

Article

Multimodal Recommendation System Based on Cross Self-Attention Fusion

Peishan Li ^{1,2,3} , Weixiao Zhan ⁴ , Lutao Gao ^{1,2,3} , Shuran Wang ^{1,2,3} and Linnan Yang ^{1,2,3,*} ¹ College of Big Data, Yunnan Agricultural University, Kunming 650201, China² Yunnan Engineering Technology Research Center of Agricultural Big Data, Kunming 650201, China³ Yunnan Engineering Research Center for Big Data Intelligent Information Processing of Green Agricultural Products, Kunming 650201, China⁴ College of Computer Science and Engineering, University of California, San Diego, CA 92093, USA

* Correspondence: 1985008@ynau.edu.cn

Abstract: Recent advances in graph neural networks (GNNs) have enhanced multimodal recommendation systems' ability to process complex user–item interactions. However, current approaches face two key limitations: they rely on static similarity metrics for product relationship graphs and they struggle to effectively fuse information across modalities. We propose MR-CSAF, a novel multimodal recommendation algorithm using cross-self-attention fusion. Building on FREEDOM, our approach introduces an adaptive modality selector that dynamically weights each modality's contribution to product similarity, enabling more accurate product relationship graphs and optimized modality representations. We employ a cross-self-attention mechanism to facilitate both inter- and intra-modal information transfer, while using graph convolution to incorporate updated features into item and product modal representations. Experimental results on three public datasets demonstrate MR-CSAF outperforms eight baseline methods, validating its effectiveness in providing personalized recommendations, advancing the field of personalized recommendation in complex multimodal environments.

Keywords: graph neural networks; multimodal recommendation systems; attention mechanism; personalized recommendation



Academic Editor: Ludovico Minati

Received: 12 December 2024

Revised: 2 January 2025

Accepted: 13 January 2025

Published: 17 January 2025

Citation: Li, P.; Zhan, W.; Gao, L.; Wang, S.; Yang, L. Multimodal Recommendation System Based on Cross Self-Attention Fusion. *Systems* **2025**, *13*, 57. <https://doi.org/10.3390/systems13010057>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The rapid growth of the mobile internet, e-commerce [1], and social media platforms [2] has created an overwhelming array of choices for online users. To address information overload and help users efficiently discover relevant products [3], recommender systems have emerged as crucial tools for personalized recommendations [4]. Traditional recommendation algorithms mainly rely on historical user behavioral data, such as browsing and purchase records, to model user–item similarities [5–9]. They are limited by their single source of information, overlooking valuable multimodal information embedded in product images and text descriptions. Recommendation algorithms that can fuse multimodal data have gradually become a hot research topic [10].

Multimodal recommendation systems enhance recommendation performance by leveraging diverse data types to supplement user interaction history [11]. The core of multimodal recommender systems lies in the fusion and modeling of heterogeneous data, which mainly include two modal forms: text and image [12]. Researchers have proposed a variety of fusion strategies, which can be mainly categorized into two main types: early fusion and late fusion [13]. Early fusion improves feature representation by combining

information from multiple modalities at the initial stage of data processing [14]; late fusion integrates the results of different modalities at the output stage of the model to improve the accuracy of decision making [15].

Recent success of transformer models has highlighted attention mechanisms' effectiveness in recommendation systems for weighting important interaction features and capturing user preferences [16–24]. However, attention mechanisms can be vulnerable to noisy data, potentially overweighting invalid signals. To address this limitation, studies have attempted to combine the attention mechanism with graph neural networks (GNNs) to better handle complex relationships and multimodal data [11]. MMGCN [25] constructs a dichotomous user–item graph for each modality and applies GNN on these graphs to learn the feature representations of users and items. MGAT [18] captures the fine-grained preferences of users for different modalities by constructing multimodal interaction graphs. It also introduces a gate attention mechanism to adaptively capture the preferences of users for different modalities during the information dissemination process. GRCN [26] identifies and prunes potential false alarm interaction edges through a graph refinement layer to optimize the structure of the user–item interaction graph. DualGNN [27] employs a dual graph structure—user–microvideo bipartite graph and user co-occurrence graph—to collaboratively learn user–item interaction graphs leveraging correlation between users. The correlation between users is used to collaboratively learn each user-specific fusion modality. EgoGCN [28] is able to adaptively extract multimodal information at the edge level and adjust features of unimodal nodes under the supervision of other modalities. CAMgr [16] focuses on using user preferences to guide model training, extracting unique features along with generic feature information, and the cross-attention mechanism focuses on improving the representation and fusion of information through user and item interaction ID features. However, the above methods may lead to insufficient learning of certain modes by the model in case of modal imbalance, thus affecting the overall recommendation effect.

Item–item isomorphisms are introduced into recommendation systems due to their ability to mine potential associations between items. HCGCN [29] and LATTICE [30] further capture the potential relationships between user behavior patterns and items through graph convolution operations. However, this structure may face challenges in dynamic recommendation environments to update delays, which affects the effectiveness of real-time recommendations. LUDP [31] explores potential relationships between items by constructing an item–item similarity graph based on multimodal features. In addition, LUDP constructs a user preference graph that captures users' preferences for different modalities through their historical interaction behavior with items. FREEDOM [10] proposed a method to freeze the item–item graph structure, which reduces computational and memory overhead by constructing the graph before training and keeping it fixed during training. FREEDOM also introduced a degree-sensitive edge pruning method for denoising interaction graphs, which rejects possible noisy edges with high probability when sampling the graphs. POWERec [8] introduced prompt-based user interest learning and weak modality augmented training, which models multimodal user interests through shared user embeddings and modality-specific prompts. Although FREEDOM [10] and POWERec [8] focus on optimizing computational and storage overheads while improving model performance, these models still need to deal with the inhomogeneity of modal information and inter-modal dependencies in real-world applications in order to model both products and users more accurately.

To solve the above challenges, we propose MR-CSAF, an adaptive multimodal recommendation algorithm based on cross-self-attention fusion. As shown in Figure 1, MR-CSAF builds on FREEDOM by adding an adaptive modality selector and a cross-self-attention

fusion module. First, MR-CSAF's adaptive modality selector dynamically adjusts the weights of different modalities to ensure that the features of each modality are learned in a balanced manner during training, even under data imbalance. MR-CSAF's adaptive modality selector is able to dynamically adjust the attention between different product image and text modalities according to the current interaction context and user preferences. This dynamic adaptation mechanism enables MR-CSAF to adaptively optimize modal contributions in different scenarios, thus improving the accuracy and robustness of the recommendation system. The cross-self-attention fusion module propagates information between text and images, transferring information within each modality through the self-attention mechanism. Using this module, we can explicitly model the interactions between text, images, and interactions within modalities. These features are later propagated and aggregated across the user-item graph through a GNN. We performed experiments on three publicly available benchmark datasets. Our experimental results show that our model outperforms existing baseline approaches.

