Spectrum-based Modality Representation Fusion Graph Convolutional Network for Multimodal Recommendation

Rongqing Kenneth Ong
Nanyang Technological University
Singapore
rongqing001@e.ntu.edu.sg

Andy W. H. Khong
Nanyang Technological University
Singapore
andykhong@ntu.edu.sg

ABSTRACT

Incorporating multi-modal features as side information has recently become a trend in recommender systems. To elucidate user-item preferences, recent studies focus on fusing modalities via concatenation, element-wise sum, or attention mechanisms. Despite having notable success, existing approaches do not account for the modality-specific noise encapsulated within each modality. As a result, direct fusion of modalities will lead to the amplification of cross-modality noise. Moreover, the variation of noise that is unique within each modality results in noise alleviation and fusion being more challenging. In this work, we propose a new Spectrum-based Modality Representation (SMORE) fusion graph recommender that aims to capture both uni-modal and fusion preferences while simultaneously suppressing modality noise. Specifically, SMORE projects the multi-modal features into the frequency domain and leverages the spectral space for fusion. To reduce dynamic contamination that is unique to each modality, we introduce a filter to attenuate and suppress the modality noise adaptively while capturing the universal modality patterns effectively. Furthermore, we explore the item latent structures by designing a new multi-modal graph learning module to capture associative semantic correlations and universal fusion patterns among similar items. Finally, we formulate a new modality-aware preference module, which infuses behavioral features and balances the uni- and multi-modal features for precise preference modeling. This empowers SMORE with the ability to infer both user modality-specific and fusion preferences more accurately. Experiments on three real-world datasets show the efficacy of our proposed model. The source code for this work has been made publicly available at https://github.com/kennethorq/SMORE.

CCS CONCEPTS

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

KEYWORDS

Multi-modal Recommendation, Graph Neural Networks, Multi-Modality Fusion

ACM Reference Format:
Rongqing Kenneth Ong and Andy W. H. Khong. 2025. Spectrum-based Modality Representation Fusion Graph Convolutional Network for Multimodal Recommendation. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining (WSDM '25), March 10–14, 2025, Hannover, Germany*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3701551.3703561

WSDM '25, March 10–14, 2025, Hannover, Germany

ACM ISBN 979-8-4007-1329-3/25/03.
https://doi.org/10.1145/3701551.3703561

1 INTRODUCTION

In the rapidly expanding world of e-commerce, recommendation systems play a critical role in assisting users to identify products of interest. The standard user browsing experience encompasses exposure to various forms of multi-modal content [20] intended to captivate and entice users. As user preferences are generally driven by a mixture of modality content, research into multi-modal recommender systems (MRSs) that leverage modalities to infer user interest has been gaining popularity. Studies have shown that the use of modalities outperforms general recommenders that rely solely on users' historical interaction [11, 15, 42, 50, 52].

The central focus of MRSs involves the integration of different modalities within the collaborative filtering (CF) framework. Earlier works such as VBPR [11] and DeepStyle [19] focus on fusing visual-specific modality by first projecting the modality features into a lower-dimensional space before combining it with the item identifier (ID) embeddings using concatenation and summation, respectively. Since user-item interactions can naturally be depicted as a bipartite graph [5, 6, 12, 23, 45, 46], graph neural networks (GNNs) have been adapted to capture the high-order connectivity of both multi-modal and behavioral interactions. To this end, LATTICE [42] constructs the latent structures associated with modalities by first constructing different views of the item graph. Thereafter, it performs modality fusion using an attention mechanism to weigh each modality. As an extension, FREEDOM [50] reduces redundancy in training the latent graphs by freezing them prior to training. More recently, to mitigate challenges associated with modality noise that may be introduced through pre-trained encoders, MGCN [41] integrates behavioral features with modality features in an attempt to reduce such noise. It then employs average pooling to fuse each of the modality features.

While existing MRSs achieve notable success in incorporating multiple modalities, they suffer from a significant drawback—amplification of cross-modality noise during fusion [39, 44]. Although it is essential to infer fusion preferences, existing works [11, 19, 42, 43] combine modalities directly without taking into account the impact of modality-specific noise, which may detrimentally affect the quality of item representations learned. Consider an illustrative case study constructed from the Amazon Clothing [10] dataset. Fig. 1(i) highlights the importance of fusion preferences, where users make purchase decisions from an image and text description. However, capturing fusion preferences via modality fusion carries the risk of modality-specific contamination. For instance, consider a pair of items in (ii), which differ in terms of their functionality—the left item being a toy hammer, while the

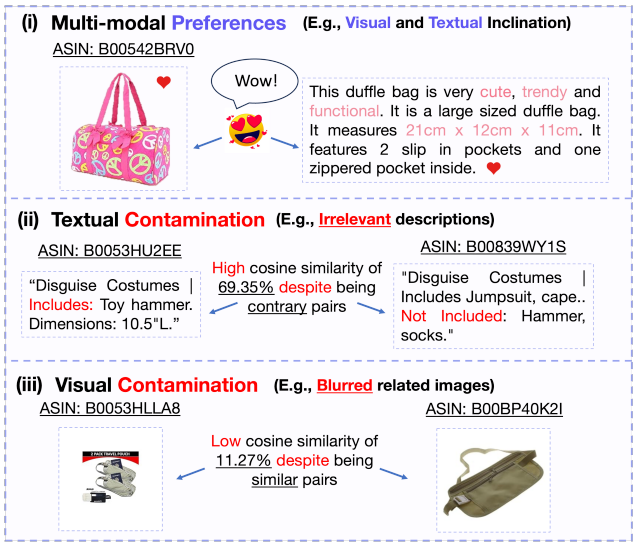

Figure 1: An illustrative example of user multi-modality preferences and issues related to modality-specific contamination. (i) User multi-modal preferences, (ii) irrelevant textual descriptions resulting in an unexpectedly high similarity score of 69.35% even though item pairs are unrelated, and (iii) blurred images resulting in a low similarity score of 11.27% even though pairs are related.

right being a jumpsuit. Due to the irrelevant description for the latter (which mentioned *Hammer*), these items yield a surprisingly high similarity score of 69.35% despite them being uncorrelated. In terms of visual contamination, as shown in Fig. 1(iii), a similar pair of items (e.g., travel pouches) yields an unexpectedly low similarity score of 11.27% when one of the images suffers from a blurring effect. As a result, the noise within each modality will be amplified further during fusion, thereby corrupting the item encoding process. Nonetheless, reducing noise contamination within each modality is highly challenging due to their unique and dynamic contamination characteristics [1, 14]. To reliably ascertain the multi-modal preferences of a user, a fusion module must be designed meticulously to mitigate the undesired effect of noise within each modality before fusion so as to capture universal patterns effectively while preserving essential uni-modal features.

In this work, we draw inspiration from the field of signal processing, which has shown to be effective for modality denoising [22, 26]. By projecting signals to the frequency domain via the Fourier transform [13], a sparse frequency spectrum is generated. This unique characteristic facilitates the acquisition of discriminative spectrum [18], which enables the distillation of critical modality features by means of effective attenuation. Furthermore, the frequency domain offers a comprehensive global perspective [26, 38], thereby enabling each spectral component to attend to all spatial domain features efficiently and effectively. Inspired by the discriminative spectral property and the global perspective by the frequency domain and to overcome the aforementioned drawbacks, we propose a new Spectrum-based Modality Representation (SMORE) fusion graph recommender that aims to capture both uni-modal and fusion preferences while concurrently suppressing modality noise originating from raw features. In particular, SMORE comprises three key components: 1. Spectrum Modality Fusion, 2. Multi-modal Graph Learning, and 3. Modality-Aware Preference module.

