Self-Supervised Adaptive Hypergraph Neural Network for Incomplete Multimodal Recommendation

Xinjie Chen*(Corresponding author), Binghang Yu , Yiheng Lou
Zhejiang Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University, Jinhua, 321004, Zhejiang, China
* E-mail of the corresponding author: chenxinjie217@gmail.com

Abstract

With the explosive growth of multimedia information, multimodal recommendation systems play a crucial role in mitigating information overload. However, a majority of existing models operate under the idealized assumption of complete modal information, largely overlooking the prevalent issue of incomplete modalities. Furthermore, their reliance on static, predefined graph structures and sparse interaction labels limits their robustness and generalization capability. To address these shortcomings, this paper proposes SAHRec, a novel **S**elf-supervised **A**daptive **H**ypergraph **Rec**ommendation framework. SAHRec introduces a data-driven dynamic learning paradigm that jointly optimizes high-order structures and node representations in an end-to-end fashion. The core of SAHRec consists of two major innovations. First, we design a differentiable hypergraph learner that adaptively constructs optimal high-order topological structures from data, moving beyond the limitations of static or heuristic-based methods. This allows the model to capture more precise and task-relevant global dependencies. Second, we introduce a modality-aware contrastive learning task as a powerful self-supervised signal. By aligning node representations derived from local and global structural views, the model is compelled to learn consistent and robust features even in an incomplete information environment, which significantly enhances its generalization. Extensive experiments conducted on three large-scale public multimodal datasets demonstrate that SAHRec, especially under extreme modality absence of up to 90%, significantly outperforms a wide range of state-of-the-art baseline methods, including strong static hypergraph models. This fully validates the effectiveness and superior robustness of our proposed approach in handling the challenging incomplete multimodal recommendation task.

Keywords: Multimodal Recommendation, Incomplete Modalities, Hypergraph Neural Network, Self-supervised Learning
DOI: 10.7176/CEIS/17-1-04
Publication date: February 28th, 2026

1. Introduction

The rapid expansion of e-commerce and multimedia sharing platforms, such as Amazon and TikTok, has resulted in an overwhelming volume of available items and content. While these platforms offer users unprecedented choices, they also create a severe "information overload" problem, where users struggle to find content that aligns with their personal interests. Multimodal recommendation systems have emerged as a core technology to address this challenge, enhancing recommendation effectiveness by leveraging rich, multimodal content (e.g., visual and textual features) to better match user preferences (Guo et al. 2024).

However, a majority of existing multimodal recommendation models operate under an idealized assumption: that all items possess complete modal information. This rarely holds in practical scenarios (Malitesta et al. 2024). In reality, some items may lack high-quality images, detailed textual descriptions, or both. This prevalent issue of incomplete modalities significantly undermines the robustness and reliability of recommendation systems. Furthermore, conventional models, including those based on graph neural networks (GNNs), are often confined to modeling pairwise interactions between users and items (He, Deng, et al. 2020). This approach fails to adequately capture the complex high-order relationships latent in the data, such as user communities with shared tastes or item clusters with similar stylistic attributes.

To mitigate these issues, hypergraph neural networks (HGNNs) have shown great promise by naturally modeling high-order relationships (Feng et al. 2019). Nevertheless, early attempts to apply HGNNs to recommendation still face two critical limitations. Firstly, the hypergraph structure is often constructed statically using predefined, heuristic rules (e.g., k-NN clustering), which may be suboptimal for the downstream recommendation task. A more desirable paradigm would be to learn the hypergraph structure dynamically and in a data-driven manner (Wei, Liang, et al. 2022). Secondly, these models heavily rely on sparse user-item interaction labels for supervision, making them vulnerable in data-scarce environments.

To address these shortcomings, this paper proposes a novel Self-supervised Adaptive Hypergraph Recommendation framework, SAHRec, designed specifically for robust recommendation under modality absence. SAHRec introduces a data-driven dynamic learning paradigm that jointly optimizes high-order structures and node representations. The framework features a differentiable hypergraph learner to adaptively construct optimal graph topologies and a modality-aware contrastive learning task that provides powerful self-supervised signals. By compelling the model to learn consistent and robust features from both local and global structural views, SAHRec significantly enhances its generalization capability in incomplete information environments. Therefore, it is worth an attempt to incorporate a novel methodology, such as SAHRec, into the state of the art of incomplete multimodal recommendation.

2. Literature review

2.1 Multimodal Recommendation

Multimodal Recommendation Systems (MRSs) have become a cornerstone of modern online platforms, aiming to alleviate information overload by integrating diverse data modalities such as images, text, and audio (Guo et al. 2024). In contrast to traditional Collaborative Filtering (CF), which relies solely on user-item interaction IDs, MRSs construct richer and more comprehensive representations of users and items. This process allows the model to capture the deep semantic attributes of items and nuanced user preferences, thereby significantly enhancing recommendation accuracy and interpretability.

The evolution of MRSs has been marked by the progressive integration of more sophisticated modeling techniques. Early works, such as Visual Bayesian Personalized Ranking (VBPR), extended classical matrix factorization by incorporating pre-trained visual features, enabling the model to capture users' aesthetic preferences (He & McAuley 2016). Subsequently, with the rise of deep learning, attention mechanisms were introduced to dynamically weigh the importance of different modalities or features, allowing for more fine-grained and context-aware user interest modeling (Chen et al. 2017).

