Information Fusion 101 (2024) 101989

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Prompt-based and weak-modality enhanced multimodal recommendation

Xue Dong [a], Xuemeng Song [b,*], Minghui Tian [b], Linmei Hu [c,*]

[a] School of Software, Shandong University, Jinan 250101, China
[b] School of Computer Science and Technology, Shandong University, Qingdao 266237, China
[c] School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China

ARTICLE INFO

Keywords:
Multimodal recommendation
Multimodal interest learning
Prompt learning

ABSTRACT

Beyond conventional recommendation systems that rely merely on user-item interaction data, multimodal recommendation systems additionally exploit the item multimodal data for boosting the recommendation performance. In this research line, late fusion-based approaches that first predict user ratings for each item modality independently and then merge these predictions for a final user rating have made significant advancements. Nevertheless, these methods still have the following two issues: (1) they utilize individual user embeddings to model user interest in different modalities, while overlooking the underlying relationship among modalities and significantly increasing the memory costs; and (2) they overlook the unreliable interest learned from certain modality, thus hindering the accurate final rating learning. To address these issues, we propose a prompt-based and weak-modality enhanced multimodal recommendation framework. It consists of two key components: (1) multimodal prompted user interest learning that adopts a single user embedding with different modality prompts to model different modality-specific user interests, and (2) weak-modality enhanced training that enhances the user interest learning in modalities where the predictions are less unreliable, ensuring well-balanced learning across all modalities. Extensive experiments on Amazon datasets have demonstrated the effectiveness of the proposed framework. The two components deployed onto existing methods help to make them more effective and efficient.

1. Introduction

Recommendation systems have become popular to help users discover their preferred online content. Traditional recommendation methods [1–4] aim to represent user interests and item properties with embeddings according to historical user-item interactions, and predict the user rating to an item by the similarity between their embeddings. Recently, considering that the multimodal information of items, e.g., images and texts, reflect the item properties and user interests from different perspectives [5–9], several researches resort to the multimodal recommendation.

Existing multimodal recommendation methods typically extract the modality features from the raw multimodal data of the item and then use the modality fusion result to supplement the user-item interaction data to predict the user rating to the item [10]. According to the timing of the modality fusion, existing methods could be roughly divided into early fusion-based [6,11,12] and late fusion-based [7,8,13] methods. The early fusion-based methods first fuse the modality features of one item into a single embedding and then based on it predict the final user rating to the item. Nevertheless, this strategy still treats the item as a whole, which cannot model the user's specific tastes on different modalities [7]. Differently, the late fusion-based methods first learn the user's rating towards each modality, i.e., the modality-specific user rating, and then predict the user final rating to the item by fusing all the modality-specific user ratings. The state-of-the-art multimodal recommendation models [9,14,15] follow the late fusion strategy. Despite their remarkable performance, existing late fusion-based multimodal recommendation methods still have the following issues:

- **Fail to efficiently model different modality-specific user interests.** Existing methods mainly utilize an individual user embedding to learn each modality-specific user interest, while overlooking the underlying relationships among multiple modality-specific user interests. Intuitively, as modalities describe the same item from different perspectives, there could be certain consistency among different modality-specific user interests. Despite several researches [16,17] have designed auxiliary techniques to model such consistency, their methods suffer from significant memory and computational overheads.
- **Fail to ensure that all the modality-specific user interests are well-learned.** Existing late-fusion based methods typically employ

* Corresponding authors.
E-mail addresses: dongxue.sdu@gmail.mail (X. Dong), sxmustc@gmail.com (X. Song), tianminghui99@gmail.com (M. Tian), hulinmei@bit.edu.cn (L. Hu).

https://doi.org/10.1016/j.inffus.2023.101989
Received 25 June 2023; Received in revised form 25 August 2023; Accepted 26 August 2023
Available online 1 September 2023

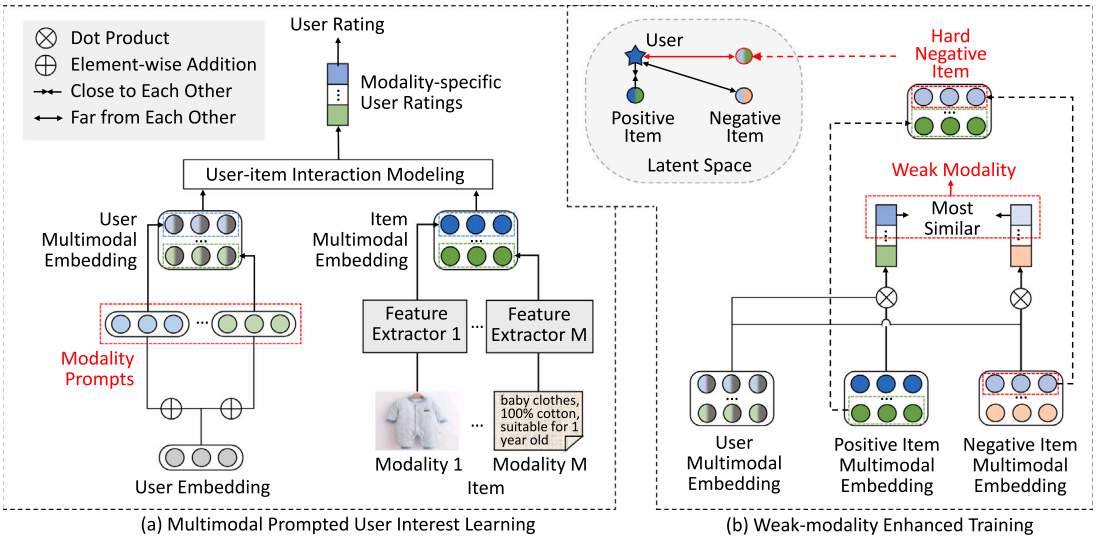

Fig. 1. The prompt-based and weak-modality enhanced multimodal recommendation framework. It consists of the multimodal prompted user interest learning (a) that adopts modality prompts to effectively and efficiently model modality-specific user interests, and weak-modality enhanced training (b) that enhances the user interest learning in the weak modality with a newly generated hard negative. The parts highlighted in red denote our key contributions.

the Bayesian personalized ranking mechanism for optimization. While this approach ensures that the fused rating of all modality-specific user ratings for the positive item surpasses that for the negative one, it does not necessarily guarantee optimal learning for each modality-specific user interest. As a result, there might exist certain modalities where user interests are inadequately captured, referred to as the 'weak modality'. This inadequacy can lead to unreliable modality-specific predictions, which in turn detrimentally influence the multimodal prediction [10].

