MSI: Multi-modal Recommendation via Superfluous Semantics Discarding and Interaction Preserving

Yi Li*
liyitunan@gmail.com
University of Chinese Academy of Sciences
Institute of Software Chinese Academy of Sciences
Beijing, USA

Qingmeng Zhu*
qingmeng@iscas.ac.cn
University of Chinese Academy of Sciences
Institute of Software Chinese Academy of Sciences
Beijing, USA

Changwen Zheng
changwen@iscas.ac.cn
Institute of Software Chinese Academy of Sciences
Beijing, China

Jiangmeng Li†
jiangmeng2019@iscas.ac.cn
Institute of Software Chinese Academy of Sciences
Beijing, China

ABSTRACT

Multi-modal recommendation aims at leveraging data of auxiliary modalities (e.g., linguistic descriptions and images) to enhance the representations of items, thereby accurately recommending items that users prefer from the vast expanse of Web-based data. Current multi-modal recommendation methods typically utilize multi-modal features to assist in learning item representations in a direct manner. However, the superfluous semantics in multi-modal features are ignored, resulting in the inclusion of excessive redundancy within the representations of items. Moreover, we disclose that multi-modal features of items rarely contain user-item interaction information. Hence, during the interaction among different item features, the user-item interaction information in ID-based representations diminishes, leading to the degeneration of recommendation performance. To this end, we propose a novel multi-modal recommendation approach, which compresses representations of extra modalities under the guidance of solid theoretical analysis and leverages two auxiliary multi-modal graphs to integrate user-item interaction information into multi-modal features. Empirical experiments on three multi-modal recommendation datasets demonstrate that our method outperforms benchmarks.

CCS CONCEPTS

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

KEYWORDS

Multi-modal learning; Recommendation; Superfluous semantics discarding; Interaction preserving

*Both authors contributed equally to this research.
†Corresponding author.

ICMR '24, June 10–14, 2024, Phuket, Thailand

ACM ISBN 979-8-4007-0619-6/24/06.
https://doi.org/10.1145/3652583.3658043

ACM Reference Format:
Yi Li, Qingmeng Zhu, Changwen Zheng, and Jiangmeng Li. 2024. MSI: Multi-modal Recommendation via Superfluous Semantics Discarding and Interaction Preserving. In *Proceedings of the 2024 International Conference on Multimedia Retrieval (ICMR '24), June 10–14, 2024, Phuket, Thailand.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3652583.3658043

1 INTRODUCTION

The fast-growing electronic commerce brings about millions of products and services. In this context, recommender systems play a crucial role in finding users' preferred offerings. Deep learning methods, such as GCN (Graph Convolution Network [18]), have been used for recommendation broadly, because of their ability to learn high-quality representations of users and items from historical interactions. However, the underutilization of excessive multi-modal content information (such as linguistic descriptions and images) of items is a long-standing challenge for multi-modal recommender systems.

With the development of multi-modal representation learning [24, 26], multi-modal recommender systems have emerged to integrate multi-modal information into conventional recommendation paradigm. As an early study, [6, 27] leverage attention mechanisms to explore relations between users' preferences and items' multi-modal features, but user-item interaction is not fully exploited. To further explore high-order connections in the user-item historical interactions, GCNs are adopted to incorporate multi-modal information into message passing process and enhance the representations of users and items. For instance, based on GCN, [40, 41] exploit the multi-modal features to enhance the representation of users and items. To boost recommendation performances further, auxiliary graphs (e.g., item-item and user-user) are adopted by [37, 43]. The mentioned GNN-based multi-modal methods [37, 40, 41, 43] make significant progress in recommender systems and adopt Bayesian Personalized Ranking (BPR) [34] loss to guide the optimization of models. But the negative samples for calculating BPR can bring wrong supervision signals into the training process. To address the issue, inspired by BYOL [11] and SimSiam [7], BM3 [45] proposes a self-supervised framework that does not require negative samples and achieves the state-of-the-art (SOTA) performance.

However, the superfluous semantic information in multi-modal features is ignored. We present an illustrative example depicted

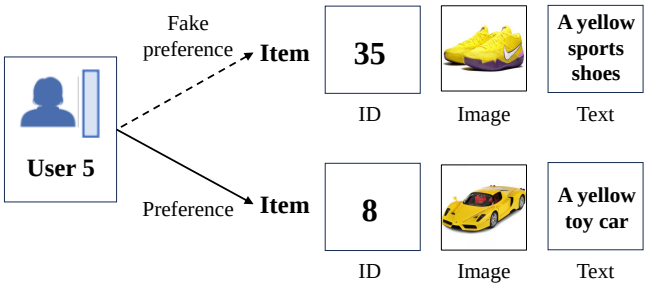

Figure 1: This Figure illustrates the existence of superfluous semantic information in multi-modal features. The superfluous semantic information originates from the superficial attributes such as shared color (yellow).

in Figure 1. Assuming that user 5 is a child who prefers kind of toys and dislikes doing sports, hence the interaction between user 5 and item 8 is observed. Nevertheless, the corresponding text and image of item 8 have identical color *"yellow"*, so representation of item 8 learned from corresponding ID, image and text contains semantics about the color *"yellow"*. Similarly, representation of item 35 includes semantics about color *"yellow"* also. The presence of the *"yellow"* semantics can lead recommendation models to infer an interaction between user 5 and item 35, which contradicts our initial assumption and represents a fake preference. We define such semantic information as **superfluous semantics**. To remedy this deficiency, we propose a superfluous semantics discarding module. In this module, we conduct theoretical analysis on how to diminish the superfluous semantics in multi-modal features from the perspective of mutual information, and compress the representation of multi-modal data to diminish redundant semantics.

On top of this, we disclose that multi-modal features contain deficient user-item interaction information. To verify our statement, based on [45], we conduct exploratory experiments that utilize three pairs of representations (user, item's ID), (user, item's text), and (user, item's image) for recommendation respectively and evaluate performances on three benchmark datasets in Table 1. We observe that leveraging representations of (user, item's ID) outperforms the compared pairs by significant margins, demonstrateing multi-modal embeddings contain deficient interaction information compared to ID embeddings. Therefore, the user-item interaction information diminishes inevitably during the interaction of different items' embeddings. Interaction preserving module is proposed to address this issue. Specifically, we construct two multi-modal graphs by replacing the initial representations of item nodes, i.e., ID-based embeddings, by textual and visual embeddings respectively. Then, we propagate and aggregate information on the graphs to obtain the informative textual and visual representations of items augmented by interaction information. Therefore, we can avoid the user-item interaction information degeneration incurred by the introduced multi-modal features.

Concretely, we propose a novel multi-modal recommendation method named **MSI**, which conducts ***M**ulti-modal recommendation by **S**uperfluous semantics discarding and **I**nstance preserving*. We further empirically demonstrate the superiority of MSI (e.g., outperforming BM3 with 3.60%, 2.34% and 2.46% in terms of Recall@20 on Baby, Sports and Elec respectively) and the effectiveness of each ingredient on multi-modal recommendation task. Our main contributions are summarized as follows:

1. We disclose the existence of superfluous semantics and demonstrate the performance degeneration incurred by the lack of interaction information in multi-modal features.

2. To tackle these issues, we propose a novel approach for multi-modal recommendation by introducing the superfluous semantics discarding and instance preserving modules.

3. Extensive comparisons demonstrate that proposed MSI achieves the state-of-the-art performance on three multi-modal recommendation datasets.

Table 1: In BM3, we leverage item representations of different modals for recommendation and evaluate the performance on benchmark multi-modal datasets by metrics Recall@10, Recall@20, NDCG@10 and NDCG@20. The results show that multi-modal features contain deficient user-item interaction information, thereby achieving degraded performance.

