Modal-aware Bias Constrained Contrastive Learning for Multimodal Recommendation

Wei Yang
Institute of Automation, Chinese Academy of Sciences
Beijing, China
weiyangvia@gmail.com

Zhengru Fang
Department of Computer Science, City University of Hong Kong
Hong Kong, China
zhefang4@cityu.edu.hk

Tianle Zhang
Institute of Automation, Chinese Academy of Sciences
Beijing, China
zhangtianle2018@ia.ac.cn

Shiguang Wu*
Institute of Automation, Chinese Academy of Sciences
Beijing, China
shiguangwuvia@gmail.com

Chi Lu
Kuaishou Technology
Beijing, China
chiluccas@gmail.com

ABSTRACT

Multimodal recommendation system has been widely used in short video platform, e-commerce platform and news media. Multimodal data contains information such as product image and product text, which is often used as auxiliary signal to improve the effect of recommendation system significantly. In order to alleviate the problems of data sparsity and noise, some researchers construct data augmentation to use self-supervised learning to help model training. These methods have achieved certain results. However, most of the work is based on data augmentation in random ways, such as random masking and random perturbation. This random method is likely to lose important information and introduce new noise, resulting in biased augmentation data. Therefore, we propose a Modal-aware Bias Constrained Contrastive Learning method (BCCL) to solve the above problems. Specifically, BCCL introduces a bias-constrained data augmentation method to ensure the quality of augmentation samples. Then the multi-modal semantic information is modeled by the designed modal awareness module. Furthermore, we propose a information alignment module to improve the sparse modal feature learning of the model. We conducted a comprehensive experiment on three real-world data sets, and the experimental results showed that the proposed BCCL outperformed all the state-of-art methods. In-depth experiments have verified the effectiveness of our proposed modules.

CCS CONCEPTS

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

*Corresponding author.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612568

KEYWORDS

Bias Constraint, Contrastive Learning, Multimodal Recommendation

ACM Reference Format:
Wei Yang, Zhengru Fang, Tianle Zhang, Shiguang Wu, and Chi Lu. 2023. Modal-aware Bias Constrained Contrastive Learning for Multimodal Recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3612568

1 INTRODUCTION

With the rapid development of the Internet, multimodal recommendation system has been widely used in short video platforms, e-commerce platforms and news media [4, 9, 19]. Compared with single attribute information, multi-modal information, including image, text and voice, conveys richer feature information. In addition, there are both consistent semantic information and modal specific semantic information between different modalities. These information reflects the user's general preferences, as well as the personalized modal preferences [1, 10, 43].

Many researchers begin to study the effective use of multimodal information to make accurate recommendations. The dominant approach is the supervised learning paradigm based on multimodal information. Some research work [3, 5, 25] first encodes multimodal data and then maps these representations to a unified semantic space via nonlinear converters. The common way to utilize multi-mode information is to connect with attribute feature as side-info signal, and learn the prediction target through feature cross model such as NFM [12]. VBPR [11] is an early multimodal recommendation approach that incorporates visual features extracted from product images into matrix factorization to reveal the visual preference. In addition, recent approaches utilize graph neural networks to fuse multimodal information and explore the connectivity of user and item interactions [20, 31, 47, 48].

However, excellent multimodal recommendation results often require rich interaction data, which is insufficient in real scenarios. In order to achieve accurate recommendations with limited user interaction, some researchers use self-supervised learning to provide supervisory signals by means of data augmentation [15, 32, 51]. Self-supervised learning enhances the embedding representation

learning by means of contrastive task, thus alleviating the problem of data sparsity. Many researches based on graph neural networks have achieved good results by using self-supervised learning. These studies have proposed various graph-based data enhancement methods, including node dropping, edge perturbation, attribute masking and so on. MMGCL [46] proposes a multi-modal representation learning framework, which aims to explicitly enhance multi-modal representation learning. As an important part of contrast learning, data augmentation determines the upper bound of representation learning effect. Accurate and effective data augmentation methods can improve information gain, otherwise the introduction of fake samples and a lot of noise will seriously damage model learning.

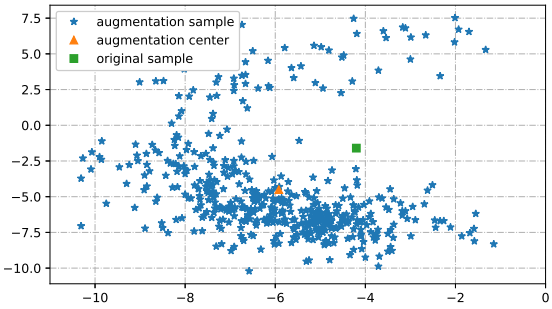

Figure 1: The distribution of node embeddings on Tiktok is generated by 500 times random masking and permutation graph augmentation. We use two dimensional vectors for representation to facilitate visualization.

Although the existing self-supervised learning methods have been widely used in multimodal recommendation systems, there are still some problems. First, much of the work is based on random masking and replacement of data augmentation, which is not always effective. Recent studies have pointed out that random data augmentation is problematic because noisy information can impair task learning [41, 52]. As shown in Fig. 1, although the center of the augmentation samples obtained by random masking is close to the original sample, it is clearly biased. Especially in some local areas, the bias is particularly obvious. Since the augmentation samples determine the information gain, the embedding learned based on this biased signal is also biased. Secondly, the general data augmentation methods are often modal-independent [16, 26], ignoring the use of complementary and mutually exclusive information between multiple modes. In addition, due to the significant differences in the amount of information contained in the data of different modalities, it is difficult for the model to fairly learn the important information of each modality. There are often distribution differences and a large amount of noise information interference between different modal data. Inaccurate modal representation can easily damage the training of the model.

Considering several important problems mentioned above, we propose a new solution. We propose a Modal-aware Bias Constrained Contrastive Learninig multimodal recommendation method (BCCL). Specifically, considering the harmful effect of biased data augmentation on model training, we propose a bias constrained method to reduce the bias of data augmentation. This method ensures that the model learns from a sample distribution with low bias and low variance. Then, considering the relationship of multimodal information, we design an intra- and inter-modal attentive extractor to learn modal-aware semantic information. Furthermore, considering the sparsity of the distribution of effective features in different modalities, we introduce the cross modal information alignment loss function to optimize the learning of modal representation. Finally, the model is trained through the joint optimization of multiple targets.

The main contributions can be summarized as follows:

- We propose a bias constraint method to reduce the bias of data augmentation, so as to help improve the accuracy of multimodal representation learning.
- We propose a modal-aware bias constrained contrastive learninig recommendation method. BCCL uses an intra- and inter-modal attention module to learn modal-aware semantic information and strengthens sparse modal representation learning based on cross modal information alignment.
- We conducted comprehensive experiments on three real data sets, and our proposed BCCL model outperformed the state-of-the-art methods. Further experiments verified the effectiveness of our designed modules.

2 RELATED WORK

2.1 Multimodal Recommendation

A lot of research work is exploring the effective use of multimodal information to help recommendation systems improve performance [9, 43, 44, 48]. Some early research methods mainly add multimodal data as auxiliary information to recommendation models [3, 5]. VBPR [11] incorporates visual features extracted from product images into matrix decomposition to reveal the visual dimensions that most affect people's behavior. VECF [3] uses the VGG module to obtain the pre-segmented image region, and uses the attention mechanism to capture user preferences in the image region. ACF [2] uses the attention awareness of the item layer and the component layer to handle recommendation tasks in the multimedia domain.

In recent years, some researches using graph neural network have been widely carried out [48, 53]. MMGCN [39] not only takes full advantage of multimodal dependencies, but also uses speaker information to model dependencies between speakers. Based on bipartite graph and user co-occurrence graph, DualGNN [33] uses the correlation between users to learn user preferences. GRCN [38] is designed to solve the implicit feedback problem. It adaptively adjusts the structure of the interaction graph according to the state of model training. MGAT [27] transmits information in a single graph, and uses the gated attention mechanism to identify the different importance scores of different patterns on user preferences.

In addition, there are some specific optimization studies to improve the effectiveness of multimodal recommendation systems. SCAHN [17] designs a semantic structure-enhanced contrastive adversarial hash network to enhance representational learning. InvRL [6] learns invariant representations to make consistent predictions of user-item interactions in a variety of environments. TopicVAE

[7] introduces topic to multimodal recommendation, and designs a decoupling presentation learning module based on VAE.

2.2 Self-supervised Learning for Recommendation

Recently, self-supervised contrast learning has been widely used in recommendation systems [15, 32, 51]. Many studies have used data augmentation combined with contrast tasks to enhance model training. CL4SRec [42] can learn more accurate user representation and reduce the problem of data sparsity simply by enhancing user interaction. $S^3 - Rec$ [54] uses mutual information maximization to fully mine the association between items and sequences.

MCPTR [21] proposes a novel contrast loss function to make different modal representations of the same item semantically similar. GHMFC [32] constructs two contrast learning modules using entity embedding representation derived from graph neural networks. These two contrast loss functions represent text-to-image and image-to-text directions respectively. Cross-CBR [22] designs a contrast learning loss based on bundle view and item view to learn graph representation. While, MICRO [51] focuses on modeling shared modal information and specific modal information.

In addition, some methods design rich ways to build samples for contrastive learning. MML [24] enhances the data set by constructing a subset of user history purchase sequences. LHBPMR [23] uses graph convolution method to select items with similar preferences to construct positive samples. MMGCL [46] designs modal edge loss and modal masking to construct positive samples. Based on two graph augmentation methods, node dropping and edge dropping, CGI [35] introduces information bottleneck contrastive learning to optimize model training. GCA [55] has constructed self-supervised contrastive learning objectives based on topology level augmentation and node attribute level augmentation. From another perspective, QRec [48] adds uniform noise to multimodal information to construct positive samples, thus improving the generalization ability of the model.

3 PRELIMINARIES

Given the user set $\mathcal{U} = \{u\}$ and the item set $\mathcal{I} = \{i\}$. Let $\mathcal{G} = \{(u, i)|u \in \mathcal{U}, i \in \mathcal{I}\}$ represent the bipartite user-item interaction graph in the recommendation system. $|\mathcal{U}|$ and $|\mathcal{I}|$ represent the sizes of the user and item nodes. We define the multimodal feature set of each item as $F^M = \{f^v, f^a, f^t\}$, where f^v represents the visual features, f^a represents the acoustic features and f^t represents the textual features. We specify a multimodal recommendation system that captures user-item relationships through modal awareness. Given the interaction graph $\mathcal{G}$ and multimodal feature set F^M, the task of our multimodal recommendation is to accurately predict whether an item will be interacted with by the user.

4 THE PROPOSED BCCL MODEL

In this section, we introduce the proposed BCCL model in detail. The overall architecture of the model is shown in Fig. 2. Firstly, bias constrained data augmentation is designed to construct high-quality samples with low bias, so as to provide accurate sample information for contrastive learning. Secondly, intra- and inter-modal attentive extractor is proposed to capture effective modal semantic information from multi-modal data. Thirdly, multi-modal contrastive learning object is proposed to assist model training. Finally, sparse adaptive enhancement module is proposed to help the model to enhance the multimodal feature learning. The model is optimized jointly under the synthesis of multiple objectives.

4.1 Bias-constrained Data Augmentation

To improve representation learning through self-supervised contrast learning, it is often necessary to construct positive and negative samples through data enhancement. Many methods have designed data augmentation methods based on interaction graphs for node and edge masking, which have achieved good results. The improvement of model effect largely depends on the quality of constructed samples. Inspired by [52], we consider constructing high-quality samples to address the limitations of the simple data enhancement method mentioned above.

Let Z denote the original representation matrix and $\hat{Z}$ denote the corresponding augmentation representation matrix. Each embedding $\hat{z}_i \in \hat{Z}$ is obtained by performing a random transfer of the original representation. Assuming that the transfer function is $f_t(\cdot)$, then we can get the following representation of the bias:

$$\Delta_{f_t(z_i)} = ||E_{\hat{z}_i \sim f_t(z_i)}(\hat{z}_i) - z_i||_2 \tag{1}$$

According to the law of large numbers, when the number of samples approaches infinity, the mean of the samples will approach the overall expected value. For simplicity, we let K represent the number of enhancements and approximate the expected value using the mean of the generated samples. So the bias of data augmentation can be expressed as:

$$\Delta_{f_t(z_i)} = ||\frac{1}{K}\sum_{k=1}^{K}\hat{z}_i^k - z_i||_2 \tag{2}$$

Then we randomly generated 500 times data augmentation by performing performing node attribute masking and edge perturbation in the original graph. We further utilize the encoder to learn a two-dimensional representation for visualization based on node embedding. The distribution of augmentation samples is shown in Fig. 1.

It can be seen that there are obvious bias in some local locations of the augmentation samples. The bias can be expressed as $\Delta_{f_t(z_i)} >> \xi$, where ξ represents the average difference between the constructed sample and the real distribution, which is theoretically a value close to 0. It is clear that the directly randomly generated samples belong to a biased set. The goal of contrast learning is to maximize the similarity between positive sample pairs and minimize the similarity between negative sample pairs. In order to ensure the effect of model learning, it is necessary to ensure that the augmentation samples are unbiased. Therefore, we need to explore a new data augmentation method to provide augmentation samples with less bias.

Considering the adverse effects of biased samples on model learning, we propose a new bias constrained data augmentation method. Suppose that the original representation matrix is $Z \in R^{n \times d}$, and the augmentation representation matrix is $\hat{Z} \in R^{n \times d}$. Then $\hat{Z}$ and Z need to satisfy a constraint as follows:

$$||Z^T Z - \hat{Z}^T \hat{Z}||_2 \leqslant \gamma Tr(Z^T Z) \tag{3}$$

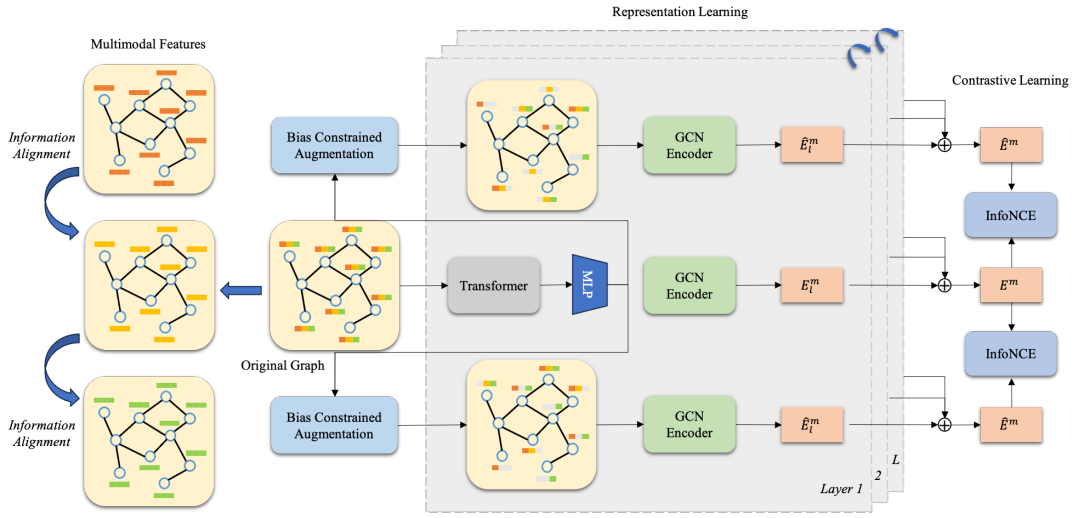

Figure 2: The components and architecture of the proposed BCCL model.

where Tr represents the trace of the matrix. γ is a threshold used to control the bias degree of $||Z^T Z - \hat{Z}^T \hat{Z}||_2$. $\hat{Z}$ is obtained from Z based on the transfer function f_t, which can be simply represented as:

$$\hat{Z} = f_t(Z, S, \psi) \tag{4}$$

where $S \in R^{n \times d}$ denotes the transfer matrix, and $\psi \in R^{n \times d}$ indicates Gaussian noise. Inspire by [21], we can regard the above augmentation process as matrix sketching. The obtained augmentation matrix $\hat{Z}$ requires a good sketch of Z, making the second-order statistics of the original and sketched matrices are similar.

Theoretically, we can do data augmentation for visual modality, acoustic mode and text modality respectively, so as to alleviate the learning difficulties caused by data sparsity. Therefore, the multimodal bias constraint loss function is expressed as follows:

$$L_c = \sum_{m \in M} max(||Z^{mT} Z^m - \hat{Z}^{mT} \hat{Z}^m||_2, \gamma Tr(Z^{mT} Z^m)) \tag{5}$$

where m represents different modalities, including visual modal v, acoustic modal a and text modal t. In practice, the sketch matrix S can be various transfer matrices such as random masking. However, considering that the key information between different modes is influential, we further introduce the modal importance perceptron to learn the sketch matrix.

4.2 Intra- and Inter-modal Attentive Extractor

Not all visual factors play the same role in attracting users, who may only care about the collar style of the product. Not all the text introduction is concerned by users, users may only care about the product style in the text introduction. The actual phenomenon shows that the user's perception of modal information is selective. Randomly masking parts of the original feature is a direct way to avoid the adverse effects of insignificant content. However, this can obscure key aspects of modal data, leading to inaccuracies in learning representations. In addition, the retention encoding of noise information and trivial information will also affect the learned representation to some extent.

In order to learn cross modal information for effective feature masking, we use transformer[28] to learn global semantic representation. Transformer is composed of stacked multi head attention modules and has been proven effective through extensive research. Specifically, we add a $[CLS]$ token vector in front of the representation sequence to learn global semantic information before entering the modal feature vector into the transformer encoder. Relevant studies [28, 45] have pointed out that $[CLS]$ token can learn the semantic information of the whole feature sequence under attention perception. Therefore, we can obtain the multimodal global semantic representation as follows:

$$H = Transformer([x_{cls}; x^v; x^a; x^t; x^f]) \tag{6}$$

where $H \in R^{5 \times k}$ represents the output of the encoder, and the vector corresponding to the $[CLS]$ token is represented as $h_{cls} \in R^{1 \times k}$. $x^m \in R^{1 \times k}$ denotes the modal representation. Specifically, since different modal representations are usually located in different feature spaces, we will follow relevant research to map them separately into the id feature space before modal information fusion.

Due to the implicit relationship between different modes, it is difficult to explicitly observe or define the roles of each mode. Therefore, we use neural networks to learn the probability of attribute masking based on multimodal global semantic representation and specific modal representation.For the nodes in the single modal graph composed of text attribute features, we can obtain the masking weight representation as follows:

$$\begin{aligned} r^t =& \pi * (sigmoid(MLP_1([x^t, x^t_{\mathcal{N}}]))) \\ &+ (1 - \pi) * (sigmoid(MLP_2([x^t, h_{cls}])) \end{aligned} \tag{7}$$

where $\pi \in [0, 1]$ denotes the hyperparameter used to control the weight between intra modal and cross modal information. MLP represents a multi-layer fully connected neural network. $x^t_{\mathcal{N}} = f_{agg}(\mathcal{N}(x^t))$ refers to the aggregated representation obtained based

on the neighborhood information of specific item. $\mathcal{N}$ indicates the set of neighbors of the specific item node in the graph. f_{agg} represents an aggregation function, which can be represented by operators such as addition and averaging.

Considering that some popular nodes are more likely to affect the training of the model, leading to biased learning. We set the hyperparameter p_n to control the overall masking probability. The probability of whether to mask nodes in the text attribute graph can be expressed as $w^t = r^t \cdot p_n$. For more granular node dimension modeling, p_n can be measured based on the degree centrality of each node in the graph.

Following [35, 55], we regard the masking matrix as one in which every element obeys the Bernoulli distribution. For the i-th node, the masking probability from the perspective of text modality can be expressed as s_i^t, which satisfies the condition of $s_i^t \backsim Bernoulli(w_i^t)$. Let W^t represent the masking parameter matrix of all nodes in the text modality. The masking matrix can be obtained according to the Bernoulli distribution as follows:

$$S^t = Sample(Bernoulli(W^t)) \tag{8}$$

where *Sample* represents the sampling function that performs sampling according to the Bernoulli distribution. The values of each element in the masking matrix can be represented as follows:

$$S_{i,j}^t = \begin{cases} 0 & \text{if } W_{i,j}^t < \eta, \\ 1 & \text{if } W_{i,j}^t \geq \eta. \end{cases} \tag{9}$$

where η indicates the probability value generated randomly. Similarly, we can obtain the visual mode masking matrix S^v and the acoustic mode masking matrix S^a respectively.

4.3 Multi-modal Contrastive Learning

In order to alleviate model learning difficulties caused by sparse data and data noise in multi-modal scenarios, self-supervised contrast learning can be adopted to improve the learning of multi-modal representation. By minimizing the distance between anchor and positive samples and maximizing the distance between anchor and negative samples, representation learning can be more effective. Based on data augmentation with bias limitation and modal information perception, Formula 4 can be rewritten as:

$$\hat{X}^m = S^m X^m + \psi^m \tag{10}$$

where $\hat{X}^m$ represent augmentation feature matrix, and S^m denotes the constructed masking matrix under modality m. Further, we can get semantic representation Z^m and $\hat{Z}^m$ based on graph encoder.

Following [35, 36], we adopt the InfoNCE [8] loss function to maximizing the lower bound of the mutual information. For mutual information on the user side, we believe that the representations of the same user in the original graph and the augmentation graph should be regarded as positive sample pairs, while the representations between two different users should be regarded as negative sample pairs. Therefore, the multimodal contrastive loss function is defined as follows:

$$L_{cl}^u = -\sum_{m \in M}^{|M|} \sum_{u \in \mathcal{U}}^{|\mathcal{U}|} log \frac{exp(Z_u \cdot \hat{Z}_u / \tau)}{\sum_{v \in \mathcal{U}} exp(Z_u \cdot \hat{Z}_v / \tau)} \tag{11}$$

where τ represents the temperature coefficient, which determines the degree of concern of contrast loss to difficult negative samples. For mutual information on the item side, we can define the contrastive loss as L_{cl}^i. We add the contrastive loss functions on both sides of user and item together as the total multimodal contrast loss function L_{cl}.

4.4 Sparse Adaptive Enhancement

The quality of multimodal data often determines the effectiveness of model learning. In fact, the key feature information in each single modal data is often sparse, which is not conducive to model learning. Especially in multimodal environments, the fusion of multiple sparse modal features increases the difficulty of model training. We've noticed that some of the research [29] in the field of computer vision has encountered similar problems. They proposed two important concepts of alignment restriction and consistency restriction to achieve similar goals. It is difficult to ensure that the feature vector of an item in each modality is a dense representation of the key information. In addition, modal data often contains a large amount of noise information, and the fusion of multiple modal noise information can easily damage the training of the model.

Many researches [30, 34] in recommendation system mainly use attention mechanism to learn the importance of different modalities, while others construct high-order intersections between different modal features. Instead, we consider introducing self-supervised cross modal alignment objectives to optimize the learning of multi-modal feature representation and minimize the distance between different modal features in low-dimensional space. We hope to introduce cross modal signal to realize cross-modal knowledge transfer and reduce the interference of noisy information. The cross modal alignment loss function is defined as follows:

$$L_a = \sum_{i=1}^{|M|-1} \sum_{j=i+1}^{|M|} g_s(W_i x_i, W_j x_j) \tag{12}$$

where x_i represents the i-th modal feature, and M denotes the multi-modal feature set as defined above. $W_i \in R^{d \times k}$ represents the mapping function used to map the feature embedding x_i into a low-dimensional space. g_s represents the similarity function. Consistent with related research work [49], we adopt cosine similarity to eliminate differences in feature distribution.

4.5 Prediction and Joint Training

Due to our design of a sample enhancement method with bias constraints, the comparison between the augmentation graph and the original graph can improve the learning effectiveness of the model. The design of cross modal alignment loss further improves the accuracy of multimodal representation. In theory, we can use any complex graph neural network as the encoder. For simplicity, we use MMGCN as the basic encoder to obtain high-level semantic representations. Based on the encoded user representation z_u and the item representation z_i of the i-th sample, we can obtain the interaction probability as:

$$\hat{y}_i = g(z_u, z_i) \tag{13}$$

where g represents the similarity function, which is replaced by the inner product. Let $y_i \in \{0, 1\}$ represent the real label of the sample.

The cross entropy loss function can be expressed as:

$$L_p = -\frac{1}{N}\sum_{i=1}^{n}[y_i log r_i + (1-y_i)log(1-r_i)] \tag{14}$$

Further, we combine bias constraint loss, multimodal contrastive loss and cross modal alignment loss for optimization. The final comprehensive loss function is expressed as follows:

$$L = L_p + \alpha L_c + \beta L_a + \mu L_{cl} \tag{15}$$

where α, β and γ represent hyperparameters, which are used to control the effects of different loss functions. By optimizing the fusion loss function, the model can achieve better performance.

5 EXPERIMENTS

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

- **RQ1** How does our BCCL model perform compared to the state-of-the-art methods?
- **RQ2** How do different designs (e.g., bias constrained module, modal attentive extractor, sparse enhancement module) affect the performance of the BCCL model?
- **RQ3** What effect does the bias constraint module bring?
- **RQ4** How do the hyperparameter settings in the model affect the model performance?

5.1 Experimental Settings

5.1.1 Datasets. We chose three real-world public data sets for the experiment, including Tiktok, Amazon-Baby, and Amazon-Sports. These three experimental data sets are also the baseline data sets used in many studies. Statistical results of the three data sets are shown in Table 1.

- **TikTok.** The data was collected from the tiktok platform, one of the world's most popular short video apps. Multimodal features include visual features, acoustic features and text features. Text embedding is encoded with Sentence-Bert.
- **Amazon.** Amazon dataset [14] contains data from Amazon e-commerce platform. We select Amazon-baby and Amazon-sports as experimental data. Based on the text extracted from product title, description, brand and classification information, text feature embedding is generated by Sentence-Bert. The 4096-dimensional product visual feature embedding is generated based on the product image.

5.1.2 Evaluation Metrics. In this paper, we mainly focus on the problem of interaction prediction, which is to effectively learn multi-modal features and attribute features to predict the probability that a user will interact with an item. Following [23, 36], we adopt three widely used indicator to evaluate top-K recommended on the accuracy of the results. The evaluation measures included Recall@K (R@K), Precision@K (P@K) and Normalized Discounted Cumulative Gain (N@K).

5.1.3 Baselines. To evaluate performance, we compare the proposed BCCL with self-supervise-based and multimodal-based recommendation models.

- **VBPR** [11] incorporates visual features extracted from product images into matrix factorization to reveal the visual dimensions that most affect people's behavior.
- **LightGCN** [13] abandons the nonlinear activation function and feature transformation process in GCN, and retains the operation of the most core aggregate neighbor node.
- **MMGCN** [39] uses the information propagation of modal-aware binary user-item graph to obtain a better user representation based on item content information.
- **GRCN** [38] can generate a detailed user-item interaction diagram for the convolution operation of the graph. The model identifies the false positive feedback in the interaction diagram and removes the corresponding noise edge.
- **SGL** [40] uses different data augmentation operators such as random node and random walk to build a contrastive learning view based on graph collaborative filtering.
- **NCL** [18] generates constrastive views by identifying semantic and structural adjacent nodes based on EM-based clustering to generate positive contrast pairs.
- **LATTICE** [50] designs a pattern-aware graph structure learning layer to learn item graph structure from multi-modal features and integrate multi-modal graphs.
- **CLCRec** [37] alleviates the cold start problem of items by contrast learning. The model builds two valid contrastive targets based on users and interacting items.
- **MMGCL** [46] introduces a negative-case sampling technique that can learn correlations between modes and ensure the effective contribution of each mode.
- **SLMRec** [26] designs data augmentation methods such as feature dropout and feature masking to improve the self-supervised learning recommendation effect.
- **MMSSL** [36] designs a modal aware interactive structural learning paradigm for data enhancement through adversarial perturbation.
- **HCGCN** [23] performs high-order graph convolutions inside user-item clusters and item-item clusters to capture various patterns belong to user behavior.

5.1.4 Parameter Settings. We randomly split each dataset into training set (80%), validation set (10%) and test set (10%). The validation set is used to select model hyperparameters, and the test set is used to evaluate the effect of the model. Considering the training time and convergence speed, the learning rate is adjusted from [$1e^{-4}$, $5e^{-4}$, $1e^{-3}$, $5e^{-3}$]. The dimension of the hidden vector is set as 64. The hyperparameters α, β and μ are searched in [$1e^{-4}$, $1e^{-3}$, $1e^{-2}$, $1e^{-1}$]. The hyperparameter π is tuned from 0.1 to 0.5. The masking hyperparameter is selected between 0.6 and 1.0. The temperature coefficient τ and threshold are searched in [$5e^{-4}$, $5e^{-3}$, $5e^{-2}$, $5e^{-1}$].

5.2 Overall Performance (RQ1)

To verify the effectiveness of our proposed BCCL model, we conducted comprehensive experiments on three datasets, and the experimental results are shown in Table 2. Analyzing the experimental results, we have the following observations:

- The BCCL model outperforms all state-of-the-art methods on three data sets and has a significant improvement. The experimental results fully verify the effectiveness of our

Table 1: Statistics of the three datasets with multimodal item Visual(V), Acoustic(A), Textual(T) information.

Dataset	User	Item	Interactions	Embedding Dim	Sparsity
Tiktok	9,319	6,710	59,541	V(128), T(768), A(128)	99.904%
Amazon-Sports	35,598	18,357	256,308	V(4,096), T(1024)	99.961%
Amazon-Baby	19,445	7,050	139,110	V(4,096), T(1024)	99.899%

Table 2: Overall performance comparison on the three datasets.

Model	Amazon-Sports			Amazon-Baby			Tiktok		
	Recall@20	Percision@20	NDCG@20	Recall@20	Percision@20	NDCG@20	Recall@20	Percision@20	NDCG@20
VBPR	0.0582	0.0031	0.0265	0.0486	0.0026	0.0213	0.0380	0.0018	0.0134
LightGCN	0.0782	0.0042	0.0369	0.0698	0.0037	0.0319	0.0653	0.0033	0.0282
MMGCN	0.0638	0.0034	0.0279	0.064	0.0032	0.0284	0.0730	0.0036	0.0307
GRCN	0.0833	0.0044	0.0377	0.0754	0.0040	0.0336	0.0804	0.0036	0.0350
LATTICE	0.0915	0.0048	0.0424	0.0829	0.0044	0.0368	0.0843	0.0042	0.0367
CLCRec	0.0651	0.0035	0.0301	0.061	0.0032	0.0284	0.0621	0.0032	0.0264
MMGCL	0.0875	0.0046	0.0409	0.0758	0.0041	0.0331	0.0799	0.0037	0.0326
SGL	0.0779	0.0041	0.0361	0.0678	0.0036	0.0296	0.0603	0.0030	0.0238
NCL	0.0765	0.004	0.0349	0.0703	0.0038	0.0311	0.0658	0.0034	0.0269
SLMRec	0.0829	0.0043	0.0376	0.0765	0.0043	0.0325	0.0845	0.0042	0.0353
MMSSL	0.0998	0.0052	0.0470	0.0962	0.0051	0.0422	0.0921	0.0046	0.0392
HCGCN	0.1032	0.0055	0.0478	0.0922	0.0048	0.0415	0.0935	0.0049	0.0412
BCCL	**0.1069**	**0.0057**	**0.0498**	**0.1002**	**0.0054**	**0.0447**	**0.0980**	**0.0052**	**0.0439**
p-value	$3.25e^{-6}$	$8.63e^{-7}$	$2.59e^{-7}$	$6.82e^{-7}$	$3.74e^{-6}$	$5.45e^{-7}$	$6.71e^{-6}$	$2.48e^{-6}$	$4.25e^{-6}$
Improvement	3.62%	3.48%	4.25%	4.16%	5.37%	5.81%	4.76%	5.39%	6.55%

model. Since BCCL ensures the quality of the augmentation samples through the bias constrained module, the model can accurately learn the feature representation. In addition, the sparse enhancement module also improves the learning effect of multimodal features.

- Self-supervised learning methods such as SGL and NCL are significantly superior to VBPR and other methods that directly use multimodal representation. Compared to these contrastive learning methods, BCCL shows a much more significant improvement. We think that this is mainly dependent on the bias-constrained task that we designed. Since most methods construct enhanced samples based on random operators, our method ensures that the model is trained on samples with low bias.
- Compared with GRCN, MMGCL and other Graph-based models, our proposed BBCL also performs significantly better. Using graph neural network can indeed make full use of information, which is better than self-supervised learning methods such as NCL. However, since BCCL not only improves the quality of training samples, but also improves the training effect of different modal features, the model can predict more accurately.

5.3 Ablation Experimental Study (RQ2)

In order to study the influence of each module on the model effect, we consider conducting the following ablation experiments. (1) We remove the bias constraint (BC) loss function to explore the effectiveness of our designed data augmentation method. (2) We remove the modal attentive (MA) extractor and replace the sketch matrix with a randomly initialized matrix. (3) We directly remove the contrast learning (CL) and bias constraint loss function. (4) We remove the information alignment (IA) loss function from the optimization objective. We conducted comprehensive ablation experiments on three data sets. The experimental results are shown in Table 3. We can observe the following conclusions:

- When we remove the constrained loss function from the optimization objective, the model's effectiveness decreases on all data sets. This experimental result directly verifies the effectiveness of the constrained data augmentation proposed by us. Since the data quality determines the upper limit of the model effect, training on the low-bias augmentation samples constructed by us can help improve the model effect.
- When we take the mode-aware attention extractor away from the model, the model's effectiveness decreases. Due to the existence of certain connections between different modal features, the attention extractor can grasp the importance of different modal information more accurately.
- Without contrast learning modules, the effect of the model is significantly reduced, which directly verifies the necessity of self-supervised learning. In addition, when we remove the information alignment loss function, the performance of the model is worse. This shows that enhanced training of sparse modal features can improve the learning ability of the model.

Table 3: Ablation study on key components of BCCL.

Model	Sports		Baby		Tiktok	
	R@20	N@20	R@20	N@20	R@20	N@20
BCCL	**0.1069**	**0.0498**	**0.1002**	**0.0447**	**0.0980**	**0.0439**
w/o-BC	0.0981	0.0477	0.0869	0.0388	0.0811	0.0365
w/o-MA	0.0995	0.0490	0.0902	0.0418	0.0854	0.0402
w/o-CL	0.0773	0.0325	0.0644	0.0376	0.0711	0.0288
w/o-IA	0.0974	0.0456	0.0847	0.0378	0.0832	0.0377

5.4 Effect of Bias Constraint Module (RQ3)

As mentioned above, the random augmentation strategy is easy to produce samples with high bias, which is not conducive to model learning. The data distribution in Fig. 1 also illustrates the bias of random augmentation. In order to solve this problem, we propose a bias constraint module to constrain the generation of enhanced samples, so as to produce high-quality samples with low bias. To verify whether the bias constraint module is really effective, the same visualization is performed on the enhanced sample generated by the bias constraint, as shown in Fig. 3. We can clearly see that the distribution of augmentation samples at this time is close to the original samples. In particular, the original large bias in some local areas has been alleviated to a large extent. Combined with the ablation experiment, it can be found that the proposed bias constraint module is indeed effective.

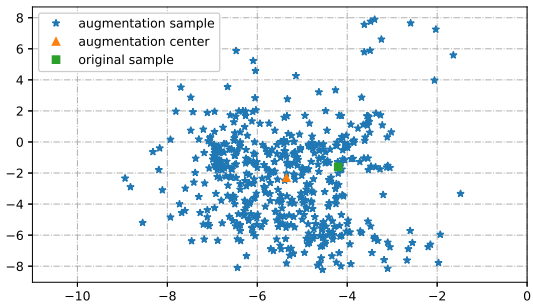

Figure 3: The distribution of node embeddings on Tiktok is generated by 500 times bias constrained graph augmentation. We use two dimensional vectors for representation to facilitate visualization.

5.5 Hyperparameter Analysis (RQ4)

In this part, we mainly analyze the sensitivity of several important parameters in the model.

- **Effect of latent dimension** d. In order to test the difference in performance of BCCL under different dimensions of embedding, d is selected from [32, 64, 128, 256, 512]. The performance across the three data sets is shown in Fig. 4. It can be seen that as the embedding dimension increases, the model will perform better on all data sets. As the embedding dimension continues to increase, the scope of effect enhancement will gradually decrease. This experimental phenomenon is consistent with other baseline models.
- **Effect of parameters** β **and** μ. Since the weights β and μ control the effects of information alignment loss and contrast loss respectively, we adjust them in [0.0001, 0.001, 0.01, 0.1, 1]. As shown in Fig. 5, models with relatively small parameter values perform better. We think it is likely that the large weight is easy to make the model training unstable. In general, β and μ are important parameters for the model, which can be adjusted flexibly to improve the model effect.

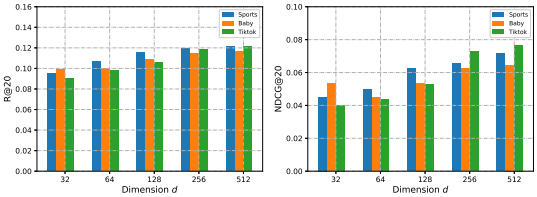

Figure 4: Impact study of hyperparameter d

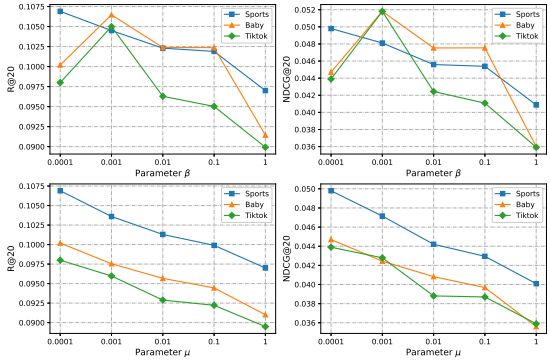

Figure 5: Impact study of hyperparameters β **and** μ

6 CONCLUSION

In this paper, we propose a modal-aware bias constrained contrastive learning method to improve the effect of multimodal recommendation systems. BCCL is composed of bias constraint module, modal awareness module and sparse enhancement module. The bias constraint module is used for data augmentation, so as to construct high-quality samples with low bias. Comprehensive experiments on multiple real-world data sets fully validate the effect of BCCL.

In the future, we will first explore multimodal interest modeling for users. Users often have different interests and preferences for information in different modalities, and the role of this multimodal interest should be explored at a finer granularity. Then we will explore more effective self-supervised and unsupervised learning methods such as diffusion and adversarial learning.

REFERENCES

[1] Yukuo Cen, Jianwei Zhang, Xu Zou, Chang Zhou, Hongxia Yang, and Jie Tang. 2020. Controllable multi-interest framework for recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2942–2951.

[2] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*. 335–344.

[3] Xu Chen, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2019. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 765–774.

[4] Yashar Deldjoo, Markus Schedl, Paolo Cremonesi, and Gabriella Pasi. 2020. Recommender systems leveraging multimedia content. *ACM Computing Surveys (CSUR)* 53, 5 (2020), 1–38.

[5] Yashar Deldjoo, Markus Schedl, and Peter Knees. 2021. Content-driven music recommendation: Evolution, state of the art, and challenges. *arXiv preprint arXiv:2107.11803* (2021).

[6] Xiaoyu Du, Zike Wu, Fuli Feng, Xiangnan He, and Jinhui Tang. 2022. Invariant Representation Learning for Multimedia Recommendation. In *Proceedings of the 30th ACM International Conference on Multimedia*. 619–628.

[7] Zhiqiang Guo, Guohui Li, Jianjun Li, and Huaicong Chen. 2022. TopicVAE: Topic-aware Disentanglement Representation Learning for Enhanced Recommendation. In *Proceedings of the 30th ACM International Conference on Multimedia*. 511–520.

[8] Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 297–304.

[9] Tengyue Han, Pengfei Wang, Shaozhang Niu, and Chenliang Li. 2022. Modality matches modality: Pretraining modality-disentangled item representations for recommendation. In *Proceedings of the ACM Web Conference 2022*. 2058–2066.

[10] Li He, Hongxu Chen, Dingxian Wang, Shoaib Jameel, Philip Yu, and Guandong Xu. 2021. Click-through rate prediction with multi-modal hypergraphs. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 690–699.

[11] 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.

[12] Xiangnan He and Tat-Seng Chua. 2017. Neural factorization machines for sparse predictive analytics. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*. 355–364.

[13] 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.

[14] Himabindu Lakkaraju, Julian McAuley, and Jure Leskovec. 2013. What's in a name? understanding the interplay between titles, content, and communities in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 7.

[15] Chenyi Lei, Shixian Luo, Yong Liu, Wanggui He, Jiamang Wang, Guoxin Wang, Haihong Tang, Chunyan Miao, and Houqiang Li. 2021. Understanding chinese video and language via contrastive multimodal pre-training. In *Proceedings of the 29th ACM International Conference on Multimedia*. 2567–2576.

[16] Xuewei Li, Aitong Sun, Mankun Zhao, Jian Yu, Kun Zhu, Di Jin, Mei Yu, and Ruiguo Yu. 2023. Multi-Intention Oriented Contrastive Learning for Sequential Recommendation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 411–419.

[17] Meiyu Liang, Junping Du, Xiaowen Cao, Yang Yu, Kangkang Lu, Zhe Xue, and Min Zhang. 2022. Semantic Structure Enhanced Contrastive Adversarial Hash Network for Cross-media Representation Learning. In *Proceedings of the 30th ACM International Conference on Multimedia*. 277–285.

[18] Zihan Lin, Changxin Tian, Yupeng Hou, and Wayne Xin Zhao. 2022. Improving graph collaborative filtering with neighborhood-enriched contrastive learning. In *Proceedings of the ACM Web Conference 2022*. 2320–2329.

[19] Shang Liu, Zhenzhong Chen, Hongyi Liu, and Xinghai Hu. 2019. User-video co-attention network for personalized micro-video recommendation. In *The World Wide Web Conference*. 3020–3026.

[20] Yixin Liu, Ming Jin, Shirui Pan, Chuan Zhou, Yu Zheng, Feng Xia, and S Yu Philip. 2022. Graph self-supervised learning: A survey. *IEEE Transactions on Knowledge and Data Engineering* 35, 6 (2022), 5879–5900.

[21] 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*. 99–108.

[22] Yunshan Ma, Yingzhi He, An Zhang, Xiang Wang, and Tat-Seng Chua. 2022. CrossCBR: Cross-view Contrastive Learning for Bundle Recommendation. *arXiv preprint arXiv:2206.00242* (2022).

[23] Zongshen Mu, Yueting Zhuang, Jie Tan, Jun Xiao, and Siliang Tang. 2022. Learning Hybrid Behavior Patterns for Multimedia Recommendation. In *Proceedings of the 30th ACM International Conference on Multimedia*. 376–384.

[24] Xingyu Pan, Yushuo Chen, Changxin Tian, Zihan Lin, Jinpeng Wang, He Hu, and Wayne Xin Zhao. 2022. Multimodal Meta-Learning for Cold-Start Sequential Recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 3421–3430.

[25] 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.

[26] Zhulin Tao, Xiaohao Liu, Yewei Xia, Xiang Wang, Lifang Yang, Xianglin Huang, and Tat-Seng Chua. 2022. Self-supervised learning for multimedia recommendation. *IEEE Transactions on Multimedia* (2022).

[27] Zhulin Tao, Yinwei Wei, Xiang Wang, Xiangnan He, Xianglin Huang, and Tat-Seng Chua. 2020. Mgat: Multimodal graph attention network for recommendation. *Information Processing & Management* 57, 5 (2020), 102277.

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

[29] Feng Wang and Huaping Liu. 2021. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2495–2504.

[30] Fangye Wang, Yingxu Wang, Dongsheng Li, Hansu Gu, Tun Lu, Peng Zhang, and Ning Gu. 2023. CL4CTR: A Contrastive Learning Framework for CTR Prediction. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 805–813.

[31] Hongwei Wang, Fuzheng Zhang, Miao Zhao, Wenjie Li, Xing Xie, and Minyi Guo. 2019. Multi-task feature learning for knowledge graph enhanced recommendation. In *The world wide web conference*. 2000–2010.

[32] Peng Wang, Jiangheng Wu, and Xiaohang Chen. 2022. Multimodal Entity Linking with Gated Hierarchical Fusion and Contrastive Training. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 938–948.

[33] Qifan Wang, Yinwei Wei, Jianhua Yin, Jianlong Wu, Xuemeng Song, and Liqiang Nie. 2021. Dualgnn: Dual graph neural network for multimedia recommendation. *IEEE Transactions on Multimedia* (2021).

[34] Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*. PMLR, 9929–9939.

[35] Chunyu Wei, Jian Liang, Di Liu, and Fei Wang. 2022. Contrastive Graph Structure Learning via Information Bottleneck for Recommendation. *Advances in Neural Information Processing Systems* 35 (2022), 20407–20420.

[36] Wei Wei, Chao Huang, Lianghao Xia, and Chuxu Zhang. 2023. Multi-Modal Self-Supervised Learning for Recommendation. In *Proceedings of the ACM Web Conference 2023*. 790–800.

[37] Yinwei Wei, Xiang Wang, Qi Li, Liqiang Nie, Yan Li, Xuanping Li, and Tat-Seng Chua. 2021. Contrastive learning for cold-start recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5382–5390.

[38] 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.

[39] 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.

[40] Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. 2021. Self-supervised graph learning for recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 726–735.

[41] Lianghao Xia, Chao Huang, Yong Xu, Jiashu Zhao, Dawei Yin, and Jimmy Huang. 2022. Hypergraph contrastive collaborative filtering. In *Proceedings of the 45th International ACM SIGIR conference on research and development in information retrieval*. 70–79.

[42] Xu Xie, Fei Sun, Zhaoyang Liu, Shiwen Wu, Jinyang Gao, Jiandong Zhang, Bolin Ding, and Bin Cui. 2022. Contrastive learning for sequential recommendation. In *2022 IEEE 38th international conference on data engineering (ICDE)*. IEEE, 1259–1273.

[43] Cai Xu, Ziyu Guan, Wei Zhao, Quanzhou Wu, Meng Yan, Long Chen, and Qiguang Miao. 2020. Recommendation by users' multimodal preferences for smart city applications. *IEEE Transactions on Industrial Informatics* 17, 6 (2020), 4197–4205.

[44] Wei Yang, Tengfei Huo, Zhiqiang Liu, and Chi Lu. 2023. based Multi-intention Contrastive Learning for Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2339–2343.

[45] Dong Yao, Zhou Zhao, Shengyu Zhang, Jieming Zhu, Yudong Zhu, Rui Zhang, and Xiuqiang He. 2022. Contrastive Learning with Positive-Negative Frame Mask for Music Representation. In *Proceedings of the ACM Web Conference 2022*. 2906–2915.

[46] Zixuan Yi, Xi Wang, Iadh Ounis, and Craig Macdonald. 2022. Multi-modal graph contrastive learning for micro-video recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1807–1811.

[47] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. Graph contrastive learning with augmentations. *Advances in neural information processing systems* 33 (2020), 5812–5823.

[48] Junliang Yu, Hongzhi Yin, Xin Xia, Tong Chen, Lizhen Cui, and Quoc Viet Hung Nguyen. 2022. Are graph augmentations necessary? simple graph contrastive learning for recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1294–1303.

[49] Junliang Yu, Hongzhi Yin, Xin Xia, Tong Chen, Jundong Li, and Zi Huang. 2023. Self-supervised learning for recommender systems: A survey. *IEEE Transactions on Knowledge and Data Engineering* (2023).

[50] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. 2021. Mining latent structures for multimedia recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3872–3880.

[51] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Mengqi Zhang, Shu Wu, and Liang Wang. 2022. Latent Structure Mining with Contrastive Modality Fusion for Multimedia Recommendation. *IEEE Transactions on Knowledge and Data Engineering* (2022).

[52] Yifei Zhang, Hao Zhu, Zixing Song, Piotr Koniusz, and Irwin King. 2022. COSTA: Covariance-Preserving Feature Augmentation for Graph Contrastive Learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2524–2534.

[53] Feng Zhao and Donglin Wang. 2021. Multimodal graph meta contrastive learning. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 3657–3661.

[54] Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 2020. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 1893–1902.

[55] Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. 2021. Graph contrastive learning with adaptive augmentation. In *Proceedings of the Web Conference 2021*. 2069–2080.