Multi-Domain Enhancement via Residual Interwoven Transfer in Cross-Domain Sequential Recommendation

Qingtian Bian
College of Computing and Data Science, Nanyang Technological University
Singapore
bian0027@e.ntu.edu.sg

Tieying Li
School of Computer Science and Engineering, Northeastern University
China
tieying@stumail.neu.edu.cn

Marcus de Carvalho
College of Computing and Data Science, Nanyang Technological University
Singapore
ivsucram@gmail.com

Jiaxing Xu
College of Computing and Data Science, Nanyang Technological University
Singapore
jiaxing003@e.ntu.edu.sg

Hui Fang
Key Laboratory of Interdisciplinary Research of Computation and Economics, School of Computing and Artificial Intelligence, Shanghai University of Finance and Economics
China
fang.hui@mail.shufe.edu.cn

Yiping Ke
College of Computing and Data Science, Nanyang Technological University
Singapore
ypke@ntu.edu.sg

Abstract

To mitigate data sparsity in Sequential Recommendation, Cross-Domain Sequential Recommendation (CDSR) exploits dynamic knowledge transfer across domains. Traditional CDSR approaches merge specific-domain sequences into mixed-domain sequences to reconnect users' dispersed interests. However, most methods rely on unidirectional transfer between mixed and specific domains on each domain task, overlooking the complex interplay between mixed-domain and domain-specific dynamics. Moreover, token-level transfer between coinciding domain sequences fails to consider inherent sequential dynamics. To address these limitations, we propose Multi-Domain Enhancement via Residual Interwoven Transfer (**MERIT**). Specifically, MERIT enhances domain representations along multiple domain-to-domain paths, leveraging the proposed extended cross-attention fusion compatible with partially overlapping sequences. To facilitate such transfers, MERIT further employs MoE networks in encoders to generate both intra-domain and inter-domain representations. In addition, by integrating stopped-gradient mixed-domain representations into specific-domain representations, MERIT enables the model to learn the residual signal of the mixed-domain information, better aligning with downstream specific-domain tasks. Extensive experiments on three real-world datasets demonstrate that MERIT consistently outperforms state-of-the-art CDSR counterparts with statistical significance. The code is available at: https://github.com/DiMarzioBian/MERIT/.

CCS Concepts

- **Information systems → Recommender systems**.

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3754589

Keywords

Recommender System, Sequential Recommendation, Cross-Domain Sequential Recommendation, Interwoven Transfer

ACM Reference Format:
Qingtian Bian, Tieying Li, Marcus de Carvalho, Jiaxing Xu, Hui Fang, and Yiping Ke. 2025. Multi-Domain Enhancement via Residual Interwoven Transfer in Cross-Domain Sequential Recommendation. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3754589

1 Introduction

Data sparsity is a significant challenge in recommender systems, where many users and items exhibit insufficient interactions. Cross-Domain Sequential Recommendation (CDSR) mitigates this issue by transferring dynamic knowledge from source to target domains. This approach can benefit both the launch of new business domains and the mutual enhancement across related domains.

CDSR assumes that users display diverse and dynamic interests, with some interests being domain-invariant while others remain domain-specific [3, 56]. For example, as illustrated in Figure 1, a user's engagement with fantasy literature, as evidenced by interactions with *Harry Potter*, *The Lord of the Rings*, and *The Hobbit*, reflects an invariant interest across the movie and book domains. In contrast, a preference for comedy, demonstrated by interactions with *Mr. Bean's Holiday* and *Johnny English 3*, is confined to the movie domain. In comparison, an interest in biography, as evidenced by interactions with *Steve Jobs: A Biography*, is specific to the book domain. While each domain sequence captures a distinct aspect of user interest, the CDSR model aims to integrate these different domain sequences holistically.

Most CDSR methods combine specific-domain sequences into mixed-domain sequences in chronological order. This approach fuses domain-invariant interests dispersed across specific-domain sequences to bridge different domains. Based on whether leveraging mixed-domain sequence and the transfer between mixed-

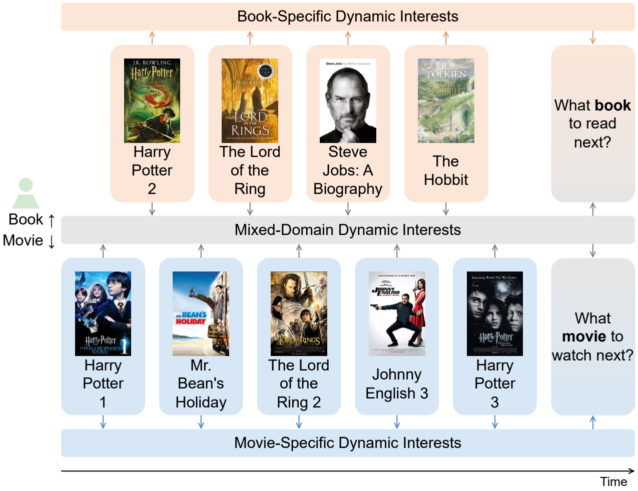


Figure 1: An example of dual-target CDSR task.

and specific-domain sequences, existing CDSR methods can be categorized into three types: no explicit transfer [34], specific-to-specific transfer [1, 2, 22, 49], unidirectional mixed-to-specific transfer [3, 7, 25, 47, 48, 51, 54], and asymmetric transfer including mixed-to-specific, target-to-mixed, and target-to-source transfers [26]. However, domain-invariant interests exist within every related domain, enabling each domain pair to use these shared interests as a bridge for mutual transfer, regardless of whether the domains are specific or mixed. Existing methods overlook this aspect, leading to suboptimal performance.

Additionally, most previous studies have been confined to token-level transfer in simple linear aggregation, where tokens represent individual interacted items. Such transfer neglects the interest dynamics within the sequences. Furthermore, the ground truth in mixed-domain and specific-domain sequences associated with the same item can be different and may belong to different domains. Performing token-level transfer in this case would introduce domain-specific knowledge into domains where it is not applicable, thus degrading performance.

To address these challenges, we propose the Multi-domain Enhancement via Residual Interwoven Transfer (**MERIT**). MERIT facilitates interwoven residual transfer by constructing domain-to-domain knowledge transfer for each possible domain pair under sequence-level fusion. Specifically, MERIT leverages MoE to instantiate multi-output FFN (moFFN), enabling each encoder to generate intra-domain representations as the main branch and inter-domain representations for transfer to other domain branches. Then, to allow the transfer of knowledge in partially overlapping sequences, MERIT proposes extended cross-attention fusion modules (ECAF) for aggregation. Eventually, MERIT aggregates stopped-gradient mixed-domain representations into specific-domain representations, enabling the learning of residual signals from the mixed-domain information and better aligning with downstream tasks.

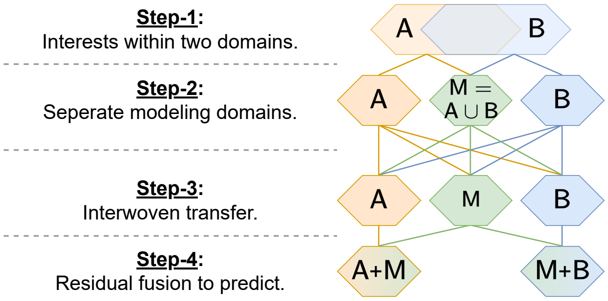


Figure 2: Framework of MERIT.

To thoroughly evaluate MERIT, we compare it with state-of-the-art (SOTA) CDSR models to assess its effectiveness, and also with Sequential Recommendation (SR) models in single- and dual-target settings to investigate the rationality of the proposed residual learning. Extensive experimental results on three real-world datasets show that MERIT outperforms SOTA CDSR baselines across all metrics with statistical significance. In short, the contributions of this paper are summarized as follows:

- We propose an interwoven residual transfer approach that enables sequence-level knowledge transfer across all possible domain pairs, thereby overcoming the limitations of unidirectional transfer methods.
- To implement this approach, we propose moFFN to generate additional inter-domain representations and ECAF modules to integrate these partially overlapped sequential representations into different intra-domain representations.
- We conduct extensive experiments on three real-world datasets to evaluate the performance of MERIT. Results show that our model surpasses state-of-the-art CDSR counterparts with statistical significance.

2 Related Work

2.1 Sequential Recommendation

Sequential Recommendation (SR) tasks aim to model interest dynamics based on the inherent chronological order of user interactions. Early works [36] employed Markov Chains to model sequential interaction transitions. With the advent of deep learning, various complex networks have been utilized to capture such sequential dynamics. These include recurrent neural networks [13, 18, 44], convolutional neural networks [40, 50, 55], and graph neural networks [9, 43, 46]. In recent years, self-attention-based models have demonstrated significant capability in SR tasks [21, 23, 37], leveraging advancements in the Transformer family [12, 41, 42].