To capture the universal modality patterns holistically for inferring user fusion preferences, SMORE performs early fusion by projecting modality features into the frequency domain using the Fourier transform. Harnessing the global perspective inherent within the frequency domain, SMORE captures the cross-modality universal patterns effectively through an efficient point-wise aggregation. Given the discriminative spectrum features, a dynamic filter is then formulated to attenuate and suppress irrelevant (noise) signals adaptively during the fusion process, ensuring that only essential sequence and spatial features are transmitted and fused.

Furthermore, we exploit the correlations between collaborative and latent structures by designing a new multi-modal graph learning module to encode high-order collaborative and relational signals from two distinct perspectives: user-item and item-item modality views, respectively. A new modality-aware preference module is also proposed to capture users' uni- and multi-modal preferences comprehensively. By injecting behavioral signals into the uni-modal and fusion features, SMORE effectively balances and achieves a more concise modeling of preferences between the uni-modal and fusion content. Experiments conducted on three real-world datasets validate the efficacy of our proposed model.

The contributions of our work are threefold:

- We propose a new spectrum-based modality fusion scheme to fuse modalities associated with different semantics effectively while suppressing the modal-specific noise from its raw content;
- We design a multi-modal graph learning module comprising modal-specific and -fusion views to capture high-order collaborative and semantically correlated signals;
- We formulate a new modality-aware preference module to capture the users' diverse uni- and multi-modal preferences explicitly, reflecting real-world scenarios.

2 RELATED WORKS

2.1 Multi-modal Recommendation

Integrating semantically rich multi-modality features into recommender systems has recently emerged as a predominant way to enhance the accuracy of recommendation systems. One of the conventional approaches to extract multi-modal features is through the use of pre-trained neural networks (e.g., Sentence Transformer [27], VGG-16 [30]). For instance, VBPR [11] employs a pre-trained convolutional neural network (CNN) to extract deep visual features corresponding to the items. Consequently, it performs modality fusion by concatenating the ID and visual embeddings to model user modality-specific preferences. VECF [3], on the other hand, leverages VGG-16 to analyze users' complex preferences for image patches by applying pre-segmentation [4, 29] and an attention model to capture key regions within images.

As user interactions naturally occur in the form of structured data, recent works utilize GNNs to capture such structural information in MRSs. Leveraging multi-modal features (e.g., visual, text, acoustic), DualGNN [34] integrates a user co-occurrence graph

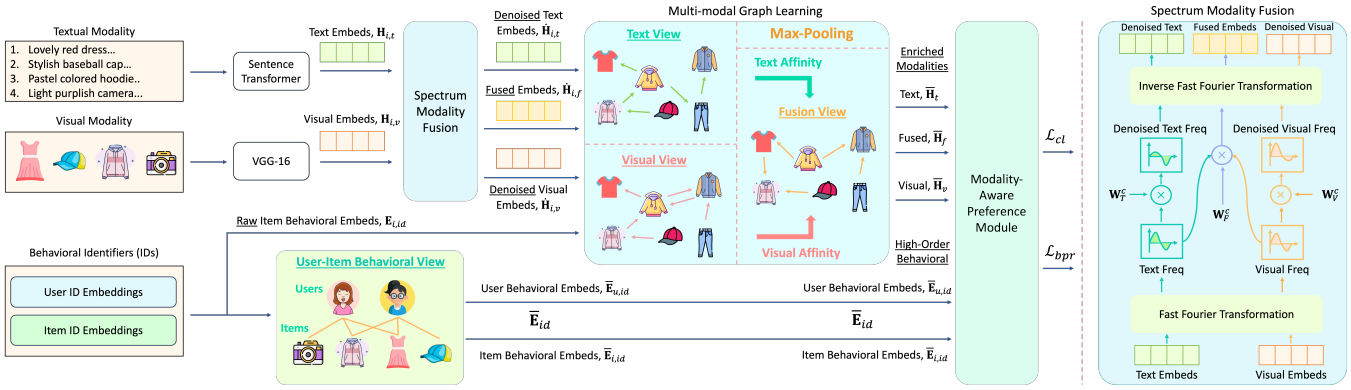

Figure 2: An illustrative overview of the proposed architecture, comprising three key components: (i) spectrum modality fusion, (ii) multi-modal graph learning, and (iii) modality-aware preference module.

and a preference learning module to model different granularities of user preferences. By introducing behavioral information into modality features, MGCN [41] incorporates an attention layer to capture the importance of different modalities. BM3 [52], on the other hand, adopts a new self-supervised learning approach that aims to reduce computational resource demands and rectify incorrect supervision signals arising from negative sampling. While proven effective, these models treat modalities independently and do not account for the fusion preferences that a user may have.

2.2 Modality Fusion Graph-based Learning

Apart from creating modalities of different views via graph neural networks, several graph-based approaches have attempted to fuse the multi-modalities directly. In essence, fusion-based approaches can be categorized into three main stages: early, intermediate, and late fusion [47]. For instance, LATTICE [42] performs early fusion by first creating similarity graphs of different modalities to preserve essential item-item connections. Thereafter, a weighted-sum fusion is applied using an importance score before the use of graph convolution to obtain the final fused representation. As an extension, FREEDOM [50] reveals the redundancy in learning item-item similarity graphs and freezes the modality graphs during fusion for better representation learning. In terms of intermediate fusion, MMGCL [40] augments multi-modal graphs with edge dropout and masking, followed by concatenation and message propagation through a graph encoder. Instead of direct concatenation, DRAGON [48] adopts the late fusion paradigm and introduces attentive concatenation to discern user preferences across varying modalities. While these fusion approaches have achieved some success, noise embedded within modalities has not been effectively suppressed, resulting in noise amplification and degradation of model performance during cross-modality fusion. In this work, we propose to fuse and denoise modalities from the frequency spectrum perspective—an effective yet under-explored area in MRSs.

3 TASK FORMULATION

We model users' implicit interaction by first defining $\mathcal{U} = \{u_j \mid 1 \leq j \leq M\}$ and $\mathcal{I} = \{i_k \mid 1 \leq k \leq N\}$ as the user and item sets, where M and N denote the number of users and items, respectively. Henceforth, we can then define an interaction matrix $\mathbf{Y} \in \mathbb{R}^{M \times N}$, where element $y_{ui} = 1$ signifies an observed user-item interaction, while $y_{ui} = 0$ indicates otherwise. For each user u and item i, the input ID embedding matrix is represented as $E_{id} \in \mathbb{R}^{d \times (|U|+|I|)}$, where d is the embedding dimension. We further denote $e_i^m \in \mathbb{R}^{d_m}$ as the modality features of each item i, where d_m is the dimension of the modalities, $m \in \mathcal{M}$ is the modality, and $\mathcal{M}$ is the set of modalities. The primary focus of this work is on visual and textual modalities, where $\mathcal{M} = \{v, t\}$ such that v and t correspond, respectively, to visual and textual modality. While two modalities are described in this work, the proposed approach can easily be extended to multiple modalities. Finally, given the interaction data and the multi-modal features of each item, our goal is to predict user preferences accurately by estimating the likelihood $\widehat{y}_{ui}$ of interaction between a user u and an item i.

