

# Contrastive Learning for Debiased Candidate Generation in Large-Scale Recommender Systems

Chang Zhou\*

DAMO Academy, Alibaba Group  
ericzhou.zc@alibaba-inc.com

Jianxin Ma\*†

DAMO Academy, Alibaba Group  
jason.mjx@alibaba-inc.com

Jianwei Zhang\*

DAMO Academy, Alibaba Group  
zhangjianwei.zjw@alibaba-inc.com

Jingren Zhou

DAMO Academy, Alibaba Group  
jingren.zhou@alibaba-inc.com

Hongxia Yang

DAMO Academy, Alibaba Group  
yang.yhx@alibaba-inc.com

## ABSTRACT

Deep candidate generation (DCG) that narrows down the collection of relevant items from billions to hundreds via representation learning has become prevalent in industrial recommender systems. Standard approaches approximate maximum likelihood estimation (MLE) through sampling for better scalability and address the problem of DCG in a way similar to language modeling. However, live recommender systems face severe exposure bias and have a vocabulary several orders of magnitude larger than that of natural language, implying that MLE will preserve and even exacerbate the exposure bias in the long run in order to faithfully fit the observed samples. In this paper, we theoretically prove that a popular choice of contrastive loss is equivalent to reducing the exposure bias via inverse propensity weighting, which provides a new perspective for understanding the effectiveness of contrastive learning. Based on the theoretical discovery, we design *CLRec*, a Contrastive Learning method to improve DCG in terms of fairness, effectiveness and efficiency in Recommender systems with extremely large candidate size. We further improve upon CLRec and propose Multi-CLRec, for accurate multi-intention aware bias reduction. Our methods have been successfully deployed in *Taobao*, where at least four-month online A/B tests and offline analyses demonstrate its substantial improvements, including a dramatic reduction in the Matthew effect.

## CCS CONCEPTS

- Information systems → Recommender systems.

## KEYWORDS

Recommender systems; candidate generation; bias reduction; inverse propensity weighting; contrastive learning; negative sampling

\*Equal contribution.

†Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD '21, August 14–18, 2021, Virtual Event, Singapore

© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-8332-5/21/08...\$15.00

<https://doi.org/10.1145/3447548.3467102>

## ACM Reference Format:

Chang Zhou, Jianxin Ma, Jianwei Zhang, Jingren Zhou, and Hongxia Yang. 2021. Contrastive Learning for Debiased Candidate Generation in Large-Scale Recommender Systems . In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21), August 14–18, 2021, Virtual Event, Singapore*. ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/3447548.3467102>

## 1 INTRODUCTION

Large-scale industrial recommender systems usually adopt a multi-stage pipeline, where the first stage, namely candidate generation, is responsible for retrieving a few hundred relevant entities from a billion scale corpus. Deep candidate generation (DCG) [14], a paradigm that learns vector representations of the entities to enable fast k-nearest neighbor retrieval [28], has become an essential part of many live industrial systems with its enhanced expressiveness.

Typical large-scale DCG models [14, 17, 32, 61] regard the problem of identifying the most relevant items to the users as estimating a multinomial distribution traversing over all items for each user, conditional on the user's past behaviors. Maximum likelihood estimation (MLE) is the conventional principle for training such models. Apparently, exact computation of the log likelihood, which requires computing softmax looping over a million- or even billion-scale collection of items, is computationally infeasible. Among the various sampling-based approximation strategies, sampled-softmax [6, 26] usually outperforms the binary-cross-entropy based approximations such as NCE [18] and negative sampling [38] at large scale.

However, the MLE paradigm and the sampling strategies mainly stem from the language modeling community, where the primary goal is to faithfully fit the observed texts. Indeed, live recommender systems are different from natural language texts in several aspects. Firstly, the training data are collected from current undergoing systems that might be sub-optimal and biased towards popular items. In such situations, some high-quality items can be under-explored in the training data, while an algorithm trained via MLE will continue to under-estimate the relevance of the under-explored items in order to faithfully fit the observed data. Secondly, the set of items for recommendation is much larger than the vocabulary of natural languages, e.g.  $\approx 100M$  items in our system compared to  $\approx 100k$  words, posing great difficulties in learning a set of fairly good representations for the entire candidate set.

In this paper, we introduce *CLRec*, a Contrastive Learning framework for debiased DCG in RECommender systems. Contrastive

learning, which constructs self-supervised tasks to improve the discriminative ability of a model, has become increasingly popular recently in the pre-training community [11, 19]. We discover that the contrastive learning paradigm has a debiasing effect that can benefit recommender systems. Specifically, our central contribution is establishing the theoretical connection between contrastive learning and inverse propensity weighting, where the latter is a well-studied technique for bias reduction [25, 41, 48]. Our theory complements the previous studies of contrastive learning [39].

Based on the theoretical discovery, an easy-to-implement framework is developed to efficiently reduce the exposure bias of a large-scale system. Our implementation maintains a fixed-size first-in first-out (FIFO) queue [19] to accumulate positive samples and their representations from the previous batches, and use the content of the queue to serve as the negative samples to be discriminated from the positive samples of the next batch. It guarantees that all items will be sampled sometime in an epoch to serve as the negative examples. More importantly, it allows us to reuse the computed results from the previous batches, e.g., saving 90% computation cost when the queue size is 10× of the batch size. As a result, we can afford to encode complex features of the negative samples even when the negative sample size, i.e., the queue size, is very large, where the rich features can improve the quality of the learned representations of the under-explored items. The queue-based framework also offers simple yet efficient implementations for complex self-supervised tasks, e.g., discriminating whether two subsequences come from the same user’s behaviors [36]. We further improve upon CLRec and propose Multi-CLRec, which employs a multi-queue design for more accurate user-intention aware bias reduction.

Our methods have been fully deployed into our live system as the default choice to serve billions of page views each day. We observe that they are capable of recommending high-quality items that are largely neglected by most current ongoing systems, and consistently outperform the previous state-of-art baselines.

## 2 OUR THEORECTICAL RESULTS ON CONTRASTIVE LEARNING IN DCG

In this section, we will reveal the connection between contrastive learning and inverse propensity weighting (IPW).

### 2.1 Problem Formulation

*Notations.* Given a dataset of user clicks  $\mathcal{D} = \{(x_{u,t}, y_{u,t}) : u = 1, 2, \dots, N, t = 1, 2, \dots, T_u\}$ , where  $x_{u,t} = \{y_{u,1:(t-1)}\}$  represents a user’s clicks prior to the  $t$ -th click  $y_{u,t}$ , and  $T_u$  denotes the number of clicks from the user  $u$ . We will drop the sub-scripts occasionally and write  $(x, y)$  in place of  $(x_{u,t}, y_{u,t})$  for conciseness. We use  $X$  to refer to the set of all possible click sequences, i.e.  $x \in X$ . Each  $y \in \mathcal{Y}$  represents a clicked item, which includes various types of features associated with the item, while  $\mathcal{Y}$  is the set of all possible items. The features of  $y$  could be in any form, e.g., the item’s unique identifier number, embeddings or raw data of its image and text description. The number of items  $|\mathcal{Y}|$  can easily reach 100 million.

*Deep Candidate Generation.* The deep candidate generation paradigm involves learning a user behavior encoder  $f_\theta(x) \in \mathbb{R}^d$  and an item encoder  $g_\theta(y) \in \mathbb{R}^d$ . The set of parameters used by each

encoder is a subset of  $\theta$ , i.e. the set of all trainable parameters in the system. It then takes  $\{g_\theta(y)\}_{y \in \mathcal{Y}}$  and builds a k-nearest-neighbor search service, e.g., Faiss [28]. As a result, given an arbitrary user behavior sequence  $x$  at serving time, we can instantly retrieve the top  $k$  items relevant to the user by finding the top  $k$  candidate  $g_\theta(y)$  similar to  $f_\theta(x)$ . Most implementations use inner product  $\phi_\theta(x, y) = \langle f_\theta(x), g_\theta(y) \rangle$  or cosine similarity as the similarity score. The typical learning procedure fits the data following the maximum likelihood estimation (MLE) principle:

$$\arg \min_{\theta} \frac{1}{|\mathcal{D}|} \sum_{(x,y) \in \mathcal{D}} -\log p_\theta(y | x), \quad \text{where} \\ p_\theta(y | x) = \frac{\exp \phi_\theta(x, y)}{\sum_{y' \in \mathcal{Y}} \exp \phi_\theta(x, y')} \quad (1)$$

The denominator of  $p_\theta(y | x)$  sums over all possible items, which is infeasible in practice and thus requires approximation, e.g., via sampling. However, the observed clicks for training are from the previous version of the recommender system. The training data thus suffer from exposure bias (i.e. missing not at random [42]) and reflect the users’ preference regarding the recommended items rather than all potential items. High-quality items that have few clicks in the training data will likely remain under-recommended by a new algorithm trained via the MLE paradigm.

### 2.2 Understanding Contrastive Learning from a Bias-Reduction Perspective

We now introduce the family of contrastive losses that we are interested in, and reveal their connection with the inverse propensity weighting (IPW) [25, 41, 48] techniques for bias reduction.

