# Deep Stable Multi-Interest Learning for Out-of-distribution Sequential Recommendation

Qiang Liu, Member, IEEE, Zhaocheng Liu, Zhenxi Zhu, Shu Wu, Senior Member, IEEE, and Liang Wang, Fellow, IEEE

Abstract—Recently, multi-interest models, which extract interests of a user as multiple representation vectors, have shown promising performances for sequential recommendation. However, none of existing multi-interest recommendation models consider the Out-Of-Distribution (OOD) generalization problem, in which interest distribution may change. Considering multiple interests of a user are usually highly correlated, the model has chance to learn spurious correlations between noisy interests and target items. Once the data distribution changes, the correlations among interests may also change, and the spurious correlations will mislead the model to make wrong predictions. To tackle with above OOD generalization problem, we propose a novel multi-interest network, named DEep Stable Multi-Interest Learning (DESMIL), which attempts to de-correlate the extracted interests in the model, and thus spurious correlations can be eliminated. DESMIL applies an attentive module to extract multiple interests, and then selects the most important one for making final predictions. Meanwhile, DESMIL incorporates a weighted correlation estimation loss based on Hilbert-Schmidt Independence Criterion (HSIC), with which training samples are weighted, to minimize the correlations among extracted interests. Extensive experiments have been conducted under both OOD and random settings, and up to 36.8% and 21.7% relative improvements are achieved respectively.

**Index Terms**—Sequential recommendation, multi-interest, out-of-distribution, stable learning.

# 1 Introduction

Sequential recommender systems aim to predict the next item(s) that a user might be interested in based on historical interactions. It has become a vital research topic of recommender systems, in scenarios such as online shopping, online video and restaurant visiting. Given historical behaviors, accurately capturing users' dynamic preferences is the core concern of sequential recommendation. Nowadays, plenty of solutions based on recurrent neural networks [1], [2], convolutional neural networks [3] and attentive networks [4], [5], [6] have been proposed.

For accurate sequential recommendation, we need to look into users' behavior history, and capture their interests on different types of items. Meanwhile, a user usually has multiple interests. For example, on an online video platform, a user may like to watch football videos, funny videos and car videos at the same time. For modeling users' multiple interests, multi-interest networks [7], [8], [9], [10] have been proposed to encode multiple interests with multiple representation vectors. However, existing multi-interest approaches ignore to consider that the interest distribution

• Qiang Liu and Zhaocheng Liu contribute equally to this work.

 Zhaocheng Liu is with Kuaishou Technology, Beijing, China. E-mail: lio.h.zen@gmail.com

 Zhenxi Zhu is with the Department of Computer Science and Technology, Nanjing University, Nanjing, China.
 E-mail: zhuzhenxi@smail.nju.edu.cn in a system is always changing. This brings multi-interest models the Out-Of-Distribution (OOD) problems [11], [12], which may inevitably harm the performances of multi-interest recommenders when data distribution is changing. To make the OOD generalization problem in multi-interest models clear, we present Def. 1-4.

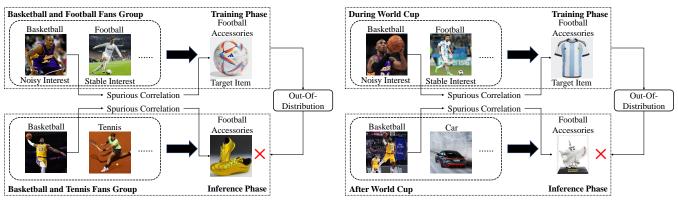
**Definition 1.** (Stable Interests) Given a target item and a user's behavior history, stable interests are those have causal relations to the target item. We denote stable interests as  $(s_0, s_1, s_2, ...)$ , and target item as y. Then, the above process can be denoted as  $(s_0, s_1, s_2, ...) \rightarrow y$ .

**Definition 2.** (Noisy Interests) Given a target item and a user's behavior history, noisy interests are those have no causal relations to the target item. We denote noisy interests as  $(n_0, n_1, n_2, ...)$ . Then, the above process can be denoted as  $(n_0, n_1, n_2, ...) \perp y$ . And there are usually unstable dependencies between stable interests and noisy interests. , which can be denoted as  $(s_0, s_1, s_2, ...) \rightarrow (n_0, n_1, n_2, ...)$ .

**Definition 3.** (Unstable Multi-interest Model) Given a limited training set  $\Omega_{train}$ , for conducting prediction  $\hat{y}$ , an unstable multi-interest model has chance to learn both causal relation  $(s_0, s_1, s_2, ...) \rightarrow \hat{y}$  and spurious correlation  $(n_0, n_1, n_2, ...) \rightarrow \hat{y}$ . Unstable multi-interest models tend to fail in OOD environments. Spurious correlations exist due to the dependencies between stable interests and noisy interests.

**Definition 4.** (Stable Multi-interest Model) Given a limited training set  $\Omega_{train}$ , for conducting prediction  $\hat{y}$ , a stable multi-interest model learns causal relation  $(s_0, s_1, s_2, ...) \rightarrow \hat{y}$ , and eliminates spurious correlations,

Qiang Liu, Shu Wu and Liang Wang are with the Center for Research on Intelligent Perception and Computing (CRIPAC), State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China. E-mail: {qiang.liu, shu.wu, wangliang}@nlpr.ia.ac.cn



(a) Training on data collected from basketball and football fans group, and (b) Training on data collected during World Cup, and inference on data inference on data collected from basketball and tennis fans group.

Fig. 1. We show two examples for demonstrating the OOD generalization problem in multi-interest recommendation models. Conventional models may mistakenly learn spurious correlations between noisy interests and target items (basketball and football accessories in the figure) during training phase, and thus mislead the model predictions during inference phase.

i.e.,  $(n_0,n_1,n_2,...) \perp \hat{y}$ . It is necessary for stable multiinterest models to generalize to different OOD environments.

The shift of interest distribution may be caused by (1) the user distribution is constantly changing with the development of a platform, and new users keep appearing; (2) in the same user group, the interest distribution is also changing, due to changing popularity trends or recommendation strategies. As shown in Fig. 1, we demonstrate some examples. The first example in Fig. 1(a) is about the distribution shift between two different user groups. And the second example in Fig 1(b) is about the distribution shift between the time periods during and after World Cup, which is a popular trend that affects user behaviors. Considering a user usually has relatively similar interests, multiple interests extracted from the user's behavior history tend to be correlated. And online systems, which tend to recommend items similar to the user's historical interests, further aggravate the dependencies between stable interests and noisy interests. With such dependencies, we have chance to learn not only causal relations between stable interests and target items, but also spurious correlations between noisy interests and target items. In both examples in Fig. 1, the dependencies between football and basketball may affect the model training process, and make multiinterest recommendation models learn the spurious correlation between basketball and football accessories. Dependencies between stable interests and noisy interests are unstable, and may change in different data distribution. Once the data distribution changes, i.e., the dependency between football and basketball in the behavior history changes, the spurious correlation will make the model produce wrong predictions during inference phase. Accordingly, we have to study stable multi-interest recommendation models which can generalize to different ODD environments.

According to above analysis and examples, to alleviate the OOD generalization problem in multi-interest recommendation models, we need to remove the dependencies between stable interests and noisy interests, which limits the model to learn the causal relations between stable interests and target items. To do this, Inverse Propensity Weighting (IPW) approaches [13], [14], [15], [16] can be adopted. However, propensity scores are hard to accurately estimate, and the variance of estimation is usually high [17]. Meanwhile, some work attempt to generate counterfactual sample for training causal models [12], [18]. But the quality of counterfactual sample generation is difficult to guarantee, and such methods lack enough explore space. Moreover, it is hard to accurately distinguish stable interests and noisy interests from the multiple interests in the model, which makes the dependencies between stable interests and noisy interests hard to remove. To tackle with above difficulties, we draw lessons from stable learning [19], [20], [21], [22]. Instead of identifying stable interests and noisy interests, we can remove the correlations among all the interests in the model. To achieve this, we can incorporate a correlation estimation loss, which can be optimized together with the main objective.