4 METHODOLOGY

The key components of SMORE encompass three core aspects: i) Spectrum Modality Fusion, ii) Multi-modal Graph Learning, and iii) Modality-aware Preference Module. Fig. 2 illustrates an overview of the proposed architecture.

4.1 Spectrum Modality Fusion

Modality fusion often confers advantages since the fused embeddings elucidate complementary and universal characteristics of different modalities. As opposed to existing works and drawing inspiration from the field of signal processing, SMORE exploits the frequency domain for dual purposes: modality fusion and denoising. With this objective, the raw multi-modal features $\mathbf{E}_{i,m}$ are first projected into a shared latent space using the multi-layer perceptron (MLP), i.e.,

$$\mathbf{H}_{i,m} = \mathbf{W}_{1,m}\mathbf{E}_{i,m} + \mathbf{b}_{1,m}, \tag{1}$$

where $\mathbf{W}_{1,m} \in \mathbb{R}^{d \times d_m}$ and $\mathbf{b}_{1,m} \in \mathbb{R}^d$ denote the projection matrix and bias vector of the MLP for each modality m, respectively. Thereafter, to convert the projected multi-modal features into the frequency domain for fusion and denoising, we utilize the discrete Fourier transform (DFT) [13] for each modality such that

$$\widetilde{\mathbf{H}}_{i,m} = \mathcal{F}_m\left(\mathbf{H}_{i,m}\right) \in \mathbb{C}^{n \times d}, \tag{2}$$

where

$$\mathcal{F}_m : \widetilde{h}_k = \sum_{j=0}^{n-1} h_j \exp\left(-\frac{2\pi \jmath}{n} jk\right), \quad 0 \le k \le n-1 \tag{3}$$

denotes the one-dimensional DFT function along the sequence and spatial dimensions of the textual and image modality, respectively, $\jmath = \sqrt{-1}$, and $\widetilde{h}_k$ denotes the modality spectrum features at frequency $2\pi k/n$, with k being the frequency-bin index. While the DFT has been widely applied for frequency conversion, quadratic complexity is incurred due to the computation of N components. Instead, we employ the fast Fourier transform (FFT), which decomposes the DFT matrix into a series of sparse matrix products [33], thereby reducing the complexity to a logarithmic scale.

With the spectral features of each modality, denoising is performed. Since dynamic noise may be present across different modalities and leveraging the discriminative spectrum generated from the FFT [18], we introduce a modality-specific dynamic filter

$$\widehat{\mathbf{H}}_{i,m} = \delta_m\left(\widetilde{\mathbf{H}}_{i,m}\right) = \mathbf{W}^c_{2,m} \odot \widetilde{\mathbf{H}}_{i,m}, \tag{4}$$

where δ_m is defined as the modal-specific transfer function of the filter, $\odot$ denotes the point-wise product, and $\mathbf{W}^c_{2,m} \in \mathbb{C}^{n\times d}$ denotes the trainable complex weight of the filter. The adaptive filter functions as a frequency selector that seeks to suppress noise-related irrelevant information. We can formulate the fusion process similarly to acquire the cross-modality fusion spectrum such that

$$\widehat{\mathbf{H}}_{i,f} = \delta_f\left(\prod_{m\in\mathcal{M}} \widetilde{\mathbf{H}}_{i,m}\right), \tag{5}$$

where, as opposed to matrix multiplication, $\mathbf{\Pi}$ is defined as the point-wise product operator, and δ_f is defined as the transfer function of the dynamic fusion filter. It is important to note that in the frequency domain, the point-wise product operation is equivalent to the circular convolution operation in the spatial domain. As a result, the rich correlations between the sequence and spatial modality (e.g., text and image) are captured, while minimizing noise contamination during fusion. More importantly, in contrast to advanced fusion methods (e.g., co-attention [35]), which require a quadratic time complexity, fusing in the frequency domain allows SMORE to achieve logarithmic runtime due to the efficient FFT and point-wise aggregation. In this aspect, we can achieve both efficiency and effectiveness in fusing modalities and denoising.

Thereafter, the spectrum of the uni-modal and fused modality features are projected back into the original feature space using the inverse discrete Fourier transform (IDFT)

$$\mathcal{F}_m^{-1} : h_j = \frac{1}{n}\sum_{k=0}^{n-1} \tilde{h}_k \exp\left(\frac{2\pi \jmath}{n} jk\right), \quad 0 \le j \le n-1, \tag{6}$$

$$\dot{\mathbf{H}}_{i,m} = \mathcal{F}_m^{-1}(\widehat{\mathbf{H}}_{i,m}) \in \mathbb{R}^{n\times d}, \quad \dot{\mathbf{H}}_{i,f} = \mathcal{F}_m^{-1}(\widehat{\mathbf{H}}_{i,f}) \in \mathbb{R}^{n\times d}. \tag{7}$$

As will be illustrated in Section 5.5, the above empowers SMORE to execute modality fusion and denoising effectively, extracting only essential uni-modal and fused features through filtering in the frequency domain.

4.2 Multi-modal Graph Learning

4.2.1 ***Item-Item Modal-Specific and Fusion Views.*** Having acquired the denoised and fused modality representations, the semantically correlated modality features can be distilled through graph convolutional operations. As highlighted in [42, 43, 50], the efficacy of a multi-modal recommender can be influenced substantially by both collaborative and semantically associated signals. These works construct individual graphs for each modality and aggregate them via learnable weights before performing message propagation on the fusion graph. By aggregating and disregarding uni-modality graph, distinct modality preferences are obscured.

In contrast, we emphasize the importance of capturing both uni-modal and fusion preferences by proposing a new multi-modal graph learning module that distinctively constructs modal-specific and fusion graphs. We first establish item-item affinities by computing the similarity of each raw modality features. Henceforth, we attain modality similarity matrix $\mathcal{S}_m$ such that the similarity between (item) row a and (item) column b entry of $\mathbf{E}_{i,m}$ is given by

$$s^m_{a,b} = \frac{\left(e^m_a\right)^{\mathrm{T}} e^m_b}{\|e^m_a\|\|e^m_b\|}. \tag{8}$$

To ensure that the uni-modal vital features are captured, we perform graph sparsification [2] by retaining K edges with the highest similarity scores such that

$$\dot{s}^m_{a,b} = \begin{cases} s^m_{a,b}, & s^m_{a,b} \in \text{top-}K_m(\{s^m_{a,c},\, c \in \mathcal{I}\}); \\ 0, & \text{otherwise}, \end{cases} \tag{9}$$

where $\dot{s}^m_{a,b}$ denotes the degree of similarity (edge weights) between items a and b in modality m. To mitigate the gradient vanishing/exploding problem [17], we then normalize the similarity matrix

$$\ddot{\mathcal{S}}_m = \mathbf{D}_m^{-1/2} \dot{\mathcal{S}}_m \mathbf{D}_m^{-1/2}, \tag{10}$$

where $\mathbf{D}_m^{-1/2}$ is defined as the degree matrix of $\dot{\mathcal{S}}_m$.