To further enhance SR models, researchers have sought various types of side information to improve performance. Common methods included introducing item attributes [20, 35, 52], user social networks [6, 11, 15], and temporal information [4, 14, 45]. With the development of LLM, recent works effectively integrate multi-modality or multimedia information, such as images, text, and video [5, 31, 32]. From this perspective, CDSR can also be seen as an extension of SR that leverages domains as side information.

2.2 Cross-Domain Sequential Recommendation

Researchers propose Cross-Domain Recommendation (CDR) to mitigate data sparsity issues by transferring knowledge from source domains to related target domains [8, 19, 24]. The Cross-Domain Sequential Recommendation (CDSR) task incorporates this mindset into SR tasks, aiming to leverage transferable dynamic knowledge to enhance performance in target domains. Early CDSR works focused on the shared-account task [16, 17, 27, 28, 38], which assumes

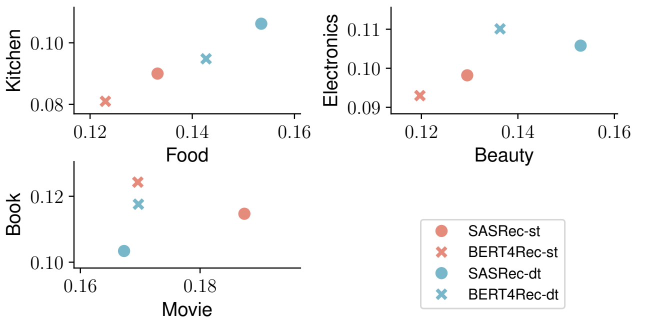


Figure 3: MRR scores of SASRec-st, BERT4Rec-st, SASRec-dt, and BERT4Rec-dt.

multiple anonymous users share the same account. However, this scenario significantly differs from the general CDSR task.

The general CDSR task assumes that the same interest may lead to different user behaviors across various domains [10]. To leverage these domain-shared interests, one widespread practice is to intersperse specific-domain sequences into mixed-domain sequences. To better bridge domains, existing methods can be categorized into several types based on whether they leverage mixed-domain sequences and the transfer directions.

Several studies focus on transferring knowledge directly between specific domains [1, 2, 22, 33, 49], while others transfer knowledge from a mixed domain to specific domains [3, 7, 25, 47, 48, 51, 53, 54]. TriCDR [26] employs an asymmetric transfer that supports mixed-to-specific, target-to-mixed, and target-to-source pathways, while SynCRec [34] introduces selectively shared experts to mitigate negative transfer between specific and mixed domains without explicit transfer mechanisms. However, none of these works reflect the interplay between specific-domain and mixed-domain sequences for various downstream tasks.

Moreover, most methods fuse intra-domain and transferred inter-domain representations at the token level, with MAN [25] being an exception that uses sequence-level fusion for mixed-to-specific transfers. However, within the Seq2Seq paradigm, token-level fusion often fails to capture full sequential dynamics, especially when targets vary across sequences and their domain affiliations are uncertain in mixed-domain sequences.

3 Preliminaries

To evaluate the effectiveness of mixed-domain sequences, we apply SASRec [21] and BERT4Rec [12] in mixed-domain sequences, denoted as SASRec-dt and BERT4Rec-dt. These variants are trained on mixed-domain sequences by calculating losses separately for each domain and summing them as supervision. We assess their performance using Mean Reciprocal Rank (MRR) on the Food-Kitchen (FK) and Beauty-Electronics (BE) datasets following [3]. The original single-domain variants are denoted as SASRec-s and BERT4Rec-s.

As shown in Figure 3, the mixed-domain variants outperform the original single-domain counterparts on both domains in the FK and BE datasets, and on the MB dataset, BERT4Rec-dt achieves close performance to BERT4Rec. These results indicate that both variants can attain closed performance even under different setups and sequential dynamics.

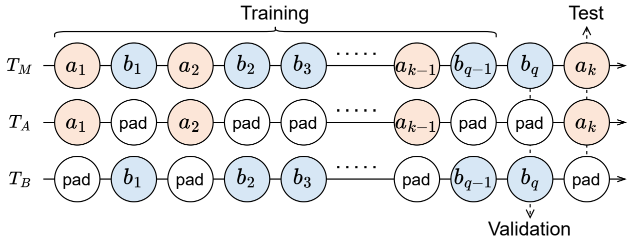


Figure 4: An example of sequences partition and splits.

4 Methodology

In this section, we propose MERIT as depicted in Figure 5. The introduction of MERIT comprises five subsections: 1) formulation of the problem; 2) sequence formulation and embedding layers; 3) sequence encoder; 4) Cross-Attention Fusion modules and interwoven transfer; and 5) model training. We will introduce each part in the following subsections.

4.1 Problem Formulation

We consider CDSR as an extension of SR that incorporates domain information. Consequently, CDSR models leveraging side information should outperform SR models; otherwise, their application would be limited. In this paper, we focus on the dual-target CDSR task and formally define it as follows:

Let $\mathcal{I}_\text{A}$ and $\mathcal{I}_\text{B}$ be the full item set of domains A and B, respectively. Given one user's two interaction sequences $T_\text{A} = (a_1, a_2, a_3, \ldots, a_k)$ within $\mathcal{I}_\text{A}$ and $T_\text{B} = (b_1, b_2, b_3, \ldots, b_q)$ within $\mathcal{I}_\text{B}$, the model is tasked with transferring knowledge between T_A and T_B to generate top-N recommendations for the next item to interact in each domain, a_{k+1} and b_{q+1}. The task can be formulated as:

Input: One user's specific-domain sequences, $T_\text{A} = (a_1, a_2, a_3, \ldots, a_k)$ and $T_\text{B} = (b_1, b_2, b_3, \ldots, b_q)$.

Output: A recommendation system that estimates the probability of this users' next items to interact, a_{k+1} and b_{q+1}.

4.2 Sequence Formulation and Embeddings

Continue on the example given in Section 4.1, we denote the union of domain A and B as domain M, and its full item set is $\mathcal{I}_\text{M} = \mathcal{I}_\text{A} \cup \mathcal{I}_\text{B}$. Following the widely used leave-one-out sequence partition [3, 21, 37, 40], the last and the penultimate interactions of each T_M are popped out as the testing and validation ground truths. Next, T_A and T_B are aligned with T_M based on the same interacted items, and any empty slots are padded to ensure that T_A and T_B are of the same length as T_M. Figure 4 provides a detailed illustration of this partitioning and alignment process.

To convert numerical indices into embeddings, each item will be embedded into one learnable vector at length d, where d denotes the dimension of embedding space. The table of item embeddings is denoted as $\mathbf{E}^{\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}| \times d}$, including all items from both domains. We also assign the positional indices for the interactions within each domain sequence separately in reversed chronological order. These positional indices are also embedded into vectors at length d, and we denote this positioned embedding table as $\mathbf{E}^{\mathcal{P}} \in \mathbb{R}^{L \times d}$, where L represents the input length. All input sequences not at length L will be truncated or padded to meet the criterion.

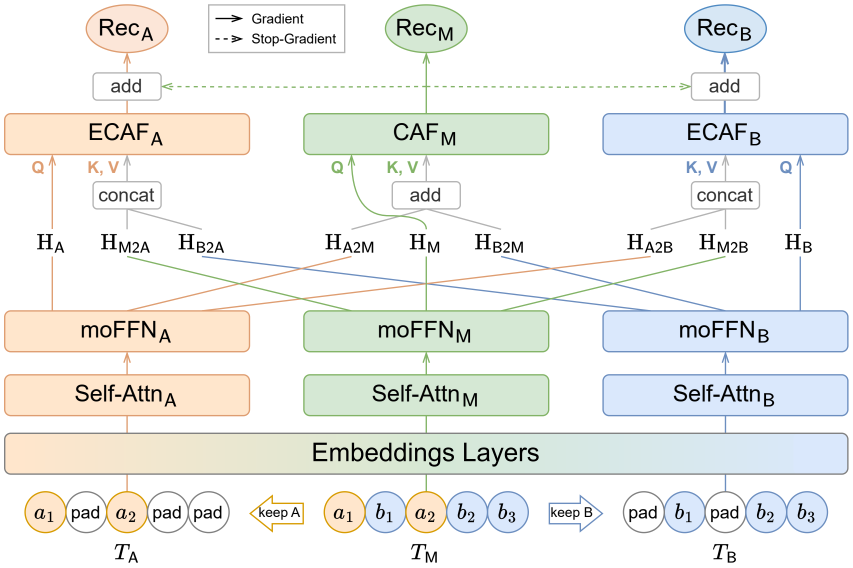


Figure 5: Overall architecture of the proposed MERIT model.

In the following subsections, we denote the embedding outputs for T_{M}, T_{A}, and T_{B} as $\mathbf{E}_{\mathrm{M}}$, $\mathbf{E}_{\mathrm{A}}$, and $\mathbf{E}_{\mathrm{B}} \in \mathbb{R}^{L \times d}$, which are the summation of the embeddings of interacted item and respective in-sequence positional embeddings.

4.3 Sequence Encoder

Inspired by SASRec [21], we instantiate one self-attention sequential encoder for each domain, mixed and specific. Specifically, each encoder consists of one self-attention layer and one multi-output Feedforward Network (moFFN).

