Explainable Multi-Modality Alignment for Transferable Recommendation

Shenghao Yang
DCST, Tsinghua University
Quan Cheng Laboratory
Beijing, China
yangshenghao@mail.tsinghua.edu.cn

Weizhi Ma*
AIR, Tsinghua University
Beijing, China
mawz@tsinghua.edu.cn

Zhiqiang Guo
DCST, Tsinghua University
Beijing, China
georgeguo.gzq.cn@gmail.com

Min Zhang*
DCST, Tsinghua University
Quan Cheng Laboratory
Beijing, China
z-m@tsinghua.edu.cn

Haiyang Wu
Machine learning platform
department, Tencent TEG
Beijing, China
gavinwu@tencent.com

Junjie Zhai
Machine learning platform
department, Tencent TEG
Beijing, China
jasonzhai@tencent.com

Chunhui Zhang
Tencent Inc.
Beijing, China
carriezhang@tencent.com

Yuekui Yang
DCST, Tsinghua University
Machine learning platform
department, Tencent TEG
Beijing, China
yuekuiyang@tencent.com

Abstract

With the development of multi-modal modeling techniques, recent recommender systems enhance transferability by incorporating cross-domain universal multi-modal data, e.g., text and image. Existing methods typically adopt pairwise alignment to alleviate the gap between modalities. However, this alignment paradigm has limitations on explainability, consistency, and expansibility, resulting in suboptimal performance. This paper proposes a novel **E**xplainable generative multi-modality **A**lignment method for transferable **Rec**ommender systems, i.e., **EARec**. Specifically, we design a two-stage framework to achieve explainable modality alignment in the source domain and recommendation based on aligned modality representations in the target domain. In the first stage, we adopt a generative task to align various modalities in parallel to a shared anchor with explainable meaning. All modalities share the same anchor to ensure a consistent direction. Additionally, we treat behavior as an independent modality to integrate task-specific information into the alignment framework. In the second stage, we compose multiple item modality representation models trained in the first stage to obtain a unified model capable of understanding various modalities simultaneously, thereby providing high-quality item modality representations for recommendations in the target domain. Benefiting from the approach of parallel modality alignment followed by model composition, the framework demonstrates flexibility in expanding new modalities. Experimental results on multiple public datasets demonstrate the superiority of EARec over baselines, and further analyses indicate the explainability and expansibility of the proposed alignment method.

Keywords

Transferable recommendation, Multi-modality alignment, Explainable alignment

ACM Reference Format:
Shenghao Yang, Weizhi Ma, Zhiqiang Guo, Min Zhang, Haiyang Wu, Junjie Zhai, Chunhui Zhang, and Yuekui Yang. 2025. Explainable Multi-Modality Alignment for Transferable Recommendation. In . ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Conventional sequential recommendation methods model item representations based on item IDs, which are non-shared across domains and limit these models' transferability. In recent years, due to its cross-domain generalizability, multi-modality information has been used in item representation learning to achieve transferable recommendation [3, 8, 11, 20].[1]

Early transferable recommendation methods typically introduce a single modality to learn cross-domain universal item transition patterns [3, 8]. Subsequent studies [11, 20] investigate harnessing more modality types, achieving sufficient performance improvement. However, as shown in Figure 1(a), different modalities usually have distinct information richness, offering varied perspectives (e.g., color or style in vision, and brand or quantity in text) on an item.

*Corresponding authors
This work is supported by the Natural Science Foundation of China (Grant No. U21B2026, 62372260).

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[1] Our work is related to 'User modeling, personalization and recommendation" track

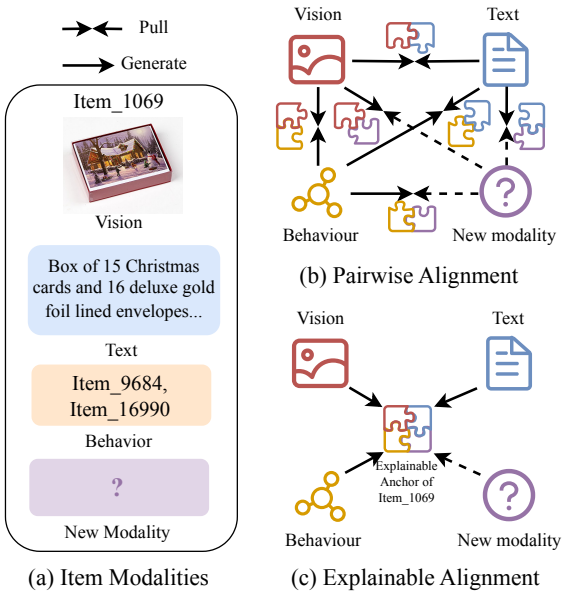

Figure 1: Illustrations of (a) various modalities of an item; (b) pairwise alignment paradigm; (c) our proposed explainable generative alignment paradigm.

How to bridge the gap between modalities to promote universal multi-modality item representation learning become a new issue.

Some recent works attempt to utilize the widespread *pairwise* alignment paradigm to achieve the alignment of modalities [11, 12, 20, 28]. As illustrated in Figure 1(b), the pairwise alignment paradigm mitigates the gap between modalities by modeling the representation consistency between two modalities. Despite notable success, this alignment paradigm has limitations in the following three aspects: *Explainability* of the alignment process. **Firstly**, cross-modal pairwise alignment is typically performed based on the latent high-dimensional representations of two modalities. The aligned results are still abstract and difficult to understand. *Consistency* of the aligning direction. **Secondly**, the pairwise alignment of multiple modalities may lead to the direction inconsistency problem. For example, in Figure 1(b), the visual modality needs to be aligned simultaneously with the textual and behavioral modalities. Inconsistent alignment directions can lead to an unstable alignment process, thereby resulting in unreliable modality representations. *Expansibility* of new modalities. **Thirdly**, as shown by the dotted line in Figure 1(b), the new modality needs to be realigned with all existing modalities, which greatly increases the complexity of training and the difficulty of adapting to new modalities.

