Does Diversity Actually Improve Long-Term Retention? An Empirical Investigation with Failure Analysis of RL-Based Diversity Control

Hyun Kang[1], Ji-Hyeong Han*[1]

[1]Department of Computer Science and Engineering, Seoul National University of Science and Technology, Seoul, Republic of Korea

Abstract

A widely held assumption in recommender systems research is that presenting diverse content leads to higher long-term user retention. We test this assumption by building a complete pipeline that learns a context-aware diversity weight λ via Soft Actor-Critic (SAC) to control a Maximal Marginal Relevance (MMR) re-ranker on top of a SASRec retriever, and by analyzing 231,829 real sessions from KuaiRand. Two findings emerge. First, the trained SAC agent collapses to a near-constant $\lambda \approx 0.505$ and is no better than the best static-λ baseline ($\Delta = 0.0101$ days); a three-stage diagnosis (Actor, Critic, environment) traces this not to a learning-algorithm failure but to a KuaiSim retention reward that barely responds to diversity. Second, and explaining why, the diversity–retention signal in the raw logs is near-null once user activity is controlled. Naive within-user correlations look strong and even flip sign across diversity metrics, but this is a session-length confound: adjusting for it collapses all three metrics to a weak, uniform faster-return association (correlation r between -0.02 and -0.035, $r^2 \lesssim 0.1\%$). Diversity's link to retention is thus real but negligible and confound-dominated—too weak for a scalar diversity reward to learn from, which is exactly why our agent collapses. Observational diversity–retention correlations, we conclude, are dominated by activity confounds and must not be taken as a reward signal without controlling for them.

Keywords

recommender systems, diversity, long-term retention, reinforcement learning, Soft Actor-Critic, maximal marginal relevance

1. Introduction

Recommender systems optimizing for long-term user retention face a fundamental tension: a recommendation that maximizes immediate engagement may not be the one that keeps the user coming back tomorrow. One widely adopted remedy is to inject *diversity* into the recommendation slate, on the intuition that a varied diet of content maintains novelty and prevents saturation [1, 2]. This diversity-retention hypothesis is so commonly assumed that diversity is routinely injected as a fixed re-ranking objective [1, 3] or folded into RL training objectives [4], treating the diversity–retention link as given.

Yet the hypothesis has rarely been tested in isolation on realistic industrial data. Does exposure to more diverse content *actually* make users return sooner and more frequently? And if so, can a reinforcement learning agent learn, from user interaction signals alone, how much diversity to inject for each user at each point in time?

In this paper, we answer both questions by building a complete, end-to-end pipeline and running it to completion:

1. **SASRec** encodes the user's interaction history and retrieves the top-100 candidate items.
2. **MMR** (Maximal Marginal Relevance) re-ranks these candidates using a single scalar $\lambda \in [0, 1]$ that interpolates between pure relevance ($\lambda = 0$) and maximum diversity ($\lambda = 1$).
3. **SAC** (Soft Actor-Critic) learns to predict λ from the user's current context, trained on retention reward signals from the KuaiSim simulation environment.

FRAME'26: Methodology First - Rethinking Research Assessment in RecSys Workshop, September 28, 2026, Minneapolis, Minnesota, USA

maxkang01@seoultech.ac.kr (H. Kang); jhhan@seoultech.ac.kr (J. Han*)

0009-0000-6207-5700 (H. Kang); 0000-0001-8391-6898 (J. Han*)

© 2026 Copyright © 2026 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-26

CEUR-WS.org/Vol-4341/paper5.pdf

Our experiments yield two findings. First, the SAC agent collapses to a nearly constant $\lambda \approx 0.505$ regardless of user context (a phenomenon we term *diversity collapse*) and is no better than the best static-λ baseline ($\Delta = 0.0101$ days under matched evaluation seeds). Second, a large-scale analysis of 231,829 real KuaiRand sessions shows *why*: once user activity (session length) is controlled, the diversity–retention association is near-null. Uncontrolled within-user correlations look strong and even change sign across diversity metrics, but that structure is a session-length confound; adjusting for it leaves only a weak, uniform, and practically negligible signal—too little for a learned controller to exploit.