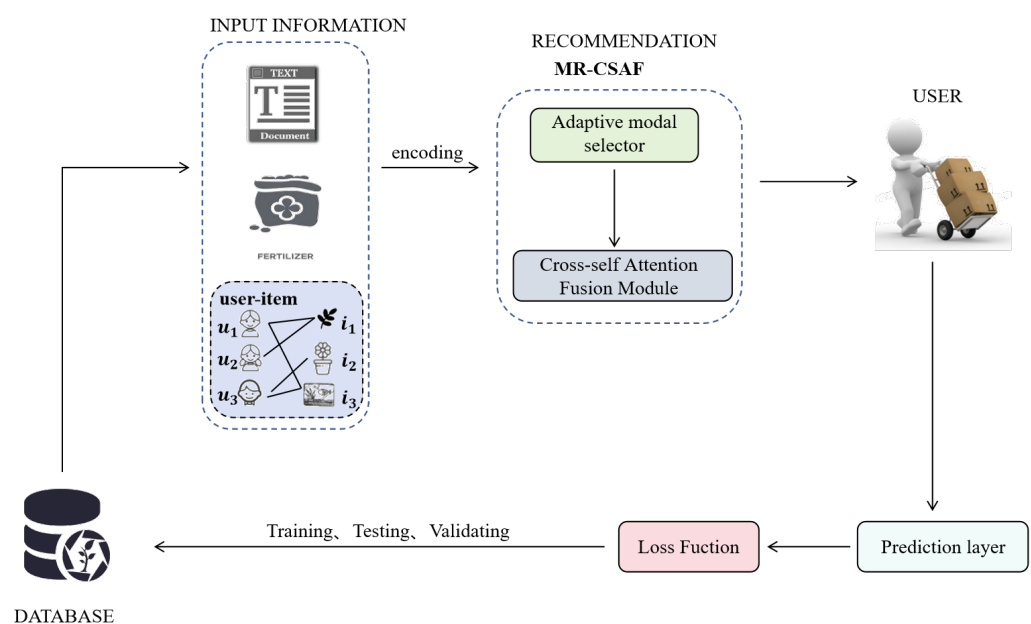


Figure 1. General diagram of MR-CSAF.

In summary, our main contributions include the following:

1. We propose a new recommendation algorithm called MR-CSAF. It leverages an adaptive modal selector and a cross-self-attention fusion mechanism that aims at accurately modeling both products and users.
2. We propose an adaptive modal selector that constructs a potential multimodal item relationship graph by dynamically adjusting modal weights, significantly enhancing the information fusion process in multimodal recommendation systems.
3. We designed a fusion module based on a cross-self-attention mechanism that explores intra- and inter-modal interactions, allowing for more comprehensive and efficient processing and fusion of multimodal information.
4. We evaluate MR-CSAF performance against several baseline approaches on three public benchmark datasets for multimodal recommendation, demonstrating superior performance compared to existing models.

The rest of this paper is organized as follows: A detailed description of our proposed model approach, MR-CSAF, is presented in Section 2. We describes the dataset and our

optimization process in Section 3. In Section 4, we conduct model-related experiments and analyze the results. Section 5 provides a brief conclusion of this paper.

2. Proposed Method

As shown in Figure 2, MR-CSAF uses Sentence2Vec [32] and DeepCNN [33] to extract features from texts and images, respectively, constructs a KNN modality-aware graph using original features for each modality m , and aggregates the modality-aware graphs using an adaptive modality selector. Next, the extracted feature information is fed into the cross-self-attention fusion module to learn the inter- and intra-modal interactions of text and images. Finally, the user-modal features are linked to the corresponding user-modal features and propagated through the user–item interaction graph to obtain the final feature representation. This representation is passed through another prediction layer tailored for downstream recommendation tasks.

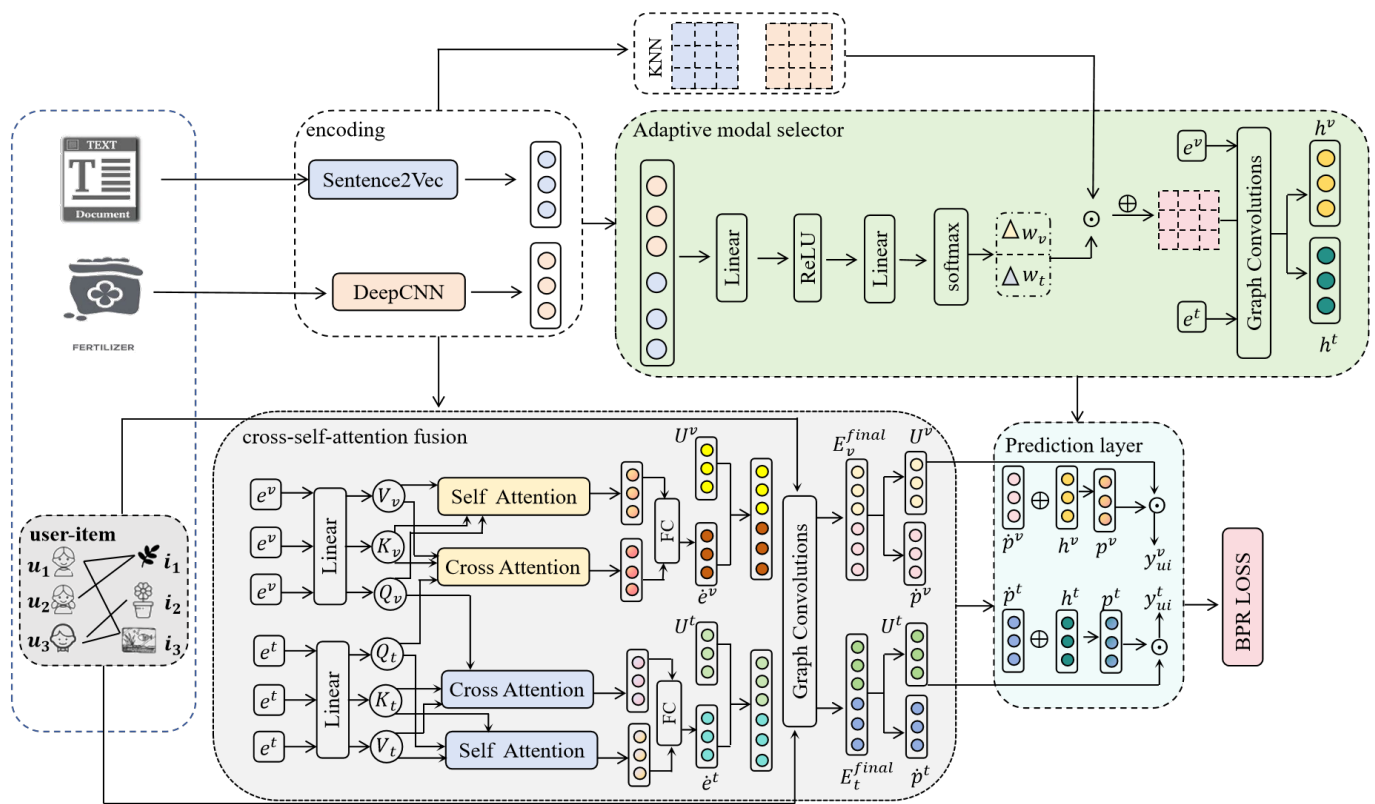


Figure 2. Internal logic diagram of MR-CSAF.

