SGP4SR: Separated-Modality Guided User Preference Learning for Multimodal Sequential Recommendation

Changhong Li[1], Zhiqiang Guo[2], Guohui Li[3*], Zhong Yang[3*], Chuhang Hong[1]

[1]School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China
[2]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[3]School of Software Engineering, Huazhong University of Science and Technology, Wuhan, China
{changhongli1, guohuili, zhongyang90, chuhanghong}@hust.edu.cn, georgeguo.gzq.cn@gmail.com

Abstract

With the booming development of multimodal data (e.g., image, text) on internet platforms, multimodal sequential recommendation methods continue to emerge. Most existing methods incorporate item modal features as auxiliary information, typically concatenating them to learn unified user representations. However, these methods directly use modal features for representation learning, neglecting the impact of inherent modal noise. We argue that internal-modal noise and cross-modal noise hinder the acquisition of more accurate user representations. To address this problem, we propose SGP4SR - Separated-modality Guided user Preference learning for multimodal Sequential Recommendation. Globally, the user preference modeling is carried out from a separated-modality perspective to alleviate cross-modal noise. Locally, for each individual modality, we use item relationship graphs and user interest centers, aggregated with ID embeddings, to replace direct modal features, thereby mitigating internal-modal noise. Finally, user representations from both separated-modality and multimodal perspectives participate in prediction independently. In experiments conducted on four real-world datasets, our method outperforms state-of-the-art approaches, achieving an average performance improvement of up to 8.84% over the best baseline. The comprehensive experiments further validate the superior noise tolerance and robustness of our method.

Code — https://github.com/changhongli1/SGP4SR

Introduction

Sequential recommender systems (Fang et al. 2020; Fan et al. 2022; Harte et al. 2023) (SRS) model user preferences by analyzing interaction sequences with temporal data. Early approaches, such as RNN-based (Hidasi et al. 2016; Hidasi and Karatzoglou 2018) and Transformer-based (Kang and McAuley 2018; Sun et al. 2019; Ma, Kang, and Liu 2019) methods, relied solely on item ID, but real-world interaction sparsity hinders the development of effective ID embeddings and thus limits recommendation performance. To address this, researchers are increasingly incorporating multimodal data (Baltrušaitis, Ahuja, and Morency 2018; Zhang et al. 2020; Zhao, Zhang, and Geng 2024)(images, texts, etc.) of items into SRS. Recent work has made significant strides: MissRec (Wang et al. 2023) captures the sequence-level multimodal synergy for better sequence representation. HM4SR (Zhang et al. 2025) enhances prediction by extracting multi-modal key item info and incorporating explicit temporal signals.

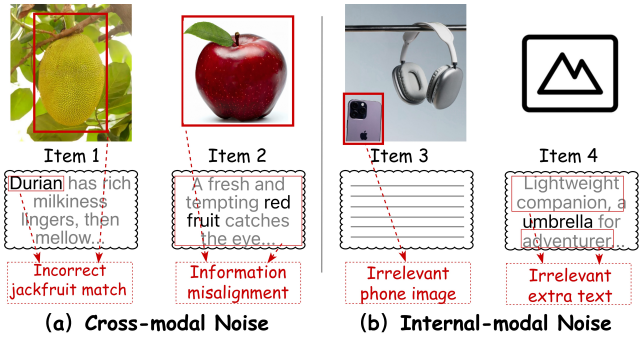

Figure 1: An illustration showing modal noise.

Although recent efforts have made significant progress, these methods often learn directly from multimodal features, which can introduce "**Modal Noise**": (1) *Cross-modal Noise: extra information generated during the fusion of embeddings from different modalities.* In real-world scenarios, modal features (such as image-text pairs) are often mismatched or inaccurate. For example, in Figure 1 (a), the description of item1 ("Durian") is incorrect compared to the image content ("jackfruit"), and the text description and the image content of item2 are similar but not completely aligned. Such inconsistencies introduce noise during multimodal fusion, making it more difficult to learn accurate representations. (2) *Internal-modal Noise: irrelevant information presented within individual modal features.* In image data, this manifests as extraneous items or background elements unrelated to the primary subject. For textual data, it includes descriptions that lack direct relevance to the recommendation task. As shown in Figure 1 (b), in the image of item3, element like the phone is irrelevant to the main subject, "headphone". Similarly, in the textual descriptions of item4, phrases such as "lightweight companion" and "adventurer" are unrelated to the recommendation task. Further-

*corresponding author.

more, internal-modal noise may cause modal information to deviate from the items themselves, further increasing the difficulty of matching and fusing between modalities, and indirectly exacerbating cross-modal noise.

To address the aforementioned issues, we propose SGP4SR - Separated-modality Guided user Preference learning for multimodal Sequential Recommendation. We design a separated-modality modeling framework for cross-modal noise and enhance input embeddings by incorporating modality relation graphs and interest cluster centers to address internal-modal noise. Specifically, (1) to filter cross-modal noise, we construct a separated-modality framework to model user representations in each modality and multi-modal representations respectively, with both participating independently in subsequent training and prediction. And (2) to filter internal-modal noise, we design a Co-occurrence guided Graph Construction Module (CGC) to construct a more accurate modal relationship graph for each modality and a Clustering Interest Perception Module (CIP) to identify users' clustered interest centers. The CGC employs item sequence co-occurrence patterns to guide graph construction, enabling nodes to concentrate on related items while filtering out irrelevant noisy connections. Additionally, the CIP leverages clustering interest centers to perceive the evolution of user interests in representation learning and guide them closer to the modal main information while keeping them away from irrelevant noise. Finally, user representations are obtained by fusing multi-modal embeddings with graph-based and clustering-based embeddings of each modality after training. In summary, the main contributions of this paper are as follows:

- We argue that the existing work fails to consider the additional noise introduced by the modal features themselves, which limits further improvements in sequential recommendation performance.
- We propose SGP4SR, a refined separate-modality framework that mitigates cross-modal noise, and incorporates the CGC and CIP modules to reduce internal-modal noise.
- Based on tests on four real-world datasets, our method achieves an average performance improvement of 20% over the strongest baseline in the best-performing dataset, demonstrating our model's superiority and its effectiveness in mitigating modal noise.