Formally, in this paper, we propose a novel multiinterest network, named DEep Stable Multi-Interest Learning (DESMIL). (1) DESMIL constructs a multi-interest extractor based on attention [23], and use it to extract multiinterest representations from input user behavior sequence. (2) Then, DESMIL selects the most important interest from the extracted ones as the representation of a user, and use it for making final predictions and constructing the main objective loss for model training. (3) Meanwhile, DESMIL incorporates a weighted correlation estimation loss. To estimate degree of correlations among the extracted interest, we adopt Hilbert-Schmidt Independence Criterion (HSIC) [24], [25] which is a widely-used non-linear independent testing statistic and has been applied for feature de-correlation [26]. We assign a weight for each sample, and obtain weighted HSIC as the weighted correlation estimation loss. DESMIL minimizes the weighted correlation estimation loss via optimizing the sample weights. (4) Then, the sample weights are also added to the main objective loss, and a weighted main objective loss is obtained. That is to say, samples with higher degrees of correlations among interests tend to have lower weights for model training, and vice versa. This makes the

multi-interest model actually being trained on a weighted training dataset, in which correlations among different interests are minimized [27], [28]. (5) Finally, the two losses, i.e., the weighted main objective loss and the weighted correlation estimation loss, are optimized iteratively, until convergence is reached. To this end, it is able for the DESMIL model to better learn the causal relations between stable interests and target items. We have conducted extensive experiments on three real-world datasets under both OOD and random experimental settings, in which DESMIL achieves promising results.

To summarize, the main contributions of this paper are listed as follows:

- We for the first time analyze the OOD generalization problem in multi-interest recommendation models, and propose to de-correlate different interests. This enables the multi-interest recommendation model to learn the causal relations between stable interests and target items, and eliminate spurious correlations between noisy interests and target items.
- We propose a novel DESMIL model, which learns stable representations for sequential recommendation, and makes stable and accurate predictions generalized to OOD environments.
- Extensive experiments show that our proposed DESMIL model outperforms several state-of-the-art sequential recommendation models by a significant margin, especially under the OOD setting.

The rest of the paper is organized as follows. In Section 2, we review some related work on sequential recommendation, deep multi-interest models and stable learning. Then we analyze the causal view of multi-interest recommendation models, and introduce the statistical criterion for nonliner correlation estimation in Section 3. Sections 4 details our proposed DESMIL model. In Section 5, we conduct empirical experiments to verify the effectiveness of DESMIL. Section 6 concludes our work.

## 2 RELATED WORK

In this section, we review some works on sequential recommendation, deep multi-interest models, and stable learning.

#### 2.1 Sequential Recommendation

Modeling users' dynamic preferences from historical behaviors is the core concern of research in sequential recommendation [29], which is a major task in recommender systems. In some traditional models [30], [31], [32], [33], Markov chain and matrix factorization are exploited to model historical behaviors. The most representative model is FPMC [30], which adopts a personalized Markov chain and train the model with a factorization model for capturing collaborative information. In recent years, various deep neural networks such as recurrent neural network [1], [2], [34], [35], [36], convolutional neural networks [3], [37] and attention-based networks [4], [5], [6], [38], [39], [40], [41] have been exploited in deep sequential recommendation models. Target-aware attention for conduction recommendation has also been studied [42], [43]. Recently, contrastive learning has been applied in sequential recommendation [44], [45], [46], [47],

for dealing with sparsity and noise in data. Meanwhile, some works [48], [49], [50] attempt to leverage variety side information for sequential recommendation.

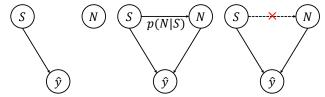
Meanwhile, causal inference have been investigated for increasing the causality and eliminate biases in recommendation [13], [14], [16], [17], [18], [51], [52], [53], [54]. Among them, biases such as exposure bias [13], [14], [53], [55] and popularity bias [17], [56] are widely studied in static recommenders. The user demographic feature shift problem has also been studied with causal representation learning for collaborative filtering [52]. For sequential recommendation, the exposure bias is also studied to deal with the missing-not-at-random problem in the user behavior history [16], [57], [58]. And some works [12], [18], [59] generate counterfactual samples in sequential user behavior history, for training models generalizable to OOD environments.

# 2.2 Deep Multi-interest Models

In real scenarios, a user may have multiple interests in the behavior history, and an overall user preference representation as in most models can hardly grasp the diverse essence of user interests [9], [60]. So, we need to extract multiple interests of a user from the behavior history for better sequential recommendation. There are some work [7], [8], [9], [10], [61], [62] studying how to effectively extract a user's multiple interests in sequential recommendation as multiple vectors. MIND [7] firstly proposes a multi-interest extractor based on the dynamic routing mechanism [63], [64], [65]. As the procedure of dynamic routing can be seen as soft-clustering, the user's historical behaviors can be grouped into different clusters. Meanwhile, a label-aware attention mechanism is proposed to effectively aggregate the multiple user preference representations in training. Besides, ComiRec [8] proposes a controllable multi-interest Framework, in which, both dynamic routing and attentive models can be adopted to extract multiple user interests. Lately, instead of implicitly generating a user's multiple interests by clustering the user behaviors, SINE [9] directly maintains a pool of conceptual prototypes to represent the all set of the user's potential interests. Then a selfattention mechanism is used to decide which prototypes are activated to the user's multiple interests. MGNM [10] proposes to combine multi-interest learning modal and graph convolutional networks. By aggregating multi-level user preferences, MGNM extracts a user's multiple interests more precisely. Meanwhile user-aware candidate matching is also studied in multi-interest models [66].

## 2.3 Stable Learning

The out-of-distribution problem [11] is a common challenge in real-world scenarios, and stable learning has become a successful way to deal with this recently. Stable learning aims to learn a stable predictive model that achieves uniformly good performance on any unknown test data [20]. The framework of most stable learning works can be divided into two steps: sample weight learning and weighted training. Specifically, sample weights are learned to decorrelate features in training data, and then weighted training is conducted to train models on weighted feature distribution, which is an approach to independent identically feature distribution. Along this strand, various de-correlation



(a) Ideal causal dia- (b) Causal diagram of (c) Removing depengram. conventional models. dency  $S \to N$ .

Fig. 2. Causal view analysis of multi-interest recommendation models.

methods [19], [20], [21], [22] have been proposed to learn sample weights and train linear stable models. Moreover, StableNet [27] proposes to adopt random Fourier features to eliminate non-linear dependencies among features in convolutional neural networks. And feature de-correlation in graph neural networks [28], [67], [68] and healthcare [69] has also been studied. Lately, Xu et al. [70] theoretically proves that the stability of least square regression and binary classification can be guaranteed with mutual independence of feature variables under mild conditions.

#### 3 PRELIMINARIES

In this section, we analyze the causal view of multi-interest recommendation models, and introduce the statistical criterion for non-liner correlation estimation.

#### 3.1 Causal View of Multi-interest Models

In Fig. 2, we illustrate the causal diagrams of multiinterest sequential recommendation.  $S = (s_0, s_1, s_2, ...),$  $N=(n_0,n_1,n_2,...)$  and  $\hat{y}$  denote stable interests, noisy interests and prediction of target item respectively. Fig. 2(a) is the ideal causal diagram, in which only the causal relation  $S \to \hat{y}$  exists. However, as shown in Fig. 2(b), dependency  $S \rightarrow N$  exists, and leads conventional multi-interest recommendation models to learn spurious correlation  $N \to \hat{y}$ . The dependency  $S \to N$  is not stable, and p(N|S) may vary in different environments. Once the data distribution changes across training phase and inference phase (validating or testing), there will be  $p_{\text{train}}(N|S) \neq p_{\text{infer}}(N|S)$ , and the spurious correlation  $N \to \hat{y}$  will mislead the model to make wrong predictions. Examples for demonstrating above phenomenon can be found in Fig. 1. Therefore, the path  $N \leftarrow S \rightarrow \hat{y}$  establishes spurious correlation  $N \rightarrow \hat{y}$ . Thus, as shown in Fig. 2(c), we need to remove  $S \to N$ , so that spurious correlation  $N \to \hat{y}$  can be eliminated, and causal relation  $S \rightarrow \hat{y}$  can be accurately learned.