2.1. Problem Definition

Let U denote the set of users and I denote the set of items. For any user $u \in U$, item $i \in I$, their corresponding modal embedding is denoted as $e_{[u,i]}^m \in \mathbb{R}^{d \times [|U|, |I|]}$ in which d is the embedding dimension, and $m \in M = \{v, t\}$ denotes the visual or textual modality.

Construct a heterogeneous user–item graph $G = (V, E)$, where the nodes are the union of all users and all products, i.e., $V = U \cup I$, the set of edges represents the interaction between users and items, denoted by $E = \{(u, i) | u \in U, i \in I\}$, where each edge connects a user and an item. By using a matrix $Q \in \mathbb{R}^{|U| \times |I|}$ based on user–item interactions, a symmetric adjacency matrix $Z \in \mathbb{R}^{|V| \times |V|}$ expression can be obtained, as shown in Equation (1):

$$Z = \begin{pmatrix} 0 & Q \\ Q^T & 0 \end{pmatrix} \quad (1)$$

In this matrix, $Z_{ui} = 1$ if user u has interacted with item i ; otherwise, $Z_{ui} = 0$.

2.2. Adaptive Modal Selector

Our adaptive modal selector constructs a perceptual graph $\tilde{S}^m$ corresponding to each modal based on the initial features of each modality. Cosine similarity, denoted as S_{ij}^m , is used to assess the similarity between two products, as shown in Equation (2).

$$S_{ij}^m = \frac{(e_i^m)^T e_j^m}{\|e_i^m\| \|e_j^m\|} \quad (2)$$

where e_i^m and e_j^m denote the features of item i and item j on modality m , respectively. For each product, we only keep their front edges that are similar, as shown in Equation (3):

$$\dot{S}^m = \begin{cases} 1 & S_{ij}^m \in \text{top} - k(S_i^m) \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

We normalize the discrete adjacency matrix S_{ij}^m to $\tilde{S}^m = (D^m)^{-\frac{1}{2}} \dot{S}^m (D^m)^{-\frac{1}{2}}$, where $D_{ii}^m = \sum_j \dot{S}_{ij}^m$, $D^m \in \mathbb{R}^{|I| \times |I|}$ is the diagonal matrix of $\dot{S}^m$.

To solve the problem of multimodal information fusion, we design an adaptive modal selector that dynamically weights between different modalities m to generate a comprehensive modal similarity matrix. Specifically, the image embedding e_i^v of the product is spliced with the text embedding e_i^t to form a joint feature representation. The modality selector module uses these common features to dynamically generate image and text modality weights for each product, as shown in Equation (4).

$$e_i = \text{concat}(e_i^v, e_i^t) \quad (4)$$

Based on the spliced features, the modality selector module generates the modal weights through a two-layer fully connected neural network, as shown in Figure 3, which ultimately generates a two-dimensional vector representing the weights of the image and text modalities of each product, respectively. Finally, the output is normalized by a softmax operation, as shown in Equation (5).

$$w = \text{softmax}(W_2 \cdot \sigma(W_1 \cdot e_i + b_1) + b_2) \quad (5)$$

where W_1 , W_2 is the learnable matrix, b_1 , b_2 is the bias term, $\sigma(\cdot)$ is the ReLU activation function, and $\text{softmax}(\cdot)$ ensures that the outputs are modal weights, generating the corresponding image and text modal weights denoted as $\omega = [\omega_v, \omega_t]$ and $\omega_v + \omega_t = 1$.

The inter-product similarity graph $S \in \mathbb{R}^{|I| \times |I|}$ is constructed by aggregating the structure of each modality, as shown in Equation (6):

$$S = \sum_{m \in M} \omega_m \tilde{S}^m \quad (6)$$

Compared to traditional fixed-weight approaches, the adaptive modal selector dynamically generates modal weights in an adaptive and learnable manner, enabling more precise fusion of multimodal information and capturing complex inter-modal interactions more effectively.

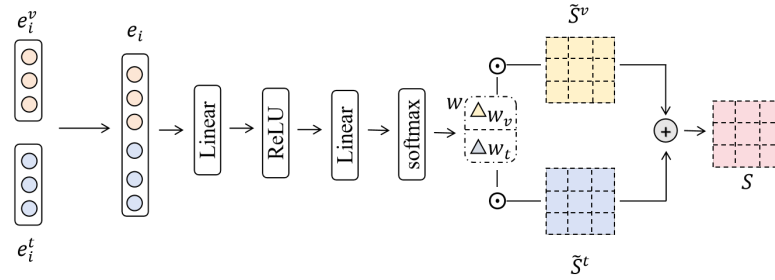


Figure 3. Internal structure of adaptive modal selector.

We apply graph convolution on the inter-product similarity graph to facilitate information propagation and aggregation, as defined in Equation (7):

$$h_i^{m^l} = \sum_{j \in N(i)} S_{ij} h_j^{m^{l-1}} \quad (7)$$

where $N(i)$ is the neighboring node of item i on graph S , and $h_i^{m^l} \in \mathbb{R}^d$ denotes the l -th layer representation of the item, $h_i^{m^0} = e_i^m$. We superimpose a convolutional layer a on graph S , and the last layer is denoted as $h^m = h^{m^a}$.

2.3. Cross-Self-Attention Fusion Module

The cross-self-attention fusion module aims to use attention mechanism to propagate information between text and images and within each modality. Using the cross-self-attention module, we can model the intra-modal and inter-modal interactions of text and images. The cross-attention mechanism uses one modality to compute the distribution of attention, which is then used to guide the information flow on the other modality. Firstly, we define the query (Q), key (K) and value (V) in the cross-self-attention mechanism using Equations (8) and (9):

$$Q_v = W_Q^v e^v, K_v = W_K^v e^v, V_v = W_V^v e^v \quad (8)$$

$$Q_t = W_Q^t e^t, K_t = W_K^t e^t, V_t = W_V^t e^t \quad (9)$$

where $W_Q^v, W_K^v, W_V^v, W_Q^t, W_K^t, W_V^t \in \mathbb{R}^{d \times d}$ is the learnable weight matrix. We then use the dot product between query and key to reflect the correlation between different modalities. Next, we obtain the cross-attention weights by normalizing them with the softmax function, as shown in Equations (10) and (11):

$$\omega_{v \rightarrow t} = \text{softmax} \left(\frac{Q_t K_v^T}{\sqrt{d}} \right) \quad (10)$$

$$\omega_{t \rightarrow v} = \text{softmax} \left(\frac{Q_v K_t^T}{\sqrt{d}} \right) \quad (11)$$