Having constructed the modality-specific graph, we adopt the max-pooling strategy to retain the highest complementary strength between different m modality graphs, thereby preserving prominent cross-modality features. The fusion affinity matrix can be defined as the max edge weights between items a and b, i.e.,

$$\ddot{\mathcal{S}}^f_{a,b} = \max_{m,m'\in\mathcal{M}}\left(\ddot{\mathcal{S}}^m_a, \ddot{\mathcal{S}}^{m'}_b\right), \quad m \ne m'. \tag{11}$$

Prior to unimodal and fusion feature propagation, we extract preference-related modality features based on behavioral guidance

$$\ddot{\mathbf{H}}_{i,m} = f^m_{gate}\left(\mathbf{E}_{i,id}, \dot{\mathbf{H}}_{i,m}\right) = \mathbf{E}_{i,id} \odot \sigma\left(\mathbf{W}_{3,m}\dot{\mathbf{H}}_{i,m} + \mathbf{b}_{3,m}\right), \tag{12}$$

$$\ddot{\mathbf{H}}_{i,f} = f^c_{gate}\left(\mathbf{E}_{i,id}, \dot{\mathbf{H}}_{i,f}\right) = \mathbf{E}_{i,id} \odot \sigma\left(\mathbf{W}_{4,f}\dot{\mathbf{H}}_{i,f} + \mathbf{b}_{4,f}\right), \tag{13}$$

where $\mathbf{W}_{(.)} \in \mathbb{R}^{d\times d}$ and $\mathbf{b}_{(.)} \in \mathbb{R}^d$ are the trainable parameters and σ is the non-linearity sigmoid gate function.

Inspired by the simplicity and efficacy of LightGCN [12], the item uni-modal $\ddot{\mathbf{H}}_{i,m}$ and fusion features $\ddot{\mathbf{H}}_{i,f}$ are propagated through a shallow light graph convolutional layer with the propagation rule being

$$\bar{\mathbf{H}}_{i,m} = \ddot{\mathcal{S}}_m \ddot{\mathbf{H}}_{i,m}, \qquad \bar{\mathbf{H}}_{i,f} = \ddot{\mathcal{S}}_f \ddot{\mathbf{H}}_{i,f}. \tag{14}$$

Likewise, we can compute the user modality features through a weighted-sum aggregation layer defined by

$$\bar{h}_{u,m} = \sum_{i\in\mathcal{N}_u} \frac{1}{\sqrt{|\mathcal{N}_u||\mathcal{N}_i|}} \bar{h}_{i,m}, \qquad \bar{h}_{u,f} = \sum_{i\in\mathcal{N}_u} \frac{1}{\sqrt{|\mathcal{N}_u||\mathcal{N}_i|}} \bar{h}_{i,m}, \tag{15}$$

where $\bar{h}_{u,m}$ and $\bar{h}_{u,f}$ are the user uni and fusion modality features, respectively, and $1/\sqrt{|\mathcal{N}_u||\mathcal{N}_i|}$ is the symmetric normalization term to avoid overscaling. It is useful to note that employing a shallow layer in the item-item view is sufficient for capturing relevant semantic associative signals since stacking multiple layers may induce undesirable high-order latent noise [41]. By concatenating $\bar{\mathbf{H}}_{u,m}$ with $\bar{\mathbf{H}}_{i,m}$, and $\bar{\mathbf{H}}_{u,f}$ with $\bar{\mathbf{H}}_{i,f}$, we can obtain the enriched

uni-modal and fusion features for both the users and items, denoted as $\overline{\mathbf{H}}_m$ and $\overline{\mathbf{H}}_f \in \mathbb{R}^{d\times(|\mathcal{U}|+|\mathcal{I}|)}$, respectively.

4.2.2 ***User-Item Behavioral View.*** The focus of this view is on encoding the high-order collaborative signals from users' historical interactions. It has been verified that the collaborative signals are highly influential in delineating users' behavioral patterns [9, 12, 21, 51]. On this basis, we recursively propagate long-range collaborative signals in the interaction graph resulting in the behavioral embedding of the users and items given by

$$\mathbf{E}_{id}^{(l)} = (\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2})\,\mathbf{E}_{id}^{(l-1)}, \quad \mathbf{A} = \begin{bmatrix} 0 & \mathbf{Y} \\ \mathbf{Y}^\top & 0 \end{bmatrix}. \tag{16}$$

Here, $\mathbf{E}_{id}^{(l-1)}$ denotes the ID embeddings at the previous layer, and $\mathbf{D}^{-1/2}$ is the diagonal degree matrix corresponding to the adjacency matrix $\mathbf{A}$. To obtain the overall high-order behavioral features of users and items, we aggregate the hidden layers by applying the mean function giving

$$\bar{\mathbf{E}}_{id} = \frac{1}{L+1}\sum_{i=0}^{L}\mathbf{E}_{id}^{(l)}, \qquad \bar{\mathbf{E}}_{id} \in \mathbb{R}^{d\times(|\mathcal{U}|+|\mathcal{I}|)}. \tag{17}$$

4.3 Modality-Aware Preference Module

With the high-order behavioral and enriched (uni-modal and fused) features $\bar{\mathbf{E}}_{id}$, $\overline{\mathbf{H}}_m$ and $\overline{\mathbf{H}}_f$, modality preferences are distilled for each user. In line with the diversity observed in real-world scenarios, a user may be inclined toward a single modality, while some may exhibit a mixture of fusion preferences. As such, we utilize complementary signals encapsulated in the fusion embeddings to weigh the uni-modal features, effectively striking a balance between the uni-modal and fusion preferences. To this end, we define

$$\alpha_m = \text{softmax}(\mathbf{p}_m^\top \tanh(\mathbf{W}_{5,m}\overline{\mathbf{H}}_f + \mathbf{b}_{5,m})) \tag{18}$$

as the modal-specific attention weights, where $p_{(\cdot)} \in \mathbb{R}^d$ is the attention vector. These weights are subsequently used to weigh the semantically associated uni-modal features to obtain the final aggregated uni-modal features given by

$$\mathbf{H}_m^* = \sum_{m\in M}\alpha_m\overline{\mathbf{H}}_m. \tag{19}$$

We next extract modality preferences derived from user collaborative information by feeding the high-order behavioral signal through a uni-modal and fusion gate function. This results in explicit uni-modal and fusion preferences given, respectively, by

$$\begin{aligned} \mathbf{Q}_m &= \psi_{pref}^m\left(\bar{\mathbf{E}}_{id}\right) = \sigma(\mathbf{W}_{6,m}\bar{\mathbf{E}}_{id} + \mathbf{b}_{6,m}), \\ \mathbf{Q}_f &= \psi_{pref}^f\left(\bar{\mathbf{E}}_{id}\right) = \sigma(\mathbf{W}_{7,f}\bar{\mathbf{E}}_{id} + \mathbf{b}_{7,f}), \end{aligned} \tag{20}$$

where σ is the non-linearity function. The overall multi-modal side features (distilled from the explicit uni-modal and fusion preferences) are then derived as

$$\mathbf{H}_s = \frac{1}{|M|}\left(\sum_{m\in M}\mathbf{H}_m^* \odot \mathbf{Q}_m\right) + \left(\mathbf{H}_f \odot \mathbf{Q}_f\right). \tag{21}$$

To maximize mutual information across high-order behavioral and modality-side information, we then incorporate an InfoNCE contrastive task [24] with

$$\mathcal{L}_{cl}^u = \sum_{u\in\mathcal{U}} -\log\frac{\exp(\bar{e}_{u,id}\cdot\bar{h}_{u,s}/\tau)}{\sum_{v\in\mathcal{U}}\exp(\bar{e}_{v,id}\cdot\bar{h}_{v,s}/\tau)} \tag{22}$$

being the user contrastive loss and τ being the hyperparameter temperature that regulates the degree of smoothness in the distribution.

