中国科学技术大学
University of Science and Technology of China

快手
KUAISHOU TECHNOLOGY

Alleviating Matthew Effect of Offline Reinforcement Learning in Interactive Recommendation

SIGIR 2023 Full Paper

Chongming Gao[1] , Kexin Huang[1] , Jiawei Chen[2,] , Yuan Zhang[3] , Biao Li[3] , Peng Jiang[3] , Shiqi Wang[4] , Zhong Zhang[5] , Xiangnan He[1,*]*

[1]University of Science and Technology of China; [2]Zhejiang University, China; [3]Kuaishou Technology Co., Ltd; [4]Chongqing University, China; [5]University of Science and Technology of China (*Corresponding author)

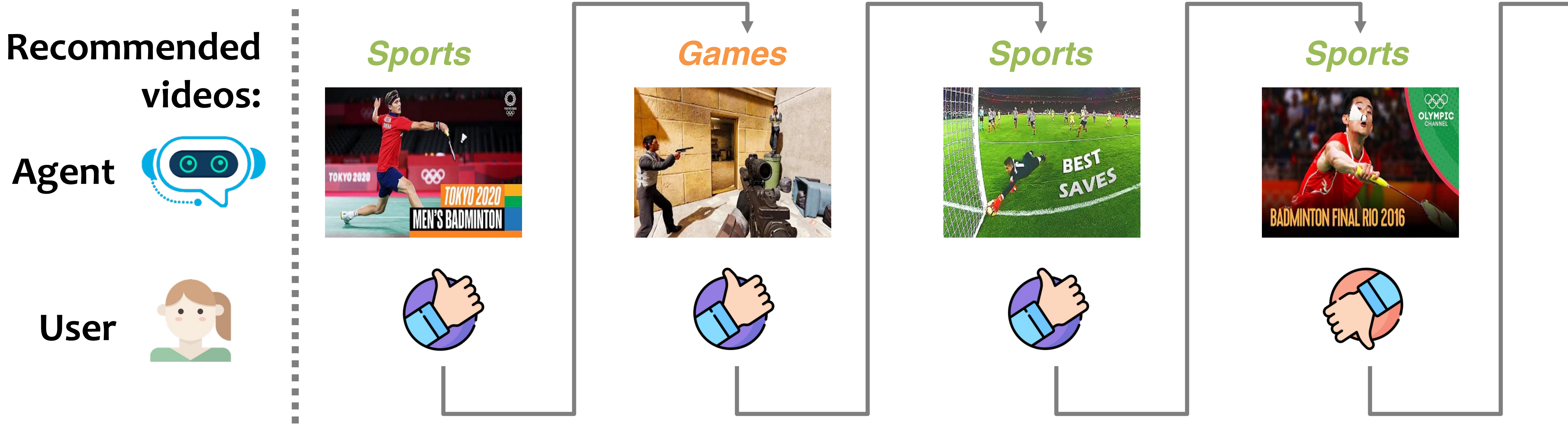
https://chongminggao.me | chongming.gao@gmail.com

❖ Outline

1. **Background and Motivation.**
    - Reinforcement Learning for Recommendation
    - Conservatism of Offline RL Induces Matthew Effect
    - Empirical Study of Matthew Effect.
2. **Proposed Method: DORL**
3. **Experiments**

1.1 Reinforcement Learning for Recommendation

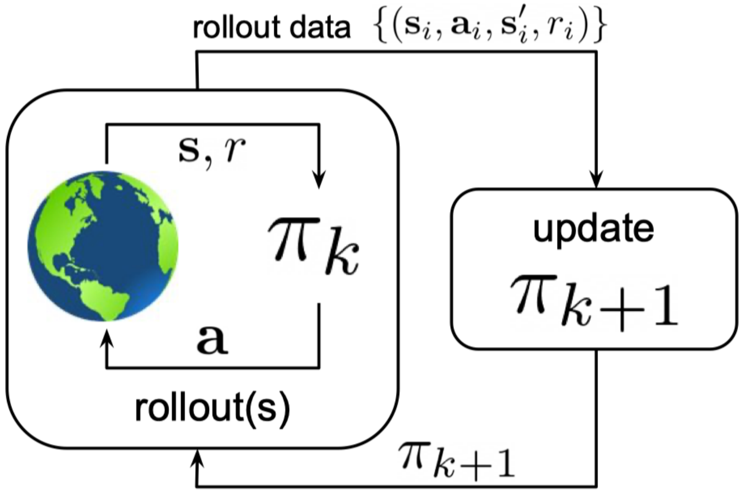
- Interactive recommendation is **the general form** of real-world recommenders *(static recommender is only a special/simplified case of IRS).*



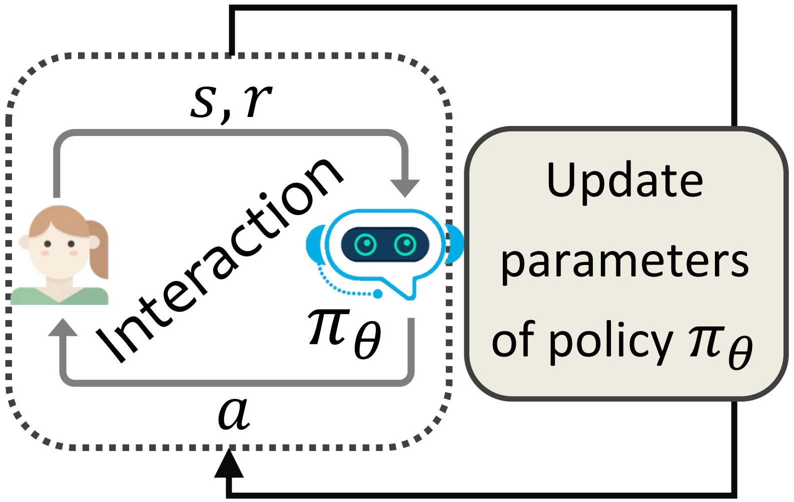
An interaction trajectory in Kuaishou, a video viewing App

1.1 Reinforcement Learning for Recommendation

- Since user satisfaction and long-term user engagement in the interactive recommendation have no *standard* answers. It is appropriate to use reinforcement learning (RL) to optimize the long-term gain.



(a) Flowchart of RL



(b) RL-based interactive recommendation

- Objective:

$$J(\pi) = \mathbb{E}_{\tau \sim p_\pi(\tau)} \left[\sum_{t=0}^{H} \gamma^t r(\mathbf{s}_t, \mathbf{a}_t) \right]$$

$$V^\pi(\mathbf{s}_t) = \mathbb{E}_{\tau \sim p_\pi(\tau|\mathbf{s}_t)} \left[\sum_{t'=t}^{H} \gamma^{t'-t} r(\mathbf{s}_{t'}, \mathbf{a}_{t'}) \right]$$

$$Q^\pi(\mathbf{s}_t, \mathbf{a}_t) = \mathbb{E}_{\tau \sim p_\pi(\tau|\mathbf{s}_t, \mathbf{a}_t)} \left[\sum_{t'=t}^{H} \gamma^{t'-t} r(\mathbf{s}_{t'}, \mathbf{a}_{t'}) \right]$$

1.1 Reinforcement Learning for Recommendation

- **Problem:** We have to use offline data to learn the policy π_θ.
- **Solution:** Use a **user model** ϕ_M in the model-based **offline reinforcement learning** (offline RL) to estimate user preferences.