4.3.1 ***Self-Attention Layer.*** To implement the self-attention mechanism, we input the same sequential embeddings as queries (Q), keys (K), and values (V) projections into the Multi-Head Attention (MHA) network [42]. For example, the mixed-domain sequential representation $\mathbf{H}_{\mathrm{M}}$ is computed from its embedding $\mathbf{E}_{\mathrm{M}}$ as follows

$$\mathbf{H}_{\mathrm{M}}^{\mathrm{SA}} = \text{Self-Attn}_{\mathrm{M}}\left(\mathbf{E}_{\mathrm{M}}\right) = \mathrm{MHA}_{\mathrm{M}}\left(\mathbf{E}_{\mathrm{M}},\ \mathbf{E}_{\mathrm{M}},\ \mathbf{E}_{\mathrm{M}}\right), \tag{1}$$

$$\mathbf{H}_{\mathrm{M}}^{\mathrm{SA}} \leftarrow \mathrm{LayerNorm}\left(\mathbf{E}_{\mathrm{M}} + \mathrm{Dropout}\left(\mathbf{H}_{\mathrm{M}}^{\mathrm{SA}}\right)\right). \tag{2}$$

Similarly, the specific-domain sequential representations $\mathbf{H}_{\mathrm{A}}$ and $\mathbf{H}_{\mathrm{B}}$ can be obtained by replacing the notation M with A or B.

4.3.2 ***moFFN.*** We argue that while representations modeled within one domain capture both domain-specific and domain-shared dynamics, the transferable inter-domain representations tend to selectively emphasize those aspects most beneficial to their target domains. To address this limitation, we leverage multi-task learning and employ CGC networks [39] as our moFFNs. Specifically, the computation pipeline for the k-th output $\mathbf{H}_k$ in the moFFN is formulated as:

$$\mathrm{Expert}\left(\mathbf{H}\right) = \left(\mathrm{Swish}\left(\mathbf{H}\mathbf{W}_1\right) \otimes \mathbf{H}\mathbf{W}_2\right)\mathbf{W}_3, \tag{3}$$

$$S_k = \left[\mathrm{Expert}_s, \mathrm{Expert}_k\right], \tag{4}$$

$$\mathrm{Gate}_k\left(\mathbf{H}\right) = \mathrm{Softmax}\left(W_k \mathbf{H}\right), \tag{5}$$

$$\mathbf{H}_k = \mathrm{LayerNorm}\left(\mathbf{H} + \mathrm{Dropout}\left(\mathrm{Gate}_k(\mathbf{H}) S_k(\mathbf{H})\right)\right), \tag{6}$$

where $W_k \in \mathbb{R}^{d \times 2}$ is the learnable matrix for the gate corresponding to the kth output, Expert_k the domain-specific expert network for k-th output, and Expert_s denotes the expert network shared across all outputs. For each expert network, we follow the feedforward network architecture from Llama [41], employing learnable matrices $\mathbf{W}_1 \in \mathbb{R}^{d \times \frac{8}{3}d}$, $\mathbf{W}_2 \in \mathbb{R}^{d \times \frac{8}{3}d}$, and $\mathbf{W}_3 \in \mathbb{R}^{\frac{8}{3}d \times d}$, with the Swish activation function defined as $\mathrm{Swish}(x) = \frac{x}{1+e^{-x}}$.

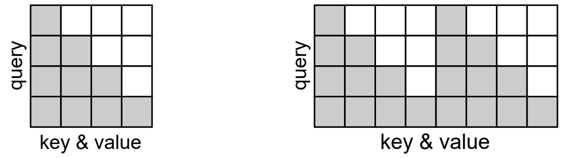
Each moFFN outputs comprise one intra-domain representation used for the domain's task and multiple inter-domain representations for transfer to other domains. In MERIT, the moFFN outputs are as follows:

- $\mathrm{moFFN}_{\mathrm{M}}$ takes $\mathbf{H}_{\mathrm{M}}^{\mathrm{SA}}$ and outputs $\mathbf{H}_{\mathrm{M}}$, $\mathbf{H}_{M2\mathrm{A}}$, and $\mathbf{H}_{M2\mathrm{B}}$.
- $\mathrm{moFFN}_{\mathrm{A}}$ takes $\mathbf{H}_{\mathrm{A}}^{\mathrm{SA}}$ and outputs $\mathbf{H}_{\mathrm{A}}$, $\mathbf{H}_{A2\mathrm{M}}$, and $\mathbf{H}_{A2\mathrm{B}}$.
- $\mathrm{moFFN}_{\mathrm{B}}$ takes $\mathbf{H}_{\mathrm{B}}^{\mathrm{SA}}$ and outputs $\mathbf{H}_{\mathrm{B}}$, $\mathbf{H}_{B2\mathrm{M}}$, and $\mathbf{H}_{B2\mathrm{A}}$.

Here, each term following the $\mathbf{H}_{S2T}$ pattern represents an inter-domain representation, signifying the output is generated from source domain S for transfer to target domain T. In contrast, each term denoted by $\mathbf{H}_{Y}$ pattern represents an intra-domain representation that directly contributes to tasks within domains Y.

4.4 Cross-Attention Fusion Modules and Interwoven Transfer

To enable sequence-level interwoven transfer where each domain transfers to every other domain, we propose Cross-Attention Fusion (CAF) modules. Recall from Section 4.2 and Figure 4 that T_{A} and T_{B} are derived from T_{M} by retaining only the interactions from their respective domains. Consequently, when fusing transferred representations from the mixed-domain sequence T_{M} with those from the coinciding specific-domain sequence $T_{A/B}$, direct token-wise addition distorts the norm of the resulting sequence. To address this, we apply token-wise addition exclusively for the mixed domain, while for specific-domain sequences, we use concatenation along the length dimension to form the transferred representations, as



Causal Mask for CAF_M Causal Mask for ECAF_A and ECAF_B

Figure 6: Illustration of exemplary 4-token cross-attention causal masks used in CAF_M and $\mathrm{ECAF}_{A/B}$. Gray cells denote visible tokens, whereas white cells represent masked tokens.

follows:

$$\mathbf{H}_{2\mathrm{M}} = \mathbf{H}_{A2\mathrm{M}+B2\mathrm{M}} = (\mathbf{H}_{A2\mathrm{M}} + \mathbf{H}_{B2\mathrm{M}}) \in \mathbb{R}^{L\times d}, \tag{7}$$

$$\mathbf{H}_{2\mathrm{A}} = \mathbf{H}_{M2\mathrm{A}||B2\mathrm{A}} = \mathrm{Concat}\left(\mathbf{H}_{M2\mathrm{A}}, \mathbf{H}_{B2\mathrm{A}}\right) \in \mathbb{R}^{2L\times d}, \tag{8}$$

$$\mathbf{H}_{2\mathrm{B}} = \mathbf{H}_{M2\mathrm{B}||A2\mathrm{B}} = \mathrm{Concat}\left(\mathbf{H}_{M2\mathrm{B}}, \mathbf{H}_{A2\mathrm{B}}\right) \in \mathbb{R}^{2L\times d}, \tag{9}$$

The pipeline for obtaining the final M-domain representation is detailed as follows:

$$\mathbf{H}_\mathrm{M}^\mathrm{F} = \mathrm{CAF}_\mathrm{M}\left(\mathbf{H}_\mathrm{M},\ \mathbf{H}_{2\mathrm{M}}\right) = \mathrm{MHA}_\mathrm{M}^{CAF}\left(\mathbf{H}_\mathrm{A},\ \mathbf{H}_{2\mathrm{M}},\ \mathbf{H}_{2\mathrm{M}}\right), \tag{10}$$

$$\mathbf{H}_\mathrm{M}^\mathrm{F} \leftarrow \mathrm{LayerNorm}\left(\mathbf{H}_\mathrm{M} + \mathrm{Dropout}\left(\mathbf{H}_\mathrm{M}^\mathrm{F}\right)\right), \tag{11}$$

$$\mathbf{H}_\mathrm{M}^F \leftarrow \mathrm{LayerNorm}\left(\mathbf{H}_\mathrm{M}^\mathrm{F} + \mathrm{Dropout}\left(\mathrm{FFN}_\mathrm{M}\left(\mathbf{H}_\mathrm{M}^\mathrm{F}\right)\right)\right), \tag{12}$$

where FFN_M adheres to the expert network architecture defined in Equation 3.

The final specific-domain representations $\mathbf{H}_\mathrm{A}^F$ and $\mathbf{H}_\mathrm{B}^F$ are obtained by substituting M with A and B, and substituting CAF_M with ECAF_A and ECAF_B, respectively. Notably, the causal masks used here are illustrated in Figure 6, we extend the causal masks in ECAF_A and ECAF_B by an additional clone on the length dimension to accommodate the length of concatenated sequences.