These findings do not reflect a broken pipeline—each component works as designed, as we verify through component-level validation and baseline comparisons (Section 4). Together they show that the diversity–retention link is too weak and confound-ridden to serve as a reward signal here: the first is established via a three-stage diagnosis (Section 5) that rules out a learning-algorithm defect, and the second (Section 6) shows, from raw behavior, that the signal itself is near-null once activity is controlled.

Contribution. Our contribution is a cautionary case study for RL-based diversity control: a learned diversity controller collapses in a realistic retention pipeline because the underlying diversity–retention signal is too weak to reward, in both the simulator and 231,829 real KuaiRand sessions. The practical lesson is that observational diversity–retention correlations are dominated by activity confounds and must not be taken as a reward signal without controlling for them—yet any pipeline that rewards a learned diversity controller, ours included, rests on exactly such an uncontrolled correlation.

2. Related Work

SASRec [5] applies multi-head self-attention to a user's interaction sequence and is a de facto retrieval backbone in industrial re-ranking pipelines [6]; we use it as our retriever, trained on KuaiRand-Pure. For diversity-aware re-ranking, Maximal Marginal Relevance (MMR) [1] trades relevance against redundancy via a fixed scalar λ, and Determinantal Point Processes (DPP) [7] provide a probabilistic framework for diverse subset selection, with fast greedy MAP inference [2] deployed at YouTube scale [3]. Both treat the diversity weight as a fixed hyperparameter; we instead *learn* it per user per context via RL.

For long-term optimization, Zou et al. [4] cast recommendation as an MDP to improve engagement, finding that myopic click-rate objectives yield suboptimal retention; Zheng et al. [8] replaced ϵ-greedy exploration with Dueling Bandit Gradient Descent to avoid near-duplicate recommendations; neither learns a diversity-control parameter. Our environment is KuaiSim [9], a simulator built on KuaiRand-Pure [10] (Kuaishou logs with randomized exposure); we use its cross-session (retention) environment and User Immediate Response Model (UIRM) checkpoint, replacing the default item selector with our SASRec–MMR pipeline.

Whereas this literature studies *how* to inject diversity, *whether* it improves long-term retention is far less examined. The closest study is Holtz et al. [11], whose Spotify field experiment found that personalized—and thus less diverse—recommendations raised engagement 28.9% while cutting individual consumption diversity 11.5%. We differ by studying short-video consumption on KuaiRand, examining within-user diversity–return correlation on natural logs rather than an intervention, and connecting the finding to RL agent failure.

3. Method

3.1. Problem Formulation

At each step t within a session, the agent observes the current user state s_t (the hidden state of KuaiSim's frozen UIRM) and selects a single item a_t to recommend. The session runs for max_step $= 5$ steps, after which the environment returns a retention reward equal to $-\text{return_day}/10$, where return_day is

the number of days until the user's next session, which the environment derives from the same UIRM. The agent's goal is to maximize the expected discounted sum of retention rewards.

Our key design choice is to *factorize* the agent's action: instead of selecting items directly from the full catalog (7,583 items), the agent outputs a scalar $\lambda \in [0, 1]$ that parameterizes an MMR re-ranker operating on a pre-computed top-100 candidate set.

3.2. SASRec Retriever

We train SASRec on KuaiRand-Pure interaction logs using RecBole [12], treating each user's click history as an item sequence. At inference time, given user history $\mathcal{H}_u$, SASRec produces a user embedding $\mathbf{e}_u$ and scores all 7,583 items, returning the top-100 candidates $\mathcal{C}_u = \{(i_1, \mathbf{e}_1), \ldots, (i_{100}, \mathbf{e}_{100})\}$, where $\mathbf{e}_k \in \mathbb{R}^{64}$ is the SASRec item embedding.

ID-space bridging. KuaiSim, RecBole, and KuaiRand use three mutually incompatible item ID spaces. We maintain two lookup tables—KuaiSim→KuaiRand and KuaiRand→RecBole—and apply them transparently at retrieval time. Retrieval runs once per session since within a 5-step session the user history changes minimally.

