

A Technique to Model Product Information and User Reviews (MPIUR)

Emna Belguith*
emna.belghith@supcom.tn
Department of Computer Science and
Networks
Higher School of Communication of
Tunis
Tunis, Tunisia
LTCI, Télécom Paris, IP Paris
Palaiseau, France

Fajrian Yunus
fajrian.yunus@telecom-paris.fr
LTCI, Télécom Paris, IP Paris
Palaiseau, France

Talel Abdesslem
talel.abdesslem@telecom-paris.fr
LTCI, Télécom Paris, IP Paris
Palaiseau, France

Abstract

Modern e-commerce platforms naturally provide three complementary modalities: textual item descriptions, item images, and user reviews. However, existing multimodal recommendation methods often omit reviews or do not integrate all three modalities within a unified framework. In this paper, we propose MPIUR, a multimodal graph-based recommendation model that encodes reviews in heterogeneous graphs, aligns image and text representations, and adapts collaborative-semantic alignment to item interaction density. Experiments on the Amazon Reviews 2023 Musical Instruments and Baby categories show that MPIUR achieves the best global nDCG@10 and Recall@10 among the evaluated methods on both datasets. It also obtains the best nDCG@10 and Recall@10 on sparse items, indicating that jointly exploiting reviews and multimodal item content is particularly useful when interaction histories are limited.

CCS Concepts

• Information systems → Recommender systems; • Computing methodologies → Neural networks.

Keywords

recommender systems, collaborative filtering, graph neural networks, multimodal learning, textual content, visual content, user reviews, sparse items

ACM Reference Format:

Emna Belguith, Fajrian Yunus, and Talel Abdesslem. 2026. A Technique to Model Product Information and User Reviews (MPIUR). In *Proceedings of GRAIL 2026: Workshop on Generative, Retrieval-augmented, and Agentic Intelligence for Personalization (GRAIL @ CIKM '26)*. ACM, New York, NY, USA, 9 pages.

*The first author's home institution is the Higher School of Communication of Tunis, but this work was conducted during her graduation internship at LTCI.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

GRAIL @ CIKM '26, Rome, Italy

© 2026 Copyright held by the owner/author(s).

1 Introduction

Recommender systems rely heavily on user-item interaction data to capture user preferences. Although collaborative filtering has proven effective for popular items, its performance degrades when interaction data is sparse — a pervasive issue in e-commerce, where many products have limited interaction histories. To compensate for this weakness, recent work has integrated multimodal item content, such as textual descriptions and images. However, existing methods either ignore user reviews or fail to integrate all three naturally available modalities — user reviews, textual item descriptions, and item images — in a unified framework.

Combining collaborative and semantic signals introduces two fundamental challenges. First, the *collaborative-semantic tradeoff*: collaborative signals capture behavioral patterns from user-item interactions and excel at modeling popular items with rich interaction histories. However, for items with few interactions, these signals become unreliable and fail to produce meaningful representations. Conversely, semantic signals derived from multimodal item content can remain informative regardless of interaction volume, but relying on them excessively can obscure behavioral differences between users or items that look similar or have similar descriptions but do not necessarily attract the same users. Existing methods either prioritize one signal over the other or apply a static fusion strategy, failing to adapt to the varying interaction density of each item.

The *modality gap* is another problem: even when both collaborative and semantic representations are available, they are learned in separate optimization processes and converge to geometrically distinct regions of the embedding space. As a result, naively combining them produces inconsistent representations — the collaborative space encodes behavioral proximity while the semantic space encodes content similarity, and the two do not naturally align. This limits the model's ability to effectively leverage both signals simultaneously.

In this paper, we propose a multimodal recommendation framework that addresses both challenges simultaneously. We construct heterogeneous graphs containing users, items, aspects, and review triplets; use review embeddings as features of user-item interactions; and fuse item images and textual descriptions by per-item attention. A popularity-weighted alignment loss places a stronger collaborative-multimodal constraint on sparse items, while the remaining objectives are applied uniformly.

The main contributions of this work are as follows:

- A unified framework jointly exploiting user reviews, textual item descriptions, and item images within a heterogeneous graph structure, where reviews serve as edge features enabling fine-grained interaction modeling.
- A Modal Encoder that fuses item visual and textual representations via per-item adaptive attention and contrastive alignment.
- An adaptive alignment objective that gives greater weight to multimodal information for items with fewer interactions.
- Experiments on two Amazon datasets demonstrating the strongest and most consistent gains on less popular items, together with competitive results on popular items.

2 Related Work

2.1 Recommender Systems

Two classical paradigms underpin recommender systems: collaborative filtering and content-based filtering.

Collaborative filtering is based on the intuition that users who have interacted with similar items in the past likely have similar preferences. A system can therefore recommend items previously selected by similar users. Interactions may take various forms, including purchases, clicks, ratings, and reviews. In this work, we focus on interactions associated with reviews, which provide richer semantic information than clicks or purchases alone. A classical collaborative-filtering technique is matrix factorization [8]. More recent approaches use graph neural networks (GNNs), representing users and items as nodes and interactions as edges [17, 18]. The main limitation of collaborative filtering is its dependence on interaction history. Performance consequently deteriorates for users or items with limited histories; the complete cold-start case occurs when no prior interaction is available.

Content-based filtering works by predicting whether a user would like an item based on the characteristics of the item itself. Items are described by their attributes, and the recommender system learns a score representing how well a particular item matches a user's preferences. Unlike collaborative filtering, content-based filtering can still recommend items even when interaction history is sparse, since it relies on item attributes rather than past interactions. Some classic techniques use keywords and TF-IDF of textual descriptions as the item attributes [7, 16]. With the rise of deep learning, item attributes are commonly represented as dense embedding vectors capturing multiple modalities [21]. In e-commerce platforms, the most naturally available and informative item attributes are textual descriptions, images, and user reviews. While textual descriptions and images characterize the product independently of any user experience, reviews are directly tied to user-item interactions: they reflect what users actually think about an item after purchasing or using it.

