STAIR: Manipulating Collaborative and Multimodal Information for E-Commerce Recommendation

Cong Xu[*], **Yunhang He**[*], **Jun Wang**[†], **Wei Zhang**[†]

East China Normal University
Shanghai, China
{congxueric, yhhe2004, wongjun, zhangwei.thu2011}@gmail.com

Abstract

While the mining of modalities is the focus of most multimodal recommendation methods, we believe that how to fully utilize both collaborative and multimodal information is pivotal in e-commerce scenarios where, as clarified in this work, the user behaviors are rarely determined entirely by multimodal features. In order to combine the two distinct types of information, some additional challenges are encountered: 1) Modality erasure: Vanilla graph convolution, which proves rather useful in collaborative filtering, however erases multimodal information; 2) Modality forgetting: Multimodal information tends to be gradually forgotten as the recommendation loss essentially facilitates the learning of collaborative information. To this end, we propose a novel approach named STAIR, which employs a novel STepwise grAph convolution to enable a co-existence of collaborative and multimodal Information in e-commerce Recommendation. Besides, it starts with the raw multimodal features as an initialization, and the forgetting problem can be significantly alleviated through constrained embedding updates. As a result, STAIR achieves state-of-the-art recommendation performance on three public e-commerce datasets with minimal computational and memory costs.

Code — https://github.com/yhhe2004/STAIR

Introduction

Recommender systems are playing an increasingly important role nowadays (Chen et al. 2019; El-Kishky et al. 2022) as the explosion of data has far exceeded the human's retrieval capabilities. Unlike traditional data mining tasks (Hoffmann 2007), recommender systems are rapidly growing to adapt to different interaction forms and data types. Consequently multimodal recommendation research (He and McAuley 2015; Wei et al. 2023a) is becoming a popular topic since textual and visual modalities are the main elements that appear in modern media. Nonetheless, their utility usually depends on the relevant recommendation scenario (McElfresh et al. 2022). For example, textual features in news recommendation and visual features in video recommendation can be highly instrumental because they are

indeed the core that attracts users. However, this is often not the case for non-content-driven recommendation scenarios such as e-commerce, where user interaction behaviors may be influenced by a variety of factors beyond modalities.

While numerous approaches (Wei et al. 2019; Zhou et al. 2023; Wei et al. 2023a) have been proposed to mine the multimodal features, the true utility of raw multimodal features in e-commerce datasets has not yet been clarified. We are to ascertain in this scenario whether the user behaviors are driven by a certain type of modality. By clustering features to infer the category of each item, the modal-specific user behavior uncertainty can be estimated via Shannon entropy. The average uncertainty across all users then can be used as an indicator. Unfortunately, we find that the behavior uncertainty under textual or visual modalities is much worse than that under pre-trained embeddings; sometimes even only slightly better than randomly generated Gaussian noise. This suggests that interaction behaviors in these e-commerce datasets are often not driven by textual or visual content. Hence these multimodal features should be used as a complement to, rather than a substitute for, collaborative information. *For this reason, we believe that how to fully utilize both collaborative and multimodal information becomes pivotal.* Unlike previous work that learns embeddings from scratch (Zhang et al. 2021; Zhou and Shen 2023), STAIR starts with multimodal features as an initialization and seeks to progressively combine these two distinct types of information together through some special designs. In practice, we have encountered two additional challenges.

Graph convolutional networks (GCNs) (Kipf and Welling 2016; He et al. 2020) have become the de facto architecture in collaborative filtering, among which LightGCN (He et al. 2020) is a pioneering work known for its simplicity and effectiveness. When these advanced techniques are applied to the modality-initialized embeddings, we, however, observe a severe *modality erasure* phenomenon: as illustrated in Figure 3, the increase in GCN layers leads to the incorporation of more collaborative information; however, this simultaneously results in the rapid erosion of modality information. In particular, unlike the over-smoothing issue (Li, Han, and Wu 2018) that often occurs after stacking multiple convolutional layers, this erasure phenomenon can be observed as soon as the embedding has been convolved. To balance the trade-off between collaborative and multimodal informa-

---

tion, we develop a stepwise convolution mechanism. Note that the generalized LightGCN over the entire embeddings is equivalent to convolving each dimension independently and then concatenating them together. By assigning a varying layer weight along the embedding dimension, the preceding dimensions will contain enough collaborative information, and conversely, the later dimensions will retain more multimodal information.

Both types of information can now co-exist within the embeddings through the proposed stepwise graph convolution. However, it is still inevitable that the multimodal features (incorporated via initialization) will be forgotten during the training process, since the training objective (*e.g.*, BPR loss (Rendle et al. 2009)) is designed to encourage the interacted pairs to be closer in the embedding space, which is actually to promote collaborative information learning. Statistically, we observe from Figure 5 that the Pearson correlation between the embeddings and the modality initialization declines rapidly. Inspired by recent advancements in constrained embedding learning (Xu et al. 2024), we enforce the most modality-similar items to be updated in a consistent manner. This actually performs a backward stepwise convolution over the descent directions, in which the layer weights vary along the embedding dimension in a manner opposite to the forward process. As a result, the later dimensions will receive greater modality enhancement compared to the preceding ones.

The main contributions of this paper can be summarized as follows.

- To the best of our knowledge, we are the first to identify the poor modal-specific user behavior uncertainty in e-commerce scenarios. We also empirically validate that the vanilla graph convolution proves useful in traditional collaborative filtering however tends to erase multimodal information, leading to sub-optimal results.

- We propose STAIR with modality initialization and forward stepwise graph convolution to address the aforementioned problems. Furthermore, the backward stepwise convolution is employed to mitigate the forgetting issue.

- STAIR achieves state-of-the-art recommendation performance over three public e-commerce datasets, with improvements ranging from 2% to 6%. More importantly, these gains can be achieved at only $1/2$ or even $1/100$ of the costs of other multimodal methods.

