Pre-Training with Transferable Attention for Addressing Market Shifts in Cross-Market Sequential Recommendation

Preprint · August 2024

DOI: 10.1145/3637528.3671698

Chen Wang
University of Illinois Chicago
28 PUBLICATIONS **213** CITATIONS
SEE PROFILE

Liangwei Yang
University of Illinois Chicago
54 PUBLICATIONS **757** CITATIONS
SEE PROFILE

Mingdai Yang
University of Illinois Chicago
14 PUBLICATIONS **58** CITATIONS
SEE PROFILE

Zhiwei Liu
University of Illinois Chicago
113 PUBLICATIONS **2,586** CITATIONS
SEE PROFILE

Pre-Training with Transferable Attention for Addressing Market Shifts in Cross-Market Sequential Recommendation

Chen Wang
University of Illinois Chicago
Chicago, Illinois, USA
cwang266@uic.edu

Ziwei Fan*
Amazon
Santa Clara, California USA
zwfan@amazon.com

Liangwei Yang
Salesforce AI Research
Palo Alto, California, USA
liangwei.yang@salesforce.com

Mingdai Yang
The University of Chicago
Chicago, Illinois, USA
frankyang@uchicago.edu

Xiaolong Liu
University of Illinois Chicago
Chicago, Illinois, USA
xliu262@uic.edu

Zhiwei Liu[†]
Salesforce AI Research
Palo Alto, California, USA
zhiweiliu@salesforce.com

Philip Yu
Tsinghua University
Beijing, China
psyu@tsinghua.edu.cn

ABSTRACT

Cross-market recommendation (CMR) involves selling the same set of items across multiple nations or regions within a transfer learning framework. However, CMR's distinctive characteristics, including limited data sharing due to privacy policies, absence of user overlap, and a shared item set between markets present challenges for traditional recommendation methods. Moreover, CMR experiences market shifts, leading to differences in item popularity and user preferences among different markets. This study focuses on cross-market sequential recommendation (CMSR) and proposes the **C**ross-market **A**ttention **T**ransferring with **S**equential **R**ecommendation (**CAT-SR**) framework to address these challenges and market shifts. CAT-SR incorporates a pre-training strategy emphasizing item-item correlation, selective self-attention transferring for effective transfer learning, and query and key adapters for market-specific user preferences. Experimental results on real-world cross-market datasets demonstrate the superiority of CAT-SR, and ablation studies validate the benefits of its components across different geographical continents. CAT-SR offers a robust and adaptable solution for cross-market sequential recommendation. The code is available at https://github.com/ChenMetanoia/CATSR-KDD/.

CCS CONCEPTS

• **Information systems → Recommender systems**.

KEYWORDS

Pre-Training, Sequential Recommendation, Self-Attention, Cross-Market Recommendation

[*] Work done before Amazon

[†] Corresponding author

KDD '24, August 25–29, 2024, Barcelona, Spain
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/10.1145/3637528.3671698

ACM Reference Format:
Chen Wang, Ziwei Fan, Liangwei Yang, Mingdai Yang, Xiaolong Liu, Zhiwei Liu, and Philip Yu. 2024. Pre-Training with Transferable Attention for Addressing Market Shifts in Cross-Market Sequential Recommendation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3637528.3671698

1 INTRODUCTION

Recommendation systems typically model historical user-item interactions to learn user and item representations. However, existing recommendation systems [12, 19, 27] usually assume localization, where data are stored to a specific country. The global expansion of companies like Amazon [3], eBay, and Spotify [1] requires expanding sale across different countries/regions, which demands the cross-market recommendation method.

The unique characteristics in the Cross-Market Recommendation (CMR) problem weaken existing recommendation methods. These characteristics include: (1) *data between markets cannot be shared* due to privacy constraints imposed by regulations such as the General Data Protection Regulation (GDPR)[1]; (2) *no overlap of users between markets*; and (3) *each market could access all items, even though some items may not be sold in the current market.*

These characteristics render cross-domain recommendation methods [9, 30, 34, 35, 40, 41], which involve transitioning between domains inapplicable. This is because they assume data centralization for joint training or the presence of overlapping users with shared interaction data. The data privacy restrictions are important and realistic problem for CMR. For example, Europe restricts the data sharing across continents, complicating the model training setting for international companies. The recent data privacy concerns related to TikTok also raise the issue of model transfer without data sharing. In CMR, a typical transfer learning protocol involves the pre-training, then fine-tune framework [2–4, 16, 17, 38].

This paper focuses on the study of Cross-Market Sequential Recommendations (CMSR). The challenges of CMSR present two distinct issues, i.e., the **shift in item popularity across markets**

[1] General Data Protection Regulation (GDPR) is a legal framework that sets guidelines for the collection and processing of personal information from individuals who live in and outside of the European Union. https://gdpr-info.eu/

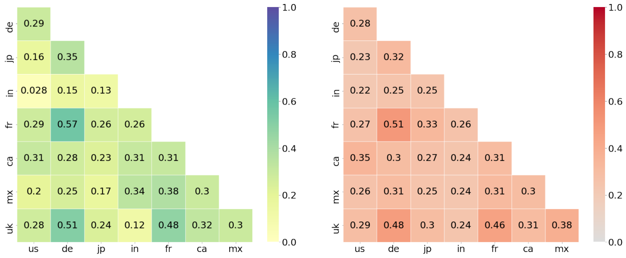

Figure 1: Comparison of similarities between eight markets including United States (us), Germany (de), Japan (jp), India (in), France (fr), Canada (ca), Mexico (mx) and United Kingdom (uk). Left heatmap showcases the item popularity similarity between two markets. Right heatmap illustrates the user preference similarity across different markets. Most low similarities show the disparities in item popularity and user preferences between markets.

and the **shift in user preferences across markets**. Firstly, the **item popularity market shift** across market describes the significant disparity of item popularity distributions between markets. The item popularity distribution reflects the global item preferences and also affects the optimization of item representations learning [25]. Although all markets share the same set of items, each market has its specific item popularity pattern due to factors such as culture and user habits. This is evident in the visualization of item popularity distributions between all pairs of markets in Fig. (1, Left), where most similarities are small (≤ 0.4), indicating a substantial popularity gap between markets. Secondly, the **shifting market preferences of users** illustrate that preferences for items are heterogeneous across markets. We measure the Spearman’s Rank Correlation of item preference ranks of users to demonstrate the similarity of user preferences in Fig. (1, Right). The maximum similarity is small, with a value of 0.51, indicating diverse user preferences across markets. The directly pre-train then fine-tune framework [3, 4, 17, 38] becomes sub-optimal in light of these challenges. To address the market shift problem in knowledge transfer, a market-aware (MA) model [2] has been proposed, explicitly modeling each market as an embedding. The item representation combines an across-market item embedding with a market embedding. Similarly, Bert4CMR [16] uses pre-training on parallel markets to learn item co-occurrences and fine-tuning on the target market to incorporate specific information.

However, existing solutions [2–4, 16, 17, 38] assume centralized data availability, which is impractical for CMR due to data sharing restrictions. While textual item representation learning [6, 14, 15] can be a solution, it neglects item popularity shifts and user preferences gaps. The shifted item popularity in the pre-trained source market limits the ability to generalize to unseen target markets. Adaptation to target market preferences becomes challenging without simultaneous access to user-item interaction data from both markets. Moreover, the issue of transferring and fine-tuning specific model parts in CMSR remains unexplored. Fine-tuning the entire model from the source market may lead to the problem of forgetting [10, 15, 29] due to insufficient target market data, causing overfitting. Directly transferring the entire model to the target market fails to capture its attributes adequately due to differences in item popularity and user preferences.