4.5 Model Training and Residual Learing

Because the optimization of the M, A, and B domains are all composed of recommendation tasks, we adopt InfoNCE [30] to compute the losses. Given the input final representation h, we denote the embedding of its ground truth item as e^+, and the embedding set of its negative samples as $\mathbf{E}^-$. Hence, the InfoNCE can be written as:

$$\mathrm{InfoNCE}\left(h,\ e^+,\ E^-\right) = -\log\frac{exp\left(h\cdot e^+/\tau\right)}{\sum_{e\in\{e^+,\ E^-\}} exp\left(h\cdot e/\tau\right)}, \tag{13}$$

where τ denotes the temperature.

For i-th interacted item of T_M, $h_{m,i}^F$ denotes its final representation from $\mathbf{H}_\mathrm{M}^F$. Its corresponding ground truth embedding is $e_{m,i}^+$. We randomly sample N_neg unobserved items from $\mathcal{I}_\mathrm{A} \setminus T_\mathrm{M}$ and another N_neg unobserved items from $\mathcal{I}_\mathrm{B} \setminus T_\mathrm{M}$ to form the negative sample set $E_{m,i}^-$. Then, this user’s M-domain loss can be formulated as follows:

$$\mathcal{L}_\mathrm{M} = \frac{1}{|T_\mathrm{M}|}\sum_{i=1}^{|T_\mathrm{M}|}\mathrm{InfoNCE}\left(h_{m,i}^F,\ e_{m,i}^+,\ E_{m,i}^-\right), \tag{14}$$

where the padding interactions in $|T_\mathrm{M}|$ are not considered.

Second, taking the same user’s A-domain training sequence T_A as an example, for its j-th interacted item, $h_{a,j}^F$ denotes its final representation from $\mathbf{H}_\mathrm{A}^F$, $e_{a,j}^+$ denotes the corresponding ground truth embedding, and $h_{M,j}^F$ denotes the corresponding final representation of the same interacted item in T_M. We only randomly sample N_neg unobserved items from $\mathcal{I}_\mathrm{A} \setminus T_\mathrm{A}$ to form the negative sample set $E_{a,j}^-$. Then, this user’s A-domain loss can be formulated as follows:

Table 1: Statistics of CDSR Datasets.

Dataset	#users	#items	#interactions	#val.	#test
Food	7,144	11,837	83,663	2,837	2,419
Kitchen		16,258	89,885	4,307	4,725
Beauty	4,474	10,379	50,329	2,086	1,875
Electronics		14,188	63,800	2,388	2,599
Movie	28,350	35,712	347,654	11,728	10,935
Book		90,958	403,147	16,622	17,415

$$\mathcal{L}_\mathrm{A} = \frac{1}{|T_\mathrm{A}|}\sum_{j=1}^{|T_\mathrm{A}|}\mathrm{InfoNCE}\left(h_{a,j}^F + \mathrm{SG}\left(h_{m,j}^F\right),\ e_{a,j}^+,\ E_{a,j}^-\right), \tag{15}$$

where SG denotes the stop-gradient operator, and the padding interactions in $|T_\mathrm{A}|$ are not considered. By replacing the notation A with B, the B-domain loss $\mathcal{L}_\mathrm{B}$ can be obtained similarly.

Recall from Section 3 that the mixed domain demonstrates its potential when solely addressing CDSR tasks. Accordingly, we leverage the final specific-domain representations, $\mathbf{H}_\mathrm{B}^F$ and $\mathbf{H}_\mathrm{B}^F$ to learn residuals that adapt mixed-domain predictions from $\mathbf{H}_\mathrm{M}^F$ to the corresponding A- and B-domain predictions. A gradient-stopping operation is applied to $\mathbf{H}_\mathrm{M}^F$ to prevent coupling, which would otherwise allow specific-domain predictions to interfere adversely with the optimization of mixed-domain predictions.

Finally, the overall loss can thus be given as:

$$Loss = \mathcal{L}_\mathrm{M} + \mathcal{L}_\mathrm{A} + \mathcal{L}_\mathrm{B}. \tag{16}$$

Additionally, the computational complexity of MERIT arises from six multi-head attention (MHA) networks, comprising three self-attention and three cross-attention modules, three FFNs, and three moFFNs. Specifically, each self-attention layer and CAF_M require $dL^2 + Ld^2$ operations, while each CAF for A and B requires $4dL^2 + 2Ld^2$ operations. Each FFN incurs a cost of df operations, and each moFFN requires $4df + 6d$ operations, where f is the internal dimension of the expert network and $f>d$. Consequently, the overall complexity is $O(dL^2 + Ld^2 + df)$. For deployment, we leverage FastAttention in PyTorch 2.2.1 to accelerate computation.

5 Experiments

In this section, we conduct extensive experiments on three publicly available real-world datasets to evaluate the effectiveness of MERIT. We aim to address the following research questions (RQ) in this section.

RQ1: Overall, how does MERIT perform against state-of-the-art CDSR models and other baseline methods?
RQ2: How do mixed- and specific-domain cross-domain representations in ECAF and CAF modules affect interwoven transfer?
RQ3: How does the proposed residual learning affect the final performance?
RQ4: How are the experts utilized in each output of MoE-FFNs?

5.1 Datasets

We conduct experiments on three SR datasets pairs, including six domains from Amazon review datasets[1] [29]: Food-Kitchen (**FK**, consisting of 'Grocery and Gourmet Food' as Food, and 'Home and Kitchen' as Kitchen), Beauty-Electronics (**BE**, with 'Beauty' as Beauty, and 'Electronics' as Electronics), and Movie-Book (**MB**, with 'Movies and TV' as Movie, and 'Books' as Book). The HVIDEO dataset [16, 27, 28, 38] is not used because it focuses on shared-account cross-domain sequential recommendation, where each account is shared by multiple individuals, differing fundamentally from general CDSR, which assumes part of user's interests is shared across domains. Similarly, the Douban dataset [57, 58] is excluded as its timestamp granularity is limited to days, making the order of interactions within the same day subjective to authors' decisions, reducing reproducibility and potentially allowing cherry-picking.

For all datasets, each review is treated as an interaction. To alleviate the influence of cold-start users and items, we first filter all non-domain-overlapping users and only retain interaction items that have been interacted with more than 10 times among domain-overlapping users. To reduce the computation load, we keep up to $L = 50$ latest interactions of all sequences, following [21], and filter all users with less than 5 interactions in each domain due to shortened sequences. The statistics of these datasets are summarized in Table 1. We run all methods with five seeds to obtain the experimental results. We use Hit Rate (HR), Normalized Discounted Cumulative Gain (NDCG), and Mean Reciprocal Rank (MRR) as our metrics. The parameters for single-target models are chosen based on the MRR score on the target domain, while those for dual-target models are selected based on the sum of MRR scores on both domains.

5.2 Baselines

We compare MERIT against four categories of baseline models, including **ST-SR** models (SASRec-st and BERT4Rec-st), **DT-SR** models (SASRec-dt and BERT4Rec-dt), **ST-CDSR** models (CD-SASRec, CD-ASR, and MGCL) and **DT-CDSR** models (C^2DSR, DREAM, and ABXI). Specifically, ST-SR models are trained independently on each domain, whereas DT-SR models are trained on both domains combined with losses computed separately for each domain to prevent bias. In contrast, ST-CDSR models are also trained on the combined domains, but evaluation metrics are computed solely on the target domain, so each ST-CDSR model is executed twice, once with domain A as the target and once with domain B. Finally, DT-CDSR models are trained and evaluated by computing metrics for both domains concurrently within the same training.

- **SASRec** [21] is an attention-based milestone SR model that implements self-attention encoder layers to capture sequential dynamics. Its single-target and dual-target versions are SASRec-st and SASRec-dt.
- **BERT4Rec** [37], built upon SASRec, enhances sequential representation through bi-directional self-attention modeling, leveraging the Cloze objective. BERT4Rec-st and BERT4Rec-dt are its single-target and dual-target versions.
- **CD-SASRec** [2] is one of the pioneering works of CDSR. It leverages self-attention to encode source-domain sequences and integrates the aggregated representation into encoding target-domain sequences.
- **CD-ASR** [1] aggregates source and target domain sequential dynamics using multiplicative attention and self-attention, respectively, to generate the final representation.
- **MGCL** [48] is a multi-view framework that harnesses contrastive learning to integrate graphical and sequential information across cross-domain and specific-domain views, enhancing the modeling of target domain dynamics.
- **C^2DSR** [7] leverages domain-based random replacement in negative samples to facilitate contrastive learning, thereby strengthening the connections between cross-domain and specific-domain representations.
- **DREAM** [51] extracts specific-domain sequential representations and adaptively injects them into cross-domain sequential modeling to enhance the capture of global interests.
- **ABXI** [3] employs task-guide alignment for specific-domain sequences and adapts domain-shared interests to specific-domain modeling.