To alleviate the above limitation of pair-wise alignment, in this paper, we propose a **E**xplainable multi-modality **A**lignment method for transferable **Rec**ommender systems, **EARec**. Specifically, we achieve EARec by solving the following three challenges:

Challenge 1: How to design the alignment framework to ensure explainability, consistency, and expansibility? We propose an explainable generative alignment method, which aligns the multimodal information of items into a unified explainable space, as shown in Figure 1(c). Considering that large language models (LLMs) can comprehend different modality inputs and generate responses, we implement the generative alignment method based on LLMs. Specifically, we fine-tune the LLM to take inputs from different item modalities and generate the same output. This shared alignment objective across modalities is referred to as "anchor", which can be any unique item content, such as title or image. The consistency of this generative target facilitates the effective alignment of different modalities in subsequent model composition. Additionally, this method shows high explainability by allowing for the evaluation of alignment quality through the generated results.

Challenge 2: How to incorporate task-specific recommendation information into the alignment process? The ultimate goal of aligning modalities is to better represent items for recommendation tasks. To incorporate recommendation-specific signals into the alignment process, we treat recommendation behavior as a modality and integrate it into the alignment framework. Additionally, we add item relation information as an auxiliary signal in the generative alignment task.

Challenge 3: How to effectively utilize multiple modalities for distinct recommendation scenarios? In various recommendation scenarios, user preferences for different modalities can vary. For instance, in e-commerce platforms, users may focus more on the color and style of items (visual modality). In information stream recommendations, users may prefer items related to those they have just viewed (behavioral modality). Therefore, we adopt an adjustable modality composition method that adaptively adjusts the weights of different modalities for different recommendation scenarios, balancing the contributions of various modalities to downstream recommendation tasks and ensuring optimal performance.

To evaluate the effectiveness of our proposed EARec, we first collect item modality data from multiple domains and construct instruction samples to fine-tune the LLM through the generative alignment task. We then composite multiple LLMs that have undergone modality alignment. Subsequently, we transfer the model capable of simultaneously understanding multiple modalities to new recommendation domains. The model is used to obtain multi-modal representations of items and these modality representations are fed into the recommendation model. Experimental results indicate that, aided by the aligned item modality representations, the performance of downstream recommendation tasks achieved significant improvements.

The main contributions of our work are summarized as follows:

- We investigate a novel multi-modality alignment paradigm to alleviate the limitations of the existing pairwise alignment approach in recommendation scenarios.
- We propose an explainable multi-modality alignment method that aligns multiple modalities into a unified explainable space through shared generative alignment objectives. We incorporate recommendation-related information during the alignment process to achieve the aligned multi-modal representations conducive to recommendation tasks.
- Experimental results demonstrate that leveraging the multi-modal representation generated by explainable multi-modality alignment can effectively enhance recommendation performance.

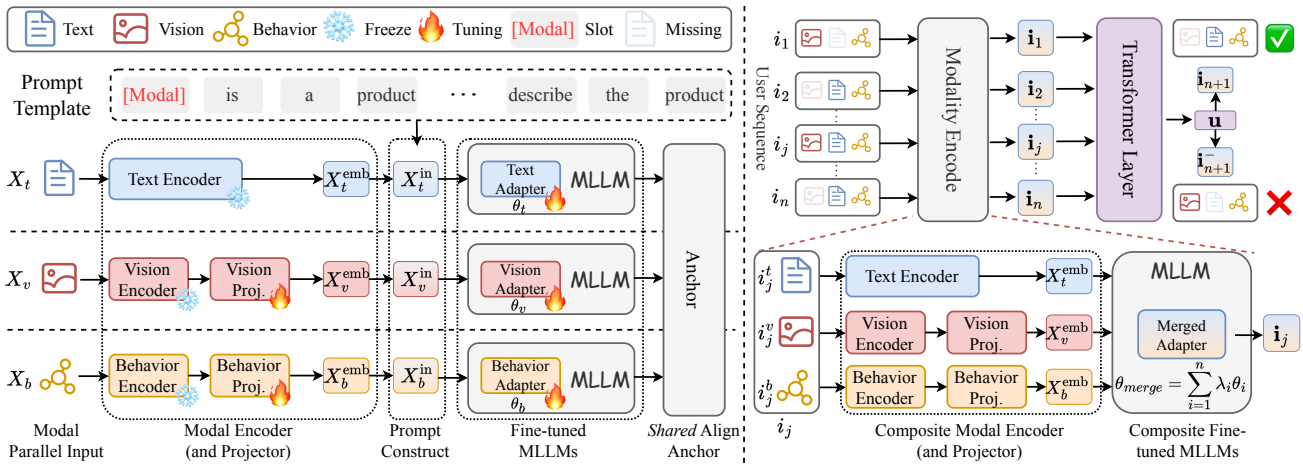

Figure 2: The overall framework of our proposed EARec, which consists of two stages for source and target domain, respectively. In stage 1, we fine-tune multiple MLLM to align three modalities, i.e., Text, Vision, and Behavior, to a shared meaningful anchor in a parallel way. In stage 2, the fine-tuned MLLMs are composited first and then serve as a modality encoder to generate the multi-modality item representations to facilitate sequential recommendation.

2 Related Work

Transferable Recommendation Transferable recommendation is a popular research area within recommendation systems, aiming to explore the effective transfer of knowledge learned from the source domain to the target domain to alleviate issues of data scarcity or cold start in the target domain. Early works often assumed an overlap of users or items between the source and target domains, using this overlap as a bridge to connect the two domains [9, 19, 23, 29]. Recently, some research [3, 7, 8, 11, 20] has begun to explore unified item representations to enhance the transferability of recommendation systems by using modality information. In particular, items are represented solely through modalities without relying on non-generalizable ID information across domains. Based on this, models can be constructed to leverage large amounts of modality data from multiple domains to learn universal item representation patterns. Then, the trained model is transferred to new domains to improve recommendation performance. However, these methods have not deeply investigated the gaps that exist between modalities, limiting the full utilization of modality information. Although MISSRec [20] and PMMRec [11] consider the issue of aligning modalities, they only adopt traditional pairwise alignment paradigms, which suffer from vague alignment direction, poor explainability, and difficulties in introducing new modalities.