Related Work

Traditional collaborative filtering   for personalized recommendation primarily retrieves the items based on a user's historical interactions (*e.g.*, clicking and purchasing behaviors). The most classic collaborative filtering approach is the matrix factorization technique (Rendle et al. 2009; Koren 2008), which aims to embed users and items separately into a unified latent space for the subsequent score calculations. With the development of deep learning, NCF (He et al. 2017) and ENMF (Chen et al. 2020) introduce additional trainable modules to further enhance interaction modeling. Note that the interaction data can be exactly represented as a bipartite graph, hence the growing research interest in graph

convolutional networks (GCNs) (Berg, Kipf, and Welling 2017; Wang et al. 2019; Yang et al. 2022). LightGCN (He et al. 2020) is a pioneering work known for its simplicity and effectiveness. Several recent studies (Wu et al. 2019; Shen et al. 2021) have clarified its superiority in incuring embedding smoothness, with some work (Guo et al. 2023) to improve this approach toward manipulating both low-frequency and high-frequency signals. However, previous efforts focusing only on the interaction data struggle to accurately model embeddings for those inactive entities. While the use of multimodal features is helpful in solving this problem, we find that existing GCNs are not sufficient to handle these multimodal features effectively.

Multimodal recommendation   have emerged as a significant area of research (He and McAuley 2015; Wang et al. 2021; Wei et al. 2023b) with the surge of various modalities (*e.g.*, text and image) in media. In scenarios where user behaviors are significantly influenced by modalities (*e.g.*, in news and video recommendations), the raw multimodal features can be readily utilized to complement collaborative signals. MMGCN (Wei et al. 2019) employs a separate GCN for each modality to boost the personalized recommendation of micro-video. In conjunction with cross-modal contrastive learning, MMSSL (Wei et al. 2023a) could achieve the state-of-the-art performance. Nevertheless, the raw features appear to be less beneficial in non-content-driven scenarios like e-commerce. Although some approaches, including LATTICE (Zhang et al. 2021) and the follow-up works (Zhou et al. 2023; Zhou and Shen 2023; Yu et al. 2023; Zhu et al. 2024), have somewhat recognized this problem and proposed some necessary modifications, we *quantitatively* justify for the first time the modal-specific user behavior uncertainty. Furthermore, we also demonstrate that the vanilla graph convolution (He et al. 2020) incuring smoothness have challenges in effectively integrating collaborative and multimodal information, leading to sub-optimal results.

Methodology

In this section, we first introduce the recommendation tasks we are interested in, and the motivation from our experimental observations. Subsequently the challenges and corresponding modifications for the co-existence of collaborative and multimodal information will be presented in detail. Figure 1 shows an overview of STAIR.

Problem definition.   Collaborative filtering (Rendle et al. 2009; He et al. 2020) typically predicts an item $i \in \mathcal{I}$ that the user $u \in \mathcal{U}$ is most likely to prefer based on historical interactions. To quantify how much a user likes an item, they are first embedded into d-dimensional vector representations, followed by the inner product usually serving as the score. The top-N candidate items then can be recommended to the user by ranking all items in descending order based on the scores.

Apart from interaction data, other side information such as multimodal features can be utilized to more adequately model item representations. They complement the lack of
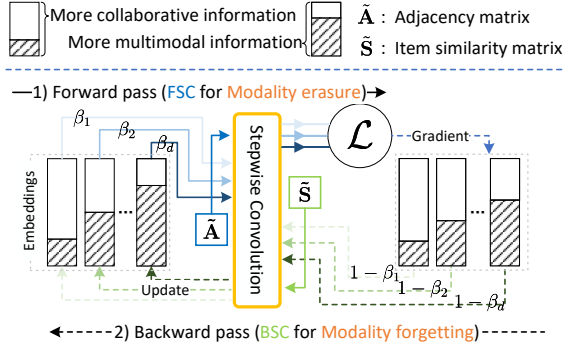
Figure 1: Overview of STAIR. 1) **Forward pass**: More multimodal information are retained in the later dimensions via FSC. 2) **Backward pass**: BSC is used to enhance multimodal information to mitigate modality forgetting problem.
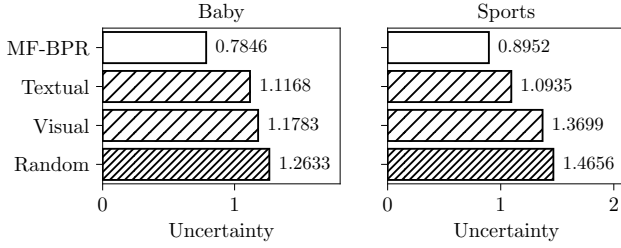

Figure 2: User behavior uncertainty under various features. **MF-BPR**: the embeddings pre-trained via MF-BPR; **Textual/Visual**: the raw multimodal features; **Random**: the randomly generated Gaussian noise.

objective characterization inherent in collaborative signals. For instance, we can infer that two items are similar if they belong to the same brand (from text descriptions) or share the same style (from thumbnails). The emergence of multimodal recommendation is to make full use of both collaborative and multimodal information. Following previous works (Zhang et al. 2021; Zhou and Shen 2023), we focus on textual and visual modalities $\mathcal{M} = \{t, v\}$ with their encoded features $\mathbf{F}^{(m)} \in \mathbb{R}^{|\mathcal{I}| \times d_m}$, $m \in \mathcal{M}$. Besides, let $\mathbf{E}^{(u)} \in \mathbb{R}^{|\mathcal{U}| \times d}$, $\mathbf{E}^{(i)} \in \mathbb{R}^{|\mathcal{I}| \times d}$ respectively represent the user/item ID embeddings, and $\mathbf{E} \in \mathbb{R}^{(|\mathcal{U}|+|\mathcal{I}|) \times d}$ the concatenation of $\mathbf{E}^{(u)}$ and $\mathbf{E}^{(i)}$.

