MENTOR: Multi-level Self-supervised Learning for Multimodal Recommendation

Jinfeng Xu[1], Zheyu Chen[2], Shuo Yang[1], Jinze Li[1], Hewei Wang[3], Edith C. H. Ngai[1]*

[1]The University of Hong Kong,
[2]The Hong Kong Polytechnic University,
[3]Carnegie Mellon University,
{jinfeng, shuo.yang, lijinze-hku}@connect.hku.hk, chngai@eee.hku.hk, zheyu.chen@connect.polyu.hk,
heweiw@andrew.cmu.edu

Abstract

As multimedia information proliferates, multimodal recommendation systems have garnered significant attention. These systems leverage multimodal information to alleviate the data sparsity issue inherent in recommendation systems, thereby enhancing the accuracy of recommendations. Due to the natural semantic disparities among multimodal features, recent research has primarily focused on cross-modal alignment using self-supervised learning to bridge these gaps. However, aligning different modal features might result in the loss of valuable interaction information, distancing them from ID embeddings. It is crucial to recognize that the primary goal of multimodal recommendation is to predict user preferences, not merely to understand multimodal content. To this end, we propose a new Multi-level sElf-supervised learNing for mulTimOdal Recommendation (MENTOR) method, which effectively reduces the gap among modalities while retaining interaction information. Specifically, MENTOR begins by extracting representations from each modality using both heterogeneous user-item and homogeneous item-item graphs. It then employs a multilevel cross-modal alignment task, guided by ID embeddings, to align modalities across multiple levels while retaining historical interaction information. To balance effectiveness and efficiency, we further propose an optional general feature enhancement task that bolsters the general features from both structure and feature perspectives, thus enhancing the robustness of our model.

Code — https://github.com/Jinfeng-Xu/MENTOR

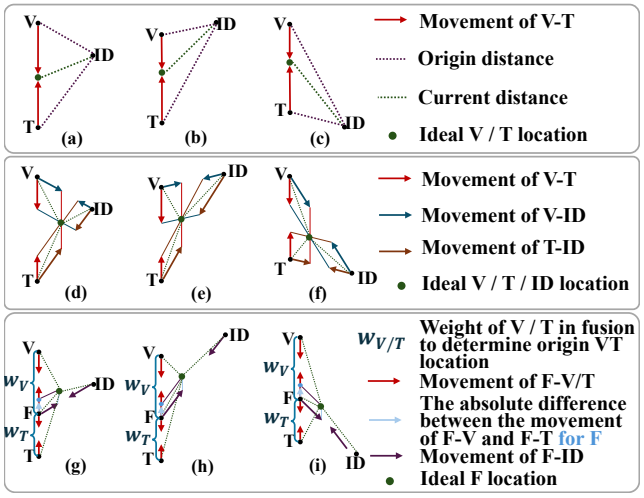


Figure 1: Motivation of multi-level alignment. (a)-(c) describe the standard modality alignment, (d)-(f) describe single modality alignment under ID guidance, and (g)-(i) describe fused (visual and textual) modality alignment under single modality revise and ID guidance. V, T, ID, and F denote visual, textual, ID, and fused modalities, respectively.

Introduction

The rapid growth of the Internet has led to significant information overload. Recommendation systems aim to alleviate information overload by simulating user preferences. However, the performance of traditional recommender systems is limited by the data sparsity problem (Xu et al. 2024c). Recent works on multimodal recommendation mitigate this limitation by utilizing multimedia information. For example, a line of work (He and McAuley 2016; Chen et al. 2017) directly leverages multimodal information as side information to improve the recommendation performance. In recent years, many traditional works (He et al. 2020; Xu et al. 2024b,a) utilize the graph convolutional network (GCN) to capture latent information between users and items. Inspired by these works, MMGCN (Wei et al. 2019) builds the user-item interaction graph for each modality separately and aggregates their prediction as the final rating prediction. DualGNN (Wang et al. 2021) builds an extra homogeneous user-user graph to explore the common user preference pattern. LATTICE (Zhang et al. 2021) and FREEDOM (Zhou and Shen 2023) introduce the item semantic graph to capture the latent semantically correlative signals. More recently, LGMRec (Guo et al. 2024) and DiffMM (Jiang et al. 2024) explore the effectiveness of hyper-graph structure and diffusion model in the multimodal recommendation, respectively.

Besides, recent traditional recommendation methods utilize self-supervised learning (SSL) to reduce label dependence. SelfCF (Zhou et al. 2023b) uses self-supervised signals to enhance recommendation performance without relying on labels. SimGCL (Yu et al. 2022) and XSimGCL (Yu

*Corresponding authors

et al. 2023) effectively combine GCN and self-supervised learning to design graph self-supervised learning. SLMRec (Tao et al. 2022) migrates SSL to the multimodal recommendation field, which effectively enhances the robustness of the model. In the multimodal recommendation field, SSL contributes to robustness enhancement while also effectively aligning features between different modalities. BM3 (Zhou et al. 2023c) and MMSSL (Wei et al. 2023) design SSL tasks to align modalities. However, we highlight that standard modality alignment may result in the loss of valuable interaction information, distancing modalities from the ID embedding. For example, Fig. 1 (b) and (c) show that the standard modality alignment leads both visual and textual modality far from ID embedding.

To tackle this problem and better align modalities, we propose a novel Multi-level sElf-supervised learNing for mulTimOdal Recommendation, named MENTOR. It introduces a novel multilevel self-supervised learning task that enhances model robustness and aligns features across different modalities without sacrificing historical interaction information. Specifically, MENTOR initially utilizes both heterogeneous user-item and homogeneous item-item graphs to extract representations for each modality. Furthermore, it establishes a novel multilevel cross-modal alignment task that effectively aligns different modalities under the direct and indirect guidance of the ID embedding, thereby preserving historical interaction information. As depicted in Fig. 1 (d)-(f), each modality is aligned with the ID embedding, which in turn indirectly aligns the fused modality with the ID embedding, effectively preventing the loss of historical interaction information that often accompanies modality alignment. Nevertheless, various modalities contribute differently to the fusion process. To optimize the alignment of the fused modality and balance the weights of different modalities, we directly align the fused modality with the ID embedding and each individual modality, as illustrated in Fig. 1 (g)-(i). To balance effectiveness and efficiency, we provide an optional general feature enhancement task to enhance the general features and robustness of our model from both feature masking and graph perturbation perspectives.

In a nutshell, we summarize our contributions as follows:

- We propose a novel framework MENTOR for the multimodal recommendation, which alleviates both data sparsity and label sparsity problems.
- We propose a novel multilevel cross-modal alignment task, which effectively aligns different modality features without losing historical interaction information.
- We develop an optional general feature enhancement task, which enhances the general features on both feature masking and graph perturbation perspectives.
- We perform comprehensive experiments on three public datasets in Amazon to validate the effectiveness of our MENTOR on both overall and component levels.