These two paradigms are not mutually exclusive. Modern recommendation methods increasingly combine both collaborative signals and semantic content to overcome their individual limitations [6, 9]. This hybrid direction motivates our work, which jointly leverages user-item interaction graphs and multimodal item representations.

2.2 Graph-based Recommendation

Graph neural networks (GNNs) have demonstrated strong performance in recommender systems by modeling user-item interactions as a bipartite graph and propagating collaborative signals through higher-order neighborhoods. By iteratively aggregating information from neighboring nodes, GNN-based methods capture complex interaction patterns that traditional collaborative-filtering approaches may fail to model. LightGCN [4] reduces model complexity by removing nonlinear transformations and feature-transformation matrices, retaining only neighborhood aggregation. Despite its simplicity, it provides strong recommendation performance and has become a widely adopted baseline. However, such methods typically operate on homogeneous graphs with untyped edges and treat all interactions equally. They therefore fail to exploit the semantic information carried by ratings and reviews associated with user-item interactions. This limitation motivates heterogeneous graph structures with typed and semantically enriched edges.

2.3 Review-based Recommendation

User reviews represent a valuable source of information for recommendation, as they contain detailed opinions about specific item aspects and directly reflect user preferences. Early studies employed deep text encoders to extract semantic information from reviews, including DeepCoNN [22], which uses parallel TextCNNs to capture semantic features, and NARRE [1], which leverages attention to score the usefulness of each review. However, these methods treat reviews as static text features, failing to capture the fine-grained aspect-level preferences they contain. More recently, LLMs have been leveraged to extract structured (aspect, opinion, sentiment) triplets from reviews, enabling richer semantic modeling at the aspect level [12].

Beyond encoding reviews, a key question is how to integrate them into the recommendation system model. TGNN [15] proposes to incorporate reviews directly into the graph structure, using review embeddings as attention weights on the edges of a heterogeneous user-item graph. This allows review-dependent weights to influence the aggregated representations. However, TGNN is optimized for rating prediction rather than ranking. In real-world recommendation scenarios, the objective is to produce a ranked list of relevant items. Whether a product is assigned a score of 4.5 or 2.5 is less important than whether that relevant products are ranked above irrelevant ones. This misalignment motivates us to inherit the heterogeneous graph structure of TGNN while replacing the loss function with the Bayesian Personalized Ranking (BPR) loss which optimizes for the ranking instead of the score.

2.4 Multimodal Recommendation

Early multimodal recommendation methods demonstrated that incorporating product features can improve recommendation accuracy beyond collaborative signals alone. VBPR [3] pioneered this direction by integrating visual features into the Bayesian Personalized Ranking framework, showing that visual item content provides meaningful signals for preference modeling. Building on this idea, BM3 [24] proposes a self-supervised framework that projects image and text features into a shared latent space and aligns them with collaborative signals. CFMM [20] further introduces a softmax-based

attention mechanism to weight image and text contributions dynamically for each item, together with modality-specific k -nearest-neighbor item graphs that capture inter-item similarities. Printf [10] integrates all three modalities—item text, item images, and user reviews—but encodes reviews separately as user features through a dimension-wise attention mechanism.

2.5 Collaborative-Semantic Alignment

A key challenge in multimodal recommendation is bridging the gap between collaborative and semantic representation spaces. Recent work has begun to address this issue from different angles. Learning Item Representations Directly from Multimodal Features [23] observes that item ID embeddings are initialized randomly while multimodal features are already informative from the start of training, causing the two spaces to diverge. They propose to initialize item embeddings directly from multimodal representations to reduce this gap from the outset. CFMM [20] introduces a fusion loss inspired by Bayesian Personalized Ranking that encourages multimodal item representations to reflect user-preference rankings, ensuring that semantic features remain useful for recommendation beyond their role as auxiliary signals. This ranking objective is applied uniformly across items. GateSID [25] instead uses a gate-regulated contrastive mechanism to adapt semantic-behavioral alignment to item popularity. Our approach combines a uniformly applied multimodal ranking loss with a separate popularity-weighted alignment objective that places a stronger constraint on sparse items.

3 Methodology

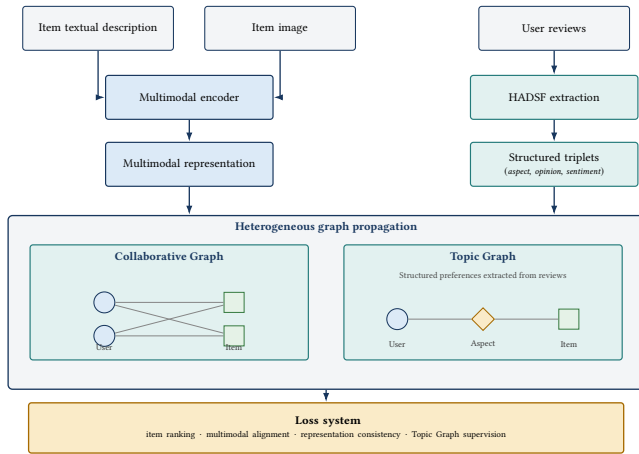


Figure 1: Overview of the proposed architecture.

Figure 1 illustrates the overall architecture of the proposed framework. Given users, items, and their associated multimodal content, the model processes three complementary modalities and integrates them through two parallel graph encoders and an adaptive loss system to produce top- K recommendations.