Datasets	Modal	R@10	R@20	N@10	N@20
Baby	Text	0.0386	0.0641	0.0196	0.0262
	Image	0.0447	0.0735	0.0226	0.0300
	ID	**0.0538**	**0.0860**	**0.0288**	**0.0370**
Sports	Text	0.0505	0.0787	0.0269	0.0341
	Image	0.518	0.0819	0.0272	0.0349
	ID	**0.0649**	**0.0973**	**0.0353**	**0.0437**
Elec	Text	0.0259	0.0400	0.0138	0.0175
	Image	0.0330	0.0495	0.0181	0.0223
	ID	**0.0437**	**0.0648**	**0.0247**	**0.0301**

2 RELATED WORK

GNN-based recommendation: GNNs have gained widespread adoption within recommender systems, facilitating the modeling of diverse relationships across various recommendation scenarios. For example, 1) collaborative filtering (e.g., LightGCN [15], LR-GCCF [3]); 2) social recommendation (e.g., GraphRec [8], KCGN [16]); 3) sequential recommendation (e.g., GCEGNN [39], SURGE [2]); 4) knowledge graph-enhanced recommender (e.g., KGAT [38]). Inspired by the extensive applications of GNNs in various domains of recommendation systems, our proposed method MSI adopts GNN as backbone to model high-order collaborative connections with assistance of multi-modal contextual information.

Multi-modal Recommendation: The early multi-modal recommendation methods learn the representation of users and items on top of the CF (Collaborative Filtering) paradigm. For example, VBPR [14] leverages the pre-trained visual features of items to make predictions. Within the BPR framework, Deepstyle [29] learns the informative representations of items with both visual and style features. Recently, models based on GNNs are receiving more and more attentions. MMGCN [41] constructs modality-specific graph to acquire the representations of users and items. GRCN [40] removes the false-positive edges and learns the representations on top of

the refined graph. DualGNN [37] and LATTICE [43] both leverage auxiliary graph for recommendation. FREEDOM [44] further researches the role of item-item graph for effective recommendation. BM3 [45] proposes a self-supervised multi-modal recommendation model without negative samples. Although BM3 [45] achieves the state-of-the-art performance, the ignoring of superfluous semantic information and the interaction degeneration limit the applications in excessive scenarios.

3 METHODOLOGY

We elaborate on our model in this section. The overall architecture is depicted in Figure 2.

The raw data includes user ID x^u and (ID, text, image) $\{x^i, m^t, m^v\}$ of item, which are encoded into embeddings by specific methods. Visual and linguistic features are encoded by methods mentioned in **Section** 4, ID embeddings of item and user are both sampled randomly from uniform distributions. We denote representations of user and item by z^u and $\{z^i, z^t, z^v\}$ respectively. In the following, we expound superfluous semantics discarding module, which diminishes recommendation-irrelevant information in $\{z^t, z^v\}$ for robust multi-modal representations $\{z_r^t, z_r^v\}$ of item.

3.1 Superfluous Semantics Discarding Module

Theoretical analysis. Inspired by the success of applying information theory [1, 9, 23, 30, 36, 42] in representation learning, we conduct theoretical analysis for recommendation from the mutual information perspective. Given item x, its representation z and recommendation label y, the definition of *sufficiency* is formulated as follows:

Definition 3.1. (*Item sufficiency for recommendation*) The representation z of item x is sufficient to predict recommendation label y if and only if $I(x; y|z) = 0$. (I stands for mutual information).

Definition 3.1 states that z is sufficient for y means representation z of item x contains enough task-relevant information to predict recommendation label y. Formally, with $I(x; y; z) = I(x; z) - I(x; z|y) = I(y; z) - I(y; z|x)$, we divide $I(x; z)$ into two components:

$$\begin{aligned} I(x; z) &= I(x; z|y) + I(y; z) - I(y; z|x) \\ &= \underbrace{I(x; z|y)}_{\text{task irrelevant}} + \underbrace{I(y; z)}_{\text{recommendation}} . \end{aligned} \tag{1}$$

$I(y; z|x) = 0$ because z is the representation of x, z is conditionally independent from any other variable once x is observed. $I(x; z|y)$ represents information shared by x and z which is irrelevant with y, while $I(y; z)$ determines the information shared by recommendation label y and representation z. Therefore, in the general supervised setting, we can learn robust representation for recommendation by minimizing $I(x; z|y)$ and maximizing $I(y; z)$.

However, in conventional recommendation, y comprises observed positive samples and randomly selected unobserved negative samples which can bring into wrong supervision signals [45]. To remedy this issue, based on a basic assumption in multi-modal learning that different modalities provide same task-relevant information, we derive further theoretical analysis which does not require y. Let m_1, m_2 denote multi-modal data corresponding to item x, we can guarantee that z is sufficient for recommendation label y as long as that z contains sufficient recommendation-relevant information which is shared by m_1 and m_2, even without knowing y. Beginning with definition of *Redundancy*, We formulate how to utilize multi-modal data to eliminate recommendation-irrelevant information.

Definition 3.2. (*Multi-modal data redundancy of item for recommendation*) m_1 is redundant with respect to m_2 for predicting the recommendation label y if and only if $I(y; m_1|m_2) = 0$.

Definition 3.2 states that information shared by recommendation label y and multi-modal data m_1 which is also contained in multi-modal data m_2. Under the condition of **mutual redundancy** (m_1 is redundant with respect to m_2 for y and vice-versa), let z_1 be the representation of m_1, and if z_1 is sufficient for m_2, i.e., $I(m_1; m_2|z_1) = 0$, then for y, z_1 is as predictive as the joint observation of the two modals, i.e., $I(m_1 m_2; y) = I(z_1; y)$. Briefly, any item representation containing all information shared by both modals is as predictive as their joint observation.

Similar to Equation 1, we divide $I(m_1; z_1)$ into two components:

$$I(m_1; z_1) = \underbrace{I(m_1; z_1|m_2)}_{\text{task irrelevant}} + \underbrace{I(m_2; z_1)}_{\text{recommendation}} . \tag{2}$$

As we can see, recommendation label y is unnecessary in Equation 2, therefore avoiding the "harmful" information brought by negative samples. $I(m_2; z_1)$ indicates shared information between m_2 and z_1, which contains recommendation information. And $I(m_1; z_1|m_2)$ represents information contained by z_1, which is unique to m_1 and is useless for predicting recommendation label y when m_2 is observed. This information is apparently irrelevant to recommendation task, and therefore, we propose to minimize such a conditional mutual information.