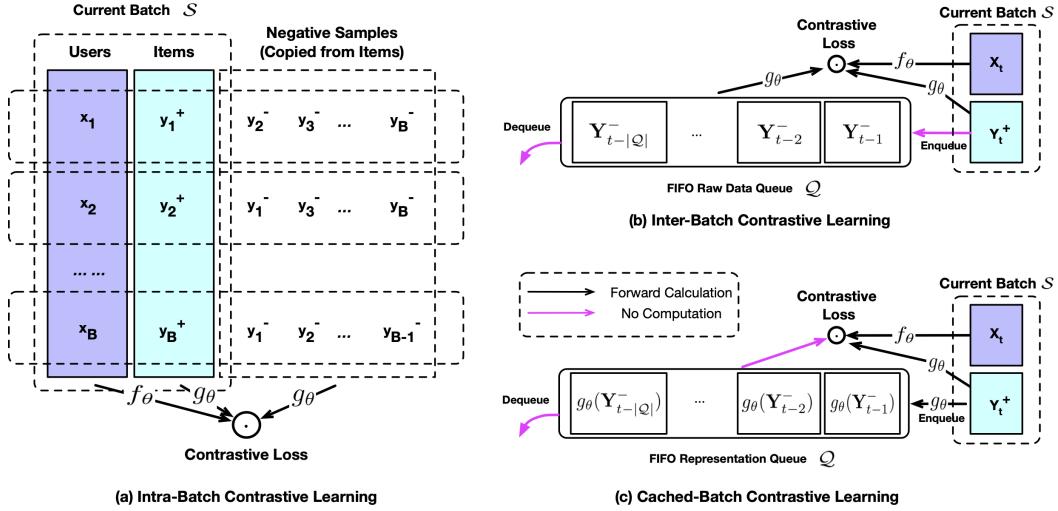
*Sampled Softmax.* The kind of contrastive loss we will investigate is similar to sampled softmax. We thus recap sampled softmax here and will show that the minor difference is crucial. There are many variants of sampeld softmax [6, 26], among which the following variant is integrated by TensorFlow [1] and commonly used:

$$\arg \min_{\theta} \frac{1}{|\mathcal{D}|} \sum_{(x,y) \in \mathcal{D}} -\log \left\{ \frac{\exp (\phi_\theta(x, y) - \log p_n(y|x))}{\exp (\phi_\theta(x, y) - \log p_n(y|x)) + \sum_{i=1}^L \exp (\phi_\theta(x, y_i) - \log p_n(y_i|x))} \right\}, \quad (2)$$

where  $\{y_i\}_{i=1}^L$  are  $L$  negative samples drawn from a pre-defined proposal distribution  $p_n(y | x)$ . Subtracting  $\log p_n(y | x)$  is necessary for it to converge to the same solution as the exact loss in Eq. (1). Most implementations assume  $p_n(y | x) = p_n(y)$  and set  $p_n(y)$  somehow proportional to the popularity of the items to improve convergence. In practice, we would draw thousands of negative samples to pair with each positive example. Sampled softmax in general outperforms other approximations such as NCE [18] and negative sampling [38] when the vocabulary is large [14, 26, 33, 37].

*Contrastive Loss.* We study the following type of contrastive loss [39, 43] under a negative sampler  $p_n(y | x)$ :

$$\arg \min_{\theta} \frac{1}{|\mathcal{D}|} \sum_{(x,y) \in \mathcal{D}} -\log \frac{\exp (\phi_\theta(x, y))}{\exp (\phi_\theta(x, y)) + \sum_{i=1}^L \exp (\phi_\theta(x, y_i))}, \quad (3)$$



**Figure 1: Three variants of the single-queue CLRec , whose implicit proposal distributions are  $p_n(y | x) = p_{\text{data}}(y)$ . The superscripts +, - mean positive and negative examples respectively. Variant (a) uses the positive examples of other instances in the present batch as the negative examples. Variant (b) creates a fixed-size FIFO queue to store the positive examples encountered in previously processed batches, and use the examples in the queue to serve as the negative examples for the present batch. Variant (c) differs from variant (b) in that the queue caches computed representations  $g_\theta(y)$  rather than raw features of  $y$ .**

where  $\{y_i\}_{i=1}^L$  are again sampled from  $p_n(y | x)$  for each  $x$  to pair with the positive sample. It no longer optimizes the MLE loss in Eq. (1), because it misses  $-\log p_n(y | x)$  and thus does not correct the bias introduced by sampling. Many efforts on contrastive learning have been focusing on designing a well-performing proposal distribution  $p_n(y | x)$  [9, 43]. InfoNCE [39] demonstrates that this loss maximizes a lower bound of the mutual information between  $x$  and  $y$  if we set the proposal distribution  $p_n(y | x)$  to be the actual data distribution  $p_{\text{data}}(y)$ , i.e. if we sample  $y$  proportional to its frequency in the dataset.

*Contrastive Learning and Exposure Bias Reduction.* The contrastive loss as shown above in Eq. (3) has recently achieved remarkable success in various fields [11, 19, 39]. Nevertheless, it still remains a question why the loss is effective. We will reveal that the contrastive loss is a sampling-based approximation of an inverse propensity weighted (IPW) loss. The IPW loss<sup>1</sup> has the form:

$$\arg \min_{\theta} \frac{1}{|\mathcal{D}|} \sum_{(x, y) \in \mathcal{D}} -\frac{1}{q(y | x)} \cdot \log p_\theta(y | x), \quad (4)$$

where  $q(y | x)$  should be the propensity score function, which represents the *impression (or exposure) distribution*, i.e., the probability that item  $y$  is recommended to user  $x$  in the previous system where we collect the training data  $\mathcal{D}$ . The idea of IPW is to model missing-not-at-random via the propensity scores in order to correct the exposure bias. We can prove that the IPW loss is optimizing  $p_\theta(y | x)$  to capture the oracle user preference even when there

<sup>1</sup>Our IPW loss is different from the previous works on debiased recommenders [42, 52]. We focus on *multinomial* propensities, i.e. whether an item is selected and recommended by a recommender out of all possible items. The previous works consider *bernoulli* propensities related with users' attention, i.e. whether a user notices a recommended item or not, and mostly deal with the position bias in ranking. We give the derivation of this multinomial IPW loss in the Appendix.

exists exposure bias. A standard implementation of the IPW loss has two steps, where the first step is to use a separate model to serve as  $q(y | x)$  and optimize it by fitting the exposure history according to the MLE principle, while the second step is to optimize  $p_\theta(y | x)$  according to Eq. (4). However, the two-stage pipeline of IPW, as well as the numerical instability and variance [42] brought by  $\frac{1}{q(y|x)}$ , makes IPW less efficient for large-scale production systems.

Fortunately, we can prove that the contrastive loss Eq. (3) is in principle optimizing the same objective as Eq. (4). And in Subsection 3 we provide simple implementations that do not require two separate steps and can avoid the instability brought by the division of  $q(y | x)$ . Our key theoretical result is as follows:

**THEOREM 1.** *The optimal solutions of the contrastive loss (Eq. 3) and the IPW loss (Eq. 4) both minimize the KL divergence from  $p_\theta(y | x)$  to  $r(y | x) = \frac{p_{\text{data}}(y|x)/q(y|x)}{\sum_{y' \in \mathcal{Y}} p_{\text{data}}(y'|x)/q(y'|x)}$ , if  $p_n(y | x)$  is set to be  $q(y | x)$ . Here  $p_{\text{data}}(y | x)$  is the data distribution, i.e. what is the frequency of  $y$  appearing in  $\mathcal{D}$  given context  $x$ .*

**PROOF.** See the Appendix. □

**Remark 1.** The implication of Theorem 1 is that the contrastive loss (Eq. 3) can approximately reduce the exposure bias if we set the proposal distribution  $p_n(y | x)$  to be the propensity score, i.e. the probability that the old systems deliver item  $y$  to user  $x$  when we were collecting the training data  $\mathcal{D}$ .

### 3 QUEUE-BASED CONTRASTIVE LEARNING FOR BIAS REDUCTION

We propose two efficient implementations of contrastive losses for bias reduction in large-scale DCG: (1) We will first present CLRec (see Figure 1), which draws negative samples from a queue that

roughly follows the non-personalized propensity score  $q(y)$ , i.e., the probability of item  $y$  being recommended by the old system under which we collect the training data, for reducing the Matthew effect. (2) We then improve upon CLRec and propose Multi-CLRec (see Figure 2), which involves multiple disentangled queues corresponding to different user intents and performs bias reduction based on the more accurate propensity score  $q(y | \text{user } x\text{'s intent})$ .

Our approaches, unlike the IPW methods that require introducing an extra model for estimating the propensity score  $q(y | x)$ , are free from the variance brought by the extra model and do not need to introduce a costly, extra stage for training the extra model.

### 3.1 CLRec: Queue-based Contrastive Learning for Reducing the Matthew Effect

The exact propensity score  $q(y | x)$  is challenging to estimate, because industrial systems involve many complicated modules and the data are sparse. CLRec thus use  $q(y)$  in place of  $q(y | x)$ , i.e. assuming  $q(y | x) \approx q(y)$ , to ease estimation and avoid small propensities. Secondly,  $q(y)$  (i.e. the exposure probability that item  $y$  is recommended to someone) has a high correlation with  $p_{data}(y)$ , i.e. the click probability that item  $y$  is being recommended and clicked by someone, because the existing system will mainly recommend items that have a high click-through rate if the system is already highly optimized. We thus further replace the impression distribution  $q(y)$  with the click distribution  $p_{data}(y)$ , i.e. assuming  $q(y) \approx p_{data}(y)$ , to ease implementation and save computation since impression data is larger than click data by an order of magnitude.

*Drawbacks of Performing Explicit Sampling.* Our analysis above shows that, by simply drawing negative samples according to  $p_{data}(y)$ , the contrastive loss (Eq. 3) can alleviate the selection bias related with the items’ impression counts in the history. However, sampling will still incur non-negligible overheads, e.g. communication costs, in a distributed environment [45]. Sampling also cannot guarantee that every item will be sampled in an epoch. We thus adopt a queue-based design [19] that avoids explicitly performing sampling, as shown in Figure 1b and Figure 1c.

*Designs of Queue-based Contrastive Learning.* We maintain a first-in first-out (FIFO) queue  $Q$ , which has a fixed capacity and can store  $|Q|$  examples. Given a new batch to process, we first enqueue the positive examples  $y$  (or their representations  $g_\theta(y)$ ) encountered in the present batch into  $Q$ . We then use the examples stored in the queue to serve as  $\{y\} \cup \{y_i\}_{i=1}^L$  (the positive example and  $L$  negative examples) when computing the denominator of the contrastive loss (Eq. 3) for the present batch. In a distributed setting, each worker maintains its own queue locally to avoid communication costs. When the queue size  $|Q|$  is equal to the batch size, our implementation is then equivalent to sampling negative examples from the present batch [12, 21, 43] (see Figure 1a). In general, we need thousands of negative samples to achieve satisfying performance. We thus use a large queue size, but with a small batch size to save memory, e.g. batch size = 256 and queue size = 2,560.

*Queue-based Caching and its Efficiency in Encoding Complex Features.* With the implementation that stores the encoded results  $g_\theta(y)$  in the queue (see Figure 1c), we can no longer back-propagate through the negative examples from the previous batches, though

we can still back-propagate through the negative examples from the present batch. As a result, we find that the total training steps required for convergence mildly increase. However, since each training step will take much less time to compute (and less communication costs in a distributed training setting), the total running time can still be greatly reduced if the features of the negative items are expensive to encode, e.g. if the features contain raw images, texts, or even structured data such as a knowledge graph.

