AM2HRec: Dual-Representation Adaptive Noise Reduction for Multi-modal Recommendation

Yang Yu, Chunna Zhang*, Shengqiang Cong, Xiaoping Yue,
Yuming Shen and Jinchi Zhao

***Abstract*—In response to the challenges of data sparsity and noise interference in multi-modal recommendation systems, this paper presents a novel multi-modal recommendation model, AM2HRec, which is based on dual-representation adaptive noise reduction techniques. The model applies noise reduction to multi-modal information that is irrelevant to user preferences by developing an adaptive decision noise reduction module. By diminishing the interference of extraneous information within multi-modal messages, the model reduces the propagation of noise at the nodes of user-item interactions, thereby preventing the contamination of the final representation. Additionally, the issue of data sparsity is addressed through the construction of both heterogeneous and homogeneous graphs for dual representation learning, which enhances the final user-item representation. To obtain information regarding each modality, AM2HRec utilizes adaptive decision-making for multi-modal fusion by assigning weights to the multi-modal data based on user preferences, thus facilitating effective integration. The experimental results demonstrate that AM2HRec outperforms existing state-of-the-art multi-modal recommendation models.**

***Index Terms*—multi-modal recommendation, dual representation learning, graph neural network, self-supervised learning**

I. Introduction

PERSONALISED recommendations [1]–[3] are essential in the context of the information explosion associated with the Internet and e-commerce. Traditional recommendation systems primarily depend on identification and classification features to connect users with items. To enhance the accuracy of these recommendation systems, there is an increasing prevalence of research focused on utilizing multi-modal information, including text, visual, and acoustic data [4], [5]. Multi-modal recommender systems utilize a variety of information—such as text, images, audio, and video related to user preferences—to generate recommendations and enhance user preference modeling, thereby increasing accuracy. Visual Bayesian Personalized Ranking (VBPR) [6] and Cross-modal Knowledge Embedding (CKE) [7] utilize multi-modal information as supplementary data to enhance recommendation performance. Furthermore, many traditional recommendation systems have begun to incorporate graph convolution networks (GCN) to capture high-order connections and improve preference features [8]–[10]. The Multi-Modal Graph Convolutional Network (MMGCN) learns modality-specific user preferences by constructing a user-item interaction graph and employing the message passing mechanism inherent in graph convolution networks to deliver more accurate recommendations. Based on MMGCN, GRCN uses multi-modal features to refine the user-item graph and prune false positive interactions. These GCN-based methods have achieved significant success and further performance improvements. To obtain better recommendation results, researchers utilize auxiliary graph structures to capture relationships between users and items. For instance, DualGNN [11] smooths the preferences of users and their neighbors through Graph Neural Networks (GNN) [12]–[14] and constructs a relationship graph among users to learn their multi-modal preferences. LATTICE [15] and FREEDOM [16] introduce item-item graphs to extract potential characteristics of items, integrate project relationships into representation learning, and improve project representation. Additionally, there is a growing trend to combine self-supervised tasks with multi-modal recommendation systems [17]–[20]. MMSSL [18] has developed a modality-aware interactive structure learning paradigm that enhances data through adversarial perturbations.

Current research in multi-modal recommendation primarily emphasizes fusion techniques, frequently neglecting the challenges associated with raw feature noise and data sparsity. This paper presents the AM2HRec model, which addresses noise interference through adaptive decision noise reduction while simultaneously enhancing user and item representations via a dual representation learning mechanism. Experiments conducted on three subsets of the Amazon Review Dataset illustrate the effectiveness of the proposed model.

II. Related Principles

A. Multimedia Recommendation

Multi-modal recommender systems derive information representations of users and items by integrating multi-modal features. As illustrated in Fig. 1, the workflow of the system includes modal feature extraction, the selection of a fusion method (Early, Intermediate, or Late fusion) for multi-modal integration [21], and subsequently employing the model to generate accurate recommendations.

Manuscript received June 25, 2024; revised December 15, 2024.

Yang Yu is a graduate student at the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, Liaoing China(222085400542@stu.ustl.edu.cn).

Chunna Zhang* is an associate professor at the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, Liaoing China(lkdzcn@ustl.edu.cn).

Shengqiang Cong is a graduate student at the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, Liaoing China(222085400551@stu.ustl.edu.cn).

Xiaoping Yue is a graduate student at the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, Liaoing China(222085400571@stu.ustl.edu.cn).

Yuming Shen is a graduate student at the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, Liaoing China(232085400119@stu.ustl.edu.cn).

Jinchi Zhao is a graduate student at the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, Liaoing China(232085400142@stu.ustl.edu.cn).

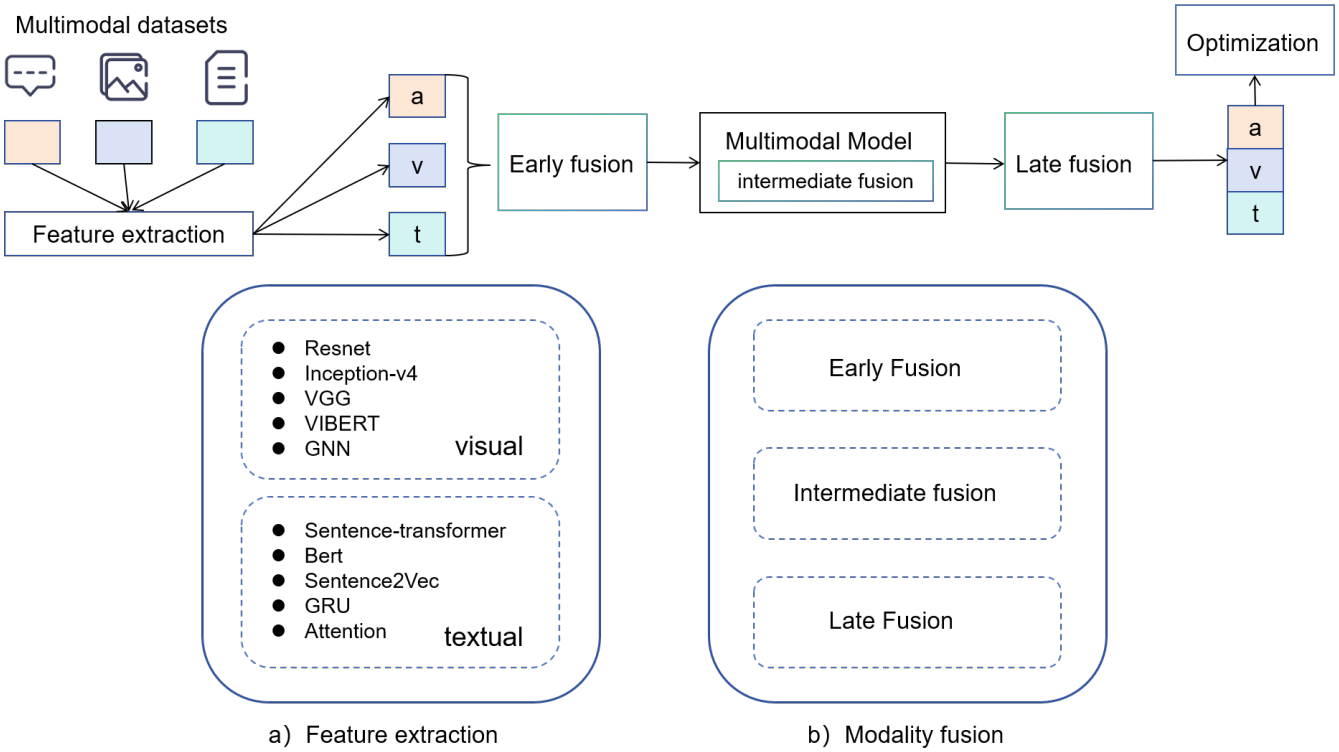


Fig. 1. Multi-modal Recommended Structure

Feature extraction aims to transform raw data into representative and interpretable modal features, which are presented as low-dimensional, easily comprehensible embeddings. The methodology for extracting visual and textual features is illustrated in Fig. 1 a).

B. User-Item Interaction Grath

The user-item interaction history graph effectively captures higher-order features of user preferences and interests by representing users and items as nodes, with interaction behaviors denoted as edges [22], [23].

The initial ID embeddings for the item are denoted as $e^0_{(u,\text{id})} \in \mathbb{R}^d$ and $e^0_{(i,\text{id})} \in \mathbb{R}^d$, where d represents the dimensionality of the embeddings.