Finally, we apply cross-attentional weights to value (V) to obtain a weighted output, as illustrated in Equations (12) and (13).

$$\Delta e^{v \rightarrow t} = \omega_{v \rightarrow t} V_v \quad (12)$$

$$\Delta e^{t \rightarrow v} = \omega_{t \rightarrow v} V_t \quad (13)$$

where $\Delta e^{v \rightarrow t}$, $\Delta e^{t \rightarrow v}$ denote the propagation of information from image to text and text to image, respectively. Finally, we update the feature embedding of this modality with the propagation information of another modality, as shown in Equations (14) and (15).

$$c^t = e^t + \Delta e^{v \rightarrow t} \quad (14)$$

$$c^v = e^v + \Delta e^{t \rightarrow v} \quad (15)$$

We summarize the process of the cross-self-attention mechanism, as shown in Equations (16) and (17).

$$\Delta e^m = \text{softmax}\left(\frac{Q_m K_m^T}{\sqrt{d}}\right) V_m \quad (16)$$

$$\tilde{e}^m = e^m + \Delta e^m \quad (17)$$

where $\Delta e^m \in \mathbb{R}^{|I| \times d}$ denotes the information propagated within the modality m . Let c^m and $\tilde{e}^m$, $m \in \{v, t\}$ denote the outputs of different modalities of the cross-attention layer and the self-attention layer, respectively. We fuse the acquired features of the self-attention mechanism with those of the cross-attention mechanism through the fully connected layer to obtain the corresponding item modal features $\dot{e}^m$, as shown in Equation (18).

$$\dot{e}^m = W(\text{concat}(\dot{e}^m + \tilde{e}^m)) + b \quad (18)$$

where W is the learnable weight matrix, and b denotes bias.

To process the graph convolution operation on the user-item graph G for information propagation and aggregation, we follow the approach adopted by FREEDOM [10]. Firstly, user-item homomodal feature embeddings are spliced to form a tensor containing the homomodal feature embeddings, as shown in Equation (19). Then a multi-layer graph convolution operation is performed using the adjacency matrix to update the embedding representation layer by layer. At each layer, we update the modal embedding by multiplying sparse matrices, as shown in Equation (20).

$$E_m^0 = \text{concat}(e_U^m, \dot{e}^m) \quad (19)$$

$$E_m^n = Z E_m^{n-1} \quad (20)$$

where e_U^m is the user feature embedding in modality m , n is the number of convolutional layers, and E_m^n is the embedding in the n -th layer. After n layers of message passing, we obtain a sequence of embeddings $E_m^{\text{all}} = \{E_m^0, \dots, E_m^n\}$. To synthesize information from different layers, we compute the mean using Equation (21):

$$E_m^{\text{final}} = \frac{1}{n} \sum_{l=0}^n E_m^l \quad (21)$$

Lastly, we split the obtained final embedding into a product embedding and user embedding, as shown in Equations (22) and (23).

$$\dot{p}^m = E_m^{\text{final}}[n_{\text{users}} :] \quad (22)$$

$$\dot{e}_U^m = E_m^{\text{final}}[: n_{\text{users}}] \quad (23)$$

2.4. Prediction Layer

Based on the modal feature data of the project obtained from Sections 2.2 and 2.3, we adopt the feature summation approach to obtain the final modal representation of the item, as shown in Equation (24).

$$p^m = h^m + \dot{p}^m \quad (24)$$

Finally, the prediction score is expressed as the dot product of the user and product embeddings in the same modality, as shown in Equation (25).

$$y_{ui}^m = (u^m)^T p_i^m \quad (25)$$

2.5. Loss Function

After obtaining the modal feature representations of all products and users, we use Bayesian personalized ranking (BPR) loss to optimize the interaction performance of the user and the item. Specifically, MR-CSAF computes the BPR loss for different modalities separately, as shown in Equation (26). The BPR loss aims to make the difference between the scores of the positive samples and those of the negative samples as large as possible, and the data it is trained on triples (u, i, j) , in which user u 's chooses product i over product j .

$$L_{BPR} = - \sum_{m \in M} \sum_{(u,i,j) \in D} \log \sigma(y_{ui}^m - y_{uj}^m) \quad (26)$$

where D is the set of training examples and $\sigma(\cdot)$ is the sigmoid function.

Moreover, MR-CSAF further transforms the original feature embeddings using a line projection layer, as shown in Equation (27). This transformation allows the model to capture more complex interactions between user features and user preferences in modality m . It serves as a regularization mechanism that ensures robust learning of the embeddings. We compute their BPR losses with user embeddings separately, as shown in Equation (28).

$$f_i^m = W_m e_i^m + b_m \quad (27)$$

$$L_{feat} = - \sum_{m \in M} \sum_{(u,i,j)} \log \sigma((u^m)^T (f_i^m - f_j^m)) \quad (28)$$

where W_m is the linear transformation matrix in modes m , and b_m is the bias vector.

The final total loss, shown in Equation (29), combines the BPR losses for the different modalities and the regularized losses for the features.

$$L_{total} = L_{BPR} + \omega L_{feat} \quad (29)$$

where the weight parameter ω is used to control the effect of feature transformation loss.

3. Experiment

3.1. Research Question

To evaluate the effectiveness and efficiency of MR-CSAF, we conduct experiments on three real-world datasets provided by Amazon to answer the following questions:

- RQ1: Can MR-CSAF outperform current mainstream multimodal recommendation algorithms on baseline tasks?
- RQ2: Do innovative components of MR-CSAF contribute to its performance?
- RQ3: How do hyperparameters affect MR-CSAF's performance?

3.2. Datasets

We conducted our experiments on three public review datasets, Garden, Baby, and Sports, provided by Amazon [34]. As detailed in Table 1, each dataset includes users' product ratings and textual information related to the products, encompassing the product name, brand, and description. Notably, beyond textual data, these datasets also offer visual information about each product, as illustrated in Table 2.

Table 1. Statistics of the datasets.

Dataset	# Users	# Items	# Interactions	# Sparsity
Garden	1686	961	13,274	99.18%
Baby	19,445	7050	160,792	99.88%
Sports	35,598	18,357	296,337	99.95%

Table 2. Introduction to the different datasets.

Dataset	Introduction	Visual Information	Text Information
Garden	Waste materials, pesticides and tools, seeds, etc.		title: Victor M231 Ultimate Flea Trap Refills, 3 Per Pack. description: These flea trap refills have a sweet odor inserted in the specially formulated sticky glue disc. Fleas do not stand a chance. Our flea traps let pet owners see the results. The non-poisonous and odorless trap refills enable safe placement of the flea trap around children and pets.
Baby	Mother and baby products, such as bottles, strollers, etc.		title: Nuby 2 Pack Soft Sipper Replacement Spout, Clear. description: 2 pack Soft Sipper Replacement Spouts fits standard neck bottles. Made from soft durable silicone. Non-drip design helps prevent leaks and spills.
Sports	Sports products such as shoes, shirts, etc.		title: Wenzel Multi Purpose Ground Mat. description: The multi-purpose ground mat is constructed of weatherproof and water resistant woven polypropylene material. Folds down to carry size with built-in handles and pockets. Be ready for sand, ground and grass protection with this handy multi purpose personal gear mat. feature: 'Easy to use', 'High quality product', 'Manufactured in China'.