Table 1: Dataset Statistics

Dataset	#User	#Item	#Interaction	Density
Baby	19,445	7,050	160,792	0.117%
Sports	35,598	18,357	296,337	0.045%
Clothing	39,387	23,033	278,677	0.031%

This task ensures the preservation of essential features distilled from behavioral and modality views. Similarly, we can obtain the item contrastive loss $\mathcal{L}_{cl}^i$ by substituting users with items as defined in Eq (22). Thereafter, the overall contrastive loss is governed by $\mathcal{L}_{cl} = \mathcal{L}_{cl}^u + \mathcal{L}_{cl}^i$.

4.4 Prediction and Optimization

By capitalizing on the refined uni-modal and complementary fused features, we acquire the final representations of the user and item

$$e_u^* = \bar{e}_{u,id} + h_{u,s}, \quad e_i^* = \bar{e}_{i,id} + h_{i,s}. \tag{23}$$

The estimated likelihood is then computed as $\widehat{y}(u,i) = e_u^{*\top}e_i$. For model optimization, we employ the BPR loss [28] to reconstruct the historical data, which prioritizes higher scores for observed items, i.e.,

$$\mathcal{L}_{bpr} = \sum_{(u,i,j)\in O} -\ln\sigma\left(\widehat{y}_{ui} - \widehat{y}_{uj}\right). \tag{24}$$

Here, $O = \{(u,i,j)|(u,i)\in O^+, (u,j)\in O^-\}$ represents the set of interactions, comprising observed O^+ and unobserved O^- interactions, and σ denotes the sigmoid function. We then perform joint optimization in conjunction with the contrastive loss such that the overall loss function is given by

$$\mathcal{L} = \mathcal{L}_{bpr} + \lambda_1\mathcal{L}_{cl} + \lambda_2||\Theta||_2^2, \tag{25}$$

where λ_1 and λ_2 regulates the influence of the contrastive task and the L2 regularization term, respectively.

5 EXPERIMENTS

We conducted an extensive set of experiments designed to address the following research questions:

- **RQ1:** How effective is the proposed SMORE architecture compared with state-of-the-art general and multi-modal models?
- **RQ2:** How do the key components and different modalities within SMORE contribute to its overall performance?
- **RQ3:** How do hyperparameter perturbations impact the overall efficacy of the proposed model?
- **RQ4:** Does spectrum-based fusion truly enhance denoising capability and capture valuable content?

5.1 Experiment Configurations

5.1.1 Datasets. In accordance with preceding works [42, 52], we perform experiments using three categories of the real-world Amazon Review datasets[1], presented by McAuley et al. [11]: (i) *Baby*, (ii) *Sports and Outdoors*, and (iii) *Clothing, Shoes and Jewelry*. For ease of reference, we label them as *Baby*, *Sports*, and *Clothing*, respectively. Offering both visual and textual insights into items, the

[1]Datasets are publicly available at http://jmcauley.ucsd.edu/data/amazon/links.html.

Table 2: Performance comparison of different recommendation models. To ascertain the stability of the results, experiments were conducted across 5 different seeds, and the improvements are statistically significant with $p < 0.01$ in a paired t-test setting.

Datasets	Metrics	General Recommenders		Multi-modal Recommenders							
		BPR	LightGCN	VBPR	MMGCN	GRCN	SLMRec	BM3	MGCN	FREEDOM	SMORE
Baby	Recall@10	0.0382	0.0453	0.0425	0.0424	0.0534	0.0545	0.0548	0.0616	0.0626	**0.0680***
	Recall@20	0.0595	0.0728	0.0663	0.0668	0.0831	0.0837	0.0876	0.0943	0.0986	**0.1035***
	NDCG@10	0.0207	0.0246	0.0223	0.0223	0.0288	0.0296	0.0297	0.0330	0.0327	**0.0365***
	NDCG@20	0.0263	0.0317	0.0284	0.0286	0.0365	0.0371	0.0381	0.0414	0.0420	**0.0457***
Sports	Recall@10	0.0417	0.0542	0.0561	0.0386	0.0607	0.0676	0.0613	0.0736	0.0724	**0.0762***
	Recall@20	0.0633	0.0837	0.0857	0.0627	0.0922	0.1017	0.0940	0.1105	0.1089	**0.1142***
	NDCG@10	0.0232	0.0300	0.0307	0.0204	0.0325	0.0374	0.0339	0.0403	0.0390	**0.0408***
	NDCG@20	0.0288	0.0376	0.0384	0.0266	0.0406	0.0462	0.0424	0.0498	0.0484	**0.0506***
Clothing	Recall@10	0.0200	0.0338	0.0281	0.0224	0.0428	0.0461	0.0418	0.0649	0.0635	**0.0659***
	Recall@20	0.0295	0.0517	0.0410	0.0362	0.0663	0.0696	0.0636	0.0971	0.0938	**0.0987***
	NDCG@10	0.0111	0.0185	0.0157	0.0118	0.0227	0.0249	0.0225	0.0356	0.0340	**0.0360***
	NDCG@20	0.0135	0.0230	0.0190	0.0153	0.0287	0.0308	0.0281	0.0438	0.0417	**0.0443***

Amazon dataset exhibits variability in the number of items per category. For pre-processing, we filter the raw data from each dataset using the 5-core setting on both users and items. The data has been summarized in Table 1. For the visual modality, we adopted the 4,096-dimensional features acquired from VGG16 [30]. For the textual modality, we employed sentence-transformers [27] to obtain a 384-dimensional text embedding from the concatenated brand, title, description, and category of each item.

5.1.2 Baselines. To verify the efficacy of SMORE, we benchmark against several state-of-the-art (STOA) recommender models. These baselines fall into two main categories: General recommenders, which focus solely on user-item interaction data to provide recommendations, and multi-modal recommenders that leverage both historical data and the multi-modal features of each item.

i) General Recommenders: The following STOA models that include STOA matrix factorization (MF) model (BPR [28]) and a graph-based model (LightGCN [12]) are chosen for comparison.

ii) Multi-modal Recommenders: To ensure robust evaluation of the proposed model, several STOA MRSs have been selected for comparison, including the MF model (VBPR[11]) and graph-based models (MMGCN [37], GRCN [36], SLMRec [31], BM3 [52], MGCN [41], FREEDOM [50]).

5.1.3 Evaluation Standards. To ensure consistency, we adhere to existing works [42, 52] and divide the interaction data into 80% for training, 10% for validating, and 10% for testing. Furthermore, we adopted the all-ranking protocol to evaluate top-K recommendation performance, using two widely-used metrics: Recall@K and NDCG@K. The results were reported for all users in the test set.

5.1.4 Implementation Details. We utilized the unified open-source MMRec framework [49] for developing the proposed model and replicating existing recommenders. For each of the selected baselines, the hyperparameters were tuned in line with the optimal configurations reported in the respective published papers. To further ensure impartiality, we complied with existing works [50, 52] and deployed the same seed across all baseline implementations and fixed the dimension of both the users and items at 64. We initialized all training parameters using the Xavier [8] technique and adopted the Adam optimizer [16]. The training process employed a fixed batch size of 2,048 and was conducted over 1,000 epochs. Early stopping was activated after 20 consecutive steps without improvement on the validation set, with Recall@20 being the indicator metric.