3.3. MMR Re-ranker

At each step t, the re-ranker selects one item from $\mathcal{C}_u$ using an MMR-style score:

$$\text{score}(i) = (1 - \lambda) \cdot \text{rel}(i) + \lambda \cdot \Big(1 - \max_{j \in \mathcal{B}_t} \cos(\mathbf{e}_i, \mathbf{e}_j)\Big), \tag{1}$$

where $\text{rel}(i)$ is a rank-based relevance proxy (linearly spaced from 1.0 at rank 1 to 0.0 at rank 100), $\mathcal{B}_t$ is the session buffer of already recommended items, and the diversity term measures dissimilarity from previously recommended items. Selected items are masked with $-\infty$ to prevent repetition. The buffer resets at the start of each session. When $\lambda = 0$, Eq. (1) reduces to greedy relevance selection. When $\lambda = 1$, it maximizes diversity from $\mathcal{B}_t$.

3.4. SAC Policy

We use Soft Actor-Critic [13] with a learnable temperature parameter. λ denotes MMR's diversity-relevance weight (following MMR convention [1]), distinct from SAC's entropy temperature α, which is learned separately. The actor is a two-layer MLP that takes s_t (UIRM hidden state, dim. 192) as input and outputs the mean μ of a Gaussian, then passed through a sigmoid to produce $\lambda \in (0, 1)$. The critic consists of two independent Q-networks (double Q-learning) taking (s_t, λ_t) as input. Temperature is updated to maintain a target entropy of -1.

Learning signal. The only training reward is the retention reward of Section 3.1 ($-\text{return_day}/10$), delivered once per 5-step session; no immediate click/watch reward is added, and the Critic regresses its discounted return via the standard SAC target. Because the action is the scalar λ, the only pathway to change this reward is the diversity of the slate—yet return_day varies mostly for reasons unrelated to diversity (activity level, content preferences, session context). The size of that controllable, diversity-driven component is exactly what Section 5.3 measures.

Training details. Replay buffer 100,000; batch size 256; learning rates 3×10^{-4}; discount factor $\gamma = 0.99$; target smoothing coefficient $\tau = 0.005$; 200,000 steps.

3.5. Baselines

We compare SAC against four baselines.

Fixed MMR (λ^*): λ fixed to the grid-search optimum over $\{0.0, 0.1, \ldots, 1.0\}$. Oracle-tuned static controller.

Greedy ($\lambda = 0$): Pure SASRec-ranked selection; no-diversity baseline.

Random λ: $\lambda \sim \mathcal{U}[0,1]$ per step, independently of user state. Stochastic lower bound.
Heuristic λ: $\lambda = 1 - \widetilde{v}$, where $\widetilde{v}$ is the variance of the top-100 candidate embeddings, min-max normalized across the batch. Clustered candidates (low variance) receive a high λ to force diversity, already-spread candidates a low λ; a non-learned but state-dependent controller.

4. Component Validation and Policy Comparison

4.1. Component-Level Validation

Running 200 randomly sampled users through the full pipeline, we verify three properties. **Embedding diversity:** pairwise cosine similarity of the top-100 candidates spans a wide range (mean 0.267, std 0.181), so the diversity kernel is not degenerate. **MMR λ-sensitivity:** across 4,800 steps, selections under $\lambda = 0$ and $\lambda = 1$ differ in **96.2%** of steps. **Change in ILD:** ILD (Intra-List Diversity), which is higher when recommended items are more dissimilar, rises **25%**, from 0.222 ($\lambda = 0$) to 0.277 ($\lambda = 1$), as intended; whether this translates to retention is examined in Section 5. These results confirm the pipeline works correctly, so any failure to learn λ must originate elsewhere.

4.2. Multi-Seed Evaluation with Confidence Intervals

To obtain a robust characterization of policy performance across user populations, we evaluate all policies across seeds $\{0, 1, 2, 3, 4\}$ and report 95% confidence intervals (t-distribution, df $= 4$) in Table 1. The paired difference between SAC and Fixed MMR is statistically significant (paired $t = 5.4$, $p = 0.006$, df $= 4$), with SAC performing slightly worse. However, the 95% CI for this paired difference (not the per-policy CIs in the table) has an upper bound of 0.016 days ($\approx$23 minutes)—negligibly small relative to the per-user return-day standard deviation of 2.6 days. All other baselines likewise fall within 0.020 days of Fixed MMR.