However, it is hard to accurately distinguish stable interests and noisy interests from the multiple interests in the model. So, it is hard to directly perform interest decorrelation between S and N. Inspired by stable learning [19], [20], [21], [22], which de-correlates all the input features via sample re-weighting, we can remove the correlations among all the interests in the recommendation model instead. That is to say, we can estimate the degree of correlation between each pair of interests, and minimize the degree of overall correlations. Moreover, considering the correlation between two interest representation vectors shall

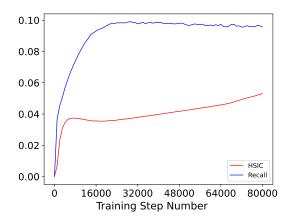


Fig. 3. The curves of HSIC value on the training set and Recall@50 value on the validation set during training ComiRec on the Book dataset. From the latter parts of the curves, we can conclude that, the excessive correlations among interests limit the performance to further increase.

be non-linear, we need a non-linear correlation estimation criterion.

## 3.2 Correlation Estimation Criterion

As we need to minimize correlations among multiple interests in the model, we have to estimate the degree of non-linear correlations between each pair of interests. To achieve this, we adopt HSIC [24], [25], which can estimate non-linear correlation between two variables and has been applied for feature de-correlation [26], as our correlation estimation criterion. HSIC is the Hilbert-Schmidt norm of the cross-covariance operator between the distributions in Reproducing Kernel Hilbert Space (RKHS).

Given two random variables U and V, the formulation of HSIC is:

$$HSIC(U,V) = \mathbb{E}_{uu'vv'}[k_u(u,u')k_v(v,v')] + \mathbb{E}_{uu'}[k_u(u,u')]\mathbb{E}_{vv'}[k_v(v,v')] - 2\mathbb{E}_{uv}[\mathbb{E}_{u'}(k_u(u,u'))\mathbb{E}_{v'}[k_v(v,v')]],$$
(1)

where  $\mathbb{E}_{uu'vv'}$  denotes the expectation over independent pairs (u,v) and (u',v') drawn from P(U,V),  $k_u$  and  $k_v$  are kernel functions. We use the Radial Basis Function (RBF) kernel which is formulated as:

$$k(u,v) = exp\left(-\frac{||u-v||_2^2}{\sigma^2}\right). \tag{2}$$

Given m samples drawn from P(U, V), the Empirical HSIC [24] is defined as

$$HSIC(U,V) = (m-1)^{-2} tr(\mathbf{K}_U \mathbf{P} \mathbf{K}_V \mathbf{P}), \tag{3}$$

where  $\mathbf{K}_U \in \mathbb{R}^{m \times m}$  and  $\mathbf{K}_V \in \mathbb{R}^{m \times m}$  have entries  $\mathbf{K}_{U_{ij}} = k(U_i, U_j)$  and  $\mathbf{K}_{V_{ij}} = k(V_i, V_j)$ ,  $\mathbf{P} = \mathbf{I} - \frac{1}{m} \mathbf{1} \mathbf{1}^{\mathbb{T}} \in \mathbb{R}^{m \times m}$  is the centering matrix, and  $\mathbf{1}$  is an  $(m \times 1)$ -dimensional vector of ones. To be noted, HSIC(U, V) = 0 if and only if  $U \perp V$ .

Furthermore, to investigate the impact of correlations among extracted interests to the training process of multi-interest recommendation models, in Fig. 3, we visualize the change of HSIC value on the training set and Recall@50 value on the validation set when training ComiRec [8] on the Book dataset [71], [72]. The HSIC value is calculated as the sum of non-linear correlations measured by HSIC

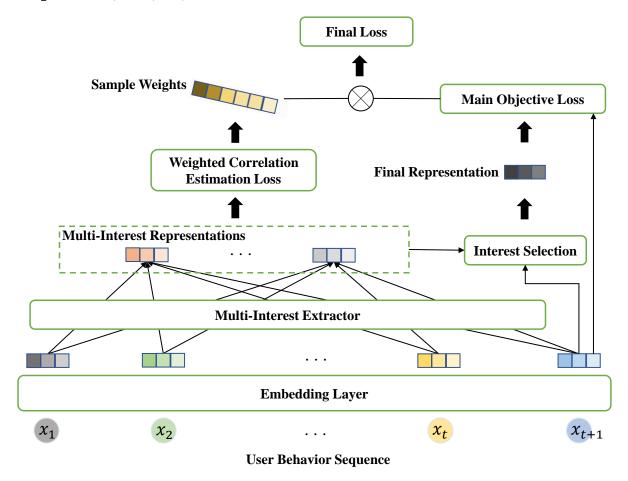


Fig. 4. The overview of our proposed DESMIL model. DESMIL extracts multi-interest representations, and selects the most important one for making final predictions and constructing the main objective loss. Meanwhile, DESMIL incorporates a weighted correlation estimation loss based on HSIC, with which training samples are weighted, to minimize the correlations among extracted interests.

between each pair of interest representations extracted in ComiRec. We can observe from the curves that, after about 2000 steps, the value of HSIC keeps increasing slowly, while the value of recall@50 stops increasing and even begins to decrease. That is to say, in the latter parts of the curves, the excessive correlations among interests limit the performance to further increase. To some extent, for multi-interest models, above observation reveals the trade-off between the correlations among interests and the model performances.

#### 4 METHODOLOGY

In this section, we formulate the problem and introduce the proposed DESMIL model in detail, and the overview of DESMIL is illustrated in Fig. 4.

# 4.1 Problem Formulation

In the setting of sequential recommendation, we have a set of users  $\mathcal{U}=\{u_1,u_2,\cdots,u_{|\mathcal{U}|}\}$  and a universe of items  $\mathcal{I}=\{i_1,i_2,\cdots,i_{|\mathcal{I}|}\}$ . For each user u, given the behavior sequence  $X^u=(x_1^u,x_2^u,\cdots,x_t^u)$  until time step t, we need to predict the user's next interaction  $x_{t+1}^u$ , where  $x_t^u\in\mathcal{I}$ . The goal of sequential recommendation is to recommend to each user a list of items that maximize her/his future needs. Meanwhile, in multi-interest models, there are multiple

interest representations for each user, and we use c to denote the number of interest representation vectors in the model.

#### 4.2 Multi-Interest Representation Extraction

Firstly, we embed items in the behavior sequence into dense vectors. Specifically, given the input sequence  $(x_1^u,\cdots,x_t^u)$ , we create an embedding matrix  $\mathbf{V}\in\mathbb{R}^{|\mathcal{I}|\times d}$  where d is the embedding dimensionality, and retrieve the input embedding matrix by applying the embedding look-up operation. Besides, to make the proposed DESMIL aware of the positions of historical items, we inject the corresponding trainable position embedding matrix [5], [23]  $\mathbf{P}\in\mathbb{R}^{t\times d}$  into the input embedding matrix. The final input embedding matrix  $\mathbf{E}_u^t\in\mathbb{R}^{t\times d}$  can be formulated as

$$\mathbf{E}_{t}^{u} = \begin{bmatrix} \mathbf{V}_{x_{1}^{u}} + \mathbf{P}_{1} \\ \vdots \\ \mathbf{V}_{x_{t}^{u}} + \mathbf{P}_{t} \end{bmatrix}.$$
 (4)

Then, we need a multi-interest extractor to generate multiple representation vectors to capture the diverse interests of a user. Considering the specific interest extraction method is not our main concern, and the main objective of this work is to learn causal representations from multi-interest representations for OOD generalization, we empirically adopt a

# **Algorithm 1** Training process of DESMIL

**Input:** Training set  $\Omega_{train}$ , and maximum training epoch Epoch.

**Output:** Model parameters  $\theta$ .

- 1: Initialize the iteration variable  $q \leftarrow 0$ .
- 2: Initialize the best iteration variable  $q_{best} \leftarrow 0$ .
- 3: Initialize sample weight  $\omega_{u,t}^{(0)}=1.0$ , for all u and t.
- 4: Initialize model parameters  $\theta^{(0)}$  via glorot uniform initializer [75].
- 5: repeat
- Draw a batch of samples  $\mathcal{B}$ . 6:
- $q \leftarrow q+1$ . Keep  $\omega_{\mathcal{B}}^{(q-1)}$  fixed and minimize  $\hat{\mathcal{L}}_{\mathrm{main}}^{(q)}$  via updating  $\theta^{(q)}$ , where  $\hat{\mathcal{L}}_{\mathrm{main}}^{(q)}$  is defined in Eq. (9).
- Keep  $heta^{(q)}$  fixed and update  $\omega^{(q)}_{\mathcal{B}}$  via minimizing  $\mathcal{L}^{(q)}_{\mathrm{corr}}$ as in Eq. (12), where  $\mathcal{L}_{\text{corr}}^{(q)}$  is defined in Eq. (11).
- Update  $q_{best} \leftarrow q$ , if better validation results achieved.
- 11: until Early stopped or maximum training epoch is reached.
- 12: **return**  $\theta^{(q_{best})}$ .