5.2 Effectiveness of SMORE (RQ1)

With reference to Table 2, comparisons with highly-competitive general and multi-modal recommenders reveal that:

The proposed model consistently outperforms all baselines, including general and multi-modal recommenders. We posit the improvement arises from SMORE's ability to capitalize on multi-modalities for inferring accurate uni-modal and fusion preferences. By leveraging the discriminative spectral characteristics and global perspective inherent in the frequency domain, universal patterns across different modalities are proficiently captured, while poor performance due to cross-modality noise is mitigated by the dynamic filter through effective attenuation and suppression. Furthermore, the multi-modal graph learning module empowers SMORE to encode high-order collaborative and semantically associated, preference-related modality features. To model the users' diverse modality preferences reliably, SMORE exploits the enriched universal fused signals to regulate the uni-modal features, ensuring that the uni-modal and fusion preferences are optimally balanced and accurately aligned with real-world scenarios.

Graph-based recommenders that fuse modalities directly are evidently less effective. In some instances, general models such as LightGCN achieve higher performance than MRSs such as MMGCN, which utilizes direct summation for fusion. This result suggests the existence of noise due to multi-modalities (as illustrated in Fig. 1) and that direct fusion can adversely impact the performance of multi-modal recommenders.

To a certain extent, indirect integration of modality features may help to reduce modality noise. Unlike VBPR, which

Table 3: Performance Comparison on multi-modalities

Datasets	Modality	R@10	R@20	N@10	N@20
Baby	Text	0.0646	0.0996	0.0341	0.0431
	Visual	0.0533	0.0854	0.0290	0.0373
	Fusion	0.0625	0.0964	0.0331	0.0418
	Full	**0.0680**	**0.1035**	**0.0365**	**0.0457**
Sports	Text	0.0727	0.1099	0.0392	0.0488
	Visual	0.0592	0.0903	0.0323	0.0404
	Fusion	0.0729	0.1091	0.0392	0.0486
	Full	**0.0762**	**0.1142**	**0.0408**	**0.0506**
Clothing	Text	0.0631	0.0945	0.0343	0.0422
	Visual	0.0443	0.0661	0.0241	0.0296
	Fusion	0.0621	0.0937	0.0342	0.0422
	Full	**0.0659**	**0.0987**	**0.0360**	**0.0443**

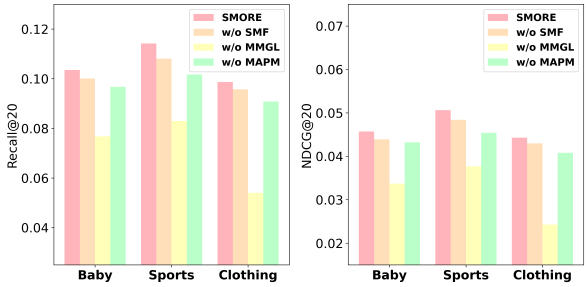

Figure 3: Ablation studies on the proposed SMORE

directly injects modality features into ID representations, GRCN exhibits moderate performance improvement by relying on modality features implicitly to enhance the interaction graph. From an alternate perspective, the two MGCN and FREEDOM baselines exhibit enhanced performance by examining the latent structures of items through various modalities. We hypothesize that such a notable performance is attributed to MGCN's uni-modal noise reduction through behavioral injection and FREEDOM's degree-aware edge pruning strategy for denoising the interaction graph. Nonetheless, these models share a common vulnerability—they do not explore the complementary and universal features effectively across different modalities. In contrast, SMORE captures both uni-modal and fusion signals explicitly and can discern users' varying degrees of preferences accurately, resulting in its superior performance.

5.3 Ablation Studies (RQ2)

To ascertain the effectiveness of each component, we segment the proposed model into four distinct variants: (i) SMORE, (ii) SMORE without spectrum modality fusion, (iii) SMORE without multi-modal graph learning, and (iv) SMORE without modality aware preference module. Results presented in Fig. 3 highlight that:

Removing any key components leads to performance deterioration. It is evident that every key component plays a significant role in SMORE, collectively contributing to its superior performance. For instance, spectrum modality fusion is responsible for fusing

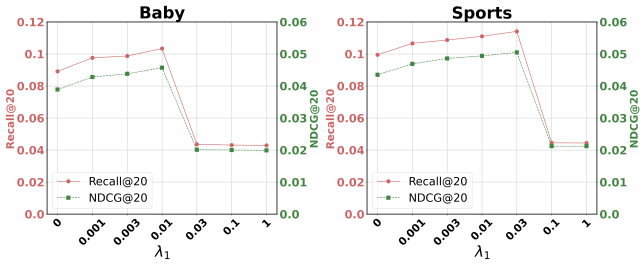

Figure 4: Variation of SMORE with λ_1

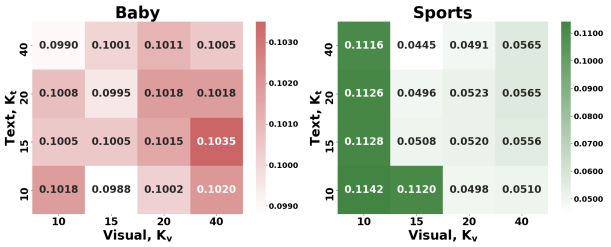

Figure 5: Variation of SMORE with K_m

and mitigating cross-modality noise through attenuation and suppression, while multi-modal graph learning and modality-aware preference modules aim to encode semantically associative universal signals and decipher accurate user preference, respectively. Hence, omitting any of these modules degrades the performance of the proposed model.

MMGL plays a quintessential role in enhancing the overall performance. We observe that, out of the four variants, the absence of MMGL results in a significant decline in performance across all datasets. This trend underscores the importance of encoding both high-order collaborative signals and the universal associative features derived from the denoised uni-modal and fused features, which collectively bolsters the overall performance of SMORE.

To further assess the impact of each modality, we perform experiments under a variety of input conditions: *text* comprising textual (sequential) information, *visual* consisting of pictorial (spatial) information, *fusion* including only the fused (complementary) information, and *full* encompassing both uni-modal and fusion information. Results tabulated in Table 3 indicate the following:

The omission of any modalities results in reduced performance. Excluding any modality reduces SMORE's capability to decipher users' diverse modality preferences. This observation provides unequivocal evidence that SMORE can leverage the uni-modal and complementary features encapsulated in the given modalities effectively while mitigating issues associated with noise contamination—a problem inherent in existing models [42, 43]. This also reliably substantiates the necessity of modeling both uni-modal and fusion preferences, which serve as a realistic representation of real-world scenarios.

