# Collaborative Filtering with Attribution Alignment for Review-based Non-overlapped Cross Domain Recommendation

Weiming Liu, Xiaolin Zheng, Mengling Hu, Chaochao Chen*
College of Computer Science and Technology, Zhejiang University, China
{21831010, xlzheng, humengling, zjuccc}@zju.edu.cn



## ABSTRACT

Cross-Domain Recommendation (CDR) has been popularly studied to utilize different domain knowledge to solve the data sparsity and cold-start problem in recommender systems. In this paper, we focus on the *Review-based Non-overlapped Recommendation* (*RNCDR*) problem. The problem is commonly-existed and challenging due to two main aspects, i.e, there are only positive user-item ratings on the target domain and there is no overlapped user across different domains. Most previous CDR approaches cannot solve the RNCDR problem well, since (1) they cannot effectively combine review with other information (e.g., ID or ratings) to obtain expressive user or item embedding, (2) they cannot reduce the domain discrepancy on users and items. To fill this gap, we propose Collaborative Filtering with Attribution Alignment model (**CFAA**), a cross-domain recommendation framework for the RNCDR problem. **CFAA** includes two main modules, i.e., *rating prediction module* and *embedding attribution alignment module*. The former aims to jointly mine review, one-hot ID, and multi-hot historical ratings to generate expressive user and item embeddings. The later includes vertical attribution alignment and horizontal attribution alignment, tending to reduce the discrepancy based on multiple perspectives. Our empirical study on Douban and Amazon datasets demonstrates that **CFAA** significantly outperforms the state-of-the-art models under the RNCDR setting.

## CCS CONCEPTS

• **Information systems** → **Recommender systems**; • **Collaborative filtering**;

## KEYWORDS

Recommendation, Domain Adaptation, Transfer Learning

**ACM Reference Format:**
Weiming Liu, Xiaolin Zheng, Mengling Hu, Chaochao Chen. 2022. Collaborative Filtering with Attribution Alignment for Review-based Non-overlapped Cross Domain Recommendation. In *Proceedings of the ACM Web Conference 2022 (WWW '22), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3485447.3512166

*Chaochao Chen is the corresponding author.


*WWW '22, April 25–29, 2022, Virtual Event, Lyon, France*

ACM ISBN 978-1-4503-9096-5/22/04…$15.00
https://doi.org/10.1145/3485447.3512166

## 1 INTRODUCTION

With the advent of digital era, more and more users participant in multiple domains (platforms) for different purposes, e.g., buying books on *Amazon* and reading news on *Flipboard* [12]. How to comprehensively utilize the cross domain information to improve the performance of recommendation systems has become a hot topic. Therefore, Cross Domain Recommendation (CDR) becomes more and more attractive for establishing highly accurate recommendation systems [5, 25, 43, 54]. Most existing CDR models assume that both the source and target domains share the same set of users, which makes it easier to transfer useful knowledge across domains based on these overlapped users. Moreover, most existing CDR models assume the existence of sufficient negative feedback, which limits their applications.

In this paper, we focus on a general problem in CDR where the target domain only has a small proportion of positive ratings but no negative ones. Meanwhile users in the source and target domains are totally non-overlapped. Besides user-item rating information, we also assume the existence of user-item reviews which are commonly used as side-information for alleviating the data sparsity problem [44, 48]. Specifically, we term this problem as *Review-based Non-overlapped Cross Domain Recommendation* (*RNCDR*), where we aim to transfer knowledge from a relative dense source domain to a sparse target domain to enhance the prediction performance, as shown in Fig. 1. The RNCDR problem is widely existed since the target domain is always suffering from the data sparsity problem. Therefore, the target domain only has a small proportion of observed positive samples without negative ones. We summarize the main challenges as follows. (1) Users/items always have diverse characteristics and preferences which should be well-exploited across domains. (2) Different domains tend to have varied behavioral patterns [15, 49], causing the existence of latent embedding attribution bias and discrepancy [22].

Although there have been previous studies on the RNCDR problem [48], they cannot solve it well. The current state-of-the-art model on RNCDR is TDAR [48] which adopts deep Domain Adversarial Neural Network (DANN) [16] to align the source and target user/item features. On the one hand, it fails to integrate the review information with one-hot ID and multi-hot historical rating [9, 34], leading to less expressive user and item embeddings [17, 24]. On the other hand, a deep adversarial network with domain discriminator is unstable and hard to train in practice [38]. Therefore, it cannot solve both challenges and leads to poor model performance. These two negative aspects make this approach difficult to achieve the RNCDR task.

To address the aforementioned issues, in this paper, we propose **CFAA**, a cross-domain recommendation framework for the RNCDR problem. In order to better model user/item embeddings and align

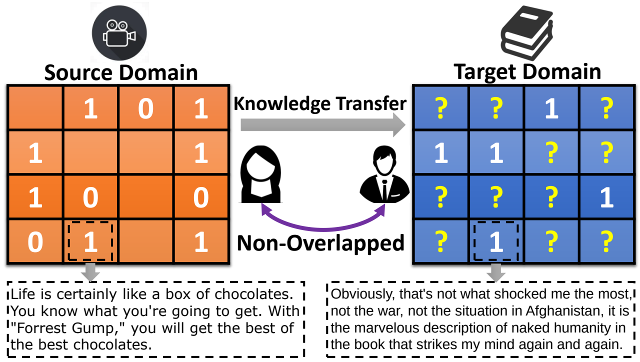



**Figure 1: The problem of Review-based Non-overlapped Cross Domain Recommendation (RNCDR).**

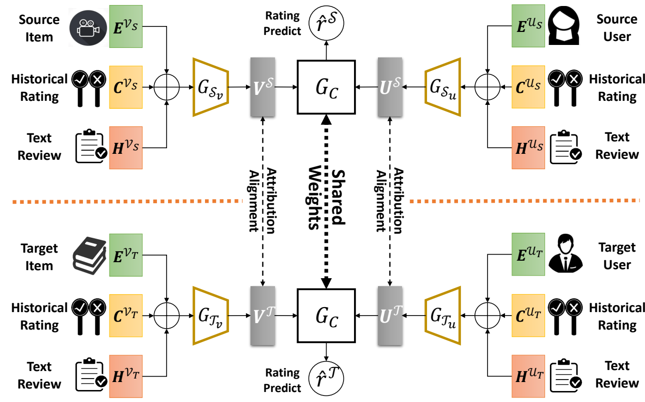



**Figure 2: The basic framework of CFAA.**

the latent embedding attributions across domains for high-quality rating predictions, we utilize two modules in **CFAA**, i.e., *rating prediction module* and *embedding attribution alignment module*. The rating prediction module aims to tackle the first challenge, capturing useful information and generating expressive user and item embeddings. For this purpose, we fuse the one-hot ID, multi-hot historical ratings, and text-based review information to better capture user and item collaborative preferences. The embedding attribution alignment module tends to tackle the second challenge, properly reducing the attribution discrepancy and transferring consistent useful knowledge across the source and target domains. Specifically, we propose dual perspectives on reducing the discrepancy, i.e., *vertical attribution alignment* with typical-sample optimal transport and *horizontal attribution alignment* with the adaptation on attribution graph. Vertical attribution alignment confines the probability distribution among the source and target domains for the corresponding attribution and horizontal attribution alignment tends to exploit the consistent relevant relationship between different attributes across domains. These two methods cooperate with each other to complete the alignment task and reduce the domain discrepancy.

We summarize our main contributions as follows: (1) We propose a novel framework, i.e., **CFAA**, for the RNCDR problem, which contains rating prediction module and embedding attribution alignment module. (2) The proposed rating prediction module can efficiently combine the text-based review information with one-hot ID and multi-hot historical rating embedding, and the embedding attribution alignment module equipped with vertical and horizontal alignment can better reduce the domain discrepancy. (3) Extensive empirical studies on Douban and Amazon datasets demonstrate that **CFAA** significantly improves the state-of-the-art models under the RNCDR setting.

## 2 RELATED WORK

**Traditional Cross Domain Recommendation.** Traditional Cross Domain Recommendation (CDR) emerges as a technique to alleviate the long-standing data sparsity problem in recommendation by assuming the same or partial proportion of overlapped user/item set across domains [20, 47]. Man et al. [28] first proposed to learn the mapping function on the overlapped users across domains. Zhao et al. [51] extended the research on the attribution aspects transfer. Some recent method [21] even exploited bidirectional latent relations between users and items to obtain more reliable knowledge. Besides, Hu et al. [17] further proposed multi-task learning strategy which stitches both source and target information together. It is noticeable that the majority of traditional CDR frameworks always assume the existence of overlapped users as the bridge for knowledge sharing across different domains [54]. However, the source and target domains may not share overlapped users in the real-world applications which limits their application scenario.

**Review-based Non-overlapped Cross Domain Recommendation.** There are some CDR models that can solve the situations where users/items are non-overlapped. Yang et al. [20] was the first to propose the widely used codebook transfer approach for non-overlapped CDR. However, it cannot be extended to implicit feedback, which strongly limits its generalization capability. Furthermore, without overlapped users or items act as bridges for knowledge transfer, the recommendation accuracy is still not satisfying [11]. Recently, some researchers [52] have propose to utilize review text of users and items to enhance the model performance. This inspires us to take advantage of these review information as the transferring bridge across different domains. For example, Wang et al. [44] used text-enhanced information for solving both overlapped and non-overlapped CDR problems with deep adversarial learning strategy. Yu et al. [48] considered a more general situation where the target domain only has positive ratings, and proposed dual adversarial alignment on both user and item embeddings. However, these models cannot combine the one-hot ID with multi-hot historical rating into the review information, leading to limited model performance. Moreover, some studies [22] have shown the existence of embedding bias and discrepancy on CDR problem since two domains may always have different expressions. This enlightens us to transfer useful review knowledge by reducing the embedding bias for better results.

