BCMIRec: Behavior Co-occurrence Enhanced Multi-Interest Dual-Graph Learning for Multimodal Recommendation

Anonymous Authors

***Abstract*—Multimodal recommendation leverages item contents such as images and texts to alleviate interaction sparsity, yet it still faces noisy implicit feedback, unreliable content similarity, and entangled user intents. To address these issues, we propose BCMIRec, a behavior co-occurrence enhanced multi-interest dual-graph framework for multimodal recommendation. BCMIRec first constructs an item co-occurrence graph from co-selected behaviors and