Multi-Modal Contrastive Pre-training for Recommendation

Zhuang Liu
liuzhuang@buaa.edu.cn
State Key Laboratory of Software Development Environment, Beihang University
Beijing, China

Yunpu Ma
cognitive.yunpu@gmail.com
Ludwig-Maximilians-Universität München
Lehrstuhl für Datenbanksysteme und Data Mining
München, Germany

Matthias Schubert
schubert@dbs.ifi.lmu.de
Ludwig-Maximilians-Universität München
Lehrstuhl für Datenbanksysteme und Data Mining
München, Germany

Yuanxin Ouyang
oyyx@buaa.edu.cn
State Key Laboratory of Software Development Environment, Beihang University
Beijing, China

Zhang Xiong
xiongz@buaa.edu.cn
Engineering Research Center of Advanced Computer Application Technology, Ministry of Education, Beihang University
Beijing, China

ABSTRACT

Personalized recommendation plays a central role in various online applications. To provide quality recommendation service, it is of crucial importance to consider multi-modal information associated with users and items, *e.g.*, review text, description text, and images. However, many existing approaches do not fully explore and fuse multiple modalities. To address this problem, we propose a multi-modal contrastive pre-training model for recommendation. We first construct a homogeneous item graph and a user graph based on the relationship of co-interaction. For users, we propose intra-modal aggregation and inter-modal aggregation to fuse review texts and the structural information of the user graph. For items, we consider three modalities: description text, images, and item graph. Moreover, the description text and image complement each other for the same item. One of them can be used as promising supervision for the other. Therefore, to capture this signal and better exploit the potential correlation of intra-modalities, we propose a self-supervised contrastive inter-modal alignment task to make the textual and visual modalities as similar as possible. Then, we apply inter-modal aggregation to obtain the multi-modal representation of items. Next, we employ a binary cross-entropy loss function to capture the potential correlation between users and items. Finally, we fine-tune the pre-trained multi-modal representations using an existing recommendation model. We have performed extensive experiments on three real-world datasets. Experimental results verify the rationality and effectiveness of the proposed method.

ICMR '22, June 27–30, 2022, Newark, NJ, USA

ACM ISBN 978-1-4503-9238-9/22/06…$15.00
https://doi.org/10.1145/3512527.3531378

CCS CONCEPTS

• **Information systems** → **Recommender systems**.

KEYWORDS

Recommender system; Multi-modal side information; Contrastive learning; Pre-training model

ACM Reference Format:
Zhuang Liu, Yunpu Ma, Matthias Schubert, Yuanxin Ouyang, and Zhang Xiong. 2022. Multi-Modal Contrastive Pre-training for Recommendation. In *Proceedings of the 2022 International Conference on Multimedia Retrieval (ICMR '22), June 27–30, 2022, Newark, NJ, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3512527.3531378

1 INTRODUCTION

To alleviate information overload on the web, recommender systems act as an indispensable tool to help users find their preferred information from massive irrelevant contents [32]. However, traditional recommender systems suffer from data sparsity and cold start problems [21]. To address these issues, multiple modalities side information, including images, texts, and videos, have been exploited to further improve recommendation performance.

As shown in Figure 1(a), an example of user-item interactions with multi-modal side information is displayed. Each item includes two modalities, *e.g.*, textural description, and image. Each user has review text information on the items they have interacted. These multi-modal side information is critical to recommender systems. For example, the visual appearance and textual descriptions play essential roles when users select products online. The review text of users can explore user interests and preferences for items.

Many approaches have been proposed to leverage the multi-modal side information associated with users and items. For example, VBPR [6] extends matrix factorization by incorporating items' visual feature. [13] proposed a review-based recommendation method, which exploits the user review data to describe users' preferences. However, they only consider a specific type of side information for the dedicated recommendation scenario. To better incorporate multi-modal information, MMGCN [29] constructs

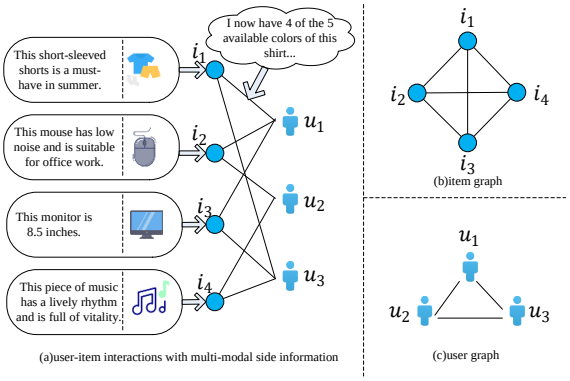

Figure 1: (a) An example user-item interactions with multi-modal side information. (b) Item graph. (c) User graph. These two graphs are built on user-item interactions (such as co-clicks or co-purchases).

modal-specific graph and conducts graph convolutional operations to capture the modal-specific user preference and item representations. [12] proposes a pre-training multi-modal graph transformer method, which learns the item representations with graph structure reconstruction and masked node feature reconstruction. However, these two methods ignore the users' review information and do not better capture the potential correlation of users and items. [14] proposes two pre-training models by constructing two single-relation graphs and multi-relation graphs, which can better capture the potential correlations of users and items. However, they can not analyze the signal of intra-modal and inter-modal in detail. Therefore, in this work, we investigate recommendation models that can, from one side, efficiently utilize the multimodality side information, capture modality-specific features, and, from another side, aggregate cross-modality information from both user and items.

To be more specific, we propose a novel multi-modal contrastive pre-training method for recommendation. Furthermore, an existing recommendation model is built on fine-tuning the pre-trained multi-modal embeddings. Illustrated in Figure 1, in addition to the description texts, images, and the review texts, we construct two homogeneous graphs based on the relationship of co-interaction. Therefore, in this paper, we consider two modalities for users: review texts and user graph, three modalities for items: description texts, images, and item graph. We apply a text encoder, an image encoder, and a graph encoder to obtain the representations for each modality, respectively. For users, we propose intra-modal aggregation and inter-modal aggregation to fuse multiple modalities. Intra-modal aggregation aims to fuse several review texts, and inter-modal aggregation is applied to obtain the muli-modal user representations. For items, we also apply inter-modal aggregation to obtain the multi-modal item representations. In addition, the description text and image complement each other for the same item. One of them can be used as promising supervision for the other. Take Figure 1(a) as an example, the description text of each item is displayed from the perspective of textual modality, and the image is from the perspective of visual modality. The semantics of these two modalities are similar. To capture this signal effectively, we propose a self-supervised contrastive learning method that aligns the textual and the visual modalities of items. After obtaining the multi-modal representations for users and items, we employ a binary cross-entropy loss function to capture the potential correlation between them.