Multi-modal Recommendation. Multi-modal information is prevalent in the interactions between recommendation systems and users, playing a crucial role in user decision-making. In recent years, various works have explored the incorporation of multi-modal information into user preference modeling. Early methods introduce modality information as auxiliary features or constructing modality-specific graphs for feature aggregation [6, 21, 24]. Some recent approaches attempt to address the issue of modality gaps and utilize self-supervised learning to achieve cross-modal alignment [14, 28]. However, these methods typically introduce modality features based on ID embeddings, thereby limiting them to single recommendation domains and lacking transferability.

Multi-modal Learning and Alignment Multi-modal learning has rapidly developed in fields such as computer vision and natural language processing, particularly with the rise of multi-modal large language models (MLLMs), which have brought transformative changes to multi-modal learning paradigms. Common MLLMs process both text modalities and a new modality simultaneously. This is typically achieved by aligning modalities with text and further performing instruction tuning with modality data [1, 13]. Another research direction explores enabling a single MLLM to handle multiple modalities beyond text. This can be realized by utilizing modality encoders with inherent alignment across various modalities [4] or by fine-tuning the MLLM with instruction data containing multiple modality inputs [27]. Recently, model composition approaches have made significant progress in alignment effectiveness [2, 16]. In these methods, the multi-modal alignment is achieved by generating text with the corresponding modality input, which can be seen as a generative modality alignment. However, the above alignment paradigm primarily focuses on general domains, and detailed research in the context of recommendation remains insufficient. RLMRec [18] is a recent work to perform generative alignment on representations of text and collaborative filtering in the recommendation scenario. Nevertheless, it generates representation rather than content and only performs alignment on two modalities, showing poor explainability and expansibility.

3 Method

3.1 Framework Overview

The overall framework of our proposed EARec is illustrated in Figure 2. Considering n different modalities $\{m_1, m_2, \ldots, m_n\}$, our goal is to align them into a unified explainable representation space,

thereby obtaining a model capable of simultaneously understanding multiple modalities. Specifically, we design a two-stage pipeline to achieve unified multi-modality alignment of items and the sequential recommendation task, respectively.

In the first stage, we develop a generative alignment method to align the inputs of different modalities into a unified explainable space. Specifically, we input various modalities into the customized multi-modal large language model (MLLM) and fine-tune it using the same generative objective. The strategy of parallel alignment ensures that the MLLM can fully understand and model each modality. Notably, we regard the item behavior information as an independent modality to integrate the recommendation-specific signal into MLLM. In the second stage, we draw inspiration from model composition methods [2] and composite multiple fine-tuned MLLMs to obtain a unified MLLM that can simultaneously understand different modalities. In particular, we merge the parameters of the MLLMs and integrate modal-specific components (i.e., modal encoders and projectors) into a unified framework. The composited MLLM is then utilized for the recommendation task. Due to its ability to simultaneously understand different item modalities, the composited MLLM can accept multiple modal inputs of items and generate a unified item representation. These representations can further be employed to derive user sequence representations, facilitating the prediction of the next item. Following previous works [20], Our stage 1 is conducted on multiple mixed source domain data, followed by the application of the trained MLLM in Stage 2 for recommendations in the target domain.

Next, we will introduce the explainable modality alignment method in Section 3.2 and the recommendation method based on aligned MLLM in Section 3.3.

3.2 Explainable Modality Alignment

3.2.1 Unified Generative Alignment. To achieve alignment between different modalities and mitigate the discrepancies that exist among them, we propose an explainable generative alignment method. Unlike the commonly used pairwise alignment methods based on contrastive learning, our approach employs explainable alignment objectives. We independently conduct the alignment processes for different modalities, training n models $\{M_1, M_2, \ldots, M_n\}$ that can understand various modalities. The advantage of this alignment approach is that it avoids alignment conflicts and exhibits good scalability for new modalities. To ensure that the subsequent composition effectively integrates the understanding capabilities of different models for different modalities, we employ the same alignment objective, referred to as an anchor point, during the alignment training process of the n models, thereby ensuring that these models comprehend the modalities within the same Explainable space.

Specifically, recognizing the potential of LLMs to understand different modalities and generate feedback, we fine-tune the LLM to learn to comprehend various modalities of items. For a given modality m, we construct instruction samples as input for the LLM, as illustrated in Figure 3, where the response portion represents the shared alignment objective throughout the alignment training process for all modalities. For the constructed prompt X_m^{in}, we denote the modality-specific slot as X_m and the modality-independent

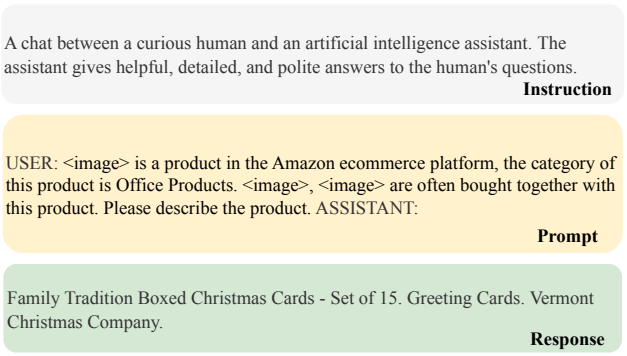

Figure 3: The instruction template of explainable generative alignment task. The response part is the shared generation objective for each modality.