To address the aforementioned challenges in the CMSR, CAT-SR includes three key components: 1) pre-training module with a novel item re-weighting function; 2) self-attention transferring; and 3) market-specific adapters for fine-tuning. In the pre-training module, we introduce a novel item re-weighting function with two goals: suppressing popular items and assigning larger weights to unpopular items. This approach helps mitigate the item popularity market shift in CMSR. For instance, the popularity of an iPhone in the US may not necessarily reflect its popularity in India. To tackle the user preferences market shift, we introduce the selective self-attention transferring mechanism. By linking the self-attention and knowledge scoring function, we ilustrate that the self-attention scores capture item-item correlations optimized based on user preferences. These correlations differ from semantic similarities derived from meta information. This mechanism efficiently handles the selection of components to transfer during target market model fine-tuning, indicating that the subset of network parameters in the self-attention module provides better generalization ability for fine-tuning in the target market. For example, regardless of the market, if a user purchases a mobile phone, there may be a high probability that they will also buy a mobile phone case or film. Lastly, we introduce market-specific MLP adapters to augment the transferred self-attention parameters, enhancing the flexibility to capture market-specific user preferences.

In summary, our technical contributions include:

1) We highlighting the distinctive characteristics and market shifts challenges of cross-market recommendations.
2) We propose a novel pre-train then fine-tune framework, CAT-SR, specifically designed for cross-market sequential recommendation. This framework introduces a novel pre-training function to address item popularity market shift and mitigate its impact on recommendations.
3) We emphasize the significance of transferring only the item self-attention module for better generalization in target model adaptation, providing valuable insights into effective knowledge transfer in CMSR.
4) We demonstrate the necessity of market-specific adapters for fine-tuning in the target market, highlighting their importance in capturing market-specific user preferences.

2 RELATED WORK

We conduct a thorough review of existing literature from key research areas, including cross-market recommendation, federated learning, transferable item representation learning and, sequential recommendation.

2.1 Cross-market Recommendation

Cross-market recommendation (CMR) has emerged as a novel recommendation problem with distinct constraints and characteristics that differentiate it from cross-domain recommendation (CDR). Both CMR and CDR share the common objective of transferring knowledge from related domains to target domains to address the issue of data sparsity. However, CMR presents unique perspectives that set it apart from CDR: (1)**privacy constraint on user-item interactions**: In CMR, localized market data are collected individually, and due to privacy concerns, user-item interactions cannot be shared between markets. (2) **shared catalog of items across markets**: Unlike CDR, where the source and target domains may have different item sets, CMR assumes that all markets share the same catalog of items. (3) **non-overlapping users between markets**: CMR assumes that there are no overlapping users between markets, meaning users in one market do not exist in other markets. (4) **inability to centralize training data**: CMR poses the constraint of not being able to centralize the training data, which is a standard assumption in existing CDR methods. This makes CMR a more challenging problem, and most existing CDR solutions may not be directly applicable to CMR.

Several methods have been proposed to address the CMR problem. FOREC [3] and MAML-CF [17] adopt a meta-learning framework, pre-training on multiple markets and fine-tuning on target markets. M3Rec [4] extends FOREC's idea by introducing a novel framework for learning item similarities. However, these models do not consider the specific characteristics of each market, which is important since each market exhibits unique traits, as shown in Fig 1. On the other hand, MA [2], Bert4CMF [16], and SGLCMR [38] proposed market-specific modules to capture market-specific patterns in general, sequence, and graph recommendations, respectively, allowing the model to be personalized for various markets to fit their distinctive characteristics. However, all of these models assume that data from all markets are easily accessible, which may not be practical in the CMR setting where data sharing is limited due to privacy concerns or other reasons.

2.2 Federated Learning

Federated Learning (FL) holds promise for cross-market recommendation systems by enabling collaborative training across decentralized edge devices or servers, ensuring data privacy, reducing communication costs, and enhancing scalability, as highlighted in Zhang et al.'s survey [32]. This method trains models locally on each device and shares only model updates with a central server, preserving privacy and supporting flexible and scalable machine learning models. While FL has not yet been widely applied to cross-market recommendations, its attributes, such as privacy-preserving joint training and model update aggregation, could be highly beneficial. Techniques from foundational FL works like FedAvg [24], which uses average loss to update the global model, and advanced models like FedDCSR [33], which employ domain-shared and domain-exclusive feature disentanglement, could inform future applications in cross-market scenarios. However, integrating FL into cross-market recommendations presents challenges, such as capturing the distinct data distributions of each market and managing increased computational demands across diverse markets. Despite these challenges, the potential for FL to enhance privacy and scalability in cross-market recommendations is significant.

2.3 Transferable Item Representation Learning

Transferable item representation learning is crucial for recommender systems, enabling dense item vector representations that facilitate generalization across markets while adapting to unique market features. In cross-domain scenarios, methods like matrix factorization (MF)[36], DDTCDR [21], and SSCDR [18] transfer item representations using different techniques, but they may overlook data privacy or market shifts. Specifically, joint-training used in MF, DDTCDR, and SSCDR may raise data privacy concerns. Additionally, despite SSCDR's adaptability to overlapping item scenarios, it may encounter negative transferring issues due to significant differences in item popularity between the source and target markets, as shown in Fig. (4). NATR [9] successfully transfers item representations without data sharing, but it may suffer from negative transferring as well. This is because NATR may not effectively account for market shifts in different item popularity and user preferences between the source and target markets. As a result, the transferred item representations may not capture the characteristics of the target market, leading to suboptimal performance in cross-market recommendation.

On the other hand, recent works [6, 14, 15] focus on leveraging textual item features for better generalization and feasibility, ensuring all markets' item representations exist in the same space. This approach supports effective transfer learning and generalization in cross-market sequential recommendation and addresses the challenges faced by methods like NATR in negative transferring.

2.4 Sequential recommendation

Sequential recommendation (SR) models the dynamic preferences of users by considering their historical interactions as a sequence. A common approach in SR is to use a sequential encoder, which can be based on either Markov chains or Recurrent Neural Networks (RNNs). Markov chain-based methods like FPMC [26] and Fossil [11] infer the next item based on a few previous interactions and model first-order transition signals. On the other hand, RNN-based methods, such as GRU4Rec [13] and HGN [23], recursively capture sequential inputs, effectively modeling sequential data.

The Transformer architecture, inspired by its successful applications in various research areas, has gained attention in SR due to its ability to model high-order item-item transitions and scalability. SASRec [19] was the first work to adopt the Transformer architecture for SR, while BERT4Rec [27] extended SASRec with bi-directional attentions. Several variants of the Transformer architecture have since been proposed to further improve SR [20]. The Transformer architecture typically includes self-attention modules, feed-forward neural networks, residual connections, and layer normalization modules. Each component plays a unique role, such as the self-attention module measuring item-item correlations within the entire sequence to capture user preferences, which differ from semantic similarities derived from item meta-features, making it particularly relevant for cross-market transfer and adaptation.

3 PROBLEM DEFINITION

In this section, we outline the problem setting of our cross-market sequential recommendation problem. Specifically, we focus on the one-to-one setting of cross-market recommendation, where we pre-train a model using abundant data from one market and fine-tune it using relatively sparse data from another market. We exclusively transfer the selective self-attention mechanism to the target market addressing data privacy concerns. In this work, we specifically consider sequential recommendation in the CMR setting, as sequential methods have been extensively investigated and found to be scalable across various recommendation tasks [19, 20, 27].