The contributions of this work can be summarized as follows:

- We propose a novel multi-modal contrastive pre-training method to fully exploit the multimodality side information of users and items. And then fine-tune the pre-trained multi-modal representations by an existing recommendation model.
- We propose intra-modal aggregation and inter-modal aggregation to fuse various modalities information and employ an alignment task based on contrastive learning for items' textual and visual modality.
- We conduct extensive experiments on three real-world datasets. The experimental results demonstrate the rationality and effectiveness of our method.

2 RELATED WORK

Multi-modal representation learning is one of the most critical problems in multi-modal applications [15]. In this section, we briefly review several lines of works closely related to ours, including multi-modal representation and multi-modal for recommendation.

Multi-modal representation. The existing multi-modal representation model can be divided into two categories: joint and coordinated [1]. Joint representations combine the various single-modal information into the same representation space. Recently, neural networks are increasingly used in the multi-modal domain [3, 25, 33], which can fuse the different modalities information into a joint representation. Besides, the probabilistic graphical models are another popular way to construct joint representations through the use of latent random variables [10, 20]. Different from joint representations, the coordinated ones learn different representations for each modality but coordinate them with constraints. For example, [16, 30] applies similarity models to minimize the distance between modalities in the coordinated space. In addition, structured, coordinated space models are employed to enforce additional constraints between the modality representations [2, 23, 26]. The type of structure enforcement is often applied with different constraints for cross-modal retrieval and image captioning.

Multi-modal for recommendation. In the field of recommender systems, massive multimedia content information of items is considered to improve recommendation performance. For example, VBPR [6] extends matrix factorization by incorporating visual features from pretrained convolutional neural networks (CNN). Visual-CLiMF [19] enhances VBPR by learning the approximate reciprocal rank instead of pairwise rank in the optimization. MMGCN [29] constructs a user-item bipartite graph in each modality and conducts graph convolutional operations, to better capture the modal-specific user preference. Following MMGCN, GRCN [28] focuses on adaptively refining the structure of the interaction graph to distill informative signals on user preference. The above methods take multimodal features as side information to integrate into

recommendation models. Recently, some multimodal pre-training methods based on graph neural networks have been proposed for recommendation. GPT-GNN [8] introduces a self-supervised attributed graph generation task, including attribute generation and edge generation, to pre-train a graph neural network so that it can capture the structural and semantic properties of the graph. [14] proposes two pre-training models based on graph convolutional networks, named GCN-P and COM-P, by considering the users' and items' side information to construct two single-relational graphs and multi-relational graphs. And then, the pre-trained model is deployed to fine-tune and enhance existing general representation-based recommender systems. Graph-BERT [31] applies transformer to learn node representations based on two tasks: graph structure reconstruction and node feature reconstruction. However, it ignores the masking operations on the nodes, which may limit the ability to aggregate the features of different nodes. To address this problem, PMGT [12] designs a masked node feature reconstruction task, which aims to reconstruct the features of masked nodes by other non-masked nodes so that it improves the recommendation performance. Unlike these existing methods, our proposed multi-modal contrastive pre-training method aims to integrate the multi-modalities information both on the user side and item side, capture modality-specific features and aggregate cross-modality information from both users and items.

3 METHODOLOGY

In this section, we first describe some preliminaries that will be used in this paper and then introduce our method in detail.

3.1 Preliminaries

Let $\mathcal{U} = \{u_1, u_2, \cdots, u_{|\mathcal{U}|}\}$ and $\mathcal{I} = \{i_1, i_2, \cdots, i_{|\mathcal{I}|}\}$ be the sets of users and items respectively, where $|\mathcal{U}|$ is the number of users, and $|\mathcal{I}|$ is the number of items. $\mathbf{R} \in \{0,1\}^{|\mathcal{U}|\times|\mathcal{I}|}$ is the user-item implicit feedback matrix. We assume that $\mathcal{G} = \{(u,i)|u \in \mathcal{U}, i \in \mathcal{I}\}$ is the undirected user-item interaction graph. An edge $y_{ui} = 1$ indicates an observed interaction between user u and item i; otherwise $y_{ui} = 0$. Beyond the interactions, we have multiple modalities for each node. For users, we consider the review texts on the items they have interacted with. For items, we consider the description text and image. We denote the modality features of user u as $\mathbf{e}_u^m \in \mathbb{R}^{d_u^m}$, the modality features of item i as $\mathbf{e}_i^m \in \mathbb{R}^{d_i^m}$, where d_u^m and d_i^m denote the dimension of the features, $m \in \mathcal{M}$ is the modality, and $\mathcal{M}$ is the set of modalities. In addition, we construct two homogeneous graphs $\mathcal{G}_u$ and $\mathcal{G}_i$ based on their co-interact relationships (such as co-clicks or co-purchase) to capture the structural information. The purpose of our proposed pre-training model is to obtain the representations of users and items that can capture the multimodality information and the graph structure. Then, the learned representations can be fine-tuned in the downstream recommendation tasks.

3.2 An Overview of the Proposed Model

The overall architecture of our proposed framework is illustrated in Figure 2. Observe that our method contains two processes: pre-training and fine-tuning. In the pre-training stage, we propose a multi-modal contrastive representation model based on both side information and the implicit feedback matrix $\mathbf{R}$. Specifically, our proposed pre-training model contains two components: user modeling and item modeling. In the component of user modeling, we first employ a text encoder to get each review text's representation and then use intra-modal aggregation to obtain the user's review embedding. Next, a graph encoder is applied to capture the structural information of the homogeneous graph $\mathcal{G}_u$. For these two different modalities information, we develop inter-modal aggregation to obtain the multi-modal representation of the user. In item modeling, we utilize text encoder, image encoder, and graph encoder to encode the description text, image, and homogeneous graph $\mathcal{G}_i$ of each item. Then, we apply inter-modal aggregation to obtain a multi-modal representation of the item. In addition, since the description text and image information complement each other for the same item, they have similar semantics. We develop a self-supervised contrastive learning method to align the representations between them. Finally, according to [14], we employ a binary cross-entropy loss function based on the feedback matrix $\mathbf{R}$ to capture the potential correlation of the target user u and its corresponding target item i. In the fine-tuning process, an existing recommendation model leverages the pre-trained user/item embeddings as initialization and fine-tunes these embeddings based on the feedback matrix $\mathbf{R}$ only.

3.3 User Modeling

User modeling aims to learn multi-modal user latent factors, denoted as $\mathbf{e}_u \in \mathbb{R}^d$ for user u, where d is the length of the embedding vector. The challenge is how to fuse information from multiple modalities inherently. In this paper, we consider two modalities for each user, *i.e.,* review texts and the homogeneous graph $\mathcal{G}_u$. To address this challenge, we first apply two encoders to encode each modality into latent representations and then employ two types of aggregation to learn users' multi-modal representation, *i.e.*, intra-modal aggregation and inter-modal aggregation. Next, we will introduce each component in detail.