simple attentive module. The attentive matrix  $\mathbf{A}_t^u \in \mathbb{R}^{c \times t}$ can be calculated as

$$\mathbf{A}_{t}^{u} = \operatorname{softmax}(\mathbf{W}_{2} \tanh(\mathbf{W}_{1}(\mathbf{E}_{t}^{u})^{\top})), \tag{5}$$

where  $\mathbf{W}_1 \in \mathbb{R}^{\hat{d} \times d}$  and  $\mathbf{W}_2 \in \mathbb{R}^{c \times \hat{d}}$  are trainable transformation matrices. Then, we obtain the multi-interest representation matrix  $\mathbf{M}^u_t \in \mathbb{R}^{c imes d}$  as

$$\mathbf{M}_{t}^{u} = \mathbf{A}_{t}^{u} \mathbf{E}_{t}^{u}. \tag{6}$$

To this end, for each user, we obtain *c* representation vectors to capture the diverse interests.

Furthermore, we adopt the interest selection strategy [8] to choose the most important interest representation from captured interests to generate the final representation of the whole behavior sequence. Given  $x_{t+1}^u$  to be predicted, and its embedding  $\mathbf{V}_{x_{t+1}^u}$ , we generate the selected representation  $\mathbf{R}_t^u \in \mathbb{R}^{1 \times d}$  as

$$\mathbf{R}_{t}^{u} = \mathbf{M}_{t}^{u}[argmax(\mathbf{M}_{t}^{u}\mathbf{V}_{x_{t+1}^{u}}^{\top}),:]. \tag{7}$$

Finally, for a batch of samples  $\mathcal{B}$  drawn from the training set  $\Omega_{train}$ , the main objective loss is formulated as

$$\mathcal{L}_{\text{main}} = -\sum_{(u,t) \in \mathcal{B}} log \left( \frac{exp(\mathbf{R}_t^u \mathbf{V}_{x_{t+1}^{-1}}^\top)}{\sum_{i \in \mathcal{I}} exp(\mathbf{R}_t^u \mathbf{V}_i^\top)} \right), \quad (8)$$

which can be implemented by the sampled softmax technique [73], [74] considering computational efficiency.

# De-correlation among Multi-Interest Representations

As discussed in Sec. 3, we adopt sample re-weighting techniques [19], [21], [27] for de-correlation among multi-interest representations. We propose an interest de-correlation regularizer that aims to estimate a weight for each sample, so that the degree of correlations among multiple interests can be minimized. Specifically, we assign a weight  $\omega_{u,t}$  for each sample. We use  $\omega_{u,t}^{(q)}$  to denote the sample weight after

calculation of the q-th training epoch, and the initial sample weight as  $\omega_{u,t}^{(0)} = 1.0$ .

For de-correlation among multi-interest representations, we re-weight the main objective losses of samples in Eq. (8), and obtain the revised weighted main objective loss at

$$\hat{\mathcal{L}}_{\text{main}}^{(q)} = -\sum_{(u,t)\in\mathcal{B}} \omega_{u,t}^{(q-1)} log\left(\frac{exp(\mathbf{R}_t^u \mathbf{V}_{x_{t+1}^\top}^\top)}{\sum_{i\in\mathcal{I}} exp(\mathbf{R}_t^u \mathbf{V}_i^\top)}\right), \quad (9)$$

which takes the samples weights optimized in the last epoch for re-weighting the model training process, and is illustrated as the final loss in Fig. 4.

Meanwhile, we need to estimate the sample weights via minimizing the degree of correlations among multiple interests. Taking the weights from the last epoch, the multiinterest representation of user u after time step t is reweighted as

$$\hat{\mathbf{M}}_{u,t}^{(q)} = \omega_{u,t}^{(q-1)} \mathbf{M}_t^u. \tag{10}$$

Then, we propose a weighted correlation estimation loss, which is based on HSIC introduced in Sec. 3.2 and used to estimate the degree of correlations among multiple interests,

$$\mathcal{L}_{\text{corr}}^{(q)} = \sum_{(u,t)\in\mathcal{B}} \sum_{j} \sum_{k} \lambda HSIC(\hat{\mathbf{M}}_{u,t}^{(q)}[j,:], \hat{\mathbf{M}}_{u,t}^{(q)}[k,:]), \quad (11)$$

where  $\lambda$  is the de-correlation importance that controls the learning process of sample weights. Then, via minimizing the degree of correlations, we optimize sample weights as

$$\omega_{\mathcal{B}}^{(q)} = \operatorname{argmin} \mathcal{L}_{\operatorname{corr}}^{(q)}.$$
 (12)

Furthermore, we alternatively minimize the weighted main objective loss  $\hat{\mathcal{L}}_{\mathrm{main}}^{(q)}$  with respect to sample weights  $\omega^{(q)}$ , and minimize the weighted correlation estimation loss  $\mathcal{L}_{\mathrm{corr}}^{(q)}$  with respect to model parameters  $\theta^{(q)}$ . Meanwhile, the detailed procedure of our proposed DESMIL model is shown in Alg. 1.

Via the above process of training, samples with higher degrees of correlations among interests will have lower weights for model training in the main objective loss, and vice versa. This makes the final model actually being trained on a weighted training dataset, in which correlations among different interests are minimized according to the loss in Eq. (11). To be noted, the sample weights are only optimized with samples in the training set, and can make the final model potentially generalize to OOD environments. The inference procedure of DESMIL is the same as conventional multi-interest model, and we do not need to estimate sample weights for samples in the validation set or the testing set.

## **EXPERIMENTS**

In this section, we perform extensive experiments to evaluate the performance of our proposed DESMIL model. We need to answer the following Research Questions (RQ):

- **RQ1:** How is the performance of DESMIL under outof-distribution environments?
- **RQ2:** How is the performance of DESMIL under ordinary in-distribution environments?

TABLE 1

Results under OOD data splitting evaluated by different metrics (%). Best performances are indicated by bold font and the strongest baselines are underlined. The improvements indicate the relative increase of DESMIL over the best baselines.

Dataset	Metric	GRU4Rec	SASRec	MIND	ComiRec	SINE	MGNM	USR	CauseRec	DESMIL	Improv.
	Recall@20	3.24	4.95	5.36	5.60	5.54	5.32	5.42	5.38	7.16	27.86%
	Recall@50	5.84	6.78	7.10	<u>7.96</u>	7.56	7.35	7.65	7.71	10.89	36.81%
	NDCG@20	2.87	3.06	3.13	3.20	3.41	3.46	3.55	3.25	4.25	19.72%
Book	NDCG@50	3.15	3.54	3.96	3.78	4.22	4.15	4.30	4.08	5.30	23.26%
	HR@20	6.93	8.25	9.98	10.25	10.17	10.26	10.31	10.48	12.42	18.51%
	HR@50	12.10	13.56	15.98	15.84	15.86	15.74	15.48	<u>16.04</u>	18.27	13.90%
	Recall@20	10.87	12.34	13.14	13.65	13.64	13.79	13.46	13.90	14.62	5.18%
	Recall@50	14.08	15.01	16.18	16.57	16.68	17.17	16.50	17.33	18.65	7.62%
	NDCG@20	9.87	11.00	11.99	12.71	12.53	12.61	12.25	12.91	13.87	7.44%
Movies and TV	NDCG@50	11.85	12.72	13.54	13.92	13.77	13.76	13.16	14.40	15.44	7.22%
	HR@20	19.25	20.11	22.36	23.32	23.27	23.17	22.88	23.44	23.77	1.41%
	HR@ 50	25.20	26.35	28.56	29.84	29.71	29.75	29.10	<u>30.05</u>	31.00	3.16%
	Recall@20	4.15	6.37	7.16	7.26	7.33	7.25	6.96	7.48	8.20	9.63%
	Recall@50	5.88	8.80	9.92	10.26	10.18	10.20	10.03	10.50	11.41	8.67%
	NDCG@20	4.31	5.26	5.90	5.82	6.02	5.92	5.80	6.15	6.82	10.89%
CDs and Vinyl	NDCG@50	5.04	6.17	6.85	6.35	6.78	6.54	6.64	6.92	7.67	10.84%
	HR@20	7.21	10.57	12.35	12.23	12.28	12.26	11.87	12.16	13.45	8.91%
	HR@50	10.87	15.60	17.36	17.05	17.34	17.24	17.10	17.15	19.36	10.63%

- RQ3: Is the DESMIL model sensitive to some important hyper-parameters?
- RQ4: How does the DESMIL model affect the training process?
- RQ5: How is the distribution of sample weights learned in the DESMIL model?