3.1 Cross-Market Recommendation

In cross-market recommendation (CMR), we have multiple markets $\mathcal{M}$ and the associated user set $\mathcal{U}$, item set $\mathcal{V}$, and user-item interactions $\mathbf{R}$, where $\mathcal{M}_i = \{\mathcal{U}^i, \mathcal{V}^i, \mathbf{R}^i\}$ for each market $\mathcal{M}_i$. In CMR, there is no user overlap between markets, *i.e.,* $\mathcal{U}^i \cap \mathcal{U}^j = \emptyset$ for any pair of market $\mathcal{M}_i$ and market $\mathcal{M}_j$. Moreover, markets share the same set of items, *i.e.,* $\mathcal{V}^i = \mathcal{V}^j$. We focus on the one-to-one learning setting, with one source market denoted as S with abundant data, and one target market denoted as T. We use the superscript to denote the data of source and target markets, which are $\mathcal{M}^S = \{\mathcal{U}^S, \mathcal{V}^S, \mathbf{R}^S\}$ for the source market and $\mathcal{M}^T = \{\mathcal{U}^T, \mathcal{V}^T, \mathbf{R}^T\}$ for the target market, respectively.

Specifically, we illustrate the workflow of our one-to-one setting in CMR, including the pre-training stage and the fine-tuning stage. We utilize the source market data $\mathcal{M}^S$ and a well-designed pre-training loss $\mathcal{L}_{\text{pre}}$ to pre-train a recommendation model with learnable parameters Θ^S as follows:

$$\Theta^S = \underset{\Theta^S}{\arg\min}\, \mathcal{L}_{\text{pre}}(\mathcal{U}^S, \mathcal{V}^S, \mathbf{R}^S). \tag{1}$$

In the fine-tuning stage, the pre-trained source market model Θ^S is transferred as the initialized point of the target market model $\Theta^T = \text{init}(\Theta^S)$. Note that the proper design of the model transferring process $\text{init}(\cdot)$ remains an essential but challenging research question to be addressed, which is firstly investigated in this work. The fine-tuning stage is formulated as follows:

$$\begin{aligned} &\underset{\Theta^T}{\arg\min}\, \mathcal{L}(\mathcal{U}^T, \mathcal{V}^T, \mathbf{R}^T, \text{init}(\Theta^T)), \\ &\qquad \text{where} \quad \text{init}(\Theta^T) = \Theta^S. \end{aligned} \tag{2}$$

The challenges in CMR include the investigations on how to transfer the pre-trained source market model, *i.e.,* the definition of $\text{init}(\cdot)$, and how to develop the pre-training loss $\mathcal{L}_{\text{pre}}$ when the source market S and the target market T have different user behavior patterns in between $\mathbf{R}^S$ and $\mathbf{R}^T$.

3.2 Sequential Recommendation

Building upon the recent advancements in sequential recommendation [6, 14, 15] that utilize textual item representations, we leverage a pre-trained language model (PLM) to encode the meta-textual information of universal items into vector representations, such as BERT [5]. These encoded textual embeddings of items are then used as inputs for sequential recommendation methods, eliminating the need for learning item embeddings from random initialization, as in methods like SASRec. One advantage of SR using textual item representations over traditional SR is that item features are universal, meaning that the recommendation model transfer in the SR becomes more feasible as all input item features are in the same PLM-encoded space.

To be specific, in SR, we have a set of users $\mathcal{U}$, and a set of items $\mathcal{V}$ with the associated textual meta information $\mathbf{X}$, *e.g.,* description and title. Each user u has a set of item interactions with timestamps, and we sort the interacted items chronologically to form the user sequence as $\mathcal{S}_u = [v_1^u, v_2^u, \ldots, v_{|\mathcal{S}^u|}^u]$. $v_i^u \in \mathcal{V}$ denotes the i-th interacted item in the user sequence $\mathcal{S}_u$. Each item v is encoded into a vector representation $\mathbf{E}_v = \text{PLM}(\mathbf{X}_v)$ and $\mathbf{E} \in \mathbb{R}^{|\mathcal{V}| \times d}$, where $\mathbf{X}_v$ denotes the item's textual information.

The SR model Θ is optimized by minimizing the next-item prediction loss $\mathcal{L}$ as follows:

$$\Theta = \underset{\Theta}{\arg\min}\, \mathcal{L}\left(\mathcal{S}_u, \mathbf{E}\right). \tag{3}$$

Again, different from the definition of SR in [7, 19, 27], the SR in our model has the fixed item feature representations $\mathbf{E}$ as inputs rather than as learnable parameters in [7, 19, 27].

Definition 3.1. **Cross-Market Sequential Recommendation**. There are multiple markets where each market has sequence interactions and items as $\mathcal{M}_i = \{\mathcal{S}^i, \mathcal{V}, \mathbf{X}\}$, where $\mathcal{S}_u^i$ denotes the user u's sequence in the market i. All markets $\mathcal{M}_i$ share the same set of items $\mathcal{V}$, and the item textual representations $\mathbf{E} = \text{PLM}(\mathbf{X})$ are shared across markets. In other words, each market has sequence interactions, items and the associated item textual representations as $\mathcal{M}_i = \{\mathcal{S}^i, \mathcal{V}, \mathbf{E}\}$. The pre-training stage defined in Eq. (1) and the fine-tuning stage in Eq. (2) still have similar formulations but with the sequence interactions and items textual representations instead. Moreover, the recommendation model is based on sequential encoder architectures, such as the building block SASRec.

4 PROPOSED FRAMEWORK

In this section, we present our framework, CAT-SR, for the cross-market sequential recommendation problem, as illustrated in Fig. (2). The framework is built upon the Transformer architecture, which serves as the building block model for the SR problem. We have developed three key components to address the challenges of market-based biased item popularity and market-specific user preferences modeling, including the pre-training with a novel designed reweighting function, the self-attention transferable operator for defining $\text{init}(\Theta^T)$ illustrated in Eq. (2), and the last market-specific adaptation module with simple MLP adapters.

4.1 Transformer as Building Block

As Transformer establishes the building block in our CMSR problem, we first introduce its ingredients for better illustrations, as shown in the major component in Fig. (2). As presented in the problem definition of the SR in Section 3.2, we have the interaction sequences $\mathcal{S}_u$ and the textual item representations $\mathbf{E}$ extracted from the meta information. For each user sequence $\mathcal{S}_u$, the sequence is first truncated by removing the earliest items if $|\mathcal{S}_u| > n$ or padded with 0s to meet the maximum sequence length n, resulting in a fixed length sequence $s = (s_1, s_2, \ldots, s_n)$. With a trainable positional embedding $\mathbf{P} \in \mathbb{R}^{n \times d}$, we have the sequence embedding matrix as:

$$\hat{\mathbf{E}}_{\mathcal{S}_u} = [\mathbf{e}_{s_1} + \mathbf{p}_{s_1}, \mathbf{e}_{s_2} + \mathbf{p}_{s_2}, \ldots, \mathbf{e}_{s_n} + \mathbf{p}_{s_n}], \tag{4}$$

where $\mathbf{e}_v$ denotes the textual item representation of the item v. Specifically, the self-attention module uses scaled dot-products between items in the sequence to infer their correlations, which