prompt portion as X_p. For the different modality inputs, we employ corresponding modal encoders for vectorization, for instance, visual modalities can be encoded using models like CLIP [17]. For modality representations that is not same with the dimensions of the LLM, we map them through the corresponding modal projectors. Formally, we express this as:

$$X_m^{emb} = [MoProj(MoEnc(X_m))], \tag{1}$$

$$X_m^{in} = [X_m^{emb}, MoEnc(X_p)], \tag{2}$$

where $MoEnc$ is the modal encoder and $MoProj$ is the mapping function for modality representations. Note that for text modalities, $MoEnc$ essentially represents the word embedding of the LLM and does not require a projector. Next, we input X_m^{in} into the LLM and train it using an autoregressive generative task, defined as:

$$\mathcal{L}_{align} = -\sum_{(x,y)\in Z}\sum_{t=1}^{|y|}\log\left(P\left(y_t \mid x, y_{<t}\right)\right), \tag{3}$$

where x comprises the instruction and prompt portions in X_m^{in}, y represents the response portion, y_t is the t-th token of y, $y_{<t}$ includes all tokens preceding the t-th token, Z denotes all instruction data. Following previous work, we only train a subset of weights within the LLM, referred to as the adapter.

Through this generative task, we align the heterogeneous modality inputs into a unified feature space comprehensible to the LLM, providing a highly explainable alignment process. We can evaluate the alignment effectiveness by comparing the feedback generated by the LLM with the actual feedback, a capability that traditional pairwise alignment methods struggle to achieve.

3.2.2 Recommendation-aware Alignment. To enhance the utility of aligned modal representations for downstream recommendation tasks, we incorporate recommendation information into the alignment process from two perspectives. On one hand, we treat item behavior as a distinct item modality, reflecting the behavioral relationships among items. Understanding this modality can significantly aid in recommendation tasks. Specifically, we utilize the item embedding from an existing recommendation system as the modal encoder for the item behavior modality, a practice that holds

practical significance and can be well integrated with widely used recommendation systems. On the other hand, we introduce relationships among items into the instructions of the alignment task. As illustrated in Figure 3, in addition to the images of the items themselves, two other images of items that share a co-purchase relationship are also included as part of the prompt to help the model better comprehend the items. This relationship among items is typically introduced in prior work through knowledge graphs, employing graph embedding techniques to optimize item representations. Our approach provides a more flexible means of incorporating this knowledge to enhance the LLM's understanding and representation of modalities.

3.3 Recommendation with Aligned Modality

3.3.1 Model Composition. Through the generative alignment training in Section 3.2, we obtain n LLMs $\{M_1, M_2, \ldots, M_n\}$ that understand different modal information. Since these modalities share the same alignment anchor points during the alignment process, these MLLMs share a common Explainable space for different modalities, allowing us to composite these models to integrate their understanding capabilities across various modalities.

Specifically, in the model composition, two components need to be addressed. One part consists of modality-related components, such as the modal encoders and projectors, which serve to vectorize modal information and map it into the LLM's Explainable space. The other part comprises modality-independent components, namely the parameters of the LLM. In the model composition, we retain the encoders and projectors for different modalities, enabling us to handle inputs containing multiple types of modal information. For the n LLMs, we merge the parameters of their respective modal adapters.

Considering that the importance of different modalities varies across different downstream recommendation scenarios, it is essential to adaptively adjust the attention given to different modalities in the item modal representation for various contexts. To achieve this, we adopt an adaptive weight model composition method. Specifically, when merging the parameters of n MLLMs, we adjust the parameter weights corresponding to different modalities, formally expressed as:

$$\theta_{merge} = \sum_{i=1}^{n} \lambda_i \theta_i, \tag{4}$$

where λ_i represents the adaptive weights. In practice, the selection of λ_i can be determined by evaluating the performance of the merged model on downstream datasets in alignment tasks.

3.3.2 Downstream Recommendation. By combining multiple MLLMs, we obtain an MLLM capable of simultaneously understanding different item modalities, which we refer to as EARec. With the introduction of behavioral modalities and item relationship knowledge, EARec can effectively encode various modalities of items and serve the downstream recommendation process. We evaluate the transfer recommendation performance of EARec in a sequence recommendation task based solely on item modality representations. Specifically, for a user's time-ordered interaction sequence $s = \{i_1, i_2, \ldots, i_t\}$, the multimodal representation of item i_j is obtained through the following formula:

$$\mathbf{i}_j = \text{EARec}([i_j^t; i_j^v; i_j^b]), \tag{5}$$

where $\text{EARec}(\cdot)$ denotes the hidden state at the last position of the model's final layer as the multimodal representation of the item. Here, i_j^t, i_j^v, i_j^b represent the behavioral, textual, and visual modalities of the item, respectively. Notably, our method is not limited to these three modalities, as the proposed alignment framework can effectively extend to new modalities. It only requires completing the generative alignment task and then incorporating the model. Additionally, for any item i_j, there may be cases where a certain modality is absent. Since EARec does not require simultaneous input of modality data during alignment, it can effectively address the issue of modality absence.

Subsequently, we follow prior work by employing a transformer layer to aggregate the item representations from user interactions to obtain sequence representations. Specifically, the input to the model is the sum of the multimodal representation of the item $\mathbf{i}_j \in \mathbb{R}^d$ and the absolute positional embedding $\mathbf{p}_j \in \mathbb{R}^d$:

$$\mathbf{f}_j^0 = \mathbf{i}_j + \mathbf{p}_j. \tag{6}$$

The entire sequence $\mathbf{F}^0 = [\mathbf{f}_1^0; \ldots; \mathbf{f}_n^0] \in \mathbb{R}^{n \times d}$ is then input into L layers of transformer layers, where the output of the $l+1$ layer is:

$$\mathbf{F}^{l+1} = \text{FFN}(\text{MHAttn}(\mathbf{F}^l)). \tag{7}$$

We take the hidden state at the last position of the L layer, $\mathbf{f}_n^L \in \mathbb{R}^d$, as the representation of the user sequence $\mathbf{u} \in \mathbb{R}^d$.