This pattern—no policy achieving a practically meaningful advantage—is precisely what one would predict if SAC has collapsed to outputting a near-constant λ. We provide a mechanistic account of why this collapse occurs in Section 5; the performance numbers here are a consequence of that collapse, not independent evidence for it.

Table 1
Multi-seed evaluation (mean return_day ↓). Δ relative to Fixed MMR ($\lambda^* = 0.1$).

Policy	Mean return_day	95% CI	Δ relative to λ^*
Fixed MMR (λ^*)	3.0825	[3.050, 3.116]	0 (ref.)
Greedy ($\lambda = 0$)	3.0854	[3.056, 3.115]	+0.0029
Heuristic λ	3.0871	[3.056, 3.118]	+0.0046
SAC (ours)	3.0925	[3.056, 3.129]	+0.0101
Random λ	3.1020	[3.064, 3.140]	+0.0195

5. Why SAC Fails: λ-Collapse Diagnosis

Section 4.2's outcome is no coincidence: the SAC agent produces $\bar{\lambda} = 0.5057 \pm 0.0041$ across 9,600 observations[1]—effectively a constant, a phenomenon we term *diversity collapse*. The issue is not the value 0.505 but that the agent fails to modulate λ with user context—the capability it was meant to learn. We diagnose *why* through a three-stage framework.

[1]KuaiSim runs 32 users per episode, up to 6 sessions per episode, and 5 steps per session; the differing counts across analyses (e.g., 9,600 observations, 4,800 steps, 5,760 sessions) are episodes × users × sessions × steps or the corresponding session counts, and differ only in the number of evaluation episodes.

5.1. Stage 1: Actor Output Distribution

The actor's pre-sigmoid output μ has mean 0.0227 and std **0.0163**, so 95% of the post-sigmoid λ fall within $[0.498,\ 0.514]$ (1.6% of the $[0, 1]$ range) across 9,600 observations: for all practical purposes a constant. This context-*invariance*, not the value 0.505, is the failure: a Critic with a context-dependent Q would move μ across states even under the same entropy target, so a std(μ) this small reflects an actor insensitive to context, not merely entropy regularization.

5.2. Stage 2: Critic Q-Landscape

Sweeping λ across 11 evenly-spaced values from 0 to 1 at 200 randomly sampled user states, Q mean moves from -0.4594 at $\lambda = 0$ to -0.4570 at $\lambda = 0.5$ to -0.4551 at $\lambda = 1$, a Q-range of only 0.0043, less than 1% of the mean Q magnitude ($|\bar{Q}| \approx 0.457$). The Critic assigns essentially the same value to all λ levels.

5.3. Stage 3: Environment Reward Sensitivity

We measure how strongly the environment's retention signal responds to diversity by the Pearson correlation r between a session's ILD (Section 4.1) and its return_day. Since a lower return_day is a faster return (Section 3.1), a positive r means higher diversity is followed by a *slower* return.

Fixing $\lambda = 0$ gives a mean ILD of 0.222 and a mean return_day of 2.968, with $r = 0.021$; fixing $\lambda = 1$ raises the mean ILD by 25% to 0.277 but shifts the mean return_day only marginally, to 2.989, with $r = 0.054$. A 25% increase in diversity therefore lengthens return_day by just 0.021 days. Both correlations are positive but tiny: r ranges from 0.021 to 0.054, which—although statistically detectable at $n \approx 5{,}760$ sessions—corresponds to $r^2 \leq 0.003$, so diversity explains at most 0.3% of the variance in return_day and is thus practically negligible. We compare the two extremes to maximize the ILD contrast; a supplementary sweep across the 11 grid points shows the same pattern, with per-λ means falling within one standard error of each other at this sample size.

Note: Stage 3 is evaluated with n=30 episodes at seed 0 using fixed-λ policies, independent of the policy comparison in Section 4.2; absolute return_day values differ from Table 1 due to the smaller episode count.

Conclusion: KuaiSim's retention reward barely responds to embedding-based diversity (Pearson $r \leq 0.054$). This is not independent evidence about the real world: KuaiSim's reward is generated by a UIRM trained on KuaiRand, so it necessarily inherits whatever diversity–retention signal that data contains—and Section 6 shows that signal is itself near-null once activity is controlled. The simulator's role here is therefore not to corroborate that finding but to explain the collapse: the agent learns inside an environment whose diversity signal is, expectedly, too small relative to other reward variance for the Critic to exploit.

