\mathbf{0.1069}_{\pm \mathbf{0.0007}}$	$\mathbf{0.0515}_{\pm \mathbf{0.0003}}$	$^{\star}\mathbf{0.1346}_{\pm \mathbf{0.0008}}$	$^{\star}\mathbf{0.0666}_{\pm \mathbf{0.0005}}$	$\mathbf{0.0951}_{\pm \mathbf{0.0009}}$	$\mathbf{0.0451}_{\pm \mathbf{0.0005}}$
		TiSASRec	$0.0954_{\pm 0.0012}$	$0.0402_{\pm 0.0004}$	$0.1036_{\pm 0.0023}$	$0.0440_{\pm 0.0011}$	$0.0906_{\pm 0.0013}$	$0.0379_{\pm 0.0003}$
		+(Ours)	$^{\dagger}\mathbf{0.1080}_{\pm \mathbf{0.0009}}$	$^{\dagger}\mathbf{0.0521}_{\pm \mathbf{0.0005}}$	$\mathbf{0.1313}_{\pm \mathbf{0.0024}}$	$^{\dagger}\mathbf{0.0649}_{\pm \mathbf{0.0013}}$	$^{\dagger}\mathbf{0.0981}_{\pm \mathbf{0.0010}}$	$^{\dagger}\mathbf{0.0467}_{\pm \mathbf{0.0004}}$
		FEARec	$0.0810_{\pm 0.0036}$	$0.0361_{\pm 0.0013}$	$0.0947_{\pm 0.0059}$	$0.0424_{\pm 0.0024}$	$0.0752_{\pm 0.0027}$	$0.0335_{\pm 0.0009}$
		+(Ours)	$\mathbf{0.0810}_{\pm \mathbf{0.0030}}$	$\mathbf{0.0385}_{\pm \mathbf{0.0016}}$	$\mathbf{0.0918}_{\pm \mathbf{0.0053}}$	$\mathbf{0.0436}_{\pm \mathbf{0.0028}}$	$\mathbf{0.0764}_{\pm \mathbf{0.0021}}$	$\mathbf{0.0363}_{\pm \mathbf{0.0011}}$
		BSARec	$0.0780_{\pm 0.0013}$	$0.0344_{\pm 0.0007}$	$0.0845_{\pm 0.0020}$	$0.0372_{\pm 0.0010}$	$0.0753_{\pm 0.0014}$	$0.0332_{\pm 0.0006}$
		+(Ours)	$\mathbf{0.0882}_{\pm \mathbf{0.0007}}$	$\mathbf{0.0418}_{\pm \mathbf{0.0002}}$	$\mathbf{0.1046}_{\pm \mathbf{0.0014}}$	$\mathbf{0.0505}_{\pm \mathbf{0.0008}}$	$\mathbf{0.0812}_{\pm \mathbf{0.0006}}$	$\mathbf{0.0381}_{\pm \mathbf{0.0001}}$
KuaiRand	Debiased	DICE	$0.0459_{\pm 0.0004}$	$0.0219_{\pm 0.0001}$	$0.0559_{\pm 0.0005}$	$0.0267_{\pm 0.0002}$	$0.0081_{\pm 0.0002}$	$0.0040_{\pm 0.0001}$
		TIDE	$0.0469_{\pm 0.0001}$	$0.0226_{\pm 0.0001}$	$0.0576_{\pm 0.0002}$	$0.0277_{\pm 0.0001}$	$0.0067_{\pm 0.0002}$	$0.0033_{\pm 0.0000}$
		MELT	$0.0483_{\pm 0.0006}$	$0.0246_{\pm 0.0003}$	$0.0612_{\pm 0.0007}$	$0.0312_{\pm 0.0004}$	$0.0000_{\pm 0.0000}$	$0.0000_{\pm 0.0000}$
		DCRec	$^{\star}0.0761_{\pm 0.0002}$	$^{\star}0.0374_{\pm 0.0001}$	$^{\star}0.0948_{\pm 0.0004}$	$^{\star}0.0464_{\pm 0.0002}$	$0.0086_{\pm 0.0007}$	$0.0037_{\pm 0.0003}$
		PAUDRec	$0.0649_{\pm 0.0007}$	$0.0311_{\pm 0.0005}$	$0.0791_{\pm 0.0011}$	$0.0380_{\pm 0.0008}$	$0.0115_{\pm 0.0009}$	$0.0054_{\pm 0.0004}$
		PAAC	$0.0665_{\pm 0.0006}$	$0.0317_{\pm 0.0003}$	$0.0797_{\pm 0.0006}$	$0.0380_{\pm 0.0003}$	$0.0166_{\pm 0.0004}$	$0.0078_{\pm 0.0003}$
		SAPID	$0.0505_{\pm 0.0007}$	$0.0259_{\pm 0.0004}$	$0.0639_{\pm 0.0009}$	$0.0329_{\pm 0.0005}$	$0.0000_{\pm 0.0000}$	$0.0000_{\pm 0.0000}$
		TPAB	$0.0829_{\pm 0.0004}$	$0.0551_{\pm 0.0005}$	$0.1018_{\pm 0.0007}$	$0.0646_{\pm 0.0006}$	$0.0091_{\pm 0.0007}$	$0.0045_{\pm 0.0004}$
		DDC	$0.0225_{\pm 0.0011}$	$0.0104_{\pm 0.0005}$	$0.0264_{\pm 0.0013}$	$0.0122_{\pm 0.0006}$	$0.0076_{\pm 0.0005}$	$0.0035_{\pm 0.0003}$
	Proposed	MF	$0.0427_{\pm 0.0006}$	$0.0218_{\pm 0.0004}$	$0.0514_{\pm 0.0005}$	$0.0265_{\pm 0.0005}$	$0.0101_{\pm 0.0008}$	$0.0043_{\pm 0.0004}$
		+(Ours)	$\mathbf{0.0672}_{\pm \mathbf{0.0002}}$	$\mathbf{0.0324}_{\pm \mathbf{0.0001}}$	$\mathbf{0.0794}_{\pm \mathbf{0.0001}}$	$\mathbf{0.0386}_{\pm \mathbf{0.0002}}$	$\mathbf{0.0209}_{\pm \mathbf{0.0011}}$	$\mathbf{0.0094}_{\pm \mathbf{0.0003}}$
		GRU	$0.0350_{\pm 0.0008}$	$0.0163_{\pm 0.0006}$	$0.0416_{\pm 0.0009}$	$0.0194_{\pm 0.0007}$	$\mathbf{0.0101}_{\pm \mathbf{0.0011}}$	$\mathbf{0.0048}_{\pm \mathbf{0.0007}}$
		+(Ours)	$\mathbf{0.0684}_{\pm \mathbf{0.0009}}$	$\mathbf{0.0336}_{\pm \mathbf{0.0006}}$	$\mathbf{0.0860}_{\pm \mathbf{0.0010}}$	$\mathbf{0.0424}_{\pm \mathbf{0.0007}}$	$0.0017_{\pm 0.0004}$	$0.0006_{\pm 0.0001}$
		SASRec	$0.0573_{\pm 0.0006}$	$0.0242_{\pm 0.0004}$	$0.0654_{\pm 0.0010}$	$0.0275_{\pm 0.0005}$	$^{\dagger}\mathbf{0.0268}_{\pm \mathbf{0.0015}}$	$^{\dagger}\mathbf{0.0116}_{\pm \mathbf{0.0007}}$
		+(Ours)	$\mathbf{0.0735}_{\pm \mathbf{0.0006}}$	$\mathbf{0.0358}_{\pm \mathbf{0.0005}}$	$\mathbf{0.0921}_{\pm \mathbf{0.0008}}$	$\mathbf{0.0449}_{\pm \mathbf{0.0006}}$	$0.0036_{\pm 0.0001}$	$0.0013_{\pm 0.0000}$
		TiSASRec	$0.0586_{\pm 0.0014}$	$0.0245_{\pm 0.0006}$	$0.0671_{\pm 0.0017}$	$0.0281_{\pm 0.0008}$	$^{\star}\mathbf{0.0272}_{\pm \mathbf{0.0007}}$	$^{\star}\mathbf{0.0118}_{\pm \mathbf{0.0004}}$
		+(Ours)	$^{\dagger}\mathbf{0.0749}_{\pm \mathbf{0.0003}}$	$^{\dagger}\mathbf{0.0363}_{\pm \mathbf{0.0002}}$	$^{\dagger}\mathbf{0.0939}_{\pm \mathbf{0.0004}}$	$^{\dagger}\mathbf{0.0456}_{\pm \mathbf{0.0003}}$	$0.0036_{\pm 0.0004}$	$0.0013_{\pm 0.0001}$
		FEARec	$0.0545_{\pm 0.0009}$	$0.0255_{\pm 0.0007}$	$0.0675_{\pm 0.0011}$	$0.0317_{\pm 0.0008}$	$\mathbf{0.0054}_{\pm \mathbf{0.0001}}$	$\mathbf{0.0023}_{\pm \mathbf{0.0001}}$
		+(Ours)	$\mathbf{0.0611}_{\pm \mathbf{0.0012}}$	$\mathbf{0.0305}_{\pm \mathbf{0.0006}}$	$\mathbf{0.0770}_{\pm \mathbf{0.0016}}$	$\mathbf{0.0385}_{\pm \mathbf{0.0007}}$	$0.0008_{\pm 0.0004}$	$0.0003_{\pm 0.0001}$
		BSARec	$0.0380_{\pm 0.0013}$	$0.0202_{\pm 0.0007}$	$0.0482_{\pm 0.0017}$	$0.0255_{\pm 0.0009}$	$0.0000_{\pm 0.0000}$	$0.0000_{\pm 0.0000}$
		+(Ours)	$\mathbf{0.0679}_{\pm \mathbf{0.0004}}$	$\mathbf{0.0333}_{\pm \mathbf{0.0003}}$	$\mathbf{0.0854}_{\pm \mathbf{0.0004}}$	$\mathbf{0.0421}_{\pm \mathbf{0.0004}}$	$\mathbf{0.0019}_{\pm \mathbf{0.0003}}$	$\mathbf{0.0007}_{\pm \mathbf{0.0001}}$