However, a critical challenge that per-sists in this domain is the issue of incomplete modalities. The majority of MRSs operate under the idealized assumption that all modal information for every item is complete and available. In real-world scenarios, this is rarely the case, as items frequently lack high-quality images or detailed descriptions. This data incompleteness severely degrades the performance and robustness of models that depend on complete information, posing a significant bottleneck for their practical deployment (Malitesta et al. 2024). Therefore, developing recommendation frameworks that are inherently robust to missing modalities is a crucial and timely research direction.

2.2 Graph Neural Networks for Recommendation

Graph Neural Networks (GNNs) have revolutionized the field of recommender systems by providing a powerful paradigm for modeling the relational structure inherent in user-item interaction data (Scarselli et al. 2008). The core idea of GNN-based recommendation is to represent the user-item interaction history as a bipartite graph and learn node (user and item) embeddings through a message-passing mechanism. In this process, each node iteratively aggregates feature information from its neighbors, effectively encoding high-order connectivity and collaborative signals into the learned representations. Models like LightGCN have demonstrated the remarkable effectiveness of this approach by simplifying the propagation process, establishing a strong baseline for CF (He, Deng, et al. 2020).

Despite their success, conventional GNNs are fundamentally limited by the structure of standard graphs, where an edge can only connect two nodes. This means they are inherently designed to model only pairwise relationships. However, many complex and valuable interactions in recommender systems are high-order in nature, such as a group of users purchasing the same item in a single transaction or a user browsing a set of related items in one session. To address this limitation, Hypergraph Neural Networks (HGNNs) have been proposed as a more general and powerful alternative (Feng et al. 2019). A hyperedge in a hypergraph can connect an arbitrary number of nodes, making it a natural tool for explicitly modeling these high-order relationships.

Recent works have begun to explore the application of HGNNs in recommendation, yet most still rely on static, heuristically defined hypergraph structures. This pre-defined topology may not be optimal for the downstream recommendation task and can be particularly vulnerable to noise when constructed from incomplete modal features. This limitation highlights the need for a mechanism that can dynamically and adaptively learn the hypergraph structure in a data-driven manner, which is a central theme of this thesis.

2.3 Self-supervised Learning for Recommendation

Self-supervised Learning (SSL) has recently emerged as a powerful paradigm to address the data sparsity and robustness issues in recommender systems. Instead of relying solely on explicit user-item interaction labels, SSL generates supervisory signals from the data itself, typically by creating augmented "views" of the data and training the model to learn representations that are invariant to these augmentations (Tao et al. 2022). Contrastive learning is a dominant approach within SSL, where the objective is to pull representations of positive pairs (e.g., two augmented views of the same node) closer together in the embedding space while pushing representations of negative pairs apart.

In the context of multimodal recommendation, SSL offers a promising way to enhance model robustness, especially under modality absence. For example, the representations of an item from its visual and textual modalities can be treated as a positive pair, compelling the model to learn modality-invariant features. By aligning representations from different modalities or structural views, the model can learn to "impute" missing information from available sources and capture the essential, shared semantics of an item (Wei, Huang, et al. 2023).

However, the uncritical application of SSL can also present challenges. The effectiveness of contrastive learning heavily depends on the quality of view augmentation and the selection of positive/negative pairs. In the complex scenario of incomplete multimodal data, designing an effective SSL task that aligns representations from both local (pairwise) and global (high-order) structures, while being aware of the reliability of different modalities, remains an open and challenging problem. This thesis conceptualizes a novel modality-aware contrastive learning task as a key mechanism to guide the learning of robust representations. We argue that a well-designed self-supervised objective, deeply integrated with a dynamic hypergraph framework, can elevate the learning process from a passive, task-driven procedure to an active, reflective partnership, thereby creating the necessary cognitive conditions for robust recommendation to occur.

3. Methodology

3.1 Overview

To address the challenges of incomplete modalities and the limitations of static, pairwise modeling, we propose a novel Self-supervised Adaptive Hypergraph Recommendation framework, named SAHRec. The core idea of SAHRec is to decouple the learning of local and global user interests, while introducing dynamic structural learning and self-supervised signals to enhance robustness and generalization. As illustrated in Figure 1 , our framework comprises three main components:

- Local Graph Embedding (LGE) Module: Operating on the fundamental user-item bipartite graph, this module is designed to independently learn decoupled local representations of users and items, capturing both collaborative and modality-specific signals from the local neighborhood.
- Adaptive Global Hypergraph (AGH) Module: This is a key innovation of our work. It introduces a differentiable hypergraph learner that moves beyond heuristic rules to adaptively and end-to-end learn the optimal hypergraph structure that captures global, high-order dependencies from the data.
- Fusion and Alignment Module: This module serves to unify the local and global representations. It employs a modality-aware contrastive learning task to align the multi-view embeddings, providing a powerful self-supervised signal, and finally fuses them for the ultimate recommendation task.

In the following sections, we will elaborate on the technical details of each component.