Motivation: Modal-Specific Behavior Uncertainty

Before the multimodal features being extensively utilized for recommender systems (Zhang et al. 2021; Zhou and Shen 2023; Wei et al. 2023a), it is necessary to thoroughly investigate the true utility of these features. Otherwise, these features may adversely affect recommendation performance particularly for e-commerce scenarios.

High user behavior uncertainty in e-commerce. Multimodal features would be instrumental for personalized recommendations if they are dominant factors appealing to users. However, users' purchasing decisions in e-commerce scenarios might be influenced by a variety of factors, not just modality-oriented. If this is true, the multimodal features should be used as a complement to, rather than a substitute for, collaborative information. To figure this out, we employ an uncertainty metric (Wang et al. 2024) to quantify how inconsistent the user behaviors are in commonly used e-commerce datasets. Specifically, we perform K-Means (Arthur and Vassilvitskii 2006) clustering on each modality $\mathbf{F}^{(m)}, m \in \mathcal{M}$ separately, assigning a unique class label to each item. Then the Shannon entropy is used to compute the uncertainty surrounding a user. Intuitively, the behaviors for a user tend to exhibit low modal-specific uncertainty if most of the corresponding interacted items fall into the same category. The average behavior uncertainty across all users then can be used to depict the utility of the modality for this scenario.

Unfortunately, we find the multimodal features are not as attractive in e-commerce datasets (*e.g.*, the widely used Baby and Sports). We illustrate in Figure 2 the respective user behavior uncertainty under textual or visual modalities. For comparison, the embeddings pretrained via MF-BPR and random noise sampled from a standard Gaussian distribution are also included. It makes sense that the pretrained embeddings have the lowest level of uncertainty because the BPR loss encourages user embeddings to align with their neighboring item embeddings. However, the multimodal features, particularly the visual modality, carry behavior uncertainty as poor as that under the random noise. This verifies our assumption about e-commerce scenarios.

Since the user behaviors are rarely driven entirely by modalities, the focus of multimodal recommendation in e-commerce should be on how to fully utilize both collaborative and multimodal information. In this work, we propose a novel approach named STAIR for this purpose. Note that item embeddings in collaborative filtering are often trained from scratch. Despite the inherent impurity of multimodal features, they can still offer a satisfactory initialization superior to random noise. Hence, STAIR starts with multimodal features as an initialization and seeks to progressively combine these two distinct types of information together through some special designs. The first problem is how to compress these dimensionally inconsistent multimodal features into the same space without a well-trained projector. We turn to the whitening technique as proposed in (Su et al. 2021; Huang et al. 2021a), which performs Singular Value Decomposition (SVD) on centralized features and retains the top-d left singular vectors for item embedding initialization. As for the user embedding, it is initialized by mean pooling its neighboring items. Certainly, this straightforward initialization fails to preserve all valuable multimodal information. Subsequent modifications pose additional challenges, which we will introduce next.

Modality Erasure and Graph Convolution

Challenge I: Modality erasure. Graph Convolutional Networks (GCNs) have become the de facto architecture (*e.g.*, LightGCN) in the field of collaborative filtering due to their superiority in feature smoothing (Shen et al. 2021).
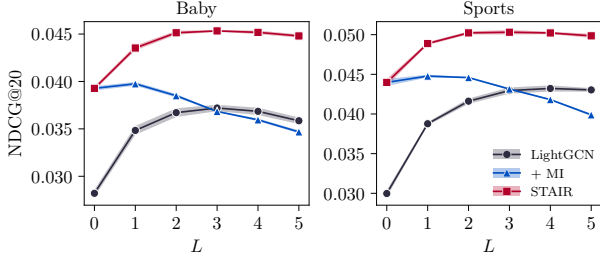
Figure 3: Modality erasure problem. Multimodal information (carried by Modality Initialization (+MI)) will be rapidly erased after a few convolution layers.

A L-layer LightGCN with equal layer weights can be formulate as follows

$$\mathbf{H} = \sum_{l=0}^{L} \frac{1}{L+1} \tilde{\mathbf{A}}^l \mathbf{E},$$

where $\mathbf{A}$ denotes the adjacency matrix and $\tilde{\mathbf{A}}$ is the corresponding matrix after symmetric sqrt normalization. Unfortunately, this technique cannot be directly applied to $\mathbf{E}$ which has incorporated multimodal information. Due to the smoothing nature of LightGCN, the valuable modality information will be rapidly erased as L increases. This can be empirically verified from Figure 3. In contrast to the oversmoothing problem (Li, Han, and Wu 2018) that typically arises when $L > 3$, the modality erasure issue promptly occurs. While both multimodal information from initialization and collaborative information from graph convolution are useful in their own ways, the challenge is how to effectively utilize both types of information, as they seem to conflict with each other. To the best of our knowledge, nobody has explicitly pointed this out, but some approaches such as LATTICE (Zhang et al. 2021) and FREEDOM (Zhou and Shen 2023) inadvertently choose to utilize two separate branches to bypass this problem. In contrast, we develop a soft transition to effortlessly fuse both.