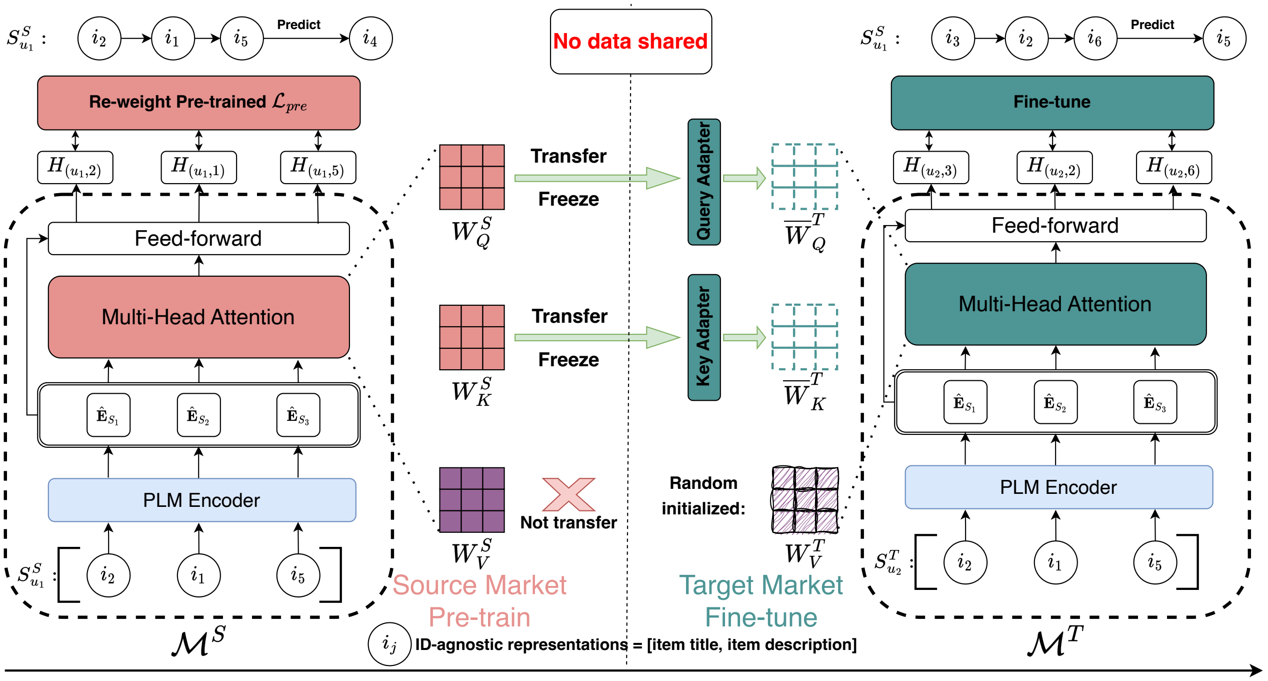

Figure 2: The proposed model framework CAT-SR consists of three major components. The first component is the source market pre-training with item popularity re-weighting to obtain the optimized self-attention for item-item correlations from the source market. The second step is to transfer the self-attention weights W^S_Q and W^S_K to the target market. The third step is the fine-tuning in the target market. We freeze $W^T_Q = W^S_Q$ and $W^T_K = W^S_K$ and update parameters via market-specific MLP query adaptor AQ and key adaptor KQ. The textual embedding of each item is generated and fixed by a shared but frozen pre-trained language model, which provides the consistent representations such that the embeddings of all items in different markets are represented in the shared latent space.

are as follows:

$$\mathrm{SA}(\hat{\mathbf{E}}_{\mathcal{S}_u}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V} = \mathrm{softmax}\left(\frac{\hat{\mathbf{E}}_{\mathcal{S}_u}\mathbf{W}_Q(\hat{\mathbf{E}}_{\mathcal{S}_u}\mathbf{W}_K)^\top}{\sqrt{d}}\right)\mathbf{V}, \tag{5}$$

where $\mathbf{Q} = \hat{\mathbf{E}}_{\mathcal{S}_u}\mathbf{W}_Q$, $\mathbf{K} = \hat{\mathbf{E}}_{\mathcal{S}_u}\mathbf{W}_K$, and $\mathbf{V} = \hat{\mathbf{E}}_{\mathcal{S}_u}\mathbf{W}_V$. As both $\mathbf{Q}$ and $\mathbf{K}$ are transformations to the same sequence items, the self-attention module SA measures learn the latent correlation between items, which are optimized based on user preferences. The point-wise feed-forward network (FFN) is further applied after the self-attention module as follows:

$$\mathbf{H}_u = \mathrm{FFN}\left(\mathrm{SA}(\hat{\mathbf{E}}_{\mathcal{S}_u})\right) = \mathrm{ReLU}\left(\mathrm{SA}(\hat{\mathbf{E}}_{\mathcal{S}_u})\mathbf{W}_1 + \mathbf{b}_1\right)\mathbf{W}_2 + \mathbf{b}_2, \tag{6}$$

where $\mathbf{H}_u$ denotes the sequence output representations. We omit the residual connection and layer normalization components for simplicity, and details can be found in [19, 28]. These components can be stacked multiple times, as shown in Fig. (2).

4.2 Source Market Pre-training

With the source market sequences and textual item representations $\mathcal{M}^S = \{\mathcal{S}^S, \mathbf{E}\}$, the pre-training on the source market is the first step in our proposed framework CAT-SR, as shown in the upper left component of Fig. (2). In the pre-training stage, we propose a re-weighted pre-training loss $\mathcal{L}_{\text{pre}}$ to mitigate the market shift from the item popularity in the source market for better generalization ability of the pre-trained model. The item popularity re-weighting pre-training is crucial as there is a significant gap in the item popularity distributions between the source and target market, which is shown in Fig. (1, Left). To properly design the pre-training loss, we need to achieve two goals simultaneously: (1) *mitigating the impact of popular items* and (2) *assigning greater significance to less popular items compared to their popular counterparts.* We achieve this by employing the proposed re-weighting function $w_{F(v)}$, which adjusts the weight allocation in the positive and negative loss calculation based on the popularity $F(v)$ of the item v. We proposed our re-weighting pre-training loss $\mathcal{L}_{\text{pre}}$, which is defined as follows:

$$\mathcal{L}_{\text{pre}} = -\sum_{\mathcal{S}_u \in \mathcal{S}^S}\sum_{t=1}^{|\mathcal{S}_u|}\left(w_{F(j^+)}\mathcal{L}_{\text{pos}} + w_{F(j^-)}\mathcal{L}_{\text{neg}}\right), \tag{7}$$

$$\mathcal{L}_{\text{pos}} = \log\left(\sigma\left(\mathbf{H}^\top_{(u,t)}\mathbf{E}_{j^+}\right)\right), \mathcal{L}_{\text{neg}} = \log\left(1 - \sigma\left(\mathbf{H}^\top_{(u,t)}\mathbf{E}_{j^-}\right)\right),$$

where $\mathcal{S}^S$ denotes the source market's sequences, $\mathbf{H}_{(u,t)}$ denotes the output at the time step t of the user sequence $\mathcal{S}_u$, $\mathbf{E}_{j^+}$ and $\mathbf{E}_{j^-}$ denote embeddings of the positively interacted item j^+ with the user u and the negatively sampled item j^-, and $\sigma(\cdot)$ is the sigmoid function. $\mathcal{L}_{\text{pre}}$ aims to make the model's output more similar to the embedding of positive items, while also making it less similar to the embedding of negative items. This is done while taking into account how popular the current item is. The weight $w_{F(v)}$ boosts the significance of less popular items while diminishing the relevance of other items throughout the backpropagation phase.

Baseline with Equal Weights. The baseline case suffers from the item popularity bias because popular items appear more in sequence interactions, where $\mathcal{L}_{\text{pre}}$ sums up all sequence interactions. In other words, popular items have more opportunities being optimized in the baseline case. Given the significant difference of item popularity between source and target markets, as shown in Fig. (4), the pre-trained model operating under an equal weight baseline demonstrates sub-optimal performance.