For the review modality, raw user reviews are processed by an LLM-based extraction mechanism to derive structured (aspect, opinion, sentiment) triplets, capturing fine-grained user preferences

beyond simple rating scores. The extracted triplets and their embeddings serve as inputs to the graph encoders. For the visual and textual item modalities, item images and textual descriptions are independently encoded by pretrained encoders and fed into a Modal Encoder, which fuses them via learned per-item attention weights. An image-text alignment loss encourages the two modalities of the same item to be mapped to nearby representations in the shared latent space.

The model employs two parallel graph encoders operating on heterogeneous graph structures. The Collaborative Graph represents each interaction by bidirectional user-item edges. Review information is used to weight message passing and is injected into the second propagation layer, producing collaborative user and item representations. In parallel, the Topic Graph connects review triplets, aspect nodes, users, and items to model fine-grained preferences at the aspect level.

The model is trained with a BPR ranking loss, an image-text alignment loss, a popularity-weighted collaborative-multimodal alignment loss, a ranking loss applied to the multimodal item representations, and a mutual-information objective that retrieves the triplet associated with each user-item interaction. The resulting relevance scores are used to rank candidate items and produce top- K recommendations.

3.1 Review Processing

Following HADSF [12], we adopt a two-stage LLM-based pipeline, implemented with Llama 3.1 [2], to extract structured semantic information from raw user reviews.

In the first stage, reviews are recursively summarized at the item level. The LLM extracts candidate themes from these summaries, after which semantically similar themes are clustered to construct a global aspect vocabulary $\mathcal{A}^*$. This vocabulary provides a consistent set of aspects across users and items.

In the second stage, each review is processed together with $\mathcal{A}^*$ and contextual information obtained from the corresponding user’s review history and the item’s review history. The LLM extracts one or more (aspect, opinion, sentiment) triplets. The complete review and each extracted triplet are encoded separately with a pretrained BERT encoder and projected to 128 dimensions. The complete-review representation $\mathbf{r}_{ui}$ is used as a user-item edge feature in the Collaborative Graph, whereas the triplet representation $\mathbf{t}_{ui}$ is used as a node feature and supervision target in the Topic Graph.

3.2 Modal Encoder

To obtain a unified item representation from visual and textual modalities, we propose a Modal Encoder that fuses pre-extracted image and text features into a shared semantic space. Item images are encoded by a pretrained ResNet50, and textual item descriptions are encoded by a pretrained MiniLM SentenceTransformer. Both encoders are frozen during training — only the projection layers are learned.

Each modality is projected into a shared latent space via a linear transformation followed by a nonlinear activation and layer normalization, producing modality-specific representations $\mathbf{h}_v$ and $\mathbf{h}_t$ for image and text, respectively.

Inspired by CFMM [20], we adopt a per-item attention mechanism to adaptively fuse the two modalities. For each item, attention weights are computed over the image and text representations and normalized via softmax:

$$[w_{img}, w_{txt}] = \text{Softmax}([\text{Query}(\mathbf{h}_v), \text{Query}(\mathbf{h}_t)]) \quad (1)$$

The final fused representation is then obtained as:

$$\mathbf{h}_{modal} = w_{img} \cdot \mathbf{h}_v + w_{txt} \cdot \mathbf{h}_t \quad (2)$$

This per-item weighting allows the model to learn different image and text contributions for different items. The learned weights indicate which projected modality is most useful to the training objective; they do not, by themselves, measure the intrinsic quality or completeness of either modality.

To enforce semantic coherence between the two modalities, we apply an InfoNCE contrastive loss between $\mathbf{h}_v$ and $\mathbf{h}_t$, detailed in Section 3.4. This objective encourages the projected image and text representations of the same item to be close in the shared latent space before they are fused into $\mathbf{h}_{modal}$. The fused modal representation is also used to initialize the item embedding before graph training, reducing the initial discrepancy between semantic and collaborative representations.

3.3 Graph Encoding

The proposed framework relies on two parallel heterogeneous graphs, both adapted from TGNN [15]. The first, referred to as the **Collaborative Graph**, models user-item interactions enriched with review embeddings as edge features. The second, referred to as the **Topic Graph**, captures fine-grained semantic preferences at the aspect level through topic nodes derived from the LLM extraction pipeline. We adapt both graphs to a ranking framework by replacing the original rating prediction objective with a BPR ranking loss. Both graphs share the same user and item nodes but differ in their edge structure and encoding mechanism.

3.3.1 Collaborative Graph Encoding. The collaborative graph connects users and items through bidirectional interaction edges, where each edge carries the BERT embedding of the associated review as a feature. For each interaction edge (u, i) , a learnable scalar p_{ui} is initialized from the review embedding and updated during training. The first-layer message from user u to item i is:

$$\mathbf{m}_{u \rightarrow i}^{(1)} = p_{ui} \mathbf{h}_u^{(0)} \quad (3)$$

The learned coefficient controls the contribution of the interaction to message passing; it should not be interpreted as an independently validated measure of review informativeness.

In the second GCN layer, the review embedding is additionally injected directly into the message:

$$\mathbf{m}_{u \rightarrow i}^{(2)} = p_{ui} \mathbf{h}_u^{(1)} + s_{ui} \mathbf{r}_{ui} \quad (4)$$

where $\mathbf{r}_{ui}$ is the review representation and s_{ui} is a learned coefficient controlling its contribution. After two GCN layers with symmetric normalization, the final collaborative representations $\mathbf{u}_{rf} \in \mathbb{R}^{128}$ and $\mathbf{i}_{rf} \in \mathbb{R}^{128}$ are obtained for users and items, respectively.