Forward Stepwise Convolution (FSC). In order to handle the aforementioned trade-off, we develop a stepwise graph convolution mechanism. This modification depends on the dimension-independent nature of the generalized LightGCN:

$$\mathbf{H} = \mathop{\|}_{j=1}^{d} \mathbf{H}_{:,j}, \quad \mathbf{H}_{:,j} = \sum_{l=0}^{L} \alpha_{jl} \tilde{\mathbf{A}}^l \mathbf{E}_{:,j}, \qquad (1)$$

where $\mathbf{E}_{:,j}$ stands for the j-th dimension of $\mathbf{E}$ and $\|$ is the concatenation operation. Moreover, in contrast to the shared layer weights, the stepwise graph convolution assigns each dimension $j = 1, \ldots, d$ at layer $l = 0, \ldots, L$ a distinct weight:

$$\alpha_{jl} = \frac{1 - \beta_j}{1 - \beta_j^{L+1}} \beta_j^l, \quad \beta_j = 0.9 \cdot \left[1 - \left(\frac{j-1}{d} \right)^\gamma \right]. \quad (2)$$

Here, γ is the sole hyperparameter that adjusts the weights of each dimension layer. There are three noteworthy designs: **1)** The value of α_{jl} is determined by the *teleport*
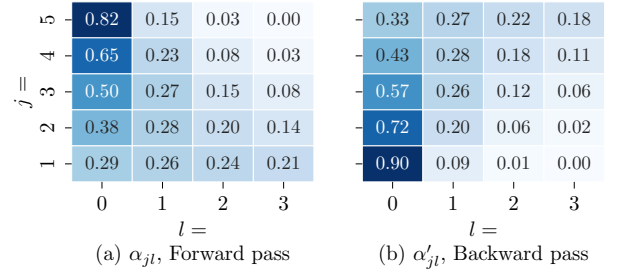

Figure 4: Stepwise layer weights under $\gamma = 1$, $d = 5$, and $L = 3$. (a) α_{jl} for forward pass; (b) α'_{jl} for backward pass.
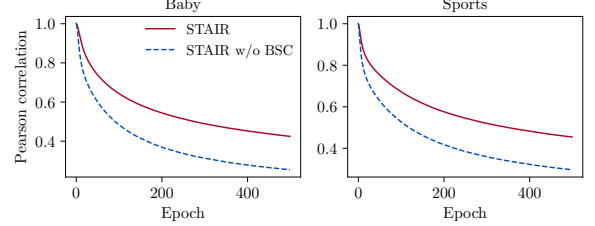

Figure 5: Modality forgetting problem. Pearson correlation coefficients between embeddings and multimodal features are averaged across all dimensions.

ratio $\beta_j \in [0, 1)$. Notice that the stepwise convolution returns the embedding itself if $\beta_j = 0$, and degenerates to a LightGCN as $\beta_j \rightarrow 1$. Note that the formulation of β_j w.r.t j is not unique and other *monotone* functions can be used. The power function is chosen here because of its ease of tuning. **2)** The layer weight α_{jl} diminishes as the layer l increases. A similar mechanism has been proposed in PageRank-inspired GCNs (Gasteiger, Bojchevski, and Günnemann 2018; Huang et al. 2021b) to mitigate oversmoothing. **3)** The normalization coefficient $(1 - \beta_j)/(1 - \beta_j^{L+1})$ is explicitly integrated to ensure that $\sum_{l=0}^{L} \alpha_{jl} = 1$ holds true. Consequently the magnitude across different dimensions remains at a consistent level.

Figure 4a illustrates a toy example. For the preceding dimensions (*e.g.*, $j = 1, 2$), the teleport ratio β_j is large and thus the layer weight is evenly distributed across each layer. The stepwise convolution for these dimensions performs similarly to the traditional LightGCN. In contrast, multimodal information can be adequately preserved for the following dimensions (*e.g.*, $j = 4, 5$), as the majority of layer weights are concentrated in the earlier convolutional layers. *In a nutshell, by adjusting the hyperparameter γ to a smaller (or larger) value, it is possible to contain more multimodal (or collaborative) information.*

Modality Forgetting and Enhancement

Challenge II: Modality forgetting. It is naïve to expect that the modality initialization only could capture all useful information inherent in multimodal features. Note that the item embeddings $\mathbf{E}^{(i)}$ is updated at step t as follows:

$$\mathbf{E}_t^{(i)} \leftarrow \mathbf{E}_{t-1}^{(i)} - \eta \cdot \Delta \mathbf{E}_t^{(i)},$$

where $\Delta \mathbf{E}_t^{(i)}$ denotes the descent direction and η is the step size (*i.e.*, learning rate). For a recommendation training objective (*e.g.*, BPR loss), $\Delta \mathbf{E}_t^{(i)}$ directs the optimization to a region that promotes item embeddings closer to their respective neighboring user embeddings. On the contrary, there is no mechanism to ensure that the initialized multimodal information can be consistently preserved, leading to a potential modality forgetting issue (Hu et al. 2024). This can be empirically validated through the Pearson correlation between trained embeddings and the initial multimodal features. As illustrated in Figure 5, the original STAIR suffers from the forgetting problem during training, and the modification introduced next can significantly mitigate this.

Backward Stepwise Convolution (BSC). Inspired by the recent advancements in constrained embedding learning (Xu et al. 2024), the original descent direction $\Delta \mathbf{E}_t^{(i)}$ is transformed so that the embeddings of items with high modality similarity are updated in a consistent manner. Following (Zhang et al. 2021; Zhou and Shen 2023), the modality similarity between two items is also depicted via the kNN graph extracted from multimodal features. Recall from the Figure 2 that the various multimodal features typically suggest differing user behavior uncertainty. It should identify more neighbors primarily from modalities of higher utility, rather than selecting the same number of neighbors for all modalities as in previous work. Formally, the k_m neighbors for item i of the modality $m \in \mathcal{M}$ are defined as follows

$$\mathcal{N}_m(i) := \text{top-}k_m\left(\left\{\frac{(\boldsymbol{f}_i^{(m)})^T \boldsymbol{f}_j^{(m)}}{\|\boldsymbol{f}_i^{(m)}\|\|\boldsymbol{f}_j^{(m)}\|}\right\}_{j=1}^{|\mathcal{I}|}\right), \quad (3)$$

where $\boldsymbol{f}_i^{(m)} \in \mathbb{R}^{d_m}$ denotes the i-th row vector of $\mathbf{F}^{(m)}$. Subsequently, the overall similarity matrix $\mathbf{S} \in \mathbb{R}^{|\mathcal{I}| \times |\mathcal{I}|}$ is constructed by aggregating the information from the different modalities:

$$S_{ij} = S_{ji} = \max(\hat{S}_{ij}, \hat{S}_{ji}), \ \hat{S}_{ij} = |\{m : j \in \mathcal{N}_m(i)\}|.$$

where $|\cdot|$ is the size of the set. For instance, if item j is adjacent to item i in both the textual and visual modalities, then $S_{ij} = 2$. Note that here we symmetrize the similarity matrix, a necessary step to ensure model convergence.