Related Works

Sequential Recommendation

Sequential recommendation (Wang et al. 2019; Chang et al. 2021; Yu et al. 2023a) aims to use the existing interaction sequences of users to predict the next most likely interacted item. Early sequential recommendations were mostly based on Markov chains (Kabbur, Ning, and Karypis 2013; He and McAuley 2016a) and neural networks (Tan, Xu, and Liu 2016; Tang and Wang 2018) to explore user preferences. Subsequently, Transformer continues to be employed in sequential recommendation scenarios due to its powerful learning capabilities. SASRec (Kang and McAuley 2018) achieves excellent improvements by using self-attention to mine potential sequence behaviors of users. In subsequent research (Ding et al. 2021; Zhou et al. 2020), many researchers have attempted to incorporate auxiliary information, such as item features, into sequence models. FDSA (Zhang et al. 2019) utilizes attention mechanisms to capture a variety of heterogeneous product features. DIF-SR (Xie, Zhou, and Kim 2022) decouples the attention calculation of various side information and item representation. UniSRec (Hou et al. 2022) uses item text to derive more transferable representations for sequences. Although these methods can partially alleviate the issue of sparse interactions, the improvement in model performance is still limited.

Multi-modal Recommendation

Multimodal information (Sarter 2006; Ngiam et al. 2011; Rahman et al. 2020) has been proven effective in alleviating data sparsity in many collaborative filtering recommendations (He and McAuley 2016b; Wei et al. 2019; Yu et al. 2023b). With the deepening of research, researchers have also begun to attempt to introduce it into sequential recommendation. In the field of multimodal sequential recommendations (Ji et al. 2023; Hu et al. 2023; Song et al. 2023), current research focuses mainly on fusion methods, pre-training, and other directions. MMMLP (Liang et al. 2023) proposes an MLP-based multimodal recommendation framework, which can explicitly learn information from various modalities. MissRec (Wang et al. 2023) utilizes pre-training and transfer learning to effectively address the cold start problem and enable efficient domain adaptation. IISAN (Fu et al. 2024) proposes a decoupled PEFT architecture with multiple modal adaptation that matches full fine-tuning performance. HM4SR (Zhang et al. 2025) introduces a hierarchical, time-aware Mixture-of-Experts framework comprising an Interactive MoE for utilizing multimodal information. Although these methods explore multimodal sequential recommendation from various perspectives and have achieved significant progress, their neglect of the noise inherent in modal features limits further improvement in performance.

Methodology

Notations and Problem Formulation

We consider an implicit recommender system that consists of a user set $\mathcal{U} = \{u_1, u_2, \ldots, u_{|\mathcal{U}|}\}$ with $|\mathcal{U}|$ users, an item set $\mathcal{X} = \{x_1, x_2, \ldots, x_{|\mathcal{X}|}\}$ with $|\mathcal{X}|$ items and a modality set $\mathcal{M} = \{V, T\}$ containing images and text. The ID embeddings of items are denoted as $\mathbf{E}^{id} = \{\mathbf{e}_1^{id}, \mathbf{e}_2^{id}, \ldots, \mathbf{e}_{|\mathcal{X}|}^{id}\} \in \mathbb{R}^{|\mathcal{X}| \times d}$, where d represents embedding dimension. The modal features of items are represented as $\mathbf{E}^{m} = \{\mathbf{e}_1^{m}, \mathbf{e}_2^{m}, \ldots, \mathbf{e}_{|\mathcal{X}|}^{m}\} \in \mathbb{R}^{|\mathcal{X}| \times d_m}$, where d_m represents the embedding dimension of modality m, $m \in \mathcal{M}$. Each user u is represented by their own interaction history sequence $\mathcal{S}^u = \{x_1, \ldots, x_i, \ldots, x_t | x_i \in \mathcal{X}\}$, $u \in \mathcal{U}$, where t represents the sequence length. Based on a given user interaction sequence $\mathcal{S}^u$, the core goal of sequential recommendation is to predict the next item that the user is most likely to interact with.

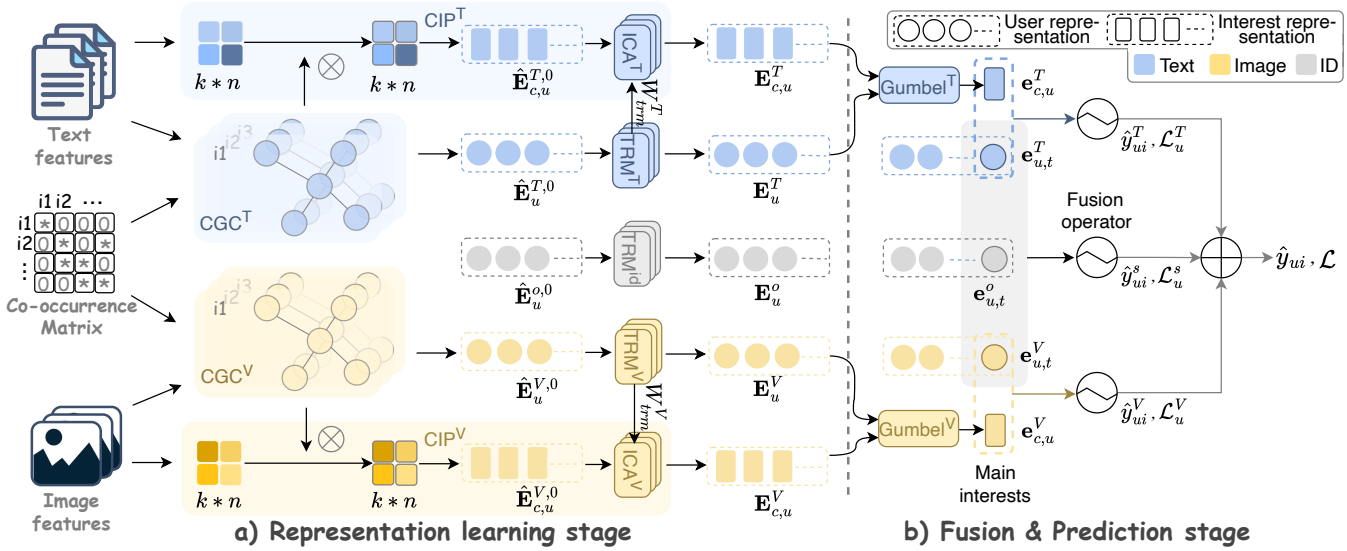

Figure 2: The overall framework of SGP4SR consists of two stages. a) Representation learning stage: Prepare and learn user representations and interest centers using modal features and ID embedding. b) Fusion & Prediction stage: Model user representations from both separated-modality and multimodal perspectives, each independently contributing to prediction.