The pre-training loss degrades to the regular cross-entropy loss adopted by existing methods [19, 20], when we enforce the $w_{F(v)} = 1$. The proposed $w_{F(v)}$ is formulated as follows:

$$w_{F(v)} = \alpha - \tanh(\beta F(v) - \beta), \tag{8}$$

where $\tanh(\cdot)$ denotes the hyperbolic tangent activation function mapping the positive input value to $[0, 1]$, α represents the upper bound of the item weight and β denotes the popularity shift. Larger values of α signify a stronger emphasis on mitigating the impact of popularity bias. This emphasis is manifested in the amplification of attention weights for items with lower popularity scores, promoting a more balanced representation of recommendations. Conversely, smaller values of α may result in a more conservative approach, where the amplification effect is less pronounced. The parameter β plays a crucial role in shaping the transformation's behavior. A higher value of β introduces a steeper slope to the tanh, leading to a sharper transition in weight adjustments. This can be interpreted as a mechanism to impose a stronger penalty on items with higher popularity scores.

Analysis of $w_{F(v)}$. We visualize the curves of the item popularity re-weighting solution $w_{F(v)}$ for items in Fig. (4). In Fig. (4), we also include the baseline $w_{F(v)} = 1$, which always assigns equal weight value of one to items, regardless of the item popularity. The proposed $w_{F(v)}$ suppresses the item weight for popular items with abundant interactions in Fig. (4). With different α and β, the curves of $w_{F(v)}$ are monotonically decreasing, with the intention that larger popularity has a smaller weight. This design addresses the item popularity shift in the source market, with the goal of obtaining a more balanced pre-trained model.

4.3 Selective Transferable Self-Attention

Transferring partial modules can help adapt the model to the target domain more effectively, which will be demonstrated in the experimental analysis Section 5.4. With the pre-trained Transformer using the data in the source market, we need to selectively transfer semantically beneficial information from the source market to the target market. As shown in the middle part of Fig. (2), we only transfer the weights of the self-attention module on items. We argue that the self-attention module captures items' semantic similarities, which are expected to be beneficial for the target market. We also argue that transferring only partial modules that capture generic features or representations is preferred over transferring the whole model, as the latter may transfer irrelevant or conflicting knowledge from the source domain, especially when the source and target markets have significant differences. As all markets share the same set of items, the item similarities are crucial in the cross-market recommendation [4]. As demonstrated in [8], the self-attention component encodes the item-item correlations. For a specific item pair (v_i, v_j) in the sequence, we expand the self-attention calculation in Eq. (5) as follows:

$$\mathrm{Att}(v_i, v_j) = \mathbf{Q}_{v_i}\mathbf{K}_{v_j}^\top = \hat{\mathbf{E}}_{v_i}\mathbf{W}_Q\mathbf{W}_K^\top\hat{\mathbf{E}}_{v_j}^\top = \hat{\mathbf{E}}_{v_i}\mathbf{W}_{QK}\hat{\mathbf{E}}_{v_j}^\top,$$

where $\mathbf{E}_{v_i}$ and $\mathbf{E}_{v_j}$ denote the item embeddings of item v_i and v_j in $\mathcal{S}_u$ respectively, $\mathbf{W}_Q \in \mathbb{R}^{d\times d}$, $\mathbf{W}_K \in \mathbb{R}^{d\times d}$ are weight matrices in self-attention, and $\mathbf{W}_{QK} = \mathbf{W}_Q\mathbf{W}_K^\top$. With the similar forms of the self-attention scores and knowledge embedding scorings from knowledge graph methods, *e.g.,* DistMult [31] and ANALOGY [22], the self-attention scaled dot-product can be interpreted as a scoring function for measuring the item-item correlation with the $\mathbf{W}_{QK}$ as the latent space mapping [8]. Moreover, the attention scores are optimized based on the user preferences in sequences, which is significantly different from the item similarities from only the item meta textual information, *i.e.,* $\mathbf{E}\mathbf{E}^\top$.

Instead of transferring the whole pre-trained Transformer, we transfer the $\mathbf{W}_{QK} = \mathbf{W}_Q\mathbf{W}_K^\top$, which encodes the item-item correlations and is referred to as the **selective self-attention transferring**. Specifically, in the $\mathrm{init}(\Theta^T) = \Theta^S$, Θ^S includes the learnable parameters in the self-attention module, point-wise feed-forward networks, and the layer normalization module. We only transfer the subset of Θ^S, *i.e.,*

$$\mathrm{init}(\mathbf{W}_Q^T) = \mathbf{W}_Q^S,\ \mathrm{init}(\mathbf{W}_K^T) = \mathbf{W}_K^S, \tag{9}$$

where $\{\mathbf{W}_Q^S, \mathbf{W}_K^S\}$ denotes the $\mathbf{W}_Q$ and $\mathbf{W}_K$ pre-trained in the source market.

4.4 Market-Specific Fine-Tuning

The self-attention transferring $\mathrm{init}(\Theta^T) = \{\mathbf{W}_Q^S, \mathbf{W}_K^S\}$ shares the item-item correlations learned from the source market, and acts as the initialization point for the item-item correlations in the target market. However, the item re-weighting pre-training loss and the selective self-attention transferring encode the domain-invariant item correlations, which is insufficient for the target market modeling due to the user preferences and item popularity distribution shifts. We propose to augment transformations on the selective transferred self-attention weights to provide better flexibility in the target market, thus adopting the domain-specific knowledge presented in the right part of Fig. (2). Specifically, while the target market are initialized with $\mathbf{W}_Q^T = \mathbf{W}_Q^S$ and $\mathbf{W}_K^T = \mathbf{W}_K^S$, we apply the MLP adapters AQ and AK to the pre-trained query and key matrices in the target domain $\mathbf{W}_Q^T$ and $\mathbf{W}_K^T$, respectively:

$$\overline{W}_Q^T = AQ(\mathbf{W}_Q^S) = \mathbf{W}_{AQ}\mathbf{W}_Q^S,\ \overline{W}_K^T = AK(\mathbf{W}_K^S) = \mathbf{W}_{AK}\mathbf{W}_K^S. \tag{10}$$

The adapters are jointly fine-tune with the target domain's data as:

$$\mathcal{L} = -\sum_{\mathcal{S}_u\in\mathcal{S}^T}\sum_{t=1}^{|\mathcal{S}_u|}\left[\log\left(\sigma\left(\mathbf{H}_{(u,t)}^\top\mathbf{E}_{j^+}\right)\right) + \log\left(1-\sigma\left(\mathbf{H}_{(u,t)}^\top\mathbf{E}_{j^-}\right)\right)\right], \tag{11}$$

Table 1: Datasets Statistics

	de	jp	in	fr	ca	mx	uk	us
# users	2,373	487	239	2,396	5,675	1,878	4,847	35,916
# items	2,210	955	470	1,911	5,772	1,645	3,392	31,125
# inter	22,247	4,485	2,015	22,905	55,045	17,095	44,515	364,339

where $\mathcal{S}^T$ denotes the target market's sequences, and $\mathbf{H}$ denotes the sequence output embeddings from the fine-tuned target model with transferred $\{\mathbf{W}_Q^S, \mathbf{W}_K^S\}$.

5 EXPERIMENTS

We extensively evaluate the effectiveness of our proposed framework CAT-SR on a large-scale cross-market dataset. Additionally, we perform ablation studies to demonstrate the superiority of our design. Our evaluation addresses the following research questions: **RQ1:** Does CAT-SR achieve better cross-market recommendations than state-of-the-art baselines? **RQ2:** Does the item popularity re-weighting module benefit? **RQ3:** Is the choice of self-attention transferring better than other components? **RQ4:** Is using MLP adapters in the fine-tuning stage necessary?

5.1 Experimental Setup

In this section, we provide an overview of the datasets, evaluation protocol, baselines, and implementation details.