3.3.2 Topic Graph Encoding. The topic graph connects review-triplet nodes to user- and item-specific aspect nodes, which are in turn connected to user and item nodes. This structure allows the model to capture fine-grained semantic preferences at the aspect level, going beyond the binary signal of whether an interaction occurred. The topic graph produces topic-level representations $\mathbf{u}_{tf}$ and $\mathbf{i}_{tf}$ for users and items, respectively, which are combined with the collaborative representations in the optimization objective.

3.4 Learning and Optimization

3.4.1 BPR Loss. The primary training objective is Bayesian Personalized Ranking (BPR) [14], which directly optimizes the relative ranking of items. For each user u , the model learns to score a positively interacted item i^+ higher than an unobserved item i^- :

$$\mathcal{L}_{BPR} = -\frac{1}{|\mathcal{B}|} \sum_{(u, i^+, i^-) \in \mathcal{B}} \log \sigma(\hat{y}_{u, i^+} - \hat{y}_{u, i^-}) \quad (5)$$

where $\hat{y}_{u, i} = \mathbf{u}_r \cdot \mathbf{i}_r$ is the score computed from the Collaborative Graph representations and σ is the sigmoid function. An L2 regularization term is added to prevent overfitting.

3.4.2 Modal Alignment Loss. To align the visual and textual representations of the same item, we apply a batch-wise cross-entropy contrastive loss between $\mathbf{h}_v$ and $\mathbf{h}_t$. For each item in the batch, its image representation must be closer to its own text representation than to those of all other items in the batch:

$$\mathcal{L}_{modal} = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \log \frac{\exp(\mathbf{h}_{v, i} \cdot \mathbf{h}_{t, i})}{\sum_{j \in \mathcal{B}} \exp(\mathbf{h}_{v, i} \cdot \mathbf{h}_{t, j})} \quad (6)$$

where $\mathbf{h}_{v, i}$ and $\mathbf{h}_{t, i}$ are the image and text representations of item i . This objective encourages the image and text representations of the same item to be close in the shared latent space.

3.4.3 Modality Gap Loss. To bridge the gap between collaborative and semantic representation spaces, we introduce a contrastive alignment loss that pulls item embeddings closer to their modal representations $\mathbf{h}_{modal}$:

$$\mathcal{L}_{MG} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} -(1 - \alpha_i) \log \frac{\exp(\mathbf{h}_i^{(0)} \cdot \mathbf{h}_{modal, i})}{\sum_{j \in \mathcal{B}} \exp(\mathbf{h}_i^{(0)} \cdot \mathbf{h}_{modal, j})} \quad (7)$$

This loss is weighted by $(1 - \alpha_i)$, where α_i is the log-normalized training degree:

$$\alpha_i = \frac{\log(1 + \deg(i))}{\log(1 + \deg_{\max})}. \quad (8)$$

Items with few interactions contribute more to this alignment, ensuring that their collaborative representations remain anchored to the semantically rich modal signal.

3.4.4 Fusion Loss. Inspired by CFMM [20], this loss ensures that the modal representations $\mathbf{h}_{modal}$ remain directly predictive of user preferences rather than serving only as an alignment target. We apply a BPR objective directly on $\mathbf{h}_{modal}$:

$$\mathcal{L}_F = \frac{1}{|\mathcal{B}|} \sum_{(u, i^+, i^-) \in \mathcal{B}} -\log \sigma(\mathbf{u}_r \cdot \mathbf{h}_{modal, i^+} - \mathbf{u}_r \cdot \mathbf{h}_{modal, i^-}) \quad (9)$$

where the objective is applied uniformly to the sampled training triples. Unlike the Modality Gap Loss, the Fusion Loss is not

weighted by item popularity; it makes the modal representations directly predictive of the ranking objective for all items.

3.4.5 MI Loss. To supervise the Topic Graph, collaborative and topic representations are first combined as $\tilde{\mathbf{u}} = \mathbf{u}_r + \mathbf{u}_t$ and $\tilde{\mathbf{i}} = \mathbf{i}_r + \mathbf{i}_t$. Their concatenation is projected into an interaction representation $\mathbf{th}_{u,i} = f([\tilde{\mathbf{u}}\|\tilde{\mathbf{i}}])$. The mutual-information objective then retrieves the review triplet associated with each user–item interaction from the triplets of the other interactions in the batch:

$$\mathcal{L}_{MI} = -\frac{1}{|\mathcal{B}|} \sum_{(u,i) \in \mathcal{B}} \log \frac{\exp(\mathbf{th}_{u,i} \cdot \mathbf{t}_{u,i})}{\sum_{(u',i') \in \mathcal{B}} \exp(\mathbf{th}_{u,i} \cdot \mathbf{t}_{u',i'})} \quad (10)$$

where $\mathbf{t}_{u,i}$ is the embedding of the corresponding extracted triplet. To avoid interfering with the primary ranking objective, this loss is introduced gradually via a linear warmup over the first training epochs.

The complete training objective is:

$$\mathcal{L} = \mathcal{L}_{BPR} + \lambda_{modal} \mathcal{L}_{modal} + \lambda_{mg} \mathcal{L}_{MG} + \mathcal{L}_F + \lambda_{MI} \mathcal{L}_{MI}. \quad (11)$$

Only $\mathcal{L}_{MG}$ contains the item-dependent factor $(1 - \alpha_i)$.

4 Experimental Settings

4.1 Datasets

We evaluate our model on two product categories from the Amazon Reviews 2023 dataset [5]: Musical Instruments and Baby. Both categories provide product images, textual descriptions, and user reviews simultaneously. Since our framework requires each user–item interaction to be associated with a review, we retain only interactions that include a review, ensuring that each edge in the heterogeneous graph carries a review embedding.

Each dataset is split into 80% for training, 10% for validation, and 10% for testing. To ensure a meaningful evaluation, we retain only users and items with at least one interaction in the training set, guaranteeing that all retained nodes have a collaborative signal during training. The dataset statistics are summarized in Table 1.