To solve these issues, we propose a PrOmpt-based and Weak-modality Enhanced multimodal Recommendation framework, namely, POWERec, shown in Fig. 1. It consists of two components: (1) the multimodal prompted user interest learning that adopts a single user embedding with different modality prompts to model different modality-specific user interests, and (2) the weak-modality enhanced training that enhances the user interest learning in the modality where the prediction is unreliable, making all the modality-specific user interests have been well-learned.

In particular, as shown in Fig. 1(a), in the first component, as same as existing methods, we extract the modality features for each item based on its multimodal data and concatenate all modality features as the item multimodal embedding. Different from previous work, for each user, we resort to a single user embedding and a few modality prompts to construct the user multimodal embeddings to model the modality-specific user interests. Intuitively, the different modality-specific user interests can be linked by the shared user embedding. Then we adopt certain user-item interaction modeling method to predict the modality-specific user ratings to the item based on the user and item multimodal embeddings, and derive the final user rating based on all modality-specific user ratings. In the second component, in addition to the original recommendation objective that enforces the final user rating to a positive item to be higher than a negative one, we further propose a weak-modality regularization, which enhances the user interest learning in the weak modality. To be more specific, as shown in Fig. 1(b), we first calculate the difference between the modality-specific user ratings to positive and negative items, and regard the modality with the smallest difference as the weak modality. We then generate a hard negative item, which is different from the positive item merely regarding the weak modality. Ultimately, the weak-modality regularization is used to enforce the final user rating to the positive item to be higher than the generated hard negative item, enhancing the user interest learning in the weak modality. This manner ensures that all the modality-specific user interests are well-learned. We conduct extensive experiments on two public datasets in Amazon and the comparison results have proven the effectiveness of the proposed framework.

Main contributions can be summarized as follows:

- We propose a multimodal prompted user interest learning method that can effectively and efficiently model the modality-specific user interests and improve the recommendation performance.
- We propose a weak-modality regularization, which enhances the user interest learning in the weak modality by punishing the user rating to the hard negative items.
- Extensive experiments have demonstrated the effectiveness of POWERec. Moreover, the two components can be easily deployed onto existing methods to improve their recommendation performance and reduce the memory costs. Our implementations are available in https://github.com/hello-dx/POWERec.

2. Related work

2.1. Multimodal recommendation

Traditional recommendation [1–3] only utilizes the historical user-item interaction data to learn the user and item embeddings that represent the user interests and item properties, respectively. Considering that the item multimodal information reflects the item properties from different perspectives, some researches resort to the multimodal recommendation [6–8,12,18,19] to enrich the user and item embedding learning. In particular, several researches [12,18] leverage the item multimodal information only as supervision to better learned the item embedding. For example, LATTICE [12] adopts the item multimodal content to construct the latent relations between items to supplement the user-item interactions. Other researches [6,7,20–22] explicitly extract the modality features from the item multimodal data as the side information to the item embedding. For example, VBPR [6] is the first attempt to leverage the item image into the recommendation, which concatenates the original item ID embedding and item visual features as the new item embedding. MMGCN [7] introduces to construct the user-item graph for each modality to learn the different modality-specific user interests. DualGNN [8] learns the user's different preferences on different modalities to help better aggregate different modality-specific user ratings to the final user rating.

Despite their effectiveness, existing methods mainly assign one user embedding for each modality to learn the modality-specific user interest, respectively, which significantly increases the memory costs, yet still not achieving as much performance gain. Moreover, they fail to ensure that the modality-specific user interests in all modalities are well-learned, which might achieve the sub-optimal performance. To settle these issues, in this work, we propose multimodal recommendation framework with the multimodal prompt and weak-modality enhancement.

2.2. Prompt learning

The prompt learning is firstly proposed to overcome the gap between the pre-training and fine-tuning [23,24]. It adds prompts into the features learned in pre-training models to improve the performance of the downstream tasks without the fine-tuning step. Early approaches [25,26] mostly incorporate manually generated discrete prompts to guide the model. Later, since the manually generated prompts are both time-consuming and trivial, other researches [27–29] turn to automatically search discrete prompts for specific tasks. Nevertheless, these methods largely depend on the quality of the generated prompts. Some recent approaches have begun to utilize the continuous learnable embeddings as the prompts [30–32] achieving the state-of-the-art performance. Inspired by them, in this paper, we add different modality prompts to the user embedding to model different modality-specific user interests. In this manner, different modality-specific user interests can be connected with a shared user embedding. Our approach is the first attempt to introduce the prompt learning into the multimodal recommendation to effectively and efficiently model the modality-specific user interests.

3. Methodology

In this section, we first brief the problem definition of the multimodal recommendation in Section 3.1. We then detail the proposed POWERec in Section 3.2.

3.1. Problem definition

Suppose that there is a set of users $\mathcal{U}$, a set of items $\mathcal{I}$, and a set of item modalities $\mathcal{M}$. Each user $u \in \mathcal{U}$ is associated with a set of his/her historical interaction items $\mathcal{I}_u$. Each item $i \in \mathcal{I}$ is associated with its multimodal features $\{\mathbf{f}_i^m | m = 1, \dots, |\mathcal{M}|\}$, where $\mathbf{f}_i^m$ denotes the item feature of the mth modality. The goal of the multimodal recommendation is to predict the user rating to an item according to the user interactions as well as the item multimodal features. Ultimately, the candidate items can be ranked according to the user ratings and the top items are recommended to the user.

3.2. Prompt-based and weak-modality enhanced multimodal recommendation framework

In this subsection, we elaborate the proposed prompt-based and weak-modality enhanced multimodal recommendation framework, termed as POWERec. It consists of two key components: (1) multimodal prompted user interest learning that adopts the modality prompts to efficiently and effectively model the modality-specific user interest to predict the user rating to an item, and (2) weak-modality enhanced training that enhances the user interest learning in the weak modality to ensure all the modality-specific user interests are well-learned.