Overview of the Model Structure

Figure 2 presents the overall structure of the model. From a global perspective, our model employs a modality-separated architecture, where each individual modality and ID-based user representations are learned independently. These representations from different modalities are only fused during the prediction stage. From a local perspective, the model can be divided into two stages: a) in the representation learning stage, a Co-occurrence Guided graph Construction Module (CGC) is designed to build modal item relation graphs, and a Clustering Interest Perception Module (CIP) is used to explore modal subject information. In the subsequent representation learning phase, the Transformer and the Interest Center Attention (ICA) Module are used to further learn the user sequence and interest center representations. b) In fusion & prediction stage, firstly, learned user representations and distinctive interest-center representations are integrated to generate single-modality user representations, as well as multimodal user representations. Subsequently, the final user prediction is achieved by integrating the prediction of these representations.

Representation Learning

Co-occurrence Guided Graph Construction Module (CGC) As analyzed in our introduction, modal features may not be suitable as direct inputs due to their inherent irrelevant noise. Therefore, in this section, we design the CGC module to replace the direct use of modal features in the form of modal relationship graphs. Unlike existing methods (Zhang et al. 2021; Zhou and Shen 2023) that build item relation graphs solely from modal features, we use co-occurrence signals that are objectively determined and implicitly contain item-attribute correlations (Liu et al. 2024) to guide modal features (relatively uncertain information affected by encoders, adapters, and other factors) in selecting more effective neighbor nodes for each item, thus constructing a more robust modal relation graph.

Co-occurrence & Modal relationship Matrix Calculation For items x_i and x_j, we count the number of times they are interacted with by the same user, denoted as $\mathcal{A}^o_{ij}$, and store this number in a matrix $\mathcal{A}^o \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$ that is initialized to be all zeros. Then, by calculating the pairwise interactions between all items, we can finally obtain the complete co-occurrence matrix $\hat{\mathcal{A}}^o$. Furthermore, we adopt the cosine similarity to calculate the semantic affinity of x_i and x_j,

$$\mathcal{A}^m_{ij} = \frac{(\mathbf{e}^m_i)^\top(\mathbf{e}^m_j)}{||\mathbf{e}^m_i||||\mathbf{e}^m_j||}, \tag{1}$$

where $\mathcal{A}^m_{ij}$ represents the semantic affinity score between items i and j in modality m. $\mathbf{e}^m_i$ and $\mathbf{e}^m_j$ are modal features of items i and j extracted from the matrix $\mathbf{E}^m$.

Modal Relationship Graph Neighbor Selection Based on $\mathcal{A}^m_{ij}$ and $\mathcal{A}^o_{ij}$, for item x_i, its semantic similarity in modality m and co-occurrence behavior information can be expressed as $\mathcal{A}^{\mathcal{G}}_{i*}, \mathcal{G} \in \{m, o\}$. Then, we select the top $H^{\mathcal{G}}$ items with the highest scores from $\mathcal{A}^{\mathcal{G}}_{i*}$, denoted as $\mathcal{N}^{\mathcal{G}}_i$, here, $H^{\mathcal{G}}$ is used to control the number of relevant items for item i.

To make more adequate use of modal features, for $\mathcal{N}^m_i$ (compared to existing methods (Zhang et al. 2021; Zhou and Shen 2023) that extract latent modal structures of items), we select a relatively larger H^m. For $\mathcal{N}^o_i$, we retain fewer items ($H^o < H^m$) to ensure all items in $\mathcal{N}^o_i$ exhibit higher co-occurrence relevance. High-relevance items in $\mathcal{N}^o_i$ can more accurately guide $\mathcal{N}^m_i$ to filter out noisy items introduced by the excessive selection of related items. Specifically, we select the common items from $\mathcal{N}^m_i$ and $\mathcal{N}^o_i$ to obtain the

modality-occurrence related items for item i,

$$\mathcal{N}_i^{m,o} = \mathcal{N}_i^m \cap \mathcal{N}_i^o, \tag{2}$$

$\mathcal{N}_i^{m,o} \in \mathbb{R}^{1\times H_i^{m,o}}$, and $H_i^{m,o}$ depends on the number of items that simultaneously exhibit high modal semantic similarity and co-occurrence behavior similarity.

Next, we choose to further expand the content of neighboring nodes. A hyperparameter H (bounded between H^o and H^m) is set to control the number of nodes in the final relationship graph. $\mathcal{N}_i^{m,o}$ contains fewer than H relevant items ($H_i^{m,o} \leq H^o < H < H^m$), we subsequently select the top $H_i^{m/o} = H - H_i^{m,o}$ highest-relevance items from the remaining candidates in $\mathcal{N}_i^m$, denoted as $\mathcal{N}_i^{m/o}$. Then, by aggregating the items from $\mathcal{N}_i^{m,o}$ and $\mathcal{N}_i^{m/o}$, we can finally determine the set $\hat{\mathcal{N}}_i^{m,o} = [\mathcal{N}_i^{m,o}, \mathcal{N}_i^{m/o}]$ with H relevant items for item i.