Firstly, we construct the user-item interaction matrix $R \in \{0,1\}^{U\times I}$, where U and I denote the number of users and items, respectively. The matrix is defined as follows:

$$R = \begin{cases} 1, & \text{if user } u \text{ interacted with item } i \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

The user-item interaction matrix R can be represented as a sparse behavior graph $G = (V, E)$, where $V = U \cup I$ denotes the set of nodes and $E = \{(u, i) \mid R_{(u,i)} = 1\}$ represents the set of edges.

Subsequently, based on the user-item interaction matrix R, we construct the adjacency matrix A. The adjacency matrix is defined as follows:

$$A = \begin{bmatrix} 0 & R \\ R^T & 0 \end{bmatrix} \tag{2}$$

Finally, we normalize the adjacency matrix A to obtain the Laplacian matrix $\hat{A}$ of the user-item graph, which is defined by the following formula:

$$\hat{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \tag{3}$$

where D denotes the diagonal degree matrix.

C. Multi-modal Feature Fusion

Multi-modal fusion enhances task performance by integrating information from diverse modalities. This process is classified into three categories: early, intermediate, and late fusion, as illustrated in Fig. 1. Early fusion combines features prior to data input; however, this approach may result in the loss of interaction information. Intermediate fusion integrates features after extraction, preserving the original information while requiring the development of appropriate strategies for combining multi-modal features. In contrast, late fusion makes final decisions based on the predictions from each modality, thereby facilitating the learning of complementary information. Techniques for multi-modal fusion include element-wise summation, attention-based approaches, and concatenation. Attention-based methods assign weights to different modalities to capture their significance, as represented in the following formula.

$$u = \sum_{m\in\{v,a,t\}} \alpha_m u_m \tag{4}$$

In this paper, the model extracts shared features from each modality while isolating exclusive features through an attention mechanism. Furthermore, the model employs adaptive decision-making to assign weights, effectively merging shared and exclusive features to optimize feature representations. Experimental results demonstrate that this approach surpasses conventional multi-modal fusion techniques.

D. Behavior-aware Fusion

Behavioral perceptual fusion [24] aims to accurately capture the characteristics of items across multiple modalities.

By learning both modality-shared and modality-exclusive features. We can more effectively integrate multi-modal information, which enhances recommendation performance.

Specifically, the modality preference P_m is initially derived from user behavior features as follows:

$$P_m = \sigma(W_1 \bar{E}_{(u,i)} + b_1) \tag{5}$$

where $W_1 \in \mathbb{R}^{d \times d}$ and $b_1 \in \mathbb{R}^d$ are learnable parameters, σ denotes the sigmoid nonlinearity that facilitates the learning of a nonlinear gate to model user modality features. $\bar{E}_{u,i}$ represents the aggregated high-order neighbor information as detailed in the method.

All modalities encompass both shared and exclusive features. Shared features are extracted from modalities through the attention mechanism [25], which calculates the modality attention weights for the final modal features $\hat{E}_m$ of users and items as detailed in the method.

$$\alpha_m = \text{softmax}(q_1^T \tanh(W_2 \hat{E}_m + b_2)) \tag{6}$$

where α_m denotes the attention weight for modality m, $q_1 \in \mathbb{R}^d$ is the attention vector, $W_2 \in \mathbb{R}^{d \times d}$ is the weight matrix, and $b_2 \in \mathbb{R}^d$ is the bias vector.

By computing attention weights for each modal feature and applying these weights, the shared modal features E_s are extracted through a weighted summation as follows:

$$E_s = \sum_{m \in M} \alpha_m \hat{E}_m \tag{7}$$

where E_s represents the shared features across modalities, while M is the set of modalities.

By subtracting the shared features E_s from the original features, we obtain the exclusive modality features E'_m for each modality as follows:

$$E'_m = \hat{E}_m - E_s \tag{8}$$

III. Method

The multi-modal recommendation model presented in this paper is depicted in Fig. 2. This model consists of three key components: adaptive decision noise reduction, bi-representational learning, and multi-modal fusion. It effectively manages the multi-modal information of projects by employing adaptive decision noise reduction to minimize noise interference. Bi-representational learning is achieved through the construction of heterogeneous and homogeneous graphs, which enhances the modality representations of both users and items. Additionally, adaptive fusion techniques are employed to integrate shared and exclusive modal features derived from behavioral sensing, thereby facilitating effective multi-modal fusion and improving the modal features of users and items for a more accurate modeling of user preferences.

A. Problem Definition

Let $U = \{u\}$ denote the user set and $I = \{i\}$ represent the item set. The embedding of user and item input IDs is denoted as $E \in \mathbb{R}^{d \times (|U||I|)}$, where d signifies the embedding dimension. The modal feature representation of each item is represented as $E_m \in \mathbb{R}^{d_m \times |I|}$, where d_m represents the feature dimension and $m \in M$ denotes the modal feature. The set of modalities is defined as $M = \{v, t\}$, with v representing the visual modality and t representing the text modality.

B. Adaptive Decision-making for Noise Reduction

Research on multi-modal recommender systems has demonstrated that modal noise issues, such as irrelevant text or distracting background images, can significantly diminish the accuracy of recommendation results. In response to this challenge, this paper proposes a noise reduction module that employs an adaptive decision-making learning mechanism. Utilizing a neural network framework. This module regulates the information flow by learning the weight distribution among modalities and generating corresponding gating signals, with the aim of enhancing the system's noise filtering capabilities. Specifically, this method assesses the importance of each modality through two consecutive linear transformation layers and generates gating parameters to control the transfer of information accordingly, thereby achieving effective noise suppression at the feature level. This strategy is anticipated to improve both the quality of recommendations and the stability of multimodal recommendation systems in the presence of modal noise.

The original features are first transformed into higher-order features. Utilizing a trainable weight matrix w_3, relevant information is extracted from the original features to relevant information more compact and representative higher-order features $\tilde{I}_m$ as follows:

$$\tilde{I}_m = W_3 I_m + b_3 \tag{9}$$

where $W_3 \in \mathbb{R}^{d \times d_m}$ and $b_3 \in \mathbb{R}^d$ denote the learnable weight matrix and bias vector, respectively.

Utilizing adaptive decision-making learning, the user behavioral feature $\bar{e}_u$ is input into the first fully connected layer of the adaptive decision-making process. Subsequently, the ReLU activation function is then applied to obtain a non-linear, sparse, and feature-selective representation h as follows:

$$h = ReLU(W_4 \bar{e}_u + b_4) \tag{10}$$

where $W_4 \in \mathbb{R}^{d_n \times d_m}$ and $b_4 \in \mathbb{R}^{d_n}$ denote the learnable weight matrix and bias vector, respectively. Additionally, d_n represents the size of the hidden layer.

The learned feature representation h is input into the second fully connected layer, where the sigmoid activation function is applied to obtain the parameter gate g as follows:

$$gate = \sigma(W_5 h + b_5) \tag{11}$$

where $W_5 \in \mathbb{R}^{1 \times d_n}$ and $b_5 \in \mathbb{R}$ denote the learnable weight matrix and bias vector, respectively. Additionally, d_n represents the size of the hidden layer.

Utilizing the parameter gate $gate$ learned from adaptive decision-making based on user behavioral characteristics, modality features are weighted to derive the denoised modality feature representation $\tilde{E}_m$ as follows:

$$\bar{I}_m = I_m \odot gate \tag{12}$$

where $\odot$ denotes element-wise multiplication, employed to multiply each modal feature by its corresponding parameter. The weighted modal features $\bar{I}_m$ represent aspects pertinent to the user's preferences, while unweighted or low-weighted features signify noise that is unrelated to the user.