3.2.1. Multimodal prompted user interest learning

In the rest of this subsection, we first detail how we derive the user and item multimodal embeddings to model the user interests and item properties, respectively, and then introduce the user-item interaction modeling, which predicts the user rating to the item based on the user and item multimodal embeddings.

Item Multimodal Embedding. As same as existing methods [7,8,33], we represent each item with a multimodal embedding derived by its multimodal content information. To map different modality features into a same latent space, we adopt the multi-layer perceptron to encode each modality feature $\mathbf{f}_i^m$ of the item i into a d-dimensional embedding $\mathbf{e}_i^m$ as follows,

$$\mathbf{e}_i^m = \underbrace{\cdots \mathrm{Tanh}(\mathbf{W}_2^m(\mathrm{Tanh}(\mathbf{W}_1^m \mathbf{f}_i^m + \mathbf{b}_1^m)) + \mathbf{b}_2^m) \cdots}_{L}, \tag{1}$$

where $\mathbf{e}_i^m \in \mathbb{R}^d$ is the embedding of the item i in the mth modality. $\mathbf{W}_l^m$ and $\mathbf{b}_l^m$ are the trainable parameters of the l-layer in the multi-layer perceptron and L is total number of layers. Tanh is the non-linear activation function.

Accordingly, concatenating all the modality embeddings, we define the item multimodal embedding $\mathbf{e}_i \in \mathbb{R}^{d \times |\mathcal{M}|}$ as follows,

$$\mathbf{e}_i = [\mathbf{e}_i^1, \dots, \mathbf{e}_i^{|\mathcal{M}|}]. \tag{2}$$

User Multimodal Embedding. As aforementioned, previous methods usually allocate multiple modality-specific embeddings for each user. This approach becomes memory-intensive with a growing user base. In contrast, drawing inspiration from prompt tuning [30,31], we perceive the learning of modality-specific user interests as various downstream tasks for the final user interest learning, and encapsulate these interests for each user by a unified base user embedding as well as a few modality-specific prompts. In particular, we first represent each user u with a single embedding $\mathbf{f}_u \in \mathbb{R}^d$ as the basic user embedding. We then introduce a few modality-specific prompt embeddings for each modality to adapt the basic user embedding to different modalities. Specifically, let $\mathbf{P}_m = [\mathbf{p}_m^1, \dots, \mathbf{p}_m^Q]$ denote the to-be-learned prompt embeddings for the mth modality, where Q is the total number of prompt embeddings for each modality. Thereafter, the basic user embedding augmented by the prompt embeddings $\mathbf{P}_m$ is expected to capture the user interest in the mth modality. Intuitively, all modality-specific interests of one user are linked by the shared basic user embedding. It is worth noting that the number of modalities is much smaller than the number of users, and thus introducing multiple prompt embeddings for each modality saves much memory compared to existing methods that assign multiple embeddings for each user.

Formally, we define the overall user multimodal embedding $\mathbf{e}_u \in \mathbb{R}^{d \times |\mathcal{M}|}$ as follows,

$$\begin{cases} \mathbf{e}_u = [\mathbf{e}_u^1, \dots, \mathbf{e}_u^{|\mathcal{M}|}], \\ \mathbf{e}_u^m = \mathbf{f}_u + \mathrm{sum}(\mathbf{P}_m), m = 1, \dots, |\mathcal{M}|, \end{cases} \tag{3}$$

where $\mathbf{e}_u^m$ is the augmented user embedding for the mth modality. $\mathrm{sum}(\mathbf{P}_m)$ refers to the element-wise summation of all N_p embeddings in $\mathbf{P}_m$. The summation operation empirically performs best among other optional operations, such as the mean pooling and max pooling.

User-item Interaction Modeling. In this module, we aim to learn the user rating to an item based on the item and user multimodal embeddings. Specifically, following previous work [7–9], we adopt the late fusion strategy. Since this part is not the focus of our work, this module can be implemented by any state-of-the-art user-item interaction modeling methods, such as BPR and LightGCN. Mathematically, this module can be formulated as follows,

$$\begin{cases} y_{u,i} = \sum_m y_{u,i}^m, \\ y_{u,i}^m \leftarrow \mathcal{F}(\mathbf{e}_u^m, \mathbf{e}_i^m), \end{cases} \tag{4}$$

where $\mathcal{F}$ is the user-item interaction modeling method. $y_{u,i}$ is the final rating of the user u to the item i, and $y^m_{u,i}$ is the modality-specific rating of the user u to the mth modality of the item i.

3.2.2. Weak-modality enhanced training

The weak-modality enhanced training consists of two training objectives: (1) the commonly used recommendation objective, i.e., BPR loss, which enforces the final user rating to a positive item to be higher than that to a negative one, and (2) the weak-modality regularization that enhances the user-item interaction modeling by regulating the user rating learning to the weak modality where the user modality-specific interest is not well-learned.

BPR Loss. Following most recommendation methods, we adopt the pair-wise BPR loss as the key objective. According to Bayesian personalized ranking mechanism [1], we first build the training set $\mathcal{D} = \{(u, i, k)\}$, where $i \in \mathcal{I}_u$ is a positive (*i.e.,* interacted) item for the user u, and $k \in \mathcal{I} \setminus \mathcal{I}_u$ is a negative item that randomly sampled from items that the user has not interacted. The training triplet (u, i, k) indicates that the user u prefers item i compared to the item k. Then the recommendation objective can be defined as follows,

$$\mathcal{L}_{bpr} = -\sum_{(u,i,k)\in\mathcal{D}} \log(\text{sigmoid}(y_{u,i} - y_{u,k})), \tag{5}$$

where $y_{u,i}$ and $y_{u,k}$ are the ratings of the user u to the positive item i and negative item k, respectively. By minimizing the above objective function, the items that the user might be interested in will be predicted a higher rating than other items.

Weak-modality Regularization. Notably, the BPR loss can only supervise the final user rating learning, but cannot optimize each modality-specific rating learning, i.e., each modality-specific interest learning. Intuitively, there could be some modality-specific interest that is not well-learned, and hence make the learned corresponding final user rating unreliable. To alleviate this issue, we design the weak-modality regularization, which first distinguishes the weak modality that the user interest is not well-learned, and then enhances the corresponding user rating learning in the weak modality.