Related Work

Multimodal Recommendation

Many recent works incorporate multimodal information to alleviate the data sparsity problem. VBPR (He and McAuley 2016) is the first attempt to utilize visual content to alleviate the data sparsity problem based on matrix factorization (Rendle et al. 2012). Moreover, many works (Chen et al. 2019; Liu et al. 2019; Chen et al. 2017) enhance the representation of items with both visual and textual modalities to mitigate the data sparsity problem further. Inspired by the traditional recommendation system, MMGCN (Wei et al. 2019) adopts GCN to construct a bipartite graph to extract the latent information in user-item interactions. GRCN (Wei et al. 2020) prunes the false-positive edges based on MMGCN to reduce the noise in the user-item bipartite graph. To explicitly mine the common preferences between users, DualGNN (Wang et al. 2021) constructs an extra user co-occurrence graph. LATTICE (Zhang et al. 2021) introduces an item semantic graph to capture the latent correlative signals between items. FREEDOM (Zhou and Shen 2023) freezing the item semantic graph based on LATTICE. LGMRec (Guo et al. 2024) and DiffMM (Jiang et al. 2024) explore the effectiveness of hyper-graph structure and diffusion model in the multimodal recommendation, respectively.

Self-supervised Learning on Recommendation

In the traditional recommendation field, SSL effectively improves the robustness of the model and mitigates the label dependency. SelfCF (Zhou et al. 2023b) and BUIR (Lee et al. 2021) generate different views to learn the representation of users and items from positive interaction, respectively. MixGCF (Huang et al. 2021) designs a general negative sampling plugin that can be directly used to train GNN-based recommender systems. MHCN (Yu et al. 2021) and SGL (Wu et al. 2021) propose to generate SSL signals via contrasting positive node pairs based on various augmentation operations. SimGCL (Yu et al. 2022) and XSimGCL (Yu et al. 2023) discard the graph augmentations and instead add uniform noises to the embedding space for creating contrastive views. In the multimodal recommendation field, SLMRec (Tao et al. 2022) proposes two SSL tasks to enhance the robustness of the model, including noise perturbation over features and multimodal pattern uncovering augmentation. It is worth noting that SSL can also be used to align features from different modalities. BM3 (Zhou et al. 2023c) simplifies the SSL task based on SLMRec, while MMSSL (Wei et al. 2023) designs a cross-modal contrastive learning task to preserve the inter-modal semantic commonality and user preference diversity jointly. However, these methods inevitably generate a large amount of noise along with modal alignment, which leads to significant attenuation of historical interaction information. In this work, we propose a multilevel cross-modal alignment task, which can effectively align different modality features while retaining historical interaction information.

Methodology

In this section, we present the MENTOR architecture and describe each component in our proposed model. Fig. 2 shows the overall architecture of MENTOR.

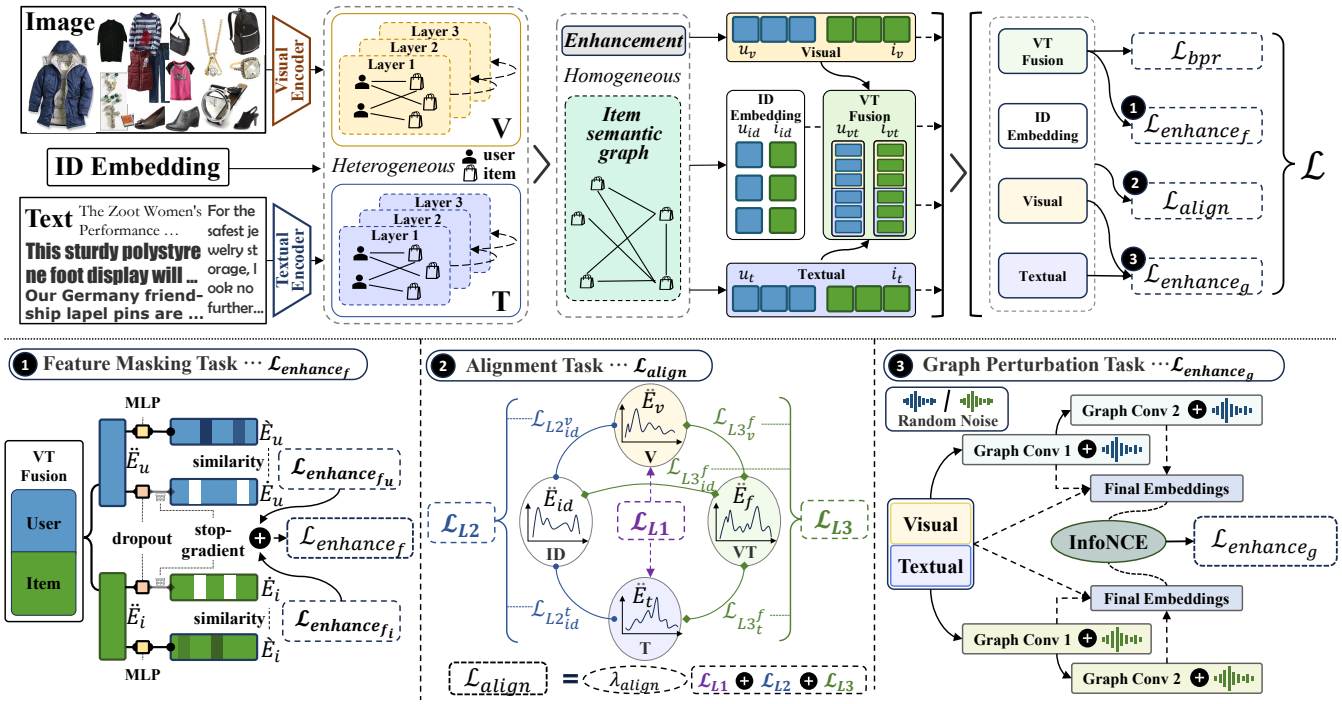


Figure 2: The architecture of our MENTOR. We first utilize both heterogeneous user-item and homogeneous item-item graphs to learn the representation of each modality. Then, we fuse visual and textual modalities. Moreover, we utilize an alignment self-supervised task (2) to align each modality without loss of interaction information. Besides, we provide an optional self-supervised task to enhance the general features on both the feature masking task (1) and the graph perturbation task (3).

Problem Definition

Let $\mathcal{U} = \{u\}$ denote the user set and $\mathcal{I} = \{i\}$ denote the item set. Then, we denote the features of each modality and the input ID embedding as $E_m = \{E_{u_m} \| E_{i_m}\} \in \mathbb{R}^{d_m \times (|\mathcal{U}|+|\mathcal{I}|)}$, where $m \in \mathcal{M}$ is the modality, $\mathcal{M}$ is the set of modalities, d_m is the dimension of the features, and $\|$ denotes concatenation operation. We only consider ID, visual, and textual modalities denoted by $\mathcal{M} = \{id, v, t\}$. However, our model can involve more modalities than these three modalities.