5.1.1 Dataset. We evaluate our proposed model using the publicly available real-world cross-market dataset called *XMarket*[2], obtained from Amazon. This dataset is designed to facilitate cross-market item recommendation and market adaptation and includes ratings, reviews, and various metadata such as item title, average rating, and item details. The dataset comprises seven electronic markets across three continents, as shown in Table1, which is consistent with CMR model FOREC [3] and MA [2]. To create item text features, we concatenate the title and description textual information for each item, which serves as the meta information in the dataset.

5.1.2 Evaluation Protocol. To evaluate our cross-market recommendation tasks, we employ the widely recognized leave-one-out (LOO) evaluation approach. This approach uses the most recent user interaction as the test item, while the second-to-last interaction serves as the validation item. To assess the model's effectiveness, we measure its ability to rank the test item against a set of negative items that the user hasn't interacted with. In line with the practices of other CMR models [2, 3, 16, 17], we sample 99 negative items for each user during evaluations to ensure equitable comparisons. As our focus is on top-N item recommendation, we rely on normalized discounted cumulative gain (nDCG@10) and Hit-Rate (HR@10) as the established evaluation metrics.

5.1.3 Baselines. In our extensive evaluation of CAT-SR's recommendation performance, we conduct a thorough comparison with various models across each target market. SASRec and S^3-Rec belong to the single-domain models, leveraging directional self-attention for capturing item correlations within sequences. Conversely, DCDCSR, SSCDR, and NATR fall under the category of CDR models, with a strong focus on learning transferable user and item representations. These models are adapted for CMR by establishing connections between the item features of two markets. Additionally, we incorporate other CMR models like MAML, FOREC, and MA into our assessment. MAML serves as a widely employed meta-learning model in the recommendation domain. FOREC, on the other hand, is a CMR model that employs a meta-learning approach to transfer knowledge from source to target markets by freezing and modifying specific layers in their architectures. MA utilizes market embeddings to adjust items for the current market and leverages data from multiple auxiliary markets for training a comprehensive recommendation system, achieving optimal performance across various settings. Notably, DCDCSR, SSCDR, MAML, FOREC, and MA possess supplementary information, granting them simultaneous access to both source and target domains, which must be considered while interpreting the evaluation results. Moreover, we further consider federated learning models like FedAvg and FedDCSR for evaluation. FedAvg uses average loss to update the global model, while the cutting-edge model FedDCSR employs a domain-shared and domain-exclusive feature disentanglement strategy for training. To fairly assess the influence of Pre-trained Language Models (PLMs) on performance, we use BERT to generate item embeddings based on UnisRec [15]. These embeddings are then applied to SASRec, UnisRec, and CAT-SR for comprehensive analysis.

5.1.4 Hyper-parameters and Grid Search. To ensure the fairness and reliability of our experiments, we implemented our proposed framework on top of RecBole [37]. For SASRec and S^{3}Rec, we conducted an extensive hyper-parameter search, exploring different values for the learning rate (0.01, 0.001, 0.0001), embedding dimension (64, 128, 384), number of layers (1, 2, 4), and number of heads (1, 2, 4). The batch size was set to 256. For DCDCSR, NATR, and SSCDR, we used the default settings provided by RecBole, with the item overlapping mode enabled. Regarding MAML, FOREC, and MA, we replicated the results using MA's publicly available code [3].

5.2 Overall Performance (RQ1)

We evaluate the proposed model against baselines on seven different markets, with the largest data set, the United States (**us**) market, used as the source domain for model pre-training. We obtain the following observations: 1) Text-enhanced sequential recommendation methods (SASRecBERT and UnisRecBERT) exhibit superior performance compared to traditional methods. This is attributed to their utilization of item texts as auxiliary features, contributing to performance enhancement. 2) Cross-market models (MAML, FOREC, MA) surpass cross-domain models (DCDCSR, SSCDR, NATR) as well as single-domain models (SASRec, S^3-Rec). The superior performance of cross-market models is attributed to their capability to address market shifts, unlike cross-domain models. Furthermore, single-domain models lack auxiliary information crucial for enhancing performance in the target market context. 3) cross-domain models do not surpass single-domain models in performance. This could be attributed to the direct transfer of item knowledge by cross-domain models, which may inadvertently lead to negative effects

[2] https://xmrec.github.io/

[3] https://github.com/samarthbhargav/efficient-xmrec

Table 2: Overall Performance Comparison Table. The best and second-best results are bold and underlined, respectively. "Improv." indicates the relative improvement ratios of the proposed approach over the best performance baselines. "*" denotes that the improvements are significant at the level of 0.05 with paired t-test. $_{\text{BERT}}$ means using BERT to generate item embedding.

	nDCG@10							HR@10						
	de	**jp**	**in**	**fr**	**ca**	**mx**	**uk**	**de**	**jp**	**in**	**fr**	**ca**	**mx**	**uk**
SASRec [19]	0.2010	0.1810	0.2234	0.1976	0.1373	0.3105	0.2529	0.3266	0.2875	0.4644	0.3322	0.2153	0.5048	0.4269
S^3-Rec [39]	0.1943	0.2094	0.2397	0.2573	0.1544	0.3321	0.2748	0.3012	0.3123	0.4817	0.3936	0.2465	0.5331	0.4532
DCDCSR [21]	0.1452	0.1561	0.1665	0.2154	0.1498	0.2734	0.2099	0.2736	0.2687	0.4200	0.3484	0.2399	0.4645	0.3612
SSCDR [18]	0.1348	0.1489	0.1074	0.1831	0.0954	0.1149	0.1465	0.2643	0.2123	0.2986	0.2446	0.1567	0.2646	0.2357
NATR [9]	0.1855	0.1737	0.1788	0.2610	0.1988	0.3331	0.2289	0.3011	0.2718	0.3567	0.4185	0.2884	0.5198	0.3984
MAML [17]	0.3048	0.1915	0.4295	0.3216	0.2938	0.5592	0.4508	0.4437	0.3162	0.5146	0.4641	0.4449	0.6560	0.5729
FOREC [3]	0.3264	0.2095	0.4383	0.3228	0.2942	**0.5664**	0.4654	0.4707	0.3367	0.5188	0.4661	0.4525	**0.6570**	0.5871
MA [2]	0.3419	0.2131	0.4678	0.3283	0.3220	0.5547	0.4563	0.4884	0.3593	0.5439	0.4737	0.4800	0.6363	0.5685
FedAvg [24]	0.2137	0.1748	0.2087	0.1950	0.1169	0.2628	0.2517	0.3287	0.2916	0.4059	0.3022	0.2090	0.4819	0.4205
FedDCSR [33]	0.4925	0.5018	0.3697	0.4244	0.3291	0.4135	0.5108	0.5974	0.5791	0.4235	0.5126	0.4157	0.4987	0.5673
$\text{SASRec}_{\text{BERT}}$	0.5464	0.4886	0.3787	0.4837	0.2808	0.3004	0.4422	0.6242	0.5863	0.4575	0.5639	0.3276	0.3919	0.5043
$\text{UnisRec}_{\text{BERT}}$	0.6342	0.6092	0.4767	0.6039	0.3883	0.2879	0.5228	0.6492	0.6241	0.5526	0.6538	0.4599	0.5628	0.6083
$\text{CAT-SR}_{\text{BERT}}$	**0.6957***	**0.6620***	**0.6145***	**0.6442***	**0.4681***	0.5338	**0.6044***	**0.7605***	**0.7444***	**0.7348***	**0.7301***	**0.5754**	0.6463	**0.6869***
Improv.	+9.70%	+8.67%	+28.91%	+6.67%	+20.55%	-	+15.61%	+17.14%	+19.28%	+32.91%	+16.70%	19.86%	-	+12.92%

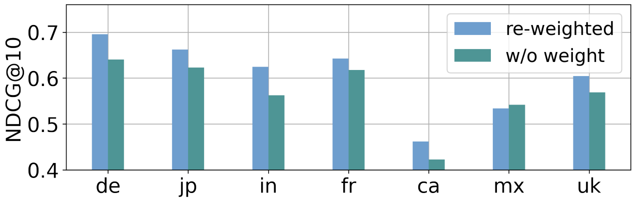

Figure 3: Performance comparison between re-weighted v.s. w/o weight on different continents.