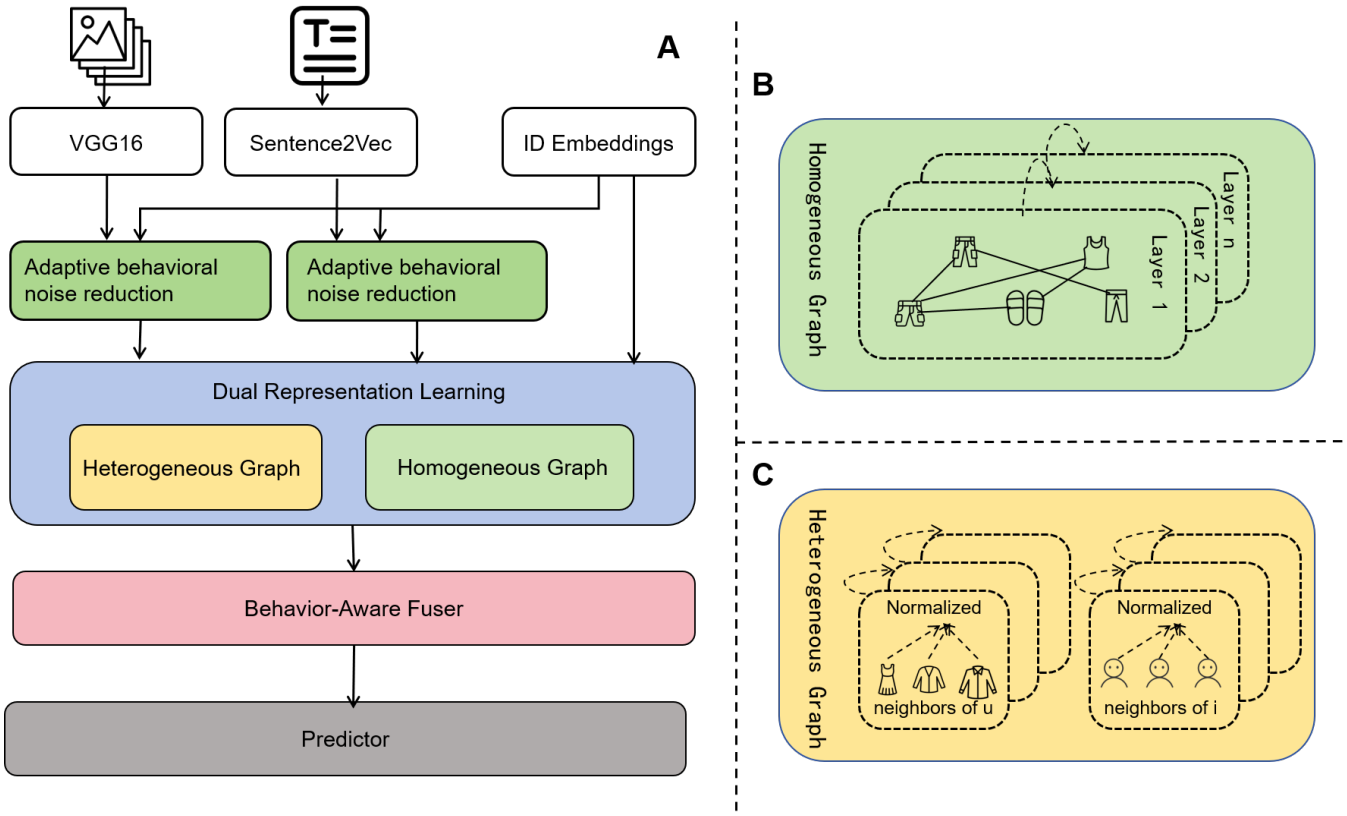


Fig. 2. A: The overall Framework B: Homogeneous Graph C: Heterogeneous Graph

C. Dual Representation Learning (DLL)

Learning the representations of users and items is crucial for the effectiveness of recommender systems. According to[15], the integration of item-item homogeneous graphs with user-item heterogeneous graphs can significantly enhance the performance of multi-modal recommendations.

1) Heterogeneous Graph: We employ graph convolution operations to propagate the embeddings of user and item IDs through the historical interaction graphs. The Laplacian matrix $\hat{A}$ of the user-item historical interaction graph is computed according to equations (3). The output of the l-th graph convolutional layer, denoted as $E^l_{U\cup I}$, can be expressed as follows:

$$E^l_{(U\cup I)} = ReLU(\hat{A}E^{(l-1)}_{(U\cup I)}\mathbf{w}^l + \mathbf{b}^l) \tag{13}$$

where $E^l_{U\cup I}$ denotes the set of user and item embeddings enhanced by the l-th convolutional layer. $E^0_{U\cup I}$ represents the initial ID embedding for both users and items, while W^l and b^l denote the trainable weight matrix and bias term associated with the l-th layer, respectively. The ReLU function serves as a non-linear activation mechanism.

Multi-layer graph convolution integrates higher-order neighborhood information at each layer. The final representations of user u and item i are obtained by aggregating information from all layers as follows:

$$\bar{E}_{(u,i)} = \frac{1}{L+1}\sum_{l=0}^{L}\bar{E}^l_{(u,i)} \tag{14}$$

where L denotes the number of layers in the graph convolutional network, and $\bar{E}_{(u,i)}$ represents the high-order neighborhood representation of user u and item i.

2) Homogeneous Graph: Based on the features of item modality, the relationships between items are evaluated using cosine similarity, which leads to the creation of an item-item similarity matrix S:

$$S_{ij} = \frac{\mathbf{f}_i^T \cdot \mathbf{f}_j}{\|\mathbf{f}_i\|\|\mathbf{f}_j\|} \tag{15}$$

where f_i and f_j denote the feature vectors of items i and j, respectively, and S_{ij} signifies the similarity between f_i and f_j.

Graph convolution operations are utilized to extract shared features among items. The item-item matrix $\tilde{S}_{ij}$ is sparsified using the k-nearest neighbors (KNN) method, retaining only the top-k most similar items. The matrix $\tilde{S}_{ij}$ is presented below:

$$\tilde{S}_{ij} = \begin{cases} S_{ij}, & \text{if } j \in \text{KNN}_k(i) \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

where $\text{KNN}_k(i)$ represents the set of k nearest neighbors associated with item i.

To address the issue of gradient explosion, the item-item similarity matrix $\tilde{S}_{ij}$ is normalized in the following manner:

$$\tilde{S} = D_m^{-1/2} S_{ij} D_m^{-1/2} \tag{17}$$

where, D_m denotes the diagonal degree matrix.

We perform a graph convolution operation on the normalized item-item similarity matrix $\hat{S}_{ij}$ to propagate the modal features of all items $\hat{I}_m$:

$$\hat{I}_m = \tilde{S}\bar{I}_m \tag{18}$$

where $\bar{I}_m$ represents the modal signature of the item following noise reduction, while $\hat{I}_m$ denotes the updated modal signature of the item.

Develop a user model feature representation by aggregating multi-modal features generated from user interactions with items. This representation can elucidate the common characteristics shared between users and the items they engage with, thereby effectively capturing the personalized preferences of users:

$$\hat{u}_m = \sum_{i \in N_u} \frac{1}{\sqrt{(|N_u||N_i|)}} \hat{i}_m \tag{19}$$

where $\hat{u}_m$ represents the modal feature of user u, while $\hat{i}_m$, derived from equations (18), denotes the modal feature of item i. Define N_u as the set of all item nodes interacted by user u, and N_i as the set of neighboring item nodes associated with item i. The quantities $|N_u|$ and $|N_i|$ respectively indicate the number of item nodes that have been interacted with by user u and the number of neighboring item nodes related to item i.

By concatenating the user modal feature $\hat{U}_m$ with the item modal feature $\hat{I}_m$, we obtain the final modal feature $\hat{E}_m$ for users and items. This resultant feature is represented in $\mathbb{R}^{d \times (|U|+|I|)}$.

D. Multi-modal Fusion

The shared modal features E_s and the exclusive modal features $\bar{E}_m$ are derived from equations (7) and (8) within the framework of behavioral perception fusion. The model presented in this paper employs an applicability fusion method to adaptively adjust the weights of various modal features based on the modal preferences P_m, which are extracted from user behavioral characteristics. The fused features are then integrated with the common modal features $\bar{E}_m$ to generate the final feature representation $E_{(U \cup I),m}$:

$$E_{(U \cup I),m} = E_s + \frac{1}{|M|} \sum_{m \in M} \acute{E}_m \odot P_m \tag{20}$$

where $E_{(U \cup I),m}$ denotes the fused modal feature representation for users and items, and $\odot$ signifies element-wise multiplication.

To facilitate the exploration of the relationship between behavioral features and multi-modal information, self-supervised auxiliary tasks have been developed. The objective is to maximize the mutual information between behavioral features and the fused multi-modal features. Mutual Information (MI) quantifies the interdependence between two random variables and measures the extent of shared information across different features. The loss function is defined as follows:

$$\begin{aligned} L_C = & \sum_{u1 \in U} \left(-\log \left(\frac{\exp\left(\frac{e_{u1,\text{mul}} \cdot \bar{e}_{u1}}{\tau}\right)}{\sum_{u2 \in U} \exp\left(\frac{e_{u2,\text{mul}} \cdot \bar{e}_{u2}}{\tau}\right)} \right) \right) \\ & + \sum_{i1 \in I} \left(-\log \left(\frac{\exp\left(\frac{e_{i1,\text{mul}} \cdot \bar{e}_{i}}{\tau}\right)}{\sum_{i2 \in I} \exp\left(\frac{e_{i2,\text{mul}} \cdot \bar{e}_{i}}{\tau}\right)} \right) \right) \end{aligned} \tag{21}$$

where $\bar{e}_u$ and $\bar{e}_i$ denote the behavioral features of user u and item i, respectively. The terms $e_{(u,\text{mul})}$ and $e_{(i,\text{mul})}$ represent the multi-modal features that are integrated from the modalities associated with user u and item i, respectively. The parameter τ denotes the temperature of the softmax function, controls the degree of smoothing in the distribution.

