mathematics

MDPI

Article

A Multimodal Graph Recommendation Method Based on Cross-Attention Fusion

Kai Li †, Long Xu †, Cheng Zhu and Kunlun Zhang *

National Key Laboratory of Information Systems Engineering, National University of Defense Technology, Changsha 410003, China; likai18@nudt.edu.cn (K.L.); xulong@nudt.edu.cn (L.X.); zhucheng@nudt.edu.cn (C.Z.)
* Correspondence: zhangkunlun@nudt.edu.cn
† These authors contributed equally to this work.

Abstract: Research on recommendation methods using multimodal graph information presents a significant challenge within the realm of information services. Prior studies in this area have lacked precision in the purification and denoising of multimodal information and have insufficiently explored fusion methods. We introduce a multimodal graph recommendation approach leveraging cross-attention fusion. This model enhances and purifies multimodal information by embedding the IDs of items and their corresponding interactive users, thereby optimizing the utilization of such information. To facilitate better integration, we propose a cross-attention mechanism-based multimodal information fusion method, which effectively processes and merges related and differential information across modalities. Experimental results on three public datasets indicated that our model performed exceptionally well, demonstrating its efficacy in leveraging multimodal information.

Keywords: multimodal graph; recommendation method; multimodal information purification; cross-attention mechanism; information fusion

MSC: 68T05

Citation: Li, K.; Xu, L.; Zhu, C.; Zhang, K. A Multimodal Graph Recommendation Method Based on Cross-Attention Fusion. *Mathematics* **2024**, *12*, 2353. https://doi.org/10.3390/math12152353

Academic Editor: Giuseppe Pirrò

Received: 27 June 2024
Revised: 23 July 2024
Accepted: 26 July 2024
Published: 28 July 2024

1. Introduction

As the storage and retrieval of multimodal information becomes increasingly easy and standardized, and with the continuous improvement in computational power, multimodal graph networks have emerged as a popular research direction in the field of information recommendation. Information recommendation tasks aim to predict information or products that users might be interested in by analyzing their historical behavior, preferences, and contextual data. This is crucial for various online service platforms, including e-commerce, news aggregation, and social media. Although traditional recommendation systems, such as collaborative filtering and content-based methods, have shown promising results, they primarily rely on explicit user feedback (e.g., ratings) or text content analysis, often neglecting multimodal information within user interactions, such as images, audio, and video content.

Despite the advancements in traditional recommendation systems like collaborative filtering and content-based methods, these approaches demonstrate limitations in effectively incorporating multimodal information within user interactions. For instance, while collaborative filtering relies heavily on explicit user feedback (e.g., ratings), it often fails to capture the nuances inherent in user-generated content such as images and videos. This was noted in research by Cinar et al., which highlighted the inadequacies of user feedback mechanisms [1]. Furthermore, traditional systems tend to overlook how multimodal signals can enhance user personalization, as discussed in the research by Lei et al. [2].

Recently, with the rapid advancement of machine learning and deep learning technologies, researchers have started exploring how to leverage the rich multimodal data in user-generated content to improve recommendation system performance. For instance,

by integrating convolutional neural networks for image processing, recurrent neural networks for sequential text processing, and self-attention mechanisms to better capture users' long-term and short-term interests. However, effectively fusing and utilizing multimodal information remains a challenging problem, involving complex tasks such as feature extraction, representation learning, and cross-modal semantic mapping of different data types.

Multimodal information contains a significant amount of noise that is irrelevant to user interests, such as background environments in images and irrelevant expressions in text. Current multimodal recommendation methods for data denoising and purification often rely on cross-domain self-supervised learning or generative adversarial networks to capture correlations between different modalities, exemplified by the MMBT model [3]. However, these methods tend to be highly dependent on large datasets and require substantial computational resources. In the realm of multimodal information fusion, the prevailing approaches typically employ simple linear parallel or series combinations to fuse different modalities with uniform or predefined weights [4–6]. This method has significant limitations, as it does not consider the diverse preferences of users, where individuals may prioritize different modalities based on personal preferences. Research documented in Hu et al. emphasized the necessity for adaptive fusion strategies to accommodate user-specific inclinations [7].

To address the highlighted issues, we present a multimodal graph recommendation model based on cross-attention fusion. Initially, we utilize a multimodal information purification and denoising method, enhanced by item and user interaction ID features, to purify and augment the image and text features corresponding to the items. Subsequently, these purified features undergo further processing and fusion through a multimodal feature fusion module employing a cross-attention mechanism.

We conducted experiments on three publicly available benchmark datasets. The results of our experiments indicated that our model surpassed the performance of existing baseline methods.

In summary, the contributions of this paper are as follows:

- We introduce a multimodal information purification and denoising approach based on item and user interaction ID features, aimed at exploring more effective ways to enhance multimodal information.
- We utilize a multimodal feature fusion module based on a cross-attention mechanism, allowing for more comprehensive and efficient processing and fusion of multimodal information.
- Our model's performance was compared against baseline methods in the field of multimodal recommendation on three public benchmark datasets, demonstrating superior performance over existing baseline models.

2. Related Work

In this section, we briefly review three types of research relevant to our work, namely multimodal graph recommendation, multimodal information denoising, and cross-attention mechanisms.

2.1. Multimodal Graph Recommendation

In recent years, the advancement of data acquisition technologies and the enhancement of computational power have led to an evolution from simple data fusion techniques to complex deep learning models capable of processing and analyzing rich multimodal data. Subsequently, we will delve into the latest research achievements in the two primary methods of multimodal graph representation learning: multimodal graph recommendation based on graph embeddings, and multimodal graph recommendation based on graph neural networks.

Traditional graph recommendation methods based on graph embeddings encompass various approaches, such as graph factorization, graph distribution representation, and graph neural embeddings. With the widespread use of multimodal data, many re-

searchers have integrated these multimodal information types into graph representation features through various feature extraction methods and aggregation techniques. He et al. proposed an extended factorization model, VBPR [8], which significantly improved the accuracy of personalized ranking methods by learning an additional layer to reveal the visual dimensions that best explain the changes in user feedback using visual features extracted from product images via deep networks. Tang et al. proposed the AMR model [9], which further improved VBPR by enhancing the robustness of the model through adversarial learning.

In the context of multimodal graph data, graph neural networks can integrate these different modalities of information to learn richer representations by performing feature fusion and propagation on the nodes and edges within the graph. Multimodal graph recommendation methods based on graph neural networks are generally classified according to the fusion methods of different modalities of information, such as direct fusion, heterogeneous graph fusion, and homogeneous graph fusion. Liu et al. proposed the PMGT [10] strategy, which uses two objectives, graph structure reconstruction and masked node feature reconstruction, to pre-train and effectively utilize multimodal information, thereby improving the accuracy of recommendation systems and click-through rate prediction. Sun et al. proposed the MKGAT [11] model, which processes information in multimodal knowledge graphs through multimodal graph attention techniques and utilizes aggregated embedding representations for recommendation, successfully enhancing the quality of feature extraction.

Although the aforementioned methods have successfully applied multimodal information in the recommendation field, there is still room for improvement in aspects such as the denoising and fusion of this information. Due to the differences in feature distribution among different modalities, we proposed a new multimodal information denoising and fusion module and achieved promising performance results.

2.2. Multimodal Information Denoising

Given the significant noise present in real-world datasets, constructing a high-quality multimodal dataset is fraught with challenges and unresolved issues. This noise can be broadly categorized into two types: single-modal noise, which stems from errors in acquisition devices, environmental background influences, or compression during transmission; and cross-modal semantic noise, which arises from poor alignment of distributions during the fusion of different modalities.

For the first category of noise, numerous mature and effective methods have been developed. One straightforward approach is to denoise multimodal data using mean fusion. Rajalingam et al. [12] decomposed images into low-frequency and high-frequency components, applying mean fusion to the low-frequency parts and full fusion to the high-frequency parts to reduce noise. However, since the degree of noise varies across different modalities, weighted averaging methods have also been introduced. Xue et al. [13] proposed a gating function that adaptively fuses multimodal data features, effectively bypassing certain noisy instance paths and mitigating the impact of noise. Another approach is joint optimization denoising, which addresses specific optimization problems using data from different modalities. Wang et al. [14] achieved excellent results by using variational models in the pixel and wavelet domains to fuse and denoise multi-focus images. Quan et al. [15] proposed the relative total variation structure analysis method (RTVSA), which integrates various features obtained from HIS and LiDAR data, effectively mitigating the impact of noise on the model.

Most multimodal tasks require training with aligned data across modalities; however, real-world data are generally weakly aligned or unaligned. Such training data can be viewed as cross-modal noise, a type of semantic-level noise. Some researchers filtered noise from a data perspective. Radenovic et al. [16] proposed the CAT filtering strategy, reducing the impact of cross-modal noise by extracting highly important data samples. Other researchers focused on model correction to identify noise. Li et al. [17] proposed

the BLIP framework, which uses a subtitler to generate subtitles and a filter to remove noise, effectively enhancing the data quality of training subtitle samples and excelling in downstream tasks. Noise regularization is another method used to reduce the impact of cross-modal noise. Huang et al. [18] proposed a pre-training framework, NLIP, which employs noise-adaptive regularization to improve and enhance weakly aligned cross-modal data.

In summary, the research on denoising multimodal information is relatively mature. However, in the recommendation domain, methods using ID embeddings of users and items for denoising still need exploration. Constraining multimodal information through attribute embedding not only effectively denoises but also focuses attention on parts with stronger user preferences, enhancing data features.

2.3. Cross-Attention-Based Fusion Methods

Cross-attention is a widely-utilized method in multimodal models for integrating information from different modalities [19–21]. It has found extensive application in computer vision and natural language processing and has demonstrated its efficacy in information fusion within the transformer architecture.

In the domain of image vision, numerous researchers have employed cross-attention to merge various types of image features. Kim et al. [22] utilized a cross-attention module to adaptively fuse local features with global dependencies, effectively preserving the unique characteristics of each modality. Tang et al. [23] introduced an image registration fusion method, SuperFusion, which leverages a cross-attention mechanism to blend complementary information from source images, thereby avoiding artifacts caused by misalignment of different image types. Xie et al. [24] enhanced the complementary representation of semantic information through a cross-attention approach and mitigated the interference noise caused by overlapping edges of aligned images.

Contemporary fusion methods effectively integrate the transformer architecture with cross-attention mechanisms to enhance model representation. Ma et al. [25] employed self-attention to fuse intra-domain information and cross-attention to combine inter-domain information, thereby fully integrating complementary information across domains. Jha et al. [26] proposed a global attention fusion network (GAF-Net), which amalgamates self-attention and cross-attention modules, preserving their distinct characteristics and demonstrating the superiority of this approach.

Despite the significant advancements of the cross-attention mechanism in these tasks, the exploitation of differential and common information between modalities remains underexplored. Consequently, this paper designed a cross-attention module to fuse cross-domain differential and common information.

3. Formulation of CAmgr

We propose a multimodal graph recommendation model based on cross-attention fusion. Figure 1 illustrates an overview of the model. Initially, we extract features from images using VGG-16 and from text using Sentence2Vec. Subsequently, we enhance and refine the multimodal information leveraging item and user ID features. Next, we construct semantic edges between items of different modalities and use a graph neural network encoder to enhance the structural information of these modal embeddings along with user-item ID embeddings. The modal features and ID features are then entered into the multimodal feature fusion module, resulting in the final fused representation. This representation is ultimately used in downstream information recommendation tasks via a prediction layer.

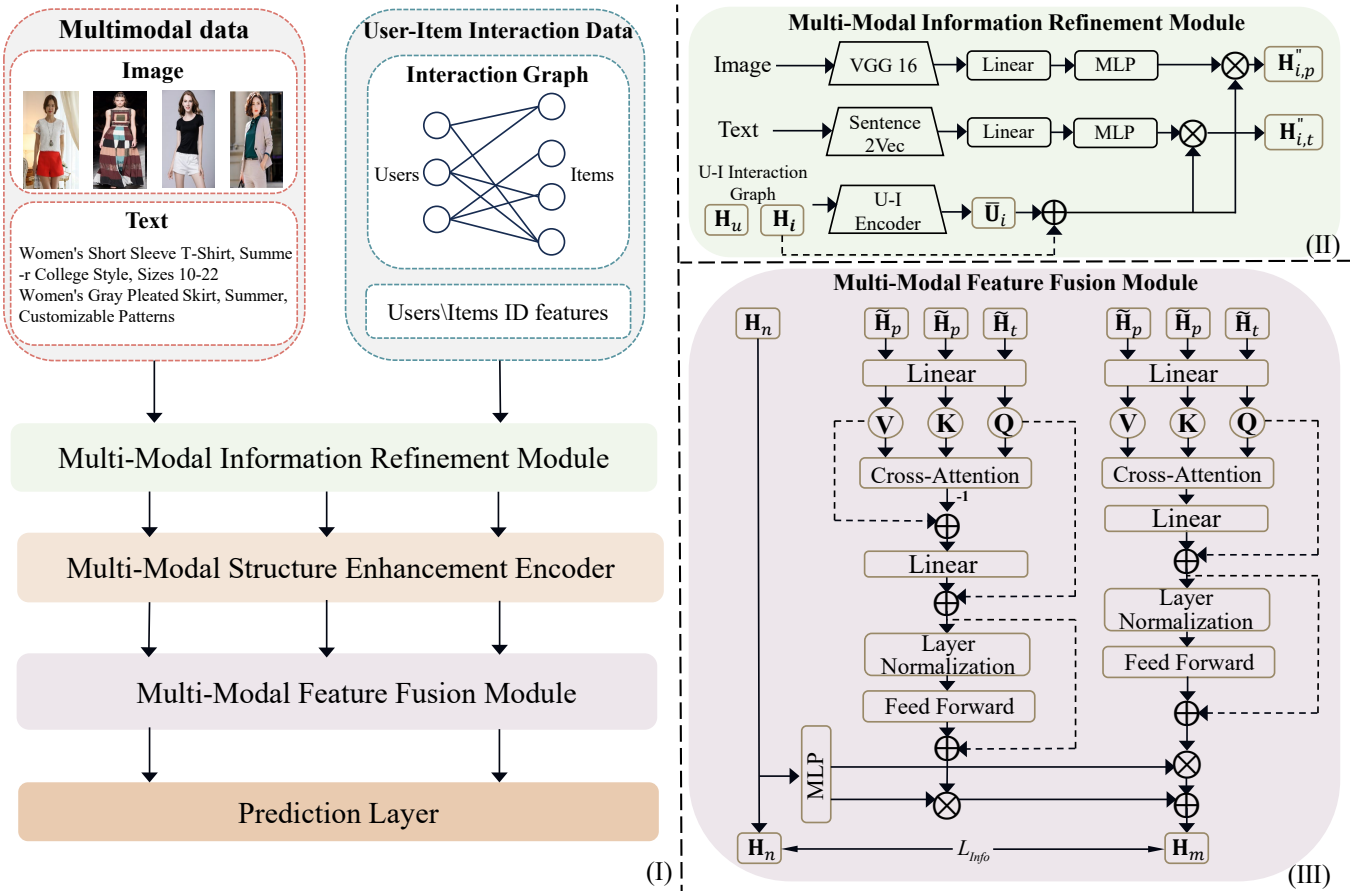

Figure 1. (**I**) The overall framework. (**II**) The multimodal information refinement method enhanced by item and interaction user ID features. (**III**) The multimodal feature fusion module based on cross-attention mechanism.

3.1. Problem Definition

Initially, let $\boldsymbol{U}$ denote the set of all users and $\boldsymbol{I}$ denote the set of all items. For any user $u \in \boldsymbol{U}$ and any item $i \in \boldsymbol{I}$, their ID embeddings are denoted as $\boldsymbol{H}_u \in \mathbb{R}^{d \times |\boldsymbol{U}|}$ and $\boldsymbol{H}_i \in \mathbb{R}^{d \times |\boldsymbol{I}|}$, where d is the embedding dimension. We then introduce the concept of feature embeddings for different modalities, focusing on image and text modalities in this study. For each item, the image and text modality features are represented as $\boldsymbol{H}_{i,p} \in \mathbb{R}^{d_p \times |\boldsymbol{I}|}$ and $\boldsymbol{H}_{i,t} \in \mathbb{R}^{d_t \times |\boldsymbol{I}|}$, with d_p and d_t being the respective embedding dimensions.

Subsequently, we define the historical interaction matrix $\boldsymbol{A} \in \mathbb{R}^{|\boldsymbol{U}| \times |\boldsymbol{I}|}$ between users and items, where $a_{mn} \in \{0, 1\}$ indicates whether user u_m has interacted with item i_n. If such an interaction has occurred, $a_{mn} = 1$; otherwise, $a_{mn} = 0$. The aim of the recommendation task is to predict the preference scores $\hat{y}_{ui}$ for various items and generate a recommendation list for each user that maximizes their satisfaction.

3.2. Multimodal Information Refinement Method Enhanced by Item and Interaction User ID Features

Although modality information contains many meaningful item feature details, it also includes a significant amount of useless noise that can mislead the model. To enhance the purification of this information and reduce its negative impact on the model, we propose a multi-modal information refinement method based on item and interaction user ID features. Specifically, we first encode the image and text information of items using VGG-16 and

Sentence2Vec models, respectively, to obtain $\boldsymbol{H}_{i,p}$ and $\boldsymbol{H}_{i,t}$. Then, we input these encodings into a linear layer for alignment, resulting in $\boldsymbol{H}_{i,p}^{'}$ and $\boldsymbol{H}_{i,t}^{'}$:

$$\boldsymbol{H}_{i,p}^{'} = \boldsymbol{W}_{1,p}\boldsymbol{H}_{i,p} + \boldsymbol{b}_{1,p}, \boldsymbol{H}_{i,t}^{'} = \boldsymbol{W}_{1,t}\boldsymbol{H}_{i,t} + \boldsymbol{b}_{1,t} \tag{1}$$

where $\boldsymbol{W}_{1,p} \in \mathbb{R}^{d \times d_p}$, $\boldsymbol{W}_{1,t} \in \mathbb{R}^{d \times d_t}$, and $\boldsymbol{b}_{1,p}, \boldsymbol{b}_{1,t} \in \mathbb{R}^{d}$ are trainable weight matrices and biases.

Next, we process the ID embeddings of items and the users who have interacted with them. For each item i, we derive a corresponding user interaction embedding matrix $\boldsymbol{U}_i \in \mathbb{R}^{d \times |\boldsymbol{I}|}$ through a historical interaction matrix:

$$\boldsymbol{U}_i = \boldsymbol{H}_u\boldsymbol{A} \tag{2}$$

Every column of $\boldsymbol{U}_i$ represents the sum of the embedding vectors of all users who interacted with item i. To normalize, we average each column by the number of users who interacted with item i. Let $\boldsymbol{B} \in \mathbb{R}^{1 \times |\boldsymbol{U}|}$ be an all-ones matrix, then we obtain the user count for each item $\boldsymbol{N}_u \in \mathbb{R}^{1 \times |\boldsymbol{I}|}$:

$$\boldsymbol{N}_u = \boldsymbol{B}\boldsymbol{A} \tag{3}$$

Thus, we can compute the average user embedding matrix for each item $\bar{\boldsymbol{U}}_i \in \mathbb{R}^{d \times |\boldsymbol{I}|}$:

$$\bar{\boldsymbol{U}}_i = \frac{1}{\boldsymbol{n}_q \cdot \boldsymbol{u}_{zq}} \tag{4}$$

where $\boldsymbol{n}_q \in \boldsymbol{N}_u$, $\boldsymbol{u}_{zq} \in \boldsymbol{U}_i$, $z = 1, \ldots, d$, $q = 1, \ldots, |\boldsymbol{I}|$.

Finally, utilizing the combined effect of item and historical interaction user ID embeddings, we refine and enhance the image and text modality information:

$$\boldsymbol{H}_{i,p}^{''} = (\boldsymbol{H}_i + \bar{\boldsymbol{U}}_i) \odot \sigma(\boldsymbol{W}_{2,p}\boldsymbol{H}_{i,p}^{'} + \boldsymbol{b}_{2,p}) \tag{5}$$

$$\boldsymbol{H}_{i,t}^{''} = (\boldsymbol{H}_i + \bar{\boldsymbol{U}}_i) \odot \sigma(\boldsymbol{W}_{2,t}\boldsymbol{H}_{i,t}^{'} + \boldsymbol{b}_{2,t}) \tag{6}$$

Here, $\boldsymbol{H}_{i,p}^{''}$ and $\boldsymbol{H}_{i,t}^{''}$ are the refined image and text embeddings; $\odot$ denotes the Hadamard product, indicating element-wise multiplication of two matrices; σ is the Sigmoid activation function; and $\boldsymbol{W}_{2,p}$, $\boldsymbol{W}_{2,t}$, $\boldsymbol{b}_{2,p}$, and $\boldsymbol{b}_{2,t}$ are trainable weight matrices and biases.

3.3. Multimodal Structure Enhanced Encoder

From previous research [27–29], it is evident that incorporating historical interactions between users and items, as well as semantic relatedness between items, can significantly enhance the performance of recommendation models. Therefore, we propose to further enhance the structural information in extracted features using a multi-modal structure-enhanced encoder.

To extract interaction information between users and items, we employ the GAT method to directly extract ID embedding features from the interaction graph:

$$\boldsymbol{H}_n^l = \sigma(\textstyle\sum_{i \in \boldsymbol{\mathcal{N}}(\boldsymbol{n})} \boldsymbol{\alpha}_{ni}^l \boldsymbol{W}^l \boldsymbol{H}_i^{l-1}) \tag{7}$$

In this bipartite graph, both users and items are represented as nodes. Hence, we use $\boldsymbol{H}_n^l$ to denote the embedding representation of node n in layer l. Here, $\boldsymbol{\mathcal{N}}(\boldsymbol{n})$ represents the set of neighboring nodes, $\boldsymbol{\alpha}_{ni}^l$ is the attention coefficient between a node and its neighbors, $\boldsymbol{W}^l$ is the weight matrix for layer l, and σ is the ReLU activation function. The attention coefficients are computed as follows, and enhance the image and text modality information:

$$\boldsymbol{\alpha}_{ni}^l = \frac{\exp\left(LeakyReLU\left(\boldsymbol{\omega}^\top \left[\boldsymbol{W}^l \boldsymbol{H}_u^{l-1} \middle| \boldsymbol{W}^l \boldsymbol{H}_i^{l-1}\right]\right)\right)}{\sum_{j \in \boldsymbol{\mathcal{N}}(\boldsymbol{n})} \exp\left(LeakyReLU\left(\boldsymbol{\omega}^\top \left[\boldsymbol{W}^l \boldsymbol{H}_u^{l-1} \middle| \boldsymbol{W}^l \boldsymbol{H}_j^{l-1}\right]\right)\right)} \tag{8}$$

The ID embeddings of items and users exchange information across layers, resulting in the final embedding representation $\boldsymbol{H}_n$ after l layers of GAT:

$$\boldsymbol{H}_n = \frac{1}{L+1}\sum_{l=0}^{L}\boldsymbol{H}_n^l \tag{9}$$

To extract multi-modal semantic structural information, we first construct a multi-modal semantic graph based on the similarity of item embeddings. We calculate the similarity matrix $\boldsymbol{S}_*$ using cosine similarity:

$$s_{f,g}^{*} = \frac{\left(\boldsymbol{h}_f^{*}\right)^{\top}\boldsymbol{h}_g^{*}}{\|\boldsymbol{h}_f^{*}\|\|\boldsymbol{h}_g^{*}\|} \tag{10}$$

where $*$ represents either the image modality p or text modality t, and $s_{f,g}^{*}$ represents the cosine similarity between the embeddings of item f and item g.

We then perform a *top-K* operation, retaining only the *top-K* similarity values and setting the rest to zero, thus obtaining the top-K similarity matrix $\boldsymbol{S}_*^{'}$:

$$s_{f,g}^{\prime *} = \begin{cases} s_{f,g}^{*}, s_{f,g}^{*} \in top-K\left(\left[s_{f,i}^{*}\right]\right), i \in \boldsymbol{I} \\ 0, \text{otherwise} \end{cases} \tag{11}$$

Here, $s_{f,g}^{\prime *}$ represents the retained semantic edge weights. Next, we apply Laplacian normalization to the similarity matrix:

$$\boldsymbol{S}_*^{''} = \boldsymbol{D}_*^{-1}\boldsymbol{S}_*^{'} \tag{12}$$

where $\boldsymbol{D}_*$ is the diagonal matrix of $\boldsymbol{S}_*^{'}$. We use a single-layer graph convolutional network to avoid introducing excessive multi-modal noise, yielding the final multi-modal item information embedding representation $\tilde{\boldsymbol{H}}_{i,*}$:

$$\tilde{\boldsymbol{H}}_{i,*} = \boldsymbol{S}_*^{''}\boldsymbol{H}_{i,*}^{''} \tag{13}$$

Finally, we aggregate and normalize the multi-modal item embeddings to obtain the user modality representation $\tilde{\boldsymbol{H}}_{u,*}$:

$$\tilde{\boldsymbol{h}}_{u,*} = \frac{1}{|\boldsymbol{\mathcal{N}}(n)|}\sum_{i\in\boldsymbol{\mathcal{N}}(n)}\tilde{\boldsymbol{h}}_{i,*} \tag{14}$$

By concatenating the user and item modality representations along the row dimension, we obtain $\tilde{\boldsymbol{H}}_* = \left[\tilde{\boldsymbol{H}}_{u,*}\middle|\tilde{\boldsymbol{H}}_{i,*}\right] \in \mathbb{R}^{d\times(|\boldsymbol{U}|+|\boldsymbol{I}|)}$.

3.4. The Multimodal Feature Fusion Module Based on Cross-Attention Mechanism

To effectively utilize the combined item feature information from various modalities and historical collaborative interaction information, we designed a multi-modal feature fusion module based on a cross-attention mechanism. Additionally, we employed a self-supervised learning method to maximize the mutual information between the fused multi-modal information and the user-item ID information, thereby leveraging user preferences to guide the model training [30–33].

Each modality's features contain both distinct and common information. Distinct information refers to the unique characteristics and details that are specific to each modality, which are not found in other modalities. Common information, on the other hand, refers to the shared or overlapping features that are present across multiple modalities. First, for the

common information part, we use a linear projection layer to convert the multi-modal features into the corresponding $\boldsymbol{K}$, $\boldsymbol{Q}$, and $\boldsymbol{V}$:

$$\boldsymbol{K} = Linear_K(\tilde{\boldsymbol{H}}_p) \tag{15}$$

$$\boldsymbol{Q} = Linear_Q(\tilde{\boldsymbol{H}}_t) \tag{16}$$

$$\boldsymbol{V} = Linear_V(\tilde{\boldsymbol{H}}_p) \tag{17}$$

where $Linear(\cdot)$ denotes the linear projection function.

Next, we illustrate the structure of our module for extracting common information between modalities. We employ a cross-attention mechanism to extract information from different modalities. To integrate the obtained common information into the modal information, we first pass it through a linear layer and then add it to $\boldsymbol{Q}$. Finally, through layer normalization, a multi-layer perceptron, and residual connections, we obtain the final embedded representation of the common modal information $\boldsymbol{H}_{com}$:

$$\boldsymbol{CA} = softmax\left(\frac{\boldsymbol{QK}^\top}{\sqrt{d_k}}\right) \tag{18}$$

$$\boldsymbol{CR} = Linear(\boldsymbol{CA}) + \boldsymbol{Q} \tag{19}$$

$$\boldsymbol{H}_{com} = MLP(LN(\boldsymbol{CR})) + \boldsymbol{CR} \tag{20}$$

Next, we analyze how to extract the distinct information of modalities, with a structure similar to the previous one. After passing through the cross-attention module, we first remove the common information from the modalities to obtain the distinct information, and then proceed with subsequent operations:

$$\boldsymbol{DA} = softmax\left(\frac{\boldsymbol{QK}^\top}{\sqrt{d_k}}\right) \tag{21}$$

$$\boldsymbol{DR} = Linear(\boldsymbol{V} - \boldsymbol{DA}) + \boldsymbol{Q} \tag{22}$$

$$\boldsymbol{H}_{dif} = MLP(LN(\boldsymbol{DR})) + \boldsymbol{DR} \tag{23}$$

Subsequently, the distinct and common information is further enhanced through the user and item ID embeddings:

$$\boldsymbol{H}_{n,com} = \sigma(\boldsymbol{W}_{2,com}\boldsymbol{H}_n + \boldsymbol{b}_{2,com}) \tag{24}$$

$$\boldsymbol{H}_{n,dif} = \sigma(\boldsymbol{W}_{2,dif}\boldsymbol{H}_n + \boldsymbol{b}_{2,dif}) \tag{25}$$

where $\boldsymbol{H}_{n,com}$ and $\boldsymbol{H}_{n,dif}$ represent the user's preference-extracted features for common and distinct modal information, respectively. $\boldsymbol{W}_{2,com}, \boldsymbol{W}_{2,dif} \in \mathbb{R}^{d\times d}$ are learnable weight matrices, $\boldsymbol{b}_{2,com}, \boldsymbol{b}_{2,dif} \in \mathbb{R}^d$ are learnable bias parameters, and σ is the Sigmoid nonlinear function.

Finally, we fuse the previously obtained features, combining the distinct and common information across modalities to obtain the final modal feature $\boldsymbol{H}_m$:

$$\boldsymbol{H}_m = \frac{1}{2}(\boldsymbol{H}_{com} \odot \boldsymbol{H}_{n,com} + \boldsymbol{H}_{dif} \odot \boldsymbol{H}_{n,dif}) \tag{26}$$

3.5. Prediction Layer

Based on the user and item embeddings $\boldsymbol{H}_n$ and modality features $\boldsymbol{H}_m$ obtained from the previous sections, we can derive the final representations of the user and item by summing these embeddings and features:

$$\boldsymbol{h}_u = \boldsymbol{h}_{u,n} + \boldsymbol{h}_{u,m}, \boldsymbol{h}_i = \boldsymbol{h}_{i,n} + \boldsymbol{h}_{i,m} \tag{27}$$

The predicted score ($\hat{y}ui$) is then calculated as the dot product of the user and item embeddings:

$$\hat{\boldsymbol{y}}_{ui} = \boldsymbol{h}_u^\top \boldsymbol{h}_i \tag{28}$$

3.6. Loss Function

In the final stage of multimodal feature fusion, we employ a self-supervised learning auxiliary training method using the InfoNCE loss [34] function to obtain L_{Info}. This loss consists of the user self-supervised loss L_u and the item self-supervised loss L_i:

$$L_{Info} = L_u + L_i \tag{29}$$

$$L_{u/i} = \sum_{u/i \in \boldsymbol{U/I}} -\log \frac{\exp(h_{u/i,m} \cdot h_{u/i,n}/\tau)}{\sum_{r/l \in \boldsymbol{U/I}} \exp(h_{r/l,m} \cdot h_{r/l,n}/\tau)} \tag{30}$$

Subsequently, we use Bayesian personalized ranking loss L_{BPR} as the main training objective and add an L2 regularization term to control model complexity, prevent overfitting, and improve generalization. The final loss function L is then given b

$$L = \alpha_1 L_{Info} + L_{BPR} + \alpha_2 \|w\|_2^2 \tag{31}$$

where α_1 and α_2 are hyperparameters that control the self-supervised loss and regularization strength, respectively, and $\|w\|_2^2$ is the squared L2 norm of the weight vector w, representing the sum of the squares of the model parameters.

4. Experiment

4.1. Research Question

To examine the performance of CAmgr, we put forward and resolved four research questions:

(RQ1) Can the proposed CAmgr model exceed other task baselines in the context of information recommendation tasks?

(RQ2) Do the innovative components within the CAmgr model make a significant contribution to its performance?

(RQ3) How does the performance of the CAmgr model fluctuate under varying hyperparameter settings?

(RQ4) How does the proposed denoising enhancement technique affect the distribution of multimodal information embeddings?

4.2. Datasets

In our work, we used three types of data from the Amazon dataset, which is commonly applied in recommendation tasks: baby, sports, and clothing.

- Baby: This dataset contains e-commerce interaction data for baby products, including multimodal data such as images and reviews.
- Sports: This dataset includes interaction data for sports products, with data spanning recent years.
- Clothing: This dataset comprises a bipartite interaction graph for various clothing items, along with reviews and interaction records from recent years.

Table 1 provides detailed statistical information about the datasets. To optimize model training, we divided each dataset into training, validation, and test sets, with a split ratio of 0.8/0.1/0.1. For the multimodal information in each dataset, we employed pre-extracted embedding features from the MGCN model [35].

Table 1. Statistics of the experimental datasets.

Datasets	Users	Items	Ratings	Density	Rating Scale
Baby	19,445	7050	160,792	0.117%	[1–5]
Sports	35,598	18,357	296,337	0.045%	[1–5]
Clothing	39,387	23,033	278,677	0.031%	[1–5]

4.3. Model Summary

We compared Camgr with nine popular baseline models in recommendation tasks. These baselines were categorized into two groups: general models that rely solely on historical interaction data (MF, LightGCN), and multimodal models that incorporate additional multimodal information (VBPR, MMGCN, GRCN, SLMRec, BM3, MICRO, and MGCN).

- MF [36]: A widely used collaborative filtering model in recommendation systems, employing matrix factorization to learn user and item representations.
- LightGCN [37]: Combines GCN with collaborative filtering, simplifying the GCN model to better suit recommendation tasks.
- VBPR [8]: Introduces visual modality information by extracting image features via CNN and integrating them with item ID embeddings for recommendations.
- MMGCN [38]: Effectively utilizes multimodal information to assist in solving ERC tasks by constructing dependencies within and across modalities.
- GRCN [39]: Investigates the impact of implicit feedback in GCN-based recommendation models and improves the user–item interaction graph structure using GAT.
- SLMRec [40]: Incorporates self-supervised learning in multimedia recommendation to capture the inherent multimodal patterns in data.
- BM3 [41]: Eliminates negative sampling in self-supervised learning to avoid introducing noisy supervision during training.
- MICRO [42]: Designs a contrastive method to fuse multimodal features, using the obtained multimodal item representations directly in collaborative filtering, for more accurate recommendations.
- MGCN [35]: Proposes using item behavior information to purify modality information and models user preferences comprehensively through a behavior-aware fusion mechanism.

4.4. Experimental Setup and Evaluation Metrics

To ensure a comprehensive and unbiased evaluation of model performance, we employed a full ranking protocol to measure the effectiveness of *top-K* recommendations. In this process, we calculated and reported the average metrics for all users in the test set, including Recall@K and NDCG@K. This method guaranteed the fairness and accuracy of the evaluation, effectively gauging the performance of recommendation systems in real-world scenarios.

We conducted 1000 iterations for all models, with an early stopping mechanism set at 20 rounds of patience. Following the settings in MICRO, Recall@20 was used as the criterion to halt training on the validation set. The model that performed best on the validation set was chosen for final testing. We utilized the Adam optimizer and configured the hyperparameters for each baseline model according to their original papers. For our CAmgr model, we adopted the Xavier initialization method with an initial dimension of 64, set the regularization coefficient in the loss function, and used a batch size of 2048. For the self-supervised task in this study, the temperature coefficient α_2 was set to 0.2. All experiments were conducted on an NVIDIA A100 40GB GPU.

5. Results and Discussion

5.1. Overall Performance

We present the performance results of the proposed multimodal graph recommendation model, CAmgr, and the baseline models in Table 2. To better illustrate the performance

of our CAmgr model, we display the results for Recall@K and NDCG@K with K set at 10 and 20.

Table 2. Experimental results on different recommendation models.

Datasets	Metrics	MF	LightGCN	VBPR	MMGCN	GRCN	SLMRec	BM3	MICRO	MGCN	CAmgr
Baby	Recall@10	0.0357	0.0479	0.0423	0.0378	0.0532	0.0540	0.0564	0.0584	0.0620	**0.0640**
	Recall@20	0.0575	0.0754	0.0663	0.0615	0.0824	0.0810	0.0883	0.0929	0.0964	**0.1056**
	NDCG@10	0.0192	0.0257	0.0223	0.0200	0.0282	0.0285	0.0301	0.0318	0.0339	**0.0382**
	NDCG@20	0.0249	0.0328	0.0284	0.0261	0.0358	0.0357	0.0383	0.0407	0.0427	**0.0437**
Sports	Recall@10	0.0432	0.0569	0.0558	0.0370	0.0559	0.0676	0.0656	0.0679	0.0729	**0.0751**
	Recall@20	0.0653	0.0864	0.0856	0.0605	0.0877	0.1017	0.0980	0.1050	0.1106	**0.1124**
	NDCG@10	0.0241	0.0311	0.0307	0.0193	0.0306	0.0374	0.0355	0.0367	0.0397	**0.0429**
	NDCG@20	0.0298	0.0387	0.0384	0.0254	0.0389	0.0462	0.0438	0.0463	0.0496	**0.0523**
Clothing	Recall@10	0.0187	0.0340	0.0280	0.0197	0.0424	0.0452	0.0421	0.0521	0.0641	**0.0695**
	Recall@20	0.0279	0.0526	0.0414	0.0328	0.0650	0.0675	0.0625	0.0772	0.0945	**0.1102**
	NDCG@10	0.0103	0.0188	0.0159	0.0101	0.0225	0.0247	0.0228	0.0283	0.0347	**0.0386**
	NDCG@20	0.0126	0.0236	0.0193	0.0135	0.0283	0.0303	0.0280	0.0347	0.0428	**0.0478**

From Table 2, it is evident that CAmgr significantly outperformed the other general and multimodal recommendation models. This indicates that our proposed method effectively utilized multimodal information. Instead of simply adding or concatenating the multimodal information, we enhance and integrate it by embedding the user and item IDs. This approach not only reduces noise but also aligns more closely with the user's focus. Subsequently, we model the interaction and semantic structure information through a multimodal structure-enhanced encoder, further enriching the feature information. Finally, we adaptively fuse the modal and attribute features using a designed cross-attention module and introduce self-supervised learning to enhance the impact of ID embeddings. Consequently, CAmgr achieved superior results across the three datasets and all metrics compared to the best baseline models.

Similarly, when incorporating multimodal auxiliary information, VBPR performed better than MF, but MMGCN was less effective than LightGCN. This could have been due to the propagation mechanism in GCN, which continuously transmits noise from the multimodal data to the user and item representations, resulting in poorer outcomes. Subsequent methods that suppress modal noise have shown gradual performance improvements. Thus, it is evident that handling noise in different modalities is crucial.

The GRCN model improves the user–item interaction graph using multimodal features, MICRO constructs a semantic auxiliary graph for items, and MGCN builds a multimodal feature purifier. These methods leverage indirectly processed multimodal features rather than directly injecting them into the model, achieving excellent performance results.

5.2. Ablation Studies

The ablation experiments in this section are divided into two parts: the first part focused on the ablation of the proposed modules, and the second part focused on the ablation of modalities. We conducted experiments on three datasets and selected Recall@20 and NDCG@20 as the metrics for the ablation experiments.

We performed ablation experiments on the model's modules to validate the effectiveness of the two methods proposed in the CAmgr model: the multimodal information purification and denoising method based on item and interacting user ID features, and the multimodal feature fusion method based on the cross-attention mechanism. Specifically, we compared CAmgr with two variants, denoted as $CAmgr_{w/o[IUD]}$ and $CAmgr_{w/o[CAFF]}$, corresponding to the multimodal information purification and denoising method based on item and interacting user ID features, and the multimodal feature fusion module based on the cross-attention mechanism, respectively. The results of CAmgr and its three variants are shown in Table 3. As seen in Table 3, our proposed CAmgr model outperformed

$CAmgr_{w/o[IUD]}$ and $CAmgr_{w/o[CAFF]}$ in all cases across the three datasets, indicating that both proposed methods could improve the accuracy of the recommendation model. Additionally, in all cases across the three datasets, removing our proposed purification and denoising method resulted in a more significant performance drop compared to removing the multimodal feature fusion module.

Table 3. Performance comparison between different variants of CAmgr.

Datasets	Modules	Recall@20	NDCG@20
Baby	CAmgr	0.1056	0.0437
	w/o IUD	0.0563	0.0188
	w/o CAFF	0.0847	0.0386
Sports	CAmgr	0.1124	0.0523
	w/o IUD	0.0612	0.0259
	w/o CAFF	0.0905	0.0341
Clothing	CAmgr	0.1102	0.0478
	w/o IUD	0.0584	0.0216
	w/o CAFF	0.0862	0.0254

To investigate the impact of different modal information inputs on the performance of the recommendation model, we conducted experiments under three different modal inputs: text information input, image information input, and a combination of text and image information inputs. As shown in Figure 2, both text and image modal inputs could effectively enhance the model's performance, but the performance improvement was more significant with image information. This indicates that effectively handling noise in multimodal information can significantly boost model performance. This may be because, in real-world scenarios, people tend to rely more on images when shopping, as visual impact is more pronounced. In contrast, textual descriptions are less direct and often contain irrelevant information, making it difficult for users to immediately find the needed information and requiring further filtering and processing of the text.

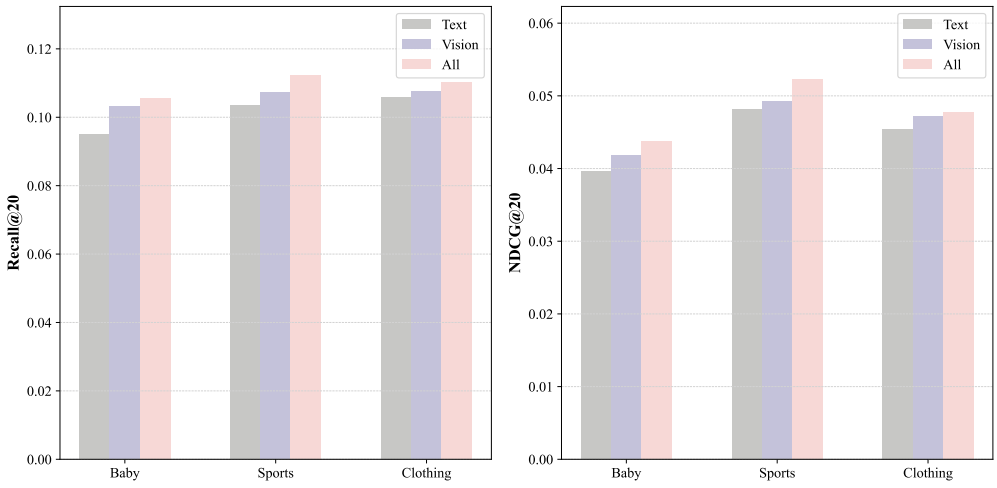

Figure 2. Performance Comparison with different modalities.

5.3. Hyperparameter Study

5.3.1. Impact of the Number of Multimodal Semantic Neighbors *k*

For the multimodal semantic neighbors of items, we selected only the top *k* features with the highest similarity to minimize the introduction of noisy neighbors. To examine the effect of neighbor quantity *k* on model performance, we adjusted the value of *k* in the proposed CAmgr, selecting {5, 10, 15, 20, and 25} as candidate values, and observed their

impact on Recall@20 and NDCG@20. As shown in the experimental results in Figure 3, the k value of 20 yielded the best results for the baby dataset, while the k value of 15 was optimal for the other datasets. These findings suggest that the number of multimodal semantic neighbors k should not be too large, as a larger value may introduce irrelevant items as neighbors, thereby increasing noise.

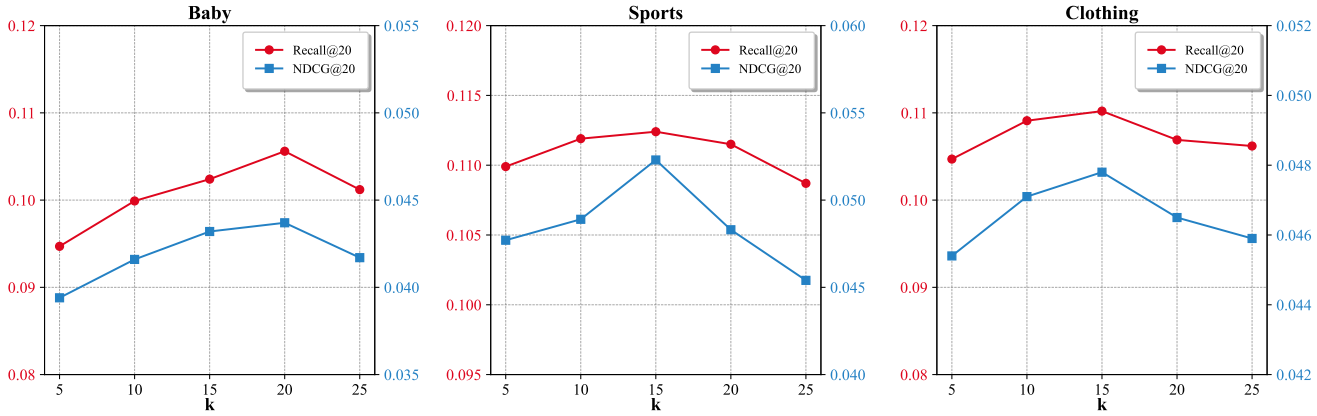

Figure 3. Performance comparison of different numbers of neighbors k.

5.3.2. Impact of the Weight of Self-Supervised Learning Loss α_1

To investigate the impact of the weight decay factor of self-supervised learning loss α_1 on model performance, we adjusted the value in the established CAmgr. The selected range for this parameter was {0.001, 0.005, 0.01, 0.05, 0.1} to examine how this hyperparameter α_1 affected the model's recommendation performance. As shown in Table 4, the value of 0.01 achieved the best performance across all three datasets, while the performance dropped sharply beyond this value. This is because self-supervised learning serves merely as an auxiliary task in our model. If the weight value α_1 is too large, the model's attention will shift towards the auxiliary task, thereby neglecting the main task, which can mislead the model.

Table 4. Performance comparison of different weights of self-supervised task α_1.

Datasets	α_1 = 0.001		α_1 = 0.005		α_1 = 0.01		α_1 = 0.05		α_1 = 0.1	
	Recall@20	NDCG@20	Recall@20	NDCG@20	Recall@20	NDCG@20	Recall@20	NDCG@20	Recall@20	NDCG@20
Baby	0.0592	0.0395	0.1035	0.0428	**0.1056**	**0.0437**	0.1010	0.0421	0.0479	0.0194
Sports	0.1006	0.0493	0.1087	0.0501	**0.1124**	**0.0523**	0.1093	0.0507	0.0373	0.0188
Clothing	0.1039	0.0447	0.1075	0.0461	**0.1102**	**0.0478**	0.0580	0.0269	0.0584	0.0263

5.4. Visual Analysis

We visually analyzed the multimodal information denoising enhancement method proposed in this paper, and visualized the modal features before denoising and the enhanced features after denoising. We randomly selected 500 items from the baby dataset, and reduced the high-dimensional modal features to a two-dimensional plane using a nonlinear dimensionality reduction method called tSNE. Then, we used two-dimensional histograms to further visualize. Figures 4 and 5 show our visual results. By analyzing the distribution of two-dimensional features, we can see that the modal information before denoising is not evenly distributed in the two-dimensional histogram, there are many outliers and many high-density regions. After noise reduction, the feature distribution is more uniform, and the color distribution of outliers and two-dimensional histograms is more balanced. By comparing FIG. 5 and FIG. 6, it can also be seen that the distribution of image features is more uniform than that of text. Therefore, in the ablation experiment of

the modal information input, it was found that the impact of image information on model performance was greater than that of text information.

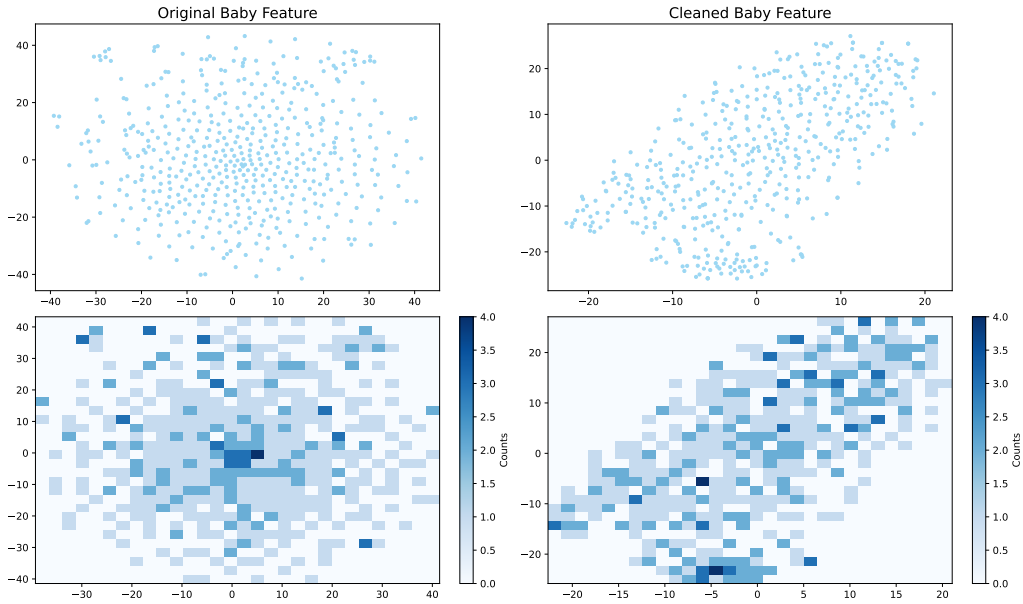

Figure 4. The distribution of representations in visual modality.

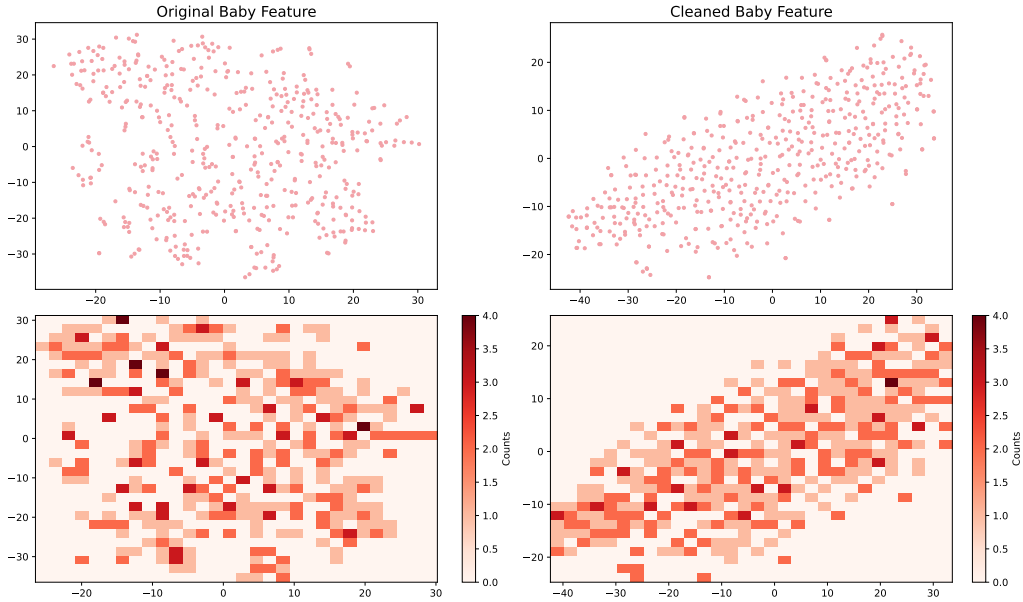

Figure 5. The distribution of representations in text modality.

6. Conclusions

In this paper, a multi-modal graph recommendation model CAmgr based on cross-attention fusion was proposed. This model significantly improves the effectiveness of multimodal information by introducing a method of purifying and denoising multimodal information based on the enhancement of items and their interactive user ID features. Through a multi-modal feature fusion module based on a cross-attention mechanism, the model can process and fuse multi-modal information more comprehensively and efficiently, so as to improve the recommendation performance. The experimental results showed that the model outperformed the existing baseline methods on three public benchmark datasets, which verified its advantages in the multi-modal recommendation task. Future work may include further optimizing the computational efficiency of the models,

as well as exploring more complex multimodal information fusion methods to address larger and more diverse datasets.

Author Contributions: Writing—original draft, K.L.; writing—review and editing, L.X.; supervision and resources, C.Z.; project administration, K.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Defense Basic Scientific Research Program (Grant No. WDZC20235250411).

Data Availability Statement: All data included in this study are available upon request through contact with the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Cinar, Y.G.; Renders, J. Adaptive Pointwise-Pairwise Learning-to-Rank for Content-based Personalized Recommendation. In Proceedings of the RecSys, Rio de Janeiro, Brazil, 25 September 2020; pp. 414–419.
2. Lei, F.; Cao, Z.; Yang, Y.; Ding, Y.; Zhang, C. Learning the User's Deeper Preferences for Multi-modal Recommendation Systems. *ACM Trans. Multim. Comput. Commun. Appl.* **2023**, *19*, 138:1–138:18. [CrossRef]
3. Serra, F.D.; Jacenków, G.; Deligianni, F.; Dalton, J.; O'Neil, A.Q. Improving Image Representations via MoCo Pre-training for Multimodal CXR Classification. In *Lecture Notes in Computer Science, Proceedings of the Medical Image Understanding and Analysis, Cambridge, UK, 27–29 July 2022*; Springer: Berlin/Heidelberg, Germany, 2022; Volume 13413, pp. 623–635.
4. Yi, J.; Chen, Z. Multi-Modal Variational Graph Auto-Encoder for Recommendation Systems. *IEEE Trans. Multim.* **2022**, *24*, 1067–1079. [CrossRef]
5. Chen, X.; Chen, H.; Xu, H.; Zhang, Y.; Cao, Y.; Qin, Z.; Zha, H. Personalized Fashion Recommendation with Visual Explanations based on Multimodal Attention Network: Towards Visually Explainable Recommendation. In Proceedings of the SIGIR, Paris, France, 21–25 July 2019; pp. 765–774.
6. Zhang, F.; Yuan, N.J.; Lian, D.; Xie, X.; Ma, W. Collaborative Knowledge Base Embedding for Recommender Systems. In Proceedings of the SIGKDD, San Francisco, CA, USA, 13–17 August 2016; pp. 353–362.
7. Hu, H.; Guo, W.; Liu, Y.; Kan, M. Adaptive Multi-Modalities Fusion in Sequential Recommendation Systems. In Proceedings of the CIKM, Birmingham, UK, 21–25 October 2023; pp. 843–853.
8. He, R.; McAuley, J.J. VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback. In Proceedings of the AAAI, Phoenix, AZ, USA, 12–17 February 2016; pp. 144–150.
9. Tang, J.; Du, X.; He, X.; Yuan, F.; Tian, Q.; Chua, T. Adversarial Training Towards Robust Multimedia Recommender System. *IEEE Trans. Knowl. Data Eng.* **2020**, *32*, 855–867. [CrossRef]
10. Liu, Y.; Yang, S.; Lei, C.; Wang, G.; Tang, H.; Zhang, J.; Sun, A.; Miao, C. Pre-training Graph Transformer with Multimodal Side Information for Recommendation. In Proceedings of the MM, Virtual, 20–24 October 2021; pp. 2853–2861.
11. Sun, R.; Cao, X.; Zhao, Y.; Wan, J.; Zhou, K.; Zhang, F.; Wang, Z.; Zheng, K. Multi-modal Knowledge Graphs for Recommender Systems. In Proceedings of the CIKM, Virtual, 19–23 October 2020; pp. 1405–1414.
12. Rajalingam, B.; Al-Turjman, F.M.; Santhoshkumar, R.; Rajesh, M. Intelligent multimodal medical image fusion with deep guided filtering. *Multim. Syst.* **2022**, *28*, 1449–1463. [CrossRef]
13. Xue, Z.; Marculescu, R. Dynamic Multimodal Fusion. In Proceedings of the CVPR, Vancouver, BC, Canada, 18–22 June 2023; pp. 2575–2584.
14. Wang, W.; Shui, P.; Feng, X. Variational Models for Fusion and Denoising of Multifocus Images. *IEEE Signal Process. Lett.* **2008**, *15*, 65–68. [CrossRef]
15. Quan, Y.; Tong, Y.; Feng, W.; Dauphin, G.; Huang, W.; Zhu, W.; Xing, M. Relative Total Variation Structure Analysis-Based Fusion Method for Hyperspectral and LiDAR Data Classification. *Remote. Sens.* **2021**, *13*, 1143. [CrossRef]
16. Radenovic, F.; Dubey, A.; Kadian, A.; Mihaylov, T.; Vandenhende, S.; Patel, Y.; Wen, Y.; Ramanathan, V.; Mahajan, D. Filtering, Distillation, and Hard Negatives for Vision-Language Pre-Training. In Proceedings of the CVPR, Vancouver, BC, Canada, 17–24 June 2023; pp. 6967–6977.
17. Li, J.; Li, D.; Xiong, C.; Hoi, S.C.H. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language understanding and Generation. In *International Conference on Machine Learning, Proceedings of the ICML, Baltimore, MD, USA, 17–23 July 2022*; Microtome Publishing: Brookline, MA, USA, 2022; Volume 162, pp. 12888–12900.
18. Huang, R.; Long, Y.; Han, J.; Xu, H.; Liang, X.; Xu, C.; Liang, X. NLIP: Noise-Robust Language-Image Pre-training. In Proceedings of the AAAI, Vancouver, BC, Canada, 20–27 February 2023; pp. 926–934.
19. Ma, J.; Zhao, J.; Jiang, J.; Zhou, H.; Guo, X. Locality Preserving Matching. *Int. J. Comput. Vis.* **2019**, *127*, 512–531. [CrossRef]
20. Zhu, H.; Ke, W.; Li, D.; Liu, J.; Tian, L.; Shan, Y. Dual Cross-Attention Learning for Fine-Grained Visual Categorization and Object Re-Identification. In Proceedings of the CVPR, New Orleans, LA, USA, 18–24 June 2022; pp. 4682–4692.

21. Praveen, R.G.; de Melo, W.C.; Ullah, N.; Aslam, H.; Zeeshan, O.; Denorme, T.; Pedersoli, M.; Koerich, A.L.; Bacon, S.; Cardinal, P.; et al. A Joint Cross-Attention Model for Audio-Visual Fusion in Dimensional Emotion Recognition. In Proceedings of the CVPR, New Orleans, LA, USA, 18–24 June 2022; pp. 2485–2494.
22. Kim, B.; Jung, H.; Sohn, K. Multi-Exposure Image Fusion Using Cross-Attention Mechanism. In Proceedings of the IEEE, Padua, Italy, 18–23 July 2022; pp. 1–6.
23. Tang, L.; Deng, Y.; Ma, Y.; Huang, J.; Ma, J. SuperFusion: A Versatile Image Registration and Fusion Network with Semantic Awareness. *IEEE CAA J. Autom. Sinica* **2022**, *9*, 2121–2137. [CrossRef]
24. Xie, H.; Zhang, Y.; Qiu, J.; Zhai, X.; Liu, X.; Yang, Y.; Zhao, S.; Luo, Y.; Zhong, J. Semantics lead all: Towards unified image registration and fusion from a semantic perspective. *Inf. Fusion* **2023**, *98*, 101835. [CrossRef]
25. Ma, J.; Tang, L.; Fan, F.; Huang, J.; Mei, X.; Ma, Y. SwinFusion: Cross-domain Long-range Learning for General Image Fusion via Swin Transformer. *IEEE CAA J. Autom. Sinica* **2022**, *9*, 1200–1217. [CrossRef]
26. Jha, A.; Bose, S.; Banerjee, B. GAF-Net: Improving the Performance of Remote Sensing Image Fusion using Novel Global Self and Cross Attention Learning. In Proceedings of the WACV, Waikoloa, HI, USA, 2–7 January 2023; pp. 6343–6352.
27. Wei, W.; Ren, X.; Tang, J.; Wang, Q.; Su, L.; Cheng, S.; Wang, J.; Yin, D.; Huang, C. LLMRec: Large Language Models with Graph Augmentation for Recommendation. In Proceedings of the WSDM, Merida, Mexico, 4–8 March 2024; pp. 806–815.
28. Wu, S.; Sun, F.; Zhang, W.; Xie, X.; Cui, B. Graph Neural Networks in Recommender Systems: A Survey. *ACM Comput. Surv.* **2023**, *55*, 97:1–97:37. [CrossRef]
29. Deldjoo, Y.; He, Z.; McAuley, J.J.; Korikov, A.; Sanner, S.; Ramisa, A.; Vidal, R.; Sathiamoorthy, M.; Kasirzadeh, A.; Milano, S. A Review of Modern Recommender Systems Using Generative Models (Gen-RecSys). *arXiv* **2024**, arXiv:2404.00579v1.
30. Zong, Y.; Aodha, O.M.; Hospedales, T.M. Self-Supervised Multimodal Learning: A Survey. *arXiv* **2023**, arXiv:2304.01008.
31. Korbar, B.; Tran, D.; Torresani, L. Cooperative Learning of Audio and Video Models from Self-Supervised Synchronization. In Proceedings of the NeurIPS, Montréal, QC, Canada, 3–8 December 2018; pp. 7774–7785.
32. Alayrac, J.; Recasens, A.; Schneider, R.; Arandjelovic, R.; Ramapuram, J.; Fauw, J.D.; Smaira, L.; Dieleman, S.; Zisserman, A. Self-Supervised MultiModal Versatile Networks. In Proceedings of the NeurIPS, Virtual, 6–12 December 2020.
33. Akbari, H.; Yuan, L.; Qian, R.; Chuang, W.; Chang, S.; Cui, Y.; Gong, B. VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text. In Proceedings of the NeurIPS, Virtual, 6–14 December 2021; pp. 24206–24221.
34. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R.B. Momentum Contrast for Unsupervised Visual Representation Learning. In Proceedings of the Computer Vision Foundation, (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 9726–9735.
35. Yu, P.; Tan, Z.; Lu, G.; Bao, B. Multi-View Graph Convolutional Network for Multimedia Recommendation. In Proceedings of the MM, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 6576–6585.
36. Koren, Y.; Bell, R.M.; Volinsky, C. Matrix Factorization Techniques for Recommender Systems. *Computer* **2009**, *42*, 30–37. [CrossRef]
37. He, X.; Deng, K.; Wang, X.; Li, Y.; Zhang, Y.; Wang, M. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In Proceedings of the SIGIR, Virtual, 25–30 July 2020; pp. 639–648.
38. Wei, Y.; Wang, X.; Nie, L.; He, X.; Hong, R.; Chua, T. MMGCN: Multi-modal Graph Convolution Network for Personalized Recommendation of Micro-video. In Proceedings of the MM, Nice, France, 21–25 October 2019; pp. 1437–1445.
39. Wei, Y.; Wang, X.; Nie, L.; He, X.; Chua, T. Graph-Refined Convolutional Network for Multimedia Recommendation with Implicit Feedback. In Proceedings of the MM, Seattle, DC, USA, 12–16 October 2020; pp. 3541–3549.
40. Tao, Z.; Liu, X.; Xia, Y.; Wang, X.; Yang, L.; Huang, X.; Chua, T. Self-Supervised Learning for Multimedia Recommendation. *IEEE Trans. Multim.* **2023**, *25*, 5107–5116. [CrossRef]
41. Zhou, X.; Zhou, H.; Liu, Y.; Zeng, Z.; Miao, C.; Wang, P.; You, Y.; Jiang, F. Bootstrap Latent Representations for Multi-modal Recommendation. In Proceedings of the WWW, Melbourne, Australia, 14–20 May 2023; pp. 845–854.
42. Zhang, J.; Zhu, Y.; Liu, Q.; Zhang, M.; Wu, S.; Wang, L. Latent Structure Mining with Contrastive Modality Fusion for Multimedia Recommendation. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 9154–9167. [CrossRef]