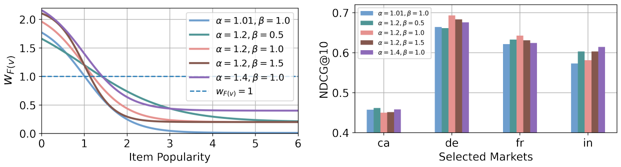

Figure 4: Left: Visualization $w_{F(v)}$ for varying parameters α and β, $F(v)$ denotes the item v's. Right:Performance differences with different α and β values in the re-weighting function for different continents.

in target markets. 4) FedDCSR significantly outperforms FedAvg, showcasing its ability to effectively separate user sequence features into domain-shared and domain-exclusive components, thus enhancing performance. However, CAT-SR consistently exceeds the performance of both federated learning models across all evaluation metrics in the seven markets. This underscores a critical distinction between cross-domain and cross-market recommendations: the importance of accurately capturing market-specific biases in cross-market scenarios. 5) CAT-SR outperforms all baselines except Mexico (**mx**), since CAT-SR eliminates the impact of the source market shifts on recommendations while acquiring abundant item self-attention knowledge from the source market. The transfer component of the model provides flexibility for downstream tasks when transferring item self-attention to the target market, preventing the dominance of source market-learned parameters. Furthermore, the adapter for query and key matrices is better aligned with the target market while preserving the item's relevance. The reason why CAT-SR's performance on Mexico is slightly worse than the best baseline, is because we cannot joint-train source and target markets so it is difficult to capture the target market distribution. While re-weighting the pre-training loss can alleviate the effects of market-shift, achieving a precise match of the target market distribution is still challenging without shared data.

5.3 Item Popularity Reweighting Effects (RQ2)

We conducted a comparison of our model's performance in different markets, with and without the inclusion of the item popularity re-weighting module in the pre-training loss objective. Fig. (3) demonstrate that the adoption of the item popularity re-weighting module significantly improves recommendation performance in all markets. Interestingly, transferring the re-weighted module to the India (**in**) market resulted in the highest improvement, even in light of the pronounced cultural differences between the United States (**us**) and India (**in**) markets. This observation underscores the model's efficacy in mitigating the impact of divergent popularity biases between these markets, effectively readjusting the item popularity bias. The negative impact of transferring from the **us** market to the Mexico (**mx**) market could be attributed to the unique characteristics of the **mx** market, which may not be well-captured by the self-attention learned from the us market. This misalignment with the preferences and behaviors of users in the mx market may have resulted in suboptimal recommendations.

We also conducted an analysis of the patterns of the popularity re-weighting function. As shown in Fig. (4), we illustrated the performance differences with various values of α and β, which are hyper-parameters of the item popularity re-weighting function in Eq. (8). Comparing four countries, we observed that Canada

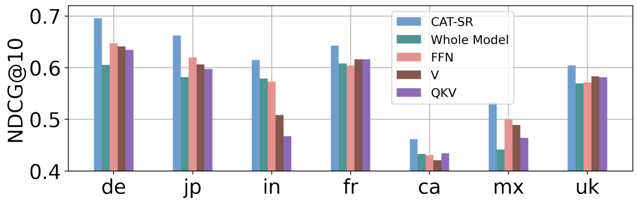

Figure 5: Performance comparison between different module transfer methods on different continents, including CAT-SR (query and key) only, transferring the whole model, the feed forward layer (FFN), (value) only, (query, key, and value)

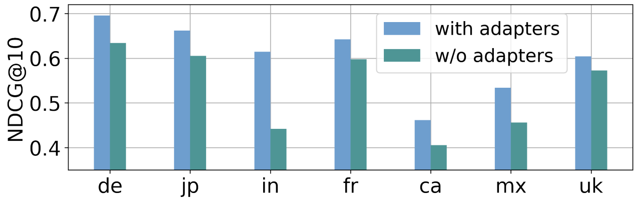

Figure 6: Performance comparison of adopting the adapters on different continents.

ca market achieved the best results when $\alpha = 1.2$ and $\beta = 0.5$, while in European markets Germany (**de**) and France (**fr**) required larger values of $\alpha = 1.2$ and $\beta = 1.5$. As demonstrated in Fig. (4), α represents the upper bound of the weight, and β determines the popularity shift weight. Considering that India (**in**) is relatively smaller markets compared to more popular markets, they may experience larger market shifts. Hence, a larger re-weighting on the item popularity is needed for these two markets to effectively address the popularity bias in recommendations.

5.4 Study of Transferable Components (RQ3)

Transferring query and key components is the most effective. We conduct five experiments in which we transferred the whole model, the feed-forward layer (FFN), or the multi-head attention module, including transferring (query, key, and value), (value) only, or (query and key) only. The experimental results show that transferring a partial model is more effective than transferring the whole model, as shown in Fig. (5). This is because transferring the entire model may cause overfitting to the source domain and negatively impact the performance on the target domain. Transferring the query and key components is the most effective approach, as the query and key components are responsible for computing the attention weights between the input sequence and the learned representations. On the other hand, the value component generates the output sequence and is less domain-specific. By transferring only the query and key components, we can leverage the knowledge learned from the source domain about computing the attention weights, particularly when the target domain has similar patterns in the input sequence as the source domain.

5.5 Market-Specific Adapters (RQ4)

We conduct an experiment to determine the usefulness of adapters during the fine-tuning stage to map the pre-trained query and key matrices to the target market, as shown in Fig. (6). Our experimental results show that adapters benefit the model. By adding an adapter such as MLP during the fine-tuning stage, the model can leverage the knowledge it gains from the source domain while adapting to the specific characteristics of the target domain. This approach effectively prevents the model from overfitting to the target domain by fixing the query and value matrices, which are responsible for generating the output sequence and may require more domain-specific knowledge. The adapter allows the model to learn new representations specific to the target domain, thereby improving performance on the target task. Overall, this approach achieves a balance between leveraging the knowledge learned from the source domain and adapting to the characteristics of the target domain, resulting in improved performance.

6 CONCLUSION

We investigate and address the challenges of cross-market sequential recommendation with our novel framework CAT-SR. It comprises three components: a novel item popularity re-weighting function for pre-training, self-attention transferring to improve generalization, and market-specific MLP adapters for user preferences adaptation. Experiments on a large dataset confirm the effectiveness of CAT-SR, and ablation studies highlight the importance of each component.

REFERENCES

[1] Ashton Anderson, Lucas Maystre, Ian Anderson, Rishabh Mehrotra, and Mounia Lalmas. 2020. Algorithmic effects on the diversity of consumption on spotify. In *Proceedings of The Web Conference 2020.* 2155–2165.

[2] Samarth Bhargav, Mohammad Aliannejadi, and Evangelos Kanoulas. 2023. Market-Aware Models for Efficient Cross-Market Recommendation. In *ECIR (1) (Lecture Notes in Computer Science, Vol. 13980).* Springer, 134–149.

[3] Hamed Bonab, Mohammad Aliannejadi, Ali Vardasbi, Evangelos Kanoulas, and James Allan. 2021. Cross-Market Product Recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management.* ACM.