E. Dual Representation Integration

User and item representations derived from heterogeneous (user interaction) and homogeneous (item-item) graphs are integrated to from a comprehensive representation that encapsulates both user interactions and item semantics. The final user and item representations are generated by aggregating the behavioral and multi-modal features, as detailed in the following formula:

$$e_u = \hat{u}_m + e_{u,\text{mul}} \tag{22}$$

$$e_i = \hat{i}_m + e_{i,\text{mul}} \tag{23}$$

where e_u and e_i denote the final representation of user u and item i, respectively. The symbols $\hat{u}_m$ and $\hat{i}_m$ represent the behavioral characteristics of the user and the item, respectively. $e_{(u,\text{mul})}$ and $e_{(i,\text{mul})}$ refer to the modal characteristics of the user and the item, which are derived from behavioral perception and multi-modal fusion.

F. Predictor and Optimization

During the prediction phase, the preference score of user u for item i is calculated by taking the inner product of their final feature representations:

$$\hat{y}_{ui} = (e_u)^T e_i \tag{24}$$

The use of the Bayesian Personalized Ranking (BPR) loss function promotes the model's capability to prioritize items that user have previously clicked on, as opposed to those they have not interacted with. To facilitate this, we construct a triplet set R, where each triplet (u, i, i') satisfies the conditions $y_{ui} = 1$ and $y_{ui'} = 0$. The BPR loss is defined as follows:

$$L_{BPR} = \sum_{(u,i,i') \in R} (-\log \sigma(\hat{y}_{ui} - \hat{y}_{ui'})) \tag{25}$$

where $\hat{y}_{ui'}$ represents the user rating for the negative sample i'. The function $\sigma(\cdot)$ denotes the logistic sigmoid function, which is defined as $\sigma(x) = \frac{1}{1+e^{-x}}$.

By integrating the loss from the self-supervised auxiliary task, the Bayesian Personalized Ranking (BPR) loss, and a regularization term, we derive the final loss function:

$$L = L_{BPR} + \lambda_C L_C + \lambda_E \|E\|_2 \tag{26}$$

where E denotes the set of model parameters, while λ_C and λ_E are hyperparameters that control the impact of the self-supervised task and L_2 regularization, respectively.

IV. Experiments

In this section, we conduct extensive experiments to evaluate the performance of the proposed model, AM2HRes, across three public datasets.

A. Experimental Settings

1) Dataset: The experiments utilize three datasets—baby, sports, and clothing—from the Amazon Review Dataset. This dataset encompasses product descriptions and image information, which serve as textual and visual features. To ensure robust analysis, five core users and projects were retained, each exhibiting a minimum of five interactions, a

TABLE II
THE GENERAL PERFORMANCE OF VARIOUS RECOMMENDATION APPROACHES REGARDING RECALL AND NDCG

Datasets	Metrics	BPR	LightGCN	VBPR	MMGCN	GRCN	MMSSL	MICRO	FREEDOM	MGCN	Ours
baby	Recall@10	0.0357	0.0479	0.0423	0.0421	0.0532	0.0613	0.0584	0.0627	0.0620	**0.0628**
	Recall@20	0.0575	0.0754	0.0663	0.0660	0.0824	0.0971	0.0929	**0.0992**	0.0964	0.0980
	NDCG@10	0.0192	0.0257	0.0223	0.0220	0.0282	0.0326	0.0318	0.0330	0.0339	**0.0339**
	NDCG@20	0.0249	0.0328	0.0284	0.0282	0.0358	0.0420	0.0407	0.0424	0.0427	**0.0431**
sports	Recall@10	0.0432	0.0569	0.0558	0.0401	0.0599	0.0673	0.0679	0.0717	0.0729	**0.0754**
	Recall@20	0.0653	0.0864	0.0856	0.0636	0.0919	0.1013	0.1050	0.1089	0.1106	**0.1133**
	NDCG@10	0.0241	0.0311	0.0307	0.0209	0.0330	0.0380	0.0367	0.0385	0.0394	**0.0410**
	NDCG@20	0.0298	0.0387	0.0384	0.0270	0.0413	0.0474	0.0463	0.0481	0.0496	**0.0508**
clothing	Recall@10	0.0206	0.0361	0.0281	0.0227	0.0421	0.0531	0.0521	0.0629	0.0641	**0.0645**
	Recall@20	0.0303	0.0544	0.0544	0.0361	0.0657	0.0797	0.0772	0.0941	0.0945	**0.0958**
	NDCG@10	0.0114	0.0197	0.0197	0.0120	0.0224	0.0291	0.0283	0.0341	0.0347	**0.0354**
	NDCG@20	0.0138	0.0243	0.0243	0.0154	0.0284	0.0359	0.0347	0.0420	0.0428	**0.0434**

methodology commonly employed in existing studies. For the text modality, text embeddings with a dimensionality of 384 were extracted by integrating the title, description, category, and brand of each item using a sentence transformer. Visual features, with a dimensionality of 4096, were derived from a pre-trained convolutional neural network. These features have been documented in the literature [26]. Table 1 provides a summary of the dataset statistics, where data sparsity is calculated as the number of interactions divided by the total number of user-item pairs.

TABLE I
STATISTICS OF THE EXPERIMENTAL DATA SET

Dataset	Users	Items	Interaction	Sparsity
baby	19,445	7,050	160,792	99.88%
sports	35,598	18,357	296,337	99.95%
clothing	39,387	23,033	278,677	99.97%

2) Baseline model: To evaluate the performance of the proposed model, we conducted a comprehensive analysis. The results were compared with those of several representative models, which can be classified into two primary categories: traditional recommendation methods and multimedia recommendation methods.

General model: BPR [27] is a classic collaborative filtering method that employs a matrix factorization framework to learn representations of users and items. LightGCN [28]: As the most widely adopted GCN-based collaborative filtering method, simplifies the design of GCN while enhancing its applicability for recommendation tasks.

Multimedia Models: MMGCN [8] constructs modality-specific graphs to learn distinct modal features, integrating these features to obtain a unified representation of users or items for prediction. GRCN [9] enhances previous GCN-based models by refining the user-item interaction graph; it leverages multi-modal features to identify and mitigate false positive interactions. MMSSL [18] employs self-supervised tasks and adversarial networks to capture information-guided user preferences in sparse interaction scenarios, thereby improving the effectiveness of recommendations. MICRO [29], an extension of the state-of-the-art LATTICE [15], enhances item representations by learning a item-item graph from multi-modal item features, thus exploring potential structures among items. FREEDOM [16] introduces a degree-sensitive edge pruning method that denoises the user-item graph by removing noise from unintended interactions. MGCN [24] designs a behavior perceptron to capture user behavioral characteristics, thereby enhancing the efficacy of modality fusion.

3) Evaluation Indicators: For a fair comparison, we follow the same evaluation setting of [15] with a random data splitting 8:1:1 on the interaction history of each user for training, validation, and testing. Besides, we follow the all-ranking protocol to evaluate the top-K recommendation performance and report the average metrics for all users in the test set: Recall@K and NDCG@K.

4) Implementation Details: In this experiment, the model was implemented using PyTorch version 1.11.0 and Python version 3.8.0. The Xavier initialization method was employed for parameter initialization, and the Adam optimizer was utilized for optimization. The embedding dimension was set to 64. For hyperparameter tuning, a grid search was conducted to identify the optimal learning rate and regularization loss weights from the set 0.0001, 0.001, 0.01, 0.1, with an initial learning rate established at 0.001. The sparsity levels for KNN neighbor counts ranged from 5 to 30. The coefficients for auxiliary loss in the self-supervised task varied between 0.001 and 0.1. An early stopping strategy was implemented, with a maximum epoch limit of 1,000; thus, training ceases when Recall@20 on the validation set does not improve over a span of 20 consecutive epochs.

B. Performance Comparison

Table 2 shows the performance comparison of the proposed AM2HRec and other baseline methods on three datasets. By analyzing the data in the table, we made several key observations:

While AM2HRec showed sub-optimal performance for the Recall@20 metric on the baby dataset, it outperformed all baseline models across all evaluation metrics—specifically Recall@10, Recall@20, NDCG@10, and NDCG@20—on both the sports and clothing datasets. Notably, AM2HRec achieved a significant performance improvement of 4.06% compared to the best baseline model in the sports dataset.