Finally, the backward stepwise convolution is used to modify the backward optimization process as follows:

$$\mathbf{E}_t^{(i)} \leftarrow \mathbf{E}_{t-1}^{(i)} - \eta \cdot \overset{d}{\underset{j=1}{\|}} \sum_{l=0}^{L} \alpha'_{jl} \tilde{\mathbf{S}}^l [\Delta \mathbf{E}_t^{(i)}]_{:,j}, \quad (4)$$

where $\tilde{\mathbf{S}}$ is the similarity matrix after symmetric sqrt normalization and

$$\alpha'_{jl} = \frac{1 - \beta'_j}{1 - \beta'^{L+1}_j} \beta'^l_j, \ \beta'_j = 1 - \beta_j. \quad (5)$$

Recall that in the forward pass, STAIR requires the preceding dimensions incorporate more collaborative information while the following dimensions preserve more multimodal information. The design of $1 - \beta_j$ in Eq. (5) serves the

Algorithm 1: STAIR training procedures.

Input: multimodal features $\mathbf{F}^{(m)}, m \in \mathcal{M}$;
          embedding matrices $\mathbf{E}^{(u)}, \mathbf{E}^{(i)}$;
Initialize $\mathbf{E}^{(i)}$ by whitening over $\mathbf{F}^{(m)}, m \in \mathcal{M}$;
Initialize $\mathbf{E}^{(u)}$ by mean pooling neighboring items;
Construct kNN similarity graph $\mathbf{S}$ based on Eq. (3);
foreach *step* t **do**
     Compute latent representations $\mathbf{H}$ via FSC;
     Compute BPR loss $\mathcal{L}_{\text{BPR}}$ over $\mathbf{H}$;
     Estimate descent direction at step t;
     Commonly update user embeddings;
     Constrainedly update item embeddings via BSC;

Output: optimized embeddings $\mathbf{E}^{(u)}, \mathbf{E}^{(i)}$.

Dataset	#Users	#Items	#Interactions	Sparsity
Baby	19,445	7,050	160,792	99.88%
Sports	35,598	18,357	296,337	99.95%
Electronics	192,403	63,001	1,689,188	99.99%

Table 1: Dataset statistics

same purpose, ensuring that the later dimensions are more significantly modality-enhanced compared to the preceding dimensions. Figure 4b illustrates how the layer weights vary across the embedding dimension in comparison to those of the forward pass.

The convergence of Eq. (4) with fixed $\beta_j \equiv \beta \in [0, 1)$ has been theoretically proved in (Xu et al. 2024). As a straightforward corollary, the convergence rate of BSC after T updates is no worse than $\mathcal{O}\big(1/((1 - \max_j \beta_j)^2 T)\big)$.

Training and Prediction

Given the latent representations $\boldsymbol{h}_u, \boldsymbol{h}_i \in \mathbb{R}^d$ obtained after the forward stepwise convolution, the inner product can be utilized to calculate the scores. We adopt the commonly used BPR loss as the training objective; that is,

$$\mathcal{L}_{\text{BPR}} = -\sum_{u \in \mathcal{U}} \sum_{i \in \mathcal{N}_u} \sum_{i' \notin \mathcal{N}_u} \log \sigma(\boldsymbol{h}_u^T \boldsymbol{h}_i - \boldsymbol{h}_u^T \boldsymbol{h}_{i'}), \quad (6)$$

where $\mathcal{N}_u$ denotes the interacted items of the user u and $\sigma(\cdot)$ represents the sigmoid function. The overall training algorithm is detailed in Algorithm 1.

Experiments

In this section, we present a thorough empirical evaluation to validate the effectiveness of our proposed approach and answer the following research questions: **RQ1:** Does the proposed STAIR outperform other collaborative filtering methods? **RQ2:** Does STAIR enjoy computational and memory efficiency compared with other multimodal models? **RQ3:** How do the various modules used in STAIR affect the recommendation performance? **RQ4:** How do the key hyperparameters affect the overall performance of STAIR?