3.3. Baseline Methods

We compared MR-CSAF with eight popular baseline models for multimodal recommendation tasks:

- VBPR [35]: Combining visual features and implicit feedback data to incorporate the visual features of goods into the recommendation system in order to improve the accuracy of personalized ranking.
- MMGCN [25]: By processing graph data at different scales, it is able to capture both local and global feature information.
- GRCN [26]: Introducing gating mechanisms and relational information to optimize the interaction modeling between users and items so as to enhance the recommendation effect.
- SLMRec [36]: Capturing both local and global feature information by processing graph data at different scales.

- LATTICE [30]: Capturing the complex relationship between user preferences and item features and learning the underlying semantic item–item structure for recommendation.
- BM3 [37]: Removing the requirement of randomly sampling negative instances required for interactions between users and items in the model and uses a latent embedding discard mechanism to perturb the original user’s and item’s embeddings.
- FREEDOM [10]: Improving recommendation accuracy by freezing the item–item graph structure and denoising the user–item interaction graph structure, while reducing memory consumption and improving computational efficiency.
- POWERec [8]: Modeling modality-specific user interests through basic user embeddings and different modality cues.

3.4. Experimental Setup and Evaluation Metrics

One NVIDIA GeForce RTX 3090 along with CUDA 11.7, Python 3.7, and PyTorch 1.13.1 were used to accelerate the computation. Following the best practice suggested by FREEDOM [10], the embedding size was set to 64 for all modeled users and products. The embedding parameters were initialized using the Xavier method and trained with the Adam optimizer, with a learning rate set to 0.001 and a batch size configured to 2048. We evaluated MR-CSAF on Recall@10, Recall@20, NDCG@10, and NDCG@20.

We used the dataset at 80% for training, 10% for validation, and 10% for testing, ensuring that the user–item interaction records in each set were randomly selected to avoid bias. We used Recall@20 on the validation set and 20 epochs as the early-stopping criterion and patient, respectively. The model that performed best on the validation test was selected for the final evaluation on the test set.

4. Results and Discussion

4.1. Overall Performance (RQ1)

We compare the effect of this model with the current mainstream eight baseline models under three datasets, as shown in Table 3. Based on Table 3, we make following observations:

1. Our experiments revealed key insights across three datasets. In the Garden dataset, which contains richer interaction data compared to the other datasets, MR-CSAF demonstrated improvements across all four evaluation metrics. These improvements can be attributed to the cross-self-attention fusion module’s ability to effectively capture both intra- and inter-modal interactions, leading to more accurate product modeling. The Baby dataset, despite its large scale of users and products, presents significant challenges due to its high sparsity. However, MR-CSAF successfully maintains its effectiveness by exploring modal interactions and accurately modeling information even with sparse data relationships. In the Sports dataset, which is the largest and also exhibits high sparsity, MR-CSAF’s adaptive modality selector and cross-attention mechanism prove particularly valuable. The adaptive modality selector dynamically adjusts modal weights to optimize the integration of image and text information, while the cross-attention mechanism enables effective multimodal fusion even under sparse data conditions, ultimately enhancing recommendation quality.

2. MR-CSAF outperformed all eight baseline models on NDCG@10, NDCG@20, Recall@10, and Recall@20 metrics. Our superior performance comes from following:

- (1) Our proposed method effectively captures multimodal information by modeling product information in each dimension and propagating and aggregating it over the graph to accurately generate user modal embeddings.
- (2) Our adaptive modal selector combines the distinct dimensional characteristics of various products and dynamically calculates weights for image and text modalities. This effectively adjusts the influence of each modality across different product recommen-

- dations, enabling adaptive adjustments within the product–product graph structure. This approach optimizes the alignment of modality contributions, enhancing the recommendation system’s flexibility and responsiveness to diverse product attributes.
- (3) Our cross-self-awareness fusion module effectively captures both cross-modal and intra-modal interactions, enabling precise fusion and modeling of product attribute features alongside modality-specific features. This approach enhances the integration of diverse modality information for a more accurate representation of product attributes.

3. In modeling multimodal feature information, graph-based multimodal recommendation models outperform the traditional matrix decomposition model VBPR [35]. This is primarily because our method can effectively capture complex relationships between users and items, enhancing contextual understanding through multimodal information. Furthermore, the graph structure better represents sparse data, improving recommendation accuracy, especially when handling high-dimensional features of users and items.

4. In graph-based multimodal recommendation models, GRCN [26] performs better than MMGCN [25], which is a result of GRCN capturing and utilizing users’ true preference information more effectively through graph structure optimization and targeted noise processing. SLMRec [36] primarily learns product feature representations through self-supervised learning; however, its reliance on individual user embeddings to simulate user features across different modalities results in weaker performance. LATTICE [30] achieves stronger results on the Garden dataset by explicitly modeling item relationships and leveraging negative samples to guide learning, which helps the model distinguish between positive and negative samples in smaller datasets. In contrast, BM3 [37] employs a self-supervised learning framework that does not depend on negative samples, making it more robust in handling larger datasets. The multimodal recommendation approaches LATTICE [30], FREEDOM [10], and POWERec [8], which all aim to enhance recommendation performance through graph structure optimization and multimodal data fusion. LATTICE [8] builds latent graph structures by dynamically assigning weights to textual and graphical modalities, effectively capturing complex relationships between items, although it may encounter efficiency challenges with large datasets. FREEDOM, in contrast, improves model efficiency by freezing the graph structure, which reduces computational complexity and noise, albeit at the cost of reduced flexibility. POWERec [8] utilizes cue learning and weak modality augmentation to efficiently model user preferences across modalities but requires extensive tuning for very large graphs, resulting in diminished performance on the Sports dataset due to its substantial data volume.

In contrast, the MR-CSAF method dynamically adjusts modal preferences across varying product dimensions based on feature information, allowing it to capture potential graph structures on the item–item graph more precisely. Additionally, MR-CSAF introduces a cross-self-attention mechanism to construct more accurate product feature representations and refines user modality feature representations by aggregating and propagating information within the user–item graph, achieving exceptional performance.

Table 3. Comparison of MR-CSAF against baselines on model efficiency. Best results are presented in bold; second-best results are underlined.