Symmetrically, we have:

$$I(m_2; z_2) = \underbrace{I(m_2; z_2|m_1)}_{\text{task irrelevant}} + \underbrace{I(m_1; z_2)}_{\text{recommendation}} . \tag{3}$$

Superfluous semantics discarding loss. Based on above theoretical analysis, we can formalize the loss function in this module. We denote textual and visual data by m_t, m_v and suppose that m_t, m_v satisfy the mutual redundancy condition for recommendation, we define an objective function for the representation z_t of m_t that discards as much information as possible without losing any recommendation-relevant information. We have shown that maximizing $I(m_v; z_t)$, and decreasing $I(z_t; m_t|m_v)$ can increase robustness of representation. Therefore, we can combine these two requirements by using a relaxed Lagrangian objective [9]:

$$\mathcal{L}_1(\theta_1; \lambda_1) = I_{\theta_1}(\mathbf{z}_t; \mathbf{m}_t \mid \mathbf{m}_v) - \lambda_1 I_{\theta_1}(\mathbf{m}_v; \mathbf{z}_t), \tag{4}$$

where θ_1 represents parameters of encoder $p_{\theta_1}(z_1|m_1)$ and λ_1 denotes Lagrangian multiplier. Symmetrically, we can define the following loss function:

$$\mathcal{L}_2(\theta_2; \lambda_2) = I_{\theta_2}(\mathbf{z}_v; \mathbf{m}_v \mid \mathbf{m}_t) - \lambda_1 I_{\theta_2}(\mathbf{m}_t; \mathbf{z}_v), \tag{5}$$

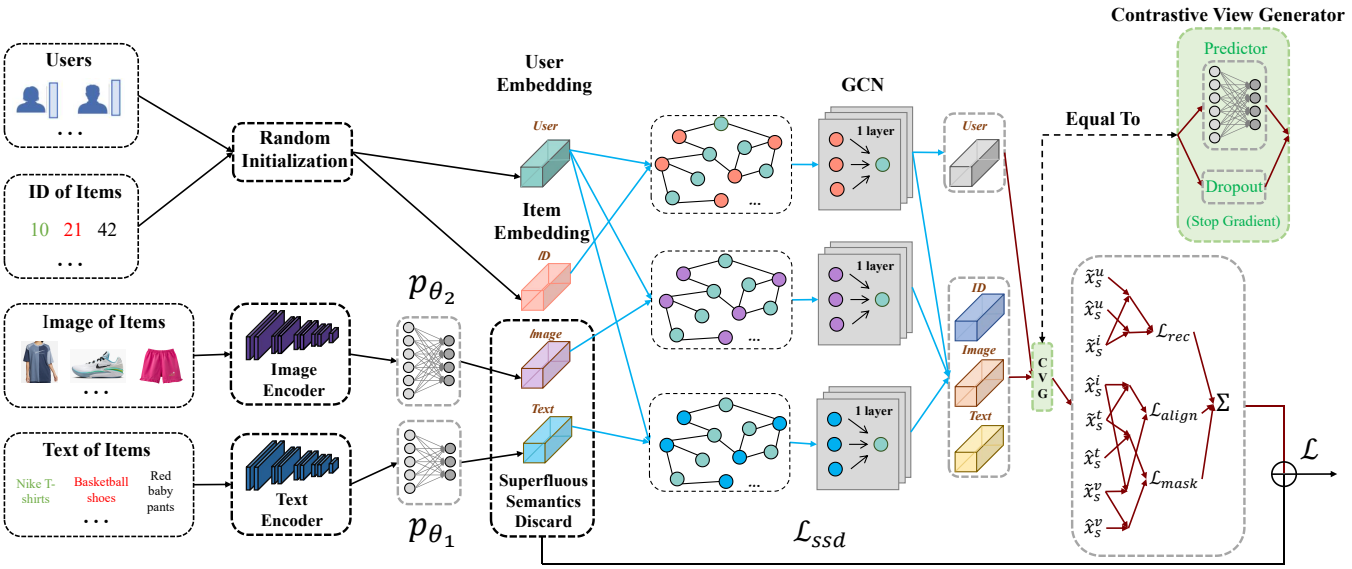

Figure 2: Architecture of MSI. $p_{\theta_1}, p_{\theta_2}$ denote $p_{\theta_1}(z_1|m_1), p_{\theta_2}(z_2|m_2)$, respectively. CVG represents Contrastive View Generator.

and the average of two loss functions $\mathcal{L}_1, \mathcal{L}_2$ is:

$$\begin{aligned}\frac{1}{2}(\mathcal{L}_1+\mathcal{L}_2) =& \frac{1}{2}(I_{\theta_1}(\mathbf{z}_t;\mathbf{m}_t \mid \mathbf{m}_v) + I_{\theta_2}(\mathbf{z}_v;\mathbf{m}_v \mid \mathbf{m}_t)) \\ &- \frac{1}{2}(\lambda_1+\lambda_2)(I_{\theta_1}(\mathbf{m}_v;\mathbf{z}_t) + I_{\theta_2}(\mathbf{m}_t;\mathbf{z}_v)).\end{aligned} \tag{6}$$

However, directly calculating $\frac{1}{2}(\mathcal{L}_1+\mathcal{L}_2)$ is problematic, we derive:

$$I_{\theta_1}(\mathbf{m}_t;\mathbf{z}_t \mid \mathbf{m}_v) \leq D_{\mathrm{KL}}\left(p_{\theta_1}(\mathbf{z}_t \mid \mathbf{m}_t) \| p_{\theta_2}(\mathbf{z}_v \mid \mathbf{m}_v)\right), \tag{7}$$

where KL denotes Kullback-Leibler divergence [20] and

$$I_{\theta_1}(\mathbf{z}_t;\mathbf{m}_v) \geq I_{\theta_1\theta_2}(\mathbf{z}_t;\mathbf{z}_v). \tag{8}$$

The derivations of Equation 7 and 8 can be found in **Appendix** A.1. In the same way, we have:

$$\begin{aligned}I_{\theta_2}(\mathbf{m}_v;\mathbf{z}_v \mid \mathbf{m}_t) &\leq D_{\mathrm{KL}}\left(p_{\theta_2}(\mathbf{z}_v \mid \mathbf{m}_v) \| p_{\theta_1}(\mathbf{z}_t \mid \mathbf{m}_t)\right) \\ I_{\theta_2}(\mathbf{m}_t;\mathbf{z}_v) &\geq I_{\theta_1\theta_2}(\mathbf{z}_t;\mathbf{z}_v).\end{aligned} \tag{9}$$

Hence, we can formalize the upper bound of the average loss as follows:

$$\begin{aligned}\frac{1}{2}(\mathcal{L}_1+\mathcal{L}_2) \leq& D_{AKL}\left(p_{\theta_1}(\mathbf{z}_t \mid \mathbf{m}_t) \| p_{\theta_2}(\mathbf{z}_v \mid \mathbf{m}_v)\right) \\ &- \frac{\lambda_1+\lambda_2}{2} I_{\theta_1\theta_2}(\mathbf{z}_t;\mathbf{z}_v),\end{aligned} \tag{10}$$

where AKL denotes average Kullback-Leibler divergence and

$$\begin{aligned}D_{AKL}(p_{\theta_1}(\mathbf{z}_t \mid \mathbf{m}_t) \| p_{\theta_2}(\mathbf{z}_v \mid \mathbf{m}_v)) = \frac{1}{2}[D_{\mathrm{KL}}(p_{\theta_1}(\mathbf{z}_t \mid \mathbf{m}_t) \\ \| p_{\theta_2}(\mathbf{z}_v \mid \mathbf{m}_v)) + D_{\mathrm{KL}}(p_{\theta_2}(\mathbf{z}_v \mid \mathbf{m}_v) \| p_{\theta_1}(\mathbf{z}_t \mid \mathbf{m}_t))].\end{aligned} \tag{11}$$

We multiply each term in the both sides of Equation 10 with $\beta = \frac{2}{\lambda_1+\lambda_2}$:

$$\begin{aligned}\frac{\beta}{2}(\mathcal{L}_1+\mathcal{L}_2) \leq& \beta D_{AKL}\left(p_{\theta_1}(\mathbf{z}_t \mid \mathbf{m}_t) \| p_{\theta_2}(\mathbf{z}_v \mid \mathbf{m}_v)\right) \\ &- I_{\theta_1\theta_2}(\mathbf{z}_t;\mathbf{z}_v).\end{aligned} \tag{12}$$

We reckon that the loss function is utilized to train encoders by adopting the back-propagation scheme, such that the *exact* value of the loss is not required, while the *relative* value of the loss is desired. We thus obtain the ultimate loss function by adopting the right side of Equation 12 as follows:

$$\mathcal{L}_{ssd} = -I_{\theta_1\theta_2}(\mathbf{z}_t;\mathbf{z}_v) + \beta D_{AKL}\left(p_{\theta_1}(\mathbf{z}_t \mid \mathbf{m}_t) \| p_{\theta_2}(\mathbf{z}_v \mid \mathbf{m}_v)\right). \tag{13}$$

The decreasing of $\mathcal{L}_{ssd}$ results in more robust multi-modal representations $\{z_r^t, z_r^v\}$. Then, the user embedding z^u and item embeddings $\{z^i, z_r^t, z_r^v\}$ are fed into the following module.