Multimodal Information Encoder

Some previous works (Zhou and Shen 2023; Zhang et al. 2021) find that both the user-item heterogeneous graph and the item-item homogeneous graph can significantly improve the performance of multimodal recommendations. Inspired by them, we propose a multimodal information encoder component to extract the representation of each modality.

User-Item Graph To capture high-order modality-specific features, we construct three user-item graphs $\mathcal{G} = \{\mathcal{G}_m | \mathcal{G}_{id}, \mathcal{G}_v, \mathcal{G}_t\}$. Each graph $\mathcal{G}_m$ maintains the same graph structure and only retains the node features associated with each modality. Formally, the user and item representations at l-th graph convolution layer can be formulated as:

$$E_{u_m}^{(l)} = \sum_{i \in \mathcal{N}_u} \frac{1}{\sqrt{|\mathcal{N}_u|}\sqrt{|\mathcal{N}_i|}} E_{i_m}^{(l-1)}, \tag{1}$$

$$E_{i_m}^{(l)} = \sum_{u \in \mathcal{N}_i} \frac{1}{\sqrt{|\mathcal{N}_u|}\sqrt{|\mathcal{N}_i|}} E_{u_m}^{(l-1)}, \tag{2}$$

where $\mathcal{N}_u$ and $\mathcal{N}_i$ denote the one-hop neighbors of u and i in $\mathcal{G}$, respectively. The final embedding for each modality is calculated by element-wise summation. Formally,

$$E_m = \{E_{u_m} \| E_{i_m}\}, \quad \bar{E}_m = \sum_{l=0}^{L} E_m^{(l)}, \tag{3}$$

where L is the number of user-item graph layers.

Item-Item Graph To extract significant semantic relations between items, we use KNN to establish the item-item graph based on the item features for each modality m. Particularly, we calculate the similarity score $S_{i,i'}^m$ between item pair $(i, i') \in \mathcal{I}$ by the cosine similarity on their modality original features f_i^m and $f_{i'}^m$.

$$S_{i,i'}^m = \frac{(f_i^m)^\top f_{i'}^m}{\|f_i^m\| \, \|f_{i'}^m\|}. \tag{4}$$

We only retain the top-k neighbors:

$$S_{i,i'}^m = \begin{cases} 1 & \text{if } S_{i,i'}^m \in \text{top-}k\,(S_i^m) \\ 0 & \text{otherwise} \end{cases}. \tag{5}$$

Then, we aggregate multilayer neighbors:

$$A_m^{(l)} = \sum_{i' \in \mathcal{N}^i} S_{i,i'}^m A_{i'_m}^{(l-1)}, \tag{6}$$

where $\mathcal{N}^i$ denotes the neighbors of item i in item-item graph. $A_{i'_m}$ is the embedding of item i' in modality m. Inspired by (Zhou and Shen 2023), we freeze each item-item graph after initialization to remove the computation costs of the item-item graph during the training phase.

Multimodel Fusion

To combine multiple modalities to mine user preferences jointly, we enhance the final embedding $\bar{E}_m$ of the user-item graph based on the final embedding $A_m^{(l)}$ of the item-item graph for each modality. The enhanced embedding $\ddot{E}_m$ can be calculated as $\bar{E}_{i_m} = \bar{E}_{i_m} + A_m^{(l)}$ and $\ddot{E}_m = \{\bar{E}_{u_m} \| \bar{E}_{i_m}\}$, where $\|$ denotes the concatenation operation.

Then, we fuse visual and textual modalities:

$$\ddot{E}_f = \{\alpha \times \ddot{E}_v \| (1-\alpha) \times \ddot{E}_t\}, \tag{7}$$

where the attention weight α is a trainable parameter which we initialize to be 0.5, $\ddot{E}_v$ and $\ddot{E}_t$ are the representation of visual and textual modalities respectively.

Multilevel Cross-Modal Alignment

The feature distributions of different modalities are extremely different. Existing methods (Wei et al. 2023; Zhou et al. 2023c) directly align different modal features, which may lose useful interaction information. We point out that the multimodal recommendation aims to predict user preferences rather than to comprehend multimodal content. Therefore, we propose a multilevel cross-modal alignment component to align modalities from the data distribution perspective using self-supervised learning. Specifically, our multilevel cross-modal alignment component encompasses three levels: the standard modality alignment level, the ID indirect guidance level, and the ID direct guidance level. The standard modality alignment level aims to reduce the disparity between visual and textual modalities. The ID indirect guidance level capitalizes on historical interaction information embedded in the ID modality to augment similar features in both visual and textual modalities. The ID direct guidance level aligns the visual and textual modalities with the fused modality to equilibrate modality weights, and directly aligns the fused modality with the ID modality. We will introduce all alignment levels, respectively.

Distribution Representation To achieve modal alignment without losing valuable modality features, we coarse-grained align the mean and covariance for each modality. Specifically, we directly calculate the mean and covariance for each modality as follows:

$$\begin{aligned} \{\mu_f, \mu_{id}, \mu_v, \mu_t\} &= \mathrm{Mean}(\ddot{E}_f, \ddot{E}_{id}, \ddot{E}_v, \ddot{E}_t), \\ \{\sigma_f, \sigma_{id}, \sigma_v, \sigma_t\} &= \mathrm{Var}(\ddot{E}_f, \ddot{E}_{id}, \ddot{E}_v, \ddot{E}_t), \end{aligned} \tag{8}$$

where $\mathrm{Mean}(\cdot)$ and $\mathrm{Var}(\cdot)$ calculate the mean and covariance for each node: $\mathbb{R}^{d_m \times (|\mathcal{U}|+|\mathcal{I}|)} \to \mathbb{R}^{(|\mathcal{U}|+|\mathcal{I}|)}$.

