

Semantic-Augmented Item Graph Denoising with Collaborative Signals for Multimodal Recommendation

1st Run Lin

*National Key Laboratory of Information Systems Engineering
National University of Defense Technology
Changsha, China
lrun@nudt.edu.cn*

2nd Xiangyu Chen

*National Key Laboratory of Information Systems Engineering
National University of Defense Technology
Changsha, China
chenxiangyu24@nudt.edu.cn*

3rd Yuzhuo Dang

*National Key Laboratory of Information Systems Engineering
National University of Defense Technology
Changsha, China
dangyuzhuo@nudt.edu.cn*

4th Zhiqiang Pan

*National Key Laboratory of Information Systems Engineering
National University of Defense Technology
Changsha, China
panzhiqiang@nudt.edu.cn*

5th Mengfan Kong

*National Key Laboratory of Information Systems Engineering
National University of Defense Technology
Changsha, China
kongmf23@nudt.edu.cn*

6th Honghui Chen*

*National Key Laboratory of Information Systems Engineering
National University of Defense Technology
Changsha, China
248856416@qq.com*

Abstract—Multimodal recommender systems (MMRS) incorporate item-side content, such as images and texts, to complement limited interaction records and improve recommendation for less-observed users or items. However, existing methods still suffer from several limitations: (1) high-level semantic information in multimodal content is difficult to explicitly capture; (2) multimodal signals are often overwhelmed by ID-based collaborative signals during training; (3) item-item graphs constructed solely based on the multimodal similarity may introduce noisy edges due to the semantic-behavior inconsistency. Driven by the above observations, we propose SAGDRec, a Semantic-Augmented Graph Denoising approach for multimodal recommendation. Specifically, we leverage an Multimodal Large Language Model (MLLM) to extract the high-level semantics from the visual and textual content, thereby enriching the multimodal semantic features. Specifically, to prevent the multimodal semantics from being overshadowed by the collaborative signals, we introduce an additional multimodal semantic branch to strengthen the learning driven by semantic cues. Moreover, we build a semantic item-item graph and refine its edges with collaborative signals, so that behaviorally inconsistent semantic neighbors are suppressed during propagation. Experiments show that SAGDRec achieves competitive performance on three public datasets.

Index Terms—Multimodal Recommender Systems, Graph Neural Networks, Multimodal Large Language Models, Multimodal Semantic Branch, Graph Denoising

I. INTRODUCTION

Recommender systems are required to estimate user preferences under increasingly rich item information, where historical interactions alone may provide an incomplete view of user

interests. In many online platforms, items are accompanied by images, titles, textual descriptions, and other content signals, which provide additional clues about item attributes. MMRS incorporate these heterogeneous signals into item modeling, allowing item representations to contain richer semantic cues and helping reduce the impact of sparse interaction data [1]–[5].

Early MMRS methods, such as VBPR [6], mainly incorporate visual or textual features into models. Recent studies such as COHESION [7] and PEARL [8] further explore homogeneous graph structures in MMRS, using connections among similar users or among related items to improve embedding quality.

Despite the substantial progress achieved by existing MMRS approaches, several critical challenges remain unsolved. First, multimodal features extracted by pre-trained encoders often lack explicit high-level semantic understanding, particularly for the visual content. Second, multimodal signals are commonly treated as auxiliary information and tend to be overwhelmed by the ID-based collaborative signals during training, leading to insufficient utilization of multimodal semantics. Moreover, item-item graphs constructed purely based on the multimodal similarity inevitably introduce noisy edges due to the semantic-behavioral inconsistency, which degrades the effectiveness of the graph-based item representation learning. Specifically, two items that are semantically similar may attract completely different user groups due to differences

in the style or the usage scenarios (e.g., casual T-shirts and moisture-wicking athletic T-shirts with similar appearances but different target users). When such false-positive edges exist in the item-item graph, graph propagation forces these behaviorally unrelated items to share similar representations, injecting irrelevant semantic signals into the item embeddings. The corrupted item embeddings may then be absorbed into user embeddings, causing ambiguous preference modeling.

Based on the above analysis, we introduce **SAGDRec**, a Semantic-Augmented Graph Denoising approach for multimodal recommendation. On the one hand, to enhance the multimodal content understanding, we leverage an MLLM to transform visual signals into high-level semantic textual representations, while extracting the core semantic information from textual signals. These two types of textual representations are then integrated and encoded by a unified text encoder, generating semantically enhanced multimodal item representations. Furthermore, to prevent these semantic signals from being overwhelmed by ID-based collaborative supervision during training, we introduce an additional multimodal semantic branch to explicitly strengthen the preference learning driven by semantic cues. Meanwhile, content-driven item links may be unreliable when semantic proximity is inconsistent with user behavior. Therefore, we use collaborative signals to calibrate the semantic graph and reduce the influence of misleading neighbors. By measuring the consistency between semantic similarity and collaborative behavioral patterns, SAGDRec suppresses edges that are semantically similar but behaviorally inconsistent, thereby improving the reliability of information propagation.

The main advances of SAGDRec are summarized below:

- We leverage an MLLM to process item images and texts, extracting high-level semantic information to construct more expressive multimodal representations of items.
- We design an additional multimodal semantic branch to alleviate the problem that multimodal signals are often overlooked during training.
- We utilize collaborative signals to denoise the item-item graph by suppressing edges with inconsistent semantic and behavioral patterns, ensuring that the information propagated by graph convolutional networks is truly beneficial for recommendation.
- Across the three public datasets, SAGDRec obtains a 6.08% average gain compared with the strongest baseline.

II. RELATED WORK

Existing studies relevant to our work can be grouped into two research lines: graph-based multimodal recommendation and large language models (LLMs)-based recommendation. These two lines of research are closely related to the core design of our proposed SAGDRec framework, as they respectively focus on graph-based modeling and leveraging LLMs for recommendation modeling, which are the two key aspects jointly addressed in this work.

A. Graph-based Multimodal Recommendation

Graph-based multimodal recommendation methods mainly differ in how they organize relational structures for multimodal propagation. Representative methods include MMGCN [9] and GRCN [10], which propagate multimodal signals on user-item graphs, while DualGNN [11], LATTICE [12], and FREEDOM [13] further exploit user-user or item-item relations for representation learning. In addition to graph structure optimization, some works also focus on improving multimodal feature modeling before or during graph construction. Then, MGCN [14] purifies modality features using item behavioral information prior to building the item-item graph and models modality-specific user preferences to guide feature fusion. Along this line, PEARL [8] further considers that raw user-item interactions may carry modality-related noise, and thus learns purified interaction graphs together with user and item affinity relations. Moreover, COHESION [7] organizes interaction-level and affinity-level graph structures within one model, and couples early modality refinement with late representation fusion to improve multimodal modeling. Additionally, DRNet [15] focuses on the internal composition of multimodal features by separating modality-aligned and modality-specific information, and then constructs cross-modal item-item graphs to capture richer item relations. From another perspective, some studies aim to alleviate semantic distortion during graph propagation. MONET [16] proposes a modality-inclusive GCN by introducing self-connections to explicitly preserve original multimodal features during linear propagation, avoiding semantic distortion caused by nonlinear aggregation. In the same way, MIRF [17] assigns modality-specific propagation depths to better capture the distinct structural patterns of different modalities. To further explore advanced graph structures, recent studies introduce hypergraphs into multimodal recommendation. LGMRec [18] uses a hypergraph-based global module to capture attribute-oriented dependencies that are difficult to obtain from local interaction patterns alone. Moreover, SEHNN [19] constructs a heterogeneous hypergraph network with customized aggregators to handle different types of information propagation.

Nevertheless, in many graph-based MMRS, content-neighbor relations are mainly determined by multimodal feature proximity. Such relations may be unreliable when items look or read alike but correspond to different user behavior patterns, causing misleading information to be propagated during graph learning.

B. LLMs-based Recommendation

Recent studies explore LLMs as semantic bridges between multimodal content and recommendation models, using their language understanding ability to convert raw item information into more structured and recommendation-oriented semantic signals. DOGE [20] leverages LLMs-guided textual information to interpret images, effectively mining cross-modal correlations between texts and images. NoteLLM-2 [21] studies LLMs-enhanced multimodal representation learning for item-to-item recommendation. Instead of directly adopting an

existing MLLM, it fine-tunes the connection between an LLM and a visual encoder, and further uses multimodal in-context learning and late fusion to reduce the tendency of the model to ignore visual content. MLLMRec [22] converts item images into semantic descriptions, fuses them with textual metadata, and further incorporates MLLM-inferred user preferences into recommendation. Similarly, MLLM-MSR [23] also transforms visual content into textual semantics and tracks the evolution of user preferences across sequential interactions.

Despite these advances, how to further incorporate LLM-derived high-level semantics into both representation learning and graph structure modeling for multimodal recommendation remains worth exploring.

III. METHODOLOGY

We first define the notation and problem setting used throughout this section. Then, we detail the SAGDRec model, which consists of four main components: multimodal representation construction module, collaborative-signal guided item-item graph construction and denoising module, additional multimodal semantic branch, and optimization objective.

Fig. 1 presents the workflow of SAGDRec. SAGDRec separates semantic extraction from recommendation training. Before model optimization, an MLLM converts item images and textual content into semantic descriptions. These semantic descriptions are further encoded by a unified text encoder, yielding semantic-enhanced multimodal item representations. In the second stage, the extracted semantic information is incorporated into a trainable recommendation framework, which consists of three modules. First, the multimodal representations are integrated with item ID embeddings to construct content-aware item representations. These semantic features are used to connect content-related items, and the resulting links are adjusted with behavioral evidence learned from historical interactions. Through graph-based aggregation, multimodal representations of both users and items are derived. Second, an additional multimodal semantic branch is introduced to prevent semantic signals from being overwhelmed by ID-based collaborative supervision during training. Third, the model fuses multimodal and ID embeddings, jointly optimized via a main recommendation loss and auxiliary losses.

In multimodal recommendation, the model needs to estimate how likely a user is to interact with an item by multimodal content. We describe this setting with the following notation. Let $\mathcal{U}$ and $\mathcal{I}$ represent the user set and the item set, respectively. Each user $u \in \mathcal{U}$ and item $i \in \mathcal{I}$ is assigned an ID embedding, denoted as e_u^{id} and e_i^{id} . In addition, item i is paired with visual and textual content, which provides extra semantic evidence beyond its ID. We denote the user-item interaction matrix as $R \in \{0, 1\}^{|\mathcal{U}| \times |\mathcal{I}|}$. For an entry $r_{u,i}$, the value 1 means that an interaction between user u and item i is recorded, while the value 0 means that such behavior is not observed. For a target item i , SAGDRec computes a matching score $\hat{y}_{u,i}$ for user u by using both past user behavior and multimodal item content. The scores assigned to all candidate items form $\hat{\mathbf{y}}_u = (\hat{y}_{u,1}, \hat{y}_{u,2}, \dots, \hat{y}_{u,|\mathcal{I}|})$. Items with larger

scores are regarded as more relevant to user u , and the final recommendation list is generated according to the descending order of these scores.

A. Multimodal Semantic Feature Construction

To enhance multimodal content understanding, we leverage an MLLM to map low-level visual signals into high-level semantic textual representations. For each item image, we generate a semantic description T_i^{img} . Meanwhile, the original item textual description is refined to extract its core semantic information, denoted as T_i^{text} . Then, the generated image descriptions and the extracted textual contents are concatenated and fed into a unified text encoder f_{text} to obtain the initial multimodal representation $e_i^{\text{mm}} \in \mathbb{R}^d$ as follows:

$$T_i^{\text{fusion}} = \text{Concat}(T_i^{\text{img}}, T_i^{\text{text}}), \quad (1)$$

$$e_i^{\text{mm}} = f_{\text{text}}(T_i^{\text{fusion}}). \quad (2)$$

Through this process, visual and textual modalities are naturally aligned in a unified semantic space, since both modalities are transformed into textual contents and encoded by the same pre-trained text encoder. This unified encoding strategy alleviates the modality gap and facilitates subsequent graph-based learning [22], [24].

B. Item-Item Graph Construction and Denoising

In order to adaptively modulate the influence of multimodal semantics during fusion with the item ID embedding, we apply MLP to the MLLM-enhanced multimodal representation e_i^{mm} to generate a gated multimodal representation:

$$e_i^{\text{mm}'} = \sigma(\text{MLP}(e_i^{\text{mm}})), \quad e_i^{\text{mm}'} \in \mathbb{R}^d. \quad (3)$$

After obtaining the gated multimodal representation $e_i^{\text{mm}'}$, we fuse it with the item ID embedding $e_i^{\text{id}} \in \mathbb{R}^d$ to construct a content-aware item representation as:

$$e_i^{\text{con}} = e_i^{\text{id}} \odot e_i^{\text{mm}'}, \quad e_i^{\text{con}} \in \mathbb{R}^d, \quad (4)$$

where $\sigma(\cdot)$ refers to the sigmoid activation, and $\odot$ indicates multiplication performed dimension by dimension. The content-aware item embeddings are further used to derive semantic connections among items. For each item, we search for nearby items in the multimodal representation space and regard them as semantic neighbors. The resulting neighbor structure forms the initial semantic item-item graph, where the pairwise similarity score $s_{i,j}$ is computed as:

$$s_{i,j} = \frac{e_i^{\text{con}} \cdot e_j^{\text{con}}}{\|e_i^{\text{con}}\| \|e_j^{\text{con}}\|}. \quad (5)$$

To keep the semantic graph concise, we filter out weak item pairs and retain K highly related items for each anchor item:

$$S_{i,j} = \begin{cases} s_{i,j}, & s_{i,j} \in \text{Top-}K(s_{i,j} \mid j \in \mathcal{I}), \\ 0, & \text{otherwise,} \end{cases} \quad (6)$$

where $\mathbf{S} \in \mathbb{R}^{|\mathcal{I}| \times |\mathcal{I}|}$ is the sparse item-item semantic similarity matrix.

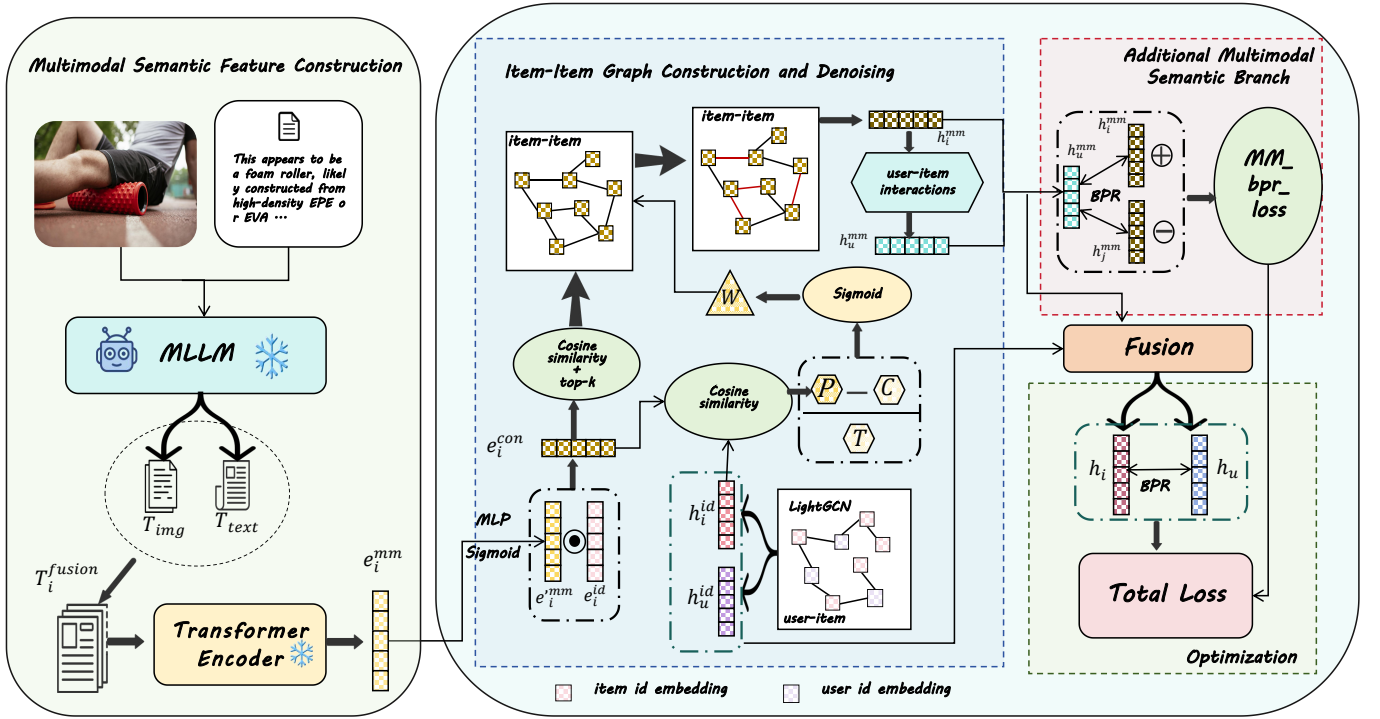


Fig. 1. The overall framework of SAGDRec.

However, semantic similarity does not necessarily imply the behavioral relevance. Specifically, item-item graphs constructed purely based on the semantic similarity may contain false-positive edges, which can mislead graph-based propagation. Semantic similarity alone may connect items with different user preference patterns. To reduce the influence of such unreliable edges, we use collaborative information from the user-item graph as behavioral guidance. The behavioral signal is obtained by running LightGCN [25] over the user-item interaction graph, which produces item embeddings encoded with collaborative patterns. We denote the concatenation of user and item ID embeddings as $E^{id} = [e_u^{id}; e_i^{id}] \in \mathbb{R}^{(|\mathcal{U}|+|\mathcal{I}|) \times d}$. The propagation can be defined as follows:

$$A = \begin{bmatrix} 0 & R \\ R^\top & 0 \end{bmatrix}, \quad (7)$$

$$H_{id}^{(l+1)} = AH_{id}^{(l)}, \quad H_{id}^{(0)} = E^{id}, \quad (8)$$

$$H_{id} = \frac{1}{L+1} \sum_{l=0}^L H_{id}^{(l)}, \quad (9)$$

where $H_{id} = [h_u^{id}; h_i^{id}] \in \mathbb{R}^{(|\mathcal{U}|+|\mathcal{I}|) \times d}$ denotes the final representation derived from collaborative propagation.

Then, we evaluate semantic-behavioral consistency by comparing the semantic vector of i with the behavior-aware vector of j . The motivation behind this design is that a reliable item-item edge should satisfy both semantic proximity and behavioral consistency. Semantic similarity indicates that two items share similar content attributes, while collaborative representations encode user preference patterns derived from historical

interactions. If an item is semantically close to another item but its collaborative representation is far from the corresponding semantic representation, this edge may reflect superficial content similarity rather than recommendation-useful relevance. Therefore, the cross-space consistency between semantic and collaborative representations can be viewed as behavioral evidence for judging whether a semantic edge should be preserved during graph propagation. Specifically, we compute the behavioral evidence as $P_{i,j}$, and introduce a temperature-controlled gating function to adjust edge weights:

$$P_{i,j} = \text{sim}(e_i^{con}, h_j^{id}), \quad (10)$$

$$w_{i,j} = \sigma \left(\frac{P_{i,j} - C}{T} \right), \quad (11)$$

$$G_{i,j} = w_{i,j} \cdot S_{i,j}, \quad (12)$$

where C controls the minimum consistency required for an edge to be strongly preserved, and T adjusts the smoothness of the gating function. Through this gating mechanism, a semantic edge receives a smaller weight when its collaborative evidence is weak, even if the two items are close in the semantic space. As a result, behaviorally inconsistent neighbors contribute less during item-item propagation, reducing the risk of injecting irrelevant semantic information into item representations.

After noise reduction processing, the similarity matrix changes from $S_{i,j}$ to $G_{i,j} \in \mathbb{R}^{|\mathcal{I}| \times |\mathcal{I}|}$. On this basis, we aggregate multimodal neighbor information for each item by

using the similarity matrix as follows:

$$h_i^{mm} = \sum_{j \in \mathcal{I}} g_{i,j} e_j^{con}, \quad h_i^{mm} \in \mathbb{R}^d, \quad (13)$$

where $g_{i,j}$ denotes the entry of the denoised item-item similarity matrix G at position (i, j) , quantifying the semantic closeness between item i and item j .

Finally, a user's multimodal vector is obtained from the denoised multimodal embeddings of previously interacted items:

$$h_u^{mm} = \sum_{j \in \mathcal{I}_u} r_{u,j} h_j^{mm}, \quad h_u^{mm} \in \mathbb{R}^d, \quad (14)$$

where $\mathcal{I}_u$ contains the items previously interacted with by user u , and h_u^{mm} is the resulting user-side multimodal semantic vector.

C. Additional Multimodal Semantic Branch

In the joint training process, ID embeddings often receive stronger optimization signals because they are directly tied to observed user-item interactions. By contrast, multimodal semantic embeddings act as auxiliary information and may not be fully trained for recommendation if they only participate through the fused branch. This imbalance can weaken the role of semantic features and make the final prediction rely excessively on collaborative ID signals. To address this issue, we introduce an additional multimodal semantic branch that explicitly enforces semantic supervision using multimodal representations. Specifically, instead of solely relying on the fused representations in the main recommendation branch, we construct an auxiliary BPR [26] loss purely based on multimodal semantic embeddings as follows:

$$\mathcal{L}_{MM-BPR} = - \sum_{(u,i,j)} \log \sigma(h_u^{mm\top} h_i^{mm} - h_u^{mm\top} h_j^{mm}), \quad (15)$$

where i corresponds to an interacted item, and j serves as the negative sample. This design encourages the model to directly learn user preferences from multimodal semantics, thereby preventing semantic information from being diluted during joint optimization.

D. Optimization Objective

We obtain the final representations by merging two complementary views: the ID-based view learned from interactions and the semantic view derived from multimodal content:

$$h_i = h_i^{mm} + h_i^{id}, \quad h_u = h_u^{mm} + h_u^{id}, \quad (16)$$

where $h_i, h_u \in \mathbb{R}^d$ denote the final representations of item i and user u , respectively.

Based on the fused representations, the main BPR loss is defined as:

$$\mathcal{L}_{BPR} = - \sum_{(u,i,j)} \log \sigma(h_u^\top h_i - h_u^\top h_j), \quad (17)$$

where triplet (u, i, j) contains one user u , an observed item i , and an unobserved item j sampled as the negative instance.

To enhance user-item alignment, we introduce a cosine similarity loss as follows:

$$\mathcal{L}_{UI} = 1 - \text{sim}(h_u, h_i), \quad (18)$$

where $\mathcal{L}_{UI}$ reduces the distance between user u and its interacted item i in the embedding space, with $\text{sim}(\cdot, \cdot)$ computed by cosine similarity.

Furthermore, we apply an InfoNCE [27] loss to align ID and multimodal representations as follows:

$$\mathcal{L}_{IM} = \text{InfoNCE}_u + \text{InfoNCE}_i, \quad (19)$$

$$\text{InfoNCE}_x = - \log \frac{\exp(h_x^{mm} \cdot h_x^{id} / \tau)}{\sum_{x'} \exp(h_x^{mm} \cdot h_{x'}^{id} / \tau)}, \quad (20)$$

where $x \in \{u, i\}$ denotes a user or an item, x' indicates all users or items in the corresponding set, and τ adjusts how strongly the contrastive objective distinguishes the positive pair from other entities. This objective encourages the multimodal representation of each entity to be close to its corresponding ID-based representation while being discriminative from others.

Model parameters are learned by jointly minimizing the following ranking and auxiliary regularization terms:

$$\mathcal{L} = \mathcal{L}_{BPR} + \alpha \mathcal{L}_{MM-BPR} + \beta \mathcal{L}_{UI} + \lambda \mathcal{L}_{IM}, \quad (21)$$

where $\mathcal{L}_{BPR}$ is the main recommendation loss, $\mathcal{L}_{MM-BPR}$ is the auxiliary multimodal semantic loss, $\mathcal{L}_{UI}$ enforces user-item alignment, and $\mathcal{L}_{IM}$ aligns ID-based and multimodal representations. The hyperparameters α , β , and λ determine how much the semantic ranking branch, the user-item alignment term, and the ID-multimodal alignment term contribute to training. The role of InfoNCE is to align two views of the same entity, making its ID-based representation and multimodal representation more consistent. However, such alignment does not directly guarantee that the multimodal representations themselves are sufficiently optimized for the ranking task. Therefore, InfoNCE focuses on representation alignment, while the semantic branch aims to preserve the recommendation capability of multimodal semantics during optimization.

IV. EXPERIMENTS

This section reports the empirical study of SAGDRec. We describe the datasets, compared methods, evaluation protocol, and implementation details, then show the empirical observations.

A. Experimental Settings

1) *Datasets*: We use three categories from the Amazon review data, namely Baby, Sports, and Clothing. Each category provides user behavior records as well as item content from visual and textual modalities, which allows us to examine whether SAGDRec can benefit from multimodal semantics under different data scales and densities. Table I gives the detailed statistics of the processed data, including the numbers of users, items, observed interactions, and density.

TABLE I
STATISTICS OF THE EXPERIMENTAL DATASETS.

Datasets	#Users	#Items	#Interactions	Density
<i>Baby</i>	19,445	7,050	160,792	0.117%
<i>Sports</i>	35,598	18,357	296,337	0.045%
<i>Clothing</i>	39,387	23,033	278,677	0.031%

For data preprocessing, we follow the 5-core setting used in prior multimodal recommendation studies [7], [12], [18]. After preprocessing, the retained user-item records are randomly partitioned into three subsets with a ratio of 8:1:1.

2) *Baselines and Evaluation Metrics*: We include both interaction-only recommenders and multimodal recommendation methods. The interaction-only group contains BPR [26] and LightGCN [25], which rely on user-item feedback without using item content. The multimodal group contains VBPR [6], MMGCN [9], GRCN [10], DualGNN [11], LATTICE [12], SLMRec [28], BM3 [29], and LGMRec [18], covering matrix-factorization-based, graph-based, and self-supervised multimodal models.

We use Recall@K and NDCG@K to measure ranking performance. Recall@K reflects whether relevant items appear in the recommendation list, while NDCG@K further considers their ranking positions. Following the evaluation setting of recent multimodal recommendation studies [7], [8], [18], we calculate the average scores over with $K = 10$ and $K = 20$.

3) *Implementation Details*: During the semantic inference stage, we employ VideoChat2-TPO [30] as the MLLM, which is built upon the Mistral-7B-Instruct-v0.2 backbone. The MLLM inference is conducted on an NVIDIA RTX 4090 GPU. Subsequently, the graph-based recommendation is conducted on an NVIDIA RTX 4070 GPU. Depending on the dataset size, the MLLM inference stage takes approximately 6 to 20 hours. Once the image descriptions and refined textual semantics are generated, they are cached and encoded as fixed semantic features for subsequent recommendation training and inference. For collaborative modeling, LightGCN is stacked with two propagation steps, while the semantic item graph uses a single convolution step. The neighborhood size is set to $K = 10$. The collaborative-signal-guided denoising mechanism is activated after a warm-up of 30 epochs. For the similarity threshold C , we carry out a grid search in $\{0.0, 0.2, 0.4, 0.6, 0.8\}$, and we test T with 0.2, 0.4, and 0.6. During training, the validation Recall@20 is monitored for model selection. Training stops when Recall@20 on the validation set remains unchanged for 30 epochs, with the epoch limit set to 1,000. All models are implemented in PyTorch with seed 999.

B. Overall Results

Table II summarizes the ranking accuracy of SAGDRec and all baselines across the three datasets. The results reveal the following findings:

First, we can see that models leveraging multimodal information consistently outperform traditional recommendation models across most datasets and metrics. This suggests that

visual and textual content can compensate for the limited supervision provided by sparse user-item interactions.

Moreover, different multimodal models exhibit varying performance across datasets. For example, DualGNN performs worse than GRCN on the Baby dataset but outperforms it on the Clothing dataset, as its user-user graph is more beneficial for larger and sparser datasets. Moreover, LGMRec achieves strong performance on Baby and Sports but shows poor performance on Clothing, likely because its complex hypergraph structure is better suited for smaller-scale datasets.

Compared with LGMRec, the strongest baseline in Table II, SAGDRec raises Recall@20 by 1.62% on Baby, 6.00% on Sports, and 10.75% on Clothing. The corresponding gains on NDCG@20 are 1.40%, 5.24%, and 11.32%. The improvement on Baby is smaller than that on the other two datasets. A possible reason is that Baby contains many items with similar visual appearances and textual descriptions. That is, the high-level semantic representations extracted by MLLM exhibit limited discriminability, leading to relatively homogeneous multimodal semantic features. Under such conditions, semantic enhancement provides less additional information beyond existing collaborative signals. This also indicates that SAGDRec is more suitable for handling large-scale and more diverse datasets.

C. Ablation Study

To examine how different designs affect the final recommendation results, we compare SAGDRec with three reduced versions: (1) **SAGDRec w/o-MLLM**, which uses the visual and textual features provided by MMRec for concatenation instead of the semantic features obtained through MLLM reasoning; (2) **SAGDRec w/o-GD**, which disables the semantic edge refinement component; (3) **SAGDRec w/o-SB**, which excludes the additional multimodal semantic branch.

Table III further reveals how the performance changes when one design is removed from SAGDRec. On Sports and Clothing, each simplified variant underperforms the full model, suggesting that the three modules improve recommendation from different aspects: semantic enrichment, edge reliability, and semantic-aware optimization. The semantic branch has the most visible impact. After removing it, the performance on Clothing decreases by 8.84% in Recall@10 and 9.05% in NDCG@10. This indicates that multimodal semantics should not only be fused with ID embeddings, but also receive direct ranking-oriented supervision. Otherwise, the optimization process may be dominated by collaborative ID signals, especially on datasets with diverse item attributes.

In addition, on the Baby dataset, using concatenated multimodal features yields slightly better performance than the MLLM-inferred features. We attribute this phenomenon to the relatively small parameter size of the adopted MLLM, which may struggle to capture fine-grained distinctions in highly specialized domains. In detail, baby-category items often exhibit highly similar appearances, where direct visual and textual features preserve subtle physical attributes. In contrast, the MLLM-based semantic abstraction may smooth

TABLE II
OVERALL PERFORMANCE OF BASELINES AND SAGDREC.

Model	Baby				Sports				Clothing			
	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20
BPR	0.0357	0.0575	0.0192	0.0249	0.0432	0.0653	0.0241	0.0298	0.0187	0.0279	0.0103	0.0126
LightGCN	0.0479	0.0754	0.0257	0.0328	0.0569	0.0864	0.0311	0.0387	0.0340	0.0526	0.0188	0.0236
VBPR	0.0423	0.0663	0.0223	0.0284	0.0558	0.0856	0.0307	0.0384	0.0281	0.0415	0.0158	0.0192
MMGCN	0.0378	0.0615	0.0200	0.0261	0.0370	0.0605	0.0193	0.0254	0.0218	0.0345	0.0110	0.0142
GRCN	0.0532	0.0824	0.0282	0.0358	0.0559	0.0877	0.0306	0.0389	0.0424	0.0650	0.0225	0.0283
DualGNN	0.0448	0.0716	0.0240	0.0309	0.0568	0.0859	0.0310	0.0385	0.0454	0.0683	0.0241	0.0299
SLMRec	0.0529	0.0775	0.0290	0.0353	0.0663	0.0990	0.0365	0.0450	0.0452	0.0675	0.0247	0.0303
LATTICE	0.0547	0.0850	0.0292	0.0370	0.0620	0.0953	0.0335	0.0421	0.0492	0.0733	0.0268	0.0330
BM3	0.0564	0.0883	0.0301	0.0383	0.0656	0.0980	0.0355	0.0438	0.0422	0.0621	0.0231	0.0281
LGMRec	0.0639	0.0989	0.0337	0.0430	0.0719	0.1068	0.0387	0.0477	0.0555	0.0828	0.0302	0.0371
SAGDRec	0.0645	0.1005	0.0344	0.0436	0.0755	0.1132	0.0405	0.0502	0.0622	0.0917	0.0338	0.0413
Improv.	0.94%	1.62%	2.08%	1.40%	5.01%	6.00%	4.65%	5.24%	12.07%	10.75%	11.92%	11.32%

out these fine-grained differences. Besides, the denoising module mainly refines the overall semantic topology among items. Such adjustment helps suppress unreliable links, but may slightly change the ordering of closely ranked candidates, causing minor NDCG variations.

D. Hyper-parameter Analysis

This part studies how SAGDRec behaves under different settings of two important parameters: the semantic branch weight α and the denoising threshold C .

1) *Impact of α* : The parameter α determines how much the auxiliary semantic branch participates in model training. Fig. 2(a) reports the Recall@10 and NDCG@10 results on Sports when α varies. The curves reach their highest values at $\alpha = 0.3$, suggesting that SAGDRec benefits from a moderate amount of semantic supervision. When α is small, the semantic branch has limited influence, and multimodal semantics are still dominated by ID-based collaborative signals, leading to suboptimal performance. As α increases, semantic supervision becomes more effective and helps the model better capture user preferences from multimodal content. However, when α exceeds 0.3, the optimization is overly biased toward semantic signals, which may weaken collaborative learning and hurt performance.

2) *Impact of Similarity Threshold C* : We also examine how the threshold C affects the graph denoising process. As presented in Fig. 2(b), increasing C from 0 to 0.6 brings consistent gains on both Recall@10 and NDCG@10. The best result is obtained at $C = 0.6$, suggesting that this value suppresses unreliable semantic edges while still preserving useful neighbor information. When C is small, the denoising effect is weak and many semantically similar but behaviorally inconsistent edges are retained, which introduces noise during graph propagation. As C increases, more noisy edges are suppressed, leading to a cleaner graph structure and improved recommendation performance. However, when C exceeds 0.6, the performance of SAGDRec degrades, indicating that overly aggressive denoising may suppress informative edges. These

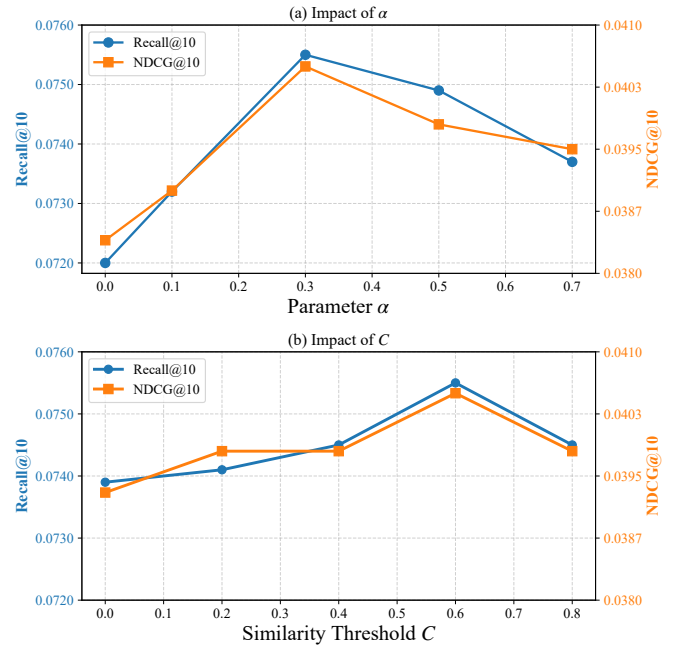


Fig. 2. Sensitivity analysis of key hyperparameters in SAGDRec. (a) Impact of the semantic branch weight α . (b) Impact of the similarity threshold C in the item-item graph denoising module.

results demonstrate that $C = 0.6$ provides an effective balance between noise removal and information preservation.

V. CONCLUSION

We explore the integration of MLLM-derived semantics and collaborative behavior in multimodal recommendation. By leveraging an MLLM, the proposed method enhances the semantic expressiveness of multimodal item representations. An additional multimodal semantic branch is introduced to prevent multimodal signals from being overshadowed by ID embeddings, while a collaborative-signal guided item-item graph denoising strategy effectively suppresses semantically

TABLE III
PERFORMANCE OF SAGDREC AND ITS VARIANTS.

Model variant	Baby		Sports		Clothing	
	R@10	N@10	R@10	N@10	R@10	N@10
SAGDRec	0.0645	0.0344	0.0755	0.0405	0.0622	0.0338
w/o-MLLM	0.0653	0.0355	0.0711	0.0387	0.0588	0.0317
w/o-GD	0.0635	0.0348	0.0737	0.0396	0.0615	0.0333
w/o-SB	0.0624	0.0342	0.0720	0.0384	0.0567	0.0308

similar but behaviorally inconsistent edges. Empirical observations from three public datasets confirm the contribution of each design, with clearer advantages on larger and more diverse datasets. Going forward, we will examine whether stronger or more efficient MLLMs can produce finer-grained item semantics and improve robustness across different recommendation scenarios.

REFERENCES

- [1] J. Huang, J. Qin, Y. Yu, and W. Zhang, "Beyond graph convolution: Multimodal recommendation with topology-aware MLPs," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 11, 2025, pp. 11 808–11 816.
- [2] J. Kim, H. Kang, S. Kim, K. Kim, and C. Park, "Disentangling and generating modalities for recommendation in missing modality scenarios," *arXiv preprint arXiv:2504.16352*, 2025.
- [3] R. K. Ong and A. W. Khong, "Spectrum-based modality representation fusion graph convolutional network for multimodal recommendation," in *Proc. 18th ACM Int. Conf. Web Search Data Mining (WSDM)*, 2025, pp. 773–781.
- [4] Y. Qi, Q. Zhang, X. Lin, X. Su, J. Zhu, J. Wang, and J. Li, "Seeing beyond noise: Joint graph structure evaluation and denoising for multimodal recommendation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 12, 2025, pp. 12 461–12 469.
- [5] Y. Wu, S. Chen, B. Wu, F. Li, and J. Gao, "Adaptive gradient masking for balancing id and mllm-based representations in recommendation," in *Advances in Neural Information Processing Systems*, vol. 38, 2026, pp. 14 331–14 357.
- [6] R. He and J. McAuley, "VBPR: Visual Bayesian personalized ranking from implicit feedback," in *Proc. AAAI Conf. Artif. Intell.*, vol. 30, no. 1, 2016.
- [7] J. Xu, Z. Chen, W. Wang, X. Hu, S.-W. Kim, and E. C. Ngai, "COHESSION: Composite graph convolutional network with dual-stage fusion for multimodal recommendation," *arXiv preprint arXiv:2504.04452*, 2025.
- [8] Y. Dang, W. Chen, Z. Pan, Y. Duan, F. Cai, and H. Chen, "Pearl: A dual-layer graph learning for multimodal recommendation," *Information Fusion*, vol. 122, p. 103168, 2025.
- [9] Y. Wei, X. d. Wang, L. Nie, X. He, R. Hong, and T.-S. Chua, "MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video," in *Proc. 27th ACM Int. Conf. Multimedia (ACM MM)*, 2019, pp. 1437–1445.
- [10] Y. Wei, X. Wang, L. Nie, X. He, and T.-S. Chua, "Graph-refined convolutional network for multimedia recommendation with implicit feedback," in *Proc. 28th ACM Int. Conf. Multimedia (ACM MM)*, 2020, pp. 3541–3549.
- [11] Q. Wang, Y. Wei, J. Yin, J. Wu, X. Song, and L. Nie, "DualGNN: Dual graph neural network for multimedia recommendation," *IEEE Trans. Multimedia*, vol. 25, pp. 1074–1084, 2021.
- [12] J. Zhang, Y. Zhu, Q. Liu, S. Wu, S. Wang, and L. Wang, "Mining latent structures for multimedia recommendation," in *Proc. 29th ACM Int. Conf. Multimedia (ACM MM)*, 2021, pp. 3872–3880.
- [13] X. Zhou and Z. Shen, "A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation," in *Proc. 31st ACM Int. Conf. Multimedia (ACM MM)*, 2023, pp. 935–943.
- [14] P. Yu, Z. Tan, G. Lu, and B.-K. Bao, "Multi-view graph convolutional network for multimedia recommendation," in *Proc. 31st ACM Int. Conf. Multimedia (ACM MM)*, 2023, pp. 6576–6585.
- [15] Y. Dang, W. Chen, Z. Pan, X. Zhang, Y. Duan, F. Cai, and H. Chen, "Dual-space feature representation learning network for multimodal recommender systems," *Advanced Engineering Informatics*, vol. 72, p. 104462, 2026.
- [16] Y. Kim, T. Kim, W.-Y. Shin, and S.-W. Kim, "MONET: Modality-embracing graph convolutional network and target-aware attention for multimedia recommendation," in *Proc. 17th ACM Int. Conf. Web Search Data Mining (WSDM)*, 2024, pp. 332–340.
- [17] J. Hu, B. Hooi, B. He, and Y. Wei, "Modality-independent graph neural networks with global transformers for multimodal recommendation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 11, 2025, pp. 11 790–11 798.
- [18] Z. Guo, J. Li, G. Li, C. Wang, S. Shi, and B. Ruan, "LGMRec: Local and global graph learning for multimodal recommendation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 8, 2024, pp. 8454–8462.
- [19] M. Xu, W. Wei, P. Yang, and H. Wu, "Semantic enhanced heterogeneous hypergraph network for collaborative filtering," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 12, 2025, pp. 12 936–12 944.
- [20] F. Meng, Z. Meng, R. Jin, R. Lin, and B. Wu, "DOGE: LLMs-enhanced hyper-knowledge graph recommender for multimodal recommendation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 12, 2025, pp. 12 399–12 407.
- [21] C. Zhang, H. Zhang, S. Wu, D. Wu, T. Xu, X. Zhao, Y. Gao, Y. Hu, and E. Chen, "Notellm-2: Multimodal large representation models for recommendation," *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, 2024.
- [22] Y. Dang, X. Zhang, Z. Pan, Y. Duan, W. Chen, F. Cai, and H. Chen, "Mllmrec: Exploring the potential of multimodal large language models in recommender systems," *arXiv preprint arXiv:2508.15304*, 2025.
- [23] Y. Ye, Z. Zheng, Y. Shen, T. Wang, H. Zhang, P. Zhu, R. Yu, K. Zhang, and H. Xiong, "Harnessing multimodal large language models for multimodal sequential recommendation," in *AAAI Conference on Artificial Intelligence*, 2024.
- [24] Y. Ji, D. Luo, C. Liu, S. Wu, J. Tong, Q. He, D. Ji, H. Lu, and Y. Ding, "Generating negative samples for multi-modal recommendation," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 6007–6016.
- [25] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, "LightGCN: Simplifying and powering graph convolution network for recommendation," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2020, pp. 639–648.
- [26] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: Bayesian personalized ranking from implicit feedback," *arXiv preprint arXiv:1205.2618*, 2012.
- [27] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.
- [28] Z. Tao, X. Liu, Y. Xia, X. Wang, L. Yang, X. Huang, and T.-S. Chua, "Self-supervised learning for multimedia recommendation," *IEEE Trans. Multimedia*, vol. 25, pp. 5107–5116, 2022.
- [29] X. Zhou, H. Zhou, Y. Liu, Z. Zeng, C. Miao, P. Wang, Y. You, and F. Jiang, "Bootstrap latent representations for multi-modal recommendation," in *Proc. ACM Web Conf. (WWW)*, 2023, pp. 845–854.
- [30] Z. Yan, Z. Li, Y. He, C. Wang, K. Li, X. Li, X. Zeng, Z. Wang, Y. Wang, Y. Qiao *et al.*, "Task preference optimization: Improving multimodal large language models with vision task alignment," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 29 880–29 892.