### 3.2 Multi-CLRec: Multiple Disentangled Queues for Intent-Aware Bias Reduction

There are tens of thousands of categories of items in our system, many of which are long-tail. As a result, approximating  $q(y | x)$  with  $q(y | \text{the category in which user } x \text{ is currently interested})$  can still incur non-negligible variance. Multi-CLRec thus aims to cluster the categories into  $H$  intentions and simultaneously learn a de-biased model with negative samples that roughly follow the smoothed propensity scores  $q(y | \text{user } x\text{'s current intention is } h)$ ,  $h = 1, \dots, H$ . In particular, Multi-CLRec uses  $H$  queues  $\{Q_h\}_{h=1}^H$  corresponding to the  $H$  intentions. We use  $H = 64$  in our system.

*Disentangled Encoding with Prototype-based Routing.* Our model (see Figure 2) borrows the idea of intention disentanglement [35], which has a built-in routing mechanism for clustering. Let  $x = \{y_{1:(T-1)}\}$  be the sequence of items clicked by user  $x$ , and the user’s ground-truth next click to predict is  $y_T$ . Let  $c_t \in \mathbb{R}^d$  be the embedding corresponding to item  $y_t$ ’s category ID. The embedding of item  $y_t$  is generated by the item encoder  $g_\theta : \mathcal{Y} \rightarrow \mathbb{R}^d$  as follows:

$$\begin{aligned} g_\theta(y_t) = & \text{MLP}_1([\text{embedding of item } y_t\text{'s unique ID}; \\ & \text{embedding of item } y_t\text{'s category ID, i.e., } c_t; \\ & \text{embedding of item } y_t\text{'s seller ID}; \\ & \text{embeddings of } y_t\text{'s tags; ...}]), \end{aligned} \quad (5)$$

where  $\text{MLP}_1(\cdot)$  represents a multilayer perceptron (MLP) whose input is the concatenation of item  $y_t$ ’s features. We then use another MLP to model the importance of each clicked item  $\{y_t\}_{t=1}^{T-1}$ :

$$\begin{aligned} p_t = & \frac{\exp(p'_t)}{\sum_{t'=1}^{T-1} \exp(p'_{t'})}, \text{ where} \\ p'_t = & \text{MLP}_2([\text{item embedding } g_\theta(y_t); \text{category embedding } c_t; \\ & \text{time gap between clicking item } y_t \text{ and item } y_T; \\ & \text{user's dwell time on item } y_t; \dots]) \text{ and } p'_t \in \mathbb{R}. \end{aligned} \quad (6)$$

Our model then uses a set of trainable parameters  $\mu_h \in \mathbb{R}^d$ ,  $h = 1, 2, \dots, H$ , to represent  $H$  intention prototypes [35], based on which each item  $y_t$  is being routed into intention  $h$  with probability  $p_{h|t}$ :

$$p_{h|t} = \frac{\exp(p'_{h|t})}{\sum_{h'=1}^H \exp(p'_{h'|t})}, \text{ where } p'_{h|t} = \frac{\langle \mu_h, c_t \rangle}{\rho \cdot \|\mu_h\| \cdot \|c_t\|}. \quad (7)$$

The temperature hyper-parameter  $\rho$  is set to  $\rho = 0.07$  following previous work [19], and  $\|\cdot\|$  stands for the  $l_2$ -norm of a vector. Based on the importance weights and the routing results, we obtain

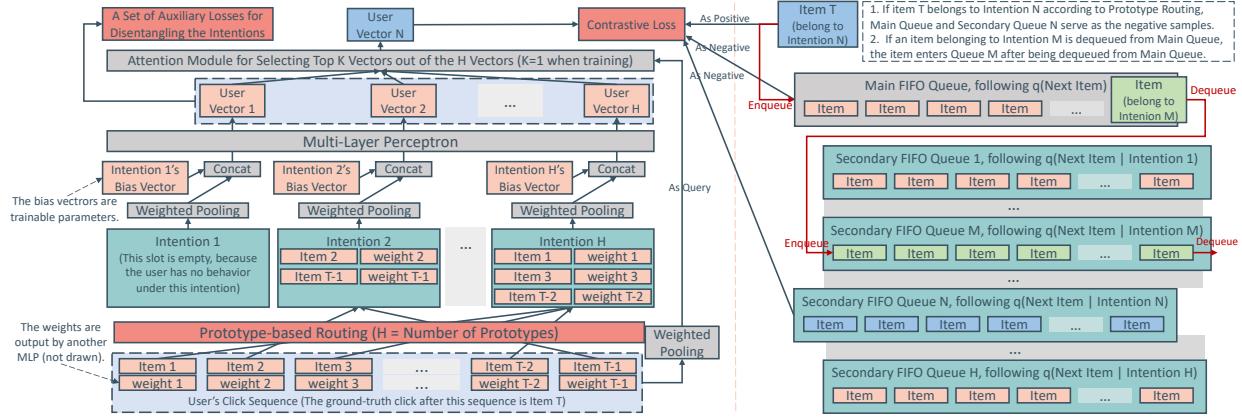


Figure 2: Multi-CLRec, a multi-queue implementation of contrastive learning for bias reduction in DCG.

$H$  intention vectors  $\{z_h\}_{h=1}^H$  for the user:

$$z_h = \text{MLP}_3([z'_h; \beta_h]), \quad z'_h = \sum_{t=1}^{T-1} p_{h|t} \cdot p_t \cdot g_\theta(y_t), \quad (8)$$

where  $\beta_h \in \mathbb{R}^d$ ,  $h = 1, 2, \dots, H$ , are trainable parameters that represent the bias vectors of the intentions, which are shared by all users. However, not all of the  $H$  intention vectors would match the user's current interest, since it is possible to have  $\sum_{t=1}^{T-1} p_{h|t} \cdot p_t \rightarrow 0$  for some  $h$ . We thus use the following attention module to select the intention vector that is most likely to match the user's interest, and set the user encoder  $f_\theta(\cdot)$ 's output to be the selected vector  $z_{h^*}$ :

$$\begin{aligned} f_\theta(x) = z_{h^*}, \quad \text{where } h^* = \arg \max_{h \in \{1, 2, \dots, H\}} w_h, \quad w_h = \frac{\exp(\tilde{w}_h)}{\sum_{h'=1}^H \exp(\tilde{w}_{h'})}, \\ \tilde{w}_h = \frac{\langle m, z_h \rangle}{\rho \cdot \|m\| \cdot \|z_h\|}, \quad m = \text{MLP}_4 \left( \sum_{t=1}^{T-1} p_t \cdot g_\theta(y_t) \right). \end{aligned} \quad (9)$$

We approximate the gradient of argmax via *straight-through* [5]. That is, we define  $\tilde{w}_h = \text{stop\_gradient}(I[h = h^*] - w_h) + w_h$ , where  $I[h = h^*] = 1$  if  $h = h^*$  and  $I[h = h^*] = 0$  otherwise. We then use  $\tilde{w}_h$  instead of  $h^*$  when computing the loss (see Equation 13).

When serving online, we can select the top- $K$  ( $1 \leq K \leq H$ ) vectors from  $\{z_h\}_{h=1}^H$  with maximum values of  $w_h$  for item retrieval, which is useful if a diverse recommendation list is preferred.

*Auxiliary Losses for Enhancing Interpretability.* The encoder architecture alone cannot guarantee an interpretable routing mechanism in our large-scale setting. We thus introduce three auxiliary losses to enhance the interpretability<sup>2</sup>. The first is to encourage  $p_{h|t}$  to be polarized, i.e.,  $p_{h^+|T} \rightarrow 1$  and  $p_{h'|T} \rightarrow 0$  for  $h' \neq h^+$ :

$$\mathcal{L}_{\text{aux},1} = -\log p_{h^+|T}, \quad \text{where } h^+ = \arg \max_{h \in \{1, 2, \dots, H\}} p_{h|T}. \quad (10)$$

The second auxiliary loss is for pushing the correct intention vector  $z_{h^+}$  closer to the target item  $y_T$  while making the incorrect

<sup>2</sup>Empirical results that demonstrate the enhanced interpretability are provided in the Appendix.

intention vectors  $\{z_{h'}\}_{h':h' \neq h^+}$ , far away from the target item:

$$\mathcal{L}_{\text{aux},2} = -\log \frac{\exp(s_{h^+,T})}{\sum_{h'=1}^H \exp(s_{h',T})}, \quad s_{h,T} = \frac{\langle z_h, g_\theta(y_T) \rangle}{\rho \cdot \|z_h\| \cdot \|g_\theta(y_T)\|}. \quad (11)$$

Our third auxiliary loss is for balancing the  $H$  intentions, such that each is responsible for a roughly equal number of items:

$$\mathcal{L}_{\text{aux},3} = \sum_{h=1}^H \frac{1}{H} \cdot (\log \frac{1}{H} - \log \pi_h), \quad \pi_h = \mathbb{E}_{\mathcal{B}}[p_{h|T}], \quad (12)$$

where  $\mathbb{E}_{\mathcal{B}}$  means that the expectation is computed based on the items encountered in the present mini-batch  $\mathcal{B}$ .

*Multi-Queue Contrastive Loss.* The main loss to optimize is:

$$\mathcal{L}_{\text{cl}} = -\log \sum_{h=1}^H \tilde{w}_h \cdot \frac{\exp \{ \cos(z_h, g_\theta(y_T)) / \rho \}}{\sum_{h'=1}^H \sum_{y' \in Q_0 \cup Q_{h^+}} \exp \{ \cos(z_{h'}, g_\theta(y')) / \rho \}}, \quad (13)$$

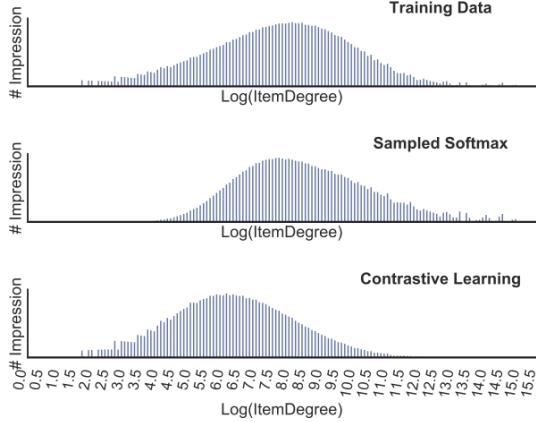
where  $\cos(x, y) = \langle x, y \rangle / (\|x\| \cdot \|y\|)$ . Here  $Q_0$  is the main queue, as described in Subsection 3.1, and  $Q_{h^+}$  is the  $h^+$ -th secondary queue (one of the  $H$  secondary queues) that stores negative samples associated with intention  $h^+ = \arg \max_h p_{h|T}$ . We update the  $1+H$  queues prior to computing the loss at each iteration as follows: (1) The positive sample  $y_T$  is enqueue into the main queue  $Q_0$  prior to computing the loss. (2) An item  $y^-$  just dequeued from the main queue  $Q_0$  is enqueue into the secondary queue  $Q_{h^-}$ , where  $h^- = \arg \max_h p_{h|y^-} = \arg \max_h \langle \mu_h, c^- \rangle / (\rho \cdot \|\mu_h\| \cdot \|c^-\|)$  and  $c^-$  is the category embedding of item  $y^-$ . (3) Items dequeued from the secondary queues  $\{Q_h\}_{h=1}^H$  are discarded.