5.4. Mechanism of Collapse

The stages chain into one causal account, read from environment to actor. Stage 3's weak reward response to diversity (Pearson $r \leq 0.054$) leaves the Critic (Stage 2) almost no value difference between high- and low-diversity actions, so its Q is flat in λ. Because the SAC actor gradient carries $\frac{\partial Q}{\partial \lambda} \cdot \frac{\partial \lambda}{\partial \theta}$, a flat Q-landscape ($\frac{\partial Q}{\partial \lambda} \approx 0$) removes it, leaving only the entropy bonus; for a sigmoid action this drives μ toward 0, so λ settles near $\sigma(0) = 0.5$, with the residual $\bar{\mu} = 0.0227$ giving the observed $\bar{\lambda} \approx 0.505$ (Stage 1). The collapse thus originates in the weak environment signal, not the learning algorithm.

6. Does Diversity Actually Help Retention?

Does diversity improve retention in *real* user behavior, beyond what the simulator captures? The simulator's near-zero response to embedding diversity (Section 5.3) could reflect either no underlying relationship or a masked one. KuaiRand's raw logs point to the former: what looks like a strong,

even sign-flipping relationship is largely a session-length confound, and once activity is controlled the association is near-null (Section 6.4)—so the simulator's flat response is faithful to the data, not a fidelity artifact.

6.1. Dataset and Setup

KuaiRand-Pure contains 2,579,779 interactions from 27,285 users over 30 days. Sessions are defined at the user×date level (minimum 3 items per session), yielding **231,829 sessions** from 26,695 users.[2] The aggregate and binned analyses (Sections 6.2–6.3) use all 231,829 sessions; the within-user analysis (Section 6.4) additionally requires ≥ 5 sessions per user.

Diversity metrics (per session): `tag_unique_ratio` (fraction of unique tags), `video_type_entropy`, `music_type_entropy`.

Retention metric: `return_gap` = days until next session (lower → faster return → better) — the real-log analog of the simulator's `return_day` (Section 3.1), computed from actual user sessions rather than KuaiSim's UIRM. It is defined only for sessions followed by another observed session; each user's final session in the 30-day window has no successor and is dropped, right-censoring the small fraction of sessions near the window's end.

6.2. Aggregate Correlations

Table 2 reports the session-level association between each diversity metric and the return gap across all 231,829 sessions.

Table 2
Session diversity vs. return gap ($n = 231{,}829$). The final column is the partial correlation controlling for session length; after this control all three metrics are weakly negative (faster return), collapsing the raw sign disagreement. $^{***}p < 0.001$.

Diversity metric	Pearson r	Spearman r	Partial r (ctrl. len)
tag_unique_ratio	$+0.088^{***}$	$+0.174^{***}$	-0.021^{***}
video_type_entropy	-0.026^{***}	-0.088^{***}	-0.014^{***}
music_type_entropy	-0.041^{***}	-0.076^{***}	-0.017^{***}

The raw Pearson and Spearman columns *disagree in sign* across metrics: `tag_unique_ratio` is positive (topic diversity → slower return) while the two entropy metrics are negative (format diversity → faster return). This apparent split is an activity confound. Session length is *negatively* correlated with `tag_unique_ratio` ($r = -0.60$; a per-item ratio that falls as sessions add repeated tags) but *positively* with the entropy metrics ($r = +0.07$ and $+0.14$), and longer, more active sessions return sooner ($r = -0.17$); a metric that co-varies with session length therefore inherits a spurious return-gap association of the matching sign. The partial-correlation column, which holds session length fixed, removes this: all three metrics then agree on a weak negative (faster-return) association (-0.014 to -0.021). The raw sign disagreement is thus a confound, not a property of diversity, a conclusion the within-user analysis (Section 6.4) confirms.