5.3 Implementation Details.

We adopt the commonly used leave-one-out strategy in the sequential recommendation, where the last and the penultimate interactions of each sequence are popped out as the testing and validation ground truths, regardless of their domains. This is to ensure a fair comparison between CDSR and SR models. The number of ground truths in the validation and testing sets categorized by respective domains is provided in Table 1. For simplicity and fairness in evaluation, we adopt the common strategy used in [21, 37, 40] by sampling 999 negative items from the ground truth's domain to form the candidate set with the ground truth item. The selection of unmentioned hyperparameters is aligned with [3].

5.4 Recommendation Performance (RQ1)

In this subsection, we quantitatively evaluate the performance of MERIT against a range of baselines. Table 2 summarizes the overall results, from which several observations emerge.

First, across all six domains from three datasets, MERIT outperforms all baselines, including state-of-the-art CDSR counterparts on every metric. Paired t-tests reveal that differences in 2 metrics are statistically significant ($p < 0.03$), and the remaining 22 of 24 metrics are highly statistically significant ($p < 0.01$). These results confirm that MERIT effectively boosts performance by leveraging residual interleaving transfer.

Second, dual-target models (DT-SR and DT-CDSR) generally outperform single-target models (ST-SR and ST-CDSR), likely because our experimental settings do not intentionally create rich-sparse domain differences that typically favor single-target models. Moreover, ST-CDSR models outperform ST-SR models on 19 of 24 metrics, while DT-SDSR models exceed DT-SR models, thereby demonstrating the effectiveness of the CDSR paradigm.

5.5 Ablation Studies (RQ2, 3)

To investigate the rationality and effectiveness of employing the interwoven transfer approach, the CAF modules, and the introduced

---

[1] https://cseweb.ucsd.edu/~jmcauley/datasets/amazon/links.html

Table 2: Recommendation Performance (RQ1): Best results are presented in bold and the runner-up is underlined. Statistical significance of pairwise differences between MERIT and the best baseline is determined using a paired t-test; p-values ≤ 0.03 are denoted by * and those ≤ 0.01 by **.

Methods	Food				Kitchen			
	HR@5	HR@10	NDCG@10	MRR	HR@5	HR@10	NDCG@10	MRR
SASRec-st	0.1930$_{\pm0.0028}$	0.2611$_{\pm0.0036}$	0.1561$_{\pm0.0021}$	0.1332$_{\pm0.0019}$	0.1241$_{\pm0.0026}$	0.1851$_{\pm0.0018}$	0.1040$_{\pm0.0005}$	0.0900$_{\pm0.0007}$
BERT4Rec-st	0.1819$_{\pm0.0035}$	0.2528$_{\pm0.0037}$	0.1462$_{\pm0.0027}$	0.1230$_{\pm0.0030}$	0.1114$_{\pm0.0040}$	0.1685$_{\pm0.0036}$	0.0926$_{\pm0.0029}$	0.0810$_{\pm0.0025}$
CD-SASRec	0.1797$_{\pm0.0079}$	0.2454$_{\pm0.0046}$	0.1421$_{\pm0.0060}$	0.1197$_{\pm0.0066}$	0.1119$_{\pm0.0067}$	0.1757$_{\pm0.0070}$	0.0946$_{\pm0.0045}$	0.0821$_{\pm0.0039}$
CD-ASR	0.1976$_{\pm0.0042}$	0.2727$_{\pm0.0052}$	0.1616$_{\pm0.0028}$	0.1368$_{\pm0.0026}$	0.1345$_{\pm0.0043}$	0.1995$_{\pm0.0044}$	0.1107$_{\pm0.0037}$	0.0941$_{\pm0.0034}$
MGCL	0.1932$_{\pm0.0041}$	0.2673$_{\pm0.0054}$	0.1523$_{\pm0.0021}$	0.1260$_{\pm0.0018}$	0.1467$_{\pm0.0009}$	0.2157$_{\pm0.0026}$	0.1203$_{\pm0.0019}$	0.1017$_{\pm0.0019}$
SASRec-dt	0.2313$_{\pm0.0034}$	0.2854$_{\pm0.0049}$	0.1797$_{\pm0.0039}$	0.1535$_{\pm0.0034}$	0.1510$_{\pm0.0037}$	0.2168$_{\pm0.0049}$	0.1248$_{\pm0.0027}$	0.1062$_{\pm0.0021}$
BERT4Rec-dt	0.2223$_{\pm0.0053}$	0.2956$_{\pm0.0030}$	0.1727$_{\pm0.0033}$	0.1427$_{\pm0.0041}$	0.1363$_{\pm0.0043}$	0.2055$_{\pm0.0052}$	0.1116$_{\pm0.0032}$	0.0948$_{\pm0.0025}$
C^{2}DSR	0.1984$_{\pm0.0072}$	0.2574$_{\pm0.0116}$	0.1546$_{\pm0.0050}$	0.1311$_{\pm0.0035}$	0.1263$_{\pm0.0051}$	0.1879$_{\pm0.0061}$	0.1051$_{\pm0.0033}$	0.0903$_{\pm0.0027}$
DREAM	0.2158$_{\pm0.0043}$	0.2771$_{\pm0.0039}$	0.1698$_{\pm0.0025}$	0.1441$_{\pm0.0021}$	0.1377$_{\pm0.0021}$	0.2045$_{\pm0.0033}$	0.1138$_{\pm0.0012}$	0.0956$_{\pm0.0006}$
ABXI	0.2498$_{\pm0.0022}$	0.3175$_{\pm0.0041}$	0.1973$_{\pm0.0025}$	0.1679$_{\pm0.0028}$	0.1737$_{\pm0.0031}$	0.2410$_{\pm0.0026}$	0.1415$_{\pm0.0012}$	0.1206$_{\pm0.0014}$
MERIT	**0.2628**$_{\pm0.0009}^{**}$	**0.3329**$_{\pm0.0033}^{**}$	**0.2105**$_{\pm0.0017}^{**}$	**0.1801**$_{\pm0.0015}^{**}$	**0.1873**$_{\pm0.0033}^{**}$	**0.2603**$_{\pm0.0037}^{**}$	**0.1522**$_{\pm0.0026}^{**}$	**0.1287**$_{\pm0.0022}^{**}$
	Beauty				Electronics			
	HR@5	HR@10	NDCG@10	MRR	HR@5	HR@10	NDCG@10	MRR
SASRec-st	0.1837$_{\pm0.0014}$	0.2597$_{\pm0.0038}$	0.1523$_{\pm0.0020}$	0.1295$_{\pm0.0016}$	0.1345$_{\pm0.0052}$	0.1894$_{\pm0.0038}$	0.1111$_{\pm0.0032}$	0.0982$_{\pm0.0033}$
BERT4Rec-st	0.1687$_{\pm0.0043}$	0.2438$_{\pm0.0043}$	0.1404$_{\pm0.0034}$	0.1197$_{\pm0.0032}$	0.1277$_{\pm0.0067}$	0.1832$_{\pm0.0069}$	0.1053$_{\pm0.0054}$	0.0930$_{\pm0.0050}$
CD-SASRec	0.1605$_{\pm0.0115}$	0.2530$_{\pm0.0166}$	0.1380$_{\pm0.0107}$	0.1162$_{\pm0.0085}$	0.1290$_{\pm0.0060}$	0.1842$_{\pm0.0030}$	0.1069$_{\pm0.0027}$	0.0948$_{\pm0.0032}$
CD-ASR	0.1661$_{\pm0.0065}$	0.2550$_{\pm0.0044}$	0.1424$_{\pm0.0033}$	0.1211$_{\pm0.0033}$	0.1355$_{\pm0.0049}$	0.1938$_{\pm0.0039}$	0.1122$_{\pm0.0033}$	0.0987$_{\pm0.0034}$
MGCL	0.1364$_{\pm0.0047}$	0.2109$_{\pm0.0078}$	0.1162$_{\pm0.0044}$	0.1001$_{\pm0.0035}$	0.1537$_{\pm0.0018}$	0.2159$_{\pm0.0042}$	0.1273$_{\pm0.0019}$	0.1118$_{\pm0.0012}$
SASRec-dt	0.2292$_{\pm0.0102}$	0.3262$_{\pm0.0032}$	0.1866$_{\pm0.0032}$	0.1530$_{\pm0.0032}$	0.1481$_{\pm0.0036}$	0.2169$_{\pm0.0041}$	0.1236$_{\pm0.0040}$	0.1058$_{\pm0.0039}$
BERT4Rec-dt	0.1970$_{\pm0.0027}$	0.3077$_{\pm0.0089}$	0.1679$_{\pm0.0057}$	0.1363$_{\pm0.0050}$	0.1519$_{\pm0.0049}$	0.2246$_{\pm0.0028}$	0.1275$_{\pm0.0031}$	0.1101$_{\pm0.0031}$
C^{2}DSR	0.1835$_{\pm0.0066}$	0.2645$_{\pm0.0034}$	0.1519$_{\pm0.0038}$	0.1290$_{\pm0.0037}$	0.1288$_{\pm0.0072}$	0.1859$_{\pm0.0063}$	0.1081$_{\pm0.0047}$	0.0960$_{\pm0.0043}$
DREAM	0.2090$_{\pm0.0047}$	0.3043$_{\pm0.0069}$	0.1742$_{\pm0.0032}$	0.1447$_{\pm0.0032}$	0.1216$_{\pm0.0040}$	0.1817$_{\pm0.0041}$	0.1023$_{\pm0.0024}$	0.0895$_{\pm0.0019}$
ABXI	0.2807$_{\pm0.0082}$	0.3835$_{\pm0.0050}$	0.2245$_{\pm0.0043}$	0.1846$_{\pm0.0038}$	0.1659$_{\pm0.0021}$	0.2389$_{\pm0.0032}$	0.1385$_{\pm0.0014}$	0.1200$_{\pm0.0019}$
MERIT	**0.3073**$_{\pm0.0074}^{**}$	**0.4054**$_{\pm0.0054}^{**}$	**0.2425**$_{\pm0.0051}^{**}$	**0.2004**$_{\pm0.0052}^{**}$	**0.1837**$_{\pm0.0032}^{**}$	**0.2619**$_{\pm0.0048}^{**}$	**0.1515**$_{\pm0.0029}^{**}$	**0.1297**$_{\pm0.0025}^{**}$
	Movie				Book			
	HR@5	HR@10	NDCG@10	MRR	HR@5	HR@10	NDCG@10	MRR
SASRec-st	0.2258$_{\pm0.0031}$	0.2961$_{\pm0.0037}$	0.1647$_{\pm0.0025}$	0.1874$_{\pm0.0027}$	0.1357$_{\pm0.0029}$	0.1789$_{\pm0.0033}$	0.1007$_{\pm0.0022}$	0.1147$_{\pm0.0023}$
BERT4Rec-st	0.2329$_{\pm0.0018}$	0.3105$_{\pm0.0012}$	0.1927$_{\pm0.0007}$	0.1696$_{\pm0.0007}$	0.1638$_{\pm0.0017}$	0.2152$_{\pm0.0013}$	0.1378$_{\pm0.0008}$	0.1243$_{\pm0.0006}$
CD-SASRec	0.2347$_{\pm0.0022}$	0.3117$_{\pm0.0026}$	0.1940$_{\pm0.0015}$	0.1709$_{\pm0.0017}$	0.1710$_{\pm0.0042}$	0.2253$_{\pm0.0043}$	0.1434$_{\pm0.0030}$	0.1285$_{\pm0.0026}$
CD-ASR	0.2352$_{\pm0.0045}$	0.3052$_{\pm0.0032}$	0.1956$_{\pm0.0027}$	0.1743$_{\pm0.0024}$	0.1622$_{\pm0.0010}$	0.2118$_{\pm0.0024}$	0.1372$_{\pm0.0010}$	0.1244$_{\pm0.0010}$
MGCL	0.2097$_{\pm0.0042}$	0.2851$_{\pm0.0040}$	0.1726$_{\pm0.0032}$	0.1509$_{\pm0.0033}$	0.1248$_{\pm0.0038}$	0.1668$_{\pm0.0049}$	0.1043$_{\pm0.0035}$	0.0946$_{\pm0.0033}$
SASRec-dt	0.2303$_{\pm0.0046}$	0.3067$_{\pm0.0043}$	0.1903$_{\pm0.0032}$	0.1673$_{\pm0.0030}$	0.1356$_{\pm0.0015}$	0.1830$_{\pm0.0014}$	0.1146$_{\pm0.0011}$	0.1034$_{\pm0.0010}$
BERT4Rec-dt	0.2317$_{\pm0.0008}$	0.3095$_{\pm0.0011}$	0.1925$_{\pm0.0015}$	0.1697$_{\pm0.0017}$	0.1547$_{\pm0.0014}$	0.2063$_{\pm0.0012}$	0.1302$_{\pm0.0009}$	0.1176$_{\pm0.0009}$
C^{2}DSR	0.2299$_{\pm0.0019}$	0.3003$_{\pm0.0026}$	0.1911$_{\pm0.0010}$	0.1700$_{\pm0.0005}$	0.1316$_{\pm0.0050}$	0.1767$_{\pm0.0050}$	0.1123$_{\pm0.0032}$	0.1025$_{\pm0.0028}$
DREAM	0.2507$_{\pm0.0068}$	0.3255$_{\pm0.0044}$	0.2082$_{\pm0.0044}$	0.1848$_{\pm0.0043}$	0.1469$_{\pm0.0037}$	0.1973$_{\pm0.0037}$	0.1237$_{\pm0.0033}$	0.1118$_{\pm0.0031}$
ABXI	0.2859$_{\pm0.0016}$	0.3682$_{\pm0.0030}$	0.2388$_{\pm0.0014}$	0.2118$_{\pm0.0011}$	0.1973$_{\pm0.0021}$	0.2571$_{\pm0.0019}$	0.1669$_{\pm0.0013}$	0.1502$_{\pm0.0014}$
MERIT	**0.2912**$_{\pm0.0011}^{**}$	**0.3723**$_{\pm0.0014}^{*}$	**0.2425**$_{\pm0.0014}^{**}$	**0.2154**$_{\pm0.0014}^{**}$	**0.2012**$_{\pm0.0018}^{**}$	**0.2615**$_{\pm0.0017}^{**}$	**0.1694**$_{\pm0.0009}^{**}$	**0.1524**$_{\pm0.0008}^{*}$