Head- and Tail-Item Performance. On Micro-Video, the proposed framework improves both Head- and Tail-item performance across all backbones, while MovieLens shows a similar pattern with a few exceptions. KuaiRand exhibits a different behavior: although All- and Head-item performance improve, Tail performance decreases for several sequential backbones. Appendix A.7 shows that this degradation is largely removed under utility-only inference, indicating that it mainly arises from the learned popularity component at inference rather than from poor Tail-item discrimination in the utility model.

Representation Uniformity. Table 2 shows substantially more dispersed user-context representations under the proposed framework across all backbones and datasets. The consistent shift across both conventional and sequential architectures indicates that the effect is not specific to a particular representation model. This is consistent with Section 3.4 and Figure 1, where separating temporal popularity from utility and enforcing centering reduces popularity-induced representation concentration.

Table 2: Uniformity(↓) of user-context representations across six recommendation backbones.

Backbone	Micro Video		MovieLens-1M		KuaiRand	
	Vanilla	Ours	Vanilla	Ours	Vanilla	Ours
MF	$-0.1349_{\pm 0.0010}$	$\mathbf{-3.8220_{\pm 0.0035}}$	$-0.0500_{\pm 0.0003}$	$\mathbf{-3.7804_{\pm 0.0027}}$	$-0.1020_{\pm 0.0006}$	$\mathbf{-3.7123_{\pm 0.0052}}$
GRU	$-0.1012_{\pm 0.0029}$	$\mathbf{-3.8029_{\pm 0.0031}}$	$-0.1036_{\pm 0.0023}$	$\mathbf{-3.5521_{\pm 0.0185}}$	$-0.1220_{\pm 0.0055}$	$\mathbf{-3.1821_{\pm 0.0073}}$
SASRec	$-0.0943_{\pm 0.0061}$	$\mathbf{-3.5652_{\pm 0.0171}}$	$-0.0759_{\pm 0.0020}$	$\mathbf{-3.0466_{\pm 0.0066}}$	$-0.0771_{\pm 0.0013}$	$\mathbf{-3.0597_{\pm 0.0185}}$
TiSASRec	$-0.0913_{\pm 0.0033}$	$\mathbf{-3.6116_{\pm 0.0111}}$	$-0.0533_{\pm 0.0025}$	$\mathbf{-2.9562_{\pm 0.0246}}$	$-0.0721_{\pm 0.0022}$	$\mathbf{-2.8812_{\pm 0.0194}}$
FEARec	$-0.0878_{\pm 0.0067}$	$\mathbf{-3.6014_{\pm 0.0247}}$	$-0.0436_{\pm 0.0033}$	$\mathbf{-3.1607_{\pm 0.0504}}$	$-0.0432_{\pm 0.0053}$	$\mathbf{-2.7468_{\pm 0.0397}}$
BSARec	$-0.0251_{\pm 0.0122}$	$\mathbf{-3.6024_{\pm 0.0136}}$	$-0.0086_{\pm 0.0050}$	$\mathbf{-3.2671_{\pm 0.0137}}$	$-0.0003_{\pm 0.0004}$	$\mathbf{-3.0626_{\pm 0.0120}}$

Role of the Centering Constraint. Figure 2 shows that the largest improvement occurs when moving from the vanilla backbone to the proposed model without centering ($\gamma = 0$), indicating that the popularity-utility decomposition drives most of the gain in both uniformity and Recall@10. Moderate centering generally preserves these improvements, supporting its role as an identification constraint that limits residual item-time effects in the utility. Excessively large γ, however, can reduce Recall@10 for some backbones by over-constraining user-specific variation.

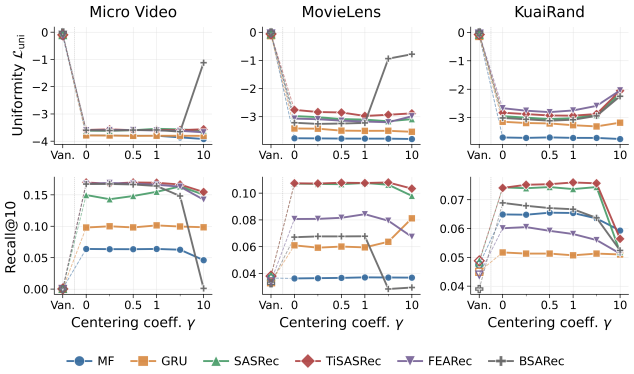


Figure 2: Effect of the centering on representation uniformity and Recall@10. Van. denotes the vanilla backbone.

Effect of Utility-Popularity Calibration Figure 3 shows that utility-only ranking ($\eta = 0$) retains substantial performance, whereas popularity-only ranking ($\eta = 1$) consistently degrades Recall@10. Performance generally peaks at intermediate η, indicating that utility and popularity provide complementary signals for predicting observed interactions, consistent with the conditional choice model in equation 1. The two components remain separately estimated and are combined only at inference.