3.2 Interaction Preserving Module

Item embeddings play an vital role in recommendation [43]. The deficiency of interaction information in multi-modal embeddings of item can impair recommendation performance. In this module, we construct two multi-modal graphs by using (z^u, z_r^t) and (z^u, z_r^v) respectively and leverages LightGCN [15] to integrate user-item interaction information into multi-modal features.

We construct the bipartite graph $\mathcal{G}$ by the given user-item interactions, where $\mathcal{G} = \{(u, r_{ui}, i) | u \in \mathcal{U}, i \in \mathcal{I}\}$. We denote user, item and node set by $\mathcal{U}, \mathcal{I}, \mathcal{V} = \mathcal{U} \cup \mathcal{I}$ respectively. $r_{ui} = 1$ represents that user u interacts with item i, otherwise $r_{ui} = 0$. The number of nodes is denoted by $|\mathcal{V}|$. The adjacency matrix is denoted by $A \in R^{|\mathcal{V}|\times|\mathcal{V}|}$ and degree matrix is denoted by D.

Moreover, we use $Z_l^m \in R^{|\mathcal{V}|\times d}(m \in (i,t,v))$ to denote embeddings of modal m at l-th layer by concating embeddings of users and items at layer l. Specifically:

$$Z_0^m = \begin{cases} Concat[z^u, z_r^m] & if \quad m \in (t,v) \\ Concat[z^u, z^i] & if \quad m = i \end{cases}. \tag{14}$$

A conventional propagation GCN [18] to calculate the hidden ID embedding Z_{l+1} at layer $l+1$ is recursively conducted as:

$$Z_{l+1} = \sigma(\hat{A} Z_l W^l), \tag{15}$$

where $\sigma(\cdot)$ is a non-linear function. $\hat{A} = \hat{D}^{-\frac{1}{2}}(A+I)\hat{D}^{-\frac{1}{2}}$ is the re-normalization of adjacency matrix A, and $\hat{D}$ is a diagonal degree matrix of $A+I$. For better recommendation, LightGCN [15] removes feature transformation W^l and non-linear activation $\sigma(\cdot)$ layers. The simplified propagation formulation in LightGCN is defined as

$$\mathrm{Z}^{l+1} = \left(\mathrm{D}^{-1/2}\mathrm{A}\mathrm{D}^{-1/2}\right)\mathrm{Z}^l, \tag{16}$$

in which the node embeddings of the $(l+1)$-th layer are only linearly aggregated from the l-th layer with $D^{-\frac{1}{2}}AD^{\frac{1}{2}}$. We use a Readout function to aggregate all representations in hidden layers to obtain final representations. Nevertheless, over-smoothing is a general problem that many GCNs suffer [4, 28], inspired by [43], a residual module to item embeddings is employed to mitigate the deficieny:

$$\begin{aligned} \mathrm{Z}^u &= \mathrm{Readout}\{Z_0^u, Z_1^u, \cdots, Z_L^u\} \\ \mathrm{Z}^m &= \mathrm{Readout}\{Z_0^m, Z_1^m, \cdots, Z_L^m\} + Z_0^m (m \in (t, v, i)), \end{aligned} \tag{17}$$

we implement $\mathrm{Readout}(\cdot)$ by conventional mean function. With information propagation and aggregation on three graphs, we can obtain $Z_{u-i} = concat[z_s^u, z_s^i]$, $Z_{u-t} = concat[z_s^{u-t}, z_s^t]$, $Z_{u-v} = concat[z_s^{u-v}, z_s^v]$. We split $Z_{u-i}, Z_{u-t}, Z_{u-v}$, then only keep user embedding z_s^u and ID, text, visual embeddings of items $\{z_s^i, z_s^t, z_s^v\}$ for subsequent module.

3.3 Multi-modal Contrastive Loss

This section is inspired by the empirical success of contrastive learning [5, 22, 25, 32]. Concretely, the negative samples in calculating BPR loss can bring wrong information and [11] proposes a self-supervised learning framework which does not require negative samples. Inspired by the stop-gradient strategy used in [11, 45], we construct following loss functions.

For a representation z_s, we employ contrastive view generator (CVG, which comprises dropout [35] and a linear predictor) to create two views as follows:

$$\begin{aligned} \hat{z_s} &= z_s \cdot Bernoulli(p) \\ \tilde{z_s} &= z_s W_p + b_p, \end{aligned} \tag{18}$$

where p is the dropout ratio and $Bernoulli(\cdot)$ represents Bernoulli distribution function [19]. W_p, b_p denote linear transformation matrix and bias respectively. We put stop-gradient on $\hat{z}_s$.

By CVG, we have $\{\hat{z}_s^u, \tilde{z}_s^u\}, \{\hat{z}_s^i, \tilde{z}_s^i\}, \{\hat{z}_s^t, \tilde{z}_s^t\}, \{\hat{z}_s^v, \tilde{z}_s^v\}$. For recommendation tasks, we want to maximize similarity between observed user-item interaction. Hence we define the recommendation loss function:

$$\mathcal{L}_{rec} = C(\hat{z}_s^u, \tilde{z}_s^i) + C(\tilde{z}_s^u, \hat{z}_s^i), \tag{19}$$

where $C(\cdot,\cdot)$ is defined by:

$$C(x, y) = 1 - \frac{x^T y}{||x||_2 ||y||_2}. \tag{20}$$

Additionally, ID, text and visual embeddings are both representations of same item from different modals. ID embeddings should close to multi-modal features intuitively. Therefore, we have alignment loss as follows:

$$\mathcal{L}_{align} = C(\tilde{z}_s^m, \hat{z}_s^i), m \in (t, v). \tag{21}$$

Furthermore, we use the following mask loss to make the multi-modal representation sparse, which is implemented by

$$\mathcal{L}_{mask} = C(\hat{z}_s^m, \tilde{z}_s^m). \tag{22}$$

Meanwhile, to avoid the over-fitting, the regularization penalty is employed. Inspired by multi-task learning, we optimize our recommendation system by the following loss function:

$$\begin{aligned} \mathcal{L} =& \mathcal{L}_{rec} + \mathcal{L}_{align} + \mathcal{L}_{mask} + \alpha \mathcal{L}_{ssd} \\ &+ \gamma \cdot (||z_s^u|| + ||z_s^i|| + ||z_s^t|| + ||z_s^v||), \end{aligned} \tag{23}$$

where α, γ are the hyper-parameters to balance the influence of $\mathcal{L}_{ssd}$ and regularization penalty.