residual learning, we introduce multiple ablated variants of specific-domain fusion (SF), mixed-domain fusion (MF), moFFN (FF), and residual learning (RL) as follows:

- SF1: Remove $\mathbf{H}_{M2A}$ from ECAF$_{\text{A}}$ and $\mathbf{H}_{M2B}$ from ECAF$_{\text{B}}$ to evaluate their contribution to the interwoven transfer.
- SF2: Remove $\mathbf{H}_{B2A}$ from ECAF$_{\text{A}}$ and $\mathbf{H}_{A2B}$ from ECAF$_{\text{B}}$ to evaluate their contribution to the interwoven transfer.
- SF3: Replace ECAF$_{\text{A}}$ with token-wise addition $\mathbf{H}_A$+$\mathbf{H}_{M2A}$ and ECAF$_{\text{B}}$ with $\mathbf{H}_B$+$\mathbf{H}_{M2B}$ to evaluate the their effectiveness.
- SF4: Remove ECAF$_{\text{A}}$ andCAF$_B$ to evaluate the rationality of transferring knowledge from domain M to domains A and B.
- FF1: Replace each moFFN with a FFN that shares the same structure as the expert network in the moFFN.

Table 3: Ablation studies in MRR. (RQ2 and RQ3)

RQ	Variants	Food	Kitchen	F+K	Beauty	Electronics	B+E	Movie	Book	M+B
-	MERIT	$\mathbf{0.1801}_{\pm0.0015}$	$\mathbf{0.1287}_{\pm0.0022}$	$\mathbf{0.3088}$	$\mathbf{0.2004}_{\pm0.0052}$	$\mathbf{0.1297}_{\pm0.0025}$	$\mathbf{0.3301}$	$\mathbf{0.2154}_{\pm0.0014}$	$\mathbf{0.1524}_{\pm0.0008}$	$\mathbf{0.3678}$
RQ2	SF1	$0.1751_{\pm0.0037}$	$\underline{0.1282}_{\pm0.0028}$	0.3033	$0.1937_{\pm0.0030}$	$0.1285_{\pm0.0026}$	0.3222	$0.2100_{\pm0.0015}$	$0.1500_{\pm0.0013}$	0.3600
	SF2	$0.1759_{\pm0.0042}$	$0.1265_{\pm0.0020}$	0.3024	$\underline{0.1971}_{\pm0.0028}$	$\underline{0.1290}_{\pm0.0018}$	$\underline{0.3261}$	$0.2130_{\pm0.0032}$	$\underline{0.1511}_{\pm0.0020}$	0.3641
	SF3	$0.1696_{\pm0.0014}$	$0.1201_{\pm0.0026}$	0.2897	$0.1909_{\pm0.0033}$	$0.1213_{\pm0.0035}$	0.3122	$0.2074_{\pm0.0016}$	$0.1471_{\pm0.0010}$	0.3545
	SF4	$0.1754_{\pm0.0030}$	$0.1259_{\pm0.0038}$	0.3013	$0.1945_{\pm0.0048}$	$0.1273_{\pm0.0024}$	0.3218	$\underline{0.2148}_{\pm0.0022}$	$0.1505_{\pm0.0011}$	$\underline{0.3653}$
	FF1	$0.1751_{\pm0.0023}$	$0.1273_{\pm0.0026}$	0.3024	$0.1945_{\pm0.0024}$	$\underline{0.1290}_{\pm0.0036}$	0.3235	$0.2101_{\pm0.0019}$	$0.1504_{\pm0.0015}$	0.3605
	MF1	$0.1696_{\pm0.0043}$	$0.1219_{\pm0.0014}$	0.2915	$0.1878_{\pm0.0015}$	$0.1183_{\pm0.0033}$	0.3097	$0.2014_{\pm0.0026}$	$0.1398_{\pm0.0011}$	0.3412
	MF2	$0.1719_{\pm0.0015}$	$0.1257_{\pm0.0022}$	0.2976	$0.1947_{\pm0.0033}$	$0.1252_{\pm0.0029}$	0.3199	$0.2135_{\pm0.0027}$	$0.1510_{\pm0.0030}$	0.3645
RQ3	RL1	$\underline{0.1796}_{\pm0.0016}$	$0.1269_{\pm0.0018}$	$\underline{0.3065}$	$0.1952_{\pm0.0051}$	$0.1247_{\pm0.0026}$	0.3199	$0.2090_{\pm0.0021}$	$0.1481_{\pm0.0009}$	0.3571
	RL2	$0.1700_{\pm0.0007}$	$0.1207_{\pm0.0017}$	0.2907	$0.1764_{\pm0.0022}$	$0.1274_{\pm0.0023}$	0.3038	$0.2008_{\pm0.0017}$	$0.1385_{\pm0.0017}$	0.3393