We minimize the total loss  $\mathcal{L}_{\text{cl}} + \lambda_1 \mathcal{L}_{\text{aux},1} + \lambda_2 \mathcal{L}_{\text{aux},2} + \lambda_3 \mathcal{L}_{\text{aux},3}$ . We simply set  $\lambda_1 = \lambda_2 = \lambda_3 = 1$ , because we find that our method is robust with respective to a wide range of coefficients.

**Remark 2.** We are implicitly performing bias reduction based on a propensity score proportional to  $q(y | \text{user } x \text{'s intention is } h) + \alpha \cdot q(y)$  for a certain smoothing factor  $\alpha > 0$ , by using the negative samples from a secondary queue  $Q_h$  and the main queue  $Q_0$ .

**Table 1:** Aggregate diversity [2], i.e. the number of distinct items recommended to a randomly sampled subset of users.

Aggregated Diversity	
sampled-softmax	10,780,111
CLRec	21,905,318



**Figure 3:** The total number of impressions of the items in a specific degree bucket vs. the logarithm of the corresponding degree. The rightmost bar is not the highest because the number of extremely popular items is small, even though each item in the bucket has a very high degree.

**Table 2:** CLRec vs. the sampling-based alternatives. We conducted these proof-of-concept live experiments in a small-traffic scenario, due to the costs of online experiments. The negative sampling baseline has been outdated and removed from our live system before this work starts.

Method	HR@50 CTR(online)	
negative sampling	7.1%	outdated
shared negative sampling	6.4%	-
sampled-softmax	17.6%	3.32%
CLRec	17.8%	3.85%

## 4 EXPERIMENT

### 4.1 Experiments in Production Environments

The online experiments have lasted for at least four months, and our algorithm serves several scenarios with different traffic volumes. Details can be found in the Appendix. The total number of items is around 100 million. We use the queue-based implementation without caching in this subsection, and we will explore settings where encoding is expensive and requires caching in Subsection 4.2.2.

**4.1.1 The Debiasing Effect in Large-Scale Production Environments.** To examine the debiasing effects of CLRec, we first conduct offline experiments and compare CLRec with sampled softmax. We report

**Table 3:** Main live experiment results conducted in one of the largest scenarios on our platform. CLRec consistently outperforms the baseline for months and has been fully deployed since Feb 2020. A method with a large popularity index tends to recommend popular items while a method with a small index is fairer to the under-exposed items.

Method	CTR	Average Dwell Time	Popularity Index
MIND	5.87%	-	0.658
CLRec	6.30%	+11.9%	0.224

**Table 4:** The user distribution of our platform usually changes significantly when the weekend arrives. When the user distribution changes significantly, Multi-CLRec has an advantage over single-queue CLRec in terms of the offline evaluation metric HitRate@50. Multi-CLRec replaced CLRec and became the new default in mid-2020, bringing a 3% relative improvement in terms of total clicks on our platform while maintaining a comparable click-through rate.

Method	Train on Weekdays' Data	
	Test on Weekdays	Test on Weekends
CLRec	17.18%	17.16%
Multi-CLRec	17.25%	17.68%

the aggregated diversity [2] in Table 1, and the distributions of the recommended items resulted from the different losses in Figure 3.

Table 1 shows that CLRec has an over 2× improvement on aggregated diversity. We see from Figure 3 that, sampled softmax tends to faithfully fit the distribution of the training data. CLRec, however, learns a relatively different distribution, which shifts towards the under-explored items and alleviates the Matthew effect.

Bias reduction not only leads to a fairer system, but also contributes to a significant improvement regarding the online performance. In Table 2, we compare CLRec with other sampling alternatives using the same implementation of the encoders. Details of the alternative methods are in the Appendix. We observe that negative sampling [38], including its variant [56] that makes the instances in a batch share the same large set of negative samples, does not perform well in our settings. CLRec’s improvement over sampled-softmax [6, 26] w.r.t. the offline metric HitRate@50 is negligible. However, CLRec achieves significant improvement regarding the click-through rate (CTR) online, which indicates that there exists discrepancy between the offline metric and the online performance.

In Table 4, we compare Multi-CLRec against CLRec. We deliberately evaluate the methods in the setting where the training and the test data have different user distributions. Multi-CLRec is more robust compared to CLRec and achieves higher performance.

**4.1.2 At Least Four Months’ Large-Scale A/B testing.** CLRec has been fully deployed into several heavy-traffic scenarios since Feb 2020, after the initial proof-of-concept ablation studies shown in Table 2. Table 3 shows our main online results conducted in these heavy-traffic scenarios, with billions of page views each day. During

**Table 5:** Training speed in terms of the number of the positive examples processed per second, and the average network traffic in a distributed environment.

Method	Samples/Sec	NetTraffic (MB/s)
sampled-softmax w/o feat.	$\approx 280k$	$\approx 700$
sampled-softmax with feat.	$\approx 130k$	$\approx 1,100$
CLRec w/o feat.	$\approx 330k$	$\approx 500$
CLRec with feat.	$\approx 280k$	$\approx 500$

**Table 6:** The benefits of encoding features for the negative samples. Most baselines that employ sampled-softmax do not encode rich features for the negative samples (though they still use features when encoding the users’ click sequences), because the number of negative samples is large and brings high costs if the features are complex. Fortunately, CLRec’s cached implementation greatly reduces the costs, as demonstrated in Table 5.

Method	HR@50
CLRec + negatives w/o features	17.4%
CLRec + negatives with features	19.4%

the at least four months’ A/B testing, CLRec has been consistently outperforming MIND [32], the previous state-of-art baseline, in terms of the fairness metrics such as aggregated diversity and average popularity index, as well as the user engagement metrics such as click-through rate and average dwell time. Compared with MIND [32], which is the previous state-of-art DCG baseline deployed in the system, CLRec tends to recommend items with a lower popularity index while being more attractive to the users. This proves CLRec’s ability of identifying high-quality items that are rarely explored by the previous systems. We further achieve a +2% relative improvement in the total click number on our platform after ensembling CLRec and MIND, compared to using MIND alone.

In mid-2020, we upgrade CLRec to Multi-CLRec, and Multi-CLRec brings a 3% relative improvement in terms of total clicks on our platform while maintaining a comparable click-through rate.

## 4.2 Computational Advantages of CLRec

**4.2.1 Training Efficiency.** Table 5 compares CLRec and sampled-softmax in terms of training speed and the network traffic required in a distributed setting. CLRec’s queue-based implementation is much more efficient than the methods that perform explicit sampling, since CLRec reuses the result computed for a positive sample shortly later when the sample is serving as a negative sample. The version of sampled-softmax that encodes features for the negative items is from [62]. This proof-of-concept experiment only uses a few features of the items, and we can deduce that the improvement regarding efficiency will be much more substantial with more complex features. Table 6 shows that encoding features for the negative samples is beneficial, which justifies the efforts spent on efficiency.

**Table 7:** Task u2i is the regular task where  $x$  is a sequence of clicks and  $y$  is the next click to be predicted. Task u2u [36] adds an auxiliary loss where  $x$  and  $y$  are both sequences from the same user (before and after a sampled timestamp), which is co-trained with task u2i. HR1@50 and Recall5@50 represent the HitRate and Recall truncated at 50 when predicting the next one and five clicks, respectively.

Task & Implementation	HR1@50	Recall5@50
CLRec-u2i	17.9%	12.1%
CLRec-u2u, cached	18.3%	12.7%
CLRec-u2u, cached + MoCo	18.2%	12.6%

**Table 8:** Results on public benchmarks preprocessed by SASRec[29] and BERT4Rec [46]. We follow BERT4Rec’s approach when constructing the test set, which penalizes false positive predictions on popular items. For fair comparison, the IPW implementation and the CLRec implementation in this table use the same Transformer [34] encoder as SASRec but with a IPW loss and a contrastive loss, respectively.

Method	Metric	ML-1M	Beauty	Steam
SASRec	HR@1	0.2351	0.0906	0.0885
BERT4Rec		0.2863	0.0953	0.0957
IPW		0.2934	0.1129	0.1381
CLRec	<b>0.3013</b>	<b>0.1147</b>	<b>0.1424</b>	
SASRec	HR@5	0.5434	0.1934	0.2559
BERT4Rec		0.5876	0.2207	0.2710
IPW		0.5985	0.2404	0.3539
CLRec	<b>0.6045</b>	<b>0.2552</b>	<b>0.3640</b>	
SASRec	HR@10	0.6629	0.2653	0.3783
BERT4Rec		0.6970	0.3025	0.4013
IPW		0.7156	0.3215	0.4924
CLRec	<b>0.7194</b>	<b>0.3423</b>	<b>0.5019</b>	

**4.2.2 Complex Pretext Task that Requires the Cached Implementation.** We now demonstrate that CLRec with a queue that caches the computed results can enable more complex pretext tasks that may improve the quality of the learned representations. We consider an auxiliary task where  $x$  and  $y$  are both sequences from the same user (before and after a sampled timestamp). The goal is to identify the correct sequence  $y$  that belongs to the same user that produces  $x$ . This auxiliary task is expensive to implement with sampled-softmax, since the negative samples are sequences and are thus expensive to encode. Fortunately, cached CLRec can implement it efficiently. Table 7 demonstrates that the auxiliary task can improve an algorithm’s ability to make long-term prediction.

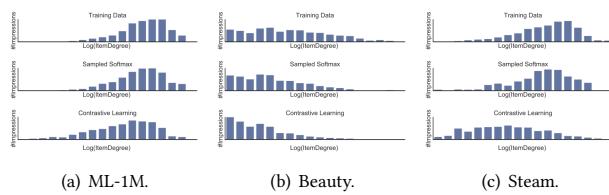
MoCo [19] proposes momentum update for stabilizing the training loss. We however observe no additional gain with MoCo.