Algorithm 1 MSI's training pipeline

Input: batch size N, user ID $\{x_1^u, x_2^u, \cdots, x_N^u\}$, item ID $\{x_1^i, x_2^i, \cdots, x_N^i\}$, item text $\{m_1^t, m_2^t, \cdots, m_N^t\}$ and item image $\{m_1^v, m_2^v, \cdots, m_N^v\}$, the model parameter θ, learning rate lr.
Initialize: the model parameter θ.
Repeat
  for $n = 1$ **to** N **to do**:
    1.Encode $\{x_n^u, x_n^i, m_n^t, m_n^v\}$ and obtain embeddings $\{z_n^u, z_n^i, z_n^t, z_n^v\}$.
    # Superfluous semantics discarding module.
    2.Calculate $\mathcal{L}_{ssd}$ and get $\{z_n^u, z_n^i, z_{r_n}^t, z_{r_n}^v\}$
    # Instance preserving module.
    3.Construct multi-modal graphs and leverage LightGCN to get $\{Z_{u-i_n}, Z_{u-t_n}, Z_{u-v_n}\}$.
    4.Split $\{Z_{u-i_n}, Z_{u-t_n}, Z_{u-v_n}\}$ and keep $\{z_{s_n}^u, z_{s_n}^i, z_{s_n}^t, z_{s_n}^v\}$.
    5.Calculate multi-modal contrastive loss $\mathcal{L}_n$.
  end for.
  Update $\theta \leftarrow \theta - lr\nabla_\theta(\mathcal{L})$ # $\mathcal{L} = \frac{1}{n}\sum \mathcal{L}_n$
Until convergence.

3.4 Top-k Recommendation

In recommendation task, to recommend items for a user, we first calculate recommendation scores between user and candidate items, a high score indicates that user prefers the item. Sequentially, we rank candidate items based on recommendation scores in descending order, and select top-k items as recommendations for the user. Since the informative representation learned by our MSI, recommendation score is calculated by inner product between $\tilde{z}_s^u$ and $\tilde{z}_s^i$:

$$Scores(user, item) = \langle \tilde{z}_s^u \cdot \tilde{z}_s^i \rangle. \tag{24}$$

4 EXPERIMENTS

Datasets. We conduct our experiments on Amazon review dataset [12], which has linguistic descriptions and corresponding images. The Amazon review can be divided by the categories of commodity and we choose three kinds of datasets denoted by Baby, Sports and Elec. In the processing of datasets, we split the interaction record

Table 2: Recommendation results on three multi-modal datasets. The "Type" indicates the method is uni-modal or multi-modal. The best result is in Bold and the second best is underlined. '-' indicates the model cannot be fitted into a Tesla V100 GPU with 32 GB memory, which is consistent with BM3 [45].

Model	Type	Baby				Sports				Elec			
		R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20
BPR	U	0.0357	0.0575	0.0192	0.0249	0.0432	0.0653	0.0241	0.0298	0.0235	0.0367	0.0127	0.0161
LightGCN		0.0479	0.0754	0.0257	0.0328	0.0569	0.0864	0.0311	0.0387	0.0363	0.0540	0.0204	0.0250
BUIR		0.0506	0.0788	0.0269	0.0342	0.0467	0.0733	0.0260	0.0329	0.0332	0.0514	0.0185	0.0232
VBPR	M	0.0423	0.0663	0.0223	0.0284	0.0558	0.0856	0.0307	0.0384	0.0293	0.0458	0.0159	0.0202
MMGCN		0.0378	0.0615	0.0200	0.0261	0.0370	0.0605	0.0193	0.0254	0.0207	0.0331	0.0109	0.0141
GRCN		0.0532	0.0824	0.0282	0.0358	0.0559	0.0877	0.0306	0.0389	0.0349	0.0529	0.0195	0.0241
DualGNN		0.0448	0.0716	0.0240	0.0309	0.0568	0.0859	0.0310	0.0385	0.0363	0.0541	0.0202	0.0248
LATTICE		0.0544	0.0848	0.0291	0.0369	0.0618	0.0947	0.0337	0.0422	-	-	-	-
BM3*		0.0538*	0.0860*	0.0288*	0.0370*	0.0649*	0.0973*	0.0353*	0.0437*	0.0437*	0.0648*	0.0247*	0.0301*
BM3		0.0564	0.0883	0.0301	0.0383	0.0656	0.0980	0.0355	0.0438	0.0437	0.0648	0.0247	0.0302
MSI	M	**0.0575**	**0.0891**	**0.0307**	**0.0389**	**0.0681**	**0.1023**	**0.0374**	**0.0462**	**0.0449**	**0.0664**	**0.0253**	**0.0309**

Table 3: Statistics of the datasets.

Datasets	Users	Items	Interactions
Baby	19445	7050	160792
Sports	35598	18357	296337
Elec	192403	63001	1689118

with a 8:1:1 ratio and form train, test and valid datasets following [37]. We treat each review rating as a record of positive user-item interactions, which is a common operation in many previous works [13, 15]. The detailed statistics of datasets are described in Table 3. As for the visual and linguistic data, we adopt 4096-dimensional features processed and published in [31] according to [43]. Simultaneously, we concatenate the title, descriptions, categories and brand to obtain the linguistic representation of each item and leverage sentence-transformers [33] to get the 384-dimensional text embeddings.

Baseline Methods. To verify the superiority of MSI, we compare our model with competitive baselines including normal CF recommendation and multi-modal recommendation models.

- **BPR** [34] bases on the matrix factorization and is optimized by a pair-wise loss.
- **LightGCN** [15] utilizes the average hidden layer embeddings for prediction with removing the nonlinear activation function and feature transformations from standard graph convolution network.
- **BUIR** [21] is a self-supervised framework based on LightGCN which does not require negative samples.
- **VBPR** [14] utilizes the visual features to learn the informative representation of items for user preference learning.
- **MMGCN** [41] designs modal-specific graphs to learn users' preference.
- **GRCN** [40] discards the false-positive edges and refines the user-item bipartite graph to gain users and items representations.
- **DualGNN** [37] constructs an additional user-user graph based on the user-item graph to obtain the representation of users and items.
- **LATTICE** [43] leverages an auxiliary item-item graph and learns representations of users and items by graph convolutional operations on both item-item and user-item interaction graph.
- **BM3** [45] proposes a novel self-supervised learning framework for multi-modal recommendation.

Implementation Details. Recall@10, Recall@20, NDCG@10 and NDCG@20 (denoted as R@10, R@20, N@10, N@20) as treated as our metrics to evaluate the performance of models. To be consistent with other existing studies [15, 43], in training, we fix the embedding size of both users and items to 64, initialize the embedding parameters with Xavier method [10] and treat Adam [17] with a learning rate of 0.001 as our optimizer. Our proposed model is implemented by PyTorch. The number of GCN layers is tuned in $\{1, 2, 3, 4\}$. The dropout rate for embedding perturbation is chosen from $\{0.3, 0.5\}$, and the regularization coefficient γ is searched in $\{0.1, 0.01\}$, the hyper-parameters $\alpha.\beta$ are both searched in $\{1e-1, 1e-2, 1e-3, 1e-4\}$. We conduct our experiments on a Tesla V100 GPU.

Superiority of Our Model. Comparison results between our model and competitive baselines are summarized in Table 2, key observations are as follows:

(1) Our MSI is inspired by BM3 [45], the 'BM3*' represents results reproduced by us in the same environment. Although the reproduced results can not catch the values reported in BM3 [45] on the baby dataset, proposed MSI remarkably outperforms both traditional recommendation methods and prevalent multi-modal recommendation methods on three datasets. Concretely, our model improves the best baselines by 3.60%, 2.34% and 2.46% in terms of Recall@20 on Baby, Sports and Elec respectively. The results verify the superiority of MSI recommendation on multi-modal recommendation datasets of various scale.

(2) Intuitively, multi-modal recommendation methods can utilize more modalities, hence resulting in higher performance. However,

Table 4: Ablation study on different ingredients of MSI.