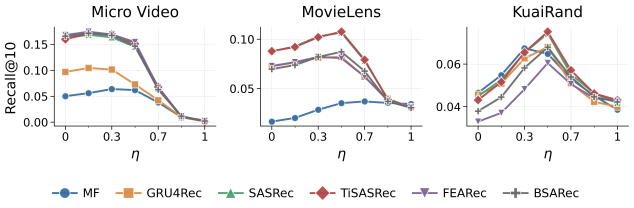


Figure 3: Effect of the interpolation weight η on Recall@10 across six recommendation backbones.

6 Conclusion

This work examined temporal popularity bias from the perspective of identifiable preference estimation. By representing timestamped interactions through conditional event intensities, we characterized the confounding between population-level item effects and user-specific utility, and showed that separating these components with a centering condition enables recovery up to ranking-equivalent shifts. We further showed that unresolved shared item-time effects can alter the user-context representations geometry, explaining popularity-induced homogenization. Guided by these results, we developed a scalable framework combining Hawkes-based popularity modeling, importance sampling, and centering. Experiments on controlled and real-world data support the theoretical analysis and show broad gains across diverse recommendation architectures.

One limitation is that the utility is identified only up to a ranking-equivalent, context-dependent shift. While this is sufficient for ranking items within a given user context, it does not directly support comparisons across user contexts, such as ranking users for a fixed item. In addition, overly concentrated popularity estimates can impair Tail ranking at inference, motivating better calibrated and more flexible temporal popularity models.

REFERENCES

Himan Abdollahpouri, R. Burke, and Bamshad Mobasher. Managing popularity bias in recommender systems with personalized re-ranking. In *The Florida AI Research Society*, 2019. URL https://api.semanticscholar.org/CorpusID:59158829.

Per K Andersen, Ornulf Borgan, Richard D Gill, and Niels Keiding. *Statistical models based on counting processes*. Springer Science & Business Media, 2012.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Yoshua Bengio and Jean-Sébastien Senécal. Adaptive importance sampling to accelerate training of a neural probabilistic language model. *IEEE Transactions on Neural Networks*, 19(4):713, 2008.

Miaomiao Cai, Lei Chen, Yifan Wang, Haoyue Bai, Peijie Sun, Le Wu, Min Zhang, and Meng Wang. Popularity-aware alignment and contrast for mitigating popularity bias. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024. URL https://api.semanticscholar.org/CorpusID:270199392.

Huiyuan Chen, Vivian Lai, Hongye Jin, Zhimeng Jiang, Mahashweta Das, and Xia Hu. Towards mitigating dimensional collapse of representations in collaborative filtering. In *Proceedings of the 17th ACM international conference on web search and data mining*, pp. 106–115, 2024.

Xinyu Du, Huanhuan Yuan, Pengpeng Zhao, Jianfeng Qu, Fuzhen Zhuang, Guanfeng Liu, Yanchi Liu, and Victor S. Sheng. Frequency enhanced hybrid attention network for sequential recommendation. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023. URL https://api.semanticscholar.org/CorpusID:258212874.

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. *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022. URL https://api.semanticscholar.org/CorpusID:251643401.

Xingzhuo Guo, Junwei Pan, Ximei Wang, Baixu Chen, Jie Jiang, and Mingsheng Long. On the embedding collapse when scaling up recommendation models. *arXiv preprint arXiv:2310.04400*, 2023.

F. Maxwell Harper, Joseph A. Konstan, and Joseph A. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5:19:1–19:19, 2016. URL https://api.semanticscholar.org/CorpusID:16619709.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 19:308–334, 1948. URL https://api.semanticscholar.org/CorpusID:121456695.

Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. *2018 IEEE International Conference on Data Mining (ICDM)*, pp. 197–206, 2018. URL https://api.semanticscholar.org/CorpusID:52127932.

Jongjin Kim and U Kang. Sequentially diversified and accurate recommendations in chronological order for a series of users. *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, 2025. URL https://api.semanticscholar.org/CorpusID:276653747.

Kibum Kim, Dongmin Hyun, Sukwon Yun, and Chanyoung Park. Melt: Mutual enhancement of long-tailed user and item for sequential recommendation. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023. URL https://api.semanticscholar.org/CorpusID:258179557.

Anastasiia Klimashevskaia, D. Jannach, Mehdi Elahi, and Christoph Trattner. A survey on popularity bias in recommender systems. *User Modeling and User-Adapted Interaction*, 34:1777 – 1834, 2023. URL https://api.semanticscholar.org/CorpusID:260378994.

Yehuda Koren, Robert M. Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42, 2009. URL https://api.semanticscholar.org/CorpusID:58370896.

Luca Lalor and Anatoliy Swishchuk. Event-based limit order book simulation under a neural hawkes process: application in market-making. *Applied Mathematical Finance*, 32(2):128–155, 2025.

Patrick J Laub, Young Lee, and Thomas Taimre. *The elements of Hawkes processes*, volume 3. Springer, 2021.

Jae-woong Lee, Seongmin Park, Mincheol Yoon, and Jongwuk Lee. uctrl: Unbiased contrastive representation learning via alignment and uniformity for collaborative filtering. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pp. 2456–2460, 2023.

Benjamin J. Lengerich, S. Tan, Chun-Hao Kingsley Chang, Giles Hooker, and Rich Caruana. Purifying interaction effects with the functional anova: An efficient algorithm for recovering identifiable additive models. In *International Conference on Artificial Intelligence and Statistics*, 2019. URL https://api.semanticscholar.org/CorpusID:207863661.

Jiacheng Li, Yujie Wang, and Julian McAuley. Time interval aware self-attention for sequential recommendation. *Proceedings of the 13th International Conference on Web Search and Data Mining*, 2020. URL https://api.semanticscholar.org/CorpusID:208780552.

Siyi Lin, Chongming Gao, Jiawei Chen, Sheng Zhou, Binbin Hu, Yan Feng, Chun Chen, and Can Wang. How do recommendation models amplify popularity bias? an analysis from the spectral perspective. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pp. 659–668, 2025.

Lingfeng Liu, Yixin Song, Dazhong Shen, Bing Yin, Hao Li, Yanyong Zhang, and Chao Wang. Rethinking popularity bias in collaborative filtering via analytical vector decomposition. *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, 2025. URL https://api.semanticscholar.org/CorpusID:283737766.

R Duncan Luce et al. *Individual choice behavior*, volume 4. Wiley New York, 1959.

Ruijia Ma, Yahong Lian, and Chunyao Song. Sub-interest-aware representation uniformity for recommender system. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 12346–12354, 2025.

Zhong Dong Ouyang, Chun-Hui Zhang, Shifu Hou, Chuxu Zhang, and Yanfang Ye. How to improve representation alignment and uniformity in graph-based collaborative filtering? In *International Conference on Web and Social Media*, 2024. URL https://api.semanticscholar.org/CorpusID:270193191.

Kijeong Park, Sungchul Hong, and Jong-June Jeon. Dynamic higher-order relations and event-driven temporal modeling for stock price forecasting. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 6048–6056, 2025.

Marian-Andrei Rizoiu, Lexing Xie, Scott Sanner, Manuel Cebrian, Honglin Yu, and Pascal Van Hentenryck. Expecting to be hip: Hawkes intensity processes for social media popularity. In *Proceedings of the 26th international conference on world wide web*, pp. 735–744, 2017.