The Standard Modality Alignment In the standard modality alignment level, as Fig. 1 (a)-(c) show, we directly align the visual and textual modality to minimize the semantic gap. The loss is defined as:

$$\mathcal{L}_{L1} = |\mu_v - \mu_t| + |\sigma_v - \sigma_t|. \tag{9}$$

ID Indirect Guidance In ID indirect guidance level, as Fig. 1 (d)-(f) show, we further align the visual modality and textual modality with ID modality, respectively. Then we fuse them together, the loss is defined as:

$$\mathcal{L}_{L2_{v-id}} = |\mu_{id} - \mu_v| + |\sigma_{id} - \sigma_v|, \tag{10}$$

$$\mathcal{L}_{L2_{t-id}} = |\mu_{id} - \mu_t| + |\sigma_{id} - \sigma_t|, \tag{11}$$

$$\mathcal{L}_{L2} = \mathcal{L}_{L2_{v-id}} + \mathcal{L}_{L2_{t-id}}. \tag{12}$$

ID Direct Alignment In ID direct guidance level, as Fig. 1 (g)-(i) show, we align the visual modality and textual modality with fused modality to balance the modality weights, respectively. Moreover, we directly align the fused modality with ID modality. The loss is defined as:

$$\mathcal{L}_{L3_{f-v}} = |\mu_f - \mu_v| + |\sigma_f - \sigma_v|, \tag{13}$$

$$\mathcal{L}_{L3_{f-t}} = |\mu_f - \mu_t| + |\sigma_f - \sigma_t|, \tag{14}$$

$$\mathcal{L}_{L3_{f-id}} = |\mu_f - \mu_{id}| + |\sigma_f - \sigma_{id}|, \tag{15}$$

$$\mathcal{L}_{L3} = \mathcal{L}_{L3_{f-v}} + \mathcal{L}_{L3_{f-t}} + \mathcal{L}_{L3_{f-id}}. \tag{16}$$

We finally get the overall multilevel cross-modal alignment loss $\mathcal{L}_{align} = \lambda_{align}(\mathcal{L}_{L1} + \mathcal{L}_{L2} + \mathcal{L}_{L3})$, where λ_{align} is the balancing hyper-parameter.

General Feature Enhancement

We further propose an optional general feature enhancement component to enhance the general feature from both the graph and feature perspectives to improve the robustness of our model. This component includes two tasks: feature masking and graph perturbation. The feature masking task generates different views from feature perspectives as shown in task (1) of Fig. 2. The graph perturbation task generates different views from the graph perspective using random noise as shown in task (3) of Fig. 2.

Feature Masking We first split $\ddot{E}$ as two sides $\ddot{E}_u$ and $\ddot{E}_i$. Then, we utilize dropout mechanism to mask out a subset of these embeddings to generate one contrastive view by $\dot{E}_u = \ddot{E}_u \cdot \mathrm{Bernoulli(p)}$ and $\dot{E}_i = \ddot{E}_i \cdot \mathrm{Bernoulli(p)}$. Then, we place stop-gradient on $\dot{E}_i$ and $\dot{E}_u$ and transfer them through MLP to construct another contrastive view by $\grave{E}_u = \ddot{E}_u W + b$ and $\grave{E}_i = \ddot{E}_i W + b$, where $W \in \mathbb{R}^{d_m \times d_m}$, $b \in \mathbb{R}^{d_m}$ denote the linear transformation matrix and bias. Formally:

$$\mathcal{L}_{enhance_f} = (1 - \mathrm{Sim}(\dot{E}_u, \grave{E}_u)) + (1 - \mathrm{Sim}(\dot{E}_i, \grave{E}_i)), \tag{17}$$

where $\mathrm{Sim}(\cdot)$ is the cosine similarity.

Graph Perturbation We follow the most commonly used augmentation methods based on the dropout mechanism in graphs (Wu et al. 2021; You et al. 2020) to construct contrastive views of structural perturbations for both visual and textual modalities. We propose a perturbed user-item graph:

$$\dagger E_{u_m}^{(l)} = \sum_{i \in \mathcal{N}_u} \frac{1}{\sqrt{|\mathcal{N}_u|}\sqrt{|\mathcal{N}_i|}} E_{i_m}^{(l-1)} + \Delta^{(l)}, \tag{18}$$

$$\dagger E_{i_m}^{(l)} = \sum_{u \in \mathcal{N}_i} \frac{1}{\sqrt{|\mathcal{N}_u|}\sqrt{|\mathcal{N}_i|}} E_{u_m}^{(l-1)} + \Delta^{(l)}, \tag{19}$$

where most parameters are the same as Eq. 1-2, and $\Delta^{(l)} \in \mathbb{R}^{d_m} \sim U(0,1)$ is a random noise vector.

The final perturbed embedding for each modality is calculated as:

$$\dagger E_m = \{\dagger E_{u_m} \| \dagger E_{i_m}\}, \quad \dagger \bar{E}_m = \sum_{l=0}^{L} \dagger E_m^{(l)}. \tag{20}$$

For both visual and textual modalities, we generate two contrastive views $\dagger \bar{E}_m^1$ and $\dagger \bar{E}_m^2$ for each modality and adopt InfoNCE (Oord, Li, and Vinyals 2018). Formally:

$$\begin{aligned}\mathcal{L}_{enhance_{g_m}} = &\sum_{u \in \mathcal{U}} -\log \frac{\exp\left(e_{u,m}^1 \cdot e_{u,m}^2/\tau\right)}{\sum_{v \in \mathcal{U}} \exp\left(e_{u,m}^1 \cdot e_{v,m}^2/\tau\right)} \\ &+ \sum_{i \in I} -\log \frac{\exp\left(e_{i,m}^1 \cdot e_{i,m}^2/\tau\right)}{\sum_{j \in I} \exp\left(e_{i,m}^1 \cdot e_{j,m}^2/\tau\right)},\end{aligned} \tag{21}$$

where $e_{u,m}^1$ and $e_{u/v,m}^2$ are the modality m features of user u/v in contrastive views $\dagger \bar{E}_m^1$ and $\dagger \bar{E}_m^2$. Besides, $e_{i,m}^1$ and $e_{i/j,m}^2$ are the modality m features of item i/j in contrastive views $\dagger \bar{E}_m^1$ and $\dagger \bar{E}_m^2$. τ is the temperature hyper-parameter. The total graph perturbation loss is calculated as $\mathcal{L}_{enhance_g} = \mathcal{L}_{enhance_{g_v}} + \mathcal{L}_{enhance_{g_t}}$. Finally, the overall general feature enhancement loss is:

$$\mathcal{L}_{enhance} = \lambda_g \mathcal{L}_{enhance_g} + \lambda_f \mathcal{L}_{enhance_f}, \tag{22}$$

where λ_g and λ_f are the balancing hyper-parameters.

Optimization

We adopt the Bayesian Personalized Ranking (BPR) loss (Rendle et al. 2012) as the basic optimization function. Essentially, BPR aims to widen the predicted preference margin between the positive and negative items for each triplet $(u, p, n) \in \mathcal{D}$, where $\mathcal{D}$ denotes the training set. The positive item p refers to the one with which the user u has interacted, while the negative item n has been randomly chosen from the set of items that the user u has not interacted with. The BPR function is defined as follows:

$$\mathcal{L}_{bpr} = \sum_{(u,p,n) \in \mathcal{D}} -\log(\sigma(y_{u,p} - y_{u,n})), \tag{23}$$

where $y_{u,p}$ and $y_{u,n}$ are the ratings of user u to the positive item p and negative item n, calculated by $\ddot{E}_u^{\mathrm{T}} \ddot{E}_p$ and $\ddot{E}_u^{\mathrm{T}} \ddot{E}_n$, respectively. σ is the Sigmoid function. The final loss is:

$$\mathcal{L} = \mathcal{L}_{bpr} + \mathcal{L}_{align} + \mathcal{L}_{enhance} + \lambda_E(\|E_v\|_2^2 + \|E_t\|_2^2), \tag{24}$$

where E_v and E_t are the model parameters. λ is a hyper-parameter to control the effect of the L_2 regularization. To trade off efficiency and effectiveness, $\mathcal{L}_{enhance}$ can be entirely or partly retained.