Finally, the prediction score for the next item is obtained by calculating the inner product between the user sequence representation and the candidate item representation. During training, we utilize cross-entropy loss to optimize the next item prediction task, defined as:

$$\mathcal{L}_{rec} = -\log \frac{\exp(\mathbf{u} \cdot \mathbf{i}_{t+1})}{\sum_{\tilde{i} \in \mathcal{I}} \exp(\mathbf{u} \cdot \tilde{\mathbf{i}})}. \tag{8}$$

During evaluation, we rank the candidate items based on the inner product scores. It is noteworthy that all parameters of the EARec model remain frozen during the training process, allowing us to offline obtain multimodal representations for all items, thereby ensuring that the downstream recommendations achieve comparable efficiency to traditional recommendation methods.

4 Experiments

In this section, we first introduce the experimental setup, followed by presenting the experimental results and analyses.

4.1 Experiment Setting

4.1.1 Datasets. In Stage 1, we utilize the item modality information from five domains to perform the explainable generative alignment tasks on multiple MLLMs. Subsequently, in Stage 2, we apply the EARec composited by these MLLMs to derive multi-modal representations of items in downstream datasets, followed by performing the sequential recommendation. Specifically:

- **Stage 1 dataset**: We select five datasets from Amazon e-commerce dataset [5, 15] for the explainable generative alignment task in stage 1 of EARec, namely "Grocery and Gourmet

Table 1: Statistics of the datasets after preprocessing. "Avg. n" denotes the average length of item sequences.

Datasets	#Users	#Items	#Image	#Inters.	Avg. n
Stage 1	1,361,408	446,975	94,151	14,029,229	13.51
- Food	115,349	39,670	29,990	1,027,413	8.91
- CDs	94,010	64,439	21,166	1,118,563	12.64
- Kindle	138,436	98,111	0	2,204,596	15.93
- Movies	281,700	59.203	8,675	3,226,731	11.45
- Home	731,913	185,552	34,320	6,451,926	8.82
Stage 2					
- Office	87,436	25,986	16,628	684,837	7.84
- Arts	45,486	21,019	9,437	395,150	8.69
- Instruments	24,962	9,964	6,289	208,926	8.37
- Movielens	610	3,650	1,846	89,664	147.99

Food", "Home and Kitchen", "CDs and Vinyl", "Kindle Store", and "Movies and TV".

- **Stage 2 dataset**: For downstream recommendations, we select three additional datasets from Amazon to evaluate EARec's transfer recommendation performance across domains, namely "Office Products", "Arts, Crafts and Sewing" and "Musical Instruments". To evaluate the transfer performance on a new platform, we select a cross-platform dataset, i.e., Movielens[2].

For all datasets, following prior work [8, 20], we remove users and items with fewer than five interactions and organize the items according to the temporal order of user interactions. We consider three item modalities: Text, Vision, and Behavior. Notably, our method can conveniently extend to accommodate any new modalities. The statistics of the datasets are summarized in Table 1.

4.1.2 Baselines. **EARec** is compared with the following baselines:

- **SASRec** [10] employs a self-attention mechanism to aggregate item ID embeddings in user sequences without incorporating additional modality information.
- **SASRec$_\mathbf{T}$** is an extension of SASRec, utilizing item textual modality information to obtain item representations instead of ID embeddings.
- **UniSRec** [8] learns cross-domain universal sequence patterns through item textual modality representations and employs MoE to adaptively adjust item representations in different domains.
- **MoRec** [25] incorporates item text modality and performs end-to-end optimization on the modality encoder and recommendation model.
- **MISSRec** [20] is based on item textual and visual modality representations, employing a multi-modal interest-aware module and cross-attention mechanisms to learn multi-interest user sequence representations.

4.1.3 Evaluation Metrics. We utilize two widely used evaluation metrics, HR@K and NDCG@K, to assess the performance of the downstream recommendation tasks. K is set to 10 and 50. Following prior work [8, 10], we adopt a leave-one-out method to split the dataset. Specifically, for a user's interaction sequence, we use the last item for testing, the second-to-last item for validation, and the remaining items for training. We obtain the ranking list through the dot product scores between the user sequence representation and all items, and we report the average results over all users.

4.1.4 Implement Details. In Stage 1, we implement the explainable multi-modality generative alignment based on the transformer library [22] and the DeepSpeed library [3]. The backbone model is Vicuna-v1.5 [4]. We select three modalities input for alignment: item description, item image, and item embedding of recommendation model, with the alignment anchor set as the item title. The text encoder is the LLM's word embedding, the vision encoder is *clip-vit-large-patch14-336* [5], and the behavior encoder utilizes the pre-trained item embedding from SASRec. We load the parameters of LLaVA's LLM part and vision projector before training for the vision modality. For behavior modality, we perform a continuous alignment on in-domain behavior data to alleviate the gap between domains. During training, we employ LoRA to efficiently fine-tune MLLM, the hyperparameter r is set to 128, and α is set to 256. The learning rate for the LoRA parameters is set at 2×10^{-4}, while the learning rate for the projector is 2×10^{-5}. We conduct experiments on four NVIDIA RTX 3090 GPUs, with a global batch size of 16.

In Stage 2, we implement downstream recommendation tasks using the RecBole library [26]. During the model composition phase, the range of modality-adaptive adjustment weights is set to [0,1], ensuring that the sum of the weights equals 1. In the downstream recommendation phase, the number of layers and heads of the transformer encoder is set to 2. For all downstream recommendation experiments, we employ the Adam optimizer and carefully search for hyperparameters, with a batch size of 2048 and NDCG@10 as the evaluation metric, employing a patience of 10 for early stopping. We adjust the learning rate within the set {0.0003, 0.001, 0.003, 0.01} and the embedding dimension within {64, 128, 300}. The code is available at:https://anonymous.4open.science/r/EARec

4.2 Performance Comparison

We compare two variants of the proposed method, EARec$_{\text{TV}}$ and EARec$_{\text{TVB}}$, with several baseline methods. The difference between the EARec$_{\text{TV}}$ and EARec$_{\text{TVB}}$ is the input modality in item encoding. The overall experimental results are shown in Table 2. From the results, several observations can be made.