Yehjin Shin, Jeongwhan Choi, Hyowon Wi, and Noseong Park. An attentive inductive bias for sequential recommendation beyond the self-attention. In *AAAI Conference on Artificial Intelligence*, 2023. URL https://api.semanticscholar.org/CorpusID:266348680.

Chenyang Wang, Yuanqing Yu, Weizhi Ma, M. Zhang, C. Chen, Yiqun Liu, and Shaoping Ma. Towards representation alignment and uniformity in collaborative filtering. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022. URL https://api.semanticscholar.org/CorpusID:250072369.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

Jiajin Wu, Bo Yang, Runze Mao, and Qing Li. Popularity-aware sequential recommendation with user desire. *Expert Syst. Appl.*, 237:121429, 2023. URL https://api.semanticscholar.org/CorpusID:261603802.

Weiqin Yang, Yue Pan, Chongming Gao, Sheng Zhou, Xiang Wang, Can Wang, and Jiawei Chen. The pitfall of scaling up: Uncovering and mitigating popularity bias amplification in scaling transformer-based recommenders. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 6044–6055, 2026.

Yuhao Yang, Chao Huang, Lianghao Xia, Chunzhen Huang, Dave Luo, and Kangyi Lin. Debiased contrastive learning for sequential recommendation. *Proceedings of the ACM Web Conference 2023*, 2023. URL https://api.semanticscholar.org/CorpusID:257636510.

Hyunsik Yoo, Ruizhong Qiu, Charlie Xu, Fei Wang, and Hanghang Tong. Generalizable recommender system during temporal popularity distribution shifts. *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, 2025. URL https://api.semanticscholar.org/CorpusID:277592299.

Junliang Yu, Hongzhi Yin, Xin Xia, Tong Chen, Li zhen Cui, and Quoc Viet Hung Nguyen. Are graph augmentations necessary?: Simple graph contrastive learning for recommendation. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021. URL https://api.semanticscholar.org/CorpusID:247940233.

Amir Hassan Zadeh and Ramesh Sharda. Modeling brand post popularity dynamics in online social networks. *Decision Support Systems*, 65:59–68, 2014.

Dan Zhang, Yangli ao Geng, Wenwen Gong, Zhongang Qi, Zhiyu Chen, Xing Tang, Shan Ying, Yuxiao Dong, and Jie Tang. Recdcl: Dual contrastive learning for recommendation. *Proceedings of the ACM Web Conference 2024*, 2024. URL https://api.semanticscholar.org/CorpusID:267311544.

Yifei Zhang, Hao Zhu, Zixing Song, Piotr Koniusz, Irwin King, et al. Mitigating the popularity bias of graph collaborative filtering: A dimensional collapse perspective. *Advances in Neural Information Processing Systems*, 36:67533–67550, 2023.

Jujia Zhao, Wenjie Wang, Xinyu Lin, Leigang Qu, Jizhi Zhang, and Tat-Seng Chua. Popularity-aware distributionally robust optimization for recommendation system. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 4967–4973, 2023.

Zihao Zhao, Jiawei Chen, Sheng Zhou, Xiangnan He, Xuezhi Cao, Fuzheng Zhang, and Wei Wu. Popularity bias is not always evil: Disentangling benign and harmful bias for recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 35(10):9920–9931, 2022.

Y. Zheng, Chen Gao, Xiang Li, Xiangnan He, Depeng Jin, and Yong Li. Disentangling user interest and conformity for recommendation with causal embedding. *Proceedings of the Web Conference 2021*, 2021a. URL https://api.semanticscholar.org/CorpusID:231984619.

Yu Zheng, Chen Gao, Xiang Li, Xiangnan He, Yong Li, and Depeng Jin. Disentangling user interest and conformity for recommendation with causal embedding. In *Proceedings of the web conference 2021*, pp. 2980–2991, 2021b.

A Appendix

A.1 Conditional Probability from Event Intensities

We denote by $N_{uv}(t) = \sum_{i=1}^{n} \mathbb{I}(u_i = u, v_i = v, t_i \leq 0)$ the counting process for interactions between user u and item v observed by time t. Summing over items gives the user-level counting process $N_u(t) = \sum_{k \in \mathcal{V}} N_{uk}(t)$. The following proposition establishes the link between item-specific event intensities and the conditional distribution of the interacted item.

Proposition 2 *Assume that the interaction process of each user has no multiple jumps to first order, i.e.,*

$$\Pr\big(N_u(t + \Delta t) - N_u(t) \geq 2 | \mathcal{F}_{t-}^u\big) = o(\Delta t).$$

Then, conditional on user u generating an interaction at time t, the associated item mark follows

$$p_0\big(v | x_u(t)\big) = \frac{\lambda_{uv}^0(t|\mathcal{F}_{t-}^u)}{\lambda_u^0(t|\mathcal{F}_{t-}^u)}.$$

Proof. For notational convenience, define the increments over $[t, t + \Delta t]$ by

$$dN_{uv}(t) = N_{uv}(t + \Delta t) - N_{uv}(t), \qquad dN_u(t) = N_u(t + \Delta t) - N_u(t).$$

From the definition of conditional intensity,

$$\Pr\big(dN_{uv}(t) = 1 | \mathcal{F}_{t-}^u\big) = \lambda_{uv}^0(t|\mathcal{F}_{t-}^u)\Delta t + o(\Delta t).$$

Since $dN_u(t) = \sum_{k \in \mathcal{V}} dN_{uk}$, the event $dN_u(t) = 1$ corresponds, up to events involving multiple jumps, to exactly one of the item-specific processes jumping during the interval. By assumption, the probability of such multiple-jump events is $o(\Delta t)$. Hence,

$$\begin{aligned}\Pr\left(dN_u(t) = 1 | \mathcal{F}_{t-}^u\right) &= \sum_{k \in \mathcal{V}} \Pr\left(dN_{uk}(t) = 1 | \mathcal{F}_{t-}^u\right) + o(\Delta t) \\ &= \left[\sum_{k \in \mathcal{V}} \lambda_{uk}^0(t|\mathcal{F}_{t-}^u)\right] \Delta t + o(\Delta t).\end{aligned}$$

Because $\mathcal{V}$ is finite, summing the item-specific remainder terms preserves the $o(\Delta t)$ order.

Likewise, conditioning on a single user-level interaction changes the numerator only through multiple-jump events, whose probability is $o(\Delta t)$. Therefore,

$$\begin{aligned}p_0\big(v | x_u(t)\big) &= \lim_{\Delta t \downarrow 0} \frac{\Pr\left(dN_{uv}(t) = 1 | \mathcal{F}_{t-}^u\right)}{\Pr\left(dN_u(t) = 1 | \mathcal{F}_{t-}^u\right)} \\ &= \frac{\lambda_{uv}^0(t|\mathcal{F}_{t-}^u)}{\sum_{k \in \mathcal{V}} \lambda_{uk}^0(t\mathcal{F}_{t-}^u)}.\end{aligned}$$

□

A.2 Uniformity Implication of Representation Homogenization

To quantify the loss of representation diversity implied by Theorem 1, let $P_{t,\rho}$ denote the distribution of normalized user-context representations $\widetilde{\mathbf{h}}_{u,t}$ at time t. Following Wang & Isola (2020), define the uniformity loss as

$$\mathcal{L}_{\text{uni}}(P_{t,\rho}) = \log \mathbb{E}_{\mathbf{h},\mathbf{h}' \sim P_{t,\rho}} \left[\exp\{-\tau \parallel \mathbf{h} - \mathbf{h}' \parallel_2^2\}\right], \qquad \tau > 0.$$

Smaller values correspond to representations that are more uniformly distributed over the unit hypersphere.