Compared to directly extracting H items from $\mathcal{A}_{i*}^m$, the calculation of $\hat{\mathcal{N}}_i^{m,o}$ based on $\mathcal{N}_i^m$ with H^m items, provides a more comprehensive consideration of the semantic information of modality m. Compared to directly selecting H^m items from $\mathcal{A}_{i*}^m$ (equivalent to $\mathcal{N}_i^m$), this method reduces the additional noise introduced by excessive neighbor selection. Therefore, for item i, we can construct its adjacent item relationship vector,

$$\hat{\mathcal{A}}_{ij}^m = \begin{cases} 1, & \mathcal{A}_{ij}^m \in \hat{\mathcal{N}}_i^{m,o}, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

by performing the above procedure for all items, we can construct the semantic affinity graph $\hat{\mathcal{A}}^m \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$ as item relation graph for modality m. It is worth mentioning that the entire calculation process of $\hat{\mathcal{A}}^m$ can be performed offline. Finally, we propagate modal relationships into ID embeddings to participate in subsequent sequential representation learning,

$$\hat{\mathbf{E}}^{\mathcal{G}} = \hat{\mathcal{A}}^{\mathcal{G}} \mathbf{E}^{id}. \tag{4}$$

Self-attention Learning Following representative methods (Kang and McAuley 2018; Sun et al. 2019), we utilize Transformer (Vaswani et al. 2017) to learn accurate and reliable sequence representations. We use it to capture long-distance dependencies in sequence embeddings. Based on $\hat{\mathbf{E}}^{\mathcal{G}}$, we introduce positional information for the user u's sequence,

$$\hat{\mathbf{E}}_u^{\mathcal{G}} = \hat{\mathbf{E}}^{\mathcal{G}}[\mathcal{S}^u] + \mathbf{P} \tag{5}$$

where $\hat{\mathbf{E}}^{\mathcal{G}}[\mathcal{S}^u] = \{\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \ldots, \hat{\mathbf{e}}_t\} \in \mathbb{R}^{t\times d}$ represents the items in the user's interaction sequence extracted from the embedding matrix $\hat{\mathbf{E}}^{\mathcal{G}}$. $\mathbf{P} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_t\} \in \mathbb{R}^{t\times d}$ is a learnable position embedding. Then, after applying operations like Dropout and LayerNorm, $\hat{\mathbf{E}}_u^{\mathcal{G}}$ can be fed into the Transformer for learning,

$$\hat{\mathbf{E}}_u^{\mathcal{G}} = \mathrm{Dropout}(\mathrm{LayerNorm}(\hat{\mathbf{E}}_u^{\mathcal{G}})), \tag{6}$$

$$\mathbf{E}_u^{\mathcal{G}} = \mathbf{Trm}^L(\hat{\mathbf{E}}_u^{\mathcal{G},0}), \tag{7}$$

Here, $\hat{\mathbf{E}}_u^{\mathcal{G},0} = \hat{\mathbf{E}}_u^{\mathcal{G}}$, which denotes the user representation at the 0-th layer.

Clustering Interest Perception Module (CIP) Clustering is an effective method to explore users' core interests (Li et al. 2019). We leverage it to make user representations closer to core preferences while farther away from internal-modal noise. It first initializes user interest centers based on modal features, and then incorporates an interest center attention mechanism (ICA) to enable the interest centers to perceive changes in user interests during learning.

Interest Center Initialization By conducting K-means clustering on modal features, we first initialize a batch of interest centers,

$$\mathbf{C}^m = \text{k-means}(\mathbf{E}^m), \tag{8}$$

where $\mathbf{C}^m \in \mathbb{R}^{k\times|\mathcal{X}|}$ represents the relationship between all items and k cluster centers. We aggregate $\hat{\mathcal{A}}^m$ to inject contextual neighbor information, thereby enhancing the receptive field of clusters,

$$\hat{\mathbf{C}}^m = \mathbf{C}^m \hat{\mathcal{A}}^m. \tag{9}$$

Interest Center Attention (ICA) Based on $\mathbf{E}^{id}$, we initialize each user's interest center representation $\hat{\mathbf{E}}_{c,u}^{m,0} = \hat{\mathbf{C}}^m \mathbf{E}^{id} \in \mathbb{R}^{k\times d}$. For each user, their interest centers will perceive the evolution of user interests during the learning process of sequence representations. The specific steps are as follows,

$$\mathbf{a}_h^l = \textsc{Softmax}\left(\frac{(\hat{\mathbf{E}}_{c,u}^{m,l-1}\mathbf{W}_h^{q,l})(\mathbf{E}_u^{m,l-1}\mathbf{W}_h^{k,l})^T}{\sqrt{d}}\right), \tag{10}$$

$$\mathbf{head}_h^l = \mathbf{a}_h^l(\mathbf{E}_u^{m,l-1}\mathbf{W}_h^{v,l}), \tag{11}$$

$$\mathbf{g}^l = [\mathbf{head}_1^l; \mathbf{head}_2^l; \ldots; \mathbf{head}_{|h|}^l]\mathbf{U}^l, \tag{12}$$

$$\hat{\mathbf{E}}_{c,u}^{m,l} = \sigma((\mathbf{g}^l\mathbf{W}_1^l + \mathbf{b}_1^l)\mathbf{W}_2^l + \mathbf{b}_2^l), \tag{13}$$

where $\hat{\mathbf{E}}_{c,u}^{m,l}$ represents the center feature of l-th layer. $\{\mathbf{W}_h^{q,l}, \mathbf{W}_h^{k,l}, \mathbf{W}_h^{v,l}\} \in \mathbb{R}^{d\times d}$ come from the multi-head attention in Eq(7) to generate the query, key and value vectors. $\mathbf{E}_u^{m,l-1}$ represents the user representation output of Eq(7) at $(l-1)$-th layer. $\mathbf{a}_h^l$ is the generated attention score of the h-th attention head. h is the number of heads. $\mathbf{U}^l \in \mathbb{R}^{d\times d}$ is a learnable parameter to integrate the attention heads in the l-th layer. After L layers of centralized attention learning, the final center representations for user u are updated as,

$$\mathbf{E}_{c,u}^m = \hat{\mathbf{E}}_{c,u}^{m,L}, \tag{14}$$

here, $\mathbf{E}_{c,u}^m$ is able to capture user u's interest in modality m.

Fusion & Prediction

In this section, we mainly complete the fusion of representations and subsequent prediction. Firstly, our fusion process is divided into separated-modality fusion and multimodal fusion. For separated-modality fusion, we aggregate the user's sequence representations with their own interest centers to guide user representations closer to the main modal information while keeping them away from irrelevant noise. For multimodal fusion, we combine the user's sequence representations learned from different modalities. Finally, the prediction relies on all user representations derived from both separated-modality and multimodal fusion, providing a comprehensive basis for accurate recommendation.