Table 1: Dataset statistics.

	Musical Instruments	Baby
Users	17 318	25 474
Items	4 886	5 492
Interactions	58 469	70 205
Sparsity	99.93%	99.95%

4.2 Evaluation Protocol and Metrics

4.2.1 Item Ranking. We use Recall@K and normalized Discounted Cumulative Gain (nDCG@K), with $K = 10$. Recall@K measures the fraction of relevant items retrieved in the top-K list, while nDCG@K additionally accounts for the ranking position of relevant items, assigning higher scores to relevant items appearing earlier in the list.

For each test interaction, the observed item is ranked against 99 items sampled from those absent from the user’s interaction

history. Because each candidate set contains one positive item, Recall@10 indicates whether that item appears in the top 10, whereas nDCG@10 additionally rewards a higher rank within the top 10.

This sampled protocol reduces evaluation cost and uses the same negative candidates for all models, but it does not constitute an exhaustive ranking over the complete item catalog.

4.2.2 HADSF Extraction Quality. We evaluate the reliability of the extracted triplets using two metrics. The Aspect Drift Rate (ADR) is the proportion of extracted aspects that do not belong to the global aspect vocabulary $\mathcal{A}^*$; lower values indicate less aspect drift. The Opinion Fidelity Rate (OFR) is the average cosine similarity between an extracted opinion and its source review; higher values indicate greater semantic consistency with the original review.

4.2.3 Triplet Ranking. To evaluate the Topic Graph independently, the topic representations $\mathbf{u}_t$ and $\mathbf{i}_t$ are concatenated and projected to obtain $\mathbf{s}_{ui}^t = f_t([\mathbf{u}_t\|\mathbf{i}_t])$. This representation is compared with the triplet embedding of every test interaction. The target triplet is treated as positive and all other test triplets as negatives; Recall@10 and nDCG@10 are then computed from the resulting ranking.

Items are further divided into two groups based on their interaction density: items with fewer than 10 interactions are considered *sparse items*, while items with 10 or more interactions are considered *popular items*. These sparse items have limited but nonzero training histories and therefore differ from the complete cold-start setting. Performance is reported separately for each group, allowing a fine-grained analysis of each model’s ability to handle items of varying popularity.

4.3 Implementation Details

All experiments are implemented in PyTorch and run on a GPU cluster. For item visual features, we use pre-extracted ResNet50 average pooling features projected to 128 dimensions via a single-layer MLP. For textual item features, we use pre-extracted embeddings from all-MiniLM-L6-v2 (384-dim), similarly projected to 128 dimensions. Both feature extractors are frozen during training — only the projection layers are learned.

The heterogeneous user–item graph is constructed with two directed edge types: (user $\rightarrow$ item) and (item $\rightarrow$ user), both enriched with the corresponding review embedding. The Collaborative Graph uses two layers and an embedding dimension of 128.

For optimization, we use Adam with a learning rate of 0.001, L2 regularization of 10^{-4} , and dropout of 0.8. The batch size is set to 512 with $n_{neg} = 5$ negative items sampled per positive interaction. Loss weights are set to $\lambda_{modal} = 0.1$, $\lambda_{mg} = 0.1$, and $\lambda_{MI} = 0.01$ with a linear warmup of 100 epochs. Training uses early stopping with a patience of 100 epochs.

Unless explicitly stated otherwise, MPIUR refers to the configuration using ResNet50 as the visual encoder. This configuration is used for the global, group-wise, ablation, and hyperparameter results; the MLLM variants are evaluated separately.

5 Results

5.1 Overall Performance

Table 2 compares MPIUR with the three baselines globally and within the sparse- and popular-item groups.

Table 2: Performance comparison in terms of nDCG@10 and Recall@10 across all item groups.

Model	Musical Instruments						Baby					
	Global		Sparse items		Popular items		Global		Sparse items		Popular items	
	N@10	R@10	N@10	R@10	N@10	R@10	N@10	R@10	N@10	R@10	N@10	R@10
LightGCN	0.1613	0.2577	0.1142	0.1943	0.2217	0.3389	0.1636	0.2984	0.1230	0.2255	0.2013	0.3660
TGNN	0.2249	0.3990	0.2289	0.4152	0.2198	0.3782	0.1977	0.3538	0.1941	0.3498	0.2010	0.3575
BM3	0.2276	0.3955	0.2183	0.3870	0.2395	0.4065	0.2025	0.3859	0.2048	0.3976	0.2004	0.3751
MPIUR	0.2344	0.4042	0.2347	0.4153	0.2340	0.3900	0.2194	0.3891	0.2384	0.4161	0.2018	0.3640

Comparison with LightGCN. LightGCN relies only on the interaction structure. MPIUR obtains substantially higher global scores on both datasets. The largest nDCG@10 differences occur for sparse items: +0.1205 on Musical Instruments and +0.1154 on Baby. MPIUR also performs better on popular items except for Recall@10 on Baby, where LightGCN obtains 0.3660 versus 0.3640.

Comparison with TGNN. TGNN is the direct graph backbone: it already uses reviews as edge features but does not include the visual and textual item modalities or the proposed loss system. We replace the original MSE loss function with the BPR loss function. MPIUR obtains higher global scores on both datasets. On sparse Musical Instruments items, Recall@10 is essentially unchanged (0.4153 versus 0.4152), while nDCG@10 increases from 0.2289 to 0.2347. The improvements are larger on sparse Baby items, where MPIUR reaches 0.2384 nDCG@10 and 0.4161 Recall@10, compared with 0.1941 and 0.3498 for TGNN. MPIUR also remains slightly better on the popular-item metrics.