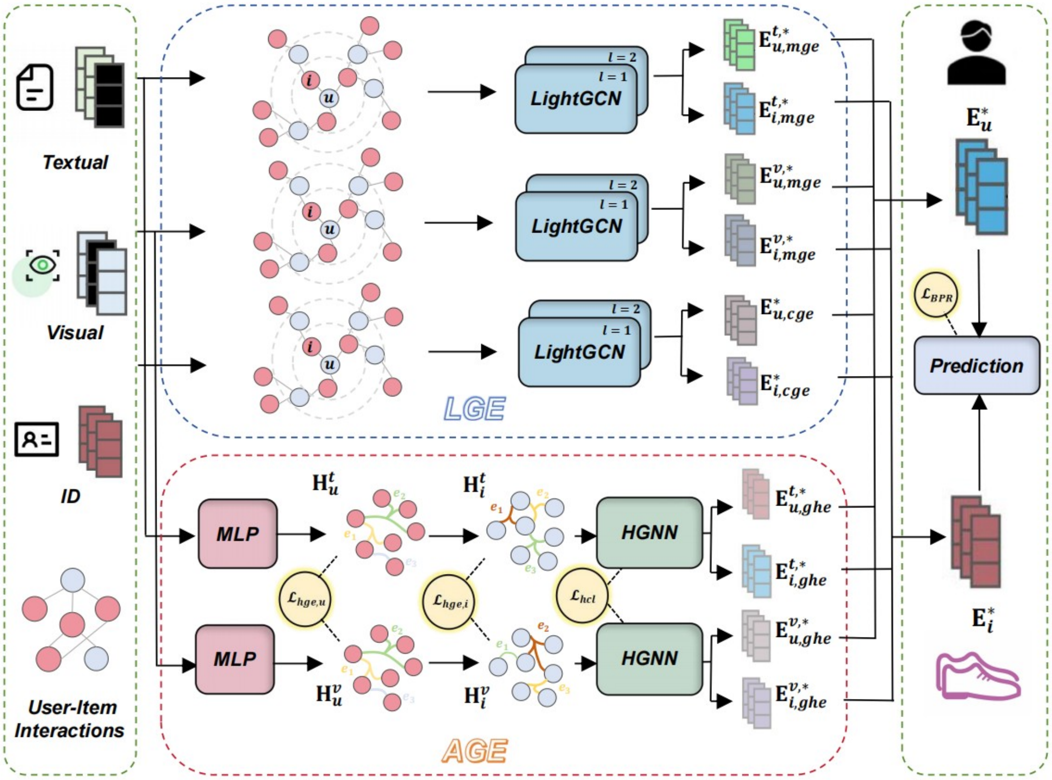


Figure 1 Self-supervised Adaptive Hypergraph Recommendation framework

3.2 Local Graph Embedding (LGE) Module

The LGE module aims to capture user preferences and item attributes reflected by the local topological structure of the user-item interaction graph. Inspired by the idea of disentangled representation learning (Guo et al. 2024), we design separate propagation channels for collaborative and modal signals. This disentangled design is crucial for handling incomplete modalities, as it prevents noisy or missing modal information from directly corrupting the pure collaborative signals. The LGE module consists of two parallel sub-modules: Collaborative Graph Embedding (CGE) and Modality Graph Embedding (MGE).

Collaborative Graph Embedding (CGE). The CGE sub-module is responsible for learning modality-agnostic, pure collaborative filtering signals from user-item interactions. It operates directly on the trainable ID embeddings of users and items. We adopt the lightweight message-passing paradigm of LightGCN (He, Deng, et al. 2020). The embedding update rule for all nodes at layer l is defined as:

$$\mathrm{E}^{(l+1)}_{\mathrm{cge}} - (\mathrm{D}^{-\frac{1}{2}}\mathrm{A}\mathrm{D}^{-\frac{1}{2}})\mathrm{E}^{(l)}_{\mathrm{cge}} \quad (1)$$

where $\mathbf{E}^{(0)}_{\mathrm{cge}}$ is the initial user and item ID embedding matrix. $\mathbf{A}$ is the adjacency matrix of the user-item bipartite graph with self-loops, and $\mathbf{D}$ is its corresponding degree matrix. After $\mathbf{L}$ layers of propagation, we aggregate the embeddings from all layers (e.g., via averaging) to obtain the final local collaborative embeddings $\mathbf{E}^{*}_{\mathrm{cge}}$, which are rich in high-order collaborative signals.

Modality Graph Embedding (MGE). The MGE sub-module focuses on learning modality-specific user preferences on the same interaction graph structure. To achieve disentanglement from collaborative signals, it uses independent modal features as input. For each modality m, we first project the raw item features $\mathbf{X}^{\mathrm{m}}$ into a unified d-dimensional space via a learnable transformation matrix $\mathbf{W}^{\mathrm{m}}$. The initial modal embeddings for users are derived by aggregating the features of their interacted items. These user and item modal embeddings, denoted as $\mathbf{E}^{\mathrm{m},(0)}_{\mathrm{mge}}$, are then fed into an identical lightweight graph convolution process as in CGE to learn the final local modal embeddings $\mathbf{E}_{\mathrm{mge}}^{\mathrm{m},*}$. This design ensures that even if one modality is severely missing, the model can still rely on the pure collaborative signals and other available modalities for effective local interest modeling.