Corollary 1 (Uniformity under Context Homogenization) *Under the conditions of Theorem1, for $\rho > B_t$,*

$$\mathcal{L}_{\text{uni}}(P_{t,\rho}) \geq -\frac{4\tau B_t^2}{\rho^2}.$$

Proof. By Theorem 1, any two representations $\mathbf{h}, \mathbf{h}'$ in the support of $P_{t,\rho}$ satisfy

$$\| \mathbf{h} - \mathbf{h}' \|_2^2 \leq \frac{4B_t^2}{\rho^2}.$$

Therefore,

$$\exp\{-\tau \| \mathbf{h} - \mathbf{h}' \|_2^2\} \geq \exp\{-\frac{4\tau B_t^2}{\rho^2}\}.$$

Taking expectation with respect to $\mathbf{h}, \mathbf{h}' \sim P_t$ gives

$$\mathbb{E}_{\mathbf{h},\mathbf{h}' \sim P_{t,\rho}}\big[\exp\{-\tau \| \mathbf{h} - \mathbf{h}' \|_2^2\}\big] \geq \exp\{-\frac{4\tau B_t^2}{\rho^2}\}.$$

Applying the logarithm to both sides yields

$$\mathcal{L}_{\text{uni}}(P_t) \geq -\frac{4\tau B_t^2}{\rho^2}.$$

□

Because the lower bound converges to zero as $\rho \to \infty$, the corollary implies

$$\mathcal{L}_{\text{uni}}(P_{t,\rho}) \to 0.$$

Thus, as the shared directional component becomes dominant, normalized user-context representations become increasingly concentrated, with complete directional collapse corresponding to a uniformity loss of zero.

A.3 Training Procedure

Algorithm 1 Stochastic Training of the Proposed Framework

Require: Observed interactions $\mathcal{D}_n$; number of auxiliary items M; time-bucket width Δ; centering coefficient γ; number of epochs E
Ensure: Popularity parameters Φ and utility parameters Ψ

```
Initialize Φ and Ψ
Construct time buckets and the corresponding active-user sets
for e = 1, ..., E do
    Shuffle the observed interactions
    for each mini-batch B = {(u_i, v_i, t_i)}_{i=1}^B do
        for each (u_i, v_i, t_i) ∈ B do
            Sample v̄_{i1}, ..., v̄_{iM} ~ Unif(V \ {v_i})
            Form C_i ← {v_i, v̄_{i1}, ..., v̄_{iM}}
            Retrieve the user context x_{u_i}(t_i) and candidate-item event histories before t_i
        end for
        Compute the importance-corrected choice loss ℓ̂_choice(Φ, Ψ) using {C_i}_{i=1}^B
        for each (u_i, v_i, t_i) ∈ B do
            Sample independently u_i^(1), u_i^(2) ~ U_Δ(t_i)
            Construct x_{u_i^(1)}(t_i) and x_{u_i^(2)}(t_i) using interactions strictly before t_i
        end for
        Compute ℓ̂_cen(Ψ) using the same candidate sets {C_i}_{i=1}^B
        L̂ ← ℓ̂_choice + γ ℓ̂_cen
        Update (Φ, Ψ) by one Adam step on ∇_{Φ,Ψ} L̂
    end for
end for
return Φ, Ψ
```

Algorithm 1 summarizes the stochastic optimization procedure used for the proposed framework. For each observed interaction (u_i, v_i, t_i), M auxiliary items are sampled uniformly from $\mathcal{V} \setminus \{v_i\}$ and combined with the observed item to form the candidate set $\mathcal{C}_i = \{v_i, \bar{v}_{i1}, \ldots, \bar{v}_{iM}\}$. The same

candidate set is used for both the importance-corrected choice loss and the centering penalty, avoiding an additional item-sampling step.

For the centering term, two reference users are sampled independently for each observed event from users active in the time bucket containing t_i. Let $u_i^{(1)}$ and $u_i^{(2)}$ denote these two draws. Their histories are reconstructed using only interactions preceding t_i, yielding the corresponding contexts $x_{u_i^{(1)}}(t_i)$ and $x_{u_i^{(2)}}(t_i)$. The two independent draws implement the product estimator introduced in Section 4.3, while preserving the temporal information available at the anchor event.

A.4 Proofs of Theoretical Results

Proposition 1 (Identifiability under Centering) *Suppose Assumption 1 holds, and* $\pi_0(v|t) > 0$ *and* $p_0(v|x_u(t)) > 0$ *for all* $v \in \mathcal{V}$. *For any* $(\widetilde{\pi}, \widetilde{f})$ *satisfying the same conditions,*

$$p_{\widetilde{\pi},\widetilde{f}}\big(v|x_u(t)\big) = p_{\pi_0,f^*}\big(v|x_u(t)\big), \quad \forall u \in \mathcal{U},\ v \in \mathcal{V},\ t \in [0,T],$$

implies

$$\widetilde{\pi}(\cdot|t) = \pi_0(\cdot|t), \qquad \widetilde{f} \in [f^*].$$

Proof. Fix a time t and suppose that $(\widetilde{\pi}, \widetilde{f})$ induces the same conditional item-choice distribution as (π_0, f^*). Using the additive representation of equation 1,

$$\log p_0\big(v|x_u(t)\big) = \log \pi_0(v|t) + f^*\big(x_u(t), v\big) + a_u(t).$$

Similary, the alternative parameterization satisfies

$$\log p_0\big(v|x_u(t)\big) = \log \widetilde{\pi}(v|t) + \widetilde{f}\big(x_u(t), v\big) + \widetilde{a}_u(t),$$

where $\widetilde{a}_u(t) = -\log \sum_{k\in\mathcal{V}} \widetilde{\pi}(k|t) \exp\big\{\widetilde{f}\big(x_u(t), k\big)\big\}$. Equating the two expressions gives

$$\widetilde{f}\big(x_u(t), v\big) - f^*\big(x_u(t), v\big) = d_v(t) + c_u(t),$$

where

$$d_v(t) = \log \pi_0(v|t) - \log \widetilde{\pi}(v|t), \qquad c_u(t) = a_u(t) - \widetilde{a}_u(t).$$

Both f^* and $\widetilde{f}$ satisfy the centering condition in Assumption 1. Summing the above identity over u therefore yields $0 = |\mathcal{U}|\, d_v(t) + \sum_{u\in\mathcal{U}} c_u(t)$. Hence, $d_v(t) = -\frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}} c_u(t)$, which does not depend on v. Denote this common value by $d(t)$. Then, $\pi_0(v|t) = e^{d(t)}\widetilde{\pi}(v|t)$. Since both $\pi_0(\cdot|t)$ and $\widetilde{\pi}(\cdot|t)$ are probability distributions,

$$1 = \sum_{v\in\mathcal{V}} \pi_0(v|t) = e^{d(t)} \sum_{v\in\mathcal{V}} \widetilde{\pi}(v|t) = e^{d(t)}.$$

Thus $d(t) = 0$, and

$$\widetilde{\pi}(v|t) = \pi_0(v|t) \qquad \forall v \in \mathcal{V}.$$

Substituting $d_v(t) = 0$ back into the previous identity gives

$$\widetilde{f}\big(x_u(t), v\big) = f^*\big(x_u(t), v\big) + c_u(t).$$

Since $c_u(t)$ is independent of v, $\widetilde{f} \in [f^*]$. Therefore, $[f^*]$ is identifiable according to Definition 1.