**Table 9: Results on public benchmarks preprocessed by HPMN [40]. We follow HPMN [40] and use AUC as the metric. The clicks in the training and the test set of each benchmark happen before and after a certain timestamp, respectively. There hence exists a distribution drift. The users also have diverse interests, i.e., the probability of a user’s two consecutive clicks belonging to the same category is low.**

Method	Amazon	UserBehavior
DNN	0.7546	0.7460
SVD++ [31]	0.7155	0.8371
GRU4Rec [21]	0.7760	0.8471
Caser [47]	0.7582	0.8745
DIEN [60]	0.7770	0.8934
RUM [13]	0.7464	0.8370
LSTM [23]	0.7765	0.8681
SHAN [55]	0.7763	0.8828
HPMN [40]	0.7809	0.9240
CLRec	0.7785	0.9278
IPW Loss + Multi-CLRec’s Encoder	0.7888	0.9319
Multi-CLRec	<b>0.7993</b>	<b>0.9353</b>

**Table 10: Aggregate diversity on public datasets, where each algorithm retrieves ten items for each test user.**

Dataset	#Users	Aggregate Diversity		
		Ground-Truth	Sample-Softmax	CLRec
ML-1M	6,040	1,883	2,348	3,020
Beauty	40,226	19,820	24,340	44,408
Steam	281,428	10,006	5,154	10,111



**Figure 4: The total number of impressions of the items a specific degree bucket vs. the logarithm of the corresponding degree. Sample softmax tends to faithfully fit the training data’s distribution, while CLRec reduces the Matthew effect.**

### 4.3 Reproducible Experiments on Public Data

In addition to the experiments on large-scale data, we also conduct quantitative experiments on several public benchmarks. The quantitative results are listed in Table 8 and Table 9, respectively.

We also conduct the same analyses as in Subsection 4.1.1 to demonstrate the debiasing effects of CLRec on the public benchmarks. We sample a sequence from each user in the training set as the input to the recommenders. Each recommender then retrieves ten items for each sequence. The results are in Table 10 and Figure 4.

Table 10 reports aggregate diversity of the items retrieved by the algorithms, as well as the aggregate diversity of the ground-truth next clicks. CLRec retrieves a more diverse set of items than sampled softmax, i.e., much more items will have a chance to be recommended after we reduce the bias in the training data.

Figure 4 visualizes the distributions of the items retrieved by each algorithm, as well as the training data’s distribution. Sample softmax tends to faithfully fit the training data’s distribution, while CLRec tends to explore the previously under-explored items.

## 5 RELATED WORK

**Deep Candidate Generation.** Deep candidate generation methods are widely deployed in industrial systems, e.g., YouTube [10, 14, 27], Taobao [32, 35, 59, 61], and Pinterest [56]. The existing methods explicitly sample negative examples from a pre-defined proposal distribution [6, 26, 38]. The proposal distribution not only affects convergence, but also has a significant impact on the performance [9]. Empirically the number of the negative samples need to be large, e.g. a few thousand ones for pairing with a positive example. Consequently, it is computationally expensive to incorporate rich features for the negative samples. The existing systems hence usually choose to not encode features for the negative examples except for simple features such as item IDs [14], even though rich features for the negative samples are demonstrated to be beneficial [4]. CLRec achieves great efficiency when encoding rich features for the negative samples by caching the computed results.

**Bias Reduction and Fairness in Recommender Systems.** Recommendation algorithms that directly fits the training data will suffer from selection bias due to the missing-not-at-random phenomenon [42], where the previous recommendation algorithms affect the training data collected. The topic of reducing the bias in training and evaluating recommender systems has been explored before [3, 10, 20, 42, 44, 52, 54, 57, 63]. However, these existing works mostly focus on small-scale offline settings, and rely on techniques less efficient for large-scale DCG. For example, most of them involve an extra stage to train a propensity score estimator. Dividing the propensity scores can also lead to high variance when small propensities exist [42]. Correcting the bias helps improve P-fairness [8], i.e. fairness towards the under-recommended products [7].

**Contrastive Learning.** Contrastive learning, which aims to learn high-quality representations via self-supervised pretext tasks, recently achieves remarkable success in various domains, e.g., speech processing [39], computer vision [11, 19, 22], graph data [50], and compositional environments [30]. The contrastive loss we investigate in this paper is a generalization of the InfoNCE loss [39]. InfoNCE is previously understood as a bound of the mutual information between two variables [39]. Our work provides a new perspective on the effectiveness of the contrastive loss, by illustrating its connection with inverse propensity weighting [25, 41, 48].

## 6 CONCLUSION

We established in theory the connection between contrastive learning and IPW, based on which we proposed CLRec and Multi-CLRec for efficient and effective bias reduction in large-scale DCG.

## REFERENCES

- [1] Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *OSDI*. 265–283.
- [2] Gediminas Adomavicius and YoungOk Kwon. 2011. Improving aggregate recommendation diversity using ranking-based techniques. *TKDE* (2011).
- [3] Qingyao Ai, Keping Bi, Cheng Luo, Jiafeng Guo, and W Bruce Croft. 2018. Unbiased learning to rank with unbiased propensity estimation. In *SIGIR*. 385–394.
- [4] Jinze Bai, Chang Zhou, Junshuai Song, Xiaoru Qu, Weiting An, Zhao Li, and Jun Gao. 2019. Personalized Bundle List Recommendation. In *WWW*. 60–71.
- [5] Yoshua Bengio, Nicholas Léonard, and Aaron C. Courville. 2013. Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation. *arXiv preprint arXiv:1308.3432* (2013).
- [6] Yoshua Bengio and Jean-Sébastien Sénelac. 2008. Adaptive importance sampling to accelerate training of a neural probabilistic language model. *IEEE Trans. Neural Networks* 19, 4 (2008), 713–722.
- [7] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, and Cristos Goodrow. 2019. Fairness in Recommendation Ranking through Pairwise Comparisons. In *SIGKDD*.
- [8] Robin Burke. 2017. Multisided Fairness for Recommendation. *arXiv:1707.00093 [cs.CY]*
- [9] Hugo Caselles-Dupré, Florian Lesaint, and Jimena Royo-Letelier. 2018. Word2vec applied to recommendation: Hyperparameters matter. In *RecSys*. ACM, 352–356.
- [10] Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H Chi. 2019. Top-k off-policy correction for a REINFORCE recommender system. In *WSDM*. 456–464.
- [11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709* (2020).
- [12] Ting Chen, Yizhou Sun, Yue Shi, and Liangjie Hong. 2017. On sampling strategies for neural network-based collaborative filtering. In *SIGKDD*.
- [13] Xu Chen, Hongteng Xu, Yongfeng Zhang, Jiaxi Tang, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2018. Sequential recommendation with user memory networks. In *WSDM*. 108–116.
- [14] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *RecSys*. ACM, 191–198.
- [15] G. Cybenko. 1989. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems* (1989).
- [16] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*. 249–256.
- [17] Mihajlo Grbovic and Haibin Cheng. 2018. Real-time personalization using embeddings for search ranking at airbnb. In *SIGKDD*.
- [18] Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*. 297–304.
- [19] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2019. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722* (2019).
- [20] Yue He, Peng Cui, Jianxin Ma, Hao Zou, Xiaowei Wang, Hongxia Yang, and Philip S. Yu. 2020. Learning Stable Graphs from Multiple Environments with Selection Bias. In *SIGKDD*.
- [21] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939* (2015).
- [22] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2018. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670* (2018).
- [23] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [24] Kurt Hornik. 1991. Approximation capabilities of multilayer feedforward networks. *Neural Networks* (1991).
- [25] Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.
- [26] Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2014. On using very large target vocabulary for neural machine translation. *arXiv preprint arXiv:1412.2007* (2014).
- [27] Manas R Joglekar, Cong Li, Jay K Adams, Pranav Khaitan, and Quoc V Le. 2019. Neural input search for large scale recommendation models. *arXiv preprint arXiv:1907.04471* (2019).
- [28] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with GPUs. *arXiv preprint arXiv:1702.08734* (2017).
- [29] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *ICDM*. IEEE, 197–206.
- [30] Thomas Kipf, Elise van der Pol, and Max Welling. 2019. Contrastive Learning of Structured World Models. *arXiv preprint arXiv:1911.12247* (2019).
- [31] Yehuda Koren. 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *SIGKDD*.
- [32] Chao Li, Zhiyuan Liu, Mengmeng Wu, Yuchi Xu, Huan Zhao, Pipei Huang, Guoliang Kang, Qiwei Chen, Wei Li, and Dik Lun Lee. 2019. Multi-interest network with dynamic routing for recommendation at Tmall. In *CIKM*.
- [33] Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, and Tony Jebara. 2018. Variational Autoencoders for Collaborative Filtering. In *WWW*.
- [34] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130* (2017).
- [35] Jianxin Ma, Chang Zhou, Peng Cui, Hongxia Yang, and Wenwu Zhu. 2019. Learning disentangled representations for recommendation. In *NeurIPS*. 5712–5723.
- [36] Jianxin Ma, Chang Zhou, Hongxia Yang, Peng Cui, Xin Wang, and Wenwu Zhu. 2020. Disentangled Self-Supervision in Sequential Recommenders. In *SIGKDD*.
- [37] Andrew McCallum, Kamal Nigam, et al. 1998. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*. Citeseer, 41–48.
- [38] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NeurIPS*. 3111–3119.
- [39] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).
- [40] Kan Ren, Jiarui Qin, Yuchen Fang, Weinan Zhang, Lei Zheng, Weijie Bian, Guorui Zhou, Jian Xu, Yong Yu, Xiaoqiang Zhu, et al. 2019. Lifelong sequential modeling with personalized memorization for user response prediction. In *SIGIR*. 565–574.
- [41] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.
- [42] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as treatments: Debiasing learning and evaluation. *arXiv preprint arXiv:1602.05352* (2016).
- [43] Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *NeurIPS*. 1857–1865.
- [44] Harald Steck. 2013. Evaluation of recommendations: rating-prediction and ranking. In *RecSys*. 213–220.
- [45] Stergios Stergiou, Zygmantas Straznickas, Rolina Wu, and Kostas Tsoutsouliklis. 2017. Distributed negative sampling for word embeddings. In *AAAI*.
- [46] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *CIKM*. 1441–1450.
- [47] Jiaxi Tang and Ke Wang. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *WSDM*. 565–573.
- [48] Steven K. Thompson. 2012. *Sampling*. Wiley.
- [49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv preprint arXiv:1706.03762* (2017).
- [50] Petar Veličković, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. 2018. Deep graph infomax. *arXiv preprint arXiv:1809.10341* (2018).
- [51] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. 2018. Cosface: Large margin cosine loss for deep face recognition. In *CVPR*. 5265–5274.
- [52] Xuanhui Wang, Michael Bendersky, Donald Metzler, and Marc Najork. 2016. Learning to rank with selection bias in personal search. In *SIGIR*. 115–124.
- [53] Nicolai Wojke and Alex Bewley. 2018. Deep cosine metric learning for person re-identification. In *WACV*. IEEE, 748–756.
- [54] Longqi Yang, Yin Cui, Yuan Xuan, Chenyang Wang, Serge Belongie, and Deborah Estrin. 2018. Unbiased offline recommender evaluation for missing-not-at-random implicit feedback. In *RecSys*. 279–287.
- [55] Haocho Ying, Fuzhen Zhuang, Fuzheng Zhang, Yanchi Liu, Guandong Xu, Xing Xie, Hui Xiong, and Jian Wu. 2018. Sequential recommender system based on hierarchical attention network. In *IJCAI*.
- [56] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton, and Jure Leskovec. 2018. Graph Convolutional Neural Networks for Web-Scale Recommender Systems. *arXiv preprint arXiv:1806.01973* (2018).
- [57] Shengyu Zhang, Dong Yao, Zhou Zhao, Tat-Seng Chua, and Fei Wu. 2021. CauseRec: Counterfactual User Sequence Synthesis for Sequential Recommendation. In *SIGIR*.
- [58] Chang Zhou, Jinze Bai, Junshuai Song, Xiaofei Liu, Zhengchao Zhao, Xiusi Chen, and Jun Gao. 2018. ATRank: An attention-based user behavior modeling framework for recommendation. In *AAAI*.
- [59] Chang Zhou, Yuqiong Liu, Xiaofei Liu, Zhongyi Liu, and Jun Gao. 2017. Scalable graph embedding for asymmetric proximity. In *AAAI*.
- [60] Guorui Zhou, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai. 2019. Deep interest evolution network for click-through rate prediction. In *AAAI*, Vol. 33. 5941–5948.
- [61] Han Zhu, Xiang Li, Pengye Zhang, Guozheng Li, Jie He, Han Li, and Kun Gai. 2018. Learning Tree-based Deep Model for Recommender Systems. In *SIGKDD*.
- [62] Rong Zhu, Kun Zhao, Hongxia Yang, Wei Lin, Chang Zhou, Baole Ai, Yong Li, and Jingren Zhou. 2019. AliGraph: a comprehensive graph neural network platform. *VLDB* 12, 12 (2019), 2094–2105.