**Domain Adaptation.** Domain adaptation technique was proposed to transfer knowledge from a well-labeled source domain to a target domain without or with less labels [42]. The most classic method is Maximum Mean Discrepancy (MMD) [4, 35] which is based on the mean statistic of samples. Moreover, some scholars proposed Correlation Alignment (CORAL) [40, 41] to align the second-order covariance statistics. More recently, ESAM [8] extended CORAL with attribution correlation congruence for solving the long-tailed item recommendation problem. However, performing statistic on mean or covariance may hard to capture the complex and high dimension data characteristics. For this, Ganin et al. [16] proposed Domain Adversarial Neural Network (DANN) which integrated a domain discriminator with adversarial training to align the embeddings across domains. Notably that previous review-based CDR models, e.g., TDAR [48] and Rec-GAN [44], also adopted adversarial training for aligning the source and target domains. However, latest researches [38] pointed out that the origin adversarial training may be unstable under some circumstances which will hurdle the performance. In this paper, we propose to enhance statistic-based embedding attribution alignment method by attribution distribution and attribution relevance alignment to exploit more expressive information on complex feature space.

# 3 MODELING FOR CFAA

First, we describe notations. We assume there are two domains, i.e., a source domain $\mathcal{S}$ and a target domain $\mathcal{T}$. There are $N_{U_S}$ and $N_{U_T}$ users in source and target domains respectively. There are $N_{V_S}$ and $N_{V_T}$ items in source and target domains respectively. Let $\boldsymbol{R}^{\mathcal{S}} \in \mathbb{R}^{N_{U_S} \times N_{V_S}}$ and $\boldsymbol{R}^{\mathcal{T}} \in \mathbb{R}^{N_{U_T} \times N_{V_T}}$ be the observed source and item rating matrices in $\mathcal{S}$ and $\mathcal{T}$ respectively. Moreover, let $\Omega^{\mathcal{S}_u}$ and $\Omega^{\mathcal{T}_u}$ be the set of user reviews. Similarly, let $\Omega^{\mathcal{S}_v}$ and $\Omega^{\mathcal{T}_v}$ be the set of item reviews. Unlike the traditional cross domain recommendation which has the assumption that both source and target domains share the same set of users, in RNCDR, the source and target users are totally non-overlapped.

Then, we introduce the overview of our proposed **CFAA** framework, as is illustrated in Fig. 2. **CFAA** model mainly has two modules, i.e., *rating prediction module* and *embedding attribution alignment module*. The rating prediction module aims to combine the review information with one-hot ID and multi-hot historical rating to generate expressive user and item embeddings. The embedding attribution alignment module is supposed to reduce the embedding attribution discrepancy across domains. We will introduce these two modules in details later.

## 3.1 Rating Prediction Module

Firstly, we provide the details of the rating prediction module. For convenience, we use the notations and calculation process in the source domain as an example. For the $i$-th user and the $j$-th item, we define their corresponding one-hot ID vectors as $\boldsymbol{X}_i^{\mathcal{U}_S}$ and $\boldsymbol{X}_j^{\mathcal{V}_S}$, respectively. Let $\boldsymbol{R}_{i*}^{\mathcal{S}}$ and $\boldsymbol{R}_{*j}^{\mathcal{S}}$ denote the historical rating for the $i$-th user and the $j$-th item. For the $i$-th user, we use $\Omega_i^{\mathcal{S}_u}$ to denote the corresponding review which includes $\Omega_i^{\mathcal{S}_u} = (\omega_{i1}^{\mathcal{S}_u}, \omega_{i2}^{\mathcal{S}_u}, \cdots, \omega_{ik}^{\mathcal{S}_u})$ sentences. Likewise, we use $\Omega_j^{\mathcal{S}_v}$ to denote the corresponding review for the $j$-th item which includes $\Omega_j^{\mathcal{S}_v} = (\omega_{j1}^{\mathcal{S}_v}, \omega_{j2}^{\mathcal{S}_v}, \cdots, \omega_{jl}^{\mathcal{S}_v})$ sentences. Notably that we adopt the sentence segmentation component (Sentencizer[1]) to split the origin document into several individual sentences in $\Omega_i^{\mathcal{S}_u}$ and $\Omega_j^{\mathcal{S}_v}$.

We adopt a trainable lookup table to exploit the user and item one-hot ID embedding as $\mathrm{LookUp}(\boldsymbol{X}_i^{\mathcal{U}}) = \boldsymbol{E}_i^{\mathcal{U}_S}$ and $\mathrm{LookUp}(\boldsymbol{X}_j^{\mathcal{V}}) = \boldsymbol{E}_j^{\mathcal{V}_S}$. We utilize the fully connected layers $F_{S_u}$ and $F_{S_v}$ to obtain user and item behavior embeddings as $F_{S_u}(\mathcal{R}_{i*}^{\mathcal{S}}) = \boldsymbol{C}_i^{\mathcal{U}_S}$ and $F_{S_v}(\mathcal{R}_{*j}^{\mathcal{S}}) = \boldsymbol{C}_j^{\mathcal{V}_S}$, respectively. Meanwhile we utilize the pre-trained BERT model's penultimate encoder layer to obtain the contextualized word embeddings for each sentence [30]. Then we average every sentence's word embeddings to generate the sentence embedding. We then average the sentence embeddings of the $i$-th user or $j$-th item as their corresponding review-based embedding $\boldsymbol{H}_i^{\mathcal{U}_S}$ and $\boldsymbol{H}_j^{\mathcal{V}_S}$. Finally, we utilize fully connected layers $G_{S_u}$ and $G_{S_v}$ to obtain the user and item general embedding as $G_{S_u}(\boldsymbol{E}^{\mathcal{U}_S} \oplus \boldsymbol{C}^{\mathcal{U}_S} \oplus \boldsymbol{H}^{\mathcal{U}_S}) = \boldsymbol{U}^{\mathcal{S}} \in \mathbb{R}^{N \times D}$ and $G_{S_v}(\boldsymbol{E}^{\mathcal{V}_S} \oplus \boldsymbol{C}^{\mathcal{V}_S} \oplus \boldsymbol{H}^{\mathcal{V}_S}) = \boldsymbol{V}^{\mathcal{S}} \in \mathbb{R}^{N \times D}$ where $N$ denotes batchsize, $D$ denotes the dimension of the latent embedding, and $\oplus$ denotes the concatenation operation. Likewise, we can also obtain the user and item general embeddings on the target domain as $\boldsymbol{U}^{\mathcal{T}} \in \mathbb{R}^{N \times D}$ and $\boldsymbol{V}^{\mathcal{T}} \in \mathbb{R}^{N \times D}$ respectively. After that, we adopt the fully connected layer $G_C$ to predict user-item ratings as $G_C(\boldsymbol{U}^{\mathcal{S}}, \boldsymbol{V}^{\mathcal{S}}) = \hat{\boldsymbol{\mathcal{R}}}^{\mathcal{S}}$ and $G_C(\boldsymbol{U}^{\mathcal{T}}, \boldsymbol{V}^{\mathcal{T}}) = \hat{\boldsymbol{\mathcal{R}}}^{\mathcal{T}}$, respectively. We further use cross entropy loss $L_C$ to minimize the prediction ratings and ground-truth ones as below:

$$L_C = -\sum_{i=1}^{N} (\mathcal{R}_i^{\mathcal{S}} \log \hat{\mathcal{R}}_i^{\mathcal{S}} + (1 - \mathcal{R}_i^{\mathcal{S}}) \log(1 - \hat{\mathcal{R}}_i^{\mathcal{S}}) + \mathcal{R}_i^{\mathcal{T}} \log \hat{\mathcal{R}}_i^{\mathcal{T}}).$$

It is noticeable that we only have positive ratings in the target domain. Through this basic loss function, the network parameters can be adjusted to the training data efficiently.

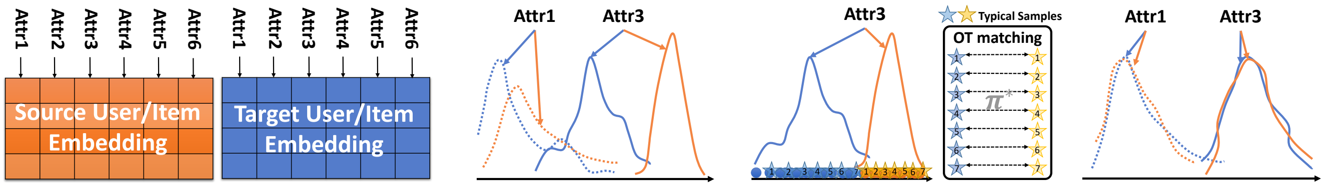
## 3.2 Embedding Attribution Alignment Module

Although the rating prediction module can provide us a simple and good baseline, it still cannot depict user and item characteristics in the source and target domains well. The main reason lies in the data sparsity across domains which causes the embedding attribution bias and discrepancy. Specifically, we denote each dimension of an embedding as a certain kind of *attribution*. The attribution discrepancy includes two aspects, i.e., *vertical probability discrepancy* and *horizontal relevance discrepancy* on each attribution, as shown in Fig. 3 and Fig. 4. For example, Attr3 in both source and target domains has different *probability distribution* as shown in Fig. 3(b). Meanwhile, Attr2 and Attr3 have different *relevance strength* across domains, indicating the existence of the relevance discrepancy, as shown in Fig. 4. These biases deteriorate the knowledge transfer across domains and may even lead to the negative transfer phenomenon [46]. Therefore, it is essential to reduce the embedding attribution discrepancy for better knowledge transfer. The embedding attribution discrepancy module consists of two main algorithms, i.e., *Vertical Attribution Alignment* and *Horizontal Attribution Alignment*. These two algorithms proceed from different perspectives, and they complement each other to solve the embedding attribution discrepancy problem.