- MF1: Replace CAF_M with token-wise addition, $\mathbf{H}_M+\mathbf{H}_{A2M}+\mathbf{H}_{B2M}$, to evaluate its effectiveness.
- MF2: Remove CAF_M to evaluate the rationality of transferring knowledge from domains A and B to M.
- RL1: Remove the stop-gradient operation on $\mathbf{H}_M^F$ in A- and B-domain task to investigate the interference caused by supervision difference.
- RL2: Remove $\mathbf{H}_M^F$ in A- and B-domain tasks to investigate the effectiveness of prediction-level residual learning.

Table 3 clearly demonstrates that every component of MERIT is essential for its overall performance. Results from SF1 and SF2 show that both mixed-to-specific and specific-to-specific transfers in the specific-domain CAF modules contribute positively. In contrast, SF3, which replaces sequence-level fusion with token-level addition, yields the second-worst performance, indicating that the distortion of coinciding sequential representations significantly hinders the model. Similarly, MF1 and MF2 reveal that sequence-level fusion is indispensable, as replacing it with token-level addition or removing CAF_M leads to notable performance declines. FF1 also demonstrates the improvement of moFFN over vanilla FFN.

Finally, the results from RL1 and RL2 confirm that prediction-level residual learning effectively decouples mixed-domain and specific-domain tasks, thereby boosting performance.

5.6 Expert Utilization in MoE-FFN (RQ4)

To quantitatively assess expert aggregation, we examine expert utilization across MoE-FFN outputs. Figure 7 presents the weight distributions for each gate on the test set, with bar heights representing mean weights and vertical dashes indicating their standard deviations.

The results indicate that intra-domain representations, $\mathbf{H}_M$, $\mathbf{H}_A$, and $\mathbf{H}_B$, exhibit fluctuations in expert utilization on the smaller FK and BE datasets. Nonetheless, overall performance remains stable as shown in Table 2, suggesting that transferred inter-domain representations help mitigate this instability. Similarly, expert utilization for $\mathbf{H}_{B2M}$ also fluctuates on the FK and BE datasets, due to the challenges in extracting transferable knowledge from lower-performing domains compared to higher-performing domains on smaller datasets. In contrast, the remaining outputs display stable expert utilization without a clear bias between shared and domain-specific experts. Furthermore, no expert receives a zero weight across all gates, indicating that the shared components between specific and shared experts are effectively captured. Notably, specific-to-specific transfers across all datasets exhibit a balanced utilization of shared and specific experts. This balance indicates a substantial overlap between mixed-to-specific and specific-to-specific transfers directed toward the same target domains. As a result, the optimization tends to assign lower gradients to specific-to-specific transfers, yielding nearly equal contributions from both expert types.

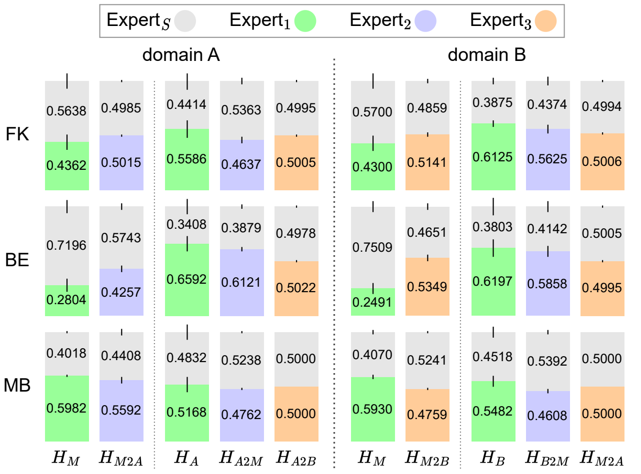


Figure 7: Expert utilization of the corresponding input tokens for testing data, separated by target domain. The height of bars and the vertical dashes indicate the mean and standard deviation of weights, respectively. (RQ4)

In summary, the expert utilization analysis confirms the effectiveness of interwoven transfer in MERIT and suggests that its stability may be influenced by dataset size.

6 Conclusion

In this paper, we introduce MERIT, a novel dual-target CDSR model that performs residual interwoven transfer between mixed- and specific-domain sequences. Experimental results on three real-world datasets demonstrate that MERIT outperforms all baselines, including state-of-the-art CDSR models, with statistical significance. We plan to explore the interplay among mixed-domain, domain-specific, and explicit domain-invariant representations in future work.

7 Acknowledgments

This research/project is supported by the Ministry of Education, Singapore under its Academic Research Fund Tier 1 (RG16/24). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Ministry of Education, Singapore. This research is also partially supported by the National Natural Science Foundation of China (Grant No. 72371148 and 72192832), and the Shanghai Rising-Star Program (Grant No. 23QA1403100).

References

[1] Nawaf Alharbi and Doina Caragea. 2021. Cross-domain attentive sequential recommendations based on general and current user preferences (cd-asr). In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. 48–55.

[2] Nawaf Alharbi and Doina Caragea. 2022. Cross-domain self-attentive sequential recommendations. In *Proceedings of International Conference on Data Science and Applications: ICDSA 2021, Volume 2*. Springer, 601–614.

[3] Qingtian Bian, Marcus Vinicius De Carvalho, Tieying Li, Jiaxing Xu, Hui Fang, and Yiping Ke. 2025. ABXI: Invariant Interest Adaptation for Task-Guided Cross-Domain Sequential Recommendation. In *THE WEB CONFERENCE 2025*.

[4] Qingtian Bian, Jiaxing Xu, Hui Fang, and Yiping Ke. 2023. CPMR: Context-Aware Incremental Sequential Recommendation with Pseudo-Multi-Task Learning. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 120–130.

[5] Shuqing Bian, Xingyu Pan, Wayne Xin Zhao, Jinpeng Wang, Chuyuan Wang, and Ji-Rong Wen. 2023. Multi-modal Mixture of Experts Represetation Learning for Sequential Recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 110–119.

[6] Chenwei Cai, Ruining He, and Julian McAuley. 2017. SPMC: Socially-aware personalized Markov chains for sparse sequential recommendation. *arXiv preprint arXiv:1708.04497* (2017).

[7] Jiangxia Cao, Xin Cong, Jiawei Sheng, Tingwen Liu, and Bin Wang. 2022. Contrastive Cross-Domain Sequential Recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 138–147.

[8] Jiangxia Cao, Xixun Lin, Xin Cong, Jing Ya, Tingwen Liu, and Bin Wang. 2022. Disencdr: Learning disentangled representations for cross-domain recommendation. In *Proceedings of the 45th International ACM SIGIR conference on research and development in information retrieval*. 267–277.

[9] Jianxin Chang, Chen Gao, Yu Zheng, Yiqun Hui, Yanan Niu, Yang Song, Depeng Jin, and Yong Li. 2021. Sequential recommendation with graph neural networks. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 378–387.

[10] Shu Chen, Zitao Xu, Weike Pan, Qiang Yang, and Zhong Ming. 2024. A Survey on Cross-Domain Sequential Recommendation. *arXiv preprint arXiv:2401.04971* (2024).

[11] Tianwen Chen and Raymond Chi-Wing Wong. 2021. An efficient and effective framework for session-based social recommendation. In *Proceedings of the 14th ACM international conference on web search and data mining*. 400–408.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[13] Tim Donkers, Benedikt Loepp, and Jürgen Ziegler. 2017. Sequential user-based recurrent neural network recommendations. In *Proceedings of the 11th ACM Conference on Recommender Systems*. 152–160.

[14] Ziwei Fan, Zhiwei Liu, Jiawei Zhang, Yun Xiong, Lei Zheng, and Philip S Yu. 2021. Continuous-time sequential recommendation with temporal graph collaborative transformer. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 433–442.

[15] Rong Gao, Jing Li, Xuefei Li, Chenfang Song, Jun Chang, Donghua Liu, and Chunzhi Wang. 2018. STSCR: Exploring spatial-temporal sequential influence and social information for location recommendation. *Neurocomputing* 319 (2018), 118–133.