6.3. Binned Analysis

Because `tag_unique_ratio` is heavily tied at its maximum (29% of sessions use a distinct tag for every item), a quartile split collapses into three unequal bins, which we report directly. Binning sessions into $(0.08, 0.625]$, $(0.625, 0.75]$ and $(0.75, 1.0]$, the mean return_gap and next-day return rate (fraction with return_gap ≤ 1) are: low, $n = 60{,}456$, 2.17 days, **76.4%**; mid, $n = 55{,}872$, 2.81 days, 62.7%; high, $n = 115{,}501$, 2.83 days, 57.4%. Low-*topic*-diversity sessions return fastest. This mirrors the raw tag

[2] All raw-data statistics reported in this section are computed from per-session records (one row per user–date session, with the diversity metrics, session length, and return gap). The per-session table and a one-command reproduction script are available at https://github.com/HyunWrites0721/diversity-retention-kuairand.

correlation but, like it, is uncontrolled for session length (Section 6.2), a confounded pattern that the activity-adjusted analysis below overturns. We therefore treat the within-user, length-adjusted analysis as the primary evidence.

6.4. Within-User Analysis

To remove between-user heterogeneity, we compute, for each user, the within-person Spearman correlation between a diversity metric and return_gap, then test whether the mean of these per-user correlations (denoted $\bar{r}$) is zero across users. We restrict to users with ≥ 5 sessions, drop those whose sessions give a degenerate correlation, and run the analysis for all three metrics—both raw and after adjusting for session length within each user (Table 3).

Table 3
Within-user diversity–retention correlation, before and after adjusting for session length (partial correlation). Negative r = diversity associated with *faster* return. The raw signs disagree across metrics; adjusting for session length collapses them to a uniform, weak, faster-return association. $^{***}p < 0.001$.

Diversity metric	Users	Raw $\bar{r}$	Length-adj. $\bar{r}$
`tag_unique_ratio`	20,018	$+0.136^{***}$	-0.035^{***}
`video_type_entropy`	11,536	-0.103^{***}	-0.021^{***}
`music_type_entropy`	19,585	-0.064^{***}	-0.021^{***}

The raw signs are a session-length artifact. The raw column looks heterogeneous—topic diversity predicts *slower* return, format diversity *faster*—but this is the same activity confound as at the aggregate level (Section 6.2). Adjusting for session length within each user collapses all three metrics to a weak *faster*-return association ($\bar{r}$ from -0.021 to -0.035): the apparent dimension-dependence is a confound, not a property of diversity.

The adjusted signal is real but negligible. The length-adjusted correlations are all significantly negative—weakly *supporting* the diversity-improves-retention hypothesis—but tiny ($r^2 \lesssim 0.1\%$), and they weaken or vanish under coarser controls and alternative retention measures (Section 7.3); being observational, we cannot even read this as causal. The robust conclusion is thus not a direction but a *magnitude*: once activity is controlled, diversity's association with return timing is near-null—which is why the embedding-based diversity signal is so weak in the simulator (Section 5.3), and why a scalar diversity controller has almost nothing to learn.

7. Discussion

7.1. Implications for RL-based Recommendation

RL-based diversity control tacitly assumes a diversity–retention signal strong enough for the Critic to learn. Our raw-data analysis (Section 6.4) shows this fails at the root: after controlling activity the signal is near-null, and before controlling it is a confound whose sign depends on the metric. A scalar diversity reward therefore has almost nothing real to latch onto—exactly what the simulator exhibits (Section 5.3) and what starves the SAC agent. The lesson is to *measure a reward signal's strength, net of obvious confounds, before optimizing it*: no optimizer can recover a signal that is not there. The three-stage diagnosis (Section 5) operationalizes this—trace a low std(μ) to a flat Q-landscape to a confound-dominated reward before blaming the optimizer.

7.2. When Might Diversity Help Retention?

We do not rule diversity out: the small activity-adjusted association points, if anywhere, toward faster return, but at these magnitudes it cannot carry a reward. Future work should look for a stronger, cleaner

signal—specific segments (new users, exploration phases), longer horizons, or interventional (A/B) manipulation to break the confound—before optimizing it.