Experiment

In this section, we conduct comprehensive experiments to evaluate the performance of our MENTOR model on three widely used real-world datasets. The following five questions can be well answered through experiment results: **RQ1**: How effective is our MENTOR compared with the state-of-the-art traditional recommendation methods and multimedia recommendation methods? **RQ2**: How do the key components of our MENTOR impact its performance? **RQ3**: Can the multilevel cross-modal alignment component effectively align different modalities? **RQ4**: How efficient is our MENTOR compared with other methods? **RQ5**: How sensitive is MENTOR with different hyper-parameters?

Dataset	# Users	# Items	# Interaction	Sparsity
Baby	19445	7050	160792	99.88%
Sports	35598	18357	296337	99.95%
Clothing	39387	23033	278677	99.97%

Table 1: Statistics of the experimental datasets.

Experimental Settings

Datasets To evaluate our proposed MENTOR in the top-N item recommendation task, we conduct extensive experiments on three widely used Amazon datasets (McAuley et al. 2015): Baby, Sports, and Clothing. These datasets provide both product descriptions and images simultaneously. Following the previous works (He and McAuley 2016; Wei et al. 2019), the raw data of each dataset are pre-processed with a 5-core setting on both items and users. Besides, we use the pre-extracted 4096-dimensional visual features and extract 384-dimensional textual features using a pre-trained sentence transformer (Zhou 2023). The statistics of these datasets are presented in Table 1.

Baselines To demonstrate the effectiveness of our proposed MENTOR, we compare it with the following state-of-the-art recommendation methods, which can be divided into two groups: traditional recommendation methods (**MF-BPR** (Rendle et al. 2012), **LightGCN** (He et al. 2020), and **LayerGCN** (Zhou et al. 2023a)) and multimedia recommendation methods (**VBPR** (He and McAuley 2016), **MMGCN** (Wei et al. 2019), **DualGNN** (Wang et al. 2021), **LATTICE** (Zhang et al. 2021), **FREEDOM** (Zhou and Shen 2023), **SLMRec** (Tao et al. 2022), **BM3** (Zhou et al. 2023c), **MMSSL** (Wei et al. 2023), **LGMRec** (Guo et al. 2024), and **DiffMM** (Jiang et al. 2024)).

Evaluation Protocols To evaluate the performance fairly, we adopt two widely used metrics: Recall@K (R@K) and NDCG@K (N@K). We report the average metrics of all users in the test dataset under both K = 10 and K = 20. We follow the popular setting (Zhou and Shen 2023) with a random data splitting 8:1:1 for training, validation, and testing.

Implementation Details We implement MENTOR and all baselines with MMRec (Zhou 2023). For the general settings, we initialized the embedding with Xavier initialization (Glorot and Bengio 2010) of dimension 64. Besides, we optimize all models with Adam optimizer (Kingma and Ba 2014). To achieve a fair evaluation, we perform a complete grid search for each baseline method following its published

Datasets	Baby				Sports				Clothing			
Model	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20
MF-BPR	0.0357	0.0575	0.0192	0.0249	0.0432	0.0653	0.0241	0.0298	0.0187	0.0279	0.0103	0.0126
LightGCN	0.0479	0.0754	0.0257	0.0328	0.0569	0.0864	0.0311	0.0387	0.0340	0.0526	0.0188	0.0236
LayerGCN	0.0529	0.0820	0.0281	0.0355	0.0594	0.0916	0.0323	0.0406	0.0371	0.0566	0.0200	0.0247
VBPR	0.0423	0.0663	0.0223	0.0284	0.0558	0.0856	0.0307	0.0384	0.0281	0.0415	0.0158	0.0192
MMGCN	0.0378	0.0615	0.0200	0.0261	0.0370	0.0605	0.0193	0.0254	0.0218	0.0345	0.0110	0.0142
DualGNN	0.0448	0.0716	0.0240	0.0309	0.0568	0.0859	0.0310	0.0385	0.0454	0.0683	0.0241	0.0299
LATTICE	0.0547	0.0850	0.0292	0.0370	0.0620	0.0953	0.0335	0.0421	0.0492	0.0733	0.0268	0.0330
FREEDOM	0.0627	0.0992	0.0330	0.0424	0.0717	0.1089	0.0385	0.0481	0.0629	0.0941	0.0341	0.0420
SLMRec	0.0529	0.0775	0.0290	0.0353	0.0663	0.0990	0.0365	0.0450	0.0452	0.0675	0.0247	0.0303
BM3	0.0564	0.0883	0.0301	0.0383	0.0656	0.0980	0.0355	0.0438	0.0422	0.0621	0.0231	0.0281
MMSSL	0.0613	0.0971	0.0326	0.0420	0.0673	0.1013	0.0380	0.0474	0.0531	0.0797	0.0291	0.0359
LGMRec	0.0639	0.0989	0.0337	0.0430	0.0719	0.1068	0.0387	0.0477	0.0555	0.0828	0.0302	0.0371
DiffMM	0.0623	0.0975	0.0328	0.0411	0.0671	0.1017	0.0377	0.0458	0.0522	0.0791	0.0288	0.0354
MENTOR$_{fg}$	0.0649	0.1011	0.0350	0.0440	0.0727	0.1094	0.0390	0.0481	0.0636	0.0949	0.0343	0.0428
MENTOR$_{f}$	0.0663	0.1037	0.0358	0.0449	0.0749	0.1126	0.0404	0.0505	0.0661	0.0981	0.0359	**0.0443***
MENTOR$_{g}$	0.0666	0.1034	0.0355	**0.0454***	0.0750	0.1129	0.0400	0.0507	0.0653	0.0977	0.0351	0.0438
MENTOR	**0.0678***	**0.1048***	**0.0362***	0.0450	**0.0763***	**0.1139***	**0.0409***	**0.0511***	**0.0668***	**0.0989***	**0.0360***	0.0441
Improv.	6.10%	5.65%	7.42%	5.58%	6.12%	4.59%	5.68%	6.24%	6.20%	5.10%	5.57%	5.48%

Table 2: Performance comparison of baselines and MENTOR in terms of Recall@K (R@K), and NDCG@K (N@K). The superscript * indicates the improvement is statistically significant where the p-value is less than 0.05.

paper to find the optimal setting. For our MENTOR, we perform a grid search on the dropout ratio in $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$, balancing hyper-parameter λ_f in $\{0.5, 1, 1.5, 2, 2.5\}$, balancing hyper-parameter λ_g in {1e-2, 1e-3, 1e-4}, temperature hyper-parameter τ in $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$, and balancing hyper-parameter λ_{align} in $\{0.1, 0.2, 0.3\}$. We fix the learning rate with 1e-4, and the number of layers in the heterogenous graph with $L = 2$. The k of top-k in the item-item graph is set as 40. For convergence consideration, the early stopping is fixed at 20. Following (Zhou 2023), we update the best record by utilizing Recall@20 on the validation dataset as the indicator.