□

Lemma 1 (Popularity-Induced Shared Component) *Suppose* $f^*\big(x_u(t), v\big) = \mathbf{h}_{u,t}^\top \mathbf{z}_v$, *and, for a fixed* t,

$$(\beta_t, \mathbf{c}_t) \in \operatorname*{arg\,min}_{\beta\in\mathbb{R}, \mathbf{c}\in\mathbb{R}^d} \sum_{v\in\mathcal{V}} \big[\log \pi_0(v|t) - \beta - \mathbf{c}^\top \mathbf{z}_v\big]^2.$$

Then the utility learned without explicitly modeling the popularity term has the projected form

$$\widehat{f}\big(x_u(t), v\big) = \big(\mathbf{h}_{u,t} + \mathbf{c}_t\big)^\top \mathbf{z}_v.$$

Proof. By the definition of the least-squares projection, the temporal log-popularity admits the decomposition

$$\log \pi_0(v|t) = \beta_t + \mathbf{c}_t^\top \mathbf{z}_v + r_{v,t},$$

where $r_{v,t} = \log \pi_0(v|t) - \beta_t - \mathbf{c}_t^\top \mathbf{z}_v$ is the projection residual. The least-squares normal equations imply

$$\sum_{v \in \mathcal{V}} r_{v,t} = 0, \qquad \sum_{v \in \mathcal{V}} r_{v,t} \mathbf{z}_v = \mathbf{0},$$

so $\mathbf{r}_t = \big(r_{v,t}\big)_{v \in \mathcal{V}}$ is orthogonal to the affine score space induced by the item representations. From the conditional choice model in equation 1, the item-dependent log-score underlying the observed interaction distribution is

$$f^*\big(x_u(t), v\big) + \log \pi_0(v|t).$$

Substituting the projection decomposition gives

$$f^*\big(x_u(t), v\big) + \log \pi_0(v|t) = \mathbf{h}_{u,t}^\top \mathbf{z}_v + \beta_t + \mathbf{c}_t^\top \mathbf{z}_v + r_{v,t} = \beta_t + \big(\mathbf{h}_{u,t} + \mathbf{c}^t\big)^\top \mathbf{z}_v + r_{v,t}.$$

The first term β_t is constant across items and therefore has no effect on the conditional item distribution. The residual $\mathbf{r}_t$ is orthogonal to the score space generated by $\{\mathbf{z}_v\}_{v \in \mathcal{V}}$. Hence, the component of the observed log-score representable by the popularity-agnostic bilinear utility model is

$$\widehat{f}\big(x_u(t), v\big) = \mathbf{h}_{u,t}^\top \mathbf{z}_v + \mathbf{c}_t^\top \mathbf{z}_v = \big(\mathbf{h}_{u,t} + \mathbf{c}_t\big)^\top \mathbf{z}_v.$$

Since $\log \pi_0(v|t)$ is shared across users, the projected coefficient $\mathbf{c}_t$ is also independent of u, yielding the same representation shift for every user context at time t.

□

Theorem 1 (Popularity-Induced Homogenization of User-Context Representations) *Define* $B_t = \max_{u \in \mathcal{U}} \parallel \mathbf{h}_{u,t} \parallel_2$. *Suppose that an item-time component absorbed into the utility induces the shifted representation*

$$\widetilde{\mathbf{h}}_{u,t} = \frac{\mathbf{h}_{u,t} + \rho \widetilde{\mathbf{c}}_t}{\parallel \mathbf{h}_{u,t} + \rho \widetilde{\mathbf{c}}_t \parallel_2}, \qquad \rho > B_t.$$

Then, for any two users $u_1, u_2 \in \mathcal{U}$,

$$\Big|\Big|\widetilde{\mathbf{h}}_{u_1,t} - \widetilde{\mathbf{h}}_{u_2,t}\Big|\Big|_2^2 \leq \frac{4B_t^2}{\rho^2}.$$

Proof. Fix a user u and write $s_{u,t} = \mathbf{h}_{u,t}^\top \widetilde{\mathbf{c}}_t$. Since $\parallel \widetilde{\mathbf{c}}_t \parallel_2 = 1$ and $\parallel \mathbf{h}_{u,t} \parallel_2 \leq B_t$, $s_{u,t} \in [-B_t, B_t]$. Decompose $\mathbf{h}_{u,t}$ into components parallel and orthogonal to $\widetilde{\mathbf{c}}_t$:

$$\mathbf{h}_{u,t} = s_{u,t} \widetilde{\mathbf{c}}_t + \boldsymbol{\zeta}_u, \qquad \widetilde{\mathbf{c}}_t^\top \boldsymbol{\zeta}_u = 0.$$

Then

$$\parallel \boldsymbol{\zeta}_u \parallel_2^2 = \parallel \mathbf{h}_{u,t} \parallel_2^2 - s_{u,t}^2 \leq B_t^2 - s_{u,t}^2.$$

Hence,

$$\widetilde{\mathbf{h}}_{u,t}^\top \widetilde{\mathbf{c}}_t = \frac{\rho + s_{u,t}}{\sqrt{(\rho + s_{u,t})^2 + \parallel \boldsymbol{\zeta}_u \parallel_2^2}} \geq \frac{\rho + s_{u,t}}{\sqrt{\rho^2 + 2\rho s_{u,t} + B_t^2}}.$$

Since $\rho > B_t$, we have $\rho + s_{u,t} > 0$. Consider the square of the right-hand side,

$$g_\rho(s_{u,t}) = \frac{(\rho + s_{u,t})^2}{\rho^2 + 2\rho s_{u,t} + B_t^2}, \qquad s_{u,t} \in [-B_t, B_t].$$

Differentiating gives

$$g_\rho'(s_{u,t}) = \frac{2(\rho + s_{u,t})(\rho s_{u,t} + B_t^2)}{(\rho^2 + 2\rho s_{u,t} + B_t^2)^2}.$$

Therefore, $g_\rho(s_{u,t})$ is minimized at $s_{u,t} = -B_t^2/\rho$. Substituting this value yields

$$\widetilde{\mathbf{h}}_{u,t}^\top \widetilde{\mathbf{c}}_t \geq \sqrt{1 - \frac{B_t^2}{\rho^2}}.$$

Now let $\theta_i = \arccos\big(\widetilde{\mathbf{h}}_{u,t}^\top \widetilde{\mathbf{c}}_t\big)$, $i = 1, 2$. The preceding bound implies

$$\theta_i \le \arccos\Big(\sqrt{1 - \frac{B_t^2}{\rho^2}}\Big).$$

By the triangle inequality on the unit sphere,

$$\arccos\big(\widetilde{\mathbf{h}}_{u_1,t}^\top \widetilde{\mathbf{h}}_{u_2,t}\big) \le \theta_1 + \theta_2 \le 2\arccos\Big(\sqrt{1 - \frac{B_t^2}{\rho^2}}\Big).$$

Since the cosine function is decreasing on $[0, \pi]$,

$$\widetilde{\mathbf{h}}_{u_1,t}^\top \widetilde{\mathbf{h}}_{u_2,t} \ge 1 - \frac{2B_t^2}{\rho^2}.$$

Finally, both shifted representations have unit norm, so

$$\| \widetilde{\mathbf{h}}_{u_1,t} - \widetilde{\mathbf{h}}_{u_2,t} \|_2^2 = 2 - 2\widetilde{\mathbf{h}}_{u_1,t}^\top \widetilde{\mathbf{h}}_{u_2,t} \le \frac{4B_t^2}{\rho^2}.$$

□

A.5 Artificial Data Experimental Details

Data Generation. We generate data with 800 users and 200 items over a 30-day observation period. Ground-truth user and item latent vectors $\mathbf{h}_u, \mathbf{z}_v \in \mathbb{R}^{16}$ are independently sampled from isotropic Gaussian distributions, and the ground-truth utility is defined as $f^*(u, v) = 3\mathbf{h}_u^\top \mathbf{z}_v$.