- [63] Hao Zou, Peng Cui, Bo Li, Zheyuan Shen, Jianxin Ma, Hongxia Yang, and Yue He. 2020. Counterfactual Prediction for Bundle Treatment. In *NeurIPS*.

## A EXPERIMENTAL SETTINGS

*Data and Hyper-parameters.* We use the data collected from the last four days on our platform as the training data and use the click data recorded in the following day as the test set. We limit the minimum length of a valid sequence to be five in the training set, because requiring the model to predict the next item of an unusually short sequence can easily introduce great noise which harms the recommendation performance. We use seven categorical features of the items, including item ID, coarse-grained and fine-grained category IDs, seller ID, seller category, brand ID, and gender preference, for our live experiments. We use the queue-based implementation without the caching mechanism on the regular tasks where the negative items consist of the seven categorical features, while using the queue-based implementation that caches the processed results when solving the complex pretext tasks due to the high costs of encoding the complex negative examples. The number of negative samples is 2,560 while the batch size is 256. Multi-CLRec use  $H = 16$  on public data and  $H = 64$  on our large-scale data. We train the model for one epoch in total on our large-scale data.

*Training Environment.* The training data contain four billion sequences of user behaviors. We train the model in a distributed TensorFlow cluster that consists of 140 workers and 10 parameter servers. Each worker is equipped with a GPU, which is either GTX 1080 Ti or Tesla P100. Each parameter server is allocated 32GB memory while each worker has 16G memory. Training CLRec and Multi-CLRec on our internal four-day data for one epoch costs less than eight hours for task u2i without caching and less than twelve hours for task u2u with caching.

*Online Serving.* We pre-compute offline the item representations via the item encoder. We then store and index the item representations using an online vector-based kNN service for fast retrieval of top relevant items at serving time. At serving time, our user encoder will infer a user vector  $f_\theta(x)$  as the request made by user  $x$  arrives and pass the user vector to the kNN service to find a few hundred relevant items  $S$ . The latency for dealing with the query is around 10 ms. The set of relevant items  $S$  is then send to be scored by a separate deep ranking model that costs much more computational power and is responsible for constructing the final recommendation list by selecting the top relevant items from  $S$ .

*Evaluation Metrics.* We report the performance in terms of the following offline and online metrics:

- HitRate@50 (HR@50) is an offline metric. We sample a subset of users and compute the metric based on the users’ click sequences to measure an algorithm’s offline performance. We pick the latest one click of a user as the label  $y$  to predict, and use the user’ sequence prior to  $y$  to serve as the input  $x$ . We run the algorithm under evaluation to retrieve top 50 items for each user via kNN searching. If the ground-truth next one click  $y$  is in the retrieved 50 items, the algorithm receives a score 1 for this user, or score 0 otherwise. The HitRate@50 of the algorithm is then the average score received by the algorithm.

- The click-trough rate (CTR) of an algorithm is the percentage of recommendations made by the algorithm that are finally clicked by the users.
- The average dwell time is the total time spent by the users on reading the details of the items clicked by them, divided by the total number of clicks on our platform.
- The aggregate diversity [2], measured on a sampled subset of users for testing, is the number of distinct items recommended to the subset of users.
- The popularity index of the items recommended by a candidate generation algorithm  $A$  is defined as  $\frac{\mathbb{E}_A[p_{\text{data}}(y)]}{\max_{A'} \mathbb{E}_{A'}[p_{\text{data}}(y)]}$ . Here  $p_{\text{data}}(y)$  is the number of times our system recommends item  $y$  to the users in the past few weeks before we deploy CLRec into the system, and  $\mathbb{E}_A[p_{\text{data}}(y)] = \sum_{y \in \mathcal{Y}} p_{\text{data}}(y)$ .  $p$ (algorithm  $A$  recommends item  $y$ ) measures how much an algorithm prefers recommending the items that are already popular. The algorithm  $A'$  that has the largest value of  $\mathbb{E}_{A'}[p_{\text{data}}(y)]$  in our system is a simple candidate generation method that does not learn vector representations, which is used as a fallback to handle the cases when the deep candidate generation methods, i.e. MIND and CLRec, fail to return the results within a time limit.

## B IMPLEMENTATION DETAILS

### B.1 Single-Queue CLRec’s Encoders

The encoders of the multi-queue implementation Multi-CLRec is described in Subsection 3.2. However, the single-queue CLRec uses a different set of user and item encoders, which is due to the fact that our large-scale experiments related with the single-queue CLRec were conducted before we invented Multi-CLRec. We thus describe in this subsection the single-queue CLRec’s encoders.

*Item Encoder.* In the single-queue CLRec, an item’s embedding is the sum of several parts, i.e. a base item embedding corresponding to its ID number, the concatenation of the item features’ embeddings, a time bucket embedding, as well as a backward positional embedding. The time bucket embedding and the positional embedding are only added when the item embedding is to be used by the user encoder  $f_\theta(x)$  for encoding a past click made by user  $x$ . The output of the item encoder  $g_\theta(y)$  for item  $y$  is the sum of only the base item embedding and the item features’ embeddings, excluding the time and positional embeddings. Notably, Multi-CLRec also use the time and positional embeddings described here for estimating each click’s importance weight.

*Time Bucket Embedding.* To encode time intervals between consecutive behaviors, we use the embeddings of bucketized time intervals, which are used by previous work [58]. To be specific, given the current timestamp  $t$  when the user query arrives and the timestamp  $t'$  when a click was recorded, the time interval  $t - t'$  is discretized and put into a bucket that corresponds to a specific range. Each bucket has its own vector representation, i.e. bucket embedding.

*Backward Positional Embedding.* We also find that the positional embedding has extra positive impacts even when the time bucket embedding is already in use. There are two ways to add positional embeddings. The forward way is to view the latest click as if it is at the anchor position, i.e. position zero, while the earliest click of the

user would then be at a varying position depending on the user’s sequence length. The backward way is, on the other hand, to view the earliest click as if it is at position zero, and the latest click would then be at a varying position when the sequence length varies. We find that the backward version of the positional embedding brings extra improvement even after the bucket embedding is used, while the forward positional embedding does not. We thus adopt the backward positional embedding. The time bucket embedding has already highlighted the recent behaviors, and the backward positional embedding may be behaving like the backward direction in a bi-directional LSTM for capturing long-range dependencies.

**B.1.2 User Encoder.** For single-queue CLRec, we employ a simplified multi-head attention (MultiAtt) module for capturing a user’s diverse interests hidden behind the user’s click sequence  $x = \{y_t\}_{t=1}^{T-1}$ . We then perform weighted head aggregation (WHA) of the multiple heads’ outputs, i.e., head embeddings. That is,  $f_\theta(x) = \text{WHA}(\text{MultiAtt}(X))$ , where  $X \in \mathbb{R}^{(T-1) \times d}$  are the  $T - 1$  clicks’ embeddings, whose specification has been described earlier.

**Multi-Head Attention.** Multi-head attention [49] achieves great success in various NLP tasks and is adopted by sequential recommendation models [29, 58]. We simplify the multi-head attention [34] for efficiency’s sake and implement the following module:

$$\text{MultiAtt}(X) = \text{SoftMax}(\text{MLP}(X)^\top, -1) X,$$

where  $X \in \mathbb{R}^{(T-1) \times d}$  are the  $T - 1$  clicks’ embeddings. MLP is a two-layer feed forward neural network whose output is of shape  $\mathbb{R}^{(T-1) \times H}$ , representing the  $H$  sets of attention scores for the  $H$  attention heads, respectively. The softmax is performed along the last axis of its input, hence the  $-1$ . The output of the above module is a set of  $H d$ -dimensional embeddings, namely  $H$  head embeddings.