### 3.2.1 *Vertical Attribution Alignment.* 

We first introduce the attribution distribution alignment. It is reasonable to assume that each dimension of user (item) embeddings has certain meaningful information like occupation, income, hobby (style, theme, brand), and etc. Although we cannot directly decipher what is the exact meaning of each dimension, deep analysis of these latent embedding

[1] https://spacy.io/api/sentencizer




(a) Initial User/Item Embedding (b) Before Vertical Attribution Alignment (c) Appling Vertical Attribution Alignment (d) After Vertical Attribution Alignment

**Figure 3: The main procedure of vertical attribution alignment. The orange and blue color denote source and target domains respectively.**

attributions can still be helpful for recommendation [8]. Each latent attribution distribution should be consistent with the source and target domains in order to enhance the model performance. However, due to the data sparsity among the CDR problem, there always exists the *vertical distribution discrepancy* between the source and target domains, as shown in Fig. 3. For example, Attr1 and Attr3 in Fig. 3(b) have different probability distribution across domains with a certain domain gap. Vertical attribution discrepancy seriously hurdles the knowledge transfer across domains, which not only raises the training difficulty but also leads to the scattered target space with domain discrepancy. Vertical attribution alignment mainly has two steps, i.e., *typical sample selection step* and *optimal transport matching step*. The former can figure out the proxy data samples and filter the irrelevant noise data, while the later can align these typical samples on the each attribution across different domains.

**Typical Sample Selection.** We first introduce the *typical sample selection algorithm*, which aims to find $K$ typical samples $\boldsymbol{M}_Z^{\mathcal{X}} \in \mathbb{R}^{K\times D}$ in both source and target domains where the $q$-th column $(\boldsymbol{M}_Z^{\mathcal{X}})_{*q} \in \mathbb{R}^{K\times 1}$ represents the typical samples on the $q$-th attribution. Let $\mathcal{X} = \{\mathcal{S}, \mathcal{T}\}$ denote the domain index, and $\boldsymbol{Z} = \{\boldsymbol{U}, \boldsymbol{V}\}$ denote the user or item set. Inspired by [2, 32], we formulate the typical sample selection optimization problem as:

$$\min \ell_q = \sum_{i=1}^{N}\sum_{j=1}^{K}(\Psi_q^{Z^{\mathcal{X}}})_{ij}\left(\boldsymbol{Z}_{iq}^{\mathcal{X}} - (\boldsymbol{M}_Z^{\mathcal{X}})_{jq}\right)^2 + \alpha\cdot\mathcal{R}\left((\Psi_q^{Z^{\mathcal{X}}})_{ij}\right)$$
$$s.t.\ (\Psi_q^{Z^{\mathcal{X}}})_i\mathbf{1} = 1, \mathcal{R}\left((\Psi_q^{Z^{\mathcal{X}}})_{ij}\right) = (\Psi_q^{Z^{\mathcal{X}}})_{ij}\log(\Psi_q^{Z^{\mathcal{X}}})_{ij}, (\Psi_q^{Z^{\mathcal{X}}})_{ij} > 0, \tag{1}$$

where $\Psi_q^{Z^{\mathcal{X}}} \in \mathbb{R}^{N\times K}$ be the similarity matrix between the data samples and the typical proxies on the $q$-th attribution. The non-negative entropy norm term $\mathcal{R}((\Psi_q^{Z^{\mathcal{X}}})_{ij}) = (\Psi_q^{Z^{\mathcal{X}}})_{ij}\log(\Psi_q^{Z^{\mathcal{X}}})_{ij}$ is set to avoid trivial solution with $\alpha$ denoting the regularization strength [2]. We will provide the optimization details on the typical sample selection algorithm in Appendix A. In short, alternatively updating $\boldsymbol{M}_Z^{\mathcal{X}}$ and $\Psi_q^{Z^{\mathcal{X}}}$ can solve Equation (1) efficiently as:

$$(\Psi_q^{Z^{\mathcal{X}}})_{ij} = \frac{\exp\left(-\zeta_{ijq}/\alpha\right)}{\sum_{k=1}^{K}\exp(-\zeta_{ikq}/\alpha)},\ (\boldsymbol{M}_Z^{\mathcal{X}})_{jq} = \frac{\sum_{i=1}^{N}(\Psi_q^{Z^{\mathcal{X}}})_{ij}\boldsymbol{Z}_{iq}^{\mathcal{X}}}{\sum_{i=1}^{N}(\Psi_q^{Z^{\mathcal{X}}})_{ij}}, \tag{2}$$

where $\zeta_{ijq} = (\boldsymbol{Z}_{iq}^{\mathcal{X}} - (\boldsymbol{M}_Z^{\mathcal{X}})_{jq})^2$. Since this problem is convex, we can obtain the stable solution of $\Psi_q^{Z^{\mathcal{X}}}$ and $\boldsymbol{M}_Z^{\mathcal{X}}$ through iterations.

**Optimal Transport Alignment.** In order to better model the attribution distribution between the source and target domains, we adopt the optimal transport technique [14]. Optimal transport is based on Kantorovich problem [1], seeking for a general coupling $\boldsymbol{\pi}_q^Z \in \mathcal{X}(\mathcal{S}_q, \mathcal{T}_q)$ between $\mathcal{S}_q$ and $\mathcal{T}_q$ on the $q$-th attribution:

$$\hat{\boldsymbol{\pi}}_q^Z = \arg\min\int \mathcal{M}((\boldsymbol{M}_Z^{\mathcal{S}})_{*q}, (\boldsymbol{M}_Z^{\mathcal{T}})_{*q})\, d\boldsymbol{\pi}_q^Z((\boldsymbol{M}_Z^{\mathcal{S}})_{*q}, (\boldsymbol{M}_Z^{\mathcal{T}})_{*q}).$$

The cost function matrix $\mathcal{M}((\boldsymbol{M}_Z^{\mathcal{S}})_{*q}, (\boldsymbol{M}_Z^{\mathcal{T}})_{*q})$ denotes the cost to move probability mass from $(\boldsymbol{M}_Z^{\mathcal{S}})_{*q}$ to $(\boldsymbol{M}_Z^{\mathcal{T}})_{*q}$ where $(\boldsymbol{M}_Z^{\mathcal{S}})_{*q}$ and $(\boldsymbol{M}_Z^{\mathcal{T}})_{*q}$ denote the typical selected samples on the $q$-th attribution for source and target domain respectively. The discrete optimal transport formulation can be expressed as:

$$\hat{\boldsymbol{\pi}}_q^Z = \arg\min\left[\langle\boldsymbol{\pi}_q^Z, \mathcal{M}\rangle_F + \epsilon\mathcal{R}(\boldsymbol{\pi}_q^Z)\right], \tag{3}$$

where $\hat{\boldsymbol{\pi}}_q^Z \in \mathbb{R}^{K\times K}$ is the ideal coupling matrix between the source typical samples $(\boldsymbol{M}_Z^{\mathcal{S}})_{*q}$ and the target typical samples $(\boldsymbol{M}_Z^{\mathcal{T}})_{*q}$. The matching matrix follows the constraint of $\mathbf{1}_K\boldsymbol{\pi}_q^Z = \mathbf{1}_K\left(\boldsymbol{\pi}_q^Z\right)^T = \frac{1}{K}\mathbf{1}_K$. The second term $\mathcal{R}(\boldsymbol{\pi}_q^Z)$ is the regularization term and $\epsilon$ is a hyper-parameter to balance the entropy regularization and matching loss. The matrix $\mathcal{M} \in \mathbb{R}^{K\times K}$ denotes the pairwise distance as $\mathcal{M}[i][j] = ||(\boldsymbol{M}_Z^{\mathcal{S}})_{iq} - (\boldsymbol{M}_Z^{\mathcal{T}})_{jq}||_2^2$. Therefore, we can calculate the optimal transport distance $d_O$ as:

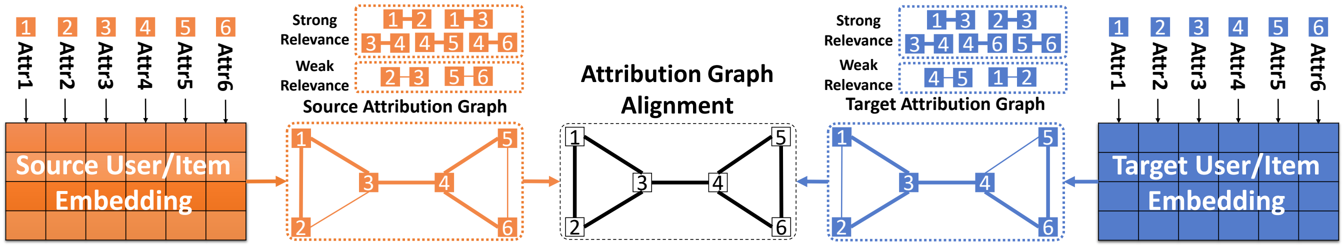
$$d_O((\boldsymbol{M}_Z^{\mathcal{S}})_{*q}, (\boldsymbol{M}_Z^{\mathcal{T}})_{*q}) = \frac{1}{K^2}\sum_{i=1}^{K}\sum_{j=1}^{K}\left[\hat{\boldsymbol{\pi}}_q^Z\right]_{ij}||(\boldsymbol{M}_Z^{\mathcal{S}})_{iq} - (\boldsymbol{M}_Z^{\mathcal{T}})_{jq}||_2^2. \tag{4}$$

In summary, we propose the attribution distribution distance alignment loss as below:

$$L_O = \frac{1}{D}\sum_{q=1}^{D}\left[d_O((\boldsymbol{M}_U^{\mathcal{S}})_{*q}, (\boldsymbol{M}_U^{\mathcal{T}})_{*q}) + d_O((\boldsymbol{M}_V^{\mathcal{S}})_{*q}, (\boldsymbol{M}_V^{\mathcal{T}})_{*q})\right]. \tag{5}$$

Take Fig. 3 for example, we first collect the typical samples of Attr3, e.g., the stars marked with numbers in Fig. 3(c), in both source and target domains. Then we adopt the Optimal Transport (OT) to align these typical samples across domains. Afterward, we minimize the distance between the matched typical samples. Finally, we can align the attribution probability distribution in Fig. 3(d).