The time-varying popularity of each item combines a long-tailed base rate and a transient temporal trend. The base popularity follows a Zipf's distribution with exponent 0.8, and each item is assigned a Gaussian trend centered at a randomly sampled time, with a width of four days.

For each user, 25 interaction times are independently sampled over the observation period and sorted chronologically. At each event time t, an item is sampled according to

$$p_\rho(v|x_u(t)) = \frac{\pi_0(v|t)^\rho \exp\{f^*(u, v)\}}{\sum_{k \notin \mathcal{S}_u(t)} \pi_0(k|t)^\rho \exp\{f^*(u, k)\}},$$

where previously interacted items are excluded. We consider $\rho \in \{0, 0.5, 1, 2, 4\}$, so that only the strength of popularity confounding changes across settings while f^* remains fixed. The resulting events are globally ordered by timestamp and chronologically split into 80% training, 10% validation, and 10% test interactions.

Model Training. For each (ρ, seed), Vanilla and the proposed model are trained on exactly the same generated interactions. Vanilla uses matrix factorization with a standard logistic contrastive objective and does not explicitly model popularity. The proposed model uses the same utility backbone together with the Hawkes popularity component and the centering penalty described in Section 4.

Both models use 32-dimensional learned embeddings and are optimized with Adam for 30 epochs with a learning rate of 0.001 and a batch size of 512. Each positive interaction is paired with seven uniformly sampled auxiliary items. For the proposed model, the centering coefficient is fixed to 1.0, and reference users for the centering term are sampled within one-day time buckets.

Evaluation. Representation concentration is measured using the uniformity loss computed from 20,000 randomly sampled pairs of ℓ_2-normalized user embeddings. Values closer to zero, indicate stronger representation concentration.

To evaluate whether the learned utility preserves the underlying preference structure, we compute Recall@10 between the model ranking and the ground-truth top-10 ranking induced solely by f^*, after excluding training interactions. For the proposed model, ranking is based only on the estimated utility f_Ψ, without the popularity component. Thus, the evaluation directly measures recovery of the popularity-adjusted preference ranking rather than prediction of the popularity-confounded interaction distribution.

A.6 Real Data Experimental Details

Datasets. Experiments are conducted on three timestamped user–item interaction datasets: Micro-Video, MovieLens-1M, and KuaiRand. MovieLens-1M represents movie recommendation, whereas Micro-Video and KuaiRand are short-video recommendation datasets. For Micro-Video and KuaiRand, interaction construction and temporal preprocessing follow the protocol of Yoo et al. (2025). In Micro-Video, a viewing event is treated as a positive interaction when the watch duration exceeds eight seconds, whereas clicks are used as positive feedback in KuaiRand. Both datasets cover a one-month observation period. For MovieLens-1M, ratings of four or five are retained and converted into positive implicit-feedback events.

All interactions are ordered by timestamp before constructing the data splits. The earliest 80% of interactions are used for training, followed by 10% for validation and the remaining 10% for testing. This chronological split preserves the temporal evolution of item popularity and prevents future interactions from being used when learning from earlier observations. Dataset statistics after preprocessing are summarized in Table 3

Dataset	# users	# items	Avg. $\lvert\mathcal{S}_{\mathbf{u}}(\mathbf{t})\rvert$	# interactions		
				Train	Validation	Test
Micro-Video	25,871	44,503	8.1	191,378	10,134	9,038
MovieLens-1M	5,331	3,522	96.5	460,089	12,984	41,260
KuaiRand	22,128	7,076	28.1	496,142	61,631	63,291

Table 3: Dataset statistics after preprocessing. Avg. $|\mathcal{S}_u(t)|$ denotes the average length of the user interaction history.

Compared Methods. The compared methods are divided into two groups: general recommendation backbones and popularity-debiasing methods. The first group covers representative architectures with different mechanisms for modeling user preference. MF (Koren et al., 2009) learns latent user and item representations through inner-product matching. For sequential recommendation, GRU encodes interaction histories recurrently, and SASRec (Kang & McAuley, 2018) models sequential dependencies using self-attention. TiSASRec (Li et al., 2020) further incorporates relative time intervals between interactions. FEARec (Du et al., 2023) combines time- and frequency-domain representations, whereas BSARec (Shin et al., 2023) introduces a frequency-aware inductive bias to balance low- and high-frequency sequential information. These models are also used as backbones for evaluating the proposed framework.

The second group consists of methods designed specifically to alleviate popularity bias. DICE (Zheng et al., 2021a) separates user interest from conformity through causal representation learning, while TIDE (Zhao et al., 2022) decomposes observed interactions into preference, item quality, and dynamic conformity effects. DCRec (Yang et al., 2023) uses conformity-aware contrastive learning, and MELT (Kim et al., 2023) addresses long-tailed user and item distributions through mutually enhanced representations. PAUDRec (Wu et al., 2023) models dynamic user desire and item popularity with causal intervention, whereas PAAC (Cai et al., 2024) transfers supervision from popular to unpopular items through popularity-aware alignment and contrastive learning. TPAB (Yoo et al., 2025) explicitly considers temporal popularity shifts, and DDC (Liu et al., 2025) mitigates the dominant popularity direction induced during collaborative filtering optimization.

Implementation Details. All models use an embedding dimension of 128 and are optimized with Adam using a batch size of $2\,048$ and a learning rate of 0.001 for 500 epochs. For each baseline, the remaining architecture-specific hyperparameters follow the search protocol reported in the corresponding original work. The backbone configuration is selected first and then kept fixed when applying the proposed framework. Hyperparameters are selected based on Recall@10 on the validation set.

For the proposed method, the Hawkes popularity model and the recommendation model use separate 128-dimensional item embeddings. The centering coefficient γ is selected from $\{0.1, 0.5, 1.0, 3.0, 10.0\}$. The importance-corrected choice loss uses 15 auxiliary items per observed interaction, sampled uniformly from $\mathcal{V} \setminus \{v_i\}$ and refreshed at every optimization step. For the cen-

tering loss, two users are independently sampled for each event from users active in the same time bucket.

Each experiment is repeated with four random seeds, and results are reported as mean ± standard deviation. All experiments are conducted on an NVIDIA A10 GPU with 64 GB of system memory under Rocky Linux 8.8, using Python 3.8.0, PyTorch 1.12.1, and CUDA 11.3.

A.7 Analysis of Tail-Item Degradation on KuaiRand

Table 1 shows a distinct pattern on KuaiRand. While the proposed framework improved All- and Head-item performance, Tail Recall@10 decreases relative to the corresponding backbone for GRU, SASRec, TiSASRec, and FEARec. This section examines whether the degradation originates from the learned utility component or from its combination with the Hawkes-based popularity component at inference.

Table 4: Utility-only ($\eta = 0$) vs. backbone, all six KuaiRand backbones, Tail Recall@10, mean ± std over the same 4 seeds.