Comparison with BM3. BM3 uses visual and textual item features but does not incorporate review content. MPIUR achieves higher global nDCG@10 and Recall@10 on both datasets and performs better on all sparse-item metrics. The pattern changes for popular items: BM3 performs better on both Musical Instruments metrics and on Baby Recall@10, whereas MPIUR obtains the higher Baby nDCG@10.

5.2 Review Extraction and Topic Graph Evaluation

Table 3 reports the quality of the triplets extracted by HADSF. The ADR is 0.0125 on Musical Instruments and 0.0126 on Baby. The corresponding OFR values are 0.9358 and 0.9375.

Table 3: Quality of the HADSF extraction.

Dataset	OFR	ADR
Musical Instruments	0.9358	0.0125
Baby	0.9375	0.0126

Table 4 evaluates the Topic Graph on triplet ranking against approximately 5,200 candidates per interaction. For a random ranking over a candidate pool of this size, the expected Recall@10 is approximately 0.0019 and the expected nDCG@10 is approximately 0.0087. Musical Instruments obtains 0.0527 Recall@10 and 0.0314 nDCG@10, while Baby obtains 0.0601 and 0.0343, respectively. Both

datasets therefore score substantially above the random-ranking baseline and obtain similar values. An error analysis shows that, in approximately 65% of failed cases, the model retrieves the correct product aspect but assigns a different opinion to it; the remaining errors mainly confuse different aspects of the same product.

Table 4: Topic Graph performance on triplet ranking.

Dataset	Recall@10	nDCG@10
Musical Instruments	0.0527	0.0314
Baby	0.0601	0.0343
Random (expected)	0.0019	0.0087

5.3 Ablation Studies

Table 5: Ablation study of the proposed model in terms of nDCG@10 on the “Musical Instruments” dataset.

Variant	Global	Sparse Items	Popular Items
Full model	0.2344	0.2347	0.2340
w/o Modality Gap Loss	0.2135	0.2126	0.2147
w/o Fusion Loss	0.2150	0.2130	0.2176
w/o Image & Text Features	0.2177	0.2176	0.2178
w/o Review Features	0.1972	0.1869	0.2104
w/o All Modalities	0.1919	0.1814	0.2053

To assess the contribution of each component, we conduct an ablation study by removing one component at a time from the full model on the “Musical Instruments” dataset. The variants are defined as follows:

- **w/o Modality Gap Loss:** the contrastive alignment between item embeddings and modal representations is removed.
- **w/o Fusion Loss:** the BPR objective applied directly on modal representations is removed, meaning $\mathbf{h}_{\text{modal}}$ is no longer explicitly guided toward the ranking objective.
- **w/o Image & Text Features:** image and text features are excluded, removing the Modal Encoder entirely, as well as the Fusion Loss, Modality Gap Loss, and Modal Alignment Loss.
- **w/o Review Features:** review embeddings are removed from the graph edges; the Topic Graph and its associated MI Loss are also removed.

- **w/o All Modalities:** all three modalities — image, text, and reviews — are removed simultaneously, reducing the model to a pure collaborative filtering baseline.

Table 5 shows that removing review features produces the largest decrease among the individual modality removals in all three groups. The sparse-item score decreases from 0.2347 to 0.1869, whereas removing image and text features decreases it to 0.2176. Removing all modalities produces the lowest scores: 0.1919 globally, 0.1814 for sparse items, and 0.2053 for popular items.

Removing the Modality Gap Loss decreases nDCG@10 to 0.2135 globally, 0.2126 for sparse items, and 0.2147 for popular items. Removing the Fusion Loss yields 0.2150, 0.2130, and 0.2176, respectively. Thus, every tested ablation performs below the full model in every reported group.

5.4 Modality Gap Analysis

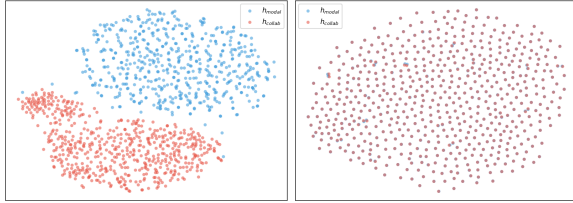


Figure 2: t-SNE visualization of collaborative and modal embeddings before and after alignment on a sample of the Musical Instruments dataset.

Figure 2 illustrates the effect of our alignment strategy on the spatial distribution of collaborative and modal embeddings. Before alignment, the two types of representations occupy visibly distinct regions. After training, the two sets of points overlap more in the two-dimensional projection. This figure is a qualitative visualization and does not measure separation in the original high-dimensional space.

5.5 Hyperparameter Sensitivity

5.5.1 Number of Negative Samples. The number of negative samples per positive interaction n_{neg} is tuned over a range from 1 to 64. We select $n_{neg} = 5$ based on validation performance. The validation score reaches its maximum at five negatives and decreases for every tested value above five.

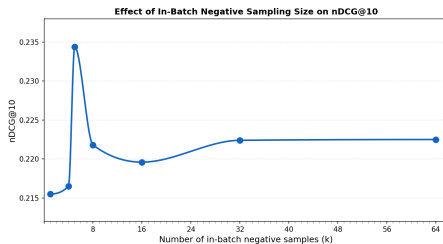


Figure 3: Sensitivity to the number of negative samples k on the Musical Instruments validation set.

5.5.2 Sparse-Item Threshold Analysis. Item interaction counts in our datasets range from 1 to nearly 500. We evaluate $T = 10$ as the operational threshold for defining sparse items. Figure 4 shows the nDCG@10 of our full model and its variant without any modality (w/o All Modalities) across item degree bins.