3.3 Adaptive Global Hypergraph (AGH) Module

This module is designed to capture global, modality-aware dependencies that transcend the local neighborhood. Its centerpiece is a differentiable hypergraph learner that adaptively constructs the hypergraph structure in an end-to-end manner.

Differentiable Hypergraph Learner. Inspired by dynamic hypergraph learning (Wei, Liang, et al. 2022), we employ a lightweight Multi-Layer Perceptron (MLP) to learn a "soft" hypergraph incidence matrix $\mathbf{H}^{\mathrm{m}}$ for each modality. For the item-side hypergraph, the learner takes the projected modal features $\widehat{\mathbf{X}}^{m}$ as input and outputs a probability distribution for each item over K^{m} latent hyperedges:

$$\mathbf{H}_{I}^{m} = Softmax(ReLU(\widehat{\mathbf{X}}^{m}\mathbf{W}_{I,1}^{m})\mathbf{W}_{I,2}^{m}) \tag{2}$$

This soft-assignment mechanism is fully differentiable, allowing the hypergraph structure to be optimized jointly with the recommendation task. The user-side hypergraph incidence matrix $\mathbf{H}_{U}^{m}$ is then derived via interaction mapping: $\mathbf{H}_{U}^{m} = \mathbf{R}\mathbf{H}_{I}^{m}$.

Regularization for Hypergraph Structure Learning. Learning the hypergraph structure solely from the downstream recommendation loss is challenging and prone to overfitting. To guide this process, we introduce a novel hypergraph learning loss $\mathcal{L}_{hge}$, which enforces that the learned structure should preserve the collaborative signals from the original interaction graph. Following the minCUT principle (Stoer & Wagner 1994), the loss is formulated as:

$$\mathcal{L}_{hge} = \sum_{m\in\mathcal{M}}(-Tr((\mathbf{H}_{U}^{m})^{T}\mathbf{S}_{U}\mathbf{H}_{U}^{m}) + ||(\mathbf{H}_{U}^{m})^{T}\mathbf{H}_{U}^{m}||_{F}^{2}) + \sum_{m\in\mathcal{M}}(-Tr((\mathbf{H}_{I}^{m})^{T}\mathbf{S}_{I}\mathbf{H}_{I}^{m}) + ||(\mathbf{H}_{I}^{m})^{T}\mathbf{H}_{I}^{m}||_{F}^{2}) \tag{3}$$

where $\mathbf{S}_{U} = \mathbf{R}\mathbf{R}^{T}$ is the user co-interaction matrix. The first term encourages similar users to be clustered into the same hyperedge, while the second term ensures structural balance by penalizing oversized hyperedges.

3.4 Fusion and Alignment Module

After obtaining decoupled local and global representations, this module aligns them using a self-supervised signal and fuses them for final prediction.

Cross-Modal Hypergraph Contrastive Learning. We posit that for a given user, their global preferences learned from different modalities should be semantically consistent. Based on this intuition, we design a cross-modal contrastive learning task. For each user u, we treat their global embeddings from the visual and textual modalities, $\mathbf{e}_{\mathrm{u,ghe}}^{\mathrm{v},*}$ and $\mathbf{e}_{\mathrm{u,ghe}}^{\mathrm{t},*}$, as a positive pair. The global embeddings of any other user serve as negative samples. We then employ the InfoNCE loss (Oord, Li, & Vinyals 2018) to maximize the agreement between positive pairs:

$$\mathcal{L}_{hcl} = \sum_{u\in\mathcal{U}} - \log\frac{\exp(s(\mathbf{e}_{u,ghe}^{v,*}, \mathbf{e}_{u,ghe}^{t,*})/\tau)}{\sum_{u\prime\in\mathcal{U}}\exp(s(\mathbf{e}_{u,ghe}^{v,*}, \mathbf{e}_{u\prime,ghe}^{t,*})/\tau)} \tag{4}$$

where $\mathbf{s}(\cdot,\cdot)$ is the cosine similarity and τ is a temperature parameter. This self-supervised task compels the model to learn modality-invariant global representations, enhancing its robustness against modality absence.

Multi-View Representation Fusion and Optimization. Finally, we fuse the local and global embeddings via a weighted summation to obtain the final user and item representations, $\mathbf{e}_{\mathrm{u}}^{*}$ and $\mathbf{e}_{\mathrm{i}}^{*}$. The overall model is trained end-to-end by jointly optimizing a multi-task objective function:

$$\mathcal{L} = \mathcal{L}_{bpr} + \lambda_{1}\mathcal{L}_{hge} + \lambda_{2}\mathcal{L}_{hcl} \tag{5}$$

where $\mathcal{L}_{bpr}$ is the standard Bayesian Personalized Ranking (BPR) loss, and λ_{1}, λ_{2} are hyperparameters balancing the contributions of the hypergraph learning and contrastive learning tasks.