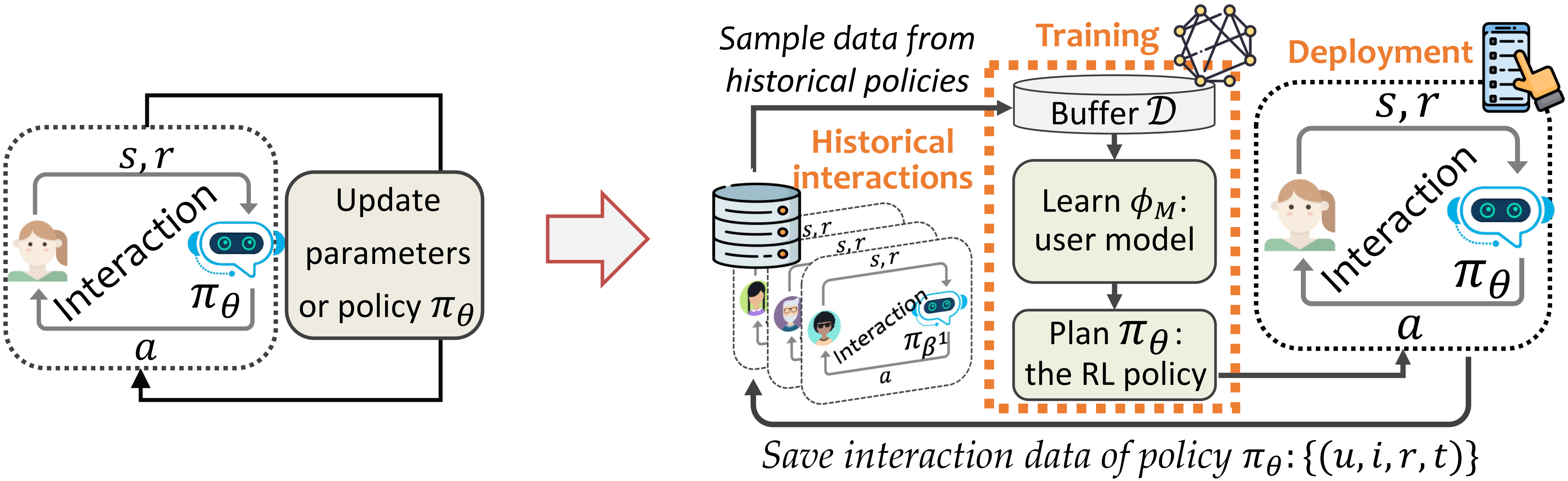


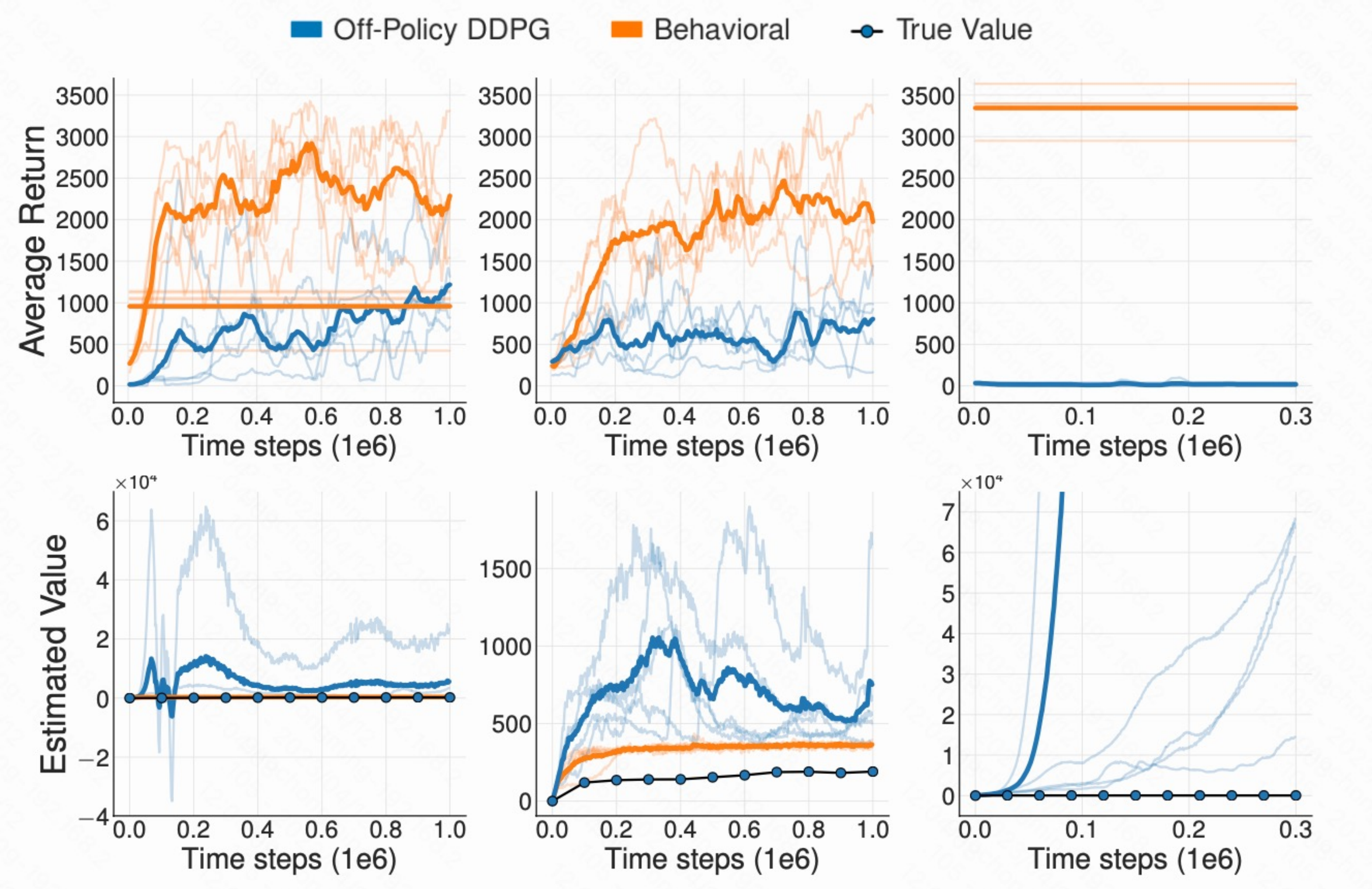
Figure: From online RL to offline model-based RL

Chongming Gao et al. CIRS: Bursting Filter Bubbles by Counterfactual Interactive Recommender System. TOIS 2023.

1.2 Conservatism of Offline RL Induces Matthew Effect

Offline RL (batched RL)

- **Challenge:** distributional shift – overestimation of the value function.



(d) Final buffer (e) Concurrent (f) Imitation

➢ Performance dropped

(How well it does?)

➢ Value overestimated

(How well it thinks it does?)

Scott Fujimoto et al. Off-Policy Deep Reinforcement Learning without Exploration. ICML '19.