Backbone	Backbone R@10	Ours @ default η	Ours @ $\eta = 0$
MF	0.0100 ± 0.0007	0.0209 ± 0.0009 (η=0.3)	$\mathbf{0.0504 \pm 0.0008}$
GRU	0.0100 ± 0.0010	0.0018 ± 0.0003 (η=0.5)	$\mathbf{0.0201 \pm 0.0006}$
SASRec	0.0268 ± 0.0013	0.0036 ± 0.0001 (η=0.5)	$\mathbf{0.0317 \pm 0.0014}$
FEARec	0.0054 ± 0.0001	0.0008 ± 0.0004 (η=0.5)	$\mathbf{0.0082 \pm 0.0003}$
BSARec	0.0000 ± 0.0000	0.0020 ± 0.0003 (η=0.5)	$\mathbf{0.0216 \pm 0.0006}$
TiSASRec	0.0277 ± 0.0012	0.0035 ± 0.0004 (η=0.5)	$\mathbf{0.0290 \pm 0.0017}$

From the final ranking score in equation 3, we re-evaluate the same trained checkpoints while varying η, without updating any model parameter. Table 4 shows that the utility-only inference ($\eta = 0$) numerically matches or exceeds the corresponding backbone for all six architectures. These results suggest that the Tail-item degradation arises mainly from adding the learned popularity component at inference, rather than poor Tail-item ranking by the utility model itself.

Table 5: Normalized entropy and top-5% probability mass for empirical 3-day popularity snapshots and the corresponding trained Hawkes popularity (SASRec+Ours, seed 1), reported as empirical → Hawkes.

Statistic	**Micro-Video**	**MovieLens**	**KuaiRand**
Normalized entropy, H_{norm}	$0.750 \rightarrow 0.822$	$0.624 \rightarrow 0.701$	$\mathbf{0.828 \rightarrow 0.242}$
Top-5% probability mass	$0.794 \rightarrow 0.621$	$0.802 \rightarrow 0.740$	$\mathbf{0.520 \rightarrow 0.994}$

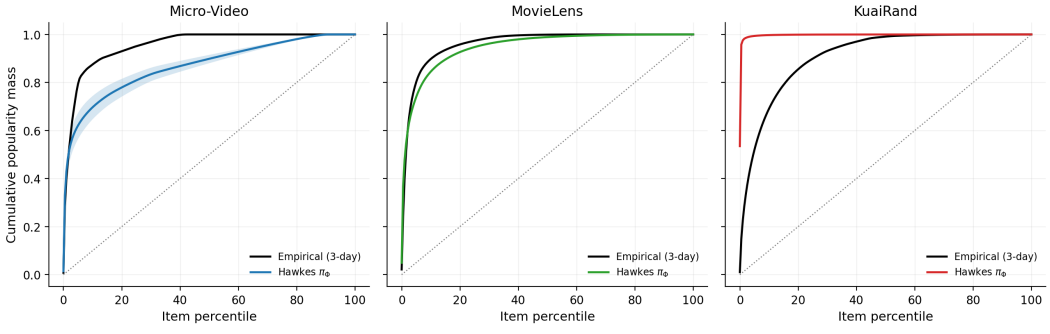


Figure 4: Time-averaged cumulative popularity mass of empirical 3-day popularity and the trained Hawkes popularity $\pi_\Phi(v \mid t)$ across the three datasets.

At daily 3-day-window snapshots, KuaiRand's own local popularity is less concentrated and more stable than Micro-Video's or MovieLens's (normalized entropy $H_{norm} = 0.828$ in Table 5) Yet, the trained Hawkes intensity $\pi_\Phi(v \mid t)$ fit to this data is extremely over-concentrated – but only on KuaiRand: H_{norm} collapses to 0.242 and top-5% mass rises from 0.520 to 0.994. While on Micro-Video and MovieLens the fitted model is mildly less concentrated than its own empirical baseline (Figure 4 and Table 5). Since this gap does not exist in the data the model was fit to, it is a genuine model-side over-concentration. Binning items by training-frequency percentile shows why: μ_v grows smoothly into the top decile on Micro-Video and MovieLens (37–45$\times$ from the 80–90th to the 99–100th percentile), but jumps discontinuously by $5{,}872.7\times$ over the same range on KuaiRand – a qualitatively different, discontinuous relationship absent from the other two datasets (Figure 5a).

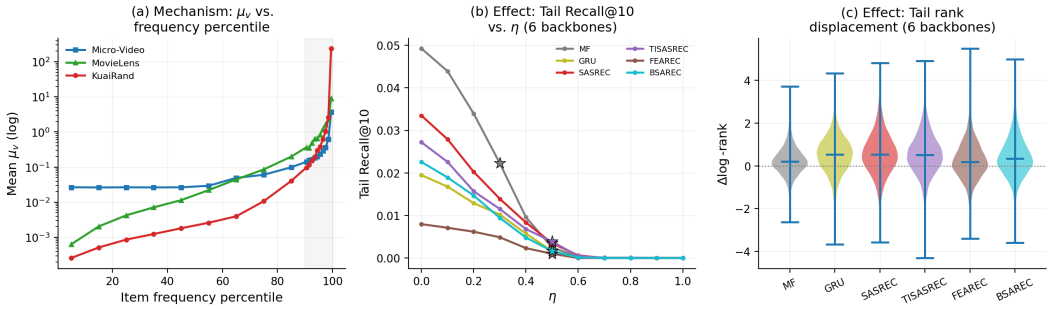


Figure 5: (a) Mean learned μ_v across item training-frequency percentiles. (b) KuaiRand Tail Recall@10 across inference-time mixing weights η for six backbones, with $\star$ indicating the trained default setting. (c) Distribution of Tail-positive log-rank displacement from the utility-only score to the combined score across six backbones.

This over-concentrated term systematically damages Tail ranking once reintroduced via $s = \eta \log \pi_\Phi(v|t) + (1-\eta) f_\Psi\big(x_u(t), v\big)$: Tail Recall@10 decreases strictly monotonically in η for all six KuaiRand backbones from $\eta = 0.1$ onward, and the true item's rank worsens for the majority of Tail test events across every backbone (Figure 5b–c). Because tail items are by construction not part of the currently-dominant set, the small over-concentrated clique occupies top-10 slots regardless of individual preference, mechanically suppressing Recall@10 on exactly this evaluation slice.

Table 6: KuaiRand Tail Recall@10 for SASRec+Ours (seed 1) with the default $\eta = 0.5$, using the same utility checkpoint f_Ψ across all variants and replacing only the popularity term.

Popularity source	Tail Recall@10
Utility only (f_Ψ alone)	0.0335
+ Learned Hawkes π_Φ	**0.0035**
+ 3-day sliding-window empirical	**0.0121**
+ Static training-frequency	**0.0140**

This vulnerability is not specific to the Hawkes parameterization: replacing it with a much simpler 3-day sliding-window or static training-frequency popularity – same utility checkpoint, same η – damages Tail Recall@10 in the same direction, though less severely (Table 6). Reintroducing any strongly concentrated, population-level popularity signal through a single global η is thus a general risk for Tail items; KuaiRand's unconstrained, unregularized background-rate parameterization simply amplifies this risk far beyond what the other two popularity sources, or the same architecture fit to the other two datasets, produce. The popularity–utility decomposition itself remains effective – the utility component alone already matches or exceeds the backbone on every architecture – so the fix belongs to the popularity specification: a bounded/regularized background rate, item-adaptive or multi-timescale decay, or an explicitly time-varying $\mu_v(t)$ are natural directions for future work.

B THE USE OF LARGE LANGUAGE MODELS

Large Language Models (LLMs) were employed exclusively for the purpose of correcting grammatical errors and enhancing the clarity of expression.