- *Weak Modality Identification.* Intuitively, given a training triplet (u, i, k), if the difference between the user ratings to the positive item i and negative item k in terms of the mth modality is small, it indicates that the model can hardly distinguish the positive item from the negative one based on the mth modality, i.e., the mth modality-specific user interest is not well-learned and should be regarded as the weak modality. Thus, for each training triplet (u, i, k), we calculate the differences between the modality-specific user ratings to the positive item and negative item in all modalities, and then define the modality with the smallest difference as the weak modality. Formally, the weak modality for each training triplet (u, i, k) can be identified as follows,

$$\begin{cases} m^*_{u,ik} = \text{argmin}([d^1_{u,ik}, \dots, d^{|\mathcal{M}|}_{u,ik}]), \\ d^m_{u,ik} = y^m_{u,i} - y^m_{u,k}, m = 1, 2, \dots, |M|, \end{cases} \tag{6}$$

  where $m^*_{u,ik}$ is the index of the weak modality for the triplet. $y^m_{u,i}$ and $y^m_{u,k}$ refer to the user ratings to the positive item i and negative item k in terms of the mth modality, respectively.
- *Weak Modality Enhancement.* To enhance the user interest learning in the specific weak modality, we introduce the hard negative item k^* for each triplet (u, i, k), which is different from the given positive item i only regarding the weak modality. Accordingly, we represent the hard negative item as follows,

$$\mathbf{e}_{k^*} = [\mathbf{e}^1_i, \dots, \mathbf{e}^{m^*}_k, \dots, \mathbf{e}^{|\mathcal{M}|}_i], \tag{7}$$

  where $\mathbf{e}_{k^*}$ is the multimodal embedding of the hard negative item for the triplet (u, i, k). Thereafter, to punish the wrongly ranked hard negatives, we define the weak-modality regularization as follows,

$$\mathcal{L}_{weak} = -\sum_{(u,i,k^*)\in\mathcal{D}} \log(\text{sigmoid}(y_{u,i} - y^*_{u,k})), \tag{8}$$

  where $y_{u,i}$ and y_{u,k^*} are the ratings of the user u to the positive item i and the generated hard negative item, respectively. As the hard negative item is different from the positive item merely in the weak modality, enforcing the user rating to the positive item to be higher than that to the hard negative item could promote the user interest learning regarding the weak modality. It is worth mentioning that with the iterative optimization for all triplets, all the weakly learned user interests can be promoted.

Ultimately, the overall objective is defined as follows,

$$\mathcal{L} = \min_{\Theta}(\mathcal{L}_{bpr} + \alpha\mathcal{L}_{weak} + \beta\|\Theta\|^2), \tag{9}$$

where $\mathcal{L}_{bpr}$ is the BPR-based recommendation objective defined in Eq. (5), $\mathcal{L}_{weak}$ is the weak-modality regularization defined in Eq. (8). α is the trade-off parameter to adjust the recommendation objective and weak-modality regularization. Θ is the set of model parameters. The last term is to avoid the over-fitting.

4. Experiments

In this section, we conducted extensive experiments to answer the following research questions:

- RQ1: How effective is the proposed POWERec?
- RQ2: How effective are two components of POWERec?
- RQ3: How do hyper-parameters affect the performance?
- RQ4: Can the weak-modality enhanced training boost the user interest learning in weak modality?

4.1. Experimental settings

In this subsection, we give the experimental settings, including the datasets, evaluation protocols, and implementation details.

Datasets. To evaluate the proposed POWERec in the task of the top-N item recommendation, we utilized the most commonly used Amazon dataset [34] and conducted extensive experiments on its two categories: Baby and Clothing. To keep the fair comparison, we closely followed the pre-processing of the dataset in the study [10], where we only kept the users and items that have more than 5 interactions. The statistics of the two datasets after pre-processing are listed in Table 1.

Table 1
Statistics of datasets.

Dataset	#User	#Item	#Interaction	Sparsity
Baby	19,445	7,050	160,792	99.88%
Clothing	39,387	23,033	278,677	99.97%

Evaluation Protocols. We randomly split each dataset with a ratio of 8 : 1 : 1 to derive the corresponding training, validation, and testing sets following the study [10]. To evaluate the effectiveness of our proposed framework in the top-N item recommendation, we adopted the following two widely-used metrics: Recall@N and NDCG@N. By default, we set $N = 10, 20$.

Implementation Details. Following most previous studies [7,8,13], we considered the item's ID embedding, visual feature, and textual feature as three modalities. We directly adopted the visual and textual features released by [10], which are a 4096-dimensional vector extracted by a CNN model [12] and a 384-dimensional vector extracted by a pre-trained sentence-transformers [35], respectively. We set the embedding size d to 64 following most studies [6,9,36]. The layer number of the multi-layer perceptron in Eq. (1) is set to 1. We adopted the state-of-the-art method LayerGCN [3] as the user-item interaction modeling method $\mathcal{F}$ in Eq. (4). We adopted the gird search to tune the hyper-parameters. In particular, we tuned the number of prompt embeddings Q in $\mathbf{P}_m$ from $[1, 2, 3, 4, 5]$, and the trade-off parameters in Eq. (9), i.e., α and β, from $[0, 0.001, 0.01, 0.1, 1]$. The best values of Q for

Table 2
Performance of the performance comparison in terms of Recall and NDCG on the two datasets. We highlight the best and second-best results in bold and underlined, respectively. We annotate the relative improvement of the proposed POWERec compared to the second-best results. Note that the results of the baselines are from the study [10].

Dataset	Method	Recall@N↑		NDCG@N↑		Parameter↓
		$N = 10$	$N = 20$	$N = 10$	$N = 20$	(MB)
Baby	VBPR	0.0423	0.0663	0.0223	0.0284	3.2
	MMGCN	0.0378	0.0615	0.0200	0.0261	3.9
	GRCN	0.0539	**0.0833**	0.0288	0.0363	4.5
	SLMRec	0.0529	0.0775	0.0290	0.0353	2.0
	DualGNN	0.0448	0.0716	0.0240	0.0309	3.7
	POWERec	**0.0545**	0.0823	**0.0299**	**0.0370**	2.0
Relative Improvement		1.1%	–	3.1%	1.9%	–
Clothing	VBPR	0.0281	0.0415	0.0158	0.0192	6.8
	MMGCN	0.0218	0.0345	0.0110	0.0142	9.0
	GRCN	0.0424	0.0662	0.0223	0.0283	9.4
	SLMRec	0.0442	0.0659	0.0241	0.0296	4.5
	DualGNN	0.0454	0.0683	0.0241	0.0299	6.4
	POWERec	**0.0462**	**0.0691**	**0.0245**	**0.0303**	4.5
Relative Improvement		1.8%	1.2%	1.7%	1.3%	–

Baby and Clothing datasets are 1 and 3, respectively. The best values of α and β are 0.01 and 1, respectively, for both datasets.