Effectiveness of MENTOR (RQ1)

Table 2 summarizes the performance of our proposed MENTOR and other baseline methods on three datasets. From the table, we find the following observations:

MENTOR achieves better performance than both traditional and multimodal recommendation methods. Specifically, our MENTOR improves the best baseline by 5.65%, 4.59%, and 5.10% in terms of Recall@20 on Baby, Sports, and Clothing, respectively. The results verify the effectiveness of our MENTOR. We attribute the enhancement that our multi-level self-supervised alignment task effectively aligns all modalities while retaining interaction information.

Leveraging multimodal information can effectively improve recommendation accuracy. Specifically, the recent multimodal recommendation methods generally outperform traditional recommendation methods in all scenarios. This demonstrates that multimodal recommendation methods can effectively mitigate the data sparsity problem by leveraging multiple modalities to jointly mine user preferences.

Our multi-level alignment task achieves significant improvement over other multimodal alignment tasks. Compared with other models (BM3 and MMSSL) that incorporate multimodal alignment tasks, MENTOR achieves around 10% improvement for all datasets without complexity techniques (e.g., DiffMM leverage diffusion modal and LGMRec construct hyper-graph).

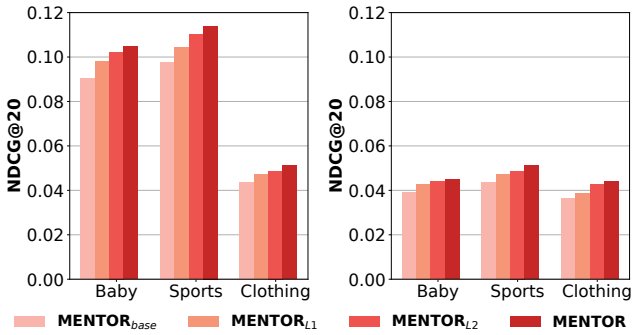


Figure 3: Effect of multilevel cross-modal alignment.

Ablation Study (RQ2)

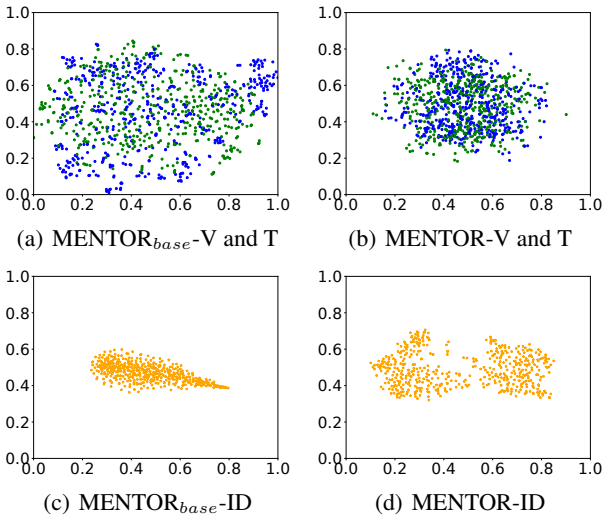
In this section, we conduct exhaustive experiments to evaluate the effectiveness of different components of MENTOR.

Effect of Multilevel Cross-Modal Alignment To investigate the effects of the multilevel cross-modal alignment component, we design the following variants of MENTOR.

- MENTOR$_{base}$, which removes the entire multilevel cross-modal alignment.
- MENTOR$_{L1}$, which retains the standard modality alignment level in the multilevel cross-modal alignment.

Dataset	Metrics	MMGCN	LATTICE	FREEDOM	LGMRec	MENTOR$_{fg}$	MENTOR
Baby	Memory (GB)	2.69	4.53	2.13	2.41	2.30	7.12
	Time (s/epoch)	4.03	3.13	2.45	4.05	3.22	6.07
Sports	Memory (GB)	3.91	19.93	3.34	3.67	3.55	8.44
	Time (s/epoch)	14.47	10.99	5.49	8.24	7.26	9.01
Clothing	Memory (GB)	4.24	28.22	4.15	4.81	4.48	12.99
	Time (s/epoch)	19.83	18.78	6.02	9.48	8.08	11.39

Table 3: Efficiency Analysis.



(a) MENTOR$_{base}$-V and T (b) MENTOR-V and T

(c) MENTOR$_{base}$-ID (d) MENTOR-ID

Figure 4: Green, blue, and orange represent the visual, textual, and ID modalities, respectively.

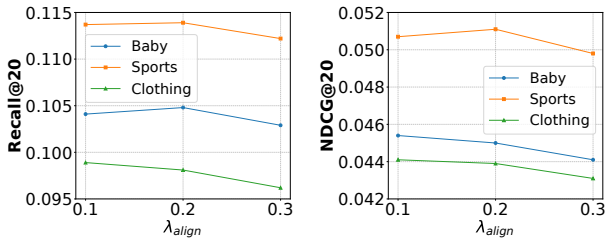
- MENTOR$_{L2}$, which retains the standard modality alignment and ID indirect guidance levels in the multilevel cross-modal alignment.

Fig. 3 shows that each level of our multilevel cross-modal alignment component leads to an obvious improvement for all datasets, and their effects can be superimposed on each other. To verify the effectiveness of ID indirect and direct guidance, we further verify the effectiveness of our multilevel cross-modal alignment by visualization analysis.

Effect of General Feature Enhancement To analyze the effectiveness of the general feature enhancement component in MENTOR, the following variants are constructed:

- MENTOR$_{fg}$: We remove the whole general feature enhancement component.
- MENTOR$_{f}$: We remove the feature masking task in the general feature enhancement component.
- MENTOR$_{g}$: We remove the graph perturbation task in the general feature enhancement component.

As illustrated in Table 2, we point out that all variants still outperform all baselines. Moreover, further performance gains can be achieved by using this component, so there is flexibility to trade off based on effectiveness and efficiency.



(a) Recall@20 (b) NDCG@20

Figure 5: Effect of the balancing hyper-parameter λ_{align}.