TABLE III
COMPARISON OF DATA FOR ABLATION STUDY OF KEY COMPONENTS IN AM2HREC

Variants		AM2HRec$_{(-A)}$		AM2HRec$_{(-U)}$		AM2HRec$_{(-I)}$		AM2HRec$_{(-2H)}$		AM2HRec$_{(-F)}$		AM2HRec	
Metrics		R@20	N@20	R@20	N@20	R@20	N@20	R@20	N@20	R@20	N@20	R@20	N@20
Dataset	baby	0.0963	0.0428	0.0650	0.0285	0.0689	0.0298	0.0427	0.0186	0.0917	0.0410	**0.0980**	**0.0431**
	sports	0.1104	0.0497	0.0911	0.0406	0.0837	0.0378	0.0655	0.0298	0.1119	0.0498	**0.1133**	**0.0508**
	clothing	0.0935	0.0422	0.0760	0.0346	0.0465	0.0215	0.0282	0.0128	0.0884	0.0401	**0.0958**	**0.0434**

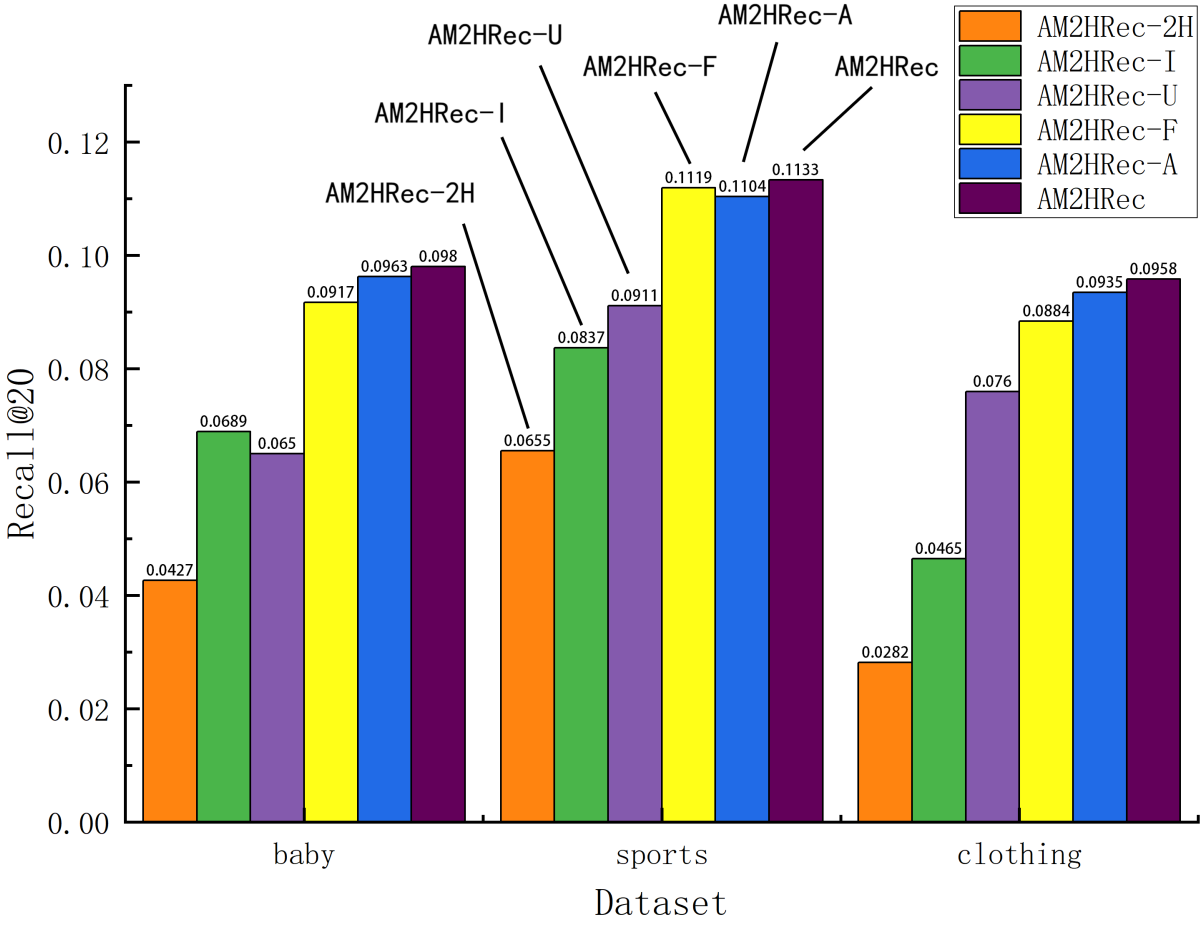


Fig. 3. Ablation Study of Key Components in AM2HRec Under the Recall@20 Metric

These results strongly support the effectiveness of the AM2HRec model, especially within the context of multi-modal recommendation tasks.

The enhanced performance of AM2HRec can be primarily attributed to two key innovations: adaptive decision noise reduction and dual representation learning. Unlike methods that simply fuse multi-modal information, AM2HRec first employs an adaptive decision noise reduction technique to effectively mitigate noise from multi-modal inputs, thereby minimizing modal interference. Furthermore, by integrating dual representation learning through both heterogeneous and homogeneous graphs, AM2HRec improves the representations of users and items, resulting in a significant enhancement in overall model performance.

Models such as BPR and LightGCN can significantly benefit from the integration of multi-modal information. For instance, VBPR achieves an average performance improvement of 27.63% over traditional BPR by incorporating visual features alongside ID embeddings. However, MMGCN often underperforms compared to LightGCN, primarily due to the introduction of noise during message propagation in Graph Convolutional Networks (GCNs), which can compromise the final representations of users and items. While models like FREEDOM and MMSSL attempt to address noise through structural noise reduction and self-supervised learning, they do not fully resolve this issue. In contrast, the adaptive decision-making approach proposed in this paper effectively minimizes multi-modal noise.

In this study, GRCN enhances user preference extraction by refining user-item graphs and utilizing graph convolutional layers. In contrast, MICRO constructs auxiliary item-item graphs to enhance item information. However, both

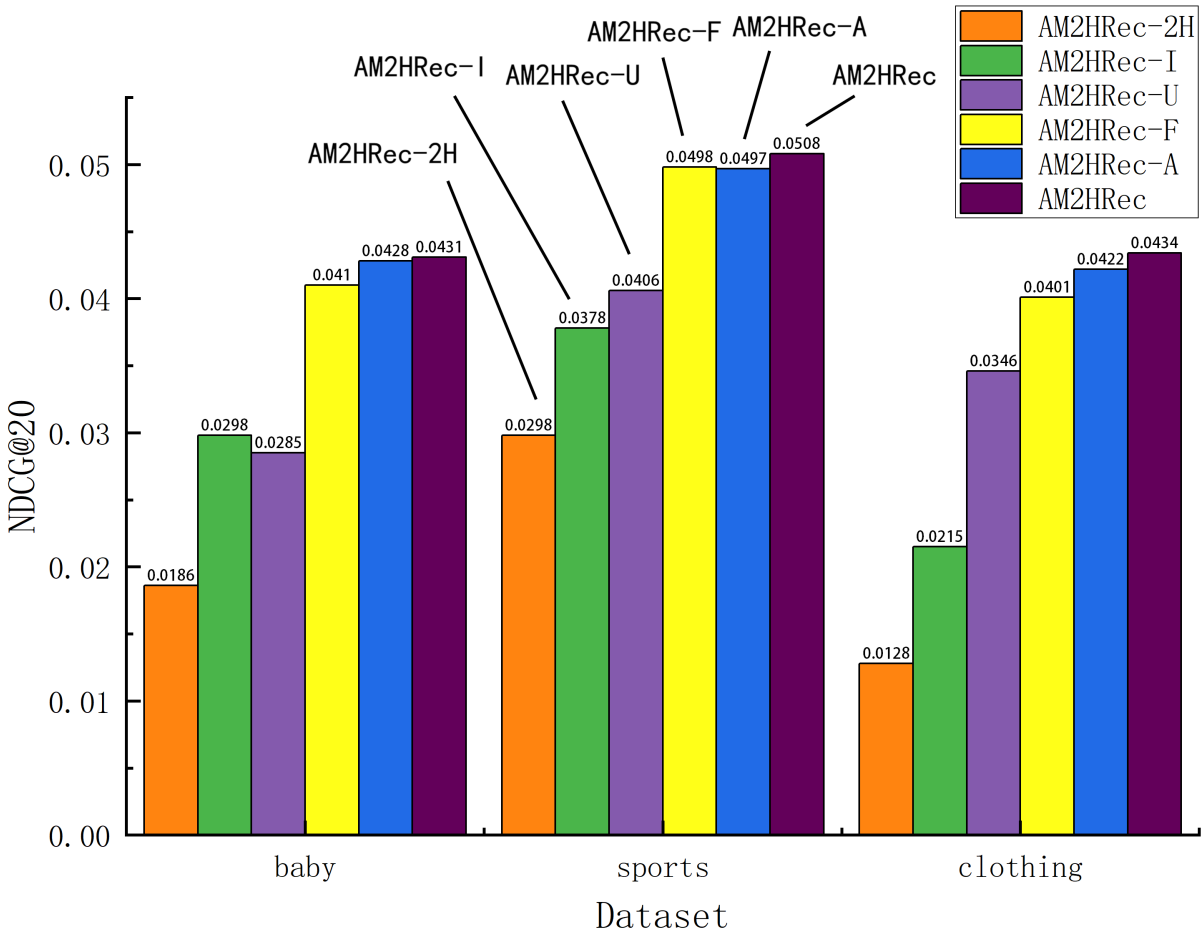


Fig. 4. Ablation Study of Key Components in AM2HRec Under the NDCG@20 Metric

methods are limited by single-representation learning, which constrains their capacity to capture a more comprehensive feature representation. This paper proposes a novel approach that optimizes user and item representations through adaptive noise reduction and bi-graph representation learning. Performance comparisons demonstrate the effectiveness of the proposed method.