Among the first three variants, SMORE demonstrates notable performance in utilizing sequential text information. It has been shown in existing studies that the use of text usually leads to a significant degradation in terms of performance due to irrelevant information [41]. On the contrary, we observe from Table 2

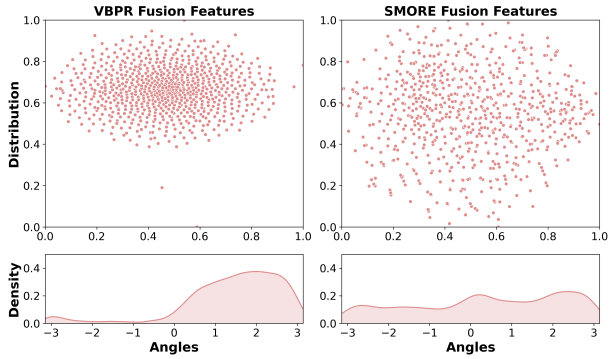

Figure 6: Distribution of fusion features for Baby Dataset

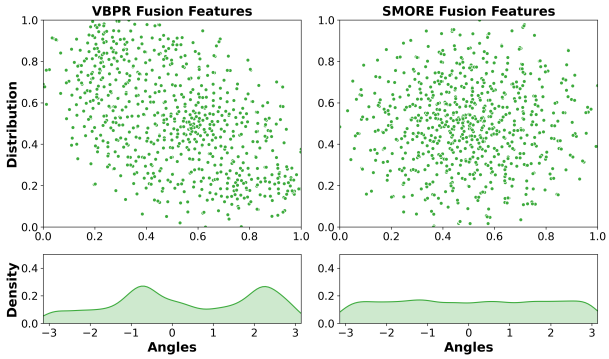

Figure 7: Distribution of fusion features for Sports Dataset

that utilizing text solely exhibits higher performance than STOA models (e.g., MGCN and FREEDOM), which utilize both modalities. This observation highlights the denoising capability of SMORE in capturing essential uni-modal features.

5.4 Selection of Key Hyperparameters (RQ3)

With reference to Fig. 4, we investigate the primary hyperparameter λ_1 of SMORE, which governs the influence of contrastive task by varying $0 \leq \lambda_1 \leq 1$. Results reveal that the omission of contrastive loss ($\lambda_1 = 0$) degrades the performance of the proposed model to a considerable degree, implying the beneficial impact of incorporating an auxiliary task for self-supervision and representation alignment. On the other hand, setting an excessively high value of $\lambda_1 = 1$ results in the most significant degradation due to the model placing undue emphasis on the auxiliary task. Notably, the best performance is achieved at $\lambda_1 = 0.01$ and 0.03 for the baby and sports datasets, respectively. This observation implies that a small value is sufficient to enhance the recommendation task.

We next assess the impact of K_m defined in (9) on SMORE by varying $10 \leq K_m \leq 40$ for each modality. With reference to Fig. 5, findings from the sports dataset indicate that a small value ($K_m = 10$) for both visual and text is sufficient to capture relevant uni-modal and fusion associative signals. However, as $K_v \geq 20$, performance degradation persists, while variations in K_t generally do not compromise the performance. On the contrary, for the Baby dataset, we noted a minor variation in trend—while setting low value ($K_t = 15$) for text coheres with the prior finding, we observe that the optimal attained performance occurs when K_v is set to a high value ($K_v = 40$). This suggests that a low value of K may not always be effective across different datasets, and setting it too low may inadvertently discard essential neighbors.

5.5 Impact of fusion in frequency domain (RQ4)

To verify the quality of the fusion features captured by the proposed model, we visualize the fusion features by first performing dimensionality reduction using t-SNE [32] to map the high-dimensional fusion features into 2-D, as illustrated in Figs. 6 and 7. The embedding distributions are then plotted using Gaussian kernel density estimation, with the unit hypersphere S^1 showing the estimation of $\arctan(y, x)$ depicted at the bottom of each figure. For comparison, we used the classical VBPR fusion model as a benchmark.

As depicted for the Baby dataset, we observe that the fusion features learned by SMORE result in a uniformly distributed structure implying that each item is characterized by its own distinct fusion semantic associations. On the other hand, we note that the distribution of VBPR is significantly condensed, highlighting the representation degeneration problem [7, 25]. This phenomenon is linked to the cross-modality noise amplification illustrated in Fig. 1, which severely corrupts and limits the expressivity of the representation after fusion. Similarly, Fig. 7 exemplifies comparable patterns that mirror previous observations. While VBPR displays a lower level of degeneration issue, the proposed model attains greater uniformity, with distributions that are evenly spread. These results clearly substantiate the capability of SMORE in fusing different modalities and mitigating the cross-modality noise in the frequency domain, validating the efficacy of our fusion approach.

6 CONCLUSION

In this work, we aim to reduce modality noise by harnessing the discriminative spectrum property and global perspective inherent in the frequency domain for modality fusion and uni-modal denoising. Specifically, the proposed multi-modal SMORE recommender effectively captures both uni-modal and fusion preferences while actively suppressing modality noise. By leveraging the discriminative modality spectrum property, we proposed an effective approach to attenuate and suppress cross-modality noise during fusion. To explore the item latent structures, we introduced a new multi-modal graph learning module to distill long-range collaborative and semantic associated universal patterns among similar items. Finally, mirroring real-world scenarios, where users often display a mixture of multi-modal preferences, we designed a modality-aware preference module that effectively balances the uni-modal and fused representations, enabling a precise capture of users' uni-modal and fusion preferences.

REFERENCES

[1] Saeed Anwar and Nick Barnes. 2019. Real image denoising with feature attention. In *Proceedings of the IEEE/CVF Int'l Conference on Computer Vision*. 3155–3164.
[2] Jie Chen, Haw-ren Fang, and Yousef Saad. 2009. Fast Approximate kNN Graph Construction for High Dimensional Data via Recursive Lanczos Bisection. *Journal of Machine Learning Research* 10, 9 (2009).
[3] Xu Chen, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2019. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In *Proceedings of the 42nd Int'l ACM SIGIR Conference on*