4. Experiment

In this section, we conduct extensive experiments on three public benchmark datasets to answer the following research questions:

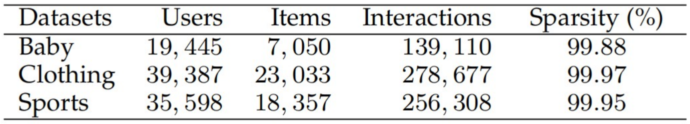
- RQ1: How does SAHRec perform compared to state-of-the-art multimodal recommendation baselines, especially under severe modality incompleteness? Holons receive instruction from and, to a certain extent, be controlled by higher level holons. The subordination to higher level holons ensures the effective operation of the larger whole.

- RQ2: How do the key components of SAHRec, particularly the adaptive hypergraph learning and the self-supervised alignment mechanism, contribute to the model's overall performance?
- RQ3: How sensitive is SAHRec to its main hyperparameters?

4.1 Experimental Settings

Datasets. To ensure a fair and comprehensive comparison, we conduct all experiments on the same three large-scale, real-world Amazon review datasets used in the previous chapter: **Baby**, **Clothing**, and **Sports**. As previously described, these datasets provide both visual and textual modalities for each item and have been preprocessed using a 5-core filtering setting. The detailed statistics of the datasets are identical to those presented in Table 1. The 384-dimensional textual features and 4096-dimensional visual features are extracted using pre-trained Sentence-BERT (Reimers 2019) and CNN (Deng 2009) models, respectively.

Table 2 Statistics of the tested datasets



Datasets	Users	Items	Interactions	Sparsity (%)
Baby	19, 445	7, 050	139, 110	99.88
Clothing	39, 387	23, 033	278, 677	99.97
Sports	35, 598	18, 357	256, 308	99.95

Evaluation Metrics. We follow the standard evaluation protocols in top-K recommendation. The performance of all models is evaluated using four widely-adopted metrics: Recall@K (R@K), Normalized Discounted Cumulative Gain (NDCG@K), Precision@K (P@K), and Mean Average Precision (MAP@K). We set K=20 for all experiments and report the average results across all test users.

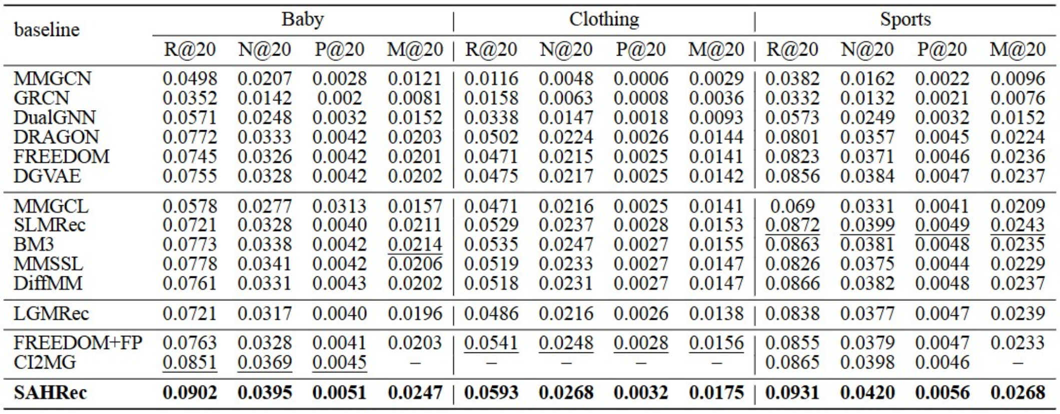
Baselines. We compare SAHRec against the same comprehensive set of baseline models detailed in the previous chapter. These baselines cover four main categories: GCN-based, SSL-based, hypergraph-based, and models specifically designed for incomplete modalities.

Implementation Details. Our proposed SAHRec model is implemented in PyTorch and trained on an NVIDIA A6000 GPU. We use the Adam optimizer for all models. For all baselines, we follow the optimal hyperparameter settings reported in their original papers to ensure a fair comparison. The key hyperparameters for SAHRec are determined via a grid search on the validation set. Specifically, the number of GCN layers L is searched in [1, 2, 3, 4]. For the AGH module, the number of hyperedges K_m is searched in [5, 10, 15, 20], the number of hypergraph convolution layers H is in [1, 2, 3, 4], and the hypergraph learning loss weight λ_1 is in {1e-4, 1e-3, 1e-2, 1e-1}. For the alignment and fusion module, the temperature τ is in [0.1, 0.2, 0.5, 1.0], the contrastive loss weight $\lambda_2$is in {1e-4, 1e-3, 1e-2, 1e-1}, and the fusion weight α is in [0.2, 0.4, 0.6, 0.8, 1.0]. The embedding size is consistently set to 64 for all models.

4.2 Performance Comparison (RQ1)

Table 2 presents the overall performance comparison of SAHRec against all baseline models under the most challenging setting of 90% modality missing rate, with the results visualized in Table 2.