#### 5.1 Experimental Datasets

We conduct experimental comparison on three public datasets collected from real-world scenarios:

- **Book Dataset**. The Book dataset is part of the Amazon Product Data<sup>1</sup> [71], [72] in the "book" category. There are 603,668 users, 367,982 items, and 8,898,041 user behaviors in total.
- **Movies and TV Dataset**. The Movies and TV dataset is part of the updated version of Amazon Review Data<sup>2</sup> [76]. There are 304,763 users, 89,590 items, and 3,506,470 user behaviors in total.
- **CDs and Vinyl Dataset**. The CDs and Vinyl dataset is also part of the updated Amazon Review Data. There are 129,237 users, 145,522 items, and 1,682,049 user behaviors in total.

Moreover, we need to conduct performance comparison not only in OOD environments for verifying OOD generalization ability, but also in in-distribution environments for verifying ordinary recommendation ability. Thus, we perform two different data splitting:

(1) The first splitting is **OOD data splitting**. As shown in examples in Fig. 1, different user groups and popularity tends result in data distribution shift, and affect the recommendation performances. Considering popularity tends are hard to identify in an offline dataset, we construct OOD data according to different user groups. In practice, we use the Jaccard similarity<sup>3</sup> of items from different users to

TABLE 2
Comparison Among DESMIL, MGNM+PW and ComiRec+PW under OOD data splitting evaluated by Recall@50 (%). Best performances are indicated by bold font.

Approach	Book	Movies and TV	CDs and Vinyl			
ComiRec+PW	8.63	17.16	10.25			
MGNM+PW	8.14	17.54	10.33			
DESMIL	10.89	18.65	11.41			

measure the similarity between users. We randomly select a user, and then iteratively select the next user with the maximum similarity to the selected users, until 50% users are selected. We obtain a set of selected users  $\mathcal{U}_1$  and a set of remaining users  $\mathcal{U}_2$ . In this way,  $\mathcal{U}_1$  and  $\mathcal{U}_2$  share extremely different distributions, in which  $\mathcal{U}_1$  can be viewed as the OOD environment of  $\mathcal{U}_2$ . Then, we randomly use 10% in  $\mathcal{U}_1$ , 10% in  $\mathcal{U}_2$  and other non-overlapping 80% in  $\mathcal{U}_2$  as the testing set, the validation set and the training set respectively. That is to say, the number of samples for training, validation and testing confirms to 8:1:1.

(2) The second splitting is **random data splitting**. We use the same splitting in previous work [8], in which samples are randomly split into training, validation and testing sets.

#### 5.2 Compared Baselines

We compare our proposed DESMIL model to the following baselines for evaluation:

- **GRU4Rec** [2]: a classic sequential recommendation model based on recurrent neural network.
- SASRec [5]: a state-of-the-art model that uses selfattention network for the sequential recommendation.
- MIND [7]: a classic multi-interest sequential network with dynamic routing for modeling users' diverse interests in the matching stage.
- ComiRec [8]: a state-of-the-art sequential network with multi-interest extraction module to generate

<sup>1.</sup> http://jmcauley.ucsd.edu/data/amazon/

<sup>2.</sup> https://nijianmo.github.io/amazon/index.html

<sup>3.</sup> https://www.learndatasci.com/glossary/jaccard-similarity/

TABLE 3

Results under random data splitting evaluated by different metrics (%). Best performances are indicated by bold font and the strongest baselines are underlined. The improvements indicate the relative increase of DESMIL over the best baselines.

Dataset	Metric	GRU4Rec	SASRec	MIND	ComiRec	SINE	MGNME	USR	CauseRec	DESMIL	Improv.
	Recall@20	3.47	4.76	5.10	5.92	5.46	6.18	5.75	5.24	7.52	21.68%
	Recall@50	6.50	7.78	7.64	9.35	8.72	<u>9.64</u>	8.60	9.36	11.06	14.73%
	NDCG@20	3.55	4.84	5.09	4.17	4.83	4.88	4.93	4.66	5.46	7.27%
Book	NDCG@50	4.42	5.74	5.97	5.47	6.04	6.19	5.95	<u>6.28</u>	7.24	15.29%
	HR@20	7.84	8.82	10.59	11.70	11.87	<u>12.70</u>	11.28	12.45	14.86	17.01%
	HR@50	12.38	13.79	15.56	18.04	18.94	20.21	17.89	20.23	21.53	6.43%
	Recall@20	13.20	14.43	14.87	15.46	15.16	<u>15.50</u>	14.76	15.30	15.76	1.68%
	Recall@50	17.66	18.27	<u>19.55</u>	18.87	19.30	19.14	19.27	19.24	20.90	6.91%
	NDCG@20	15.07	14.49	<u>15.80</u>	14.73	15.57	15.44	14.90	15.10	15.54	-
Movies and TV	NDCG@50	16.21	16.72	<u>17.23</u>	16.17	16.64	16.68	16.81	16.83	17.36	0.75%
	HR@20	22.67	23.25	25.34	25.87	25.12	25.76	24.75	<u>25.94</u>	26.42	1.85%
	HR@50	29.54	30.43	32.93	33.68	33.30	<u>33.95</u>	32.90	33.90	34.80	2.50%
	Recall@20	4.39	6.92	7.55	7.96	7.69	8.03	7.35	7.77	8.75	8.97%
	Recall@50	6.07	8.52	10.32	11.23	10.93	11.32	10.57	11.12	12.09	6.80%
	NDCG@20	4.81	6.44	7.93	6.84	7.28	7.46	7.12	7.51	7.86	-
CDs and Vinyl	NDCG@50	5.42	7.10	8.88	8.01	8.34	8.68	8.18	8.57	8.91	0.34%
	HR@20	8.47	12.86	14.28	14.35	14.55	14.60	14.08	14.49	15.73	7.74%
	HR@50	11.79	16.29	19.38	20.26	20.71	20.83	19.80	20.66	21.89	5.09%

multiple user interests and aggregation module to obtain top-N items. We use the SA setting of ComiRec which is described as ComiRec-SA in the original paper.

- SINE [9]: a state-of-the-art multi-interest model that maintains a pool of conceptual prototypes to represent the all set of a user's potential interests, and uses self-attention to decide which prototypes are activated to the user's multiple interests.
- MGNM [10]: a state-of-the-art multi-interest model that combines with graph convolutional networksbased recommenders.
- USR [16]: a state-of-the-art debiasing sequential recommendation model with latent confounders in an IPW-based framework, for dealing with exposure bias in user interaction history.
- CauseRec [12]: a state-of-the-art sequential network that performs contrastive user representation learning to model the counterfactual data distribution for generalizing to OOD environments.

We compare above baselines with DESMIL under both OOD and random data splitting. Moreover, we additionally consider an IPW-based debiasing method called Permutation Weighting (PW) [15]. We perform PW on two representative multi-interest models ComiRec and MGNM, then obtain ComiRec+PW and MGNM+PW for performance comparison with DESMIL in OOD environments.

#### 5.3 Experimental Settings

In this subsection, we introduce some details of our experimental settings.