Separated-modality Fusion Using Gumbel-Softmax, we can identify the most important interest centers $\mathbf{e}_{c,u}^m \in \mathbb{R}^{1\times d}$ for user u,

$$\mathbf{e}_{c,u}^m = \text{GUMBEL\_SOFTMAX}(\mathbf{E}_u^m, \mathbf{E}_{c,u}^m), \tag{15}$$

then, we can optimize the seperated-modality representation of user u,

$$\widetilde{\mathbf{e}}_u^m = \mathbf{e}_{u,t}^m + \alpha^m \mathbf{e}_{c,u}^m, \tag{16}$$

here, $\mathbf{e}_{u,t}^m = \mathbf{E}_u^m[-1]$ refers to the representation corresponding to the last item in the user sequence. Since it reflects the user's most recent interaction, it is well-suited to represent the user. α^m is a hyperparameter used to control the weight of the interest center. $\widetilde{\mathbf{e}}_u^m$ will focus more on the main modal information while staying away from irrelevant information, and at the same time, it will model the representation of user u under modality m in greater depth.

Multimodal Fusion Based on the user representations learned from each modality, we fuse the sequence representations from multiple modalities,

$$\mathbf{e}_u^s = \sum_{m\in\mathcal{M}} \rho_m \cdot \mathbf{e}_{u,t}^m + \mathbf{e}_{u,t}^o, \tag{17}$$

where $\mathbf{e}_{u,t}^m$ and $\mathbf{e}_{u,t}^o$ are the last (t-th) item representations in user u's sequence. ρ_m is a hyperparameter. We set $\sum_{m\in\mathcal{M}} \rho_m = 1$.

Prediction Optimization After obtaining the user sequence representation, we use the user representation and ID embeddings of items to calculate the prediction score $\hat{y}_{ui}^s$ of user u and item x_i,

$$\hat{y}_{ui}^s = \mathbf{e}_u^s (\mathbf{e}_i^{id})^\top, \tag{18}$$

where $\mathbf{e}_i^{id}$ is the ID embedding of item x_i. To further enhance the expressive capability of separated-modality representations, from a single-modality perspective, we achieve the independent prediction in each modality based on $\widetilde{\mathbf{e}}_u^m$,

$$\hat{y}_{ui}^m = \widetilde{\mathbf{e}}_u^m (\hat{\mathbf{e}}_i^m)^\top, \tag{19}$$

here, $\hat{\mathbf{e}}_i^m$ comes from embedding $\hat{\mathbf{E}}^{\mathbf{m}}$, and $\hat{y}_{ui}^m$ is the predicted score of item i for user u in modality m. This further enhances the influence of user preferences under a single modality. Then, the final prediction score $\hat{y}_{ui}$ of user u and item x_i is calculated as,

$$\hat{y}_{ui} = \hat{y}_{ui}^s + \sum_{m\in\mathcal{M}} \rho_m \cdot \hat{y}_{ui}^m. \tag{20}$$

Subsequently, following other sequential recommendations (Ji et al. 2023; Wang et al. 2023), we use cross entropy loss (Zhang and Sabuncu 2018; Yaoshiang and Wookey 2019) as the recommendation loss, which can minimize the negative logarithmic likelihood of the ground truth value for correctly recommending the next item. Thus, our overall loss function can be expressed as,

$$\mathcal{L} = \frac{1}{|\mathcal{U}|} \sum_{u\in\mathcal{U}} \mathcal{L}_u^s + \sum_{m\in\mathcal{M}} \rho_m \cdot \mathcal{L}_u^m, \tag{21}$$

here, the calculations of $\mathcal{L}_u^m$ and $\mathcal{L}_s^m$ are based on cross-entropy loss and can be expressed as,

$$\mathcal{L}_u^m = -\sum_{x_i\in\mathcal{X}} y_{ui} \log(\hat{y}_{ui}^m), \tag{22}$$

where y_{ui} is the ground-truth binary interaction value. The loss $\mathcal{L}_u^s$ can be implemented in similar manner.

Computational complexity

In our work, the computational complexity of the training process mainly comes from two parts: graph convolution and attention computation. Specifically, the computational complexity of graph convolution (Eq.(4)) is $\mathcal{O}(|\mathcal{X}|^2 d)$. The computational cost of attention (Eq.(7) and Eq.(10)–Eq.(13)) is $\mathcal{O}((t^2+k^2)d + (t+k)d^2)$. Notably, the item relationship graph can all be computed offline. Overall, the complexity of our work is comparable to that of existing Transformer-based multimodal sequential recommendation methods.

Dataset	#Users	#Items	#Inters	Avg.n	Sparsity
Pantry	13,102	4,899	113,861	8.691	0.9982
Baby	19,446	7,051	141,347	7.269	0.9990
Clothing	39,388	23,034	239,290	6.075	0.9997
Office	87,436	25,986	684,837	7.840	0.9997

Table 1: Statistics of four evaluation datasets.

Experiment

Experimental Setup

Datasets and Preprocessing We conduct evaluation experiments on four publicly available benchmark datasets from the widely-used Amazon platform[1], which contains reviews from millions of Amazon customers. We collect (a) *Prime Pantry* (**Pantry**), (b) *Baby* (**Baby**), (c) *Clothing, Shoes and Jewelry* (**Clothing**), and (d) *Office Products* (**Office**) to train and evaluate our method. Table 1 summarizes the statistical results of these four datasets. Following previous works (Wang et al. 2023; Zhou 2023), for **Pantry** and **Office** datasets, we use pre-trained CLIP-B/32 to extract textual and visual features with 512 dimensions. For **Baby** and **Clothing** datasets, we use the 4096-dimensional visual features that have been extracted and published, and we extract textual embeddings by concatenating the title, descriptions, categories, and brand of each item and utilizing pre-trained sentence-transformers to obtain 384-dimensional sentence embeddings.

Evaluation Protocols The performance of our SGP4SR on the testing set is evaluated by two commonly used protocols: Recall@N focuses on how many correct items are recommended, while NDCG@N accounts for the ranking quality of correct items. We truncate the ranked list by setting N to $\{10, 50\}$. After training, the learned recommendation model can generate a ranked top-N list from all items to evaluate the two protocols.