Datasets	Metrics	VBPR	MMGCN	GRCN	SLMRec	LATTICE	BM3	FREEDOM	POWERec	MR-CSAF
Garden	NDCG@10	0.0547	0.0655	0.0758	0.0747	<u>0.0849</u>	0.0835	0.0791	0.0748	0.0948
	NDCG@20	0.0709	0.0826	0.0945	0.0922	0.1022	<u>0.1034</u>	0.0961	0.0914	0.1136
	Recall@10	0.1030	0.1155	0.1361	0.1345	<u>0.1571</u>	0.1429	0.1376	0.1262	0.1637
	Recall@20	0.1651	0.1823	0.2090	0.2019	<u>0.2242</u>	0.2199	0.2026	0.1910	0.2379

Table 3. Cont.

Datasets	Metrics	VBPR	MMGCN	GRCN	SLMRec	LATTICE	BM3	FREEDOM	POWERec	MR-CSAF
Baby	NDCG@10	0.0223	0.0220	0.0282	0.0285	0.0292	0.0301	<u>0.0330</u>	0.0311	0.0344
	NDCG@20	0.0284	0.0282	0.0358	0.0357	0.0370	0.0383	<u>0.0424</u>	0.0398	0.0444
	Recall@10	0.0423	0.0421	0.0532	0.0540	0.0547	0.0564	<u>0.0627</u>	0.0579	0.0641
	Recall@20	0.0663	0.0660	0.0824	0.0810	0.0850	0.0883	<u>0.0992</u>	0.0918	0.1033
Sports	NDCG@10	0.0307	0.0209	0.0306	0.0374	0.0335	0.0355	<u>0.0385</u>	0.0300	0.0394
	NDCG@20	0.0384	0.0270	0.0389	0.0462	0.0421	0.0438	<u>0.0481</u>	0.0377	0.0498
	Recall@10	0.0558	0.0401	0.0559	0.0676	0.0620	0.0656	<u>0.0617</u>	0.0565	0.0734
	Recall@20	0.0856	0.0636	0.0877	0.1017	0.0953	0.0980	<u>0.1089</u>	0.0863	0.1130

4.2. Ablation Studies (RQ2)

In this section, we decouple the proposed model and evaluate the contribution of each component to the recommendation accuracy. We choose Recall@20 and NDCG@20 as metrics for the ablation experiments and validate them on three datasets, as shown in Table 4. Based on the architecture of MR-CSAF, we designed the following variant of MR-CSAF:

- Adaptive modal selector (AMS): AMS enhances the Ours model by incorporating adaptive modality selectors but does not utilize the cross-self-attention mechanism to manage embeddings.
- Cross-Self-Attention Fusion (CSAF): CSAF learns interactions between different modalities. Within the same modality, it only uses the cross-self-attention fusion mechanism.

Table 4. Ablation study on a different dataset.

Datasets	Modules	NDCG@20	Recall@20
Garden	AMS	0.1128	0.2378
	CSAF	0.1077	0.2256
	MR-CSAF	0.1136	0.2379
Baby	AMS	0.0433	0.1014
	CSAF	0.0435	0.1012
	MR-CSAF	0.0444	0.1033
Sports	AMS	0.0496	0.1123
	CSAF	0.0486	0.1106
	MR-CSAF	0.0498	0.1130

The results demonstrate that our proposed MR-CSAF consistently outperforms both AMS and CSAF across all three datasets, indicating that both methods enhance the accuracy of model recommendations. Moreover, removing AMS significantly impacts recommendation performance more than removing CSAF in all cases across the datasets. In multimodal recommendation systems, different modalities contain rich and diverse information, making it essential to select and fuse these modalities effectively. In scenarios with sparse data, an adaptive modality selector optimizes information utilization, ensuring the model can still make accurate recommendations despite limited available data. By selectively concentrating on specific modalities, the adaptive modality selector helps alleviate performance degradation caused by imbalanced modal information. Additionally, the cross-self-attention fusion mechanism enhances feature representation by assigning varying weights to different features, enabling the model to prioritize critical information during decision making.

Although both CSAF and CAMgr employ cross-attention mechanisms (CAMgr-CAM), their approaches differ significantly. CSAF focuses on capturing both inter-modal interactions between products and intra-modal information transfers, propagating these refined representations through the user–product graph to enhance both user and product features. In contrast, CAMgr-CAM primarily optimizes the model using user feature information,

processing user–item interactions to identify user interests and match them with product features. Our experimental results, shown in Table 5, demonstrate CSAF’s superior ability to capture complex modal interactions and effectively propagate these refined representations through graph structures.

Table 5. Comparison of CSAF and CAmgr-CAM.

Datasets	Modules	NDCG@20	Recall@20
Baby	CAmgr-CAM	0.0386	0.0847
	CSAF	0.0435	0.1012
Sports	CAmgr-CAM	0.0341	0.0905
	CSAF	0.0486	0.1106

In cross-self-attention, all modalities interact with each other, which can introduce noise from irrelevant modalities. In contrast, the modality selector learns weights based on the specific product embedding, allowing it to assign lower weights to irrelevant modalities. This mechanism prevents noisy modalities from interfering with the decision-making process, thereby enhancing the accuracy of the recommendations.

4.3. Hyperparameter Study (RQ3)

In this subsection, we evaluate the following two key hyperparameters: the feature dimensions d_{hidden} defined in Equation (5) and the weight parameters ω defined in Equation (29).

4.3.1. Impact of Hidden_dim Size in Adaptive Modal Selectors

When tuning the parameters of the adaptive modality selector, we experimentally set the hidden layer sizes d_{hidden} to candidate values of 8, 16, 32, 64, and 128. We then assessed their impact on Recall@20 and NDCG@20, as illustrated in the results of Figure 4. Our findings indicate that a hidden layer size of 16 yields the best performance on the Garden, Baby and Sports datasets. Compared to other hidden layer dimension sizes, $d_{hidden} = 16$ may provide the model with the appropriate expressiveness to capture the complex inter-modal relationships in these three datasets more effectively, while avoiding the risk of overfitting associated with too high a dimension.

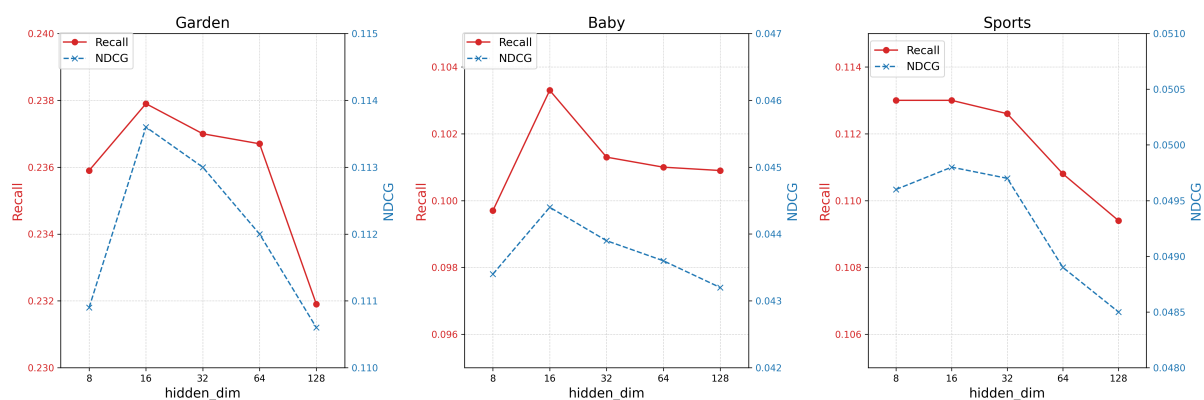


Figure 4. Performance of MR-CSAF with different hidden layer sizes d_{hidden} on the Baby and Garden datasets.