1.2 Conservatism of Offline RL Induces Matthew Effect

Offline RL (batched RL)

- **Challenge:** distributional shift – overestimation of the value function.
- **Solution:** Conservatism or Pessimism

$$\begin{cases} \pi_{k+1} = \operatorname{argmax}_\pi \langle \pi, Q_k \rangle \\ Q_{k+1} = r - b + \gamma P \langle \pi_{k+1}, Q_k \rangle \end{cases} \Leftrightarrow \begin{cases} \pi_{k+1} = \operatorname{argmax}_\pi \langle \pi, Q'_k - b \rangle \\ Q'_{k+1} = r + \gamma P \langle \pi_{k+1}, Q'_k - b \rangle \end{cases}$$

Shideh Rezaeifar et al. Offline Reinforcement Learning as Anti-Exploration. AAAI 22.

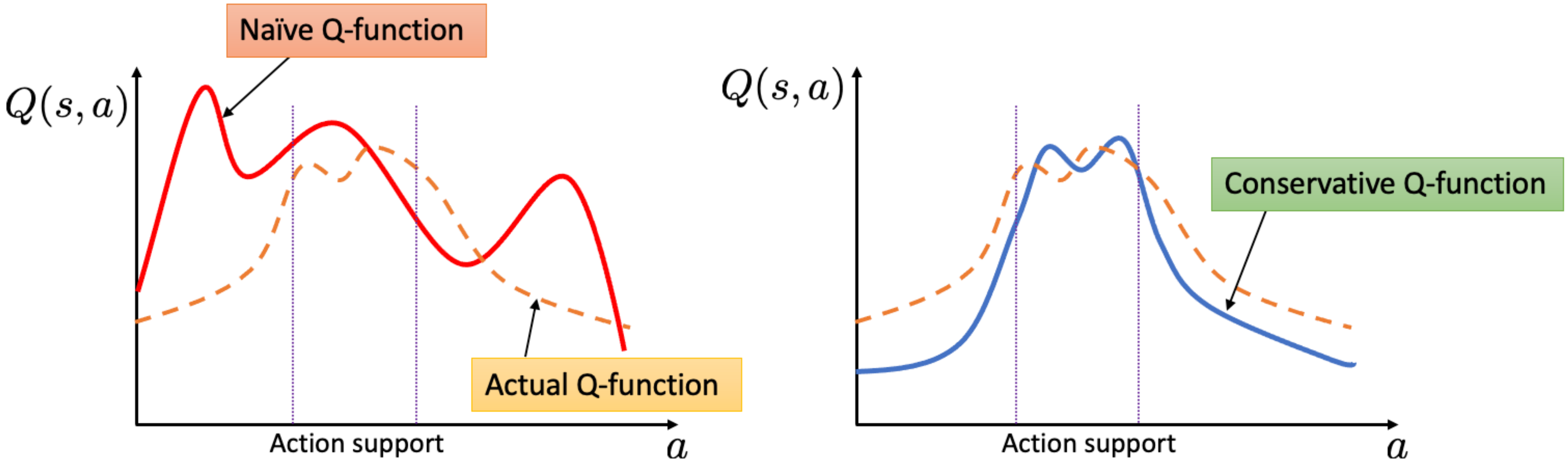


Figure from Aviral Kumar' blog: https://bair.berkeley.edu/blog/2020/12/07/offline/

1.2 Conservatism of Offline RL Induces Matthew Effect

Offline RL (batched RL)

- **Challenge:** distributional shift – overestimation of the value function.
- **Solution:** Conservatism or Pessimism
- Problem **in Recommendation:** Matthew Effect:

“The rich gets richer and the poor gets poorer.”

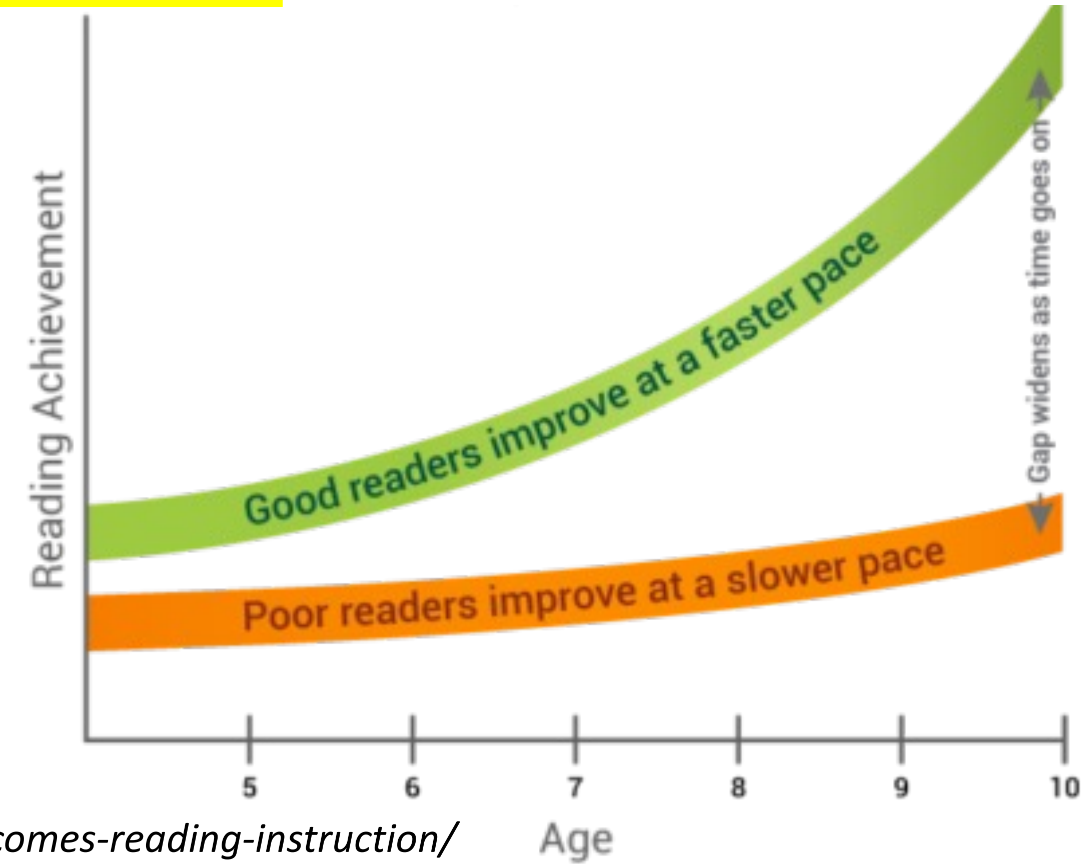
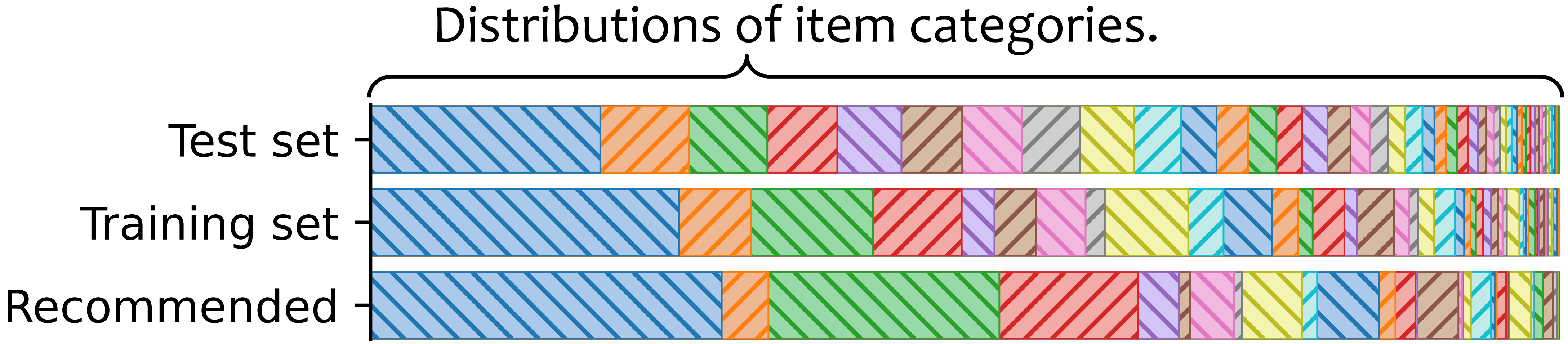