Research and Development in Information Retrieval. 765–774.
[4] Yang Chen, Yueqi Duan, Runzhong Zhang, and Yap-Peng Tan. 2024. Adaptive Margin Contrastive Learning for Ambiguity-aware 3D Semantic Segmentation. In *2024 IEEE Int'l Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.
[5] Hong Wei Chun, Rongqing Kenneth Ong, and Andy W. H. Khong. 2024. Reasonable Sense of Direction: Making Course Recommendations Understandable with LLMs. In *2024 IEEE 67th International Midwest Symposium on Circuits and Systems (MWSCAS)*. IEEE, 1408–1412.
[6] Enyan Dai, Wei Jin, Hui Liu, and Suhang Wang. 2022. Towards robust graph neural networks for noisy graphs with sparse labels. In *Proceedings of the Fifteenth ACM Int'l Conference on Web Search and Data Mining*. 181–191.
[7] Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Representation degeneration problem in training natural language generation models. *arXiv preprint arXiv:1907.12009* (2019).
[8] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth Int'l Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, 249–256.
[9] Li He, Xianzhi Wang, Dingxian Wang, Haoyuan Zou, Hongzhi Yin, and Guandong Xu. 2023. Simplifying graph-based collaborative filtering for recommendation. In *Proceedings of the sixteenth ACM Int'l Conference on Web Search and Data Mining*. 60–68.
[10] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th Int'l Conference on World Wide Web*. 507–517.
[11] Ruining He and Julian McAuley. 2016. VBPR: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.
[12] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd Int'l ACM SIGIR Conference on Research and Development in Information Retrieval*. 639–648.
[13] Michael Heideman, Don Johnson, and Charles Burrus. 1984. Gauss and the history of the fast Fourier transform. *IEEE ASSP Magazine* 1, 4 (1984), 14–21.
[14] Saeed Izadi, Darren Sutton, and Ghassan Hamarneh. 2023. Image denoising in the deep learning era. *Artificial Intelligence Review* 56, 7 (2023), 5929–5974.
[15] Yungi Kim, Taeri Kim, Won-Yong Shin, and Sang-Wook Kim. 2024. MONET: Modality-Embracing Graph Convolutional Network and Target-Aware Attention for Multimedia Recommendation. In *Proceedings of the 17th ACM Int'l Conference on Web Search and Data Mining*. 332–340.
[16] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[17] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
[18] An Lao, Qi Zhang, Chongyang Shi, Longbing Cao, Kun Yi, Liang Hu, and Duoqian Miao. 2024. Frequency spectrum is more effective for multimodal representation and fusion: A multimodal spectrum rumor detector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 18426–18434.
[19] Qiang Liu, Shu Wu, and Liang Wang. 2017. Deepstyle: Learning user preferences for visual recommendation. In *Proceedings of the 40th Int'l ACM SIGIR Conference on Research and Development in Information Retrieval*. 841–844.
[20] Yunshan Ma, Xiaohao Liu, Yinwei Wei, Zhulin Tao, Xiang Wang, and Tat-Seng Chua. 2024. Leveraging multimodal features and item-level user feedback for bundle construction. In *Proceedings of the 17th ACM Int'l Conference on Web Search and Data Mining*. 510–519.
[21] Kelong Mao, Jieming Zhu, Xi Xiao, Biao Lu, Zhaowei Wang, and Xiuqiang He. 2021. UltraGCN: ultra simplification of graph convolutional networks for recommendation. In *Proceedings of the 30th ACM Int'l Conference on Information & Knowledge Management*. 1253–1262.
[22] Michael Michelashvili and Lior Wolf. 2019. Speech denoising by accumulating per-frequency modeling fluctuations. *arXiv preprint arXiv:1904.07612* (2019).
[23] Rongqing Kenneth Ong, Wei Qiu, and Andy W. H. Khong. 2023. Quad-Tier Entity Fusion Contrastive Representation Learning for Knowledge Aware Recommendation System. In *Proceedings of the 32nd ACM Int'l Conference on Information and Knowledge Management*. 1949–1959.
[24] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).
[25] Ruihong Qiu, Zi Huang, Hongzhi Yin, and Zijian Wang. 2022. Contrastive learning for representation degeneration problem in sequential recommendation. In *Proceedings of the fifteenth ACM Int'l Conference on Web Search and Data Mining*. 813–823.
[26] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. 2021. Global filter networks for image classification. *Advances in Neural Information Processing Systems* 34 (2021), 980–993.
[27] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).
[28] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618* (2012).
[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th Int'l conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 234–241.
[30] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
[31] Zhulin Tao, Xiaohao Liu, Yewei Xia, Xiang Wang, Lifang Yang, Xianglin Huang, and Tat-Seng Chua. 2022. Self-supervised learning for multimedia recommendation. *IEEE Transactions on Multimedia* (2022).
[32] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 11 (2008).
[33] Charles Van Loan. 1992. *Computational frameworks for the fast Fourier transform*. SIAM.
[34] Qifan Wang, Yinwei Wei, Jianhua Yin, Jianlong Wu, Xuemeng Song, and Liqiang Nie. 2021. Dualgnn: Dual graph neural network for multimedia recommendation. *IEEE Transactions on Multimedia* 25 (2021), 1074–1084.
[35] Wei Wei, Chao Huang, Lianghao Xia, and Chuxu Zhang. 2023. Multi-modal self-supervised learning for recommendation. In *Proceedings of the ACM Web Conference 2023*. 790–800.
[36] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2020. Graph-refined convolutional network for multimedia recommendation with implicit feedback. In *Proceedings of the 28th ACM Int'l Conference on Multimedia*. 3541–3549.
[37] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM Int'l Conference on Multimedia*. 1437–1445.
[38] Lanling Xu, Zhen Tian, Bingqian Li, Junjie Zhang, Daoyuan Wang, Hongyu Wang, Jinpeng Wang, Sheng Chen, and Wayne Xin Zhao. 2024. Sequence-level Semantic Representation Fusion for Recommender Systems. In *Proceedings of the 33rd ACM Int'l Conference on Information and Knowledge Management*. 5015–5022.
[39] Shuo Yang, Zhaopan Xu, Kai Wang, Yang You, Hongxun Yao, Tongliang Liu, and Min Xu. 2023. Bicro: Noisy correspondence rectification for multi-modality data via bi-directional cross-modal similarity consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19883–19892.
[40] Zixuan Yi, Xi Wang, Iadh Ounis, and Craig Macdonald. 2022. Multi-modal graph contrastive learning for micro-video recommendation. In *Proceedings of the 45th Int'l ACM SIGIR Conference on Research and Development in Information Retrieval*. 1807–1811.
[41] Penghang Yu, Zhiyi Tan, Guanming Lu, and Bing-Kun Bao. 2023. Multi-view graph convolutional network for multimedia recommendation. In *Proceedings of the 31st ACM Int'l Conference on Multimedia*. 6576–6585.
[42] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. 2021. Mining latent structures for multimedia recommendation. In *Proceedings of the 29th ACM Int'l Conference on Multimedia*. 3872–3880.
[43] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Mengqi Zhang, Shu Wu, and Liang Wang. 2022. Latent structure mining with contrastive modality fusion for multimedia recommendation. *IEEE Transactions on Knowledge and Data Engineering* (2022).
[44] Qingyang Zhang, Yake Wei, Zongbo Han, Huazhu Fu, Xi Peng, Cheng Deng, Qinghua Hu, Cai Xu, Jie Wen, Di Hu, et al. 2024. Multimodal fusion on low-quality data: A comprehensive survey. *arXiv preprint arXiv:2404.18947* (2024).
[45] Kai Zhao, Yukun Zheng, Tao Zhuang, Xiang Li, and Xiaoyi Zeng. 2022. Joint learning of e-commerce search and recommendation with a unified graph neural network. In *Proceedings of the Fifteenth ACM Int'l Conference on Web Search and Data Mining*. 1461–1469.
[46] Yilun Zheng, Sitao Luan, and Lihui Chen. 2024. What is missing in homophily? disentangling graph homophily for graph neural networks. *arXiv preprint arXiv:2406.18854* (2024).
[47] Hongyu Zhou, Xin Zhou, Zhiwei Zeng, Lingzi Zhang, and Zhiqi Shen. 2023. A comprehensive survey on multimodal recommender systems: Taxonomy, evaluation, and future directions. *arXiv preprint arXiv:2302.04473* (2023).
[48] Hongyu Zhou, Xin Zhou, Lingzi Zhang, and Zhiqi Shen. 2023. Enhancing dyadic relations with homogeneous graphs for multimodal recommendation. *arXiv preprint arXiv:2301.12097* (2023).
[49] Xin Zhou. 2023. Mmrec: Simplifying multimodal recommendation. In *Proceedings of the 5th ACM Int'l Conference on Multimedia in Asia Workshops*. 1–2.
[50] Xin Zhou and Zhiqi Shen. 2023. A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation. In *Proceedings of the 31st ACM Int'l Conference on Multimedia*. 935–943.
[51] Xin Zhou, Aixin Sun, Yong Liu, Jie Zhang, and Chunyan Miao. 2023. Selfcf: A simple framework for self-supervised collaborative filtering. *ACM Transactions on Recommender Systems* 1, 2 (2023), 1–25.
[52] Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. 2023. Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM Web Conference 2023*. 845–854.