Visualization Analysis (RQ3)

To further verify the effectiveness of our multilevel cross-modal alignment, we visualize the distribution of the representation. Fig. 4 shows the impact of our multilevel cross-modal alignment component in the 2-dimension perspective. The two models of our comparison are MENTOR$_{base}$ and MENTOR. Specifically, we randomly sample 500 items from the Baby dataset. Then, we use t-SNE (Van der Maaten and Hinton 2008) to map embedding to the 2-dimension space. We observe that the textual modality distribution of MENTOR$_{base}$ is relatively more discrete than the visual modality distribution of MENTOR$_{base}$. Besides, the textual and visual modality distributions of MENTOR are closer to each other compared with MENTOR$_{base}$. Moreover, the distributions of visual and textual modalities of MENTOR are more similar to ID modality. Therefore, we attribute this effective alignment to the guidance of ID modality, which retains significant interaction information. We measure the average distance between the visual and textual modalities, which are 0.44 for MENTOR$_{base}$ and 0.18 for MENTOR.

Efficiency Study (RQ4)

In this section, we compare the efficiency of MENTOR with baselines. Specific statistical memory and time are shown in Table 3. The light version MENTOR$_{fg}$ achieves outstanding performance with low computation costs. The entire MENTOR provides superior performance without seriously compromising training speed.

Hyper-parameter Analysis (RQ5)

The Balancing Hyper-parameter λ_{align} Fig. 5(a) and Fig. 5(b) show the performance trends of MENTOR with different settings of λ_{align}. We observe that the optimal

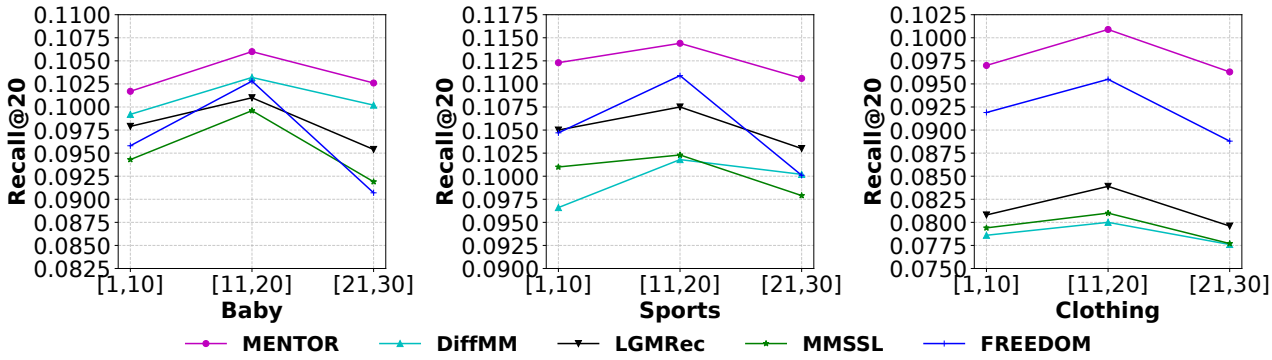
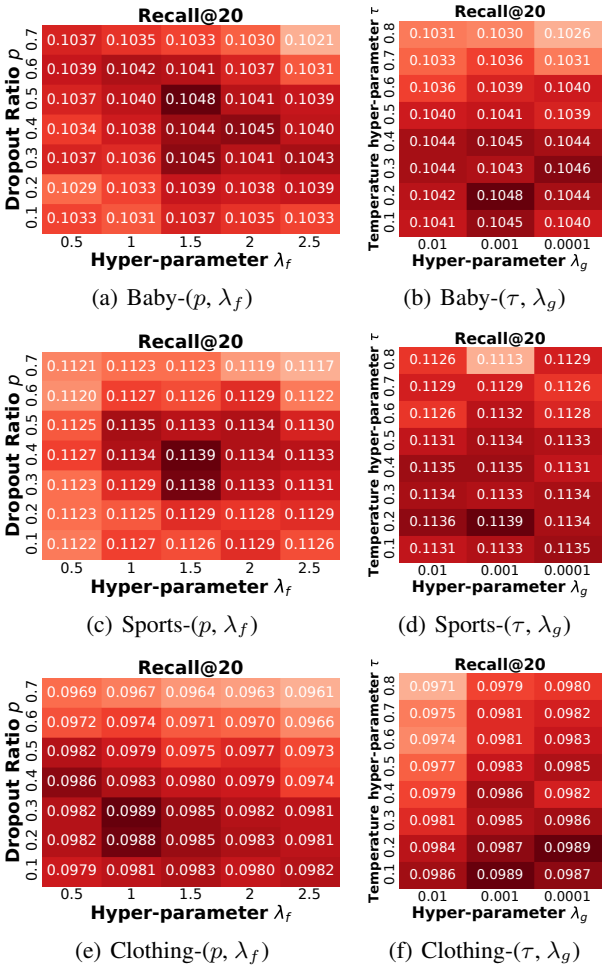


Figure 6: Sparsity degree analysis on three datasets.



(a) Baby-(p, λ_f) (b) Baby-(τ, λ_g)

(c) Sports-(p, λ_f) (d) Sports-(τ, λ_g)

(e) Clothing-(p, λ_f) (f) Clothing-(τ, λ_g)

Figure 7: Performance of MENTOR with respect to different hyper-parameter pairs (p,λ_f) and (τ,λ_g).

λ_{align} on Baby and Sports datasets is 0.1, while the optimal λ_{align} on Clothing datasets is 0.2.

The Pair of Hyper-parameters p and λ_f As Fig. 7(a)-7(e) shown, we find that hyper-parameters λ_f and p influence each other, so we need to select them in pairs. For Baby and Sports datasets, the optimal values of (λ_f, p) pair are (1.5, 0.5) and (1.5, 0.4). Moreover, for Clothing dataset, the optimal value of (λ_f, p) pair is (1, 0.3). A possible reason for these results is that the Clothing dataset is more sparse than the other two datasets, which makes it more sensitive to dropout operations.

The Pair of Hyper-parameters λ_g and τ The balancing hyper-parameter λ_g and the temperature hyper-parameter τ jointly control the feature masking task in the general feature enhancement component. Fig. 7(b)-7(f) shows that the best performances are achieved with (λ_g, τ) = (1e-3, 0.2) on Baby and Sports datasets. For Clothing dataset, the optimal value of (λ_g, τ) is (1e-3, 0.1) or (1e-4, 0.2).

Different Data Sparsity

We further test the effectiveness of MENTOR with different data sparsity settings on all three datasets. We choose the four latest and best-performed models as baselines, including MMSSL, FREEDOM, LGMRec, and DiffMM. We split each dataset into three sub-datasets based on users' interacted item numbers in the training set. Fig. 6 shows that MENTOR provides consistently more powerful and robust performance than all baselines on all datasets with different degrees of sparsity. We attribute this outstanding and stable performance to our well-designed self-supervised tasks.