*3.2.2 Horizontal Attribution Alignment.* Then we introduce the horizontal attribution alignment algorithm. Previous researches have pointed out that aligning the corresponding relevant attribution relationship, e.g., adopting the correlation alignment with covariance matrix in ESAM [8], can better enhance the model performance. However, covariance is hard to capture the complex and nonlinear hidden relationships between different attributions under the RNCDR problem [6]. Therefore, we propose horizontal attribution alignment with attribution subspace modelling and attribution graph alignment methods. We first use attribution subspace modelling to build the attribution graph, exploiting the hierarchical




(a) Building up source attribution graph (b) Attribution graph alignment (c) Building up target attribution graph

**Figure 4: The main procedure of horizontal attribution alignment.**

and topological structure between different attributions. Then we propose attribution graph alignment to align attribution graphs in source and target domains with Wasserstein distance metric.

**Attribution Subspace Modelling.** SLIM (Sparse LInear Methods) [10, 33] have been widely adopted in recommendation systems due to its good performance. It computes the item-item relations with statistical learning using the corresponding coefficient matrix. Meanwhile, one can even adopt it to measure the embedding attribution-attribution relations according to the following optimization problem:

$$\min_{B_Z^{\mathcal{X}}} \frac{1}{2}\left\|Z^{\mathcal{X}} - Z^{\mathcal{X}} B_Z^{\mathcal{X}}\right\|_2^2 + \nu \left\|B_Z^{\mathcal{X}}\right\|_*, s.t.\ \mathrm{diag}\left(B_Z^{\mathcal{X}}\right) = 0, \quad (6)$$

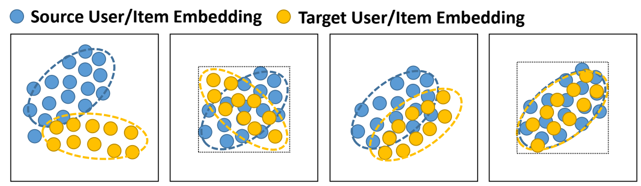
where $\mathcal{X} = \{\mathcal{S}, \mathcal{T}\}$ denotes the domain index, $Z = \{U, V\}$ denotes the user and item set, $\mathrm{diag}(B_Z^{\mathcal{X}}) = 0$ is a constant that avoids the trivial solution, $\nu$ is the balance hyper parameter, and $||\cdot||_*$ is the nuclear-norm which can make the matrix to become low rank [10]. The low rank constraint can also enhance the robustness and generalization of $B_Z^{\mathcal{X}}$. Considering that $||B_Z^{\mathcal{X}}||_* = \mathrm{Tr}\left(\sqrt{(B_Z^{\mathcal{X}})^T B_Z^{\mathcal{X}}}\right)$, the original optimization problem Equation (6) can be rewritten as:

$$\begin{aligned} &\min_{B_Z^{\mathcal{X}}, \Phi} \frac{1}{2}\left\|Z^{\mathcal{X}} - Z^{\mathcal{X}} B_Z^{\mathcal{X}}\right\|_2^2 + \nu \mathrm{Tr}\left((B_Z^{\mathcal{X}})^T \Phi B_Z^{\mathcal{X}}\right) \\ &s.t.\ \mathrm{diag}\left(B_Z^{\mathcal{X}}\right) = 0, \Phi = \left(B_Z^{\mathcal{X}} (B_Z^{\mathcal{X}})^T\right)^{-\frac{1}{2}}. \end{aligned} \quad (7)$$

Alternatively updating $B_Z^{\mathcal{X}}$ and $\Phi$ can solve Equation (7) efficiently. For $B_Z^{\mathcal{X}}$, it has the following closed-form solution:

$$\left(B_Z^{\mathcal{X}}\right)_{ij} = \begin{cases} 0, & i = j \\ -\dfrac{\Theta_{ij}}{\Theta_{jj}}, & \text{Others.} \end{cases} \quad (8)$$

where $\Theta = ((Z^{\mathcal{X}})^T Z^{\mathcal{X}} + \nu(\Phi + \Phi^T))^{-1}$. After we have updated $B_Z^{\mathcal{X}}$, we fix it as a constant and update $\Phi$ through the equality constraint $\Phi = (B_Z^{\mathcal{X}} (B_Z^{\mathcal{X}})^T)^{-\frac{1}{2}}$. By utilizing the iterative updating method until it converges, we can obtain the results of $B_Z^{\mathcal{X}}$ and $\Phi$. The optimization details will be provided in Appendix B. Notably that $B_Z^{\mathcal{X}}$ may be asymmetric and contain negative values, we build up the attribution graph by taking the average value of $|B_Z^{\mathcal{X}}|$ and $|(B_Z^{\mathcal{X}})^T|$ which can be depicted as $A_Z^{\mathcal{X}} = (|B_Z^{\mathcal{X}}| + |(B_Z^{\mathcal{X}})^T|)/2$, where $(A_Z^{\mathcal{X}})_{ij}$ depicts the attribution similarity on the $i$-th and $j$-th attribution [23]. Meanwhile $A_Z^{\mathcal{X}}$ can be viewed as the graph adjacent matrix where each node denotes the corresponding attribution. Therefore, we can establish the attribution graph to represent the topology structure among these attributions, as shown in Fig. 4(a) and Fig. 4(c) in the source and target domains.




(a) Initial (b) Vertical Attribution Alignment (c) Horizontal Attribution Alignment (d) Embedding Attribution Alignment

**Figure 5: The illustrations of attribution alignment.**

**Attribution Graph Alignment.** After we obtain the attribution graph through the adjacent matrix $A_Z^{\mathcal{X}}$, we can match the source and target attribution graph across domains in Fig. 4(b). To start with, we first calculate the corresponding Laplacian matrix $L_Z^{\mathcal{S}} = D_Z^{\mathcal{S}} - A_Z^{\mathcal{S}}$ and $L_Z^{\mathcal{T}} = D_Z^{\mathcal{S}} - A_Z^{\mathcal{S}}$ through the origin adjacent matrix where $D_Z^{\mathcal{S}}$ and $D_Z^{\mathcal{T}}$ denote the degree matrix on graph.

According to previous research [37], a graph can be interpreted as a multivariate normal distribution. Specifically, the mean of the multivariate normal distribution is zero and the covariance is the inverse Laplacian matrix $(L_Z^{\mathcal{S}})^+$ and $(L_Z^{\mathcal{T}})^+$, shown as below:

$$\mathbb{P}(\mathcal{G}_Z^{\mathcal{S}}) = \mathcal{N}(0, (L_Z^{\mathcal{S}})^+), \quad \mathbb{P}(\mathcal{G}_Z^{\mathcal{T}}) = \mathcal{N}(0, (L_Z^{\mathcal{T}})^+). \quad (9)$$

The above formulation has been further used in many graph inference algorithms to represent the graph characteristics. Therefore, we can calculate the attribution graph distance and reduce the discrepancy based on the formulation. We adopt the Wasserstein distance [27] to measure the distance between different Gaussians distributions, which is shown as:

$$\begin{aligned} &d_W(\mathcal{N}(0, (L_Z^{\mathcal{S}})^+), \mathcal{N}(0, (L_Z^{\mathcal{T}})^+)) \\ &= \mathrm{Tr}\left((L_Z^{\mathcal{S}})^+ + (L_Z^{\mathcal{T}})^+ - 2((L_Z^{\mathcal{S}})^+)^{\frac{1}{2}} (L_Z^{\mathcal{T}})^+ ((L_Z^{\mathcal{S}})^+)^{\frac{1}{2}})^{\frac{1}{2}}\right). \end{aligned} \quad (10)$$

Meanwhile the Wasserstein distance on graph is more sensitive than traditional Frobenius norm distance which can detect more subtle modification [29]. Therefore we propose the attribution map distance alignment loss as below:

$$L_A = d_W\left(\mathbb{P}(\mathcal{G}_U^{\mathcal{S}}), \mathbb{P}(\mathcal{G}_U^{\mathcal{T}})\right) + d_W\left(\mathbb{P}(\mathcal{G}_V^{\mathcal{S}}), \mathbb{P}(\mathcal{G}_V^{\mathcal{T}})\right). \quad (11)$$

The relevance between different attributions will finally meet the consensus through attribution graph alignment, as shown in Fig. 4(b). In summary, vertical and horizontal attribution alignment are both indispensable, as shown in Fig. 5. With the vertical attribution distribution alignment, one can just align the marginal distribution across domains, as shown in Fig. 5(b). Otherwise, the horizontal attribution alignment can only align the attribution relationship but cannot reduce the distribution discrepancy as shown in Fig. 5(c). Therefore, two methods can complement each other and work together to complete the task, as shown in Fig. 5(d).

### 3.3 Putting Together

The total loss of **CFAA** could be obtained by combining the losses of the rating prediction module and the embedding attribution alignment module. That is, the loss of **CFAA** is given as:

$$L_{\mathbf{CFAA}} = L_C + \lambda_O L_O + \lambda_A L_A, \tag{12}$$

where $\lambda_O$ and $\lambda_A$ are hyper-parameters to balance different type of losses. By doing this, **CFAA** can not only model the source user-item interactions, but also reduce the embedding bias and discrepancy on user and item across domains.

## 4 EMPIRICAL STUDY

In this section, we conduct experiments on several real-world datasets to answer the following questions: (1) **RQ1**: How does our approach perform compared with the state-of-the-art CDR methods? (2) **RQ2**: How do the vertical and horizontal attribution alignments contribute to performance improvement? (3) **RQ3**: How does the performance of **CFAA** vary with different values of the hyper-parameters?