First, recommendation methods incorporating modal information generally outperform traditional methods, i.e., SASRec. This demonstrates that introducing modal information effectively enhances item representation and improves recommendation tasks. Second, although introducing more modal information and using a pairwise alignment approach to reduce gaps between modalities, MISSRec still underperforms MoRec, which only uses a single modality. It indicates that the pair-wise alignment does not integrate the modalities effectively and impairs the recommendation performance. Third, the benefits of transferable recommendation baseline models are not pronounced in cross-platform datasets, i.e.,

[2]https://grouplens.org/datasets/movielens/latest/

[3]https://github.com/microsoft/DeepSpeed

[4]https://huggingface.co/lmsys/vicuna-7b-v1.5

[5]https://huggingface.co/openai/clip-vit-large-patch14-336

Table 2: Downstream recommendation performance of different models. The best, second-best and third-best performances are denoted in bold, underlined, and wavy-line fonts, respectively. The subscript "T", "V", and "B" denote the Text, Vision, and Behavior modality used in the item encoding. The superscripts * and ** indicate $p \leq 0.05$ and $p \leq 0.01$ for the paired t-test of the best EARec variant vs. the best baseline.

Setting		Baselines					Ours		
Dataset	Metric	SASRec	SASRec_{T}	$\text{UniSRec}_{\text{T}}$	MoRec_{T}	$\text{MISSRec}_{\text{TV}}$	EARec_{TV}	$\text{EARec}_{\text{TVB}}$	Improv.
Office	HR@10	0.1064	0.1043	0.1046	0.1096	0.1038	0.1210	**0.1234****	12.59%
	HR@50	0.1641	0.1709	0.1751	0.1794	0.1701	0.1973	**0.1981****	10.42%
	NDCG@10	0.0710	0.0640	0.0627	0.0673	0.0666	0.0707	**0.0713**	0.42%
	NDCG@50	0.0835	0.0785	0.0780	0.0825	0.0808	0.0873	**0.0876****	4.91%
Arts	HR@10	0.1074	0.1078	0.1099	0.1101	0.1119	0.1224	**0.1244****	11.17%
	HR@50	0.1986	0.2050	0.2118	0.2127	0.2100	0.2329	**0.2330****	9.54%
	NDCG@10	0.0571	0.0613	0.0602	0.0637	0.0625	0.0664	**0.0671****	5.34%
	NDCG50	0.0769	0.0825	0.0823	0.0860	0.0836	0.0905	**0.0908****	5.58%
Instruments	HR@10	0.1126	0.1175	0.1087	0.1229	0.1201	0.1241	**0.1252****	1.87%
	HR@50	0.2087	0.2224	0.2079	0.2278	0.2218	0.2336	**0.2362****	3.69%
	NDCG@10	0.0618	0.0690	0.0622	0.0717	**0.0771**	0.0667	0.0727	-
	NDCG@50	0.0826	0.0917	0.0837	0.0944	**0.0988**	0.0909	0.0967	-
Movielens	HR@10	0.0967	0.0803	0.0721	0.0557	0.0885	0.0984	**0.1033****	6.83%
	HR@50	0.2852	0.2705	0.2705	0.2246	0.2361	**0.2984****	0.2852	4.63%
	NDCG@10	0.0419	0.0352	0.0308	0.0249	0.0393	0.0440	**0.0466****	11.22%
	NDCG@50	0.0826	0.0761	0.0740	0.0617	0.0703	**0.0868****	0.0858	5.08%

Table 3: Analysis on modality expansibility in Office dataset. "w/o" denotes removing the alignment of specific modality. The best and second-best performances are denoted in bold and underlined fonts, respectively. "Improvement" denotes the performance gain of EARec compared to "w/o All"

Datasets	Office		Movielens	
Metrics	HR@10	NDCG@10	HR@10	NDCG@10
EARec	**0.1234**	**0.0713**	**0.1033**	**0.0466**
- w/o Behavior	0.1212	0.0704	0.0984	0.0422
- w/o Vision	0.1204	0.0691	0.0967	0.0446
- w/o Text	0.1200	0.0692	0.0951	0.0422
- w/o All	0.1189	0.0699	0.1000	0.0428
Improvement	+3.78%	+2.00%	+3.30%	+8.88%

Movielens, suggesting the limitation of only performing transferable learning from the same platform.

The two variants of our method, EARec_{TV} and $\text{EARec}_{\text{TVB}}$, achieve the best overall performance. Specifically, EARec_{TV} outperforms MISSRec in most cases. This indicates that, compared to traditional pairwise alignment methods, our proposed explainable generative alignment method is more effective in incorporating modalities. Furthermore, $\text{EARec}_{\text{TVB}}$ achieves better performance than EARec_{TV} by utilizing the behavior modality for item representation, demonstrating the effectiveness of our method in expanding to new modalities and showcasing the potential to enhance recommendation performance through further incorporating modalities. The performance improvements on the Movielens dataset indicate that our method has learned more generalizable modality representations, leading to better transfer recommendation performance.

4.3 Analysis of Modality Expansibility

In this section, we analyze the modality expansibility of EARec. We evaluate the impact of incorporating a new modality and the overall effect of aligning multiple modalities on the final performance. Specifically, we compare EARec with four of its variants: (1) **w/o Behavior**, (2) **w/o Vision**, (3) **w/o Text**, and (4) **w/o All**, i.e., the original model before the integration of all three modalities. To ensure a capacity for modality handling, we directly adopt LLaVA for **w/o All** variants.