We optimized the proposed POWERec with Adam optimizer [37] and used the learning rate of $1e^{-3}$. We set the mini-batch sizes to 2048 for all the datasets, and trained the model with 1000 epochs and the early stop strategy. The results are selected according to NDCG@20 on the validation set.

4.2. Performance comparison (RQ1)

In order to evaluate the effectiveness of the proposed POWERec, we selected the following multimodal recommendation methods,

- **VBPR** [6] is the first attempt that considers the visual features in the recommendation. In our context, for the fair comparison, we extended VBPR by further involving the item textual feature to learn the user interests.
- **MMGCN** [7] constructs a user-item interaction graph for each modality and utilizes the graph convolutional techniques to learn the user interests for each item modality. Ultimately, the user and item embeddings used to predict the final user rating are derived by aggregating all the modality-specific embeddings.
- **GRCN** [13] constructs a user-item interaction graph for each modality similar to MMGCN. Differently, it identifies the false-positive edges in the graph and cuts off such edges to refine the user-item interaction graph.
- **SLMRec** [9] utilizes the self-supervised learning techniques in the graph-based models to uncover the hidden signals from the data with contrastive loss.
- **DualGNN** [8] utilizes the modality features to learn the different modality-specific user interests, and explicitly models the user different attentions over different modalities to achieve a better recommendation.

The results of the baselines and POWERec on two datasets are shown in Table 2, where the best and second-best results are highlighted in bold and underlined, respectively. For a comprehensive comparison, we also listed the parameters of each method in the "Parameter" column. From the results in Table 2, we have the following observations.

(1) POWERec achieves the best performance compared with existing multimodal recommendation methods on most metrics. There might be two reasons. (1) Existing methods mainly utilize different user embeddings to model modality-specific user interests independently, while overlooking the underlying relations among different modality-specific user interests. Differently, we leverage a shared basic user embedding with different modality prompts to model the modality-specific user interests, which can help to link different modality-specific user interests. And (2) we enhance the user interest learning in the weak modality by the weak-modality enhanced training, which helps alleviate the bad effect of the weak-modality on the final user interest learning and hence promote the recommendation performance.

(2) POWERec not only achieves better performance than most existing methods, but also has fewer parameters. This proves that the proposed POWERec is an effective and lightweight solution for the multimodal recommendation. This is because that we leverage a shared single user embedding with different modality prompts to model different modality-specific user interests, instead of assigning multiple user embeddings like existing methods. Overall, the effective and lightweight POWERec has greater practical application potential, especially in scenarios with tremendous users.

(3) SLMRec has comparable parameters with our POWERec, but performs worse than ours. This is because that SLMRec utilizes a single user embedding to model different modality-specific user interests, while failing to distinguish the user's specific interests in different modalities and hence only achieving the sub-optimal performance. In contrast, POWERec introduces different modality prompts to adapt the single user embedding to different modalities, which can capture the user's different specific interests in different modalities.

4.3. Ablation study (RQ2)

To further evaluate the effectiveness of each component in the proposed POWERec: the multimodal PrOmpted user interest learning (PO) and Weak-modality Enhanced training (WE), we design the following variants.

- **PORec**. We removed the WE component in POWERec. In particular, PORec predicts the user rating to an item as same as POWERec, but is optimized with the original recommendation objective defined in Eq. (5).
- **BaseRec**. We removed both the PO and WE components in POWERec. In particular, we left out the modality prompts in the user multimodal embedding and only utilized the basic user embedding to learn the different modality-specific user interests. Meanwhile, BaseRec is also optimized with only the original recommendation objective defined in Eq. (5).

Table 3
Impact of the proposed multimodal prompted user interest learning (MP), and weak-modality enhanced training (WE). We highlight the best results in bold of each group of methods.

Dataset	Method	Recall@N		NDCG@N		Parameter↓
		$N = 10$	$N = 20$	$N = 10$	$N = 20$	(MB)
Baby	BaseRec	0.0500	0.0802	0.0266	0.0344	2.0
	PORec	0.0530	0.0820	0.0288	0.0362	2.0
	POWERec	**0.0545**	**0.0823**	**0.0299**	**0.0370**	2.0
	DualGNN	0.0448	0.0716	0.0240	0.0309	3.7
	PODualGNN	0.0550	0.0871	0.0297	0.0379	2.5
	POWEDualGNN	**0.0579**	**0.0918**	**0.0311**	**0.0398**	2.5
Clothing	BaseRec	0.0387	0.0582	0.0209	0.0259	4.5
	PORec	0.0451	0.0683	0.0236	0.0295	4.5
	POWERec	**0.0462**	**0.0691**	**0.0245**	**0.0303**	4.5
	DualGNN	0.0454	0.0683	0.0241	0.0299	6.4
	PODualGNN	0.0468	0.0711	0.0257	0.0318	3.8
	POWEDualGNN	**0.0503**	**0.0753**	**0.0272**	**0.0336**	3.8

Besides, the two key components can also be easily deployed into existing multimodal recommendation methods. Without losing generality, we deployed the PO and WE components into the best baseline, i.e., DualGNN, and designed the following extensions.

- **PODualGNN**. We added the PO component into DualGNN. In particular, we replaced the multiple user embeddings in DualGNN with our defined user multimodal embedding, i.e., a single user embedding with different modality prompts. Then, we predict the user rating to an item according to the original DualGNN. PODualGNN is optimized with only the recommendation objective.
- **POWEDualGNN**. We added both the PO and WE components into DualGNN. In particular, we replaced the multiple user embeddings in DualGNN with the proposed user multimodal embedding and utilized the proposed weak-modality enhanced training to optimize the POWEDualGNN, which aims to enhance the user interest learning in the weak modality.

The ablation study results on Baby and Clothing datasets are shown in Table 3, where the best performance is highlighted in bold. We have the following observations.