### 4.1 Datasets and Tasks

We conduct extensive experiments on two popularly used real-world datasets, i.e., *Douban* and *Amazon*. First, the **Douban** dataset [53, 55] has three domains, i.e., Book, Music, and Movie. Second, the **Amazon** dataset [31, 51] has five domains, i.e., Movies and TV (Movie), Books (Book), CDs and Vinyl (Music), Instant Videos (Video) and Clothes (Clothes). Both datasets have user-item ratings and reviews. The detailed statistics of these datasets after pre-process are shown in Appendix C. We select the relative large datasets (e.g., **Amazon Movie**, **Douban Book**) as the source domains and the rest as the target domains. We remove the users and items less than 30 records to increase the density following existing research in the source domain [48]. Meanwhile we also delete some part of interactions in the target datasets to make them more sparse. For each datasets, we binarize the ratings to 0 and 1. Specifically, we take the ratings higher or equal to 4 as 1 and others as 0. Therefore we conduct several tasks that transferring the useful knowledge from the source to the target domains. Notably that it includes both easy and hard tasks, e.g., the transfer task between **Amazon Movie** → **Amazon Video** is easy since they are rather similar, while **Amazon Movie** → **Amazon Clothes** is hard because they are different.

### 4.2 Experiment Settings

We randomly divide the observed source and target data into training, validation, and test sets with a ratio of 8:1:1. Users and items are selected to be both non-overlapped across domains. We set batch size $N = 256$ for both the source and target domains. The latent embedding dimension is set to $D = 300$. For the English reviews on Amazon, we apply the pretrained transformer based on RoBERTa-Large for semantic textual similarity task [26]. For the Chinese reviews on Douban, we apply the pretrained Chinese BERT [13]. We set $K = \frac{N}{2}$ for typical sample selection in attribution distribution alignment. For **CFAA** model, we set the balance hyper-parameters as $\lambda_O = 0.5$ and $\lambda_A = 0.8$. We set the hyper-parameters $\alpha = 0.1$ and $\nu = 0.1$ for solving the typical sample selection method and attribution subspace modelling respectively. For all the experiments, we perform five random experiments and report the average results. We choose Adam [18] as optimizer, and adopt Hit Rate@$k$ (HR@$k$), Recall@$k$, and NDCG@$k$ [45] as the ranking evaluation metrics with $k = 10$.

### 4.3 Baseline

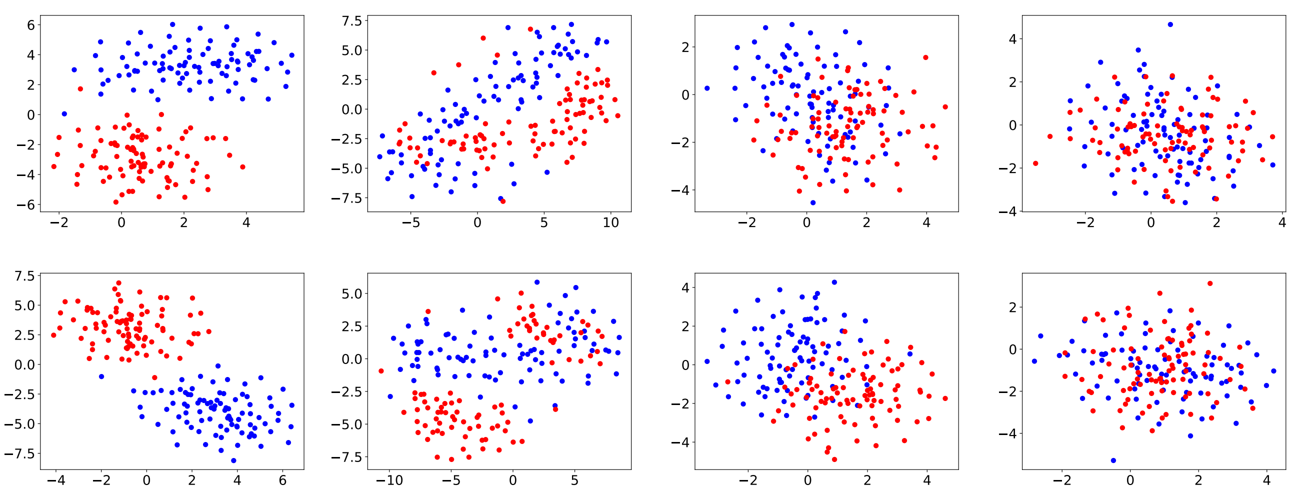
We compare our proposed **CFAA** with the following state-of-the-art recommendation models. (1) **DeepCoNN** [52] Deep Cooperative Neural Networks (DeepCoNN) is the first deep collaborative model to leverage both user and item textual features from reviews for recommendation. (2) **NARRE** [7] Neural Attentional Rating Regression with Explanations (NARRE) utilizes two parallel CNNs with attention mechanism to extract review-level information for recommendation. (3) **ESCOFILT** [36] Extractive Summarization-Based Collaborative Filtering (ESCOFILT) is a BERT-based state-of-the-art collaborative filtering model based on user and item review information. (4) **Rec-DAN** [44] Discriminative Adversarial Networks for Recommender System (Rec-DAN) adopts adversarial training strategy to align the joint user-item textual features to transfer useful knowledge. (5) **TDAR** [48] Text-enhanced Domain Adaptation (TDAR) is the state-of-the-art reviewed-based non-overlapped CDR model which adopts adversarial training strategy to user and item embeddings respectively with text memory network. (6) **ESAM** [8] Entire Space Adaptation Model (ESAM) adopts attribute correlation alignment to improve long-tail recommendation performance by suppressing inconsistent distribution between displayed and non-displayed items. (7) **DARec** [50] Deep Domain Adaptation for Cross-Domain Recommendation via Transferring Rating Patterns (DARec) adopts adversarial training strategy to extract and transfer knowledge patterns for shared users across domains. Note that the original **ESAM** and **DARec** models cannot be directly applied to RNCDR tasks, and thus we adopt the same rating prediction module as **CFAA** for them. Besides, for a fair comparison, all the models use the same types of data and pre-processing methods during experiments.

### 4.4 Recommendation Performance (RQ1)

**Results and discussion.** The comparison results on Douban and Amazon datasets are shown in Table 1. Note that **CFAA**-Base represents the model that only adopts the rating prediction module for collaborative filtering without embedding attribution alignment. From them, we can find that: (1) Only adopting the single target domain information (e.g., **DeepCoNN**) cannot obtain satisfying results under the RNCDR settings due to the data sparsity problem. (2) **CFAA**-Base outperforms the previous reviewed-based recommendation model (e.g., **ESCOFILT**), indicating that adopting BERT-based review embedding with one-hot ID and multi-hot historical rating information can efficiently enhance the model performance. (3) Although **ESAM** achieves better performance than **CFAA**-Base due to its attribution correlation congruence, it still fails to capture the nonlinear and complex topology structure among different attributions. (4) Although **DARec** with gradient reverse layer can obtain good performance, the unstable adversarial training may hurdle the model to obtain more accurate results [38]. As a result, it will eventually cause coarsely matching across domains.

**Table 1: Experimental results on Douban and Amazon datasets.**

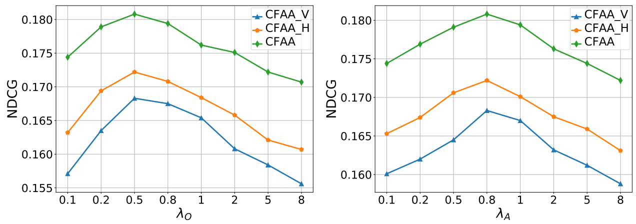
| | (Amazon) Movie→Video | | | (Amazon) Movie→Music | | | (Amazon) Movie→Clothes | | | (Amazon) Book→Video | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | HR | Recall | NDCG | HR | Recall | NDCG | HR | Recall | NDCG | HR | Recall | NDCG |
| DeepCoNN | .1459 | .1338 | .0944 | .1476 | .1588 | .0625 | .1039 | .1135 | .0517 | .1274 | .1816 | .0798 |
| NARRE | .1538 | .1413 | .1032 | .1573 | .1707 | .0948 | .1346 | .1270 | .0654 | .1491 | .1902 | .0870 |
| Rec-GAN | .1601 | .1692 | .1065 | .1787 | .1876 | .1193 | .1560 | .1402 | .0726 | .1597 | .1969 | .1033 |
| TDAR | .1782 | .1786 | .1197 | .1915 | .2043 | .1485 | .1674 | .1461 | .0802 | .1709 | .2050 | .1167 |
| ESCOFILT | .1896 | .1977 | .1304 | .2024 | .2115 | .1567 | .1751 | .1504 | .0893 | .1835 | .2198 | .1281 |
| DARec | .2044 | .2220 | .1651 | .2462 | .2417 | .1709 | .1846 | .1708 | .0995 | .2096 | .2427 | .1589 |
| ESAM | .2067 | .2275 | .1696 | .2408 | .2391 | .1680 | .1870 | .1715 | .1012 | .2113 | .2402 | .1561 |
| **CFAA**-Base | .2013 | .2168 | .1569 | .2316 | .2301 | .1624 | .1832 | .1686 | .0965 | .2048 | .2349 | .1522 |
| **CFAA**-V | .2083 | .2263 | .1683 | .2440 | .2514 | .1736 | .1865 | .1713 | .1004 | .2133 | .2486 | .1679 |
| **CFAA**-H | .2105 | .2205 | .1722 | .2391 | .2468 | .1705 | .1891 | .1735 | .1028 | .2167 | .2513 | .1705 |
| **CFAA** | **.2190** | **.2414** | **.1808** | **.2552** | **.2637** | **.1901** | **.1944** | **.1798** | **.1110** | **.2250** | **.2605** | **.1816** |
| | (Amazon) Book→Music | | | (Amazon) Book→Clothes | | | (Douban) Movie→Music | | | (Douban) Book→Music | | |
| | HR | Recall | NDCG | HR | Recall | NDCG | HR | Recall | NDCG | HR | Recall | NDCG |
| DeepCoNN | .1169 | .1624 | .0706 | .0939 | .1085 | .0439 | .1868 | .1540 | .0787 | .1485 | .1363 | .0571 |
| NARRE | .1382 | .1696 | .0933 | .1148 | .1137 | .0504 | .2037 | .1669 | .0944 | .1576 | .1580 | .0962 |
| Rec-GAN | .1599 | .1817 | .1228 | .1261 | .1206 | .0613 | .2204 | .1782 | .1193 | .1798 | .1636 | .0984 |
| TDAR | .1708 | .2032 | .1414 | .1465 | .1291 | .0688 | .2365 | .2016 | .1328 | .1913 | .1792 | .1109 |
| ESCOFILT | .1784 | .2108 | .1570 | .1587 | .1369 | .0745 | .2396 | .2103 | .1399 | .2057 | .1904 | .1250 |
| DARec | .1897 | .2295 | .1840 | .1694 | .1501 | .0873 | .2562 | .2415 | .1638 | .2203 | .2101 | .1405 |
| ESAM | .1915 | .2321 | .1909 | .1712 | .1530 | .0897 | .2589 | .2366 | .1544 | .2219 | .2133 | .1393 |
| **CFAA**-Base | .1860 | .2265 | .1761 | .1670 | .1488 | .0854 | .2530 | .2327 | .1502 | .2194 | .2075 | .1368 |
| **CFAA**-V | .1924 | .2340 | .1912 | .1699 | .1522 | .0910 | .2628 | .2410 | .1713 | .2250 | .2146 | .1421 |
| **CFAA**-H | .1913 | .2319 | .1876 | .1725 | .1543 | .0941 | .2605 | .2424 | .1690 | .2236 | .2137 | .1434 |
| **CFAA** | **.2021** | **.2493** | **.2015** | **.1782** | **.1595** | **.1016** | **.2715** | **.2548** | **.1821** | **.2321** | **.2210** | **.1496** |