The comparison results are shown in Table 3, from which several observations can be drawn. First, the incorporation of each new modality improves the model's performance, with the Vision and Text modalities contributing relatively more to the improvements. Second, the original model, without the proposed explainable generative alignment across the three modalities, performs significantly worse, demonstrating the effectiveness of the proposed method.

4.4 Analysis of Explainable Alignment

In this section, we analyze the explainability of the proposed multi-modality alignment method. EARec aligns information from different modalities into a unified explainable space through generative alignment, allowing the alignment results to be compared with the anchor points to demonstrate the effectiveness of the alignment. Since we use item titles as anchor points, we evaluate the quality of the alignment by computing the RougeL score between the generated alignment results and the anchors. Additionally, we

Figure 4: The generative results of the EARec and LLaVA on various modality inputs. The key aspect of specific modality captured by EARec is highlighted with red font.

further analyze the correlation between the alignment results and downstream recommendation performance.

We illustrate the generative result of the alignment from modalities to anchor in Figure 4. By comparing the generative results of EARec and LLaVA, we find that EARec generates the response closer to the anchor text, as reflected by its higher RougeL score. More importantly, EARec captures item-specific characteristics unique to different modalities. In the visual modality, EARec provides text related to "Ink Cartridges" and "Pen Cases", demonstrating a deeper understanding of the vision features. In the text modality, EARec extracts finer-grained item categories from the description, i.e., "Fountain Pens", instead of the broader "Office Products" given by LLaVA. Most notably, for outputs of the behavior modality, EARec generates text of a complementary item, i.e., "Bottle of Black Ink". This indicates that EARec can capture the item relation knowledge, which is highly beneficial for recommendation.

Furthermore, we present the correlation between explainability and recommendation performance. As shown in Figure 5, we present the alignment performance for the three modalities and the corresponding downstream recommendation results. Several insights can be drawn from the figure. First, applying generative alignment to the vanilla model, i.e., LLaVA, significantly improves the RougeL scores of the generated results (e.g., from 0.3282 to 0.4337 for the Text modality), demonstrating the effectiveness of

Table 4: Modality Adaptable Adjustment weights for model composition in Office dataset. The best and second-best performances are denoted in bold and underlined, respectively.

Text	Vision	Behavior	HR@10	HR@50	NDCG@10	NDCG@50
33%	33%	33%	0.1187	0.1900	0.0674	0.0830
20%	40%	40%	0.1185	0.1908	0.0703	0.0861
15%	42.5%	42.5%	0.1192	0.1919	0.0703	0.0861
10%	45%	45%	0.1193	<u>0.1921</u>	<u>0.0722</u>	<u>0.0880</u>
5%	47.5%	47.5%	<u>0.1194</u>	0.1919	0.0721	0.0879
5%	5%	90%	**0.1234**	**0.1981**	**0.0713**	**0.0876**

the alignment task. Notably, before aligning the Behavior modality, LLaVA was entirely unable to understand this modality, resulting in a near-zero RougeL score. Second, when the models for the three modalities are composited (i.e., LLaVA+TVB), the composite model shows further improvements in RougeL scores for the text and vision modalities, reflecting the model composition effectively integrates the model's ability to understand the three modalities. Nevertheless, this trend does not appear in behavior modality. We speculate the reason is the overfitting of the LLaVA+B variant since the training data of behavior modality is relatively insufficient compared to the other two modalities. Third, the alignment results' RougeL scores are generally proportional to the recommendation performance. This demonstrates that evaluating the alignment results can help us select the most suitable model for downstream recommendation tasks, highlighting the explainability of our model.

4.5 Analysis of Modality Adaptable Adjustment

In this section, we analyze the impact of the proposed modality-adaptive adjustment method on model composition. We conducted experiments using the Office dataset, adjusting the weights of the parameters associated with the three modalities in the MLLM to modify the model's understanding of different modalities, and we compared the corresponding downstream recommendation performance. The experimental results are shown in Table 4.

From these results, we observe that as the weight of the behavior modality parameters increases, the model's performance steadily improves. This enhancement can be attributed to two primary reasons. First, since the understanding of the behavior modality is more complex than that of the text and vision modalities, increasing the weights of the parameters related to the behavior modality emphasizes the model's capability to comprehend it. Second, the behavior modality is specific to the recommendation task and contains more information beneficial for recommendations, making the prominence of this modality effective in improving the model' s recommendation performance.

5 Conclusion

In this paper, we propose EARec, a novel explainable generative multi-modality alignment method for transferable recommender systems. Addressing the limitations of conventional pairwise alignment strategies, EARec leverages a two-stage pipeline to unify the alignment of diverse modalities and enhance sequential recommendation. It aligns multiple modalities to a shared anchor

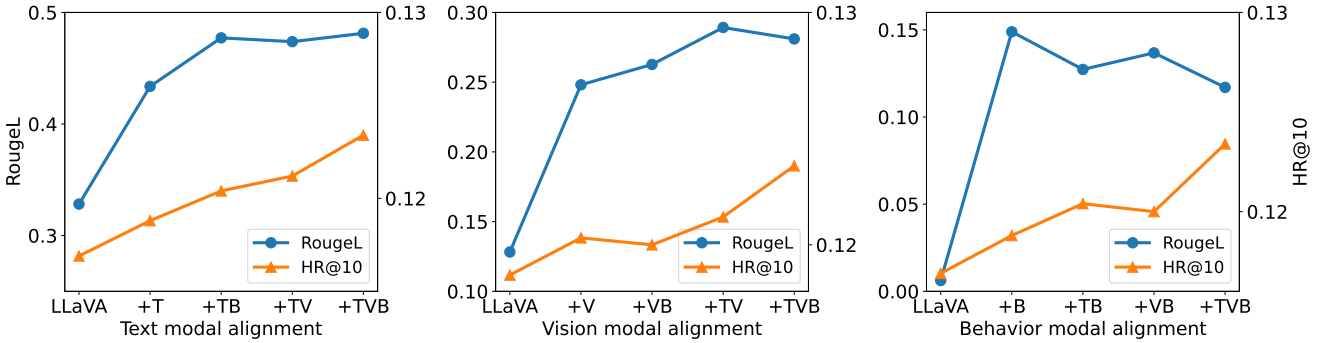

Figure 5: RougeL of generative alignment vs. Downstream recommendation task in Office dataset. "+T","+V", and "+B" denote performing an alignment on Text, Vision, and Behavior modalities, respectively.

with explainable meaning, ensuring consistent alignment across modalities and incorporating behavior-related information as an independent modality. In the second stage, we composite aligned modality encoders to enable effective transfer to the target domain for improved recommendation performance. Experimental results on multiple datasets demonstrate the effectiveness of EARec, and further analysis shows its high explainability and expansibility.