Table 3 Performance comparison of baselines with the 90% missing rate

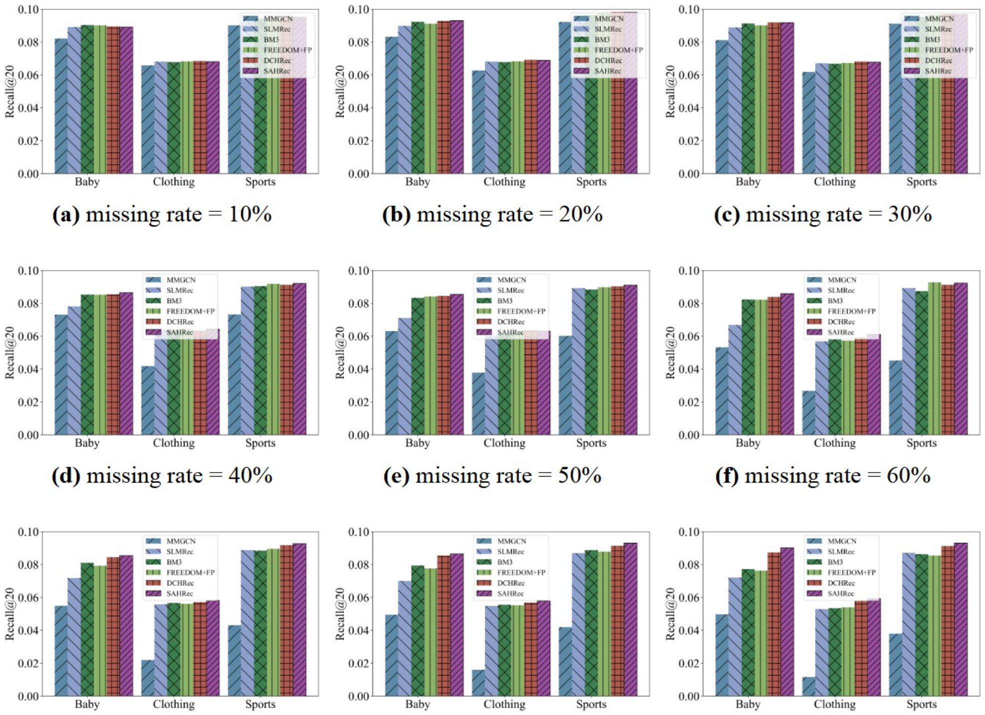


baseline	Baby				Clothing				Sports			
	R@20	N@20	P@20	M@20	R@20	N@20	P@20	M@20	R@20	N@20	P@20	M@20
MMGCN	0.0498	0.0207	0.0028	0.0121	0.0116	0.0048	0.0006	0.0029	0.0382	0.0162	0.0022	0.0096
GRCN	0.0352	0.0142	0.002	0.0081	0.0158	0.0063	0.0008	0.0036	0.0332	0.0132	0.0021	0.0076
DualGNN	0.0571	0.0248	0.0032	0.0152	0.0338	0.0147	0.0018	0.0093	0.0573	0.0249	0.0032	0.0152
DRAGON	0.0772	0.0333	0.0042	0.0203	0.0502	0.0224	0.0026	0.0144	0.0801	0.0357	0.0045	0.0224
FREEDOM	0.0745	0.0326	0.0042	0.0201	0.0471	0.0215	0.0025	0.0141	0.0823	0.0371	0.0046	0.0236
DGVAE	0.0755	0.0328	0.0042	0.0202	0.0475	0.0217	0.0025	0.0142	0.0856	0.0384	0.0047	0.0237
MMGCL	0.0578	0.0277	0.0313	0.0157	0.0471	0.0216	0.0025	0.0141	0.069	0.0331	0.0041	0.0209
SLMRec	0.0721	0.0328	0.0040	0.0211	0.0529	0.0237	0.0028	0.0153	0.0872	0.0399	0.0049	0.0243
BM3	0.0773	0.0338	0.0042	0.0214	0.0535	0.0247	0.0027	0.0155	0.0863	0.0381	0.0048	0.0235
MMSSL	0.0778	0.0341	0.0042	0.0206	0.0519	0.0233	0.0027	0.0147	0.0826	0.0375	0.0044	0.0229
DiffMM	0.0761	0.0331	0.0043	0.0202	0.0518	0.0231	0.0027	0.0147	0.0866	0.0382	0.0048	0.0237
LGMRec	0.0721	0.0317	0.0040	0.0196	0.0486	0.0216	0.0026	0.0138	0.0838	0.0377	0.0047	0.0239
FREEDOM+FP	0.0763	0.0328	0.0041	0.0203	0.0541	0.0248	0.0028	0.0156	0.0855	0.0379	0.0047	0.0233
CI2MG	0.0851	0.0369	0.0045	–	–	–	–	–	0.0865	0.0398	0.0046	–
SAHRec	**0.0902**	**0.0395**	**0.0051**	**0.0247**	**0.0593**	**0.0268**	**0.0032**	**0.0175**	**0.0931**	**0.0420**	**0.0056**	**0.0268**

The results clearly demonstrate that our proposed SAHRec model achieves state-of-the-art performance, outperforming all baseline models across all datasets and on all evaluation metrics. For instance, compared to the strongest baseline in each dataset, SAHRec shows significant improvements. This powerfully validates the superiority of our approach in handling the challenging task of recommendation with severe modality incompleteness.