[1]http://jmcauley.ucsd.edu/data/amazon/links.html

Datasets	Metric	Bert4Rec	SASRec	FDSA	UniSRec	SASRecF	MMMLP	MissRec	IISAN	HM4SR	SGP4SR	improv.
Pantry	R@10	0.0308	0.0501	0.0395	0.0693	0.0557	0.0531	<u>0.0779</u>	0.0518	0.0552	**0.0802**	2.95%
	N@10	0.0152	0.0214	0.0209	0.0311	0.0243	0.0273	<u>0.0365</u>	0.0248	0.0235	**0.0380**	4.11%
	R@50	0.1030	0.1322	0.1151	0.1827	0.1488	0.1382	**0.1875**	0.1437	0.1481	<u>0.1833</u>	-2.29%
	N@50	0.0305	0.0394	0.0370	0.0556	0.0441	0.0454	<u>0.0598</u>	0.0442	0.0432	**0.0599**	0.17%
Baby	R@10	0.0254	0.0444	0.0431	<u>0.0534</u>	0.0464	0.0425	0.0519	0.0456	0.0482	**0.0596**	11.61%
	N@10	0.0121	0.0192	0.0207	0.0240	0.0203	0.0213	<u>0.0247</u>	0.0222	0.0208	**0.0283**	14.57%
	R@50	0.0792	0.1227	0.1250	0.1465	0.1267	0.1165	<u>0.1480</u>	0.1261	0.1274	**0.1538**	3.92%
	N@50	0.0227	0.0282	0.0380	0.0424	0.0373	0.0369	<u>0.0437</u>	0.0392	0.0376	**0.0477**	9.15%
Clothing	R@10	0.0119	0.0225	0.0215	0.0403	0.0311	0.0219	<u>0.0422</u>	0.0232	0.0279	**0.0541**	28.20%
	N@10	0.0057	0.0109	0.0106	0.0172	0.0136	0.0107	<u>0.0201</u>	0.0113	0.0126	**0.0261**	29.85%
	R@50	0.0318	0.0583	0.0572	0.1013	0.0735	0.0574	<u>0.1151</u>	0.0645	0.0748	**0.1243**	7.99%
	N@50	0.0097	0.0137	0.0182	0.0303	0.0226	0.0182	<u>0.0348</u>	0.0201	0.0225	**0.0411**	18.10%
Office	R@10	0.0825	0.1229	0.1076	<u>0.1280</u>	0.1231	0.1103	0.1275	0.1161	0.1216	**0.1307**	2.11%
	N@10	0.0634	0.0812	0.0868	0.0831	0.0801	0.0870	0.0856	<u>0.0875</u>	0.0808	**0.0876**	0.11%
	R@50	0.1227	0.1827	0.1665	<u>0.2016</u>	0.1879	0.1503	0.2005	0.1732	0.1835	**0.2055**	1.93%
	N@50	0.0721	0.0930	0.0987	0.0991	0.0938	0.0956	<u>0.1012</u>	0.0997	0.0941	**0.1037**	2.47%

Table 2: Performance comparisons of SGP4SR and other baselines on four datasets. The best result is in boldface and the second best is underlined. Improvement is obtained between SGP4SR and the best result in baselines.

Baselines We compare our SGP4SR with the following competitive methods, divided into three groups: 1) ID-based sequential recommendations: **SASRec** (Kang and McAuley 2018), **Bert4Rec** (Sun et al. 2019). 2) Text-based sequential recommendations: **FDSA** (Zhang et al. 2019), **UniSRec** (Hou et al. 2022). 3) Multimodal-based sequential recommendations: **SASRecF** (**SASRec** with multimodal features), **MMMLP** (Liang et al. 2023), **MissRec** (Wang et al. 2023), **IISAN** (Fu et al. 2024) and **HM4SR** (Zhang et al. 2025).

Experimental Details Our model is implemented in PyTorch[2]. For each user sequence in all datasets, we select the last item to construct the test set and the one before it for the validation set. The remaining items are included in the training set. For a fair comparison, we optimize all models via the Adam (Kingma and Ba 2014) optimizer with the fixed embedding size 300 and the mini-batch size 512. In addition, we search the learning rate from $\{1e^{-4}, 1e^{-3}, \ldots, 1e^{-1}\}$, the neighbor number H in modality-aware relation graph from $\{2, 4, 6, 8, 10, 12, 14, 16, 18, 20\}$, the co-occurrence neighbor parameter μ^o from $\{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$, the modal neighbor parameter μ^m from $\{1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6\}$, the center number k from $\{0, 10, 20, 30, 40, 50, 60, 70, 80\ 90, 100\}$. An early stopping mechanism with a patience of 10 is applied to alleviate overfitting problems.

Performance Comparison

Table 2 reports the performance comparisons of SGP4SR and all baselines in terms of Recall and NDCG on four datasets. From the table, we have the following observations:

- SGP4SR nearly achieves substantial improvements over all baselines across four datasets. Compared with the suboptimal MissRec, SGP4SR achieves an average performance improvement of 8.84% across the four datasets, and particularly in Clothing, the performance gain reaches approximately 20%.
- Compared to methods such as MMMLP and HM4SR, which do not explicitly account for modal noise, SGP4SR demonstrates significant advantages across all datasets. This implicitly illustrates the impact of modal noise on model performance and highlights the effectiveness of noise reduction in enhancing preference learning.
- Compared to ID-based sequential recommendations, text-based or multimodal-based sequential recommendations demonstrate significant performance advantages. The results indicate the effectiveness of introducing modal features in modeling representations of users and items.
- Among the multimodal-based sequential recommendation methods, MMMLP which uses a lightweight MLP as the backbone network performs inferior to other methods that utilize Transformer architectures as the backbone network in most cases. This demonstrates the advantages of self-attention in sequential recommendation.