	Baby				Sports				Electronics			
	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	0.0417	0.0638	0.0227	0.0284	0.0586	0.0885	0.0324	0.0401	0.0372	0.0557	0.0208	0.0256
LightGCN	0.0543	0.0851	0.0293	0.0372	0.0633	0.0956	0.0349	0.0432	0.0393	0.0579	0.0224	0.0272
JGCF	0.0563	0.0874	0.0305	0.0385	0.0664	0.0999	0.0369	0.0456	0.0425	0.0618	0.0244	0.0294
MMGCN	0.0353	0.0580	0.0189	0.0248	0.0331	0.0529	0.0179	0.0230	0.0217	0.0339	0.0117	0.0149
LATTICE	0.0562	0.0875	0.0308	0.0388	0.0683	0.1024	0.0379	0.0466	-	-	-	-
BM3	0.0559	0.0866	0.0306	0.0385	0.0647	0.0980	0.0358	0.0444	0.0417	0.0631	0.0233	0.0289
FREEDOM	<u>0.0649</u>	<u>0.0991</u>	0.0346	0.0434	<u>0.0715</u>	<u>0.1088</u>	0.0383	0.0479	<u>0.0427</u>	<u>0.0647</u>	<u>0.0239</u>	<u>0.0295</u>
MMSSL	0.0595	0.0929	<u>0.0350</u>	<u>0.0442</u>	0.0667	0.1001	<u>0.0390</u>	<u>0.0482</u>	-	-	-	-
STAIR	**0.0674**	**0.1042**	**0.0359**	**0.0453**	**0.0743**	**0.1117**	**0.0407**	**0.0503**	**0.0440**	**0.0663**	**0.0245**	**0.0302**
▲%	3.85%	5.21%	2.55%	2.61%	3.91%	2.66%	4.36%	4.36%	3.12%	2.53%	2.60%	2.34%
p-value	4.48e-2	4.41e-4	2.43e-2	2.69e-4	2.82e-5	8.29e-3	4.78e-3	2.52e-3	3.52e-4	1.68e-4	2.26e-4	1.93e-4
▲% over LightGCN	24.00%	22.44%	22.31%	21.82%	17.39%	16.83%	16.54%	16.40%	11.92%	14.56%	9.49%	11.30%
▲% over FREEDOM	3.85%	5.21%	3.58%	4.47%	3.91%	2.66%	6.05%	5.01%	3.12%	2.53%	2.60%	2.34%

Table 2: Overall performance comparison. The best results are marked in **bold** and the second-best results are <u>underlined</u>. Paired t-test is performed over 5 independent runs for evaluating p-value (≤ 0.05 indicates statistical significance). Symbol ▲% stands for the relative improvement against the baselines. '-' indicates that the method cannot be performed with a RTX 3090 GPU.

Experiment Setup

Datasets. We consider in this paper three commonly used e-commerce datasets obtained from Amazon reviews, including Baby, Sports, and Electronics. As suggested by (Zhang et al. 2021; Zhou and Shen 2023), we filter out users and items with less than 5 interactions, and Table 1 presents the dataset statistics after preprocessing. Each dataset contains item thumbnails and text descriptions (*e.g.*, title, brand). Following (Zhou et al. 2023), the 4,096-dimensional visual features published in (Ni, Li, and McAuley 2019), and the 384-dimensional sentence embeddings published in (Zhou 2023) are used for experiments.

Baselines. To validate the effectiveness of STAIR, we compare it with several recommendation models. Firstly, MF-BPR (Rendle et al. 2009) and two GCN-based methods, LightGCN (He et al. 2020) and JGCF (Guo et al. 2023), are chosen as benchmarks in terms of the traditional collaborative filtering that focuses on only the interaction data. Besides, we consider five multimodal collaborative filtering methods, of which LATTICE (Zhang et al. 2021) and FREEDOM (Zhou and Shen 2023) are the two most similar efforts that also use kNN graphs, while BM3 (Zhou et al. 2023) and MMSSL (Wei et al. 2023a) are the two focusing on contrastive learning.

Evaluation. We retrieve the top-N ($N = 10, 20$) candidate item list by ranking the scores in descending order. Subsequently, two widely used metrics, Recall and Normalized Discounted Cumulative Gain (NDCG), are employed to evaluate the quality of the recommendations. The former quantifies the proportion of relevant items within the top-N recommended candidates to all relevant items present in the test set, while the latter considers the ranking positions, attributing higher scores to more prioritized items. We refer to the average of these metrics across all users as Recall@N (or R@N) and NDCG@N (or N@N), respectively.

	Time/Epoch (Second)			GPU Memory (MB)		
	Baby	Sports	Elec∼	Baby	Sports	Elec∼
MF-BPR	1.18s	1.79s	8.33s	468M	664M	1,494M
LightGCN	1.25s	1.99s	9.08s	478M	684M	1,676M
JGCF	1.31s	1.99s	11.30s	510M	742M	2,060M
MMGCN	2.99s	7.84s	47.15s	1,266M	2,142M	8,530M
LATTICE	2.11s	11.68s	-	1,664M	5,928M	-
BM3	1.52s	3.23s	20.81s	1,032M	2,088M	6,464M
FREEDOM	1.68s	3.45s	19.41s	1,034M	2,096M	6,484M
MMSSL	27.70s	156.80s	-	3,048M	10,656M	-
STAIR	1.45s	2.65s	10.23s	490M	696M	1,738M

Table 3: Computational and memory costs.

Implementation Details. For fairness (Zhang et al. 2021; Zhou and Shen 2023), we fix the embedding dimension to 64 and the number of convolutional layers to 3 for both FSC and BSC processes. As suggested in (Xu et al. 2024), AdamW is employed as the optimizer for training STAIR, whose learning rate is searched from {1e-4, 5e-4, 1e-3, 5e-3} and the weight decay in the range of [0, 1]. The exponent γ that controls the changing rate of the layer weights can be chosen from {0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5}. In addition, we adjust the number of neighbors k_m from {1, 3, 5, 10, 20} for each modality $m \in \mathcal{M}$ separately. For the baselines, we determined the optimal learning rate and weight decay in a manner analogous to STAIR, and searched for other hyper-parameters following the suggestions of the original papers. For all methods, we report the results on the best checkpoint identified by the validation NDCG@20 metric over 500 epochs.

Overall Performance Comparison (RQ1, RQ2)

We respectively present the overall performance and efficiency comparison in Table 2 and 3, from which we have the following key observations:

- In traditional collaborative filtering approaches, the performance of personalized recommendation continues to improve as the utilization of collaborative information (within interaction data) becomes more and more adequate. In particular, GCNs play a crucial role in this progress. However, analogous conclusions cannot be readily inferred from e-commerce multimodal recommendations: MMGCN performs graph convolution separately on each modality but fails to achieve satisfactory results, even inferior to MF-BPR that does not use multimodal features. This corroborate our argument that how to fully utilize both types of information becomes pivotal in e-commerce multimodal recommendation.