**Weighted Heads Aggregation (WHA).** After the simplified multi-head attention module, there are  $H$  head embeddings  $\{z_h\}_{h=1}^H$ . We use a WHA module to aggregate the  $H$  head embeddings into one single user embedding. WHA first produces a global feature vector  $m \in \mathbb{R}^d$ , based on which another weighted attention is performed to produce the final user embedding  $f_\theta(x) \in \mathbb{R}^d$  for user  $x$ :

$$m = \frac{\sum_{h=1}^H z_h}{H}, \quad \alpha_h = \frac{\exp(z_h^\top A m)}{\sum_{h'=1}^H \exp(z_{h'}^\top A m)}, \quad f_\theta(x) = \sum_{h=1}^H \alpha_h z_h,$$

where  $A \in \mathbb{R}^{d \times d}$  is a learnable parameter. We find WHA to be more effective than mean-pooling or concatenation.

## B.2 Normalization and Initialization

WHA and contrastive learning both seem to be sensitive to the similarity metrics used. We thus  $l2$ -normalize the head embeddings before sending them into WHA and normalize the output of WHA as well, which empirically improves training convergence and recommendation diversity. We use the cosine similarity with temperature  $\rho > 0$  for contrastive learning, i.e.  $\phi_\theta(x, y) = \frac{\langle f_\theta(x), g_\theta(y) \rangle}{\rho \|f_\theta(x)\| \|g_\theta(y)\|}$ , where  $\rho = 0.07$  following previous work [19].

Some efforts [35, 51, 53] have been spent on analyzing the cosine similarity related with  $l2$ -normalization as well as the positive effects brought by  $l2$ -normalization. According to [53] which assumes

a normal distribution, cosine softmax will force the embeddings to form more compact clusters, which makes kNN search more easy.

The embeddings should use normal initialization [16] instead of uniform initialization if  $l2$ -normalization is in use, so that the initial embeddings are uniformly spread over the hyper-sphere, which prevents the loss from oscillating at the beginning of training.

## B.3 Loss Functions and Sampling Strategies

**Negative sampling** [38]. We sample  $L$  examples to pair with each positive click from a proposal distribution  $q(y)$ . The proposal distribution  $q(y)$  is proportional to the item’s degree, i.e.,  $\text{degree}(y)^{0.75}$  as recommended in [38]. We use a distributed version of the sampling strategy based on the alias method [45], which is provided by AliGraph [62]. We tune  $L$  from 8 to 2,560 and report the best.

**Shared negative sampling** [56]. This variant aims to increase the efficiency of the original implementation of negative sampling. It makes the positive examples in the present batch share the same set of  $L$  negative samples, rather than sampling a different set of  $L$  negative samples for each positive instance. Sampled softmax, as well as CLRec, similarly let the positive examples in a batch share the same set of negative samples. However, this implementation of negative sampling with sharing still uses the binary cross-entropy loss used by the original negative sampling implementation.

**Sampled softmax** [26]. We use the sampled softmax implemented in Tensorflow [1], which subtracts the correction term to make the sampled loss approximately optimize the same objective as the full loss. We set the number of negative examples used by sample softmax per batch to be the same as the queue size of CLRec.

**CLRec & Multi-CLRec.** The batch size is 256. The queue size of CLRec is 2,560. The queue size of Multi-CLRec is 1,280, leading to 2,560 negative items in total as it draws samples from both a main queue and an additional secondary queue.

## B.4 Complex Pretext Tasks

In task u2u [36],  $x$  and  $y$  are both sequences from the same user, before and after a sampled timestamp, respectively. Intuitively, pre-training a recommender to solve task u2u may improve a recommender’s ability to make long-term prediction. We co-train the same encoders to solve both task u2u and the original u2i task where  $x$  and  $y$  are a sequence and the sequence’s next click. Task u2i is still required, since we ultimately need the recommender to recommend items rather than sequences at serving time. The batch size is 256. We use one separate queue of size 2,560 for each task. We limit the sequence length of  $y$  in task u2u to be less than ten. We also investigate whether Momentum Contrast (MoCo) [19] may improve performance, and vary the momentum parameter in the range of  $\{0.9, 0.99, 0.999, 0.9999\}$  and report the best result.

## C ADDITIONAL EMPIRICAL RESULTS

In this section, we provide additional empirical results that demonstrate the interpretability of Multi-CLRec and justify some designs of our encoders.



**Figure 5: Multi-CLRec’s user encoder can select and output top- $K$  vectors at serving time. We show here the top- $K$  vectors (for  $K = 7$ ) do represent the user’s  $K$  different intentions, owing to the auxiliary losses for intention disentanglement.**

### C.1 Results about Multi-CLRec’s Encoders

In Figure 5, we demonstrate that the top- $K$  vectors selected by Multi-CLRec’s user encoder do represent a user’s  $K$  different intentions, which is mostly due to the auxiliary losses for intention disentanglement. We note that the case shown here is *not* cherry-picked, but a randomly sampled case.

### C.2 Results about CLRec’s Encoders

In this subsection, we provide some interesting empirical findings that motivates the design of our encoders when developing the single-vector CLRec, even though the findings are not strictly related to contrastive learning or bias reduction. Each of the following

experiments should be taken as the analysis for each component individually, which are conducted during different time periods. Considering the costs of deployment into a live system, we deploy a model online only if it has at least similar performance with the previous baselines, and mark a method’s online performance as “-” in the table if it is not deployed.

Table 11 then demonstrate the usefulness of combining time bucket embedding and backward positional embedding.

We then illustrate in Table 12 that the encoders can benefit from the combination of weighted head aggregation (WHA) and the cosine similarity, where the results show that cosine similarity is indeed needed for reaching a higher HitRate.

We also find that combining WHA and cosine can endow a single-vector user encoder with the ability to generate a diversified recommendation list that reflects the different intents behind a user’s behavior sequence (see Figure 6). Without WHA, we observe that the retrieved item set can be easily dominated by one single intent, showing limited diversity.

The previous methods, especially those use dot product instead of cosine similarity, easily lead to a much skewer interest distribution of the top- $k$  retrieved items. These previous methods tend to under-explore items from the less popular interests. Many have also noticed this issue and proposed their solutions, such as using multiple interest vectors [32, 35] or multiple interest sub-models [61]. Our result (Figure 6), however, indicates that a single-vector model can sometimes be sufficient for generating a diverse recommendation list.

**Table 11: Effects of the forward positional embedding, backward positional embedding, and time bucket embedding (TBE). We conducted these ablation studies about the encoders in a small-traffic scenario.**

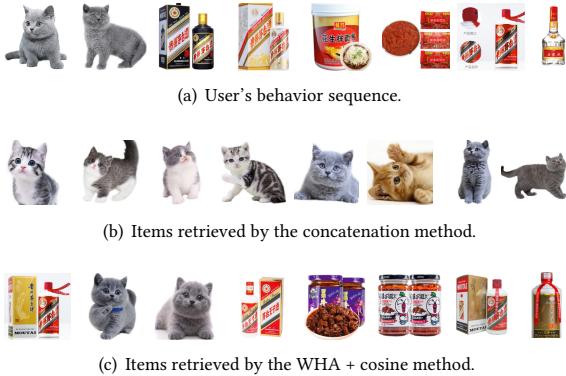
Method	HR@50 (Offline)	CTR (Online)
FwdPosEmb	18.5%	-
BackPosEmb	18.2%	2.98%
TimeBucketEmb	18.9%	-
FwdPositionEmb + TBE	16.4%	-
BackPosEmb + TBE	19.1%	3.25%

**Table 12: Performance of the different head aggregation strategies used by the single-vector CLRec’s user encoder.**

Method	HR@50 (Offline)	CTR (Online)
MeanPooling	15.1%	3.10%
MeanPooling + Cosine	18.1%	-
MeanPooling + MLP	10.5%	-
MeanPooling + MLP + Cosine	17.7%	-
Concatenation + MLP	17.8%	3.15%
WHA + Cosine	19.7%	3.21%

## D THEORECTICAL ANALYSES AND PROOFS

In this section, we provide the derivation of the IPW loss for a multinomial policy, along with the proof of our theorem.



**Figure 6: It is possible for a single-vector user encoder to generate a diverse recommendation list, as long as the model is properly designed, e.g., if the model combines the weighted head aggregation (WHA) strategy with the cosine similarity. Nevertheless, the single-vector CLRec is not as controllable and reliable as the multi-vector Multi-CLRec when it comes to preserving the multiple intentions in a balanced manner.**

## D.1 Derivation of Inverse Propensity Weighting for Multinomial Policies

In this subsection, we derive the IPW loss for DCG, where we represent the recommendation policy as a multinomial distribution. We assume that the recommender system is a sequential recommender system where a user will receive only one recommendation  $y$  when the user's state becomes  $x$ . However, we note that it is easy to verify that our conclusions still hold when user state  $x$  receives  $K$  recommendations, where  $K \geq 1$ . We will also focus on a single user state  $x$  for conciseness.

**D.1.1 Training with Unbiased Data.** Ideally, we should collect the training data that do not suffer from selection bias. This is equivalent to collect training data under a random recommendation policy  $\pi_{\text{uni}}$  that always recommends a random item in a uniform way, i.e., the random policy should not be aware of the value of  $x$  and  $y$ .

Let the oracle user preference be  $p(\text{click} = 1 | x, y)$ , which is a bernoulli distribution and represents how likely user  $x$  will click item  $y$  if we recommend item  $y$  to the user. Assuming that user  $x$  clicks one item recommended by the random policy  $\pi_{\text{uni}}$ , the click data collected under  $\pi_{\text{uni}}$  will be:

$$p_{\pi_{\text{uni}}}(y|x) = \frac{p(\text{click} = 1 | x, y)}{\sum_{y'} p(\text{click} = 1 | x, y')}, \quad (14)$$

which is the oracle click data distribution for unbiased training, since the uniform random policy  $\pi_{\text{uni}}$  does not introduce any selection bias. Note that  $p_{\pi_{\text{uni}}}(y|x)$  is a multinomial distribution rather than a multivariate bernoulli distribution.