(1) PORec that utilizes the modality prompts to adapt the basic user embedding to different modality-specific user interests outperforms BaseRec that only utilizes the basic user embedding. This verifies the benefit of introducing the modality prompts to capture the user's specific interest in different modalities.
(2) PODualGNN not only achieves the better performance than DualGNN but also has few parameters. This might be because that DualGNN assigns multiple user embeddings for each user to model the different modality-specific user interests, which makes the user's modality-specific interests learned independently and increases the number of parameters.
(3) Additionally equipped with the WE component, POWERec and POWEDualGNN outperform PORec and PODualGNN on both datasets, respectively. This indicates that there does exist the modality in which the user interest is not well-learned and our proposed weak-modality enhanced training component could enhance the user interest learning in the weak modality, whereby ensuring that all the modality-specific user interests are well-learned and achieving the better performance.
(4) The successful deployments of the two components on DualGNN reflects their great potential in boosting existing multimodal recommendation models' performance with fewer parameters.

4.4. Hyper-parameter discussion (RQ3)

In this subsection, we evaluated the following two key hyper-parameters: the trade-off parameter α defined in Eq. (9), and the number of the prompt embeddings Q defined in Eq. (3).

Trade-off Parameter α adjusts the weight between the original recommendation objective and weak-modality regularization, where a larger α indicates the higher contribution of the weak-modality regularization in optimization. Specifically, we fixed the other hyper-parameter Q to the most suitable value for each dataset, i.e., 3 and 1 for Baby and Clothing datasets, respectively, and tuned α from $\{0, 0.001, 0.01, 0.1, 1\}$. Without losing generality, we showed Recall@10 and NDCG@10 results of POWERec on the two datasets in Fig. 2(a) and (b). As can be seen, along with the trade-off parameter α increasing, the performance first rises significantly until it achieves the best performance with the most suitable $\alpha = 0.01$ on both datasets, and then decreases. This demonstrates the necessity of appropriately adding the weak-modality regularization to enhance the user interest learning in the weak modality. However, when α becomes excessively large, it will make the model focused on distinguishing the hard negative items, while overlooking the original negative items, which makes the model failed to learn the overall user interest and hence destroys the recommendation performance.

Number of Prompt Embeddings Q controls the number of the prompt embeddings, which encode the user's specific interests for different modalities. Intuitively, the larger the Q, the more specific interest should be learned for different modalities. Specifically, we fixed the trade-off parameter α to the most suitable value 0.01 for both datasets, and tuned the number of the prompt embeddings Q from $\{1, 2, 3, 4, 5\}$. We showed the Recall@10 and NDCG@10 of POWERec on the two datasets in Fig. 2(c) and (d). From the results, we can see that the performance first increases along with the number Q increasing, and then decreases in Baby dataset, while the performance in Clothing dataset keeps decreasing. This might be because that the difference between different modality-specific user interests in Baby dataset is larger than that in Clothing dataset. For example, in Baby dataset, the item's text description can indicate the item's suitable age of babies, which can be hard to reflect from the item's image. Accordingly, the user's interest learned by the textual modality could be very different from that by the visual modality. Therefore, in Baby dataset, it needs more prompt embeddings to capture such differences among different modality-specific user interests.

4.5. Visualization (RQ4)

In this subsection, we visually demonstrated the effectiveness of the proposed weak-modality enhanced training from the following two perspectives. On the one hand, as aforementioned, the difference between the modality-specific user ratings to the positive and negative items indicates whether the modality-specific user interest is well-learned. Therefore, we investigated the modality-specific user rating difference between positive and negative items to show whether there exists the weak modality. On the other hand, when the modality-specific user interest is not well-learned, the model should hardly distinguish the

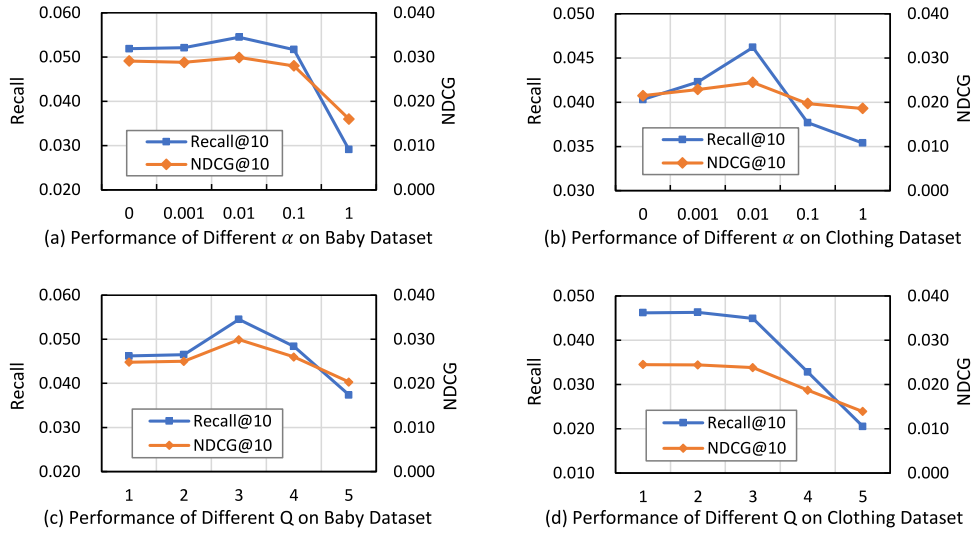

Fig. 2. Performance of the proposed POWERec with respect to different hyper-parameters on both datasets.