The superiority of SAHRec stems from its novel design. Unlike conventional methods that rely on static or heuristic graph structures, SAHRec's adaptive hypergraph learner can discover a data-driven, task-optimal high-order structure in an end-to-end manner. Furthermore, the introduction of modality-aware contrastive learning provides a powerful self-supervised regularization signal. This mechanism compels the model to learn consistent and modality-invariant representations, which is crucial for enhancing generalization capability in the face of severe information loss.

To further validate the robustness of SAHRec, we conducted experiments across a spectrum of increasing missing rates, with the results visualized in Figure 2. The figure shows that while all models' performance degrades as modality absence becomes more severe, SAHRec's performance curve is consistently the highest and the flattest among all methods. This demonstrates its exceptional stability and robustness, highlighting that its performance gap over baselines widens as the data becomes sparser and more challenging.



(a) missing rate = 10%

(b) missing rate = 20%

(c) missing rate = 30%

(d) missing rate = 40%

(e) missing rate = 50%

(f) missing rate = 60%

(g) missing rate = 70%

(h) missing rate = 80%

(i) missing rate = 90%

Figure 2 Performance about the comparison with different missing rates

4.3 Ablation Study (RQ2)

To thoroughly investigate the contribution of each key component in SAHRec, we conducted an extensive ablation study. We designed several variants of the SAHRec model by removing or replacing specific modules. The results are summarized in Table 3.

Table 4 Ablation study results on Baby, Clothing, and Sports.

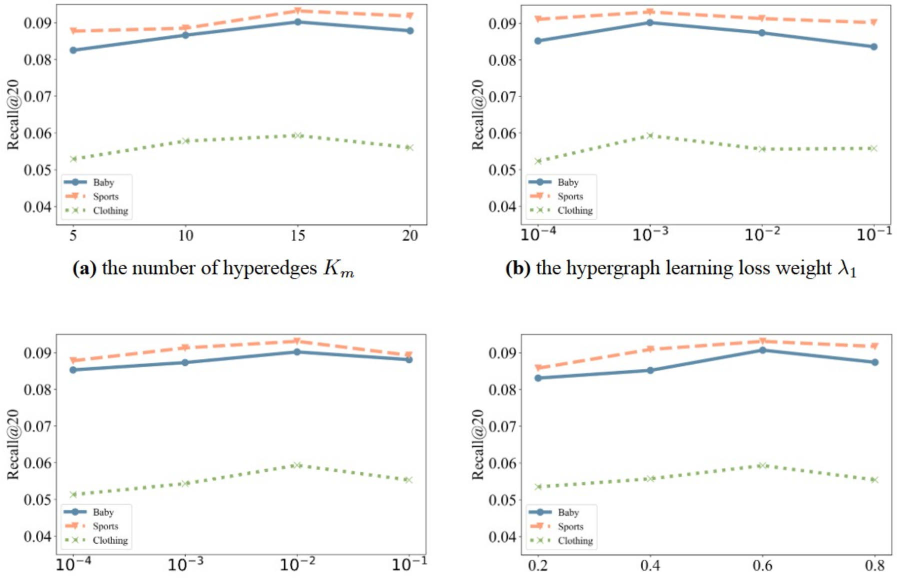
Metric	Baby				Clothing				Sports			
	R@20	N@20	P@20	M@20	R@20	N@20	P@20	M@20	R@20	N@20	P@20	M@20
w/o LGE	0.0865	0.0379	0.0049	0.0238	0.0569	0.0253	0.0031	0.0169	0.0905	0.0408	0.0054	0.0260
w/o AGH	0.0812	0.0355	0.0045	0.0221	0.0544	0.0241	0.0029	0.0157	0.0879	0.0395	0.0051	0.0249
w/ Static-H	0.0858	0.0375	0.0048	0.0235	0.0571	0.0255	0.0031	0.0170	0.0910	0.0411	0.0054	0.0261
w/o HGL	0.0871	0.0383	0.0049	0.0240	0.0575	0.0258	0.0031	0.0171	0.0914	0.0413	0.0054	0.0262
w/o SSL	0.0881	0.0388	0.0050	0.0241	0.0582	0.0262	0.0032	0.0172	0.0919	0.0415	0.0055	0.0264
SAHRec	**0.0902**	**0.0395**	**0.0051**	**0.0247**	**0.0593**	**0.0268**	**0.0032**	**0.0175**	**0.0931**	**0.0420**	**0.0056**	**0.0268**

The key findings are as follows: (1) Removing either the local (w/o LGE) or the global (w/o AGH) module leads to a significant performance drop, confirming the necessity of modeling both local and global dependencies. (2) Replacing the adaptive hypergraph learner with a static, k-NN-based one (w/ Static-H) results in inferior performance, which validates the superiority of the dynamic, data-driven structure learning approach. (3) Removing the hypergraph learning regularization loss (w/o HGL) also hurts performance, proving its essential role in guiding the structure learning process. (4) Most notably, removing the self-supervised contrastive learning task (w/o SSL) causes one of the most significant performance drops, highlighting the crucial contribution of the self-supervised signal to the model's robustness.

In summary, the ablation results systematically verify that the decoupled local-global architecture, the data-driven adaptive hypergraph learning, and the self-supervised alignment mechanism are all indispensable components of the SAHRec framework.