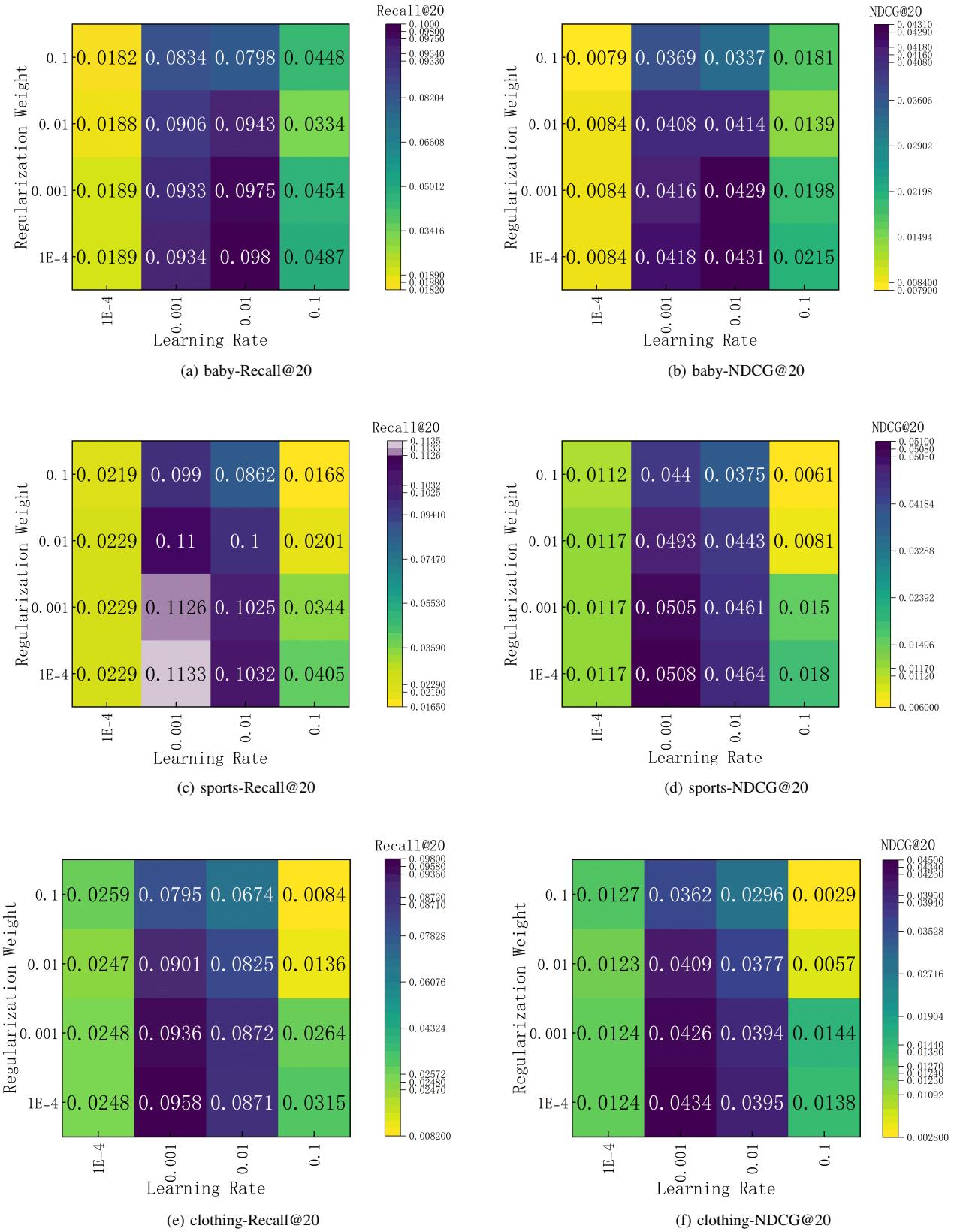
C. Ablation Study

In this section, we conduct exhaustive experiments to evaluate the effectiveness of different components of AM2HRec.

1) Effect of Different Components of the Model: To quantitatively evaluate the contribution of each component in the AM2HRec model, this study designed a series of model variants, each omitting a specific module. Specifically:AM2HRec$_{(-A)}$ removes the adaptive decision-making denoising module, allowing modal features without denoising to be directly input into the model.AM2HRec$_{(-U)}$ eliminates the user-item interaction graph, propagating modal features solely on the item-item interaction graph. AM2HRec$_{(-I)}$ excludes the item-item interaction graph, limiting the propagation of modal features to the user-item interaction graph. AM2HRec$_{(-2H)}$ omits the dual representation learning module, using the denoised modal features directly for modal fusion. Finally, AM2HRec$_{(-F)}$ removes the adaptive modal fusion module, employing a simple modal splicing method to integrate modal information. By comparing the performance differences between these variants and the original AM2HRec model, this study aims to provide an in-depth analysis of the specific impact of each module on the model's overall performance, thereby offering guiding insights for model optimization.

Based on the data presented in Table 3 and the bar charts illustrated in Figures 3 and 4, we can conduct a comprehensive analysis comparing the performance of the three datasets. The experimental findings indicate a significant improvement in performance when employing dual representation learning compared to single representation learning. Concurrently learning representations from both the user-item heterogeneous graph and the item-item homogeneous graph facilitates a more holistic understanding of user-item relationships. This methodology enables the extraction of richer features, ultimately leading to more accurate recommendations.

Experimental data for the AM2HRec$_{(-F)}$ variant indicate that the dynamic fusion strategy, which adjusts modal feature weights based on user behavior, outperforms the straightforward method of merging modal features. This validates the effectiveness of the adaptive fusion approach in integrating multi-modal information and enhancing the performance of recommender system.



(a) baby-Recall@20

(b) baby-NDCG@20

(c) sports-Recall@20

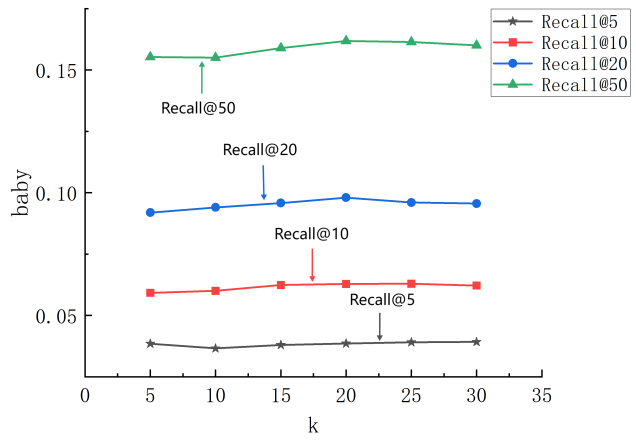
(d) sports-NDCG@20

(e) clothing-Recall@20

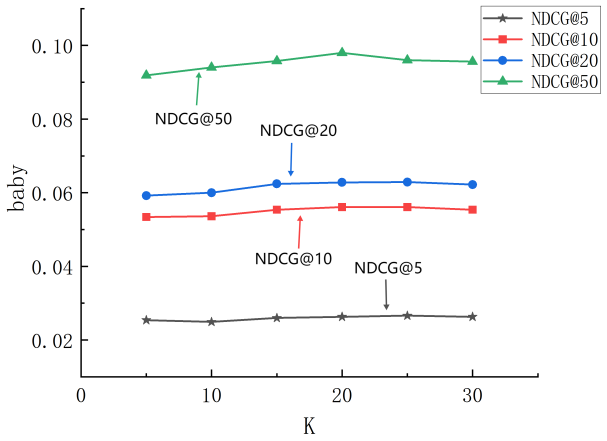
(f) clothing-NDCG@20

Fig. 5. Performance of AM2HRec at Different Learning Rates and Regularization Loss Weight Values. Darker Colors Indicate Higher Performance Metrics