Multi-modal Encoder. To obtain the review representations, we utilize the pre-trained Transformer [22] with the architecture modifications described in [18] to extract the features of each review text. Each review text sequence is bracketed with *[SOS]* and *[EOS]* tokens. Furthermore, we take the activations of the last layer of the transformer at the *[EOS]* token as the feature representation of each review text which is layer normalized.

As for the graph modality, we first construct a homogeneous user graph $\mathcal{G}_u$ based on the user-item interactions. Specifically, if two users have commonly interacted items, we build an edge between these two users. Let S_u denote the set of items which the user u has interacted with, then the user graph can be formally formulated as $\mathcal{G}_u = \{(u_i, u_j)|u_i \in \mathcal{U}, u_j \in \mathcal{U}, \text{and} \quad S_{u_i} \cap S_{u_j} \neq \emptyset\}$.

To capture the structural information of the user graph $\mathcal{G}_u$, we consider the LightGCN [7] model as graph encoder. It is the light version of GCN, including only the most essential component in GCN - neighborhood aggregation. Let L denotes the number of GCN layers, $\mathcal{N}_u$ denotes the neighborhood of node u, then the graph convolution operation in LightGCN is defined as:

$$\mathbf{e}_u^{(l+1)} = \sum_{u' \in \mathcal{N}_u} \frac{1}{\sqrt{|\mathcal{N}_u|}} \mathbf{e}_{u'}^{(l)}, \tag{1}$$

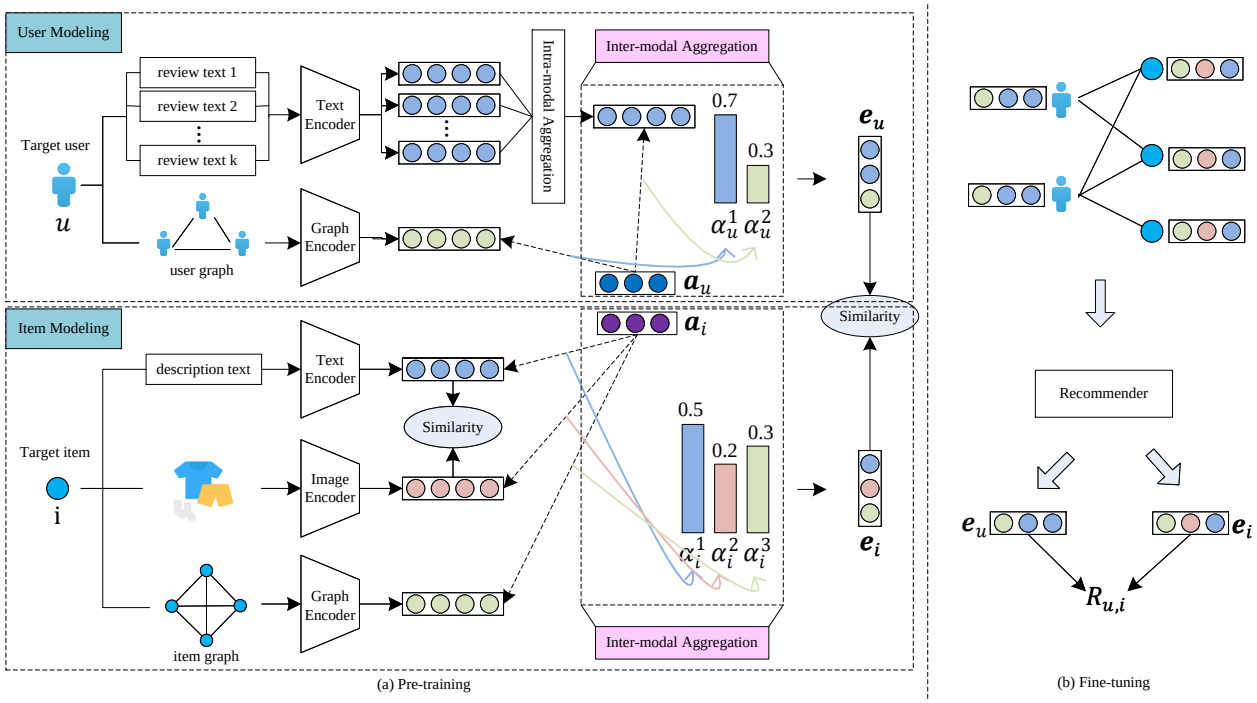

Figure 2: The overall architecture of the proposed framework.

where l is the index of GCN layers, $\frac{1}{\sqrt{|\mathcal{N}_u|}}$ is the normalization term. After L layers LightGCN, we further combine the embeddings obtained at each layer to form the final representation of a node in the user graph:

$$\mathbf{e}_u^* = \frac{1}{L+1}\sum_{l=0}^{L}\mathbf{e}_u^{(l)}, \tag{2}$$

where $\mathbf{e}_u^{(0)}$ is the embedding at the 0-th layer, which is the trainable model parameters and denotes the graph modality information.

Intra-modal Aggregation. For each user, there will be several items that they have interacted with. For each interacted item, there will be a piece of review text. Here, we assume that a user u interacts with k items $\{i_1, i_2, \cdots, i_k\}$ and define the corresponding review texts as $\{r_1, r_2, \cdots, r_k\}$. We can obtain the representation of each review text by the text encoder, which leads to:

$$\mathbf{e}_{\mathbf{r}_i} = f_{text\_encoder}(r_i), \quad i \in \{1, 2, \cdots, k\} \tag{3}$$

where $\mathbf{e}_{\mathbf{r}_i}$ is the representation of review r_i.

To get the text modality representation for user u, we need to aggregate all the user u's review representations $\{\mathbf{e}_{\mathbf{r}_1}, \mathbf{e}_{\mathbf{r}_2}, \cdots, \mathbf{e}_{\mathbf{r}_k}\}$, we define it as intra-modal aggregation. Specifically, it can be abstracted as:

$$\mathbf{e}_u^1 = AGG\left(\mathbf{e}_{\mathbf{r}_i}, i \in \{1, 2, \cdots, k\}\right). \tag{4}$$

The AGG is an aggregation function. Many operators can be used, such as max aggregator, weighted sum aggregator, LSTM aggregator, bilinear interaction aggregator, etc. In this paper, we employ the average pooling operation on these representations. It assumes that different reviews have the same contributions to the text modality representation for $\mathbf{e}_u^1$. In our experiments, we find that this operator can lead to good performance in general. Thus we do not design a special component to optimize the AGG function. We leave the analysis of fine-grained sentiment of the reviews as future work.

Inter-modal Aggregation. By intra-modal aggregation, we obtain the text modality representation $\mathbf{e}_u^1$ for user u. Similarly, we can get the graph modality representation by the graph encoder:

$$\mathbf{e}_u^2 = f_{graph\_encoder}(\mathcal{G}_u) \tag{5}$$

In addition, in order to obtain the multi-modal representation for user u, we have to aggregate these two modality representations $\mathbf{e}_u^1$ and $\mathbf{e}_u^2$. We define it as inter-modal aggregation. Moreover, different modalities' representations usually lie in different spaces. So, we need to project all modality representations into a common latent vector space. Specifically, for modality m, we design a modal-specific mapping matrix $\mathbf{W}_{um} \in \mathbb{R}^{d_u^m \times d}$ to transform its representation $\mathbf{e}_u^m$ into common space as follows:

$$\mathbf{e'}_u^m = \mathbf{W}_{um}^\top \cdot \mathbf{e}_u^m + \mathbf{b}_u^m, \quad m \in \{1, 2\}, \tag{6}$$

where $\mathbf{e'}_u^m$ is the projected feature of modality m for user u, $\mathbf{b}_u^m \in \mathbb{R}^{d\times 1}$ denotes as vector bias. Next, we employ an attention mechanism to aggregate messages from different modalities:

$$\mathbf{e}_u = \sum_{m\in\{1,2\}} \alpha_u^m \cdot \mathbf{e'}_u^m, \tag{7}$$

where α_m denotes the attention value of modality m. It can be calculated as follows:

$$\alpha_u^m = \frac{exp\left(ReLU\left(\mathbf{a}_u^\top \cdot \mathbf{e'}_u^m\right)\right)}{\sum_{j\in\{1,2\}} exp\left(ReLU\left(\mathbf{a}_u^\top \cdot \mathbf{e'}_u^j\right)\right)}, \tag{8}$$

where $\mathbf{a_u} \in \mathbb{R}^{d\times 1}$ is the attention vector for users.

3.4 Item Modeling

As shown in the below part of Figure 2 in the pre-training stage, item modeling is used to learn multi-modal item latent factors, denoted as $\mathbf{e}_i \in \mathbb{R}^d$ for item i. In this paper, items are associated with three modalities, which contain description text, image, and homogeneous graph $\mathcal{G}_i$. Similarly, we first apply three encoders to obtain the corresponding modality representations and then fuse these modalities by inter-modal aggregation. In addition, we develop a self-supervised contrastive learning method to align the textual modality and visual modality.

Multi-modal Encoder. Similar to user modeling, we also select the pre-trained Transformer and LightGCN to encode an item i's description text and homogeneous graph $\mathcal{G}_i$, respectively:

$$\mathbf{e}_i^1 = f_{text\_encoder}(t), \quad \mathbf{e}_i^3 = f_{graph\_encoder}(\mathcal{G}_i) \tag{9}$$

where t is the item i's description text, $\mathbf{e}_i^1$ denotes the item i's textual modality representation, $\mathbf{e}_i^3$ denotes graph modality representation. And the construction way of $\mathcal{G}_i$ is similar to that of $\mathcal{G}_u$. Namely, $\mathcal{G}_i = \{(i_a, i_b)|i_a \in \mathcal{I}, i_b \in \mathcal{I}, \text{and} \quad S_{i_a} \cap S_{i_b} \neq \emptyset\}$, where S_i denotes the set of users that interact with item i.

To extract the features of the image, we consider CLIP [17] as the visual encoder, whose parameters are pre-trained based on the dataset of 400 million (image, text) pairs collected by [17]. Formally, the representation of visual modality for item i can be defined as:

$$\mathbf{e}_i^2 = f_{image\_encoder}(img), \tag{10}$$

where img denotes item i's image information.

Contrastive Inter-modal Alignment. Since the item i's description text and image refer to similar content, and they are different ways to describe the item i, one of them can be used as promising supervision for the other, so we make these two modality representations as similar as possible. Moreover, we develop a self-supervised contrastive inter-modal alignment (CIMA) task to align these two representations. It is worth noting that we ignore the graph modality in this task since it aims to capture the local graph structure and needs to consider the item i's neighborhoods. They have different semantics in the embedding space.

The contrastive inter-modal alignment task aims to make $\mathbf{e}_i^1$ and $\mathbf{e}_i^2$ close to each other in the learned embedding space if they are from the same item, otherwise far away. Given a batch of N items, we can obtain N (text, image) pairs, CIMA is trained to predict which of the $N \times N$ possible (text, image) pairings across a batch actually occurred. To do this, CIMA aims to maximize the cosine similarity of $\mathbf{e}_i^1$ and $\mathbf{e}_i^2$ of the N real pairs in the batch while minimizing the other $N^2 - N$ incorrect pairings. We optimize a cross-entropy loss over these similarity scores.

$$\mathcal{L}_1 = -\sum_{C_{p,q}\in \mathbf{C}} C_{p,q} \log s\left(\mathbf{e}_{i_p}^1, \mathbf{e}_{i_q}^2\right) + \left(1 - C_{p,q}\right) \log\left(1 - s\left(\mathbf{e}_{i_p}^1, \mathbf{e}_{i_q}^2\right)\right), \tag{11}$$

where $\mathbf{C} \in \mathbb{R}^{N\times N}$ is a diagonal matrix, p, q is the index of $\mathbf{C}$, if $p = q$, $C_{p,q} = 1$, otherwise, $C_{p,q} = 0$. $s\left(\mathbf{e}_{i_p}^1, \mathbf{e}_{i_q}^2\right)$ is the scaled pairwise cosine similarity function, namely, $s\left(\mathbf{e}_{i_p}^1, \mathbf{e}_{i_q}^2\right) = \mathbf{e}_{i_p}^1 \cdot \mathbf{e}_{i_q}^2 / (\tau \cdot \|\mathbf{e}_{i_p}^1\| \|\mathbf{e}_{i_p}^2\|)$, τ denotes a temperature parameter.

Inter-modal aggregation. By multi-modal encoder, we can obtain three modality representations $\mathbf{e}_i^1, \mathbf{e}_i^2, \mathbf{e}_i^3$ for item i. We also apply inter-modal aggregation to obtain the multi-modal representation $\mathbf{e}_i$. Similar to user modeling, we first design a modal-specific mapping matrix $\mathbf{W}_{im} \in \mathbb{R}^{d_i^m \times d}$ to transform each modality representation into common space as follows:

$$\mathbf{e'}_i^m = \mathbf{W}_{im}^\top \cdot \mathbf{e}_i^m + \mathbf{b}_i^m, \quad m \in \{1, 2, 3\}, \tag{12}$$

where $\mathbf{e'}_i^m$ is the projected feature of modality m for item i, $\mathbf{b}_i^m \in \mathbb{R}^{d\times 1}$ is the bias vector. Then, we apply attention mechanism to aggregate different modalities:

$$\mathbf{e}_i = \sum_{m\in\{1,2,3\}} \alpha_i^m \cdot \mathbf{e'}_i^m, \tag{13}$$

where α_i^m is the attention score of modality m. It can be calculated as follows:

$$\alpha_i^m = \frac{exp\left(ReLU\left(\mathbf{a}_i^\top \cdot \mathbf{e'}_i^m\right)\right)}{\sum_{j\in\{1,2,3\}} exp\left(ReLU\left(\mathbf{a}_i^\top \cdot \mathbf{e'}_i^j\right)\right)}, \tag{14}$$

where $\mathbf{a_i} \in \mathbb{R}^{d\times 1}$ is the attention vector for items.

3.5 Model Optimization

So far, we have obtained a multi-modal representation for each user and item. To make representations that capture the potential correlation between users and their interacted items, according to [14], we apply a binary cross-entropy (BCE) loss function to make the target user u and the target item i as similar as possible:

$$\mathcal{L}_2 = -\sum_{R_{i,j}\in\mathbf{R}} R_{i,j} \cdot \log \hat{R}_{i,j} + \left(1 - R_{i,j}\right) \cdot \log\left(1 - \hat{R}_{i,j}\right), \tag{15}$$

where $\hat{R}_{i,j}$ is the predicted score, which is calculated by the embedding dot product $\hat{R}_{i,j} = \mathbf{e}_u^\top \mathbf{e}_i$.

To effectively learn the parameters in the pre-training stage, we need to specify an overall objective function. In this paper, we have formulated two components of the loss function: $\mathcal{L}_1$ aims to align the items' textual representation and visual representation. $\mathcal{L}_2$ is to capture the correlation between the target user and target item. We form a linear combination of the two components to obtain a joint loss function. That leads to

$$\mathcal{L} = \lambda \cdot \mathcal{L}_1 + (1 - \lambda) \cdot \mathcal{L}_2, \tag{16}$$

where λ is a coefficient that balances between the two losses. We employ Adam [11] as the optimizer. Its main advantage is that the learning rate can be self-adaptive during the training phase, which eases the pain of choosing a proper learning rate.

3.6 Fine-tuning with Existing Recommendation Model

Most of the existing recommendation models randomly initialize the embedding of users and items from a uniform distribution [24], or the Xavier distribution [7, 27], etc. However, this initialization method lacks prior knowledge, which leads to the instability of the model [14], falls into the locally-optimal solutions, and then deteriorates the recommendation performance. To address this issue, we propose to initialize the users' and items' embeddings from the output of our pre-training model. Then we further fine-tune these embeddings with the recommendation model's own parameter optimizer. Specifically, we first apply our proposed pre-training model to obtain the multimodal representation of users and items, and then feed them to the existing recommendation model as the parameter initialization. In this paper, we select LightGCN based on MixGCF [9] to further fine-tune these embeddings with the interactions **R** only, which is the state-of-the-art GNN-based recommendation model.

4 EXPERIMENTS

In this section, we conduct experiments to evaluate the performance of our proposed model on three public datasets. Specifically, we aim to answer the following research questions:

- **RQ1**: How does our model perform compared with the state-of-the-art pre-training methods, such as PMGT, GCN-P, etc.?
- **RQ2**: What is the contribution of various modalities (*e.g.*, textual, visual, and graph structural) to the overall performance?
- **RQ3**: How do different training objectives $\mathcal{L}_1$ and $\mathcal{L}_2$ influence the recommendation performance?
- **RQ4**: How robust is our pre-training model with respect to different loss weights β?

4.1 Experimental Settings

4.1.1 Datasets. To evaluation the effectiveness of our proposed pre-training models, we conduct experiments on three real-world datasets: Amazon (include "Video Games" and "Toys and Games"), Sharee. These datasets vary significantly in their domains, size, and sparsity. A detailed description can be found in Appendix A.

For each dataset, we randomly select 80% of the historical interactions of each user and assign them to the training set. The remainder comprises the test set. The statistics of the processed datasets are summarized in Table 1. We build the item graph and user graph based on users' interactions in the training set. That is, to build item graph, if two items i_1 and i_2 interacted by the common user, we construct an edge between i_1 and i_2. Similarly, to build the user graph, if two users u_1 and u_2 have interacted with the same item, we construct an edge between u_1 and u_2.

4.1.2 Evaluation Protocols. To evaluate the quality of the recommendation models, we adopt two widely-used ranking-based metrics: Recall@k and NDCG@k, which are computed by the all-ranking protocol - all items that are not interacted by a user are the candidates. Specifically, Recall@k measures the average number of items that the users interact with that are ranked among the top-k ranking list. Moreover, NDCG@k considers the hit position of the items and gives a higher score if the hit items are in the top positions. In this work, we report Recall and NDCG with $k = 20, 40, 60$. For all these metrics, the higher the value, the better the performance.

Table 1: Statistics of the datasets.

Dataset	#Users	#Items	#Interactions	Sparsity
VG	55217	17389	472586	99.95%
TG	208143	78698	1754420	99.99%
Sharee	43719	34654	500000	99.97%

4.1.3 Baseline Methods. To verify the effectiveness of our proposed pre-training model, we compare it with the following representative baseline methods: **Random**, **CLIP** [17], **MMGCN** [29], **GPT-GNN** [8], **Graph-BERT** [31], **PMGT** [12], and **GCN-P** [14]. A short description of all baselines is given in Appendix B.

4.1.4 Implementation Details. In the experiments, we first pre-train the embeddings for users and items on the training set, and then fine-tune them in the recommendation model. In the pre-training stage, we set the dimensionality of the review text and graph modalities d_u^1 and d_u^2 as 512, 64, respectively, for users. For items, we set the textual, visual, and graph modalities' dimension d_i^1, d_i^2 and d_i^3 to 512, 512, and 64, respectively. Moreover, the dimensionality of multi-modal representation d is set to 64. For graph modality, the number of LightGCN layers L is set to 2, and we use Xavier initializer [4] to initialize the graph-specific user/item embeddings. We fix the batch size to 2048 for all baselines and our method. Grid search is applied to choose the learning rate and the loss weight λ over the ranges $\{10^{-4}, 10^{-3}, 10^{-2}\}$ and $\{0.2, 0.4, 0.6, 0.8\}$. In most cases, the optimal values are 10^{-3} and 0.2, respectively. The temperature parameter τ is set to 0.1. In the fine-tuning process, we set the batch size to 2048 for LightGCN, the candidate size M in MixGCF is set to 64. We train 1000 epochs for the recommendation model to converge. All other hyper-parameters are set according to the suggestions from the settings specified in the original publications.

4.2 Performance Comparison (RQ1)

To evaluate the effectiveness of our proposed multi-modal contrastive pre-training method, we take the pre-trained user and item representations as initialization to train the LightGCN based on MixGCF recommendation model. Table 2 summarizes the best results of all considered baselines on the three datasets. We have the following observations:

The random initialization method shows the worst performance on three datasets. This indicates it lacks prior knowledge and can not well guide the recommendation model to converge to the local optimal value, further limiting the performance. Compared with the random initialization, initializing the recommendation model with pre-trained representations usually achieves better performance. This demonstrates that the pre-training strategies can provide more refined representations for each user and item in the downstream recommendation task, further improving the performance.

CLIP is a text-image pair pre-training model that captures text and image correlation by an alignment task. We can find that textual and visual information can effectively improve the recommendation

Table 2: A comparison of the overall performance among all considered baselines on three datasets.

Datasets	Metrics	Random	CLIP	MMGCN	GPT-GNN	Graph-BERT	PMGT	GCN-P	**OURS**
VG	Recall@20	0.1571	0.1712	0.1703	0.1659	0.1696	0.1718	0.1916	**0.1939**
	Recall@40	0.2300	0.2470	0.2441	0.2404	0.2446	0.2490	0.2663	**0.2712**
	Recall@60	0.2827	0.3015	0.2950	0.2918	0.2961	0.3018	0.3161	**0.3228**
	NDCG@20	0.0743	0.0814	0.0823	0.0789	0.0811	0.0814	0.0954	**0.0957**
	NDCG@40	0.0904	0.0983	0.0987	0.0955	0.0977	0.0988	0.1120	**0.1130**
	NDCG@60	0.1006	0.1088	0.1086	0.1054	0.1077	0.1090	0.1217	**0.1230**
TG	Recall@20	0.0876	0.0967	0.0966	0.0931	0.1034	0.1024	0.1025	**0.1067**
	Recall@40	0.1233	0.1347	0.1347	0.1301	0.1392	0.1423	0.1374	**0.1450**
	Recall@60	0.1478	0.1616	0.1615	0.1560	0.1622	0.1693	0.1611	**0.1712**
	NDCG@20	0.0433	0.0475	0.0475	0.0458	0.0530	0.0509	0.0529	**0.0534**
	NDCG@40	0.0512	0.0560	0.0559	0.0540	0.0612	0.0597	0.0607	**0.0619**
	NDCG@60	0.0559	0.0611	0.0611	0.0589	0.0659	0.0649	0.0653	**0.0670**
Sharee	Recall@20	0.1059	0.1135	0.1106	0.1090	0.1074	0.1077	0.1382	**0.1392**
	Recall@40	0.1685	0.1762	0.1724	0.1706	0.1684	0.1711	0.2090	**0.2106**
	Recall@60	0.2157	0.2266	0.2208	0.2192	0.2156	0.2160	0.2586	**0.2627**
	NDCG@20	0.0533	0.0572	0.0557	0.0550	0.0535	0.0538	0.0719	**0.0723**
	NDCG@40	0.0694	0.0735	0.0716	0.0712	0.0693	0.0701	0.0903	**0.0908**
	NDCG@60	0.0800	0.0848	0.0826	0.0819	0.0800	0.0804	0.1017	**0.1027**

performance. In comparison, MMGCN constructs modal-specific user-item bipartite graphs to obtain users' and items' multi-modal representations. However, it slightly underperforms CLIP. One potential reason is that in our three datasets, the text and image of the same item are more similar. We should make their representations as close as possible instead of dividing modalities.

GPT-GNN, Graph-BERT, and PMGT are three GNN-based deep pre-training methods, and they employ GNN to aggregate the neighbor information in the homogeneous item graph, which can obtain a better representation for each item. Specifically, PMGT outperforms GPT-GNN and Graph-BERT in most cases. This demonstrates the effectiveness of PMGT in exploiting the item graph structure and item features. GCN-P is the state-of-the-art pre-training model for recommendation, which outperforms all other consider baselines.

Our proposed method consistently yields the best performance on all three datasets for various metrics. In particular, our method improves over the strongest baselines GCN-P *w.r.t.* Recall@20, Recall@40, and Recall@60 by 4.10%, 5.53%, and 6.27% in TG, respectively. By intra-modal aggregation, inter-modal aggregation, and aligning the textual modality and visual modality for items, our method is capable of capturing multiple modalities for users and items. In contrast, GCN-P ignores the alignment and multi-modal aggregation tasks. This indicates that multi-modalities information is critical for improving the recommendation performance.

4.3 Effect of Modalities (RQ2)

To explore the effects of different modalities, we compare the experimental results on different modalities over all the three datasets, as shown in Figure 3. It shows the performance in terms of Recall@20 and NDCG@20 for our method. It is noted that the Sharee dataset does not have review text for users, so we only compare the other three modalities.

We observe that the method with multi-modal information outperforms the variants that only consider single modality information on all three datasets. In addition, the performance with respect to NDCG@20 has a similar trend as Recall@20. The graph modality is the most effective among the other modalities. It makes sense since graph modality information can well capture the potential interaction between users and items and better reflect user preferences. Compared with the visual modality, the textual modality provides more critical information for recommendation on VG and TG datasets. This is reasonable since they are product datasets, the textual description highly related to the products and more detailed than visual modality. However, for the micro-video dataset - Sharee, the visual modality offers important cues than the textual modality since the texts are of low quality. That is, the descriptions are noisy, incomplete, and even irrelevant to the micro-video content. This indicates that different modalities have different effects on different datasets. In general, multi-modal representations can lead to better recommendation performance.

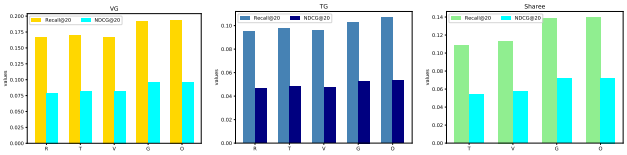

Figure 3: Performance of our proposed pre-training model considering different modality information on three datasets. R denotes the reviews information for users. T, V denotes the textual and visual modality information for items. G denotes the graph modality information. O denotes original model considering all the modality information.

4.4 Ablation Study (RQ3)

In the proposed pre-training method, we mainly design two objectives for learning the representations of all the users and items. We now check whether the recommendation performance improvements are actually the result of these two components. To answer **RQ3**, we conduct an ablation study to analyze their impacts. Table 3 shows the performance each variant on all the three datasets in terms of various metrics. The percentages in the subscript brackets indicate the relative decline of each variant compared to our complete pre-training method.

Table 3: Results of ablation studies. "w/o CIMA" indicates training the model without $\mathcal{L}_1$. "w/o BCE" indicates training the model without $\mathcal{L}_2$.

Datasets	Metrics	w/o CIMA	w/o BCE
VG	Recall@20	$0.1898_{(-2.11\%)}$	$0.1729_{(-10.83\%)}$
	Recall@40	$0.2671_{(-1.51\%)}$	$0.2499_{(-7.85\%)}$
	Recall@60	$0.3181_{(-1.46\%)}$	$0.3022_{(-6.38\%)}$
	NDCG@20	$0.0922_{(-3.66\%)}$	$0.0821_{(-14.21\%)}$
	NDCG@40	$0.1094_{(-3.19\%)}$	$0.0992_{(-12.21\%)}$
	NDCG@60	$0.1193_{(-3.01\%)}$	$0.1093_{(-11.14\%)}$
TG	Recall@20	$0.1051_{(-1.50\%)}$	$0.0974_{(-8.72\%)}$
	Recall@40	$0.1431_{(-1.31\%)}$	$0.1367_{(-5.72\%)}$
	Recall@60	$0.1695_{(-0.99\%)}$	$0.1637_{(-4.38\%)}$
	NDCG@20	$0.0525_{(-1.69\%)}$	$0.0477_{(-10.61\%)}$
	NDCG@40	$0.0610_{(-1.45\%)}$	$0.0564_{(-8.89\%)}$
	NDCG@60	$0.0662_{(-1.19\%)}$	$0.0616_{(-8.06\%)}$
Sharee	Recall@20	$0.1316_{(-5.46\%)}$	$0.1143_{(-17.89\%)}$
	Recall@40	$0.2020_{(-4.08\%)}$	$0.1805_{(-14.29\%)}$
	Recall@60	$0.2535_{(-3.50\%)}$	$0.2274_{(-13.44\%)}$
	NDCG@20	$0.0678_{(-6.22\%)}$	$0.0576_{(-20.33\%)}$
	NDCG@40	$0.0861_{(-5.18\%)}$	$0.0746_{(-17.84\%)}$
	NDCG@60	$0.0979_{(-4.67\%)}$	$0.0854_{(-16.85\%)}$

The most obvious observation from Table 3 is that no matter which component we remove, it will degrade the recommendation performance to varying degrees. It indicates that these two components can capture different information for users and items from different perspectives, which has a positive effect on the improvement of recommendation performance. In particular, when we train the model without $\mathcal{L}_2$, the performance degradation is more serious; the relative declines are 14.21%, 10.61%, and 20.33% with respect to NDCG@20 on the datasets of VG, TG, and Sharee, respectively. This shows that compared with the contrastive inter-modal alignment task, effectively capturing the potential correlations between users and items is of more positive significance to the improvement of recommendation performance. On the other hand, with the increase of k, the degrading range of Recall@k and NDCG@k is smaller, which shows that our method is more sensitive to small k.

4.5 Parameter Sensitivity Analysis (RQ4)

In this subsection, we examine the robustness with respect to the influential hyper-parameter: loss weight β. We analyze the loss weight β by fixing the remaining hyper-parameters at their optimal value.

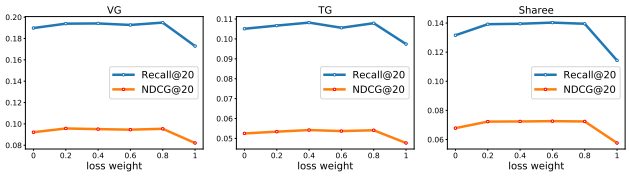

Figure 4: Performance of our method w.r.t. different loss weights β on three datasets.

Figure 4 shows the Recall@20 and NDCG@20 for our proposed method with the loss weight β varying from 0 to 1 by step 0.2. Especially, for $\beta = 0$, there will be no $\mathcal{L}_1$, for $\beta = 1$, no $\mathcal{L}_2$. The most obvious observation from Figure 4 is that the Recall@20 and NDCG@20 have the same trend as β increases: They both increase steadily up to specific high values of β. If we continue to increase β further, the performance begins to drop. This shows that we can obtain better performance by selecting β in an appropriate interval. These two loss functions promote each other. No matter which loss is considered only, the optimal recommendation performance can not be obtained. Specifically, we can obtain the best performance when setting β to 0.8, 0.4, 0.6 on VG, TG, and Sharee datasets, respectively.

5 CONCLUSIONS

In this paper, we introduced a multi-modal contrastive pre-training method for recommendation. In particular, we first constructed a homogeneous user graph and item graph based on the relationship of co-interaction. Then, we applied different encoders to encode different modalities. For users, we employed intra-modal aggregation to obtain the representations of review texts and then utilized inter-modal aggregation to obtain the users' multi-modal representations. For items, we proposed a contrastive inter-modal alignment task to make the representations of textual and visual modalities as similar as possible, and then employed inter-modal aggregation to obtain the items' multi-modal representations. Finally, we applied a binary cross-entropy loss to capture the correlation between users and items. We fine-tuned the pre-trained multi-modal representations by an existing recommendation model. The superiority of the proposed method has been validated on three real-world datasets. Specifically, our method improves over the strongest baselines GCN-P *w.r.t.* Recall@20, Recall@40, and Recall@60 by 4.10%, 5.53%, and 6.27% in TG, respectively. Further analyses are provided towards the rationality of each designed component, modalities, and the robustness of influential hyper-parameter.

ACKNOWLEDGMENTS

This work was partially supported by the National Natural Science Foundation of China (No.61977002), the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A and the State Key Laboratory of Software Development Environment of China (No. SKLSDE-2022ZX-14). The authors of this work take full responsibilities for its content. We thank the anonymous reviewers for their insightful comments and suggestions on this paper.

REFERENCES

[1] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443.

[2] Yue Cao, Mingsheng Long, Jianmin Wang, Qiang Yang, and Philip S Yu. 2016. Deep visual-semantic hashing for cross-modal retrieval. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1445–1454.

[3] Zhiyong Cheng, Xiaojun Chang, Lei Zhu, Rose C Kanjirathinkal, and Mohan Kankanhalli. 2019. MMALFM: Explainable recommendation by leveraging reviews and images. *ACM Transactions on Information Systems (TOIS)* 37, 2 (2019), 1–28.

[4] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 249–256.

[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*. 507–517.

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

[8] Ziniu Hu, Yuxiao Dong, Kuansan Wang, Kai-Wei Chang, and Yizhou Sun. 2020. Gpt-gnn: Generative pre-training of graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1857–1867.

[9] Tinglin Huang, Yuxiao Dong, Ming Ding, Zhen Yang, Wenzheng Feng, Xinyu Wang, and Jie Tang. 2021. MixGCF: An Improved Training Method for Graph Neural Network-based Recommender Systems. (2021).

[10] Yelin Kim, Honglak Lee, and Emily Mower Provost. 2013. Deep learning for robust feature generation in audiovisual emotion recognition. In *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 3687–3691.

[11] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[12] Yong Liu, Susen Yang, Chenyi Lei, Guoxin Wang, Haihong Tang, Juyong Zhang, Aixin Sun, and Chunyan Miao. 2021. Pre-training Graph Transformer with Multimodal Side Information for Recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 2853–2861.

[13] Yong Liu, Susen Yang, Yinan Zhang, Chunyan Miao, Zaiqing Nie, and Juyong Zhang. 2021. Learning Hierarchical Review Graph Representations for Recommendation. *IEEE Transactions on Knowledge and Data Engineering* (2021).

[14] Zaiqiao Meng, Siwei Liu, Craig Macdonald, and Iadh Ounis. 2021. Graph neural pre-training for enhancing recommendations using side information. *arXiv preprint arXiv:2107.03936* (2021).

[15] Liqiang Nie, Xuemeng Song, and Tat-Seng Chua. 2016. Learning from multiple social networks. *Synthesis lectures on information concepts, retrieval, and services* 8, 2 (2016), 1–118.

[16] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. 2016. Jointly modeling embedding and translation to bridge video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4594–4602.

[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. *arXiv preprint arXiv:2103.00020* (2021).

[18] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[19] Sujoy Roy and Sharath Chandra Guntuku. 2016. Latent factor representations for cold-start video recommendation. In *Proceedings of the 10th ACM conference on recommender systems*. 99–106.

[20] Nitish Srivastava, Ruslan Salakhutdinov, et al. 2014. Multimodal learning with deep Boltzmann machines. *J. Mach. Learn. Res.* 15, 1 (2014), 2949–2980.

[21] Zhu Sun, Qing Guo, Jie Yang, Hui Fang, Guibing Guo, Jie Zhang, and Robin Burke. 2019. Research commentary on recommendations with side information: A survey and research directions. *Electronic Commerce Research and Applications* 37 (2019), 100879.

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

[23] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2015. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361* (2015).

[24] Mengting Wan, Di Wang, Jie Liu, Paul Bennett, and Julian McAuley. 2018. Representing and recommending shopping baskets with complementarity, compatibility and loyalty. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 1133–1142.

[25] Shuhui Wang, Yangyu Chen, Junbao Zhuo, Qingming Huang, and Qi Tian. 2018. Joint global and co-attentive representation learning for image-sentence retrieval. In *Proceedings of the 26th ACM international conference on Multimedia*. 1398–1406.

[26] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. 2015. On deep multi-view representation learning. In *International conference on machine learning*. PMLR, 1083–1092.

[27] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural graph collaborative filtering. In *Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval*. 165–174.

[28] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2020. Graph-refined convolutional network for multimedia recommendation with implicit feedback. In *Proceedings of the 28th ACM International Conference on Multimedia*. 3541–3549.

[29] 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.

[30] Ran Xu, Caiming Xiong, Wei Chen, and Jason Corso. 2015. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29.

[31] Jiawei Zhang, Haopeng Zhang, Congying Xia, and Li Sun. 2020. Graph-bert: Only attention is needed for learning graph representations. *arXiv preprint arXiv:2001.05140* (2020).

[32] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Mengqi Zhang, Shu Wu, and Liang Wang. 2021. Latent Structures Mining with Contrastive Modality Fusion for Multimedia Recommendation. *arXiv preprint arXiv:2111.00678* (2021).

[33] Zheng Zhang, Lizi Liao, Minlie Huang, Xiaoyan Zhu, and Tat-Seng Chua. 2019. Neural multimodal belief tracker with adaptive attention for dialogue systems. In *The World Wide Web Conference*. 2401–2412.

A DATASETS

We conduct experiments on the following three real-world datasets:

- **Amazon Datasets**: This is a set of product review datasets crawled from Amazon.com, which are widely used for product recommendation[5]. They are split into separate datasets according to the top-level product categories on Amazon. In this work, we adopt the following 5-core review subsets for studying the item recommendation, such that each of the remaining users and items has five reviews each, *i.e.*, "Video Games" and "Toys and Games" (short for VG and TG, respectively). The metadata of a product includes its text description and the URL of its image[1], which are used to extract the textual and visual features of the product, respectively. For each user, we utilize the review texts to initialize its features. Following [12], we convert all the observed review ratings to be positive interactions and filter out the products that are not included in the metadata files.
- **Sharee Dataset**: Sharee (now renamed Lemon8) is a stream of interest information under ByteDance for the Japanese market, which is a benchmark for the Japanese version of Xiaohongshu. It has three first-level menus: homepage, release, and personal homepage. Users can find the content they are interested in on the homepage. We take the content clicked by the user as the positive interactions, its title, and cover image are used to extract the textual and visual features. We select user interaction data for 30 days from July 1, 2021, to July 30, 2021, filter out users and items with less than 10 interactions, and randomly sample 500000 pieces of interaction data as our Sharee dataset.

[1] https://nijianmo.github.io/amazon/index.html

B BASELINES

To verify the effectiveness of our proposed pre-training model, we compare it with the following representative baseline methods:

- **Random**: The user and item embeddings are randomly initialized for the recommendation model.
- **CLIP** [17] is a pre-training model based on image-text pairs, which jointly trains an image encoder and text encoder to maximize the cosine similarity if the image and text embeddings of all the real pairs. In this paper, we construct an image-text pair for each item, then utilize CLIP to obtain the image and text embeddings, and employ the average pooling operation to initialize the items.
- **MMGCN** [29] is a multi-modal graph convolution network framework, which can yield modal-specific representations of users and items to capture user preferences better.
- **GPT-GNN** [8] employs the attribute generation and edge generation tasks to pre-train the graph neural network model, which can capture the inherent dependency between node attributes and graph structure during the generative process.
- **Graph-BERT** [31] applies graph-transformer to pre-train the graph neural network model based on the node attribute reconstruction and graph structure recovery tasks. However, it does not employ masking operations on the nodes.
- **PMGT** [12] is a pre-trained multi-modal graph transformer model, which learns item representations by considering both item side information and their relationships. Different from Graph-BERT, it designs a masked node feature reconstruction task to obtain more refined embeddings.
- **GCN-P** [14] constructs user-user graph and item-item graph from the users' and items' side information, respectively, to pre-train node representations by using graph convolutional networks. And then fine-tune them using an existing general representation-based recommendation model.