Conclusion

In this paper, we propose a novel self-supervised learning framework in multimodal recommendation, named MENTOR, for multimodal recommendation. After learning the representation of all modalities by our powerful multimodal information encoder. MENTOR introduces a tailored multilevel cross-modal alignment task to align different modalities on data distribution while retaining historical interaction information. Moreover, MENTOR devises an optional general feature enhancement contrastive learning task from both the feature and graph perspectives to improve the model robustness. Our extensive experimental results on several widely used datasets show that MENTOR achieves significant accuracy improvement compared with the state-of-the-art multimodal recommendation methods.

Acknowledgements

This work was supported by the Hong Kong UGC General Research Fund no. 17203320 and 17209822, and the project grants from the HKU-SCF FinTech Academy.

References

Chen, J.; Zhang, H.; He, X.; Nie, L.; Liu, W.; and Chua, T.-S. 2017. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*, 335–344.

Chen, X.; Chen, H.; Xu, H.; Zhang, Y.; Cao, Y.; Qin, Z.; and Zha, H. 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*, 765–774.

Glorot, X.; and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 249–256. JMLR Workshop and Conference Proceedings.

Guo, Z.; Li, J.; Li, G.; Wang, C.; Shi, S.; and Ruan, B. 2024. LGMRec: Local and Global Graph Learning for Multimodal Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 8454–8462.

He, R.; and McAuley, J. 2016. VBPR: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

He, X.; Deng, K.; Wang, X.; Li, Y.; Zhang, Y.; and Wang, M. 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*, 639–648.

Huang, T.; Dong, Y.; Ding, M.; Yang, Z.; Feng, W.; Wang, X.; and Tang, J. 2021. Mixgcf: An improved training method for graph neural network-based recommender systems. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 665–674.

Jiang, Y.; Xia, L.; Wei, W.; Luo, D.; Lin, K.; and Huang, C. 2024. DiffMM: Multi-Modal Diffusion Model for Recommendation. *Proceedings of the 32ed ACM International Conference on Multimedia*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lee, D.; Kang, S.; Ju, H.; Park, C.; and Yu, H. 2021. Bootstrapping user and item representations for one-class collaborative filtering. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 317–326.

Liu, S.; Chen, Z.; Liu, H.; and Hu, X. 2019. User-video co-attention network for personalized micro-video recommendation. In *The world wide web conference*, 3020–3026.

McAuley, J.; Targett, C.; Shi, Q.; and Van Den Hengel, A. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, 43–52.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Rendle, S.; Freudenthaler, C.; Gantner, Z.; and Schmidt-Thieme, L. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*.

Tao, Z.; Liu, X.; Xia, Y.; Wang, X.; Yang, L.; Huang, X.; and Chua, T.-S. 2022. Self-supervised learning for multimedia recommendation. *IEEE Transactions on Multimedia*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, Q.; Wei, Y.; Yin, J.; Wu, J.; Song, X.; and Nie, L. 2021. Dualgnn: Dual graph neural network for multimedia recommendation. *IEEE Transactions on Multimedia*.

Wei, W.; Huang, C.; Xia, L.; and Zhang, C. 2023. Multi-Modal Self-Supervised Learning for Recommendation. In *Proceedings of the ACM Web Conference 2023*, 790–800.

Wei, Y.; Wang, X.; Nie, L.; He, X.; and Chua, T.-S. 2020. Graph-refined convolutional network for multimedia recommendation with implicit feedback. In *Proceedings of the 28th ACM international conference on multimedia*, 3541–3549.

Wei, Y.; Wang, X.; Nie, L.; He, X.; Hong, R.; and Chua, T.-S. 2019. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM international conference on multimedia*, 1437–1445.

Wu, J.; Wang, X.; Feng, F.; He, X.; Chen, L.; Lian, J.; and Xie, X. 2021. Self-supervised graph learning for recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 726–735.

Xu, J.; Chen, Z.; Li, J.; Yang, S.; Wang, H.; and Ngai, E. C. 2024a. AlignGroup: Learning and Aligning Group Consensus with Member Preferences for Group Recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2682–2691.

Xu, J.; Chen, Z.; Li, J.; Yang, S.; Wang, W.; Hu, X.; and Ngai, E. C.-H. 2024b. FourierKAN-GCF: Fourier Kolmogorov-Arnold Network–An Effective and Efficient Feature Transformation for Graph Collaborative Filtering. *arXiv preprint arXiv:2406.01034*.

Xu, J.; Chen, Z.; Ma, Z.; Liu, J.; and Ngai, E. C. 2024c. Improving Consumer Experience With Pre-Purify Temporal-Decay Memory-Based Collaborative Filtering Recommendation for Graduate School Application. *IEEE Transactions on Consumer Electronics*.

You, Y.; Chen, T.; Sui, Y.; Chen, T.; Wang, Z.; and Shen, Y. 2020. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33: 5812–5823.

Yu, J.; Xia, X.; Chen, T.; Cui, L.; Hung, N. Q. V.; and Yin, H. 2023. XSimGCL: Towards extremely simple graph contrastive learning for recommendation. *IEEE Transactions on Knowledge and Data Engineering*.

Yu, J.; Yin, H.; Li, J.; Wang, Q.; Hung, N. Q. V.; and Zhang, X. 2021. Self-supervised multi-channel hypergraph convolutional network for social recommendation. In *Proceedings of the web conference 2021*, 413–424.

Yu, J.; Yin, H.; Xia, X.; Chen, T.; Cui, L.; and Nguyen, Q. V. H. 2022. Are graph augmentations necessary? simple graph contrastive learning for recommendation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, 1294–1303.

Zhang, J.; Zhu, Y.; Liu, Q.; Wu, S.; Wang, S.; and Wang, L. 2021. Mining latent structures for multimedia recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*, 3872–3880.

Zhou, X. 2023. MMRec: Simplifying Multimodal Recommendation. *arXiv preprint arXiv:2302.03497*.

Zhou, X.; Lin, D.; Liu, Y.; and Miao, C. 2023a. Layer-refined graph convolutional networks for recommendation. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, 1247–1259. IEEE.

Zhou, X.; and Shen, Z. 2023. A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 935–943.

Zhou, X.; Sun, A.; Liu, Y.; Zhang, J.; and Miao, C. 2023b. Selfcf: A simple framework for self-supervised collaborative filtering. *ACM Transactions on Recommender Systems*, 1(2): 1–25.

Zhou, X.; Zhou, H.; Liu, Y.; Zeng, Z.; Miao, C.; Wang, P.; You, Y.; and Jiang, F. 2023c. Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM Web Conference 2023*, 845–854.