Model	Baby				Sports				Elec			
	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20
MSI	**0.0575**	**0.0891**	**0.0307**	**0.0389**	**0.0681**	**0.1023**	**0.0374**	**0.0462**	**0.0449**	**0.0664**	**0.0253**	**0.0309**
w/o S	0.0566	0.0883	0.0300	0.0381	0.0671	0.0990	0.0361	0.0441	0.0445	0.0658	0.0250	0.0305
w/o I	0.0547	0.0879	0.0290	0.0276	0.0651	0.0980	0.0359	0.0439	0.0441	0.0652	0.0249	0.0303

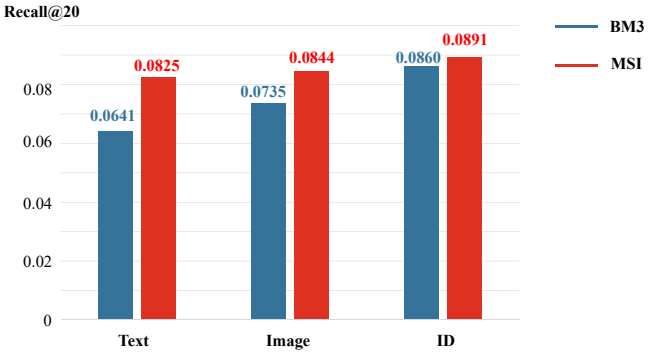

Figure 3: The recommendation performance achieved by MSI and BM3 with respect to representations of different pairs in terms of Recall@20.

as reported in the Table 2, not all multi-modal recommendation models outperform the uni-modal recommendation models without leveraging modal features. We speculate that the reason lies in the way leveraging multi-modal features. As stated in **Section** 1, the initial multi-modal features lack of interaction information, MMGCN, GRCN and DualGNN simply fuse ID embeddings with multi-modal embeddings, which causes the degeneration of interaction. Meanwhile, MMGCN, GRCN and DualGNN ignore the noisy information contained in multi-modal features which is demonstrated by the ablation study in the following. Our proposed MSI can alleviate these deficiencies. MSI avoids the interaction degeneration by integrating interaction information into multi-modal features, and discard the superfluous semantics under solid theoretical analysis. Therefore MSI achieves the performance which is superior to the multi-modal recommendations mentioned above.

The ablation study on ingredients of MSI. To fully understand the role each ingredient plays, we conduct an ablation study on three datasets which is shown in Table 4. We train an ablation model without superfluous semantics discarding (w/o S), a model without interaction preserving (w/o I) and a module without both (w/o S & I) which equals to BM3 [45]. We observe that the performance of the model drops (compared with complete model) regardless of which module is removed, indicating that both modules are helpful for recommendation task. Comparing these two modules, the removal of the interaction preserving module has a greater impact on performance, which suggests encoding user-item interaction information into multi-modal representations is productive for recommendation.

The ablation study of loss functions in MSI: To explore the role that different loss functions play, we train three variants of

Table 5: Ablation study on loss functions of MSI.

Datasets	Variants	R@10	R@20	N@10	N@20
Baby	MSI w/o a&m	0.0523	0.0843	0.0279	0.0361
	MSI w/o a	0.0559	0.0879	0.0295	0.03780
	MSI w/o m	0.0543	0.0863	0.0289	0.0376
	MSI	0.0575	0.0891	0.0307	0.0389
Sports	MSI w/o a&m	0.0625	0.0953	0.0351	0.0435
	MSI w/o a	0.0639	0.0963	0.0361	0.0437
	MSI w/o m	0.0658	0.0971	0.0365	0.0439
	MSI	0.0681	0.1023	0.0374	0.0462
Elec	MSI w/o a&m	0.0438	0.0649	0.0244	0.0298
	MSI w/o a	0.0405	0.0609	0.0222	0.0272
	MSI w/o m	0.0417	0.0625	0.0231	0.0283
	MSI	0.0449	0.0664	0.0253	0.0309

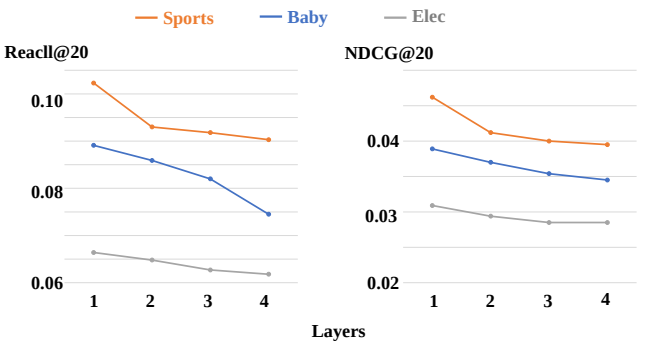

Figure 4: Top-20 recommendation accuracy of MSI varies with the number of layers L.

MSI for analysis. MSI w/o a & m denotes the MSI without $\mathcal{L}_{align}$ and $\mathcal{L}_{mask}$ (It is noteworthy that MSI w/o a&m equals to MSI w/o v&t). MSI w/o a and MSI w/o m denote the MSI without $\mathcal{L}_{align}$ and $\mathcal{L}_{mask}$ respectively. The experimental results are shown in Table 5. The $\mathcal{L}_{align}$ and $\mathcal{L}_{mask}$ can enhance the model performance of MSI on Baby and Sports datasets. However, MSI w/o a&m utilizing only $\mathcal{L}_{align}$ or only $\mathcal{L}_{mask}$ impairs recommendation accuracy on Elec dataset. Furthermore, the effectiveness of $\mathcal{L}_{align}$ and $\mathcal{L}_{mask}$ also varies with the datasets. The importance of $\mathcal{L}_{align}$ is superior to $\mathcal{L}_{mask}$ in the Baby dataset. while the importance of $\mathcal{L}_{align}$ is inferior to $\mathcal{L}_{mask}$ in the Sports dataset. According to the observations of ablation studies on multi-modal features and the loss function, we disclose that the recommendation accuracy on the

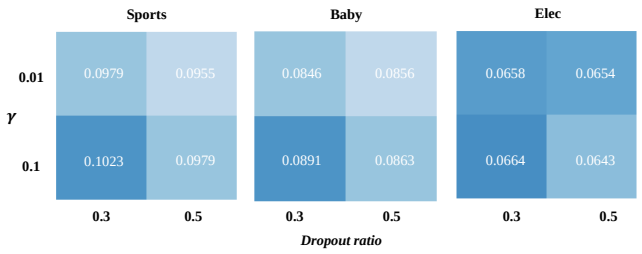

Figure 5: The performance achieved by MSI with respect to different combinations of the dropout ratio and λ on three datasets. Darker background indicates better recommendation accuracy.

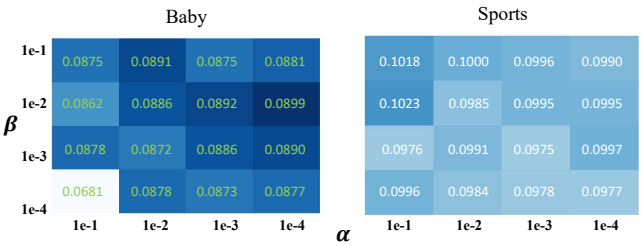

Figure 6: The Performance achieved by MSI with respect to different combinations of α and β on Baby and Sports datasets. Darker background indicates better recommendation accuracy.

large dataset such as Elec performs a disparate pattern with the small-scale datasets i.e., Baby and Sports). That is, the separate textual feature, visual features, $\mathcal{L}_{align}$ or $\mathcal{L}_{mask}$ in MSI shows no enhancement in recommendation accuracy on the Elec dataset.