- LATTICE and FREEDOM successfully circumvent this dilemma via two separate branches, achieving significantly better results than LightGCN. However, this approach hinders an optimal trade-off between collaborative and multimodal information, often resulting in too much reliance on one side. In contrast, STAIR employs a much soft transition to effortlessly fuse both. The stepwise convolution allows for the co-existence of collaborative and multimodal features along the dimension. Consequently, STAIR not only surpasses LightGCN by at least 10%, but also achieves performance improvements ranging from 2% to 6% when compared to FREEDOM. Moreover, as shown in Table 3, either the learnable similarity matrix of LATTICE or the particular sampling strategy employed by FREEDOM incur non-negligible costs, whereas the cost of STAIR is only on par with that of LightGCN.

- In contrast to BM3 and MMSSL which emphasize modality learning primarily through contrastive learning, the multimodal denoising and enhancement in STAIR are accomplished through a kNN graph and the BSC. While the former is also capable of mitigating the modality forgetting problem; BSC is arguably better adapted to STAIR as it can optimize the embeddings in line with the forward pass. In addition, we have found that the data augmentation and subsequent computation required by MMSSL can be quite expensive. When dealing with larger datasets such as Electronics, the computational and memory requirements make MMSSL impossible to implement in real-world recommendations. In contrast, STAIR has shown to be significantly more efficient.

Ablation Study (RQ3)

We show the ablation experiments with LightGCN as the baseline in Figure 6. First, modality initialization (MI) not only contributes to the final recommended performance, but also improves the robustness to various random seeds. Second, the usefulness of the forward stepwise convolution (FSC) differs in various datasets. This can be attributed to the observation that user behavior uncertainty on Sports is lower than that on Baby, leading to a less severe modality erasure issue. Finally, the backward stepwise convolution (BSC) is essential due to its capability in adaptive modality enhancement to mitigate the forgetting problem.
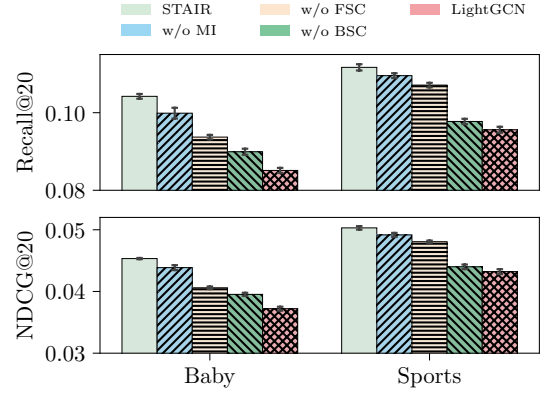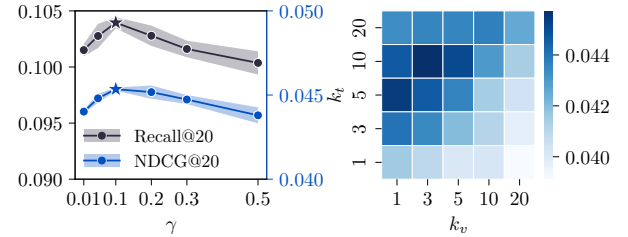

Figure 6: STAIR ablation experiments.

Figure 7: Impact of γ and k_m on the Baby dataset.

Hyperparameter Sensitivity Analysis (RQ4)

Figure 7 shows the hyperparameter sensitivity on the Baby dataset. **1)** According to Eq. (2), a larger γ implies more emphasis on collaborative information. It seems that more multimodal information should be retained during the convolution process, as a slightly lower value of γ performs better. This is because partial collaborative information can be directly learned through the recommendation loss, rather than solely depending on the convolution. **2)** It can be observed that the textual modality is more effective in reflecting users' preferences compared to the visual modality. This is reasonable since the textual features are encoded based on the less noisy attributes such as title, brand, and category. This also provides justification for why previous efforts often assign a small weight to the visual modality (Zhou and Shen 2023).

Conclusion and Future Work

We first identify the poor modal-specific user behavior uncertainty in e-commerce scenarios, and propose modality initialization, forward and backward stepwise graph convolutions for the co-existence of collaborative and multimodal information. Extensive experiments validate the superiority of STAIR in recommendation accuracy and efficiency.

Note that although STAIR achieves state-of-the-art performance in e-commerce multimodal recommendation, it may not fully mine the raw multimodal features in content-driven scenarios such as news and video recommendation (Wu et al. 2020). A promising direction in the future is to boost STAIR to benefit from raw multimodal features adaptively.

Acknowledgments

References

Arthur, D.; and Vassilvitskii, S. 2006. k-means++: The advantages of careful seeding. Technical report, Stanford.

Berg, R. v. d.; Kipf, T. N.; and Welling, M. 2017. Graph convolutional matrix completion. *arXiv preprint arXiv:1706.02263*.

Chen, C.; Zhang, M.; Zhang, Y.; Liu, Y.; and Ma, S. 2020. Efficient neural matrix factorization without sampling for recommendation. *ACM Transactions on Information Systems (TOIS)*, 38(2): 1–28.

Chen, Q.; Zhao, H.; Li, W.; Huang, P.; and Ou, W. 2019. Behavior sequence transformer for e-commerce recommendation in alibaba. In *International Workshop on Deep Learning Practice for High-dimensional Sparse Data*, 1–4.

El-Kishky, A.; Markovich, T.; Park, S.; Verma, C.; Kim, B.; Eskander, R.; Malkov, Y.; Portman, F.; Samaniego, S.; Xiao, Y.; et al. 2022. Twhin: Embedding the twitter heterogeneous information network for personalized recommendation. In *International Conference on Knowledge Discovery & Data Mining (KDD)*, 2842–2850.