Modality Mismatch Noise Experiment

To further verify the robustness and capability of mitigating cross-modal noise of the proposed method, we randomly select 5 groups of items with different proportions from the Baby dataset and randomly replace their image modality features with those of other items to simulate the modality mismatch noise in real-world scenarios. Three representative multimodal methods (SASRecF, MissRec and HM4SR) are presented for comparison with SGP4SR. It can be observed in Figure 3 that all baseline methods exhibit performance degradation to varying degrees, while SGP4SR remains stable with a slight decline. For instance, at a replacement ratio of 50%, the Recall@10 scores of each baseline declined by 23.06%, 29.48%, and 11.41%, while the NDCG@10 scores dropped by 32.02%, 29.15%, and

[2]https://pytorch.org

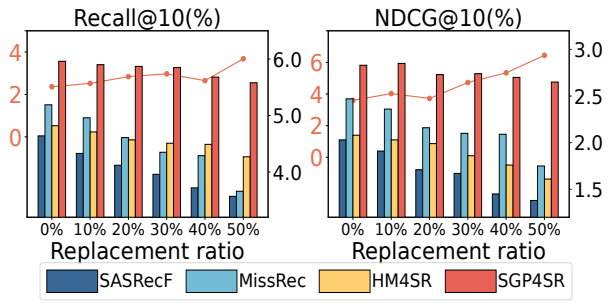

Figure 3: Modality mismatch noise experiment by replacing image modality features on the **Baby** dataset. The line curve (with scales corresponding to the left axis) represents the performance ratio of SGP4SR to HM4SR.

Datasets	Baby		Clothing		Office	
Methods	**R@10**	**N@10**	**R@10**	**N@10**	**R@10**	**N@10**
w/o COO	0.0565	0.0262	0.0529	0.0256	0.1293	0.0855
w/o CGC	0.0435	0.0224	0.0267	0.0142	0.1158	0.0862
w/o CIP	0.0547	0.0281	0.0369	0.0192	0.1214	0.0869
w/o MUL	0.0527	0.0264	0.0475	0.0239	0.1248	0.0859
w/o C2C	0.0435	0.0217	0.0245	0.0133	0.1079	0.0787
SGR4SR	**0.0596**	**0.0283**	**0.0541**	**0.0261**	**0.1307**	**0.0876**

Table 3: The effectiveness of different variants of SGP4SR.

22.60%, respectively. In contrast, our method experienced only about a 6% decrease in both metrics. Additionally, as the replacement ratio grows, the performance advantage of SGP4SR over HM4SR tends to increase, even though HM4SR is less affected by modality mismatch noise than other baselines. This clearly illustrates the substantial advantages of our separate-modality based approach in enhancing noise resistance and robustness.

Ablation Studies

Table 3 shows the results of ablation studies of SGP4SR on Baby, Clothing and Office datasets. Specifically, **w/o COO** represents the item relationship graph construction without the guidance of co-occurrence signals. **w/o CGC** removes CGC module and directly uses original modal features as input for user sequence representation learning. **w/o CIP** removes CIP module while retaining only sequential representation learning. **w/o MUL** refers to abandoning separated-modality framework by fusing the two modal features and feeding them into the original single-modality modeling process to learn user representations. **w/o C2C** denotes the simultaneous removal of CGC and CIP. We can observe:

- **w/o CGC** and **w/o CIP** cause significant performance degradation, showing the effectiveness of the two key components and their ability in mitigating the internal-modal noise. Furthermore, **w/o C2C** exhibits more significant performance degradation, which indicates the synergistic effect of CGC and CIP in handling internal-modal noise.

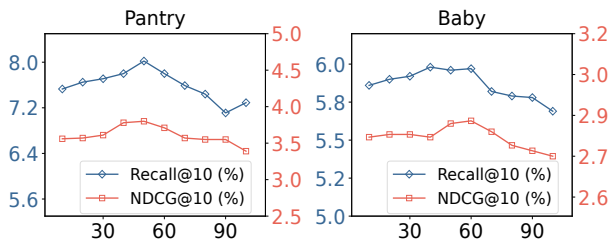

Figure 4: The impact of different center number k.

- **w/o COO** also causes a certain degree of performance degradation, which demonstrates the capability of co-occurrence signals in guiding the construction of modal relationship graphs.
- **w/o MUL** causes a significant performance degradation. It indicates the effectiveness of the separated-modality framework in mitigating cross-modal noise.

Key Parameter Experiment

Each module of SGP4SR involves different hyperparameters such as μ^o, H, k, etc., and their settings will affect performance to a certain extent. Due to space constraints, we only present the experimental results of k in the main text.

Impact of the Center Number k To further verify the effect of the center number k, we report the performance of SGP4SR with various values of k in Figure 4. We can observe that the performance is optimal when k increases to the range of $50-60$. Continuing to increase k of centers may instead introduce information irrelevant to the main subject. In practice, we set $k = 50, 50, 60, 60$ on Pantry, Baby, Clothing and Office, respectively.

Conclusion

In this work, we propose a novel SGP4SR, which models user representations from a separated-modality perspective to mitigate cross-modal noise and incorporates the CGC and CIP modules to construct item relationships and identify user interest centers respectively, thereby alleviating internal-modal noise. SGP4SR achieved substantial performance improvements over baseline methods across four real-world datasets, demonstrating its superiority in sequential recommendation. Subsequent modal noise experiments further confirmed its strong noise resistance and enhanced robustness. For future work, we plan to utilize item multi-modal data to explore interpretable recommendation results, so as to understand modal noise more intuitively.

Acknowledgements

We would like to express our sincere gratitude to the reviewers for their constructive comments. This work was partially supported by the National Natural Science Foundation of China under Grant Nos. 62302180 and 62272176, and the Equipment Pre-research Joint Fund Project of the Ministry of Education of China under Grant No. 8091B02072302.

References

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443.

Chang, J.; Gao, C.; Zheng, Y.; Hui, Y.; Niu, Y.; Song, Y.; Jin, D.; and Li, Y. 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.

Ding, H.; Ma, Y.; Deoras, A.; Wang, Y.; and Wang, H. 2021. Zero-shot recommender systems. *arXiv preprint arXiv:2105.08318*.

Fan, Z.; Liu, Z.; Wang, Y.; Wang, A.; Nazari, Z.; Zheng, L.; Peng, H.; and Yu, P. S. 2022. Sequential recommendation via stochastic self-attention. In *Proceedings of the ACM web conference 2022*, 2036–2047.

Fang, H.; Zhang, D.; Shu, Y.; and Guo, G. 2020. Deep learning for sequential recommendation: Algorithms, influential factors, and evaluations. *ACM Transactions on Information Systems (TOIS)*, 39(1): 1–42.