4.4 Hyperparameter Analysis (RQ3)

We analyzed the sensitivity of SAHRec to its four key new hyperparameters: the number of hyperedges K_m, the hypergraph learning loss weight λ_1, the contrastive learning loss weight λ_2, and the global fusion weight α. As shown in Figure 3 , the model's performance generally shows a pattern of first increasing and then decreasing as each hyperparameter value changes. This indicates that a proper balance is required. For instance, too few hyperedges are insufficient to capture complex relations, while too many may introduce noise. Similarly, the loss weights λ_1 and λ_2 need to be carefully tuned to balance the multi-task learning objectives. The optimal value for the fusion weight α is typically in the middle range, suggesting that the final user preference is a result of the combined effect of both local behavioral patterns and global semantic dependencies.



(a) the number of hyperedges K_m

(b) the hypergraph learning loss weight λ_1

(c) the contrastive learning loss weight λ_2

(d) the global fusion weight α

Figure 3 Hyperparameter analysis on different datasets

5. Conclusion

In this paper, we tackled the critical challenge of incomplete modalities in multimodal recommendation. Existing methods often fail in real-world scenarios due to their reliance on complete data, static graph structures, and sparse supervision. To address these limitations, we proposed SAHRec, a Self-supervised Adaptive Hypergraph Recommendation framework that enhances recommendation accuracy and robustness through a dynamic, data-driven learning paradigm.

The SAHRec model integrates three core innovations. First, a decoupled Local Graph Embedding module isolates collaborative signals from modality-specific preferences to prevent noise propagation. Second, an Adaptive Global Hypergraph module employs a differentiable learner to discover optimal high-order structures in an end-to-end fashion, moving beyond static, heuristic-based methods. Third, a modality-aware contrastive learning task provides a powerful self-supervised signal, compelling the model to learn consistent and robust representations by aligning local and global views.

Extensive experiments on three large-scale datasets demonstrated the superiority of SAHRec. Our model significantly outperformed a wide range of state-of-the-art baselines across all evaluation metrics, particularly under extreme modality absence of up to 90%. In-depth ablation studies further verified the indispensable role of each proposed component. In conclusion, by shifting the paradigm from static to dynamic and self-supervised learning, this research provides a novel and effective solution for building more robust and intelligent recommendation systems.

Future Work. While SAHRec has shown promising results, several avenues for future research remain open. First, exploring more efficient and interpretable mechanisms for adaptive hypergraph construction, especially for web-scale graphs, would be a valuable direction. Second, extending our framework to incorporate temporal dynamics, capturing the evolution of user interests and item attributes over time, presents an interesting challenge. Finally, investigating the application of our dynamic learning and self-supervised principles to other domains, such as cross-domain recommendation and fairness-aware modeling, could further broaden the impact of this work.

References

GUO Z, LI J, LI G, et al. LGMRec: local and global graph learning for multimodal recommendation[C] // Proceedings of the AAAI Conference on Artificial Intelligence : Vol 38. 2024 : 8454–8462.

MALITESTA D, ROSSI E, POMO C, et al. Do We Really Need to Drop Items with Missing Modalities in Multimodal Recommendation?[C] // Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. 2024 : 3943–3948.

HE X, DENG K, WANG X, et al. Lightgcn: Simplifying and powering graph convolution network for recommendation[C] // Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval. 2020 : 639– 648.

FENG Y, YOU H, ZHANG Z, et al. Hypergraph neural networks[C] // Proceedings of the AAAI conference on artificial intelligence : Vol 33. 2019 : 3558–3565.

WEI C, LIANG J, BAI B, et al. Dynamic hypergraph learning for collaborative filtering[C] // Proceedings of the 31st ACM international conference on information & knowledge management. 2022 : 2108 – 2117.

HE R, MCAULEY J. VBPR: visual bayesian personalized ranking from implicit feedback[C] // Proceedings of the AAAI conference on artificial intelligence : Vol 30. 2016.

CHEN J, ZHANG H, HE X, et al. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention[C] // Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval. 2017 : 335 – 344.

SCARSELLI F, GORI M, TSOI A C, et al. The graph neural network model[J]. IEEE transactions on neural networks, 2008, 20(1) : 61–80.

TAO Z, LIU X, XIA Y, et al. Self-supervised learning for multimedia recommendation[J]. IEEE Transactions on Multimedia, 2022, 25 : 5107–5116.

WEI W, HUANG C, XIA L, et al. Multi-modal self-supervised learning for recommendation[C] // Proceedings of the ACM Web Conference 2023. 2023 : 790–800.

OORD A V D, LI Y, VINYALS O. Representation learning with contrastive predictive coding[J]. arXiv preprint arXiv:1807.03748, 2018.

STOER M, WAGNER F. A simple min cut algorithm[C] // European Symposium on Algorithms. 1994 : 141 – 147.

REIMERS N, GUREVYCH I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks[J]. arXiv preprint arXiv:1908.10084, 2019.

DENG J, DONG W, SOCHER R, et al. Imagenet: A large-scale hierarchical image database[C] // 2009 IEEE conference on computer vision and pattern recognition. 2009 : 248 – 255.