The effectiveness of interaction preserving. To demonstrate that interaction preserving module can integrate user-item interaction into multi-modal features. We conduct experiments on Baby dataset, which is depicted in Figure 3. Same with Table 1, we use representations of three pairs (user, item's ID), (user, item's text), (user, item's image) for recommendation and select Recall@20 to evaluate the performance. The textual and visual features in BM3 contain deficient interaction information whereas MSI integrates interaction information into multi-modal features and enhances their performance markedly. As we can see, MSI improves the performance by 28.7%, 14.8%, 3.4% in terms of Recall@20 respectively. Hence, interaction preserving module can boost the recommendation performance by avoiding the diminishing of interaction information in multi-modal features.

Sensitivity study of hyper-parameters. We conduct a hyper-parameter sensitivity study with regard to recommendation performance in terms of Recall@20. At least two datasets are chosen to evaluate the performance of MSI under different hyperparameter settings. The following five hyper-parameters are considered, i.e., the number of GCN layers L, the ratio of embedding dropout, the regularization coefficient γ, the coefficient α and β.

The number of GCN layers in MSI varies within {1, 2, 3, 4}. Figure 4 illustrates the performance trends of MSI corresponding to different settings of L. As shown in Figure 4, on three datasets, we obtain the best result when layer equals to 1. Furthermore, MSI performs relatively slight performance degradation with the increase of layers. We speculate the reason is the over-smoothing of users and items representations.

As mentioned in **Implementation Details**, we search the dropout ratio in {0.3, 0.5} and the regularization coefficient γ in {0.1, 0.01}. Figure 5 reveals the performance with various combinations of dropout ratio and regularization coefficient λ in terms of Recall@20. MSI achieves the best performance with 0.3 dropout ratio and λ = 0.1.

Our hyperparameter α, β are searched in {1e-1, 1e-2, 1e-3, 1e-4} both. Figure 6 reveals the performance achieved by MSI with different combinations of α and β in terms of Recall@20. It is worth noting that even MSI achieves the best performance with $\{\alpha = 1e-2, \beta = 1e-1\}$ and $\{\alpha = 1e-1, \beta = 1e-2\}$ on the Baby and Sports datasets in terms of Recall@20 respectively, the model performance drops dramatically in terms of Recall@10, NDCG@10 and NDCG@20. Therefore, for the consideration of comprehensive performance, the eventual combinations of hyper-parameters on Baby, Sports, Elec datasets are {layer = 1, dropout ratio = 0.3, λ = 0.1, α = 0, 1, β = 0.001}, {layer = 1, dropout ratio = 0.3, λ = 0.1, α = 0.01, β = 0.0001}, {layer = 1, dropout ratio = 0.3, λ = 0.1, α = 0.001, β = 0.0001} respectively.

5 CONCLUSION

In this paper, we disclose the existence of superfluous semantic information in the multi-modal features and discover the user-item interaction information degeneration challenging benchmark by the introduced motivating experiments. To tackle the issues, we propose a novel multi-modal recommendation method, namely MSI, which utilizes superfluous semantics discarding module to diminish the task-irrelevant information and interaction preserving module to integrate interaction information into multi-modal features. Extensive experiments demonstrate the superiority of MSI over benchmarks for multi-modal recommendation task.

6 ACKNOWLEDGEMENT

This work is supported by the National Funding Program for Postdoctoral Researchers, Grant No. GZC20232812, Scientific and Technological Innovation 2030 - "New Generation Artificial Intelligence", No. 2022ZD0116407, the CAS Project for Young Scientists in Basic Research, Grant No. YSBR-040, the Youth Innovation Promotion Association CAS, No. 2021106, 2022 Special Research Assistant Grant Project, No. E3YD5901, and the China Postdoctoral Science Foundation, No. 2023M743639.

REFERENCES

[1] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. 2016. Deep Variational Information Bottleneck. *CoRR* abs/1612.00410 (2016). arXiv:1612.00410 http://arxiv.org/abs/1612.00410

[2] Jianxin Chang, Chen Gao, Yu Zheng, Yiqun Hui, Yanan Niu, Yang Song, Depeng Jin, and Yong Li. 2021. Sequential recommendation with graph neural networks. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 378–387.

[3] Lei Chen, Le Wu, Richang Hong, Kun Zhang, and Meng Wang. 2020. Revisiting graph based collaborative filtering: A linear residual graph convolutional network approach. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 27–34.

[4] Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. 2020. Simple and Deep Graph Convolutional Networks. In *Proceedings of the 37th*

International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event (Proceedings of Machine Learning Research, Vol. 119). PMLR, 1725–1735. http://proceedings.mlr.press/v119/chen20v.html

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[6] Xu Chen, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2019. Personalized Fashion Recommendation with Visual Explanations based on Multimodal Attention Network: Towards Visually Explainable Recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer (Eds.). ACM, 765–774. https://doi.org/10.1145/3331184.3331254

[7] Xinlei Chen and Kaiming He. 2021. Exploring Simple Siamese Representation Learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 15750–15758. https://doi.org/10.1109/CVPR46437.2021.01549

[8] Wenqi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. 2019. Graph neural networks for social recommendation. In *The world wide web conference*. 417–426.

[9] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. 2020. Learning Robust Representations via Multi-View Information Bottleneck. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=B1xwcyHFDr

[10] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 249–256.

[11] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. 2020. Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html

[12] 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 international conference on world wide web*. 507–517.

[13] 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.

[14] Ruining He and Julian J. McAuley. 2016. VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, Dale Schuurmans and Michael P. Wellman (Eds.). AAAI Press, 144–150. http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11914

[15] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yong-Dong Zhang, and Meng Wang. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 639–648. https://doi.org/10.1145/3397271.3401063

[16] Chao Huang, Huance Xu, Yong Xu, Peng Dai, Lianghao Xia, Mengyin Lu, Liefeng Bo, Hao Xing, Xiaoping Lai, and Yanfang Ye. 2021. Knowledge-aware coupled graph neural network for social recommendation. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 4115–4122.

[17] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[18] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=SJU4ayYgl

[19] Solomon Kullback. 1935. On the Bernoulli distribution. (1935).

[20] Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics* 22, 1 (1951), 79–86.

[21] Dongha Lee, SeongKu Kang, Hyunjun Ju, Chanyoung Park, and Hwanjo Yu. 2021. Bootstrapping User and Item Representations for One-Class Collaborative Filtering. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 1513–1522. https://doi.org/10.1145/3404835.3462935

[22] J Li et al. [n. d.]. MetaMask: Revisiting Dimensional Confounder for Self-Supervised Learning." arXiv, Sep. 16, 2022. Accessed: Nov. 30, 2022.

[23] Jiangmeng Li, Hang Gao, Wenwen Qiang, and Changwen Zheng. 2023. Information theory-guided heuristic progressive multi-view coding. *Neural Networks* 167 (2023), 415–432.

[24] Jiangmeng Li, Wenwen Qiang, Changwen Zheng, Bing Su, Farid Razzak, Ji-Rong Wen, and Hui Xiong. 2022. Modeling multiple views via implicitly preserving global consistency and local complementarity. *IEEE Transactions on Knowledge and Data Engineering* (2022).

[25] Jiangmeng Li, Wenwen Qiang, Changwen Zheng, Bing Su, and Hui Xiong. 2022. Metaug: Contrastive learning via meta feature augmentation. In *International Conference on Machine Learning*. PMLR, 12964–12978.

[26] Yi Li, Qingmeng Zhu, Hao He, Ziyin Gu, and Changwen Zheng. 2023. MOC: Multimodal Sentiment Analysis via Optimal Transport and Contrastive Interactions. In *International Conference on Neural Information Processing*. Springer, 439–451.