## 5.3.1 Parameter Configuration

The embedding size of items is 64. According to the best performances of ComiRec, the batch size for the Book dataset is 1024, while for the other two datasets is 128. The number of negative samples for sampled softmax loss is 10. All models use early stopping based on the Recall@50

on the validation set. The de-correlation importance and the number of interests are tuned in the range of  $\{0.01,0.1,1.0,10.0,100.0\}$  and the range of  $\{2,4,6,8\}$ , respectively. We use the Adam optimizer with learning rate lr=0.001 for optimization.

#### 5.3.2 Evaluation Metrics

We use the top-p Recall, Normalized Discounted Cumulative Gain (NDCG), and Hit Rate (HR) to evaluate performances of all the compared models. We select p=20,50 in our experiments. The three metrics measure the model performances with different criteria. Recall@p is defined as the fraction of relevant items found in the top p recommended items. NDCG@p further considers the normalization of gains and the ranking of correctly recommended items, where items with higher relevance affect the final score more. HR@p is defined as the proportion of top p recommended items found in the testing set.

#### 5.4 Results under OOD Data Splitting (RQ1)

In Tab. 1, we illustrate the experimental comparison under the OOD data splitting. Among the compared baselines, CauseRec performs the best, achieves best performances on 12 out of 18 target metrics. This shows that, to a certain extent, training with counterfactual samples can improve the model's OOD generalization ability. USR performs better than single-interest models, and is competitive with multiinterest models. Considering the recommendation backbone in USR is a simple single-interest GRU, this shows the effectiveness of IPW-based approaches. Among multiinterest models, simple models, i.e., MIND and ComiRec, seem to have relatively better performances. This indicates that, models with simpler structures tend to perform better in different OOD environments. Moreover, our proposed DESMIL model clearly outperforms all the compared baselines by large margins.

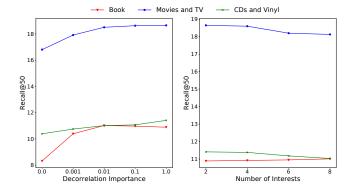


Fig. 5. Hyper-parameter study of de-correlation importance coefficient and number of interests under OOD data splitting measured by Recall(0.50) (%).

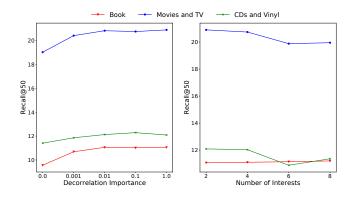


Fig. 6. Hyper-parameter study of de-correlation importance coefficient and number of interests under random data splitting measured by Recall(50) (%).

Meanwhile, we conduct comparison among DESMIL, MGNM+PW and ComiRec+PW under the OOD data splitting in Tab. 2. We can observe that, among the compared baselines, ComiRec+PW performs better on the Book dataset, and MGNM+PW performs better on the CDs and Vinyl dataset and the Movies and TV dataset. Compared with results in Tab. 1, MGNM+PW and ComiRec+PW perform better than MGNM and ComiRec respectively, and are competitive with CauseRec. This further shows the effectiveness of IPW-based approaches [13], [14], [15] for OOD generalization. And DESMIL still significantly outperforms above two PW-augmented approaches. These results and observations strongly demonstrate the effectiveness and stability of DESMIL in OOD environments.

## 5.5 Results under Random Data Splitting (RQ2)

The experimental comparison under the random data splitting is shown in Tab. 3. We can observe that, results in Tab. 1 are commonly lower than those in Tab. 3, which shows distribution shift leads to more difficult tasks. Overall speaking, attention-based SASRec performs better than RNN-based GRU4Rec, and multi-interest models have better performances than both of them. With the help of graph convolutional networks, MGNM is the best one among the compared multi-interest models. Meanwhile, our proposed DESMIL model constantly outperforms all the compared baselines, except evaluated by NDCG@20 on the Movies

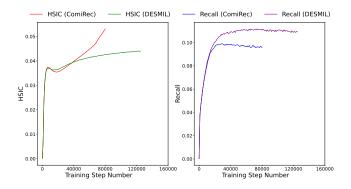


Fig. 7. The curves of HSIC on the training set and Recall@50 on the validation set when training ComiRec and DESMIL on Book. Compared with ComiRec, DESMIL shows slighter correlations among multiple interests measured by HSIC, and better performance measured by Recall@50.

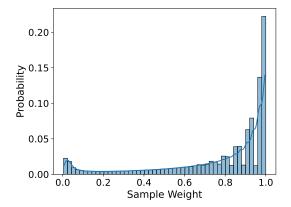


Fig. 8. The distribution of sample weights in the DESMIL model trained on the Book dataset.

and TV dataset and the CDs and Vinyl dataset. On the Book dataset, and evaluated by Recall and HR, the improvements of DESMIL are significant. In real applications, Recall is often considered the most important metric as it can best reflect the model performance facing an enormous candidate set of items and almost equally important but limited exposure positions. These results further demonstrate the effectiveness of our proposed DESMIL model. Moreover, comparing with results in Tab. 1 and Tab. 3, improvements achieved by DESMIL under OOD settings are much more significant than those under random settings. This indicates that, DESMIL is a multi-interest model suitable to OOD generalization, but still has great performances under regular data splitting.

## 5.6 Hyper-parameter Sensitivity (RQ3)

To investigate the stability of our proposed model to the hyper-parameters, we conduct hyper-parameter sensitivity study. In Fig. 5 and Fig. 6, we illustrate the sensitivity of decorrelation importance coefficient  $\lambda$  and number of interests c, under the OOD data splitting and the random data splitting respectively. Results in the figures are evaluated by the Recall@50 metric. We can observe that, c does not affect the performances of DESMIL very much. Moreover, when  $\lambda=0.0$ , performances of DESMIL drop significantly. And  $\lambda=0.0$  indicates DESMIL without interest de-correlation,

which can also be viewed as the ablation study. This shows the importance of interest de-correlation in DESMIL. Meanwhile, when  $\lambda \in [0.01, 0.1, 1.0]$ , the performances of DESMIL stay relatively stable, which shows that we do not have too much burden for hyper-parameter tuning in practice. In our other experiments, we simply set  $\lambda = 1.0$  and c = 2.

# 5.7 Visualization (RQ4 and RQ5)

In Figure 7, we visualize the change of HSIC on the training set, and Recall on the validation set when training DESMIL and ComiRec on the Book dataset. Both DESMIL and ComiRec use early stopping and their training terminates at different steps, which results in the different lengths of curves shown in the figure. To be noted, DESMIL performs optimization of HSIC by sample re-weighting in the training phase, while the calculation of HSIC, which is shown in the figure, is not weighted. Different from DESMIL, ComiRec does not control the optimization of HSIC, i.e., correlations among interests, during training. During the first 10000 steps, the HSIC and Recall of both models quickly increase. Then, the HSIC of ComiRec continues to increase rapidly, while the HSIC of DESMIL grows relatively smoothly. This makes it possible for DESMIL to update more steps and obtain better performances. In a word, via minimizing the weighted correlation estimation loss based on HSIC in Eq. (11), we can break the trade-off relation between the correlations among interests and the model performances introduced in Fig. 3, and alleviate the dependencies between stable interests and noisy interests which may mislead the model to learn spurious correlations.

Moreover, in Figure 8, we illustrate the probability distribution of sample weights leaned in DESILL on the Book dataset, in the form of histograms. Sample weights in the figure are mostly in the values from 0.8 to 1.0, with some located near the value of 0.0. The values near 1.0 indicate slight changes in sample weights, while the values near 0.0 indicate sharp changes of sample weights in the main objective loss function. This shows that, most samples in the Book dataset do not require specific de-correlation operations, while a small part of samples are indeed marginalized.

#### 6 Conclusion

In this paper, we investigate the OOD generalization problem in multi-interest models, for accurate and stable sequential recommendation. To achieve this, we propose a novel multi-interest recommendation model called DESMIL. DESMIL uses an attentive module to extract multiple interests, and selects the most important one for conducting final predictions. Then, a weighted correlation estimation loss is incorporated, to alleviate the correlations among different extracted interests in the training set. The DESMIL model can learn stable representations in sequential recommendation, and make stable and accurate predictions generalized to OOD environments. Extensive experimental results under both OOD and random settings strongly demonstrate that our proposed DESMIL model is a promising sequential recommendation model.