Figure from: https://www.phonicbooks.co.uk/2017/06/04/matthew-effect-comes-reading-instruction/

1.3 Empirical Study of Matthew Effect

❑ **Visualization:**

- A DeepFM model train on the KuaiRand-Pure dataset



❑ **Observation: Matthew effect occurs in Recommendation!**

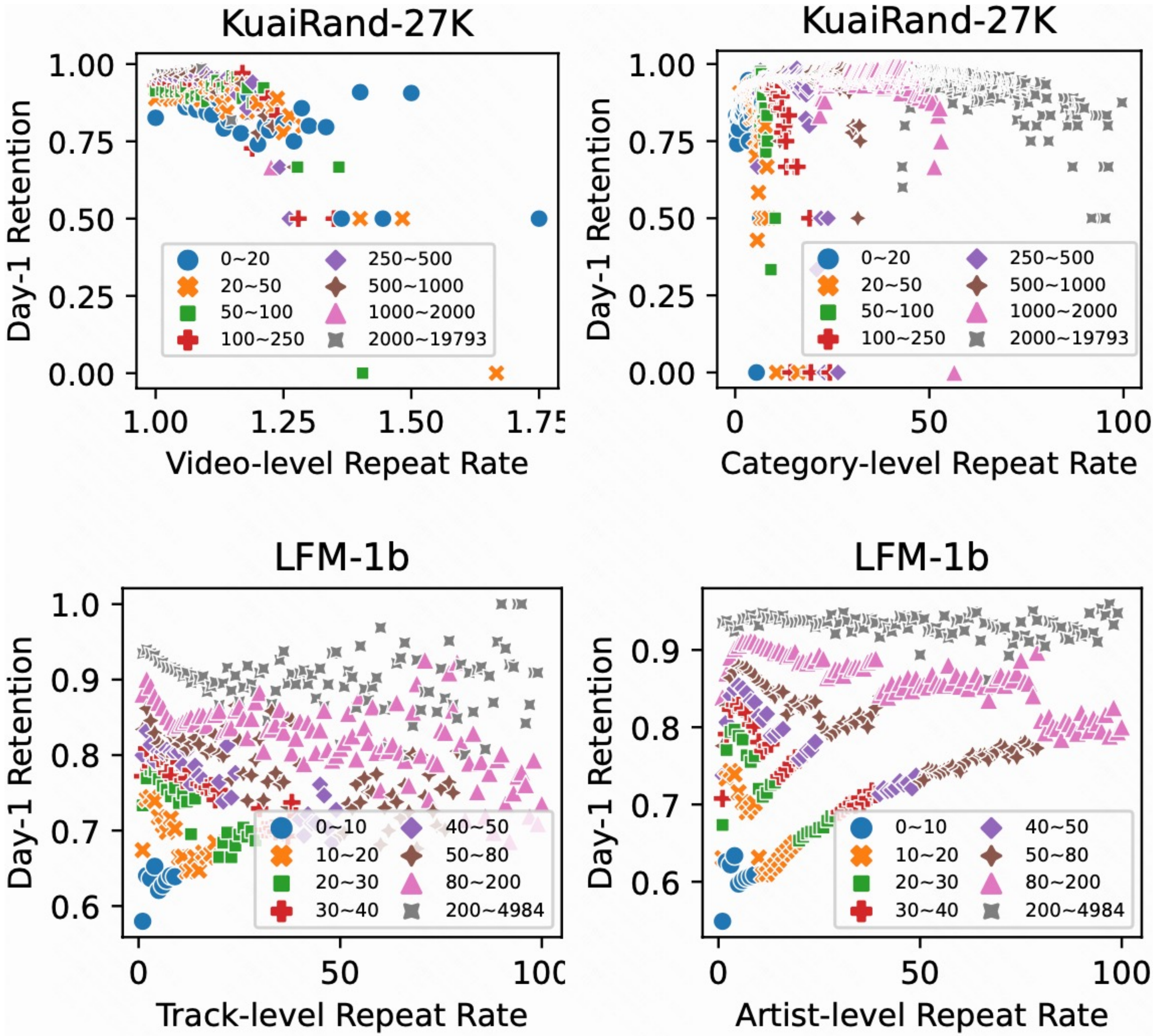
1.3 Empirical Study of Matthew Effect

❑ **Visualization:**

- The relationship between Day-1 Retention and item- level and category-level repeat rates in two datasets.
- The item-level (or category-level) repeat rate of a user viewing videos on a certain day is defined as:

$$\frac{\text{the number of viewing events}}{\text{the number of unique videos (or unique categories)}}.$$

❑ **Observation: Matthew effect hurts user experience!**



❖ Outline

1. Background and Motivation.
2. Proposed Method: DORL
    - State-of-the-art Offline RL Framework: MOPO
    - Problem: Accelerating Matthew Effect
    - Alleviate Matthew Effect by Penalizing Entropy
    - Framework of DORL
3. Experiments

3.1 State-of-the-art Offline RL Framework: MOPO

❑ Conservatism of MOPO: Penalizing uncertainty.

- A state-of-the-art general Model-based Offline Policy Optimization framework (MOPO) introduce a penalty function $p(s, a)$ on the estimated reward $\hat{r}(s, a)$ as:

$$\tilde{r}(s, a) = \hat{r}(s, a) - \lambda p(s, a)$$

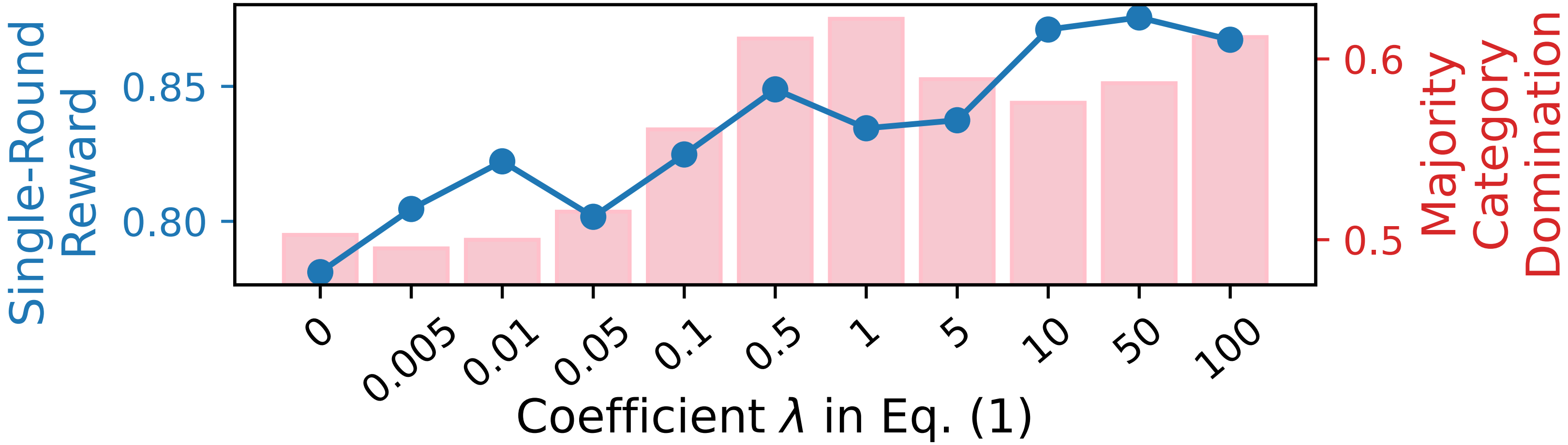
- The implementation of $p(s, a)$ in MOPO is the uncertainty P_U of the dynamics model, i.e., $p(s, a) = P_U$.

Tianhe Yu et al. MOPO: Model-based offline policy optimization. NeurIPS 2020.

3.2 Problem: Accelerating Matthew Effect

❑ **Problem** of Conservatism of MOPO in Recommendation:

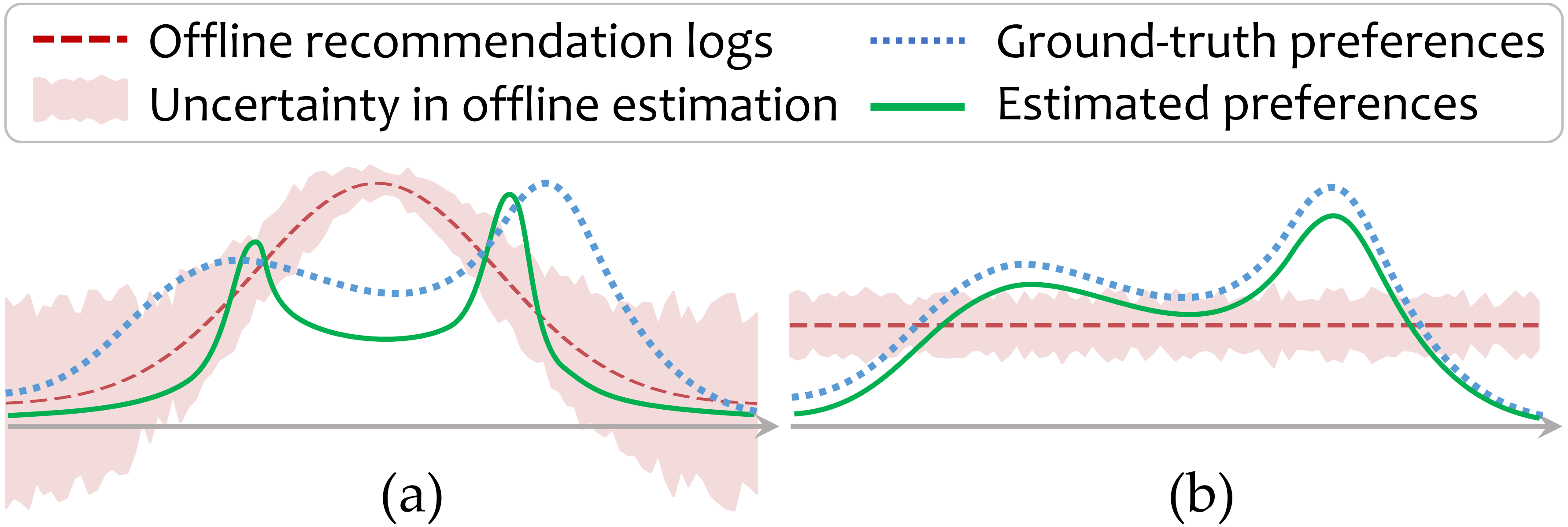
$$\tilde{r}(s,a) = \hat{r}(s,a) - \lambda p(s,a) \qquad (1)$$



- The single-round reward (the blue line) improves as expected, due to that the model pays more attention to items in the training distribution.
- The Majority Category Domination (the red bars) increases, which indicates a severer Matthew effect.

3.3 Alleviate Matthew Effect by Penalizing Entropy

❑ An intuitive example:



- **(a) Gaussian recommendation policy:** Since previous policy (Red line) cannot precisely reflect users' ground-truth preferences (blue line), the learned preferences (green line) are prone to be biased towards popular items (Matthew effect).
- **(b) Uniform recommendation policy** Under a random policy (Red line) the policy can capture ground-truth user preferences (blue line) and thus can produce an unbiased estimated preferences (green line).

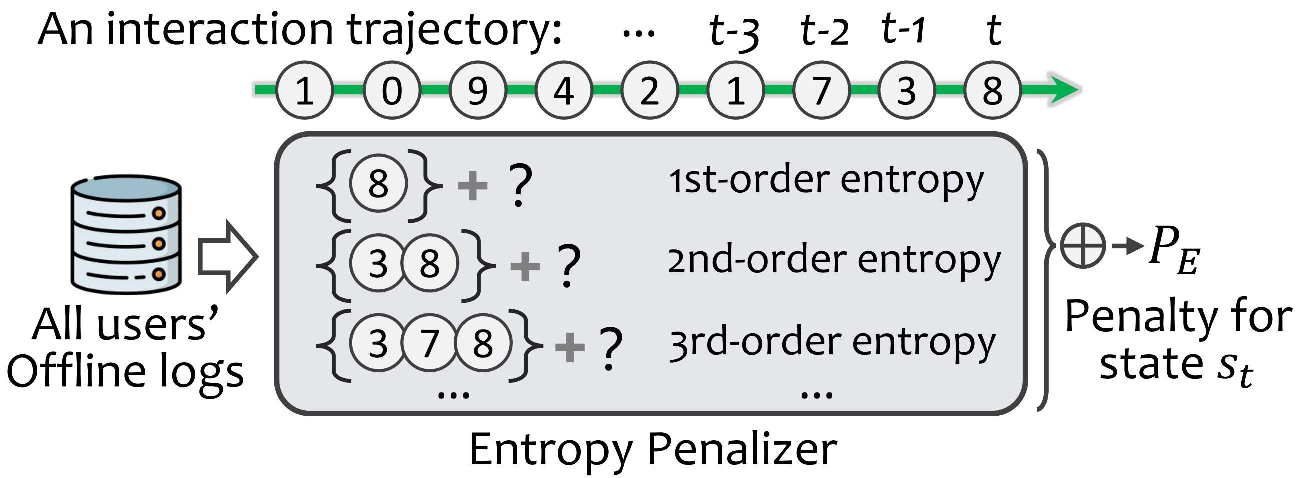
3.3 Alleviate Matthew Effect by Penalizing Entropy

❑ Our Solution: Penalizing entropy of the behavior policy

- We add a term P_E in modified reward to penalize actions that lead to less diverse states.