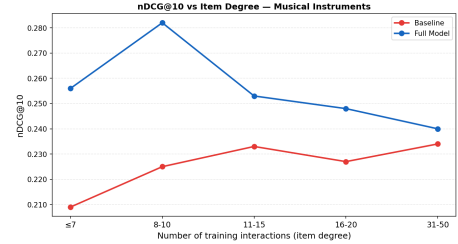


Figure 4: nDCG@10 of the full model and the variant without modalities across item-degree bins on Musical Instruments.

The full model outperforms the variant without modalities in every displayed bin. The largest observed gap occurs in the 8–10 bin. The gap is smaller in every bin above 10 interactions and reaches its minimum in the 31–50 bin. We use 10 interactions as the operational threshold for the group-wise evaluation.

6 VLM-based Visual Encoding Experiment

Large Language Models (LLMs) and Multimodal LLMs (MLLMs) have demonstrated a strong ability to capture rich semantic information that pretrained encoders such as CNNs often fail to expose, particularly attributes that are visually implicit but not explicitly stated in item metadata [19, 23]. Motivated by this, recent work has explored replacing traditional visual/textual feature extractors with MLLM-based pipelines to obtain item representations more directly aligned with the recommendation objective, rather than relying on generic embeddings [19, 23]. Building on this direction, we investigate whether replacing the frozen CNN-based image encoder in our Modal Encoder with an MLLM can yield item representations that better capture recommendation-relevant visual semantics.

We replace the frozen ResNet50 image encoder with VLM-based pipelines using Qwen2.5-VL-7B-Instruct [13] and LLaVA-13B [11]. Item images are first converted into textual descriptions, and the resulting descriptions are encoded with the same MiniLM text encoder used elsewhere in the framework. The remainder of the Modal Encoder architecture is unchanged.

Prompt Design. The prompt is designed to elicit visual attributes that are implicit in the product image but absent from the item’s original textual metadata, allowing the generated description to serve as a complementary signal rather than a redundant one. To reduce hallucinations, the model is explicitly instructed to describe only information that can be inferred from the image itself. The prompt is further contextualized with the item’s category ({DATASET}), guiding the model toward domain-specific attributes instead of producing a generic description. Finally, the prompt follows an instructional format by explicitly specifying the aspects to cover—core function, key features, target user, style and design, and quality and materials—to encourage consistent descriptions

across items. The output is also limited to 100 words, producing concise captions suitable for downstream text encoding.

Provide a brief product description based on the image, focusing on the key features and benefits relevant to the {DATASET} category.

- **Core Function:** What is the primary purpose of this product?
- **Key Features:** What are the standout features that set it apart from similar products?
- **Target User:** Who is the ideal user or buyer for this product?
- **Style and Design:** What are the product’s overall appearance, color, and design?
- **Quality & Materials:** What can be inferred about the materials or build quality?

Answer in one paragraph of no more than 100 words.

Table 6: Global nDCG@10 comparison between CNN-based and VLM-based visual encoding.

Dataset	ResNet50	Qwen2.5-VL	LLaVA-13B
Musical Instruments	0.2344	0.2381	0.2316
Baby	0.2194	0.2113	0.2091

Table 6 reports the results. On Musical Instruments, Qwen2.5-VL obtains a small improvement over ResNet50, whereas LLaVA-13B remains below the ResNet50 baseline. On Baby, ResNet50 outperforms both VLM variants. Qwen2.5-VL is better than LLaVA-13B on both datasets.

7 Limitations

Several limitations should be considered when interpreting the results. First, preprocessing retains only interactions associated with a textual review. Purchases, ratings, and other interactions without comments are excluded even though they contain useful collaborative information, which may bias the data toward users who write reviews frequently. Second, every retained item must have at least one training interaction; the model therefore addresses sparse items rather than the complete item cold-start setting. Third, the additional modalities enrich primarily the item representations and do not provide comparably rich information about the users.

The ranking formulation also treats every observed interaction as a positive example regardless of its rating. Consequently, one-star and five-star interactions contribute equally to the BPR objective. Moreover, the evaluation ranks each positive item against 99 sampled negatives rather than the complete catalog. Although all models use the same candidates, full-catalog evaluation is required to confirm the findings under a more exhaustive protocol. Finally, the VLM experiment shows that generated visual descriptions are domain-dependent and cannot systematically replace CNN features.

8 Conclusion

This paper introduced MPIUR, a graph-based recommendation framework that jointly exploits user reviews, textual item descriptions, and item images. It combines a review-enriched Collaborative Graph, an aspect-level Topic Graph, and an attention-based Modal

Encoder. Its objectives align collaborative and semantic representations while giving stronger alignment weight to items with limited interaction histories.

On the Amazon Reviews 2023 Musical Instruments and Baby categories, MPIUR obtains the best global nDCG@10 and Recall@10 among the evaluated methods (Table 2). Its strongest and most consistent performances occur for sparse items. We interpret this pattern as evidence that semantic content is most useful as a complement when collaborative evidence is limited. The ablation results (Table 5) further suggest that the three modalities provide complementary information: reviews yield the largest individual gain, particularly for sparse items, while image and text features provide smaller gains across all groups. The larger sparse-item degradation after removing the Modality Gap Loss is also consistent with the intended popularity-aware alignment mechanism. We also observe that the removal of modality gap loss or fusion loss affects the performance worse than the removal of image & text features. It shows that extra input modalities do not necessarily improve performance if they are not “aligned” to each other. Unintuitively, the removal of review features degrade the performance on sparse items worse than on popular items, even though we expect that the prediction for the sparse items rely more on the image and text features. A possible reason is that for sparse items, each review strongly affects the prediction, thus their removal affects the performance badly.