(a) **CFAA**-Base (b) DARec (c) ESAM (d) **CFAA**

**Figure 6: The t-SNE visualization of user latent embeddings on Douban Movie → Douban Music (first row) and item latent embeddings (second row). The user (item) latent embeddings in the source domain are shown with red dots and that in the target domain are shown with blue dots.**

Meanwhile although **DARec** and **TDAR** both utilizes adversarial training, **DARec** equipped with more expressive user and item embeddings can obtain more delightful results. (5) **CFAA** consistently achieves the best performance, which proves that embedding




(a) Effects of $\lambda_O$ (b) Effects of $\lambda_A$

**Figure 7: (a)-(b) show the effect of hyper-parameters $\lambda_O$ and $\lambda_A$ on model performance.**

attribution alignment module with attribution distribution and relevance alignment can significantly improve the prediction accuracy. Notably that our proposed **CFAA** can enhance the performance when source and target domains are both similar (e.g., **Amazon Movie → Amazon Video**) and different (e.g., **Amazon Movie → Amazon Clothes**).

**Visualization.** To better show the user embeddings across domains, we visualize the t-SNE embeddings [19] for **CFAA**-Base, **DARec**, **ESAM**, and **CFAA**. The results of **Douban Movie → Douban Music** are shown in Fig. 6(a)-(d). The first and second row denote user and item embeddings, respectively. From it, we can see that (1) **CFAA**-Base cannot reduce the embedding bias and discrepancy on both users and items across domains, leading to insufficient knowledge transfer, as shown in Fig. 6(a). (2) **ESAM** and **DARec** can marginally align user and item embedding attributions to a certain extent, but there still exists domain discrepancy which causes negative transfer, as shown in Fig. 6(b)-(c). (3) **CFAA** with attribution distribution and relevance alignment can better match users and items across domains, as shown in Fig. 6(d). The visualization results illustrates the validity of our model.

## 4.5 Analysis (RQ2 and RQ3)

**Ablation.** To study how does each module of **CFAA** contribute on the final performance, we compare **CFAA** with its several variants, including **CFAA-V** and **CFAA-H**. **CFAA-V** only adopts the vertical attribution (i.e., attribution distribution) alignment while **CFAA-H** only adopts the horizontal attribution (i.e., attribution relevance) alignment. The comparison results are shown in Table 1. From it, we can observe that (1) **CFAA-V** and **CFAA-H** always get more accurate output predictions than **CFAA**-Base, which indicates that reducing embedding attribution alignment is essential. (2) However, **CFAA-V** and **CFAA-H** still cannot achieve the best results against **CFAA**. Simply aligning vertical embedding probability distribution on **CFAA-V** will sometimes neglect the attribution relevance and cause negative transfer. Likewise, only concentrating on the horizontal attribution alignment will sometimes ignore the attribution discrepancy across domains. Overall, the above ablation study demonstrates that our proposed embedding alignment module is effective in solving the RNCDR problem.

**Distribution Discrepancy.** The domain adaptation theory [3] suggests proxy $\mathcal{A}$-distance as a measure of cross-domain discrepancy. We adopt $d_{\mathcal{A}}(\mathcal{S}, \mathcal{T}) = 2(1 - 2\epsilon(h))$ to analysis the distance between two domains, where $\epsilon(h)$ is the generalization error of a linear classifier $h$ that discriminates the source domain $\mathcal{S}$ and the target domain $\mathcal{T}$ [3]. Table 2 demonstrates the domain discrepancy on **Amazon Movie → Amazon Music** and **Douban Movie → Douban Music** tasks using **CFAA**-Base, **CFAA-V**, **CFAA-H**, **CFAA**, and several baseline methods. From it, we can conclude that: (1) The large number of $d_{\mathcal{A}}$ on **CFAA**-Base indicates the existence of embedding attribution discrepancy between the source and target domains. (2) Most of the current baselines (e.g.,**ESAM**) can reduce $d_{\mathcal{A}}$ but they still cannot obtain lower $d_{\mathcal{A}}$ than **CFAA**, indicating their limitations on domain adaptation. (3) By adopting the embedding attribution alignment methods, **CFAA-V** and **CFAA-H** can reduce the embedding attribution discrepancy according to the $d_{\mathcal{A}}$ Furthermore, we can observe that **CFAA** achieves the lowest $d_{\mathcal{A}}$ distance, because it considers both vertical and horizontal embedding attribution alignment. Lower $d_{\mathcal{A}}$ can always obtain better results since more useful knowledge can be transferred from source to target domains, which is consistent with the previous ablation study.

**Table 2: The results on $d_{\mathcal{A}}$ for domain discrepancy.**

| | (Amazon) Movie→Music | | (Douban) Movie→Music | |
|---|---|---|---|---|
| | User | Item | User | Item |
| DARec | 1.5441 | 1.5294 | 1.4823 | 1.5045 |
| ESAM | 1.5082 | 1.5310 | 1.4656 | 1.4780 |
| **CFAA**-Base | 1.7503 | 1.7354 | 1.7372 | 1.7215 |
| **CFAA**-V | 1.4125 | 1.4496 | 1.3839 | 1.3742 |
| **CFAA**-H | 1.4039 | 1.4363 | 1.3950 | 1.3831 |
| **CFAA** | **1.3548** | **1.3635** | **1.3084** | **1.2996** |

**Effect of hyper-parameters.** We finally study the effects of hyper-parameters on model performance. For **CFAA**, we vary $\lambda_O$ and $\lambda_A$ in $\{0.1, 0.2, 0.5, 0.8, 1, 2, 5, 8\}$ and report the results in Fig. 7(a)-(b) on **Amazon Movie → Amazon Music**. Fig. 7(a)-(b) show the bell-shaped curves, indicating that choosing the proper hyper-parameters to balance the rating prediction loss and embedding attribution alignment loss can effectively improve the model performance. Empirically, we choose $\lambda_O = 0.5$ and $\lambda_A = 0.8$.

## 5 CONCLUSION

In this paper, we propose Collaborative Filtering with Attribution Alignment model for solving review-based non-overlapped cross domain recommendation (**CFAA**), which includes the *rating prediction module* and the *embedding attribution alignment module*. We innovatively adopt horizontal and vertical attribution alignment to better reduce the embedding discrepancy from different perspectives. Vertical distribution alignment utilizes the typical sample selection with optimal transport to make them consistent across domains. Horizontal relevance alignment applies the subspace modelling with attribution graph alignment to reduce the discrepancy. We also conduct extensive experiments to demonstrate the superior performance of our proposed **CFAA** on several datasets and tasks.

## ACKNOWLEDGMENTS

This work was supported in part by the National Key R&D Program of China (No.72192823 and No.62172362).

# REFERENCES


[1] Sigurd Angenent, Steven Haker, and Allen Tannenbaum. 2003. Minimizing flows for the Monge–Kantorovich problem. *SIAM journal on mathematical analysis* 35, 1 (2003), 61–97.

[2] Liang Bai and Jiye Liang. 2020. Sparse Subspace Clustering with Entropy-Norm. In *ICML*. PMLR, 561–568.

[3] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2007. Analysis of representations for domain adaptation. In *NIPS*. 137–144.

[4] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. 2006. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics* 22, 14 (2006), e49–e57.

[5] Iván Cantador, Ignacio Fernández-Tobías, Shlomo Berkovsky, and Paolo Cremonesi. 2015. Cross-domain recommender systems. In *Recommender systems handbook*. Springer, 919–959.

[6] Chao Chen, Zhihong Chen, Boyuan Jiang, and Xinyu Jin. 2019. Joint domain alignment and discriminative feature learning for unsupervised deep domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 3296–3303.