[4] Jiangxia Cao, Xin Cong, Tingwen Liu, and Bin Wang. [n. d.]. Item Similarity Mining for Multi-Market Recommendation. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval.*

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1).* Association for Computational Linguistics, 4171–4186.

[6] Hao Ding, Yifei Ma, Anoop Deoras, Yuyang Wang, and Hao Wang. 2021. Zero-Shot Recommender Systems. *CoRR* abs/2105.08318 (2021). arXiv:2105.08318 https://arxiv.org/abs/2105.08318

[7] Ziwei Fan, Zhiwei Liu, Alice Wang, Zahra Nazari, Lei Zheng, Hao Peng, and Philip S Yu. 2022. Sequential recommendation via stochastic self-attention. In *Proceedings of the ACM Web Conference 2022.* 2036–2047.

[8] Ziwei Fan, Zhiwei Liu, Chen Wang, Peijie Huang, Hao Peng, and S Yu Philip. 2022. Sequential Recommendation with Auxiliary Item Relationships via Multi-Relational Transformer. In *2022 IEEE International Conference on Big Data (Big Data).* IEEE, 525–534.

[9] Chen Gao, Xiangning Chen, Fuli Feng, Kai Zhao, Xiangnan He, Yong Li, and Depeng Jin. 2019. Cross-domain Recommendation Without Sharing User-relevant Data. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, Ling Liu, Ryen W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia (Eds.). ACM, 491–502. https://doi.org/10.1145/3308558.3313538

[10] Bowen Hao, Jing Zhang, Hongzhi Yin, Cuiping Li, and Hong Chen. 2021. Pre-training graph neural networks for cold-start users and items representation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining.* 265–273.

[11] Ruining He and Julian McAuley. 2016. Fusing similarity models with markov chains for sparse sequential recommendation. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE, 191–200.

[12] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 639–648.

[13] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939* (2015).

[14] Yupeng Hou, Zhankui He, Julian J. McAuley, and Wayne Xin Zhao. 2022. Learning Vector-Quantized Item Representation for Transferable Sequential Recommenders. *CoRR* abs/2210.12316 (2022). https://doi.org/10.48550/arXiv.2210.12316 arXiv:2210.12316

[15] Yupeng Hou, Shanlei Mu, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. [n. d.]. Towards Universal Sequence Representation Learning for Recommender Systems. In *KDD ’22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

[16] Zheng Hu and Fuji Ren. 2023. Bert4CMR: Cross-Market Recommendation with Bidirectional Encoder Representations from Transformer. *CoRR* abs/2305.15145 (2023).

[17] HyeoungGuk Kang, Donghoon Lee, and Hyunsouk Cho. 2023. Outlier-aware Cross-Market Product Recommendation. In *BigComp*. IEEE, 120–123.

[18] SeongKu Kang, Junyoung Hwang, Dongha Lee, and Hwanjo Yu. [n. d.]. Semi-Supervised Learning for Cross-Domain Recommendation to Cold-Start Users. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*.

[19] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 197–206.

[20] Jiacheng Li, Yujie Wang, and Julian McAuley. 2020. Time interval aware self-attention for sequential recommendation. In *Proceedings of the 13th international conference on web search and data mining*. 322–330.

[21] Pan Li and Alexander Tuzhilin. 2020. DDTCDR: Deep Dual Transfer Cross Domain Recommendation. In *WSDM ’20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, James Caverlee, Xia (Ben) Hu, Mounia Lalmas, and Wei Wang (Eds.). ACM, 331–339. https://doi.org/10.1145/3336191.3371793

[22] Hanxiao Liu, Yuexin Wu, and Yiming Yang. 2017. Analogical inference for multi-relational embeddings. In *International conference on machine learning*. PMLR, 2168–2178.

[23] Chen Ma, Peng Kang, and Xue Liu. 2019. Hierarchical gating networks for sequential recommendation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 825–833.

[24] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*. PMLR, 1273–1282.

[25] Ruihong Qiu, Zi Huang, Hongzhi Yin, and Zijian Wang. 2022. Contrastive learning for representation degeneration problem in sequential recommendation. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 813–823.

[26] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*. 811–820.

[27] 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 *Proceedings of the 28th ACM international conference on information and knowledge management*. 1441–1450.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[29] Chen Wang, Yueqing Liang, Zhiwei Liu, Tao Zhang, and Philip S. Yu. 2021. Pre-training Graph Neural Network for Cross Domain Recommendation. In *Third IEEE International Conference on Cognitive Machine Intelligence, CogMI 2021, Atlanta, GA, USA, December 13-15, 2021*. IEEE, 140–145. https://doi.org/10.1109/CogMI52975.2021.00026

[30] Yaqing Wang, Chunyan Feng, Caili Guo, Yunfei Chu, and Jenq-Neng Hwang. [n. d.]. Solving the Sparsity Problem in Recommendations via Cross-Domain Item Embedding Based on Co-Clustering. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11-15, 2019*, J. Shane Culpepper, Alistair Moffat, Paul N. Bennett, and Kristina Lerman (Eds.).

[31] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2014. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575* (2014).

[32] Chen Zhang, Yu Xie, Hang Bai, Bin Yu, Weihong Li, and Yuan Gao. 2021. A survey on federated learning. *Knowledge-Based Systems* 216 (2021), 106775.

[33] Hongyu Zhang, Dongyi Zheng, Xu Yang, Jiyuan Feng, and Qing Liao. 2024. FedDCSR: Federated cross-domain sequential recommendation via disentangled representation learning. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*. SIAM, 535–543.

[34] Yinan Zhang, Yong Liu, Peng Han, Chunyan Miao, Lizhen Cui, Baoli Li, and Haihong Tang. 2020. Learning Personalized Itemset Mapping for Cross-Domain Recommendation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, Christian Bessiere (Ed.). ijcai.org, 2561–2567. https://doi.org/10.24963/ijcai.2020/355

[35] Cheng Zhao, Chenliang Li, Rong Xiao, Hongbo Deng, and Aixin Sun. [n. d.]. CATN: Cross-Domain Recommendation for Cold-Start Users via Aspect Transfer Network. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.).

[36] Lili Zhao, Sinno Jialin Pan, and Qiang Yang. 2017. A unified framework of active transfer learning for cross-system recommendation. *Artif. Intell.* 245 (2017), 38–55. https://doi.org/10.1016/j.artint.2016.12.004

[37] Wayne Xin Zhao, Shanlei Mu, Yupeng Hou, Zihan Lin, Yushuo Chen, Xingyu Pan, Kaiyuan Li, Yujie Lu, Hui Wang, Changxin Tian, Yingqian Min, Zhichao Feng, Xinyan Fan, Xu Chen, Pengfei Wang, Wendi Ji, Yaliang Li, Xiaoling Wang, and Ji-Rong Wen. 2021. RecBole: Towards a Unified, Comprehensive and Efficient Framework for Recommendation Algorithms. In *CIKM*. ACM, 4653–4664.

[38] Xinping Zhao and Yingchun Yang. 2022. SGLCMR: Self-supervised Graph Learning of Generalized Representations for Cross-Market Recommendation. In *IJCNN*. IEEE, 1–8.

[39] Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 2020. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 1893–1902.

[40] Feng Zhu, Chaochao Chen, Yan Wang, Guanfeng Liu, and Xiaolin Zheng. [n. d.]. DTCDR: A Framework for Dual-Target Cross-Domain Recommendation. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019*.

[41] Feng Zhu, Yan Wang, Chaochao Chen, Guanfeng Liu, Mehmet A. Orgun, and Jia Wu. 2018. A Deep Framework for Cross-Domain and Cross-System Recommendations. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, Jérôme Lang (Ed.). ijcai.org, 3711–3717. https://doi.org/10.24963/ijcai.2018/516