$$\tilde{r}(s,a) = \hat{r}(s,a) \underbrace{- \lambda_1 P_U}_{\text{Penalty on Uncertainty}} \underbrace{- \lambda_2 P_E}_{\text{Penalty on Entropy}}$$

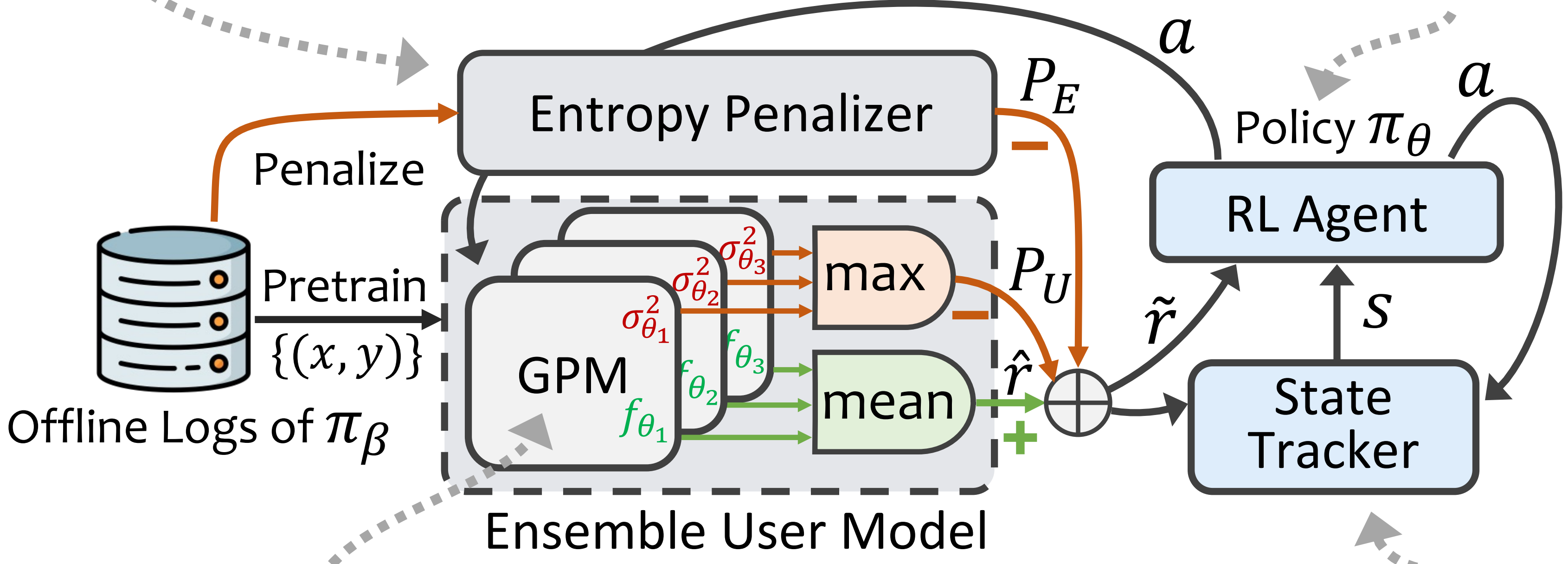
- We define P_E to be the summation of k-order entropy of the behavior policy:



3.4 Framework of DORL

❑ Our Solution: Penalizing entropy of the behavior policy

- **Entropy Penalizer**: introduced in the last page.
- **Policy** π_θ: implemented as an actor-critic algorithm.



- **Objective** of the k-th Gaussian probabilistic model (**GPM**)

$$\mathcal{L}(\theta_k) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2\sigma^2_{\theta_k}(x_i)} \| y_i - f_{\theta_k}(x_i) \|^2 + \frac{1}{2}\log \sigma^2_{\theta_k}(x_i)$$

- **State tracker**: a naive average layer

$$\vec{s}_{t+1} := \frac{1}{N}\sum_{n=t-N+1}^{t} [\vec{e}_{a_n} \oplus \tilde{r}_n]$$

❖ Outline

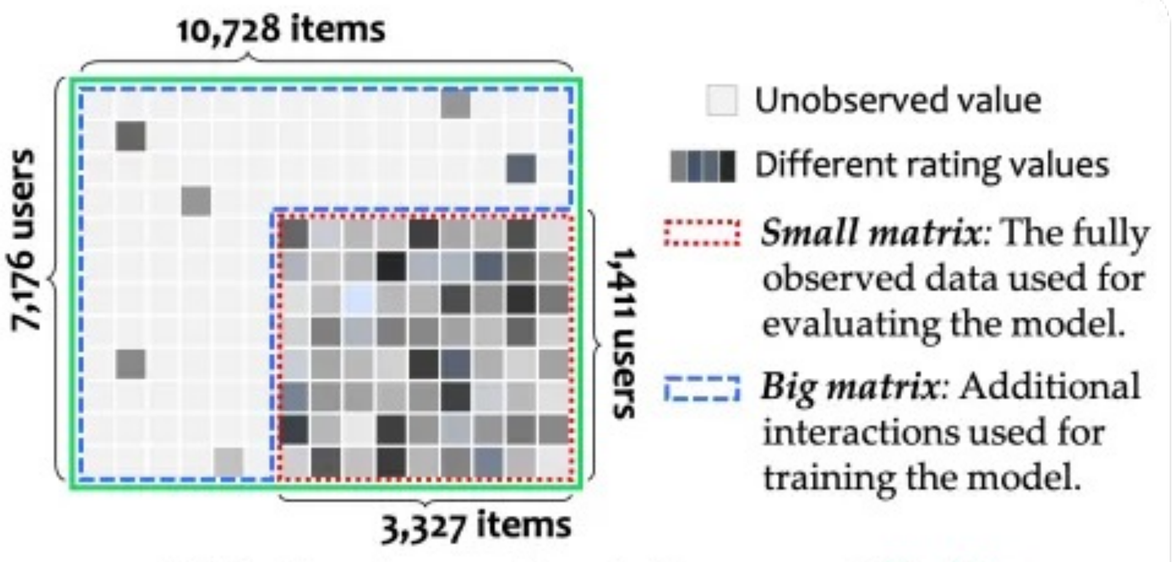
1. Background and Motivation.
2. Proposed Method: DORL
3. Experiments
    - Experimental Setup
    - Performance Comparison
    - More Analysis