# **REFERENCES**

- [1] F. Yu, Q. Liu, S. Wu, L. Wang, and T. Tan, "A dynamic recurrent model for next basket recommendation," in *Proceedings of the International ACM SIGIR conference on Research and Development in Information Retrieval*, 2016, pp. 729–732.
- [2] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk, "Session-based recommendations with recurrent neural networks," 2016.
- J. Tang and K. Wang, "Personalized top-n sequential recommendation via convolutional sequence embedding," in Proceedings of the ACM International Conference on Web Search and Data Mining, 2018, pp. 565–573.
- [4] Q. Liu, Y. Zeng, R. Mokhosi, and H. Zhang, "Stamp: short-term attention/memory priority model for session-based recommendation," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 1831–1839.
- [5] W.-C. Kang and J. McAuley, "Self-attentive sequential recommendation," in *IEEE International Conference on Data Mining*, 2018, pp. 197–206.
- [6] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang, "Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer," in *Proceedings of the ACM International* Conference on Information & Knowledge Management, 2019, pp. 1441– 1450.
- [7] C. Li, Z. Liu, M. Wu, Y. Xu, H. Zhao, P. Huang, G. Kang, Q. Chen, W. Li, and D. L. Lee, "Multi-interest network with dynamic routing for recommendation at tmall," in *Proceedings of the ACM International Conference on Information & Knowledge Management*, 2019, pp. 2615–2623.
- [8] Y. Cen, J. Zhang, X. Zou, C. Zhou, H. Yang, and J. Tang, "Controllable multi-interest framework for recommendation," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 2942–2951.
- [9] Q. Tan, J. Zhang, J. Yao, N. Liu, J. Zhou, H. Yang, and X. Hu, "Sparse-interest network for sequential recommendation," in Proceedings of the ACM International Conference on Web Search and Data Mining, 2021, pp. 598–606.
- [10] Y. Tian, J. Chang, Y. Niu, Y. Song, and C. Li, "When multi-level meets multi-interest: A multi-grained neural model for sequential recommendation," in *Proceedings of the International ACM SIGIR* conference on Research and Development in Information Retrieval, 2022, pp. 1632–1641
- [11] Z. Shen, J. Liu, Y. He, X. Zhang, R. Xu, H. Yu, and P. Cui, "Towards out-of-distribution generalization: A survey," *arXiv preprint arXiv:2108.13624*, 2021.
- [12] S. Zhang, D. Yao, Z. Zhao, T.-S. Chua, and F. Wu, "Causerec: Counterfactual user sequence synthesis for sequential recommendation," in *Proceedings of the International ACM SIGIR conference on Research and Development in Information Retrieval*, 2021, pp. 367–377.
- [13] T. Schnabel, A. Swaminathan, A. Singh, N. Chandak, and T. Joachims, "Recommendations as treatments: Debiasing learning and evaluation," in *International Conference on Machine Learning*, 2016, pp. 1670–1679.
- [14] X. Wang, R. Zhang, Y. Sun, and J. Qi, "Doubly robust joint learning for recommendation on data missing not at random," in *International Conference on Machine Learning*, 2019, pp. 6638–6647.
- [15] D. Arbour, D. Dimmery, and A. Sondhi, "Permutation weighting," in *International Conference on Machine Learning*. PMLR, 2021, pp. 331–341.
- [16] Z. Wang, S. Shen, Z. Wang, B. Chen, X. Chen, and J.-R. Wen, "Unbiased sequential recommendation with latent confounders," in *Proceedings of the ACM Web Conference*, 2022, pp. 2195–2204.
- [17] Y. Zhang, F. Feng, X. He, T. Wei, C. Song, G. Ling, and Y. Zhang, "Causal intervention for leveraging popularity bias in recommendation," in *Proceedings of the International ACM SIGIR conference on Research and Development in Information Retrieval*, 2021, pp. 11–20.
- [18] Z. Wang, J. Zhang, H. Xu, X. Chen, Y. Zhang, W. X. Zhao, and J.-R. Wen, "Counterfactual data-augmented sequential recommendation," in *Proceedings of the International ACM SIGIR conference on Research and Development in Information Retrieval*, 2021, pp. 347–356.
- [19] Z. Shen, P. Cui, T. Zhang, and K. Kunag, "Stable learning via sample reweighting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 5692–5699.
- [20] K. Kuang, P. Cui, S. Athey, R. Xiong, and B. Li, "Stable prediction across unknown environments," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 1617–1626.

- [21] K. Kuang, R. Xiong, P. Cui, S. Athey, and B. Li, "Stable prediction with model misspecification and agnostic distribution shift," in Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 4485–4492.
- [22] K. Kuang, H. Zhang, R. Wu, F. Wu, Y. Zhuang, and A. Zhang, "Balance-subsampled stable prediction across unknown test data," ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 16, no. 3, pp. 1–21, 2021.
- [23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.
- [24] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf, "Measuring statistical dependence with hilbert-schmidt norms," in *The Annual Conference on Learning Theory*, 2005, pp. 63–77.
- [25] A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, A. J. Smola et al., "A kernel statistical test of independence." in Advances in Neural Information Processing Systems, 2007, pp. 585–592.
- [26] H. Bahng, S. Chun, S. Yun, J. Choo, and S. J. Oh, "Learning debiased representations with biased representations," in *Interna*tional Conference on Machine Learning, 2020, pp. 528–539.
- [27] X. Zhang, P. Cui, R. Xu, L. Zhou, Y. He, and Z. Shen, "Deep stable learning for out-of-distribution generalization," in *Proceedings of* the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5372–5382.
- [28] S. Fan, X. Wang, C. Shi, P. Cui, and B. Wang, "Generalizing graph neural networks on out-of-distribution graphs," *arXiv preprint arXiv:2111.10657*, 2021.
- [29] H. Fang, D. Zhang, Y. Shu, and G. Guo, "Deep learning for sequential recommendation: Algorithms, influential factors, and evaluations," ACM Transactions on Information Systems (TOIS), vol. 39, no. 1, pp. 1–42, 2020.
- [30] S. Rendle, C. Freudenthaler, and L. Schmidt-Thieme, "Factorizing personalized markov chains for next-basket recommendation," in Proceedings of the ACM Web Conference, 2010, pp. 811–820.
- [31] R. He and J. McAuley, "Fusing similarity models with markov chains for sparse sequential recommendation," in *IEEE Interna*tional Conference on Data Mining, 2016, pp. 191–200.
- [32] R. He, W.-C. Kang, and J. McAuley, "Translation-based recommendation," in *Proceedings of the ACM Conference on Recommender* Systems, 2017, pp. 161–169.
- [33] B. Hidasi and D. Tikk, "General factorization framework for context-aware recommendations," *Data Mining and Knowledge Dis*covery, vol. 30, no. 2, pp. 342–371, 2016.
- [34] Q. Liu, S. Wu, L. Wang, and T. Tan, "Predicting the next location: A recurrent model with spatial and temporal contexts," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016.
- [35] Q. Liu, S. Wu, D. Wang, Z. Li, and L. Wang, "Context-aware sequential recommendation," in *IEEE International Conference on Data Mining*, 2016, pp. 1053–1058.
- [36] Q. Liu, S. Wu, and L. Wang, "Multi-behavioral sequential prediction with recurrent log-bilinear model," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 6, pp. 1254–1267, 2017.
- [37] J. Wang, Q. Liu, Z. Liu, and S. Wu, "Towards accurate and interpretable sequential prediction: A cnn & attention-based feature extractor," in *Proceedings of the ACM International Conference on Information & Knowledge Management*, 2019, pp. 1703–1712.
- [38] Y. Luo, Q. Liu, and Z. Liu, "Stan: Spatio-temporal attention network for next location recommendation," in *Proceedings of the ACM Web Conference*, 2021, pp. 2177–2185.
- [39] J. Li, Y. Wang, and J. McAuley, "Time interval aware self-attention for sequential recommendation," in *Proceedings of the ACM International Conference on Web Search and Data Mining*, 2020, pp. 322–330.
- [40] C. Hsu and C.-T. Li, "Retagnn: Relational temporal attentive graph neural networks for holistic sequential recommendation," in *Proceedings of the ACM Web Conference*, 2021, pp. 2968–2979.
- [41] Z. Fan, Z. Liu, Y. Wang, A. Wang, Z. Nazari, L. Zheng, H. Peng, and P. S. Yu, "Sequential recommendation via stochastic self-attention," in *Proceedings of the ACM Web Conference*, 2022, pp. 2036–2047.
- [42] G. Zhou, X. Zhu, C. Song, Y. Fan, H. Zhu, X. Ma, Y. Yan, J. Jin, H. Li, and K. Gai, "Deep interest network for click-through rate prediction," in *Proceedings of the ACM SIGKDD International Con*ference on Knowledge Discovery & Data Mining, 2018, pp. 1059–1068.
- [43] G. Zhou, N. Mou, Y. Fan, Q. Pi, W. Bian, C. Zhou, X. Zhu, and K. Gai, "Deep interest evolution network for click-through rate prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, pp. 5941–5948.