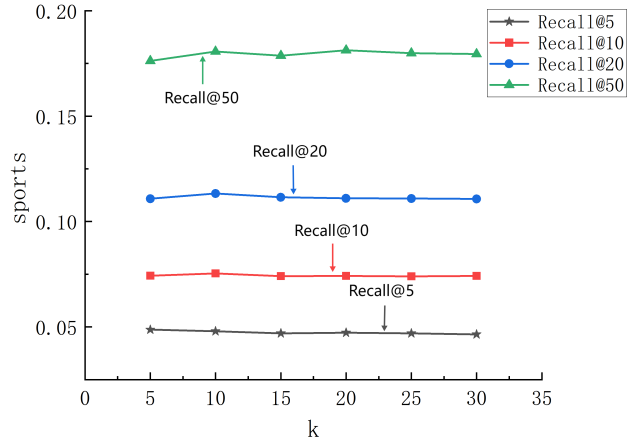
In the AM2HRec$_{(-A)}$ variant, the use of raw features without adaptive decision denoising leads to a decline in model performance. This suggests that noise in the raw modal information interferes with user-item representation, thereby reducing the model's effectiveness. Adaptive decision-making noise reduction techniques process raw modal information to effectively eliminate noise and enhance overall model performance.



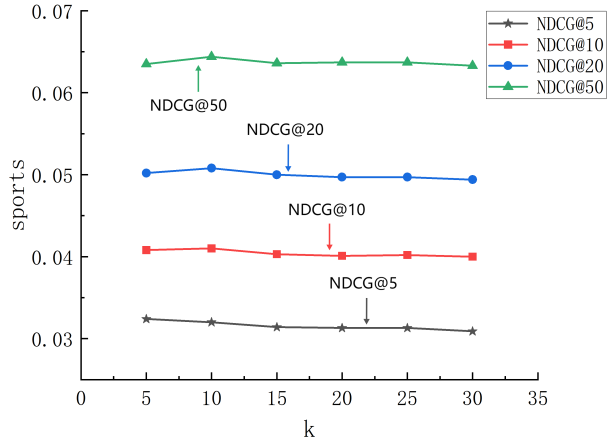
(a) baby-Recall@k



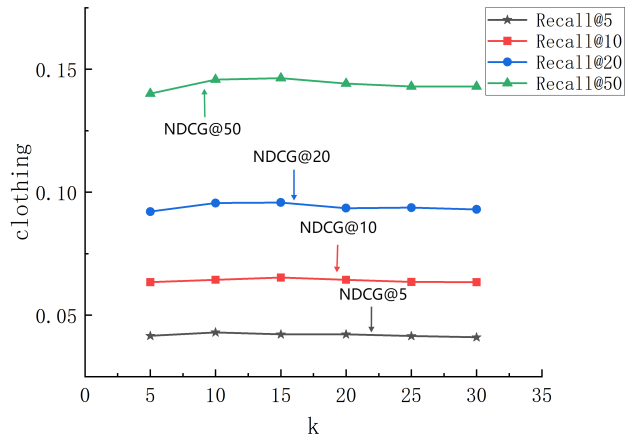
(b) baby-NDCG@k



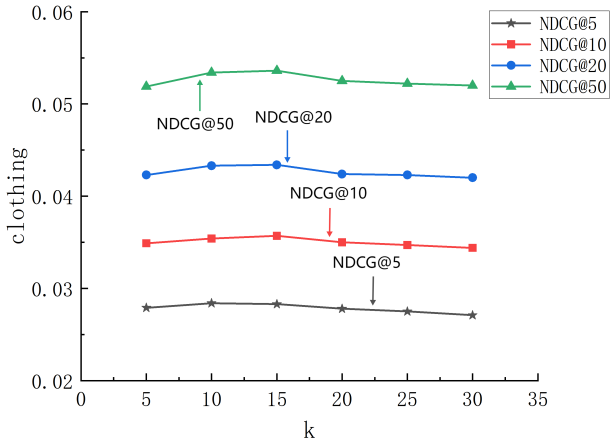
(c) sports-Recall@k



(d) sports-NDCG@k



(e) clothing-Recall@k

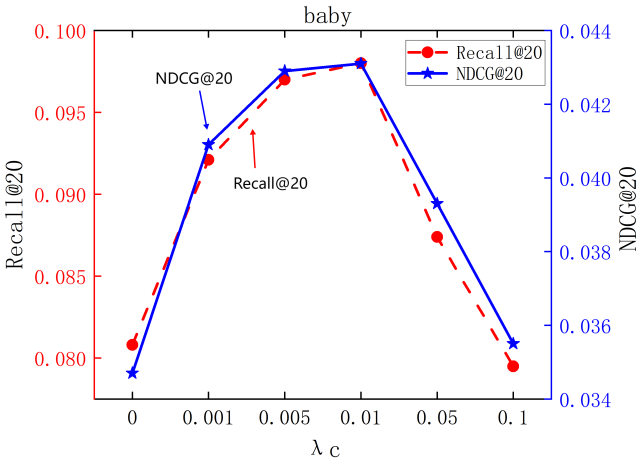


(f) clothing-NDCG@k

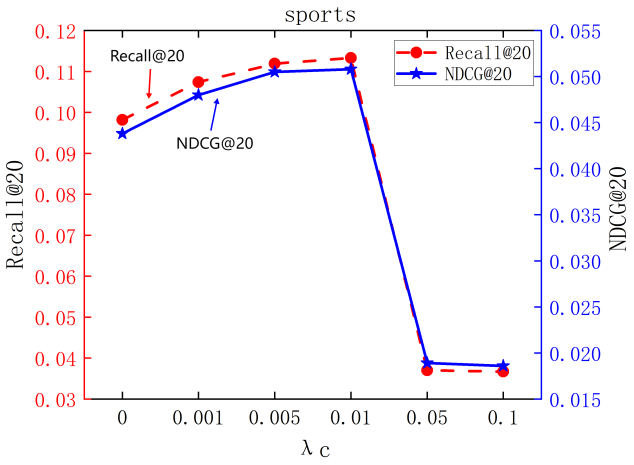
Fig. 6. Comparative Analysis of Various KNN-k Values Based on the Recall@K Metric

D. Hyper-parameter Sensitivity Study

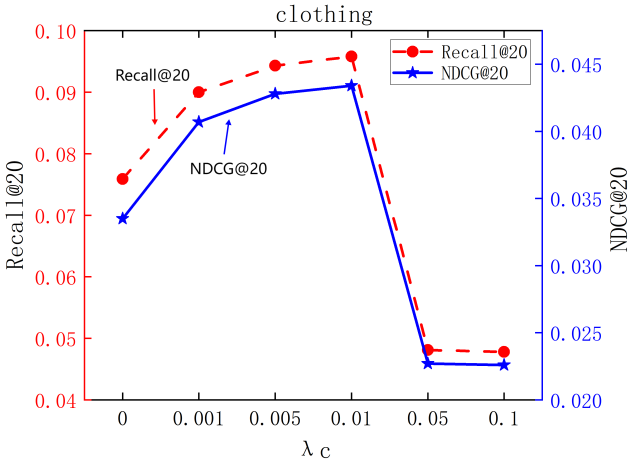
1) The Pair of Hyper-parameters Learning Rate and Regularization Loss Weight: To conduct a comprehensive exploration of the hyperparameter settings for the AM2HRec model, we performed sensitivity analyses on the evaluation metrics Recall@20 and NDCG@20. We varied the learning rate and regularization loss weight values within the ranges of 0.0001, 0.001, 0.01, 0.1. Fig. 5 illustrates the performance



(a) λ_C-baby



(b) λ_C-sports



(c) λ_C-clothing

Fig. 7. Weights of Self-supervised Task λ_C

of the AM2HRec model across the baby, sports, and clothing datasets, showcasing various combinations of learning rates and regularization weights.

The experimental results demonstrate that, for the baby dataset, the model attains optimal performance with a learning rate of 0.01 and a regularization loss weight of 1e-4. Additionally, when the learning rate is maintained within the range of 1e-3, 1e-2, the model exhibits stable performance, further validating the effectiveness and stability of the AM2HRec model. These findings provide clear guidance for hyperparameter selection aimed at optimizing the AM2HRec model.