7.3. Limitations

Single platform. We use only KuaiRand (short video); other modalities may differ. **Session-level diversity.** Effects may act over longer (weekly) timescales our metric misses. **Simulator fidelity.** KuaiSim's retention signal derives from a UIRM trained on the same data; a sharper test needs explicit A/B manipulation. **Diversity operationalization.** Section 5.3 uses embedding-based ILD, Section 6 content-category metrics; we assume but do not show these correspond. **Signal fragility.** The activity-adjusted association is not only tiny but specification-sensitive—it weakens under a log-length control and the tag effect washes out for a binary "returned within two days" outcome—so we claim no robust direction. **Narrow diversity range.** The re-ranker selects 5 of 100 relevance-filtered items, so the diversity span is small. **Causal direction.** The within-user correlation (Section 6.4) is observational; a latent state (e.g., fatigue, mood) could drive both, which only randomizing diversity could rule out.

8. Conclusion

Our SASRec–MMR–SAC pipeline works at the component level, yet the learned diversity controller collapses to a near-constant λ and is no better than the static-λ baselines. A three-stage diagnosis (Section 5) traces this to a KuaiSim retention reward that barely responds to diversity—the environment, not the optimizer. Section 6 shows why: in the raw logs the diversity–retention signal is near-null once user activity is controlled. Naive within-user correlations look strong and even flip sign across metrics, but that is a session-length confound; the activity-adjusted association is uniformly small (correlation r between -0.02 and -0.035). No diversity signal here is strong or stable enough to reward.

We do not invalidate diversity as an objective, nor generalize beyond KuaiRand. Yet the failure we document is a general risk: a diversity bonus added to an RL reward, as in the pipeline we built, optimizes exactly this kind of confounded, near-null signal. We caution against reading observational diversity–retention correlations as effect sizes—they are activity-confounded and can flip sign with the metric or outcome—and against rewarding diversity without checking that a signal survives such confounds. Future work should establish a genuine, confound-free signal (in specific segments, over longer horizons, or via A/B diversity manipulation) before optimizing it.

Acknowledgement

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2024-00407295).

Declaration on Generative AI

The authors used Claude (Anthropic) for editing assistance (prose clarity, LaTeX formatting) and take full responsibility for this publication's content.

References

[1] Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. *SIGIR*.

[2] Chen, L., Zhang, G., & Zhou, E. (2018). Fast greedy MAP inference for determinantal point process to improve recommendation diversity. *NeurIPS*.

[3] Wilhelm, M., Ramanathan, A., Bonomo, A., Jain, S., Chi, E. H., & Gillenwater, J. (2018). Practical diversified recommendations on YouTube with determinantal point processes. *CIKM*.

[4] Zou, L., Xia, L., Ding, Z., Song, J., Liu, W., & Yin, D. (2019). Reinforcement learning to optimize long-term user engagement in recommender systems. *KDD*.
[5] Kang, W.-C., & McAuley, J. (2018). Self-attentive sequential recommendation. *ICDM*.
[6] Covington, P., Adams, J., & Sargin, E. (2016). Deep neural networks for YouTube recommendations. *RecSys*.
[7] Kulesza, A., & Taskar, B. (2012). Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*.
[8] Zheng, G., Zhang, F., Zheng, Z., Xiang, Y., Yuan, N. J., Xie, X., & Li, Z. (2018). DRN: A deep reinforcement learning framework for news recommendation. *WWW*.
[9] Zhao, K., Liu, S., Cai, Q., Zhao, X., Liu, Z., Zheng, D., Jiang, P., & Gai, K. (2023). KuaiSim: A comprehensive simulator for recommender systems. *NeurIPS*.
[10] Gao, C., Li, S., Zhang, Y., Chen, J., Li, B., Lei, W., Jiang, P., & He, X. (2022). KuaiRand: An unbiased sequential recommendation dataset with randomly exposed videos. *CIKM*.
[11] Holtz, D., Carterette, B., Chandar, P., Nazari, Z., Cramer, H., & Aral, S. (2020). The engagement-diversity connection: Evidence from a field experiment on Spotify. *EC*.
[12] Zhao, W. X., Mu, S., Hou, Y., Lin, Z., Chen, Y., Pan, X., Li, K., Lu, Y., Wang, H., Tian, C., Min, Y., Feng, Z., Fan, X., Chen, X., Wang, P., Ji, W., Li, Y., Wang, X., & Wen, J.-R. (2021). RecBole: Towards a unified, comprehensive and efficient framework for recommendation algorithms. *CIKM*.
[13] Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *ICML*.