- [44] K. Zhou, H. Wang, W. X. Zhao, Y. Zhu, S. Wang, F. Zhang, Z. Wang, and J.-R. Wen, "S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization," in *Proceedings of the ACM International Conference on Information & Knowledge Management*, 2020, pp. 1893–1902.
- [45] X. Xie, F. Sun, Z. Liu, S. Wu, J. Gao, B. Ding, and B. Cui, "Contrastive learning for sequential recommendation," arXiv preprint arXiv:2010.14395, 2020.
- [46] Z. Liu, Y. Chen, J. Li, P. S. Yu, J. McAuley, and C. Xiong, "Contrastive self-supervised sequential recommendation with robust augmentation," arXiv preprint arXiv:2108.06479, 2021.
- [47] Y. Chen, Z. Liu, J. Li, J. McAuley, and C. Xiong, "Intent contrastive learning for sequential recommendation," in *Proceedings of the ACM Web Conference*, 2022, pp. 2172–2182.
- [48] Q. Cui, S. Wu, Q. Liu, W. Zhong, and L. Wang, "Mv-rnn: A multiview recurrent neural network for sequential recommendation," IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 2, pp. 317–331, 2018.
- pp. 317–331, 2018.
  [49] C. Liu, X. Li, G. Cai, Z. Dong, H. Zhu, and L. Shang, "Non-invasive self-attention for side information fusion in sequential recommendation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, pp. 4249–4256.
- [50] Y. Xie, P. Zhou, and S. Kim, "Decoupled side information fusion for sequential recommendation," in *Proceedings of the International* ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 1611–1621.
- [51] S. Mu, Y. Li, W. X. Zhao, J. Wang, B. Ding, and J.-R. Wen, "Alleviating spurious correlations in knowledge-aware recommendations through counterfactual generator," in *Proceedings of the International ACM SIGIR conference on Research and Development in Information Retrieval*, 2022, pp. 1401–1411.
- [52] W. Wang, X. Lin, F. Feng, X. He, M. Lin, and T.-S. Chua, "Causal representation learning for out-of-distribution recommendation," in *Proceedings of the ACM Web Conference*, 2022, pp. 3562–3571.
- [53] H. Wang, T.-W. Chang, T. Liu, J. Huang, Z. Chen, C. Yu, R. Li, and W. Chu, "Escm2: Entire space counterfactual multi-task model for post-click conversion rate estimation," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 363–372.
- [54] W. Wang, F. Feng, X. He, H. Zhang, and T.-S. Chua, "Clicks can be cheating: Counterfactual recommendation for mitigating clickbait issue," in Proceedings of the International ACM SIGIR conference on Research and Development in Information Retrieval, 2021, pp. 1288– 1297.
- [55] S. Guo, L. Zou, Y. Liu, W. Ye, S. Cheng, S. Wang, H. Chen, D. Yin, and Y. Chang, "Enhanced doubly robust learning for debiasing post-click conversion rate estimation," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 275–284.
  [56] T. Wei, F. Feng, J. Chen, Z. Wu, J. Yi, and X. He, "Model-agnostic
- [56] T. Wei, F. Feng, J. Chen, Z. Wu, J. Yi, and X. He, "Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system," in *Proceedings of the ACM SIGKDD International* Conference on Knowledge Discovery & Data Mining, 2021, pp. 1791– 1800.
- [57] K. Damak, S. Khenissi, and O. Nasraoui, "Debiasing the cloze task in sequential recommendation with bidirectional transformers," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2022, pp. 273–282.
- [58] C. Xu, J. Xu, X. Chen, Z. Dong, and J.-R. Wen, "Dually enhanced propensity score estimation in sequential recommendation," in Proceedings of the ACM International Conference on Information & Knowledge Management, 2022, pp. 2260–2269.
- [59] M. Yang, Q. Dai, Z. Dong, X. Chen, X. He, and J. Wang, "Top-n recommendation with counterfactual user preference simulation," in Proceedings of the ACM International Conference on Information & Knowledge Management, 2021, pp. 2342–2351.
- [60] N. Liu, Q. Tan, Y. Li, H. Yang, J. Zhou, and X. Hu, "Is a single vector enough? exploring node polysemy for network embedding," in *Proceedings of the ACM SIGKDD International Conference* on Knowledge Discovery & Data Mining, 2019, pp. 932–940.
- [61] W. Chen, P. Ren, F. Cai, F. Sun, and M. de Rijke, "Improving end-to-end sequential recommendations with intent-aware diversification," in *Proceedings of the ACM International Conference on Information & Knowledge Management*, 2020, pp. 175–184.
- [62] J. Ma, C. Zhou, P. Cui, H. Yang, and W. Zhu, "Learning disentangled representations for recommendation," 2020
- gled representations for recommendation," 2020.
  [63] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," arXiv preprint arXiv:1710.09829, 2017.

- [64] G. E. Hinton, S. Sabour, and N. Frosst, "Matrix capsules with em routing," in *International Conference on Learning Representations*, 2018
- [65] G. E. Hinton, A. Krizhevsky, and S. D. Wang, "Transforming autoencoders," in *International Conference on Artificial Neural Networks*, 2011, pp. 44–51.
- [66] Z. Chai, Z. Chen, C. Li, R. Xiao, H. Li, J. Wu, J. Chen, and H. Tang, "User-aware multi-interest learning for candidate matching in recommenders," in *Proceedings of the International ACM SIGIR* conference on Research and Development in Information Retrieval, 2022, pp. 1326–1335.
- pp. 1326–1335.
  [67] S. Fan, X. Wang, C. Shi, K. Kuang, N. Liu, and B. Wang, "Debiased graph neural networks with agnostic label selection bias," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
  [68] H. Li, X. Wang, Z. Zhang, and W. Zhu, "Ood-gnn: Out-of-
- [68] H. Li, X. Wang, Z. Zhang, and W. Zhu, "Ood-gnn: Out-ofdistribution generalized graph neural network," IEEE Transactions on Knowledge and Data Engineering, 2022.
- [69] Y. Luo, Z. Liu, and Q. Liu, "Deep stable representation learning on electronic health records," in *IEEE International Conference on Data Mining*, 2022.
- [70] R. Xu, P. Cui, Z. Shen, X. Zhang, and T. Zhang, "Why stable learning works? a theory of covariate shift generalization," arXiv

- preprint arXiv:2111.02355, 2021.
- [71] J. McAuley, C. Targett, Q. Shi, and A. Van Den Hengel, "Image-based recommendations on styles and substitutes," in *Proceedings of the International ACM SIGIR conference on Research and Development in Information Retrieval*, 2015, pp. 43–52.
- [72] R. He and J. McAuley, "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering," in *Proceedings of the ACM Web Conference*, 2016, pp. 507–517.
- [73] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for youtube recommendations," in *Proceedings of the ACM Conference* on Recommender Systems, 2016, pp. 191–198.
- [74] S. Jean, K. Cho, R. Memisevic, and Y. Bengio, "On using very large target vocabulary for neural machine translation," arXiv preprint arXiv:1412.2007, 2014.
- [75] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Interna*tional Conference on Artificial Intelligence and Statistics, 2010, pp. 249– 256.
- [76] J. Ni, J. Li, and J. McAuley, "Justifying recommendations using distantly-labeled reviews and fine-grained aspects," in *Proceedings* of the Conference on Empirical Methods in Natural Language Processing, 2019, pp. 188–197.