[27] Fan Liu, Zhiyong Cheng, Changchang Sun, Yinglong Wang, Liqiang Nie, and Mohan S. Kankanhalli. 2019. User Diverse Preference Modeling by Multimodal Attentive Metric Learning. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, Laurent Amsaleg, Benoit Huet, Martha A. Larson, Guillaume Gravier, Hayley Hung, Chong-Wah Ngo, and Wei Tsang Ooi (Eds.). ACM, 1526–1534. https://doi.org/10.1145/3343031.3350953

[28] Meng Liu, Hongyang Gao, and Shuiwang Ji. 2020. Towards Deeper Graph Neural Networks. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, Rajesh Gupta, Yan Liu, Jiliang Tang, and B. Aditya Prakash (Eds.). ACM, 338–348. https://doi.org/10.1145/3394486.3403076

[29] Qiang Liu, Shu Wu, and Liang Wang. 2017. DeepStyle: Learning User Preferences for Visual Recommendation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White (Eds.). ACM, 841–844. https://doi.org/10.1145/3077136.3080658

[30] Saeid Motiian, Marco Piccirilli, Donald A. Adjeroh, and Gianfranco Doretto. 2016. Information Bottleneck Learning Using Privileged Information for Visual Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 1496–1505. https://doi.org/10.1109/CVPR.2016.166

[31] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. 188–197.

[32] Wenwen Qiang, Jiangmeng Li, Changwen Zheng, Bing Su, and Hui Xiong. 2022. Interventional contrastive learning with meta semantic regularizer. In *International Conference on Machine Learning*. PMLR, 18018–18030.

[33] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 3980–3990. https://doi.org/10.18653/v1/D19-1410

[34] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, June 18-21, 2009*, Jeff A. Bilmes and Andrew Y. Ng (Eds.). AUAI Press, 452–461. https://www.auai.org/uai2009/papers/UAI2009_0139_48141db02b9f0b02bc7158819ebfa2c7.pdf

[35] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1 (2014), 1929–1958. https://doi.org/10.5555/2627435.2670313

[36] Zhibin Wan, Changqing Zhang, Pengfei Zhu, and Qinghua Hu. 2021. Multi-View Information-Bottleneck Representation Learning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 10085–10092. https://ojs.aaai.org/index.php/AAAI/article/view/17210

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

[38] Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. 2019. Kgat: Knowledge graph attention network for recommendation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 950–958.

[39] Ziyang Wang, Wei Wei, Gao Cong, Xiao-Li Li, Xian-Ling Mao, and Minghui Qiu. 2020. Global context enhanced graph neural networks for session-based recommendation. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*. 169–178.

[40] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2020. Graph-Refined Convolutional Network for Multimedia Recommendation with Implicit Feedback. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann (Eds.). ACM, 3541–3549. https://doi.org/10.1145/3394171.3413556

[41] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal Graph Convolution Network for Personalized Recommendation of Micro-video. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, Laurent Amsaleg, Benoit Huet, Martha A. Larson, Guillaume Gravier, Hayley Hung, Chong-Wah Ngo, and Wei Tsang Ooi (Eds.). ACM, 1437–1445. https://doi.org/10.1145/3343031.3351034

[42] Junchi Yu, Tingyang Xu, Yu Rong, Yatao Bian, Junzhou Huang, and Ran He. 2021. Graph Information Bottleneck for Subgraph Recognition. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=bM4Iqfg8M2k

[43] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. 2021. Mining Latent Structures for Multimedia Recommendation. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo César, Florian Metze, and Balakrishnan Prabhakaran (Eds.). ACM, 3872–3880. https://doi.org/10.1145/3474085.3475259

[44] Xin Zhou. 2022. A Tale of Two Graphs: Freezing and Denoising Graph Structures for Multimodal Recommendation. *CoRR* abs/2211.06924 (2022). https://doi.org/10.48550/arXiv.2211.06924 arXiv:2211.06924

[45] Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. 2023. Bootstrap Latent Representations for Multimodal Recommendation. In *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, Ying Ding, Jie Tang, Juan F. Sequeda, Lora Aroyo, Carlos Castillo, and Geert-Jan Houben (Eds.). ACM, 845–854. https://doi.org/10.1145/3543507.3583251

A APPENDIX

A.1 The derivation of formulation

A.1.1 Derivation of equation 7.

$$
\begin{aligned}
& I_{\theta_1}\left(\mathbf{m}_t ; \mathbf{z}_t \mid \mathbf{m}_v\right) \\
& =\mathbb{E}_{\substack{\boldsymbol{m}_t, \boldsymbol{m}_v \sim p\left(\mathbf{m}_t, \mathbf{m}_v\right) \\ \boldsymbol{z} \sim p_{\theta_1}\left(\mathbf{z}_t \mid \mathbf{m}_t\right)}}\left[\log \frac{p_\theta\left(\mathbf{z}_t=\boldsymbol{z} \mid \mathbf{m}_t=\boldsymbol{m}_t\right)}{p_\theta\left(\mathbf{z}_t=\boldsymbol{z} \mid \mathbf{m}_v=\boldsymbol{m}_v\right)}\right] \\
& =\mathbb{E}_{\substack{\boldsymbol{m}_t, \boldsymbol{m}_v \sim p\left(\mathbf{m}_t, \mathbf{m}_v\right) \\ \boldsymbol{z} \sim p_{\theta_1}\left(\mathbf{z}_t \mid \mathbf{m}_t\right)}}[\log \frac{p_{\theta_1}\left(\mathbf{z}_t=\boldsymbol{z} \mid \mathbf{m}_t=\boldsymbol{m}_t\right)}{p_{\theta_2}\left(\mathbf{z}_v=\boldsymbol{z} \mid \mathbf{m}_v=\boldsymbol{m}_v\right)} \\
& \frac{p_{\theta_2}\left(\mathbf{z}_v=\boldsymbol{z} \mid \mathbf{m}_v=\boldsymbol{m}_v\right)}{p_{\theta_1}\left(\mathbf{z}_t=\boldsymbol{z} \mid \mathbf{m}_v=\boldsymbol{m}_v\right)}] \\
& =D_{\mathrm{KL}}\left(p_{\theta_1}\left(\mathbf{z}_t \mid \mathbf{m}_t\right) \| p_{\theta_2}\left(\mathbf{z}_v \mid \mathbf{m}_v\right)\right)- \\
& D_{\mathrm{KL}}\left(p_{\theta_1}\left(\mathbf{z}_v \mid \mathbf{m}_t\right) \| p_{\theta_2}\left(\mathbf{z}_v \mid \mathbf{m}_v\right)\right) \\
& \leq D_{\mathrm{KL}}\left(p_{\theta_1}\left(\mathbf{z}_t \mid \mathbf{m}_t\right) \| p_{\theta_2}\left(\mathbf{z}_v \mid \mathbf{m}_v\right)\right) .
\end{aligned}
\tag{25}
$$

A.1.2 Derivation of equation 8.

$$
\begin{aligned}
I_{\theta_1}(\mathbf{z}_t ; \mathbf{m}_v) &= I_{\theta_1\theta_2}(\mathbf{z}_t ; \mathbf{z}_v \mathbf{m}_v) - I_{\theta_1\theta_2}(\mathbf{z}_t ; \mathbf{z}_v | \mathbf{m}_v) \\
&= I_{\theta_1\theta_2}(\mathbf{z}_t ; \mathbf{z}_v \mathbf{m}_v) \\
&= I_{\theta_1\theta_2}(\mathbf{z}_t ; \mathbf{z}_v) + I_{\theta_1\theta_2}(\mathbf{z}_t ; \mathbf{m}_v | \mathbf{z}_v) \\
&\geq I_{\theta_1\theta_2}(\mathbf{z}_t ; \mathbf{z}_v)
\end{aligned}
\tag{26}
$$