To learn an unbiased recommendation policy, we need to optimize  $p_\theta(y|x)$  on the click data collected under the uniform random policy. In other words, the unbiased loss function should be:

$$R(\theta|x) = - \sum_y p_{\pi_{\text{uni}}}(y|x) \log p_\theta(y|x). \quad (15)$$

**D.1.2 Training with Biased Data: Inverse Propensity Weighting.** However, random exploration is expensive in a real-world recommender system. In practice, the click data  $\mathcal{D}_\pi$  for training a new policy is collected under a biased old recommendation policy  $\pi$ . Let  $q_\pi(y | x)$  be the probability that the old policy recommends item  $y$  to a user in state  $x$ . Note that  $q_\pi(y | x)$  is a multinomial distribution, i.e.  $\sum_{y'} q_\pi(y' | x) = 1$ .

The generating process of the biased dataset  $\mathcal{D}_\pi$  is as follows. Policy  $\pi$  makes a recommendation, i.e. draws a one-hot impression vector  $O$  from the multinomial distribution  $q_\pi(y | x)$ , whose  $y$ th element  $O_y = 1$  if the recommendation is item  $y$ , and  $O_y = 0$  otherwise. On the other hand, user  $x$  is associated with a multi-hot vector  $C$  representing the user's preference regarding all the items, whose  $y$ th element  $C_y$  is drawn from the bernoulli distribution  $p(\text{click} = 1 | x, y)$ ,  $y = 1, 2, 3, \dots, |\mathcal{Y}|$ . Here  $C_y = 1$  if the user will click item  $y$  after being recommended item  $y$ ,  $C_y = 0$  otherwise.

A naïve estimator that does not take selection bias into account typically learns a new biased policy  $p_\theta(y|x)$  by minimizing the following biased loss:

$$\hat{R}_{\text{naive}}(\theta|\mathcal{D}_\pi) = - \sum_y O_y C_y \log p_\theta(y|x), \quad (16)$$

Since  $O$  and  $C$  are independent, the expectation of the above loss is

$$\begin{aligned} \mathbb{E}_{\mathcal{D}_\pi} [\hat{R}_{\text{naive}}(\theta|\mathcal{D}_\pi)] &= - \sum_y \mathbb{E}[O_y] \mathbb{E}[C_y] \log p_\theta(y|x) \\ &= - \sum_y q_\pi(y|x) p(\text{click} = 1 | x, y) \log p_\theta(y|x). \end{aligned}$$

The naïve estimator is thus not optimizing the ideal loss as described in Eq. (15).

Now let us analyze the IPW estimator, which minimizes the following loss:

$$\hat{R}_{\text{IPW}}(\theta|\mathcal{D}_\pi) = - \sum_y \frac{1}{q(y|x)} O_y C_y \log p_\theta(y|x). \quad (17)$$

Note that  $p_{\pi_{\text{uni}}}(y|x) \propto p(\text{click} = 1 | x, y)$  according to Eq. (14). We thus have:

$$\begin{aligned} \mathbb{E}_{\mathcal{D}_\pi} [\hat{R}_{\text{IPW}}(\theta|\mathcal{D}_\pi)] &= - \sum_y \frac{q_\pi(y|x)}{q(y|x)} p(\text{click} = 1 | x, y) \log p_\theta(y|x) \\ &\propto - \sum_y \frac{q_\pi(y|x)}{q(y|x)} p_{\pi_{\text{uni}}}(y|x) \log p_\theta(y|x). \end{aligned}$$

We can see that, if we set  $q(y|x)$  to  $q_\pi(y|x)$ , the IPW estimator is then an unbiased estimation that minimizes the oracle unbiased loss Eq. (15).

**D.1.3 Bernoulli Propensities vs. Multinomial Propensities.** We note that our formulation is different from the existing literature on debiasing a recommender [3, 10, 42, 44, 52, 54]. Most of the existing methods assume that the recommendations  $O$  are drawn from a multivariate bernoulli policy, i.e.,  $O_y$  follows an independent bernoulli distribution  $q_\pi(\text{recommend} = 1 | x, y)$  [42, 52]. Here  $O_y \in \{0, 1\}$  denotes whether item  $y$  is recommended by the policy. We take a different approach and instead assume that  $O$  is drawn from a multinomial policy  $q_\pi(y|x)$  rather than a multivariate bernoulli one.

In other words, the difference lies in whether we should design a candidate generation policy as a multinomial model or as a

multivariate bernoulli model. We choose to design the policy as a multinomial one, due to the following reasons:

- *The multinomial formulation brings superior performance.* There are many empirical results that report better performance with a multinomial candidate generation method than a multivariate bernoulli one, especially with a large set of items for recommendation [14, 26, 32, 33]. Our empirical results in Table 2, where the negative sampling baselines follow the multivariate bernoulli formulation, also verify this finding. Similarly, the natural language processing (NLP) community also report better results with a multinomial classifier, especially when the number of classes increases [37].
- *The multinomial formulation is more consistent with the real-world production environments.* The multivariate bernoulli formulation implicitly assumes that the number of recommendations requested by a user can be modeled as part of the recommendation policy  $\pi$ , e.g., the expectation of the number of recommendations is  $\sum_y q_\pi(\text{recommend} = 1 \mid x, y)$  with the multivariate bernoulli formulation. However, this assumption does not hold in many recommender systems, where the number of recommendations received by a user is decided by the user rather than the system. For example, the user can request more recommendations by scrolling down the page or stop receiving any new recommendation by leaving the page. On the contrary, the multinomial policy  $q_\pi(y \mid x)$  does not attempt to model the number of recommendations requested by a user. Rather, the multinomial policy  $q_\pi(y \mid x)$  only models which one item it should recommend if the user explicitly requests the system to make one recommendation.

## D.2 Proof of Theorem 1

**THEOREM 2.** *The optimal solutions of the contrastive loss (Eq. 3) and the IPW loss (Eq. 4) both minimize the KL divergence from  $p_\theta(y \mid x)$  to  $r(y \mid x) = \frac{p_{\text{data}}(y|x)/q(y|x)}{\sum_{y' \in \mathcal{Y}} p_{\text{data}}(y'|x)/q(y'|x)}$ , if  $p_n(y \mid x)$  is set to be  $q(y \mid x)$ . Here  $p_{\text{data}}(y \mid x)$  is the data distribution, i.e. what is the frequency of  $y$  appearing in  $\mathcal{D}$  given context  $x$ .*

**PROOF.** We now give a proof sketch on the theorem. We will focus on one training instance, i.e. one sequence  $x \in \mathcal{X}$ . The IPW

loss (Eq. 4) for training sample  $x$  is

$$\begin{aligned} - \sum_{y:(x,y) \in \mathcal{D}} \frac{\log p_\theta(y \mid x)}{q(y \mid x)} &\propto - \sum_{y \in \mathcal{Y}} \frac{p_{\text{data}}(y \mid x)}{q(y \mid x)} \log p_\theta(y \mid x) \\ &\propto - \sum_{y \in \mathcal{Y}} r(y \mid x) \log p_\theta(y \mid x) = D_{\text{KL}}(r \parallel p_\theta) + \text{const.w.r.t. } \theta. \end{aligned} \quad (18)$$

The IPW loss is thus minimizing the Kullback–Leibler (KL) divergence from  $p_\theta(y \mid x)$  to  $r(y \mid x)$ .

Let us now focus on the contrastive loss for the training sample  $(x, y)$ . Let  $C = \{y\} \cup \{y_i\}_{i=1}^L$ , where  $y$  is the positive example and  $\{y_i\}_{i=1}^L$  are the  $L$  negative samples drawn from  $q(y \mid x)$ . Note that  $C$  is a multi-set where we allow the same item to appear multiple times. The contrastive loss (Eq. 3) for  $x$  equals to

$$\begin{aligned} - \sum_{y:(x,y) \in \mathcal{D}} \sum_C q(C \mid x, y) \log \frac{\exp(\phi_\theta(x, y))}{\sum_{y' \in C} \exp(\phi_\theta(x, y'))} \\ - \sum_{y \in \mathcal{Y}} \sum_C q(C \mid x, y) p_{\text{data}}(y \mid x) \log \frac{\exp(\phi_\theta(x, y))}{\sum_{y' \in C} \exp(\phi_\theta(x, y'))}, \end{aligned} \quad (19)$$

where  $q(C \mid x, y) = \prod_{i=1}^L q(y_i \mid x)$  if  $y \in C$  or  $q(C \mid x, y) = 0$  if  $y \notin C$ , since by definition  $C$  must include  $y$  if we know that the positive example is  $y$ .

Let  $q(C \mid x) = \prod_{y' \in C} q(y' \mid x)$ . We then have  $q(C \mid x, y) = \frac{q(C \mid x)}{q(y \mid x)}$  if  $C$  includes  $y$ . As a result, we can see that the contrastive loss for training sample  $x$  is proportional to

$$\begin{aligned} - \sum_{y \in \mathcal{Y}} \sum_{C:y \in C} \frac{q(C \mid x)}{q(y \mid x)} p_{\text{data}}(y \mid x) \log \frac{\exp(\phi_\theta(x, y))}{\sum_{y' \in C} \exp(\phi_\theta(x, y'))} \\ = \mathbb{E}_{q(C \mid x)} \left[ - \sum_{y \in C} \frac{p_{\text{data}}(y \mid x)}{q(y \mid x)} \log \frac{\exp(\phi_\theta(x, y))}{\sum_{y' \in C} \exp(\phi_\theta(x, y'))} \right] \\ = \mathbb{E}_{q(C \mid x)} \left[ D_{\text{KL}}(r^C \parallel p_\theta^C) \right] + \text{const.w.r.t. } \theta. \end{aligned} \quad (20)$$

Here we use  $r^C$  and  $p_\theta^C$  to refer to the probability distributions  $r^C(y \mid x) = \frac{p_{\text{data}}(y \mid x)/q(y \mid x)}{\sum_{y' \in C} p_{\text{data}}(y' \mid x)/q(y' \mid x)}$  and  $p_\theta^C(y \mid x) = \frac{\exp(\phi_\theta(x, y))}{\sum_{y' \in C} \exp(\phi_\theta(x, y'))}$ , whose supports are  $C \subset \mathcal{Y}$ . Since we are minimizing the KL divergence under all possible  $C \subset \mathcal{Y}$ , the global optima will be the ones that make  $p_\theta(y \mid x)$  equal to  $r(y \mid x)$  for all  $y \in \mathcal{Y}$  if  $\phi_\theta(x, y)$  is expressive enough to fit the target distribution arbitrarily close. Note that  $\phi_\theta(x, y)$  is indeed expressive enough since we implement it as a neural network, due to the universal approximation theorem [15, 24]. The two losses hence have the same global optima.  $\square$