4.3.2. Effect of Feature Transformation Loss ω

We set the MR-CSAF feature transformation loss to the following values: 0.0, 1×10^{-5} , 1×10^{-4} , 1×10^{-3} , 1×10^{-2} , 1×10^{-1} . This was completed to investigate how this hyperparameter affects the model’s recommendation performance. Figure 5 illustrates

its impact on Recall@20 and NDCG@20 across three public datasets. Notably, the value of $1e-05$ achieves the best performance on all datasets. However, increasing this value leads to a decline in performance. A larger loss may hinder the model's ability to fully leverage useful features within the dataset. In this scenario, the contribution of the feature transformation loss L_{feat} to the total loss function L_{total} becomes significantly larger, causing the model to concentrate excessively on feature transformation during training. This heightened focus can make the model more sensitive to noise and outliers in the original data. Conversely, a value that is too small may undermine the model's generalization ability, resulting in poor performance on unseen test data.

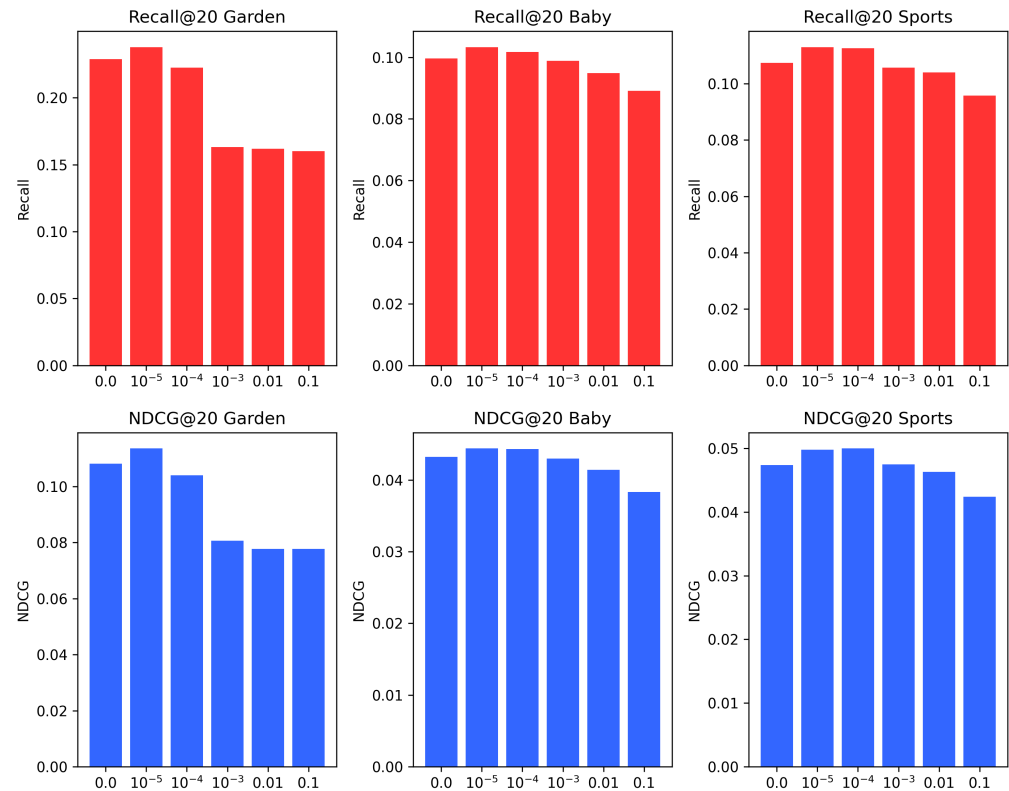


Figure 5. Performance of MR-CSAF with different feature transformation loss ω on the Baby, Garden and Sports datasets.

5. Conclusions

In this paper, we propose MR-CSAF, an adaptive multimodal recommendation algorithm based on cross-attentional fusion. This algorithm significantly improves the accuracy of information fusion by introducing an adaptive modality selector, which dynamically constructs a multimodal item relationship graph. Additionally, it improves the relevance of recommendation results by capturing complex inter- and intra-modal interactions through the cross-self-attention mechanism, thereby boosting recommendation performance. MR-CSAF outperformed eight state-of-the-art methods across three public benchmark datasets, demonstrating its potential in multimodal recommendation systems.

However, the current implementation of MR-CSAF is slow to train due to the increased number of parameters, resulting from the adaptive modality selector. In a future study, we plan to explore techniques that can reduce the number of parameters in the selector, such as model compression, model distillation, and lightweight self-attention mechanisms. Furthermore, we plan to investigate our model's generalization performance on larger and more diverse datasets to further enhance the applicability of MR-CSAF in practical scenarios.

Author Contributions: Conceptualization, P.L. and W.Z.; methodology, P.L.; validation, P.L.; formal analysis, S.W.; investigation, L.G.; resources, L.G.; data curation, S.W.; writing—original draft preparation P.L.; writing—review and editing, P.L. and W.Z.; visualization, P.L.; supervision, L.Y.; project administration, L.Y.; funding acquisition, L.Y. All authors have read and agreed to the published version of the manuscript.

Funding: Financial support was received for the research, authorship, and/or publication of this article. This study was supported by the Yunnan Provincial Science and Technology Major Project (No.202202AE090008), titled “Application and Demonstration of Digital Rural Governance Based on Big Data and Artificial Intelligence”.

Institutional Review Board Statement: Not involving humans or animals.

Data Availability Statement: The data for this study can be downloaded at <http://jmcauley.ucsd.edu/data/amazon/links.html> (accessed on 18 January 2024).