Fu, J.; Ge, X.; Xin, X.; Karatzoglou, A.; Arapakis, I.; Wang, J.; and Jose, J. M. 2024. IISAN: Efficiently adapting multimodal representation for sequential recommendation with decoupled PEFT. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 687–697.

Harte, J.; Zorgdrager, W.; Louridas, P.; Katsifodimos, A.; Jannach, D.; and Fragkoulis, M. 2023. Leveraging large language models for sequential recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, 1096–1102.

He, R.; and McAuley, J. 2016a. Fusing similarity models with markov chains for sparse sequential recommendation. In *2016 IEEE 16th international conference on data mining (ICDM)*, 191–200.

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

Hidasi, B.; and Karatzoglou, A. 2018. Recurrent neural networks with top-k gains for session-based recommendations. In *Proceedings of the 27th ACM international conference on information and knowledge management*, 843–852.

Hidasi, B.; Quadrana, M.; Karatzoglou, A.; and Tikk, D. 2016. Parallel recurrent neural network architectures for feature-rich session-based recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, 241–248.

Hou, Y.; Mu, S.; Zhao, W. X.; Li, Y.; Ding, B.; and Wen, J.-R. 2022. Towards universal sequence representation learning for recommender systems. In *Proceedings of SIGKDD*, 585–593.

Hu, H.; Guo, W.; Liu, Y.; and Kan, M.-Y. 2023. Adaptive multi-modalities fusion in sequential recommendation systems. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 843–853.

Ji, W.; Liu, X.; Zhang, A.; Wei, Y.; Ni, Y.; and Wang, X. 2023. Online distillation-enhanced multi-modal transformer for sequential recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 955–965.

Kabbur, S.; Ning, X.; and Karypis, G. 2013. Fism: factored item similarity models for top-n recommender systems. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 659–667.

Kang, W.-C.; and McAuley, J. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*, 197–206. IEEE.

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

Li, C.; Liu, Z.; Wu, M.; Xu, Y.; Zhao, H.; Huang, P.; Kang, G.; Chen, Q.; Li, W.; and Lee, D. L. 2019. Multi-interest network with dynamic routing for recommendation at Tmall. In *Proceedings of the 28th ACM international conference on information and knowledge management*, 2615–2623.

Liang, J.; Zhao, X.; Li, M.; Zhang, Z.; Wang, W.; Liu, H.; and Liu, Z. 2023. Mmmlp: Multi-modal multilayer perceptron for sequential recommendations. In *Proceedings of the ACM Web Conference*, 1109–1117.

Liu, Y.; Zhang, X.; Zou, M.; and Feng, Z. 2024. Attribute simulation for item embedding enhancement in multi-interest recommendation. In *Proceedings of the 17th ACM international conference on web search and data mining*, 482–491.

Ma, C.; Kang, P.; and Liu, X. 2019. Hierarchical gating networks for sequential recommendation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 825–833.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; Ng, A. Y.; et al. 2011. Multimodal deep learning. In *ICML*, volume 11, 689–696.

Rahman, W.; Hasan, M. K.; Lee, S.; Zadeh, A.; Mao, C.; Morency, L.-P.; and Hoque, E. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2020, 2359.

Sarter, N. B. 2006. Multimodal information presentation: Design guidance and research challenges. *International journal of industrial ergonomics*, 36(5): 439–445.

Song, K.; Sun, Q.; Xu, C.; Zheng, K.; and Yang, Y. 2023. Self-supervised multi-modal sequential recommendation. *arXiv preprint arXiv:2304.13277*.

Sun, F.; Liu, J.; Wu, J.; Pei, C.; Lin, X.; Ou, W.; and Jiang, P. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*, 1441–1450.

Tan, Y. K.; Xu, X.; and Liu, Y. 2016. Improved recurrent neural networks for session-based recommendations. In *Proceedings of the 1st workshop on deep learning for recommender systems*, 17–22.

Tang, J.; and Wang, K. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of WSDM*, 565–573.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, J.; Zeng, Z.; Wang, Y.; Wang, Y.; Lu, X.; Li, T.; Yuan, J.; Zhang, R.; Zheng, H.-T.; and Xia, S.-T. 2023. Missrec: Pre-training and transferring multi-modal interest-aware sequence representation for recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 6548–6557.

Wang, S.; Hu, L.; Wang, Y.; Cao, L.; Sheng, Q. Z.; and Orgun, M. 2019. Sequential recommender systems: challenges, progress and prospects. *arXiv preprint arXiv:2001.04830*.

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.

Xie, Y.; Zhou, P.; and Kim, S. 2022. Decoupled side information fusion for sequential recommendation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, 1611–1621.

Yaoshiang, H.; and Wookey, S. 2019. The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. *IEEE access*, 8: 4806–4813.

Yu, J.; Yin, H.; Xia, X.; Chen, T.; Li, J.; and Huang, Z. 2023a. Self-supervised learning for recommender systems: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 36(1): 335–355.

Yu, P.; Tan, Z.; Lu, G.; and Bao, B.-K. 2023b. Multi-view graph convolutional network for multimedia recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 6576–6585.

Zhang, C.; Yang, Z.; He, X.; and Deng, L. 2020. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3): 478–493.

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*, MM '21, 3872–3880.

Zhang, S.; Chen, L.; Shen, D.; Wang, C.; and Xiong, H. 2025. Hierarchical Time-Aware Mixture of Experts for Multi-Modal Sequential Recommendation. In *Proceedings of the ACM on Web Conference 2025*, 3672–3682.

Zhang, T.; Zhao, P.; Liu, Y.; Sheng, V. S.; Xu, J.; Wang, D.; Liu, G.; Zhou, X.; et al. 2019. Feature-level deeper self-attention network for sequential recommendation. In *Proceedings of IJCAI*, 4320–4326.

Zhang, Z.; and Sabuncu, M. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31.

Zhao, F.; Zhang, C.; and Geng, B. 2024. Deep multimodal data fusion. *ACM computing surveys*, 56(9): 1–36.

Zhou, K.; Wang, H.; Zhao, W. X.; Zhu, Y.; Wang, S.; Zhang, F.; Wang, Z.; and Wen, J.-R. 2020. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *Proceedings of the 29th ACM international conference on information & knowledge management*, 1893–1902.

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

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.