[16] Lei Guo, Li Tang, Tong Chen, Lei Zhu, Quoc Viet Hung Nguyen, and Hongzhi Yin. 2021. DA-GCN: A domain-aware attentive graph convolution network for shared-account cross-domain sequential recommendation. *arXiv preprint arXiv:2105.03300* (2021).

[17] Lei Guo, Jinyu Zhang, Li Tang, Tong Chen, Lei Zhu, and Hongzhi Yin. 2022. Time interval-enhanced graph neural network for shared-account cross-domain sequential recommendation. *IEEE Transactions on Neural Networks and Learning Systems* 35, 3 (2022), 4002–4016.

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

[19] Guangneng Hu, Yu Zhang, and Qiang Yang. 2018. Conet: Collaborative cross networks for cross-domain recommendation. In *Proceedings of the 27th ACM international conference on information and knowledge management*. 667–676.

[20] Jin Huang, Wayne Xin Zhao, Hongjian Dou, Ji-Rong Wen, and Edward Y Chang. 2018. Improving sequential recommendation with knowledge-enhanced memory networks. In *The 41st international ACM SIGIR conference on research & development in information retrieval*. 505–514.

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

[22] Chenglin Li, Mingjun Zhao, Huanming Zhang, Chenyun Yu, Lei Cheng, Guoqiang Shu, Beibei Kong, and Di Niu. 2022. RecGURU: Adversarial learning of generalized user representations for cross-domain recommendation. In *Proceedings of the fifteenth ACM international conference on web search and data mining*. 571–581.

[23] 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.

[24] Pan Li and Alexander Tuzhilin. 2020. Ddtcdr: Deep dual transfer cross domain recommendation. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 331–339.

[25] Guanyu Lin, Chen Gao, Yu Zheng, Jianxin Chang, Yanan Niu, Yang Song, Kun Gai, Zhiheng Li, Depeng Jin, Yong Li, et al. 2024. Mixed attention network for cross-domain sequential recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. 405–413.

[26] Haokai Ma, Ruobing Xie, Lei Meng, Xin Chen, Xu Zhang, Leyu Lin, and Jie Zhou. 2024. Triple sequence learning for cross-domain recommendation. *ACM Transactions on Information Systems* 42, 4 (2024), 1–29.

[27] Muyang Ma, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Lifan Zhao, Peiyu Liu, Jun Ma, and Maarten de Rijke. 2022. Mixed information flow for cross-domain sequential recommendations. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 16, 4 (2022), 1–32.

[28] Muyang Ma, Pengjie Ren, Yujie Lin, Zhumin Chen, Jun Ma, and Maarten de Rijke. 2019. π-net: A parallel information-sharing network for shared-account cross-domain sequential recommendations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 685–694.

[29] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 43–52.

[30] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[31] Xingyu Pan, Yushuo Chen, Changxin Tian, Zihan Lin, Jinpeng Wang, He Hu, and Wayne Xin Zhao. 2022. Multimodal meta-learning for cold-start sequential recommendation. In *Proceedings of the 31st ACM international conference on information & knowledge management*. 3421–3430.

[32] Yunzhu Pan, Chen Gao, Jianxin Chang, Yanan Niu, Yang Song, Kun Gai, Depeng Jin, and Yong Li. 2023. Understanding and modeling passive-negative feedback for short-video sequential recommendation. In *Proceedings of the 17th ACM conference on recommender systems*. 540–550.

[33] Chung Park, Taesan Kim, Taekyoon Choi, Junui Hong, Yelim Yu, Mincheol Cho, Kyunam Lee, Sungil Ryu, Hyungjun Yoon, Minsung Choi, et al. 2023. Cracking the Code of Negative Transfer: A Cooperative Game Theoretic Approach for Cross-Domain Sequential Recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 2024–2033.

[34] Chung Park, Taesan Kim, Hyungjun Yoon, Junui Hong, Yelim Yu, Mincheol Cho, Minsung Choi, and Jaegul Choo. 2024. Pacer and Runner: Cooperative Learning Framework between Single-and Cross-Domain Sequential Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2071–2080.

[35] Ahmed Rashed, Shereen Elsayed, and Lars Schmidt-Thieme. 2022. Context and attribute-aware sequential recommendation via cross-attention. In *Proceedings of the 16th ACM Conference on Recommender Systems*. 71–80.

[36] Steffen Rendle. 2010. Factorization machines. In *2010 IEEE International Conference on Data Mining*. IEEE, 995–1000.

[37] 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.

[38] Wenchao Sun, Muyang Ma, Pengjie Ren, Yujie Lin, Zhumin Chen, Zhaochun Ren, Jun Ma, and Maarten De Rijke. 2021. Parallel split-join networks for shared account cross-domain sequential recommendations. *IEEE Transactions on Knowledge and Data Engineering* (2021).

[39] Hongyan Tang, Junning Liu, Ming Zhao, and Xudong Gong. 2020. Progressive layered extraction (ple): A novel multi-task learning (mtl) model for personalized recommendations. In *Proceedings of the 14th ACM Conference on Recommender Systems*. 269–278.

[40] Jiaxi Tang and Ke Wang. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*. 565–573.

[41] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[43] Ziyang Wang, Wei Wei, Gao Cong, Xiao-Li Li, Xian-Ling Mao, and Minghui Qiu. 2020. Global context enhanced graph neural networks for session-based recommendation. In *Proceedings of the 43rd international ACM SIGIR Conference on Research and Development in Information Retrieval*. 169–178.

[44] Chao-Yuan Wu, Amr Ahmed, Alex Beutel, Alexander J Smola, and How Jing. 2017. Recurrent recommender networks. In *Proceedings of the 10th ACM International Conference on Web Search and Data Mining*. 495–503.

[45] Jibang Wu, Renqin Cai, and Hongning Wang. 2020. Déjà vu: A contextualized temporal attention mechanism for sequential recommendation. In *Proceedings of The Web Conference 2020*. 2199–2209.

[46] Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. 2019. Session-based recommendation with graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 346–353.

[47] Zitao Xu, Xiaoqing Chen, Weike Pan, and Zhong Ming. [n. d.]. Heterogeneous Graph Transfer Learning for Category-aware Cross-Domain Sequential Recommendation. In *THE WEB CONFERENCE 2025*.

[48] Zitao Xu, Weike Pan, and Zhong Ming. 2023. A Multi-view Graph Contrastive Learning Framework for Cross-Domain Sequential Recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 491–501.

[49] Zitao Xu, Weike Pan, and Zhong Ming. 2024. Transfer learning in cross-domain sequential recommendation. *Information Sciences* 669 (2024), 120550.

[50] An Yan, Shuo Cheng, Wang-Cheng Kang, Mengting Wan, and Julian McAuley. 2019. CosRec: 2D convolutional neural networks for sequential recommendation. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 2173–2176.

[51] Xiaoxin Ye, Yun Li, and Lina Yao. 2023. DREAM: Decoupled Representation via Extraction Attention Module and Supervised Contrastive Learning for Cross-Domain Sequential Recommender. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 479–490.

[52] Xu Yuan, Dongsheng Duan, Lingling Tong, Lei Shi, and Cheng Zhang. 2021. Icaisr: Item categorical attribute integrated sequential recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 1687–1691.

[53] Tianzi Zang, Yanmin Zhu, Ruohan Zhang, Chunyang Wang, Ke Wang, and Jiadi Yu. 2023. Contrastive Multi-View Interest Learning for Cross-Domain Sequential Recommendation. *ACM Transactions on Information Systems* 42, 3 (2023), 1–30.

[54] Dongyi Zheng, Hongyu Zhang, Jianyang Zhai, Lin Zhong, Lingzhi Wang, Jiyuan Feng, Xiangke Liao, Yonghong Tian, Nong Xiao, and Qing Liao. 2025. FedCSR: A Federated Framework for Multi-Platform Cross-Domain Sequential Recommendation with Dual Contrastive Learning. In *Proceedings of the 31st International Conference on Computational Linguistics*. 8699–8713.

[55] Lei Zheng, Vahid Noroozi, and Philip S Yu. 2017. Joint deep modeling of users and items using reviews for recommendation. In *Proceedings of the tenth ACM international conference on web search and data mining*. 425–434.

[56] Xiaolin Zheng, Jiajie Su, Weiming Liu, and Chaochao Chen. 2022. DDGHM: dual dynamic graph with hybrid metric training for cross-domain sequential recommendation. In *Proceedings of the 30th ACM International Conference on Multimedia*. 471–481.

[57] Feng Zhu, Chaochao Chen, Yan Wang, Guanfeng Liu, and Xiaolin Zheng. 2019. Dtcdr: A framework for dual-target cross-domain recommendation. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 1533–1542.

[58] Feng Zhu, Yan Wang, Chaochao Chen, Guanfeng Liu, and Xiaolin Zheng. 2020. A graphical and attentional framework for dual-target cross-domain recommendation.. In *IJCAI*, Vol. 21. 39.