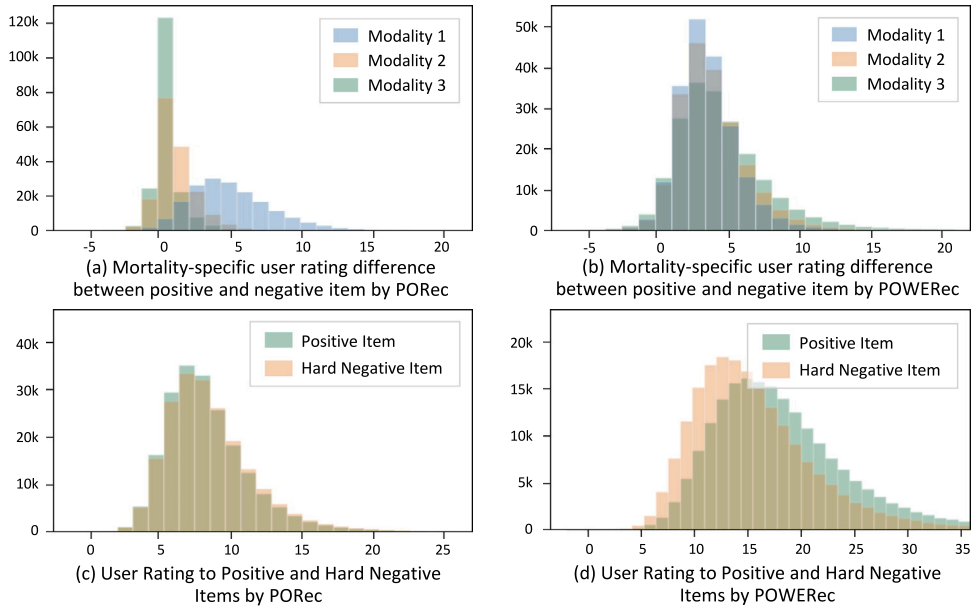

Fig. 3. Visualization regarding the weak-modality enhanced training component. (a) and (b) show the statistics of the modality-specific user rating difference between positive and negative items derived by PORec and POWERec, respectively. The three modalities are the item's ID embedding, visual feature, and textual feature, respectively. (c) and (d) show the user ratings to positive and hard negative items by PORec and POWERec, respectively.

positive item and the generated hard negative item. Therefore, we investigated the user ratings to the positive item and the generated hard negative item.

On modality-specific user rating difference between positive and negative items. Without losing generality, we visualized the statistics of the differences between the modality-specific user ratings to the positive and negative items predicted by our POWERec and its variant PORec that does not contain the weak-modality enhanced training component on Clothing dataset. The results are shown in the form of histogram in Fig. 3(a) and (b). In particular, the x-axis refers to the difference between the modality-specific user ratings to positive and negative items, while the y-axis refers to the total number of the training triplets within the corresponding bin. The three modalities are the item's ID embedding, visual feature, and textual feature, respectively. As can be seen from the results of PORec in Fig. 3(a), the differences between the user ratings to positive and negative items in the modality 1 are large, which indicates that the user interest in this modality is well-learned. This might be because that the modality 1 refers to the item ID embedding, which is a randomly initialized vector similar to the user embedding. Thus, the user-item interactions in this modality are easier to grasp. However, the differences between the user ratings to positive and negative items in the modalities 2 and 3 are concentrated near 0, suggesting that the user interests are not been well-learned in these modalities. This might be because that the modalities 2 and 3, i.e., the item visual and textual features, have larger gaps with the user embedding, whereby the interactions are harder to learn. Differently, from the results of POWERec in Fig. 3(b), we can see that the differences between the user ratings to positive and negative items in all three modalities are almost larger than 0. This implies that with the help of the weak-modality promoted training, POWERec is able to enhance the user interest learning in the weak modality and hence promote that all modality-specific user interests can be well-learned.

On the user ratings over positive and hard negative items. We visualized the statistics of the user ratings to the positive item and the hard negative item predicted by PORec and POWERec on Clothing dataset in Fig. 3(c) and (d), respectively. As can be seen from the results of PORec in Fig. 3(c), overall, the user ratings to the positive item and the generated hard negative item are almost the same, which indicates that PORec can hardly distinguish the positive and hard negative items. This confirms again that there exists the weak modality. In contrast, as shown in Fig. 3(d), with the weak-modality enhanced training, POWERec can distinguish the positive item and hard negative item for the user to some extent. This proves that the proposed weak-modality enhanced training can help correct the wrongly ranked hard negatives to enhance the user interest learning in the weak modality.

5. Conclusion

In this paper, we propose a prompt-based and weak-modality enhanced multimodal recommendation framework, termed as POWERec. Different from existing methods, we propose to effectively and efficiently model modality-specific user interests with a single shared basic user embedding and different modality prompts. In addition, to avoid the negative effect caused by the weak modality, we design the weak-modality regularization to enhance the user interest learning in the weak modality. Extensive experiments on two public datasets in Amazon have demonstrated the effectiveness of the proposed POWERec. Besides, POWERec can be easily deployed into existing multimodal recommendation methods, which not only improves their recommendation performance but also reduces the memory costs. Nevertheless, the current POWERec assumes that the generated hard negative items are truly negative, while failing to consider the noise when the hard negative item is false negative. Therefore, in the future, we plan to distinguish whether the generated hard negative item is truly negative to alleviate the negative impact caused by the false negatives, which could further improve the recommendation performance.

CRediT authorship contribution statement

Xue Dong: Conceptualization, Methodology, Validation, Writing. **Xuemeng Song:** Writing – review & editing, Funding acquisition. **Minghui Tian:** Investigation, Validation, Writing – review & editing. **Linmei Hu:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request

Acknowledgments

This work is supported by National Natural Science Foundation of China, No.: 62376137; Shandong Provincial Natural Science Foundation, No.: ZR2022YQ59; and National Science Foundation of China, NSFC, No. 62276029.