[7] Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2018. Neural attentional rating regression with review-level explanations. In *WWW*. 1583–1592.

[8] Zhihong Chen, Rong Xiao, Chenliang Li, Gangfeng Ye, Haochuan Sun, and Hongbo Deng. 2020. Esam: Discriminative domain adaptation with non-displayed items to improve long-tail performance. In *SIGIR*. 579–588.

[9] Weiyu Cheng, Yanyan Shen, Yanmin Zhu, and Linpeng Huang. 2018. DELF: A Dual-Embedding based Deep Latent Factor Model for Recommendation.. In *IJCAI*, Vol. 18. 3329–3335.

[10] Yao Cheng, Liang Yin, and Yong Yu. 2014. Lorslim: Low rank sparse linear methods for top-n recommendations. In *ICDM*. IEEE, 90–99.

[11] Paolo Cremonesi and Massimo Quadrana. 2014. Cross-domain recommendations without overlapping data: Myth or reality?. In *Proceedings of the 8th ACM Conference on Recommender systems*. 297–300.

[12] Jamie Cui, Chaochao Chen, Lingjuan Lyu, Carl Yang, and Wang Li. 2021. Exploiting Data Sparsity in Secure Cross-Platform Social Recommendation. *Advances in Neural Information Processing Systems* (2021).

[13] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101* (2019).

[14] Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. 2018. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *ECCV*. 447–463.

[15] Ali Mamdouh Elkahky, Yang Song, and Xiaodong He. 2015. A multi-view deep learning approach for cross domain user modeling in recommendation systems. In *WWW*. 278–288.

[16] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research* 17, 1 (2016), 2096–2030.

[17] 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.

[18] D. Kingma and J. Ba. 2014. Adam: A Method for Stochastic Optimization. *Computer Science* (2014).

[19] Van Der Maaten Laurens and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, 2605 (2008), 2579–2605.

[20] B. Li, Y. Qiang, and X. Xue. 2009. Can movies and books collaborate?: cross-domain collaborative filtering for sparsity reduction. *Sun Yat-sen University* 38, 4 (2009), 2052–2057.

[21] Pan Li and Alexander Tuzhilin. 2021. Dual Metric Learning for Effective and Efficient Cross-Domain Recommendations. *IEEE Transactions on Knowledge and Data Engineering* (2021).

[22] Siqing Li, Liuyi Yao, Shanlei Mu, Wayne Xin Zhao, Yaliang Li, Tonglei Guo, Bolin Ding, and Ji-Rong Wen. 2021. Debiasing Learning based Cross-domain Recommendation. In *KDD*. 3190–3199.

[23] Xiang Li, Ben Kao, Siqiang Luo, and Martin Ester. 2018. Rosc: Robust spectral clustering on multi-scale data. In *WWW*. 157–166.

[24] Ying Li, Jia-Jie Xu, Peng-Peng Zhao, Jun-Hua Fang, Wei Chen, and Lei Zhao. 2020. ATLRec: An attentional adversarial transfer learning network for cross-domain recommendation. *Journal of Computer Science and Technology* 35, 4 (2020), 794–808.

[25] Weiming Liu, Jiajie Su, Chaochao Chen, and Xiaolin Zheng. 2021. Leveraging Distribution Alignment via Stein Path for Cross-Domain Cold-Start Recommendation. *Advances in Neural Information Processing Systems* 34 (2021).

[26] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[27] Anton Mallasto and Aasa Feragen. 2017. Learning from uncertain curves: The 2-Wasserstein metric for Gaussian processes. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 5665–5674.

[28] Tong Man, Huawei Shen, Xiaolong Jin, and Xueqi Cheng. 2017. Cross-Domain Recommendation: An Embedding and Mapping Approach.. In *IJCAI*. 2464–2470.

[29] Hermina Petric Maretic, Mireille EL Gheche, Giovanni Chierchia, and Pascal Frossard. 2019. GOT: An optimal transport framework for graph comparison. *arXiv preprint arXiv:1906.02085* (2019).

[30] Derek Miller. 2019. Leveraging BERT for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165* (2019).

[31] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In *EMNLP-IJCNLP*. 188–197. https://doi.org/10.18653/v1/D19-1018

[32] Feiping Nie, Xiaoqian Wang, and Heng Huang. 2014. Clustering and Projected Clustering with Adaptive Neighbors. In *KDD (KDD '14)*. 977–986.

[33] Xia Ning and George Karypis. 2011. Slim: Sparse linear methods for top-n recommender systems. In *ICDM*. IEEE, 497–506.

[34] Rong Pan, Yunhong Zhou, Bin Cao, Nathan N Liu, Rajan Lukose, Martin Scholz, and Qiang Yang. 2008. One-class collaborative filtering. In *ICDM*. IEEE, 502–511.

[35] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. 2011. Domain Adaptation via Transfer Component Analysis. *IEEE Transactions on Neural Networks* 22, 2 (2011), 199–210.

[36] Reinald Adrian Pugoy and Hung-Yu Kao. 2021. Unsupervised Extractive Summarization-Based Representations for Accurate and Explainable Collaborative Filtering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2981–2990.

[37] Havard Rue and Leonhard Held. 2005. *Gaussian Markov random fields: theory and applications*. CRC press.

[38] Rui Shu, Hung H Bui, Hirokazu Narui, and Stefano Ermon. 2018. A dirt-t approach to unsupervised domain adaptation. *arXiv preprint arXiv:1802.08735* (2018).

[39] Harald Steck. 2019. Embarrassingly shallow autoencoders for sparse data. In *WWW*. 3251–3257.

[40] Baochen Sun and Kate Saenko. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*. Springer, 443–450.

[41] B. Sun and K. Saenko. 2016. Deep CORAL: Correlation Alignment for Deep Domain Adaptation. *Springer International Publishing* (2016).

[42] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. 2018. A survey on deep transfer learning. In *International conference on artificial neural networks*. Springer, 270–279.

[43] Yanchao Tan, Carl Yang, Xiangyu Wei, Yun Ma, and Xiaolin Zheng. 2021. Multi-Facet Recommender Networks with Spherical Optimization. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE, 1524–1535.

[44] Cheng Wang, Mathias Niepert, and Hui Li. 2019. Recsys-dan: discriminative adversarial networks for cross-domain recommender systems. *IEEE transactions on neural networks and learning systems* 31, 8 (2019), 2731–2740.

[45] Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. 2019. Kgat: Knowledge graph attention network for recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 950–958.

[46] Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. 2019. Characterizing and avoiding negative transfer. In *CVPR*. 11293–11302.

[47] Weike, Pan, , , Qiang, and Yang. 2013. Transfer learning in heterogeneous collaborative filtering domains. *Artificial Intelligence* (2013).

[48] Wenhui Yu, Xiao Lin, Junfeng Ge, Wenwu Ou, and Zheng Qin. 2020. Semi-supervised collaborative filtering by text-enhanced domain adaptation. In *KDD*. 2136–2144.

[49] Bowen Yuan, Jui-Yang Hsia, Meng-Yuan Yang, Hong Zhu, Chih-Yao Chang, Zhenhua Dong, and Chih-Jen Lin. 2019. Improving ad click prediction by considering non-displayed events. In *CIKM*. 329–338.

[50] Feng Yuan, Lina Yao, and Boualem Benatallah. 2019. DARec: Deep domain adaptation for cross-domain recommendation via transferring rating patterns. *arXiv preprint arXiv:1905.10760* (2019).

[51] Cheng Zhao, Chenliang Li, Rong Xiao, Hongbo Deng, and Aixin Sun. 2020. CATN: Cross-Domain Recommendation for Cold-Start Users via Aspect Transfer Network. 229–238. https://doi.org/10.1145/3397271.3401169

[52] Lei Zheng, Vahid Noroozi, and Philip S Yu. 2017. Joint deep modeling of users and items using reviews for recommendation. In *WSDM*. 425–434.

[53] Feng Zhu, Chaochao Chen, Yan Wang, Guanfeng Liu, and Xiaolin Zheng. 2019. DTCDR: A framework for dual-target cross-domain recommendation. In *CIKM*. 1533–1542.

[54] Feng Zhu, Yan Wang, Chaochao Chen, Jun Zhou, Longfei Li, and Guanfeng Liu. 2021. Cross-Domain Recommendation: Challenges, Progress, and Prospects. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*. 4721–4728.

[55] Feng Zhu, Yan Wang, Jun Zhou, Chaochao Chen, Longfei Li, and Guanfeng Liu. 2021. A unified framework for cross-domain and cross-system recommendations. *IEEE Transactions on Knowledge and Data Engineering* (2021).

# A TYPICAL SAMPLE SELECTION METHOD

As mentioned in Section 3.2.1, the typical sample selection algorithm is given by:

$$\min \ell_q = \sum_{i=1}^{N}\sum_{j=1}^{K}(\boldsymbol{\Psi}_q^{Z^X})_{ij}\left(\boldsymbol{Z}_{iq}^X - (\boldsymbol{M}_Z^X)_{jq}\right)^2 + \alpha \cdot \mathcal{R}\left((\boldsymbol{\Psi}_q^{Z^X})_{ij}\right)$$
$$s.t.\ (\boldsymbol{\Psi}_q^{Z^X})_i\mathbf{1} = 1, \mathcal{R}\left((\boldsymbol{\Psi}_q^{Z^X})_{ij}\right) = (\boldsymbol{\Psi}_q^{Z^X})_{ij}\log(\boldsymbol{\Psi}_q^{Z^X})_{ij}, (\boldsymbol{\Psi}_q^{Z^X})_{ij} > 0. \tag{13}$$

We now provide the optimization details on the typical-proxies algorithm. Alternatively updating $\boldsymbol{M}_Z^X$ and $\boldsymbol{\Psi}_q^{Z^X}$ can solve Equation (13) efficiently.

**Update $\boldsymbol{\Psi}_q^{Z^X}$.** We first fix $\boldsymbol{M}_Z^X$ and update $\boldsymbol{\Psi}_q^{Z^X}$. By using Lagrangian multiplier to minimize the objective function, we have:

$$\ell_Q = \sum_{i=1}^{N}\sum_{j=1}^{K}(\boldsymbol{\Psi}_q^{Z^X})_{ij}\zeta_{ijq} + \alpha \cdot \mathcal{R}((\boldsymbol{\Psi}_q^{Z^X})_{ij}) + \chi\sum_{i=1}^{N}\left(\sum_{j=1}^{K}(\boldsymbol{\Psi}_q^{Z^X})_{ij} - 1\right), \tag{14}$$

where $\zeta_{ijq} = (\boldsymbol{Z}_{iq}^X - (\boldsymbol{M}_Z^X)_{jq})^2$ and $\chi$ is the Lagrangian multiplier. Taking the differentiation of Equation (14) w.r.t. $(\boldsymbol{\Psi}_q^{Z^X})_{ij}$ and setting it to 0, we obtain:

$$\frac{\partial \ell_Q}{\partial(\boldsymbol{\Psi}_q^{Z^X})_{ij}} = \zeta_{ijq} + \alpha(\log(\boldsymbol{\Psi}_q^{Z^X})_{ij} + 1) + \chi = 0. \tag{15}$$

By solving and simplifying Equation (15), we have:

$$(\boldsymbol{\Psi}_q^{Z^X})_{ij} = \exp\left(-\frac{\alpha+\chi}{\alpha}\right)\exp\left(-\frac{\zeta_{ijq}}{\alpha}\right). \tag{16}$$

Meanwhile, taking $\sum_{j=1}^{K}(\boldsymbol{\Psi}_q^{Z^X})_{ij} = 1$ into Equation (16), we have:

$$\sum_{j=1}^{K}\exp\left(-\frac{\alpha+\chi}{\alpha}\right)\exp\left(-\frac{\zeta_{ijq}}{\alpha}\right) = \exp\left(-\frac{\alpha+\chi}{\alpha}\right)\sum_{j=1}^{K}\exp\left(-\frac{\zeta_{ijq}}{\alpha}\right) = 1. \tag{17}$$

That is,

$$\exp\left(-\frac{\alpha+\chi}{\alpha}\right) = \frac{1}{\sum_{j=1}^{K}\exp\left(-\frac{\zeta_{ijq}}{\alpha}\right)}. \tag{18}$$

Thus, the final solution of $(\boldsymbol{\Psi}_q^{Z^X})_{ij}$ is given by:

$$(\boldsymbol{\Psi}_q^{Z^X})_{ij} = \frac{\exp\left(-\zeta_{ijq}/\alpha\right)}{\sum_{k=1}^{K}\exp(-\zeta_{ikq}/\alpha)}. \tag{19}$$

**Update $\boldsymbol{M}_Z^X$.** After we have updated $\boldsymbol{\Psi}_q^{Z^X}$, we fix it as a constant and update $\boldsymbol{M}_Z^X$. Thus, Equation (14) becomes

$$\min_{\boldsymbol{M}_Z^X}\sum_{i=1}^{N}\sum_{j=1}^{K}(\boldsymbol{\Psi}_q^{Z^X})_{ij}||\boldsymbol{Z}_{iq}^X - (\boldsymbol{M}_Z^X)_{jq}||_2^2. \tag{20}$$

Taking the differentiation of Equation (20) w.r.t. $\boldsymbol{M}_Z^X$ and setting it to 0, we can update $\boldsymbol{M}_Z^X$ as:

$$(\boldsymbol{M}_Z^X)_{jq} = \frac{\sum_{i=1}^{N}(\boldsymbol{\Psi}_q^{Z^X})_{ij}\boldsymbol{Z}_{iq}^X}{\sum_{i=1}^{N}(\boldsymbol{\Psi}_q^{Z^X})_{ij}}. \tag{21}$$

We can obtain the stable solution of $\boldsymbol{\Psi}_q^{Z^X}$ and $\boldsymbol{M}_Z^X$ through several iterations.

**Table 3: Statistics on Douban and Amazon datasets.**

| Datasets | Items | Users | Interactions | Density |
|---|---|---|---|---|
| Amazon Movie (S) | 50,052 | 123,960 | 1,697,532 | 0.027% |
| Amazon Book (S) | 43,168 | 95,643 | 1,032,019 | 0.025% |
| Douban Movie (S) | 34,893 | 151,258 | 1,278,401 | 0.024% |
| Douban Book (S) | 38,776 | 111,270 | 965,041 | 0.022% |
| Amazon Music (T) | 7,710 | 11,053 | 106,188 | 0.124% |
| Amazon Video (T) | 1,580 | 4,555 | 11,137 | 0.155% |
| Amazon Clothes (T) | 21,554 | 35,669 | 89,176 | 0.012% |
| Douban Music (T) | 3,562 | 11,278 | 2,1451 | 0.053% |

# B ATTRIBUTION SUBSPACE MODELLING

As mentioned in Section 3.2.2, the attribution subspace modelling algorithm is given by:

$$\min_{\boldsymbol{B}_Z^X,\boldsymbol{\Phi}}\frac{1}{2}\left\|\boldsymbol{Z}^X - \boldsymbol{Z}^X\boldsymbol{B}_Z^X\right\|_2^2 + \nu\mathrm{Tr}\left((\boldsymbol{B}_Z^X)^T\boldsymbol{\Phi}\boldsymbol{B}_Z^X\right)$$
$$s.t.\ \mathrm{diag}\left(\boldsymbol{B}_Z^X\right) = 0, \boldsymbol{\Phi} = \left(\boldsymbol{B}_Z^X(\boldsymbol{B}_Z^X)^T\right)^{-\frac{1}{2}}. \tag{22}$$

Alternatively updating $\boldsymbol{B}_Z^X$ and $\boldsymbol{\Phi}$ can solve Equation (22) efficiently.

We first fix $\boldsymbol{\Phi}$ and update $\boldsymbol{B}_Z^X$. By using Lagrangian multiplier to minimize the objective function, we have:

$$\ell_B = \frac{1}{2}\left\|\boldsymbol{Z}^X - \boldsymbol{Z}^X\boldsymbol{B}_Z^X\right\|_2^2 + \nu\mathrm{Tr}\left((\boldsymbol{B}_Z^X)^T\boldsymbol{\Phi}\boldsymbol{B}_Z^X\right) + \boldsymbol{\gamma}\mathrm{diag}(\boldsymbol{B}_Z^X). \tag{23}$$

Taking the differentiation of Equation (23) w.r.t. $\boldsymbol{B}_Z^X$ and setting it to 0, we obtain:

$$\frac{\partial \ell_B}{\partial \boldsymbol{B}_Z^X} = (\boldsymbol{Z}^X)^T(\boldsymbol{Z}^X\boldsymbol{B}_Z^X - \boldsymbol{Z}^X) + \nu\Xi\boldsymbol{B}_Z^X + \mathrm{diagMat}(\boldsymbol{\gamma}) = 0, \tag{24}$$

where $\boldsymbol{\Phi} + \boldsymbol{\Phi}^T = \Xi$. diagMat$(\cdot)$ denotes the diagonal matrix. By solving and simplifying Equation (24), we have:

$$\boldsymbol{B}_Z^X = ((\boldsymbol{Z}^X)^T\boldsymbol{Z}^X + \nu\Xi)^{-1}((\boldsymbol{Z}^X)^T\boldsymbol{Z}^X - \mathrm{diagMat}(\boldsymbol{\gamma})). \tag{25}$$

Here, we suppose that $\nu$ is always sufficient large and $((\boldsymbol{Z}^X)^T\boldsymbol{Z}^X + \nu\Xi)^{-1}$ is invertible [39]. We define $\boldsymbol{\Theta} = ((\boldsymbol{Z}^X)^T\boldsymbol{Z}^X + \nu\Xi)^{-1}$ and substitute it into Equation (24):

$$\boldsymbol{B}_Z^X = \boldsymbol{I} - \boldsymbol{\Theta}\cdot\mathrm{diagMat}(\nu\mathbf{1} + \boldsymbol{\gamma}). \tag{26}$$

The unknown value of $\boldsymbol{\gamma}$ can be solved by the diagonal constraint $\mathrm{diag}(\boldsymbol{B}_Z^X) = 0$. We can obtain that $\boldsymbol{\gamma} = \mathbf{1}\oslash\mathrm{diag}(\Xi)$, where $\oslash$ denotes the elementwise division. Thus, the final solution of $\boldsymbol{B}_Z^X$ is given by:

$$\left(\boldsymbol{B}_Z^X\right)_{ij} = \begin{cases} 0, & i = j \\ -\dfrac{\boldsymbol{\Theta}_{ij}}{\boldsymbol{\Theta}_{jj}}, & \text{Others.}\end{cases} \tag{27}$$

After we have updated $\boldsymbol{B}_Z^X$, we fix it as a constant and update $\boldsymbol{\Phi}$ through the equality constraint $\boldsymbol{\Phi} = (\boldsymbol{B}_Z^X(\boldsymbol{B}_Z^X)^T)^{-\frac{1}{2}}$. Utilizing this iteration method until it converges, we can obtain the results of $\boldsymbol{B}_Z^X$ and $\boldsymbol{\Phi}$.

# C DATASETS AND TASKS

We conduct extensive experiments on two popularly used real-world datasets, i.e., *Douban* and *Amazon*. The **Douban** dataset includes Book, Music, and Movie and the **Amazon** dataset has five domains, i.e., Movies and TV (Movie), Books (Book), CDs and Vinyl (Music), Instant Videos (Video) and Clothes (Clothes). The detailed statistics of these datasets after pre-process are shown in Table 3.