4.1 Experimental Setup

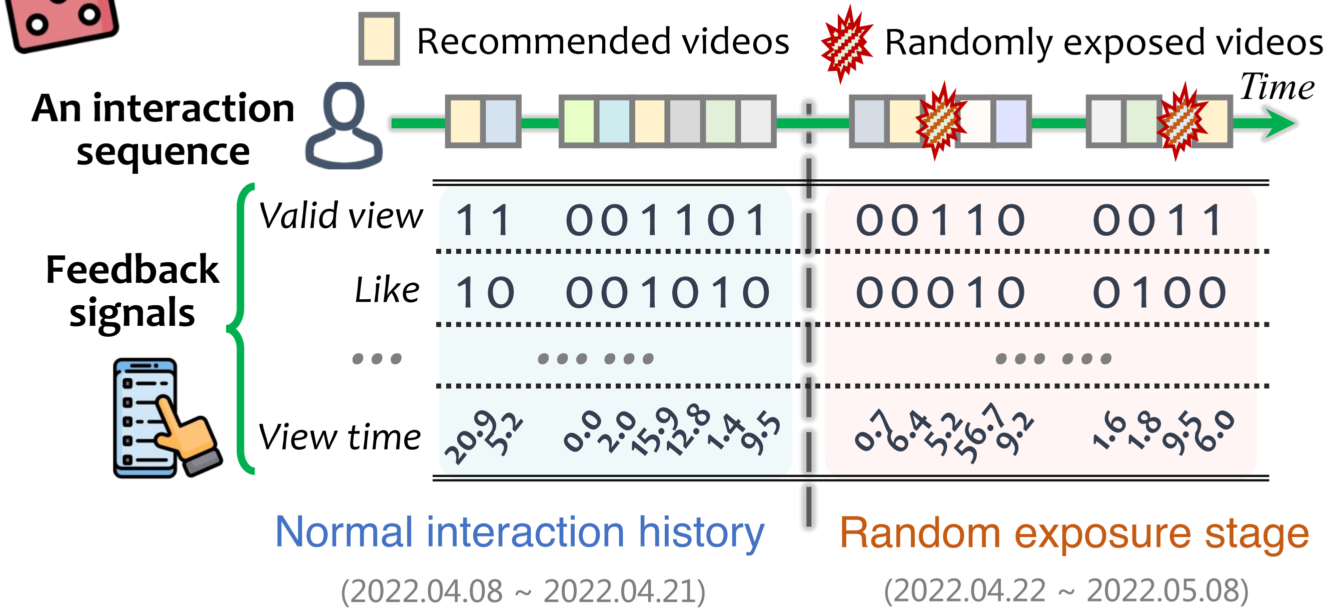
❑ Datasets

Datasets	Usage	#Users	#Items	#Interactions	#Categories
KuaiRec	Train	7,176	10,728	12,530,806	31
	Test	1,411	3,327	4,676,570	31
KuaiRand	Train	27,285	7,551	1,436,609	46
	Test	27,285	7,583	1,186,059	46

KuaiRec *(https://kuairec.com)*



KuaiRand *(https://kuairand.com)*



4.1 Experimental Setup

Exit mechanism:

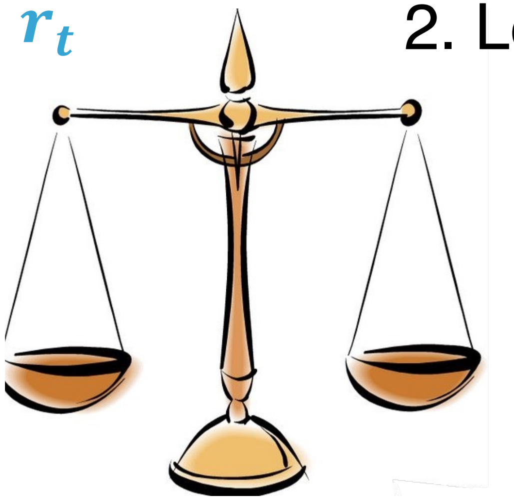
- **"Feel bored then quit"**: If the recommended item is similar to previous recommended ones, the interactive process terminates.

Evaluation metric:

Accumulated reward $= \sum_{t=1}^{T} \boldsymbol{r_t}$, *which requires:*

1. High single-round reward $\boldsymbol{r_t}$

2. Long trajectory length T

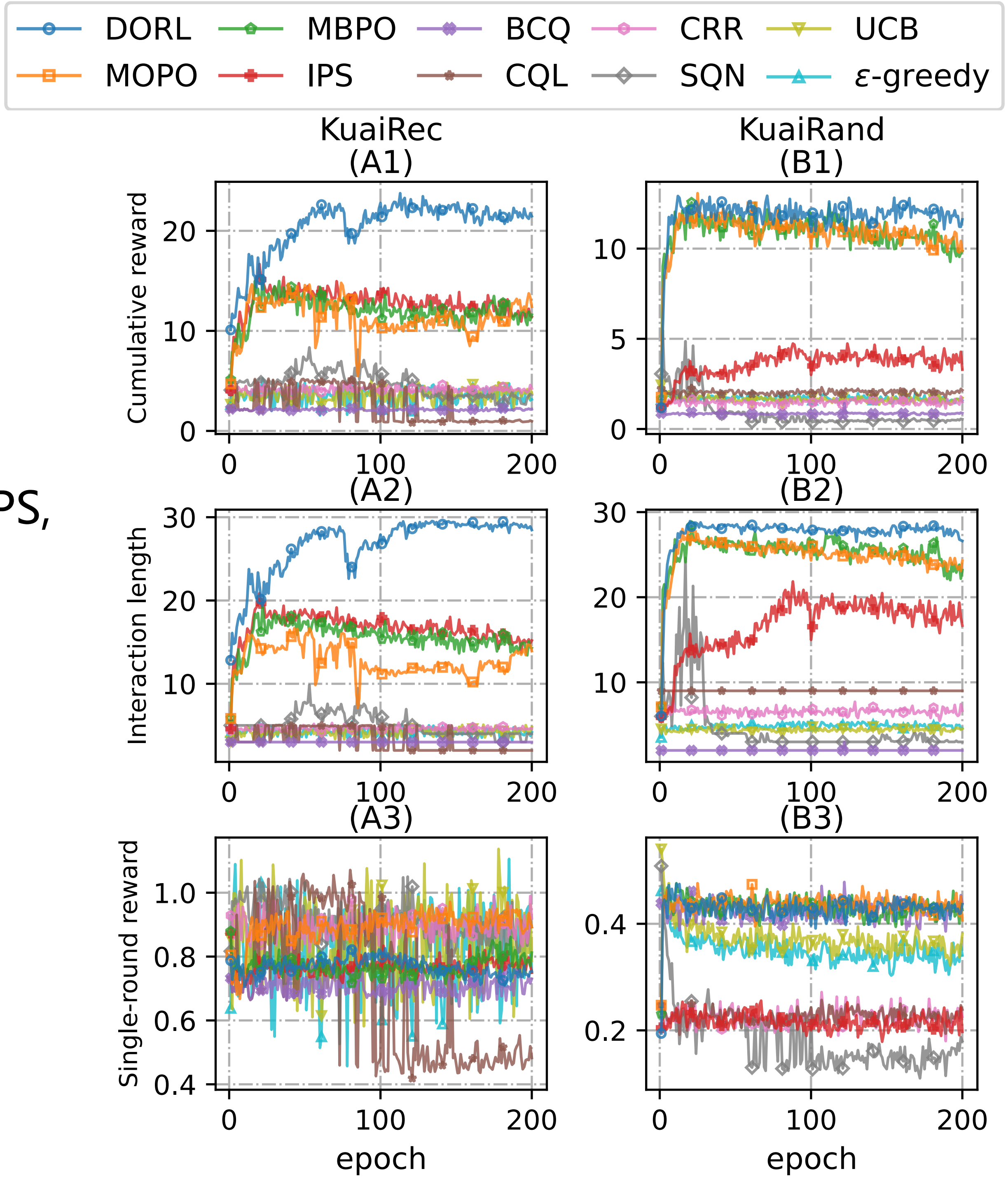


Find a balance

4.2 Results

Overall performance

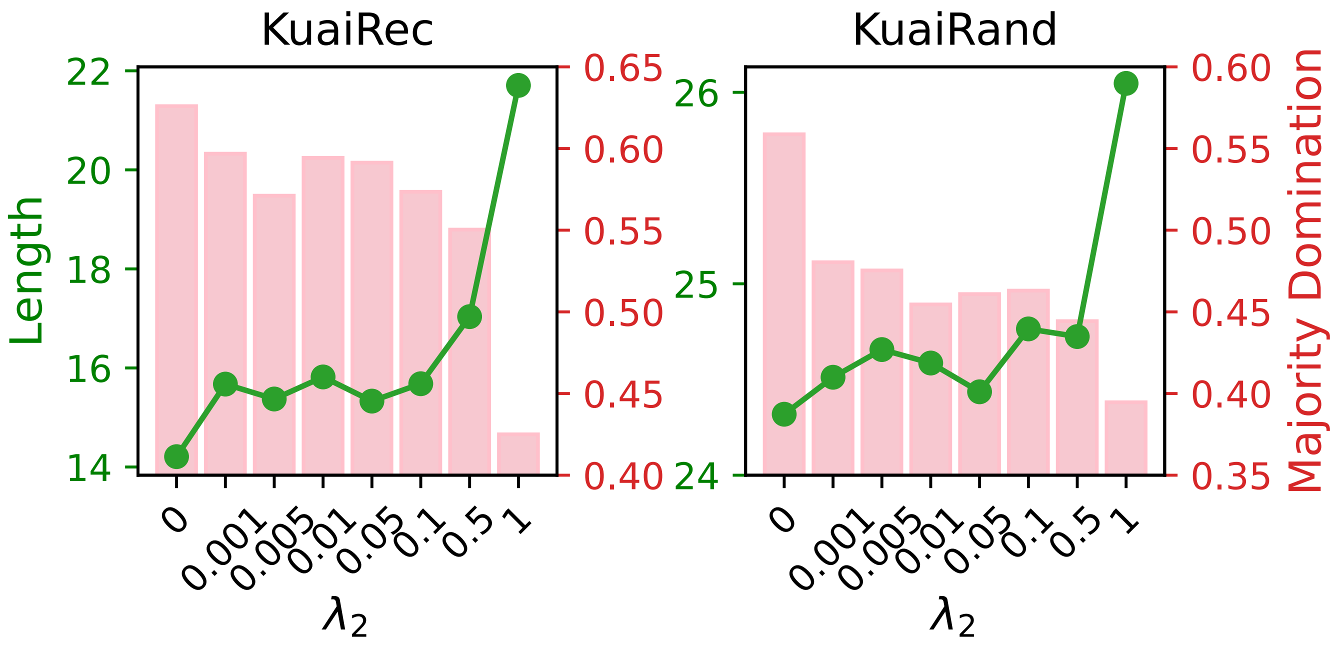
1. Our proposed DORL (blue line) shows the best performance in terms of the cumulative reward and the interaction length.
2. The four model-based RL methods (MBPO, IPS, MOPO, and DORL) significantly outperform the four model-free RL methods (SQN, CRR, CQL, and BCQ) with respect to trajectory length and cumulative rewards.
3. Single round reward: MOPO > MBPO. But MOPO neglects less popular items, which makes trajectory length: MOPO < MBPO. DORL remedies the situation!



4.3 Results on Alleviating Matthew Effect

☐ Varying the strength of penalty through different λ_2

$$\tilde{r}(s,a) = \hat{r}(s,a) - \lambda_1 P_U - \lambda_2 P_E$$

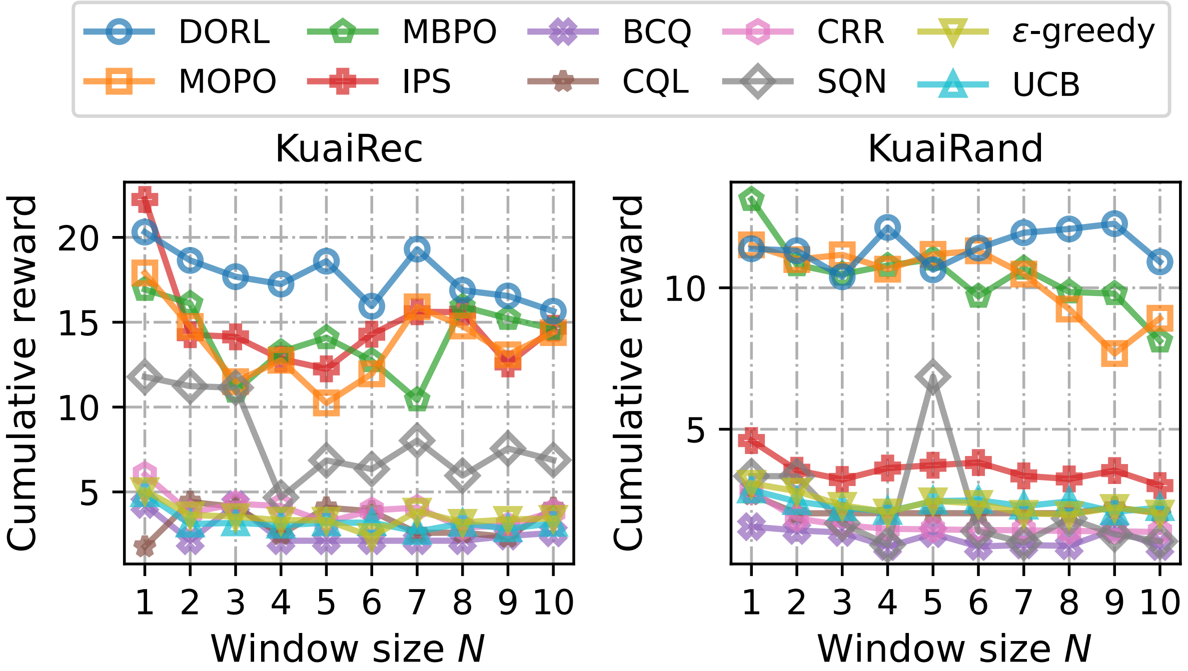


- Increasing λ_2 can diversify the outputs, which extends the interaction length (green lines) and results in **high user satisfaction**.
- Increasing λ_2 reduces Majority Category Domination (the red bars), **alleviating the Matthew effect**.

4.4 Results under Different Ending Conditions

☐ Varying User Sensitivity:

- The interaction will end if there are one similar item in previous *N* results.



(As the value of *N* increases, the user's tolerance for similar content decreases)

- DORL outperforms all other policies, which demonstrates the **robustness** of DORL in different environments.

❖ Conclusion and Discussion

☐ Contributions:

- We point out that conservatism in offline RL can incur the Matthew effect in recommendation. We show this phenomenon in existing methods and how it hurts user satisfaction.
- We propose the DORL model that introduces a counterfactual exploration in offline data.
- We demonstrate the effectiveness of DORL in an interactive recommendation setting, where alleviating the Matthew effect increases users' long-term experience.

☐ Future work:

- To develop recommender systems as decision makers rather than preference fitters.
- When fitting user interests is not a bottleneck anymore, researchers could consider higher-level goals, such as pursuing users' long-term satisfaction or optimizing social utility.

中国科学技术大学
University of Science and Technology of China

快手
KUAISHOU TECHNOLOGY

Thanks

Chongming Gao | 高崇铭
chongming.gao@gmail.com