References

[1] S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, BPR: Bayesian personalized ranking from implicit feedback, in: Conference on Uncertainty in Artificial Intelligence, AUAI, 2009, pp. 452–461.
[2] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, M. Wang, LightGCN: Simplifying and powering graph convolution network for recommendation, in: Conference on Research and Development in Information Retrieval, ACM, 2020, pp. 639–648.
[3] X. Zhou, D. Lin, Y. Liu, C. Miao, Layer-refined graph convolutional networks for recommendation, CoRR abs/2207.11088, 2022, arXiv:2207.11088.
[4] Y. Himeur, A. Alsalemi, A. Al-Kababji, F. Bensaali, A. Amira, C. Sardianos, G. Dimitrakopoulos, I. Varlamis, A survey of recommender systems for energy efficiency in buildings: Principles, challenges and prospects, Inf. Fusion 72 (2021) 1–21.
[5] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, A. Hussain, Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions, Inf. Fusion 91 (2023) 424–444.
[6] R. He, J.J. McAuley, VBPR: Visual Bayesian personalized ranking from implicit feedback, in: Conference on Artificial Intelligence, AAAI, 2016, pp. 144–150.
[7] Y. Wei, X. Wang, L. Nie, X. He, R. Hong, T. Chua, MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video, in: Conference on Multimedia, ACM, 2019, pp. 1437–1445.
[8] Q. Wang, Y. Wei, J. Yin, J. Wu, X. Song, L. Nie, DualGNN: Dual graph neural network for multimedia recommendation, IEEE Trans. Multim. 25 (2023) 1074–1084.
[9] T. Zhulin, X. Liu, Y. Xia, X. Wang, L. Yang, X. Huang, T.-S. Chua, Self-supervised learning for multimedia recommendation, IEEE Trans. Multim. (2023).
[10] H. Zhou, X. Zhou, Z. Zeng, L. Zhang, Z. Shen, A comprehensive survey on multimodal recommender systems: Taxonomy, evaluation, and future directions, CoRR, 2023, abs/2302.04473.
[11] S. Liu, Z. Chen, H. Liu, X. Hu, User-video co-attention network for personalized micro-video recommendation, in: The World Wide Web Conference, ACM, 2019, pp. 3020–3026.
[12] J. Zhang, Y. Zhu, Q. Liu, S. Wu, S. Wang, L. Wang, Mining latent structures for multimedia recommendation, in: Conference on Multimedia, ACM, 2021, pp. 3872–3880.
[13] Y. Wei, X. Wang, L. Nie, X. He, T. Chua, Graph-refined convolutional network for multimedia recommendation with implicit feedback, in: Conference on Multimedia, ACM, 2020, pp. 3541–3549.
[14] H. Zhou, X. Zhou, Z. Shen, Enhancing dyadic relations with homogeneous graphs for multimodal recommendation, CoRR, 2023, abs/2301.12097.
[15] X. Zhou, H. Zhou, Y. Liu, Z. Zeng, C. Miao, P. Wang, Y. You, F. Jiang, Bootstrap latent representations for multi-modal recommendation, in: Proceedings of the ACM Web Conference, ACM, 2023, pp. 845–854.
[16] Z. Yi, X. Wang, I. Ounis, C. MacDonald, Multi-modal graph contrastive learning for micro-video recommendation, in: International Conference on Research and Development in Information Retrieval, ACM, 2022, pp. 1807–1811.
[17] T. Han, P. Wang, S. Niu, C. Li, Modality matches modality: Pretraining modality-disentangled item representations for recommendation, in: The ACM Web Conference, ACM, 2022, pp. 2058–2066.
[18] D. Pan, X. Li, X. Li, D. Zhu, Explainable recommendation via interpretable feature mapping and evaluation of explainability, in: International Joint Conference on Artificial Intelligence, IJCAI, 2020, pp. 2690–2696.
[19] C. Chen, B. Song, J. Guo, T. Zhang, Multi-dimensional shared representation learning with graph fusion network for session-based recommendation, Inf. Fusion 92 (2023) 205–215.
[20] J. Guo, Y. Zhou, P. Zhang, B. Song, C. Chen, Trust-aware recommendation based on heterogeneous multi-relational graphs fusion, Inf. Fusion 74 (2021) 87–95.
[21] T. Kim, Y. Lee, K. Shin, S. Kim, MARIO: Modality-aware attention and modality-preserving decoders for multimedia recommendation, in: Conference on Information & Knowledge Management, ACM, 2022, pp. 993–1002.
[22] X. Liu, Z. Tao, J. Shao, L. Yang, X. Huang, EliMRec: Eliminating single-modal bias in multimedia recommendation, in: Conference on Multimedia, ACM, 2022, pp. 687–695.
[23] Y. Gu, X. Han, Z. Liu, M. Huang, PPT: Pre-trained prompt tuning for few-shot learning, in: Proceedings of the Association for Computational Linguistics, ACL, 2022, pp. 8410–8423.
[24] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Comput. Surv. 55 (9) (2023) 195:1–195:35.
[25] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D.M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: Neural Information Processing Systems, 2020.
[26] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P.J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (2020) 140:1–140:67.

[27] T. Gao, A. Fisch, D. Chen, Making pre-trained language models better few-shot learners, in: Proceedings the Association for Computational Linguistics, ACL, 2021, pp. 3816–3830.
[28] Z. Jiang, F.F. Xu, J. Araki, G. Neubig, How can we know what language models know, Trans. Assoc. Comput. Linguistics 8 (2020) 423–438.
[29] T. Shin, Y. Razeghi, R.L.L. IV, E. Wallace, S. Singh, AutoPrompt: Eliciting knowledge from language models with automatically generated prompts, in: Conference on Empirical Methods in Natural Language Processing, ACL, 2020, pp. 4222–4235.
[30] B. Lester, R. Al-Rfou, N. Constant, The power of scale for parameter-efficient prompt tuning, in: Conference on Empirical Methods in Natural Language Processing, ACL, 2021, pp. 3045–3059.
[31] Z. Liang, H. Hu, C. Xu, J. Miao, Y. He, Y. Chen, X. Geng, F. Liang, D. Jiang, Learning neural templates for recommender dialogue system, in: Conference on Empirical Methods in Natural Language Processing, ACL, 2021, pp. 7821–7833.
[32] X. Guo, B. Li, H. Yu, Improving the sample efficiency of prompt tuning with domain adaptation, in: Association for Computational Linguistics: EMNLP, ACL, 2022, pp. 3523–3537.
[33] F. Xiao, L. Deng, J. Chen, H. Ji, X. Yang, Z. Ding, B. Long, From abstract to details: A generative multimodal fusion framework for recommendation, in: International Conference on Multimedia, ACM, 2022, pp. 258–267.
[34] J. Ni, J. Li, J.J. McAuley, Justifying recommendations using distantly-labeled reviews and fine-grained aspects, in: Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing, Association for Computational Linguistics, 2019, pp. 188–197.
[35] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks, in: Conference on Empirical Methods in Natural Language Processing, ACL, 2019, pp. 3980–3990.
[36] X. Du, Z. Wu, F. Feng, X. He, J. Tang, Invariant representation learning for multimedia recommendation, in: Conference on Multimedia, ACM, 2022, pp. 619–628.
[37] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: International Conference on Learning Representations, 2015, pp. 1–15.