The results on popular items require a more cautious interpretation. BM3 remains better on three of the four popular-item metrics, which may indicate that review information becomes redundant or noisy when an item already has a rich interaction history. This explanation remains a hypothesis because the present experiments do not directly measure review redundancy or quality by popularity. Similarly, the VLM experiment (Table 6) shows that generated visual descriptions are not uniformly superior to CNN features: Qwen2.5-VL improves the performance on Musical Instruments, whereas ResNet50 CNN remains better on Baby. Finally, the HADSf scores (Table 3) suggest high opinion fidelity and low aspect drift, while the Topic Graph error analysis (Section 5.2) suggests that the model identifies aspects more reliably than exact opinions. Lastly, the fact that the Topic Graph scores are considerably higher than the random-ranking baseline (Table 4) suggests that the graph captures a non-random retrieval signal despite the much larger candidate pool.

The present study remains limited to reviewed interactions, sampled ranking evaluation, and items observed during training. Future work will reintegrate interactions without reviews, incorporate rating signals, enrich user-side representations, combine CNN and VLM visual features, and extend the framework to complete item cold-start.

Acknowledgments

The authors thank the academic supervisors at SUP’COM for their valuable feedbacks.

References

- [1] Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2018. Neural Attentional Rating Regression with Review-level Explanations. In *Proceedings of the 2018 World Wide Web Conference*. 1583–1592.
- [2] Abhimanyu Dubey et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI]

- [3] Ruining He and Julian McAuley. 2016. VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.
- [4] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 639–648.
- [5] Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging Language and Items for Retrieval and Recommendation. arXiv:2403.03952 [cs.IR]
- [6] Jun Hu, Bryan Hooi, Bingsheng He, and Yinwei Wei. 2025. Modality-independent graph neural networks with global transformers for multimodal recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 11790–11798.
- [7] Michal Kompan and Mária Bieliková. 2010. Content-based news recommendation. In *International Conference on Electronic Commerce and Web Technologies*. Springer, 61–72.
- [8] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.
- [9] Jianing Li, Chaoqun Yang, Guanhua Ye, and Quoc Viet Hung Nguyen. 2024. Graph neural networks with deep mutual learning for designing multi-modal recommendation systems. *Information Sciences* 654 (2024), 119815.
- [10] Hao-Lun Lin, Jyun-Yu Jiang, Ming-Hao Juan, and Pu-Jen Cheng. 2023. printf: Preference Modeling based on user reviews with item images and textual information via graph learning. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 1431–1440.
- [11] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *NeurIPS*.
- [12] Zheng Nie and Peijie Sun. 2026. HADSF: Aspect Aware Semantic Control for Explainable Recommendation. In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*. 1–11. doi:10.1145/3773966.3778013
- [13] Qwen Team. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923 [cs.CV]
- [14] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. 452–461.
- [15] Jie Shuai, Le Wu, Kun Zhang, Peijie Sun, Richang Hong, and Meng Wang. 2023. Topic-enhanced Graph Neural Networks for Extraction-based Explainable Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1–11. doi:10.1145/3539618.3591776
- [16] Jong-Soo Sohn, Un-Bong Bae, and In-Jeong Chung. 2013. Contents recommendation method using social network analysis. *Wireless Personal Communications* 73, 4 (2013), 1529–1546.
- [17] Qifan Wang, Yinwei Wei, Jianhua Yin, Jianlong Wu, Xuemeng Song, and Liqiang Nie. 2021. Dualgnn: Dual graph neural network for multimedia recommendation. *IEEE Transactions on Multimedia* 25 (2021), 1074–1084.
- [18] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM international conference on multimedia*. 1437–1445.
- [19] Yuyang Ye, Zhi Zheng, Yishan Shen, Tianshu Wang, Hengruo Zhang, Peijun Zhu, and Hui Xiong. 2025. Harnessing Multimodal Large Language Models for Multimodal Sequential Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. doi:10.1609/aaai.v39i12.33426
- [20] Chong Zhang and ZhiCai Zhang. 2025. Cross-fusion activates deep modal integration for multimedia recommendation. *PLOS ONE* 20, 7 (2025), e0327663. doi:10.1371/journal.pone.0327663
- [21] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Mengqi Zhang, Shu Wu, and Liang Wang. 2022. Latent structure mining with contrastive modality fusion for multimedia recommendation. *IEEE Transactions on Knowledge and Data Engineering* 35, 9 (2022), 9154–9167.
- [22] Lei Zheng, Vahid Noroozi, and Philip S. Yu. 2017. Joint Deep Modeling of Users and Items Using Reviews for Recommendation. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. 425–434.
- [23] Xin Zhou, Xiaoxiong Zhang, Dusit Niyato, and Zhiqi Shen. 2025. Learning Item Representations Directly from Multimodal Features for Effective Recommendation. arXiv:2505.04960 [cs.IR] doi:10.48550/arXiv.2505.04960
- [24] Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. 2023. Bootstrap Latent Representations for Multimodal Recommendation. In *Proceedings of the ACM Web Conference 2023*. 845–854. doi:10.1145/3543507.3583251
- [25] Hai Zhu, Yantao Yu, Lei Shen, Bing Wang, and Xiaoyi Zeng. 2026. GateSID: Adaptive Gating for Semantic-Collaborative Alignment in Cold-Start Recommendation. arXiv:2603.22916 [cs.IR] doi:10.48550/arXiv.2603.22916

GenAI-Usage Disclosure

Generative AI assistants (Claude by Anthropic as well as Mistral by Mistral AI) was used in a limited, non-authorial capacity during the preparation of this work: for minor code corrections and debugging assistance, for LaTeX formatting and compilation troubleshooting, and for language/wording suggestions during manuscript editing. The research design, model architecture, experiments, result analysis, and scientific claims presented in this paper are entirely the authors' own work; no AI-generated text, code, or results were incorporated without review and validation by the authors.