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

References

1. Ding, Y.; Ma, Y.; Wong, W.K.; Chua, T.S. Modeling instant user intent and content-level transition for sequential fashion recommendation. *IEEE Trans. Multimed.* **2008**, *24*, 142–149. [\[CrossRef\]](#)
2. Wu, L.; Chen, L.; Hong, R.; Fu, Y.; Xie, X.; Wang, M. A hierarchical attention model for social contextual image recommendation. *IEEE Trans. Knowl. Data Eng.* **2019**, *32*, 1854–1867. [\[CrossRef\]](#)
3. Yan, C.; Liu, L. Recommendation Method Based on Heterogeneous Information Network and Multiple Trust Relationship. *Systems* **2023**, *11*, 169. [\[CrossRef\]](#)
4. Ma, T.; Huang, L.; Lu, Q.; Hu, S. Kr-gcn: Knowledge-aware reasoning with graph convolution network for explainable recommendation. *ACM Trans. Inf. Syst.* **2023**, *41*, 1–27. [\[CrossRef\]](#)
5. Rendle, S.; Freudenthaler, C.; Gantner, Z.; Schmidt-Thieme, L. BPR: Bayesian personalized ranking from implicit feedback. *arXiv* **2012**, arXiv:1205.2618.
6. 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.
7. Zhou, X.; Lin, D.; Liu, Y.; Miao, C. Layer-refined graph convolutional networks for recommendation. In Proceedings of the 2023 IEEE 39th International Conference on Data Engineering (ICDE), Anaheim, CA, USA, 3–7 April 2023; pp. 1247–1259.
8. Dong, X.; Song, X.; Tian, M.; Hu, L. Prompt-based and weak-modality enhanced multimodal recommendation. *Inf. Fusion* **2024**, *101*, 101989. [\[CrossRef\]](#)
9. Molaie, M.M.; Lee, W. Economic corollaries of personalized recommendations. *J. Retail. Consum. Serv.* **2022**, *68*, 103003. [\[CrossRef\]](#)
10. Zhou, X.; Shen, Z. A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 935–943.
11. Zhou, H.; Zhou, X.; Zeng, Z.; Zhang, L.; Shen, Z. A comprehensive survey on multimodal recommender systems: Taxonomy, evaluation, and future directions. *arXiv* **2023**, arXiv:2302.04473.
12. Liu, Q.; Hu, J.; Xiao, Y.; Zhao, X.; Gao, J.; Wang, W.; Li, Q.; Tang, J. Multimodal recommender systems: A survey. *ACM Comput. Surv.* **2024**, *57*, 1–17. [\[CrossRef\]](#)
13. Gandhi, A.; Adhvaryu, K.; Poria, S.; Cambria, E.; Hussain, A. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Inf. Fusion* **2023**, *91*, 424–444. [\[CrossRef\]](#)
14. Gadzicki, K.; Khamsehashari, R.; Zetzsche, C. Early vs late fusion in multimodal convolutional neural networks. In Proceedings of the 2020 IEEE 23rd International Conference on Information Fusion (FUSION), Rustenburg, South Africa, 6–9 July 2020; pp. 1–6.
15. Wang, Y.; Xu, X.; Yu, W.; Xu, R.; Cao, Z.; Shen, H.T. Combine early and late fusion together: A hybrid fusion framework for image-text matching. In Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 5–9 July 2021; pp. 1–6.
16. Li, K.; Xu, L.; Zhu, C.; Zhang, K. A Multimodal Graph Recommendation Method Based on Cross-Attention Fusion. *Mathematics* **2024**, *12*, 2353. [\[CrossRef\]](#)
17. Wang, R.; Wu, Z.; Lou, J.; Jiang, Y. Attention-based dynamic user modeling and deep collaborative filtering recommendation. *Expert Syst. Appl.* **2022**, *188*, 116036. [\[CrossRef\]](#)
18. Tao, Z.; Wei, Y.; Wang, X.; He, X.; Huang, X.; Chua, T.S. Mgat: Multimodal graph attention network for recommendation. *Inf. Process. Manag.* **2020**, *57*, 102277. [\[CrossRef\]](#)

19. Hu, Z.; Cai, S.M.; Wang, J.; Zhou, T. Collaborative recommendation model based on multi-modal multi-view attention network: Movie and literature cases. *Appl. Soft Comput.* **2023**, *144*, 110518. [\[CrossRef\]](#)
20. He, Q.; Liu, S.; Liu, Y. Optimal Recommendation Models Based on Knowledge Representation Learning and Graph Attention Networks. *IEEE Access* **2023**, *11*, 19809–19818. [\[CrossRef\]](#)
21. Liu, F.; Chen, H.; Cheng, Z.; Liu, A.; Nie, L.; Kankanhalli, M. Disentangled multimodal representation learning for recommendation. *IEEE Trans. Multimed.* **2022**, *25*, 7149–7159. [\[CrossRef\]](#)
22. Liu, F.; Cheng, Z.; Sun, C.; Wang, Y.; Nie, L.; Kankanhalli, M. User diverse preference modeling by multimodal attentive metric learning. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 1526–1534.
23. Wu, C.; Wu, F.; Qi, T.; Zhang, C.; Huang, Y.; Xu, T. Mm-rec: Visiolinguistic model empowered multimodal news recommendation. In Proceedings of the 45th international ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, 11–15 July 2022; pp. 2560–2564.
24. Xun, J.; Zhang, S.; Zhao, Z.; Zhu, J.; Zhang, Q.; Li, J.; He, X.; He, X.; Chua, T.S.; Wu, F. Why do we click: Visual impression-aware news recommendation. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, 20–24 October 2021; pp. 3881–3890.
25. Wei, Y.; Wang, X.; Nie, L.; He, X.; Hong, R.; Chua, T.S. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 1437–1445.
26. Wei, Y.; Wang, X.; Nie, L.; He, X.; Chua, T.S. Graph-refined convolutional network for multimedia recommendation with implicit feedback. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 3541–3549.
27. Wang, Q.; Wei, Y.; Yin, J.; Wu, J.; Song, X.; Nie, L. Dualgnn: Dual graph neural network for multimedia recommendation. *IEEE Trans. Multimed.* **2021**, *25*, 1074–1084. [\[CrossRef\]](#)
28. Chen, F.; Wang, J.; Wei, Y.; Zheng, H.T.; Shao, J. Breaking isolation: Multimodal graph fusion for multimedia recommendation by edge-wise modulation. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 385–394.
29. Mu, Z.; Zhuang, Y.; Tan, J.; Xiao, J.; Tang, S. Learning hybrid behavior patterns for multimedia recommendation. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 376–384.
30. Zhang, J.; Zhu, Y.; Liu, Q.; Wu, S.; Wang, S.; Wang, L. Mining latent structures for multimedia recommendation. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, 20–24 October 2021; pp. 3872–3880.
31. Lei, F.; Cao, Z.; Yang, Y.; Ding, Y.; Zhang, C. Learning the user’s deeper preferences for multi-modal recommendation systems. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *19*, 1–18. [\[CrossRef\]](#)
32. Arora, S.; Liang, Y.; Ma, T. A simple but tough-to-beat baseline for sentence embeddings. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
33. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26–30 June 2016; pp. 770–778.
34. Linden, G.; Smith, B.; York, J. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet Comput.* **2003**, *7*, 76–80. [\[CrossRef\]](#)
35. 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.
36. Tao, Z.; Liu, X.; Xia, Y.; Wang, X.; Yang, L.; Huang, X.; Chua, T. Self-Supervised Learning for Multimedia Recommendation. *IEEE Trans. Multimed.* **2023**, *25*, 5107–5116. [\[CrossRef\]](#)
37. 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.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.