Gasteiger, J.; Bojchevski, A.; and Günnemann, S. 2018. Predict then propagate: Graph neural networks meet personalized pagerank. In *International Conference on Learning Representations (ICLR)*.

Guo, J.; Du, L.; Chen, X.; Ma, X.; Fu, Q.; Han, S.; Zhang, D.; and Zhang, Y. 2023. On Manipulating Signals of User-Item Graph: A Jacobi Polynomial-based Graph Collaborative Filtering. In *International Conference on Knowledge Discovery & Data Mining (KDD)*, 602–613.

He, R.; and McAuley, J. 2015. VBPR: Visual bayesian personalized ranking from implicit feedback.

He, X.; Deng, K.; Wang, X.; Li, Y.; Zhang, Y.; and Wang, M. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *International Conference on Research and Development in Information Retrieval (SIGIR)*, 639–648.

He, X.; Liao, L.; Zhang, H.; Nie, L.; Hu, X.; and Chua, T.-S. 2017. Neural collaborative filtering. In *International Conference on World Wide Web (WWW)*, 173–182.

Hoffmann, H. 2007. Kernel PCA for novelty detection. *Pattern Recognition (PR)*, 40(3): 863–874.

Hu, H.; Liu, Q.; Li, C.; and Kan, M.-Y. 2024. Lightweight modality adaptation to sequential recommendation via correlation supervision. In *European Conference on Information Retrieval (ECIR)*, 123–139.

Huang, J.; Tang, D.; Zhong, W.; Lu, S.; Shou, L.; Gong, M.; Jiang, D.; and Duan, N. 2021a. WhiteningBERT: An easy unsupervised sentence embedding approach. In *Findings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 238–244.

Huang, Q.; He, H.; Singh, A.; Lim, S.-N.; and Benson, A. 2021b. Combining label propagation and simple models out-performs graph neural networks. In *International Conference on Learning Representations (ICLR)*.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*.

Koren, Y. 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *International Conference on Knowledge Discovery & Data Mining (KDD)*, 426–434.

Li, Q.; Han, Z.; and Wu, X.-M. 2018. Deeper insights into graph convolutional networks for semi-supervised learning. In *Conference on Artificial Intelligence (AAAI)*.

McElfresh, D.; Khandagale, S.; Valverde, J.; Dickerson, J.; and White, C. 2022. On the generalizability and predictability of recommender systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 4416–4432.

Ni, J.; Li, J.; and McAuley, J. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 188–197.

Rendle, S.; Freudenthaler, C.; Gantner, Z.; and Schmidt-Thieme, L. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Uncertainty in Artificial Intelligence (UAI)*.

Shen, Y.; Wu, Y.; Zhang, Y.; Shan, C.; Zhang, J.; Letaief, B. K.; and Li, D. 2021. How powerful is graph convolution for recommendation? In *International Conference on Information & Knowledge Management (CIKM)*, 1619–1629.

Su, J.; Cao, J.; Liu, W.; and Ou, Y. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

Wang, F.; Liu, Y.; Liu, K.; Wang, Y.; Medya, S.; and Yu, P. S. 2024. Uncertainty in graph neural networks: A survey. *arXiv preprint arXiv:2403.07185*.

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*, 25: 1074–1084.

Wang, X.; He, X.; Wang, M.; Feng, F.; and Chua, T.-S. 2019. Neural Graph Collaborative Filtering. In *International Conference on Research and Development in Information Retrieval (SIGIR)*, SIGIR '19.

Wei, W.; Huang, C.; Xia, L.; and Zhang, C. 2023a. Multimodal self-supervised learning for recommendation. In *Proceedings of the ACM Web Conference (WWW)*, 790–800.

Wei, Y.; Liu, W.; Liu, F.; Wang, X.; Nie, L.; and Chua, T.-S. 2023b. Lightgt: A light graph transformer for multimedia recommendation. In *International Conference on Research and Development in Information Retrieval (SIGIR)*, 1508–1517.

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 *International Conference on Multimedia (MM)*, 1437–1445.

Wu, F.; Qiao, Y.; Chen, J.-H.; Wu, C.; Qi, T.; Lian, J.; Liu, D.; Xie, X.; Gao, J.; Wu, W.; et al. 2020. Mind: A large-scale dataset for news recommendation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 3597–3606.

Wu, F.; Souza, A.; Zhang, T.; Fifty, C.; Yu, T.; and Weinberger, K. 2019. Simplifying graph convolutional networks. In *International Conference on Machine Learning (ICML)*, 6861–6871.

Xu, C.; Wang, J.; Wang, J.; and Zhang, W. 2024. Graph-enhanced optimizers for structure-aware recommendation embedding evolution. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yang, M.; Li, Z.; Zhou, M.; Liu, J.; and King, I. 2022. Hicf: Hyperbolic informative collaborative filtering. In *Conference on Knowledge Discovery & Data Mining (KDD)*, 2212–2221.

Yu, P.; Tan, Z.; Lu, G.; and Bao, B.-K. 2023. Multi-view graph convolutional network for multimedia recommendation. In *International Conference on Multimedia (MM)*, 6576–6585.

Zhang, J.; Zhu, Y.; Liu, Q.; Wu, S.; Wang, S.; and Wang, L. 2021. Mining latent structures for multimedia recommendation. In *International Conference on Multimedia (MM)*, 3872–3880.

Zhou, X. 2023. MMRec: Simplifying multimodal recommendation. *arXiv preprint arXiv:2302.03497*.

Zhou, X.; and Shen, Z. 2023. A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation. In *International Conference on Multimedia (MM)*, 935–943.

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

Zhu, X.; Jin, M.; Zhang, H.; Meng, C.; Zhang, D.; and Li, X. 2024. Modeling Domains as Distributions with Uncertainty for Cross-Domain Recommendation. In *International Conference on Research and Development in Information Retrieval (SIGIR)*, 2517–2521.