Acknowledgments

To Robert, for the bagels and explaining CMYK and color spaces.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems* 35 (2022), 23716–23736.

[2] Chi Chen, Yiyang Du, Zheng Fang, Ziyue Wang, Fuwen Luo, Peng Li, Ming Yan, Ji Zhang, Fei Huang, Maosong Sun, and Yang Liu. 2024. Model Composition for Multimodal Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 11246–11262. https://doi.org/10.18653/v1/2024.acl-long.606

[3] Hao Ding, Yifei Ma, Anoop Deoras, Yuyang Wang, and Hao Wang. 2021. Zero-shot recommender systems. *arXiv preprint arXiv:2105.08318* (2021).

[4] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15180–15190.

[5] Ruining He and Julian McAuley. 2016. Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. In *Proceedings of the 25th International Conference on World Wide Web (WWW '16)*. International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/2872427.2883037

[6] Ruining He and Julian McAuley. 2016. VBPR: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 30.

[7] Yupeng Hou, Zhankui He, Julian McAuley, and Wayne Xin Zhao. 2022. Learning Vector-Quantized Item Representation for Transferable Sequential Recommenders. *arXiv preprint arXiv:2210.12316* (2022).

[8] Yupeng Hou, Shanlei Mu, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. 2022. Towards Universal Sequence Representation Learning for Recommender Systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 585–593.

[9] 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.

[10] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*. IEEE, 197–206.

[11] Youhua Li, Hanwen Du, Yongxin Ni, Pengpeng Zhao, Qi Guo, Fajie Yuan, and Xiaofang Zhou. 2024. Multi-modality is all you need for transferable recommender systems. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 5008–5021.

[12] Yang Li, Qi'Ao Zhao, Chen Lin, Jinsong Su, and Zhilin Zhang. 2024. Who To Align With: Feedback-Oriented Multi-Modal Alignment in Recommendation Systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 667–676.

[13] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems* 36 (2024).

[14] Kang Liu, Feng Xue, Dan Guo, Peijie Sun, Shengsheng Qian, and Richang Hong. 2023. Multimodal graph contrastive learning for multimedia-based recommendation. *IEEE Transactions on Multimedia* 25 (2023), 9343–9355.

[15] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-based Recommendations on Styles and Substitutes. arXiv:1506.04757 [cs.CV]

[16] Artemis Panagopoulou, Le Xue, Ning Yu, Junnan Li, Dongxu Li, Shafiq Joty, Ran Xu, Silvio Savarese, Caiming Xiong, and Juan Carlos Niebles. 2023. X-instructblip: A framework for aligning x-modal instruction-aware representations to llms and emergent cross-modal reasoning. *arXiv preprint arXiv:2311.18799* (2023).

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[18] Xubin Ren, Wei Wei, Lianghao Xia, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. 2024. Representation learning with large language models for recommendation. In *Proceedings of the ACM on Web Conference 2024*. 3464–3475.

[19] Ajit P Singh and Geoffrey J Gordon. 2008. Relational learning via collective matrix factorization. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 650–658.

[20] Jinpeng Wang, Ziyun Zeng, Yunxiao Wang, Yuting Wang, Xingyu Lu, Tianxiang Li, Jun Yuan, Rui Zhang, Hai-Tao Zheng, and Shu-Tao Xia. 2023. MISSRec: Pre-training and transferring multi-modal interest-aware sequence representation for recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 6548–6557.

[21] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM international conference on multimedia*. 1437–1445.

[22] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*. 38–45.

[23] Chuhan Wu, Fangzhao Wu, Tao Qi, Jianxun Lian, Yongfeng Huang, and Xing Xie. 2020. Ptum: Pre-training user model from unlabeled user behaviors via self-supervision. *arXiv preprint arXiv:2010.01494* (2020).

[24] Penghang Yu, Zhiyi Tan, Guanming Lu, and Bing-Kun Bao. 2023. Multi-view graph convolutional network for multimedia recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 6576–6585.

[25] Zheng Yuan, Fajie Yuan, Yu Song, Youhua Li, Junchen Fu, Fei Yang, Yunzhu Pan, and Yongxin Ni. 2023. Where to Go Next for Recommender Systems? ID-vs.

Modality-based recommender models revisited. *arXiv preprint arXiv:2303.13835* (2023).

[26] Wayne Xin Zhao, Shanlei Mu, Yupeng Hou, Zihan Lin, Yushuo Chen, Xingyu Pan, Kaiyuan Li, Yujie Lu, Hui Wang, Changxin Tian, et al. 2021. Recbole: Towards a unified, comprehensive and efficient framework for recommendation algorithms. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 4653–4664.

[27] Zijia Zhao, Longteng Guo, Tongtian Yue, Sihan Chen, Shuai Shao, Xinxin Zhu, Zehuan Yuan, and Jing Liu. 2023. Chatbridge: Bridging modalities with large language model as a language catalyst. *arXiv preprint arXiv:2305.16103* (2023).

[28] Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. 2023. Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM Web Conference 2023*. 845–854.

[29] 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.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009