2) Effects of the Number of Item Neighbor K: In constructing the item-item homogeneity graph, we investigated the impact of varying the number of item neighbors, denoted as k, on model performance. The objective is to minimize interference from unrelated items by selecting the k most similar items. We employed Recall@20 and NDCG@20 as evaluation metrics across different values of k. Fig. 6 (g)-(l) illustrates the variations in performance metrics across three distinct datasets as a function of different k values. A comprehensive analysis of these experimental results reveals that the optimal value of k is dependent on the dataset. Specifically, the model achieves peak performance on the baby dataset when k is set to 20. In contrast, optimal performance is observed with k values of 10 and 15 for the sports and clothing datasets, respectively. This finding highlights the importance of adjusting the number of item neighbors based on the characteristics of each dataset to maximize recommendation performance. Through meticulous parameter tuning, we can construct more effective homogeneous graphs, thereby enhancing the accuracy and relevance of multi-modal recommendations.

3) Effects of the Weight of Self-Supervised Task λ_C: We explore the impact of the self-supervised task weight λ_C on model performance. Analyzing three different datasets, we obtained consistent findings, which are presented in Fig. 7 We have observed that combining the optimization of self-supervised auxiliary tasks with core recommendation tasks can improve the overall performance of the model.

Specifically, we determined that the optimal value of λ_C is approximately 0.01, and when λ_C exceeds this threshold, the performance of the model is significantly reduced. This implies that a moderate λ_C helps to enhance the learning

effect of the main recommendation task. On the contrary, if λ_C is set too high, the model may place too much emphasis on auxiliary tasks, thereby affecting the guidance of self-supervised tasks and resulting in compromised model performance.

Therefore, choosing an appropriate weight for the self-supervised loss term is critical to ensure the effectiveness of the model's recommendation capabilities.

V. Conclusion

In this paper, we propose the AM2HRec model, a multi-modal recommendation framework that utilizes dual-representation adaptive denoising. Initially, the framework preprocesses the original multi-modal data features through an adaptive decision-making denoising module, which alleviates the detrimental effects of noise features on information nodes. Subsequently, the model engages in dual representation learning on heterogeneous and isomorphic graphs to capture high-order modal features and the intricate semantic relationships between items, thereby improving the representation of both users and items. Furthermore, the AM2HRec model employs an adaptive decision-making method to integrate multi-modal data and optimize the modal fusion process. Experimental results across three datasets demonstrate that the AM2HRec model significantly enhances the performance of recommendation systems. Looking ahead, we intend to incorporate self-supervised tasks and large language models to further improve the capabilities of multi-modal data processing, thereby addressing issues related to data sparsity and cold start problems.

References

[1] Nagagopiraju Vullam, Sai Srinivas Vellela, Venkateswara Reddy, M Venkateswara Rao, Khader Basha SK, and D Roja. Multi-Agent Personalized Recommendation System in E-Commerce Based on User. In *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pages 1194–1199, 2023. IEEE.

[2] Yitong Pang, Lingfei Wu, Qi Shen, Yiming Zhang, Zhihua Wei, Fangli Xu, Ethan Chang, Bo Long, and Jian Pei. Heterogeneous Global Graph Neural Networks for Personalized Session-Based Recommendation. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 775–783, 2022.

[3] Chuhan Wu, Fangzhao Wu, Yongfeng Huang, and Xing Xie. Personalized news recommendation: Methods and challenges. *ACM Transactions on Information Systems*, 41(1):1–50, 2023. ACM New York, NY.

[4] Kang Liu, Feng Xue, Dan Guo, Le Wu, Shujie Li, and Richang Hong. Megcf: Multimodal entity graph collaborative filtering for personalized recommendation. *ACM Transactions on Information Systems*, 41(2):1–27, 2023. ACM New York, NY.

[5] Yun Li, Shuyi Liu, Xuejun Wang, and Peiguang Jing. Self-supervised deep partial adversarial network for micro-video multimodal classification. *Information Sciences*, 630:356–369, 2023. Elsevier.

[6] Zengmao Wang, Haifeng Xia, Shuai Chen, and Gang Chun. Joint representation learning with ratings and reviews for recommendation. *Neurocomputing*, 425:181–190, 2021. Elsevier.

[7] Yi-Hong Lu, Chang-Dong Wang, Pei-Yuan Lai, and Jian-Huang Lai. PKAT: Pre-training in Collaborative Knowledge Graph Attention Network for Recommendation. In *2023 IEEE International Conference on Data Mining (ICDM)*, pages 448–457, 2023. IEEE.

[8] Desheng Cai, Shengsheng Qian, Quan Fang, Jun Hu, and Changsheng Xu. Adaptive anti-bottleneck multi-modal graph learning network for personalized micro-video recommendation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 581–590, 2022.

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

[10] Fan Liu, Zhiyong Cheng, Lei Zhu, Zan Gao, and Liqiang Nie. Interest-aware message-passing GCN for recommendation. In *Proceedings of the web conference 2021*, pages 1296–1305, 2021.

[11] Qifan Wang, Yinwei Wei, Jianhua Yin, Jianlong Wu, Xuemeng Song, and Liqiang Nie. Dualgnn: Dual graph neural network for multimedia recommendation. *IEEE Transactions on Multimedia*, 25:1074–1084, 2021. IEEE.

[12] Chen Gao, Yu Zheng, Nian Li, Yinfeng Li, Yingrong Qin, Jinghua Piao, Yuhan Quan, Jianxin Chang, Depeng Jin, Xiangnan He, et al. A survey of graph neural networks for recommender systems: Challenges, methods, and directions. *ACM Transactions on Recommender Systems*, 1(1):1–51, 2023. ACM New York, NY, USA.

[13] Yue Teng and Kai Yang. Research on Enhanced Multi-head Self-Attention Social Recommendation Algorithm Based on Graph Neural Network. *IAENG International Journal of Computer Science*, vol. 51, no. 7, pp 754-764, 2024.

[14] Yaochen Xie, Zhao Xu, Jingtun Zhang, Zhengyang Wang, and Shuiwang Ji. Self-supervised learning of graph neural networks: A unified review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2412–2429, 2022. IEEE.

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

[16] Xin Zhou and Zhiqi Shen. A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 935–943, 2023.

[17] Xin Zhou, Aixin Sun, Yong Liu, Jie Zhang, and Chunyan Miao. Selfcf: A simple framework for self-supervised collaborative filtering. *ACM Transactions on Recommender Systems*, 1(2):1–25, 2023. ACM New York, NY.

[18] Wei Wei, Chao Huang, Lianghao Xia, and Chuxu Zhang. Multi-modal self-supervised learning for recommendation. In *Proceedings of the ACM Web Conference 2023*, pages 790–800, 2023.

[19] Ke Wang, Yanmin Zhu, Tianzi Zang, Chunyang Wang, Kuan Liu, and Peibo Ma. Multi-aspect Graph Contrastive Learning for Review-Enhanced Recommendation. *ACM Transactions on Information Systems*, 42(2):1–29, 2023. ACM New York, NY, USA.

[20] Junliang Yu, Xin Xia, Tong Chen, Lizhen Cui, Nguyen Quoc Viet Hung, and Hongzhi Yin. XSimGCL: Towards extremely simple graph contrastive learning for recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 2023. IEEE.

[21] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations & Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions. *ACM Computing Surveys*, 2023. ACM New York, NY.

[22] Ruxing Li, Dan Yang, and Xi Gong. Heterogeneous Graph Contrastive Learning with Attention Mechanism for Recommendation. *Engineering Letters*, vol. 32, no. 10, pp 1930–1938, 2024.

[23] Haibo Hu, Dan Yang, and Yu Zhang. DPRec: Social Recommendation Based on Dynamic User Preferences. *IAENG International Journal of Computer Science*, vol. 50, no. 3, pp 980-987, 2023.

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

[25] Yanan Wang, Jianming Wu, and Keiichiro Hoashi. Multi-attention fusion network for video-based emotion recognition. In *2019 International Conference on Multimodal Interaction*, pages 595–601, 2019.

[26] Xin Zhou. Mmrec: Simplifying multimodal recommendation. In *Proceedings of the 5th ACM International Conference on Multimedia in Asia Workshops*, pages 1–2, 2023.

[27] Shan Gao, Junwei Jin, Bicao Li, Cuijuan Lou, and Yuying Jiang. Pairwise Preference over Multi-type Implicit Feedback Confidence Based Bayesian Personalized Ranking for Collaborative Filtering. In *2023 IEEE 11th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, volume 11, pages 1989–1992, 2023. IEEE.

[28] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 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*, pages 639–648, 2020.

[29] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Mengqi Zhang, Shu Wu, and Liang Wang. Latent structure mining with contrastive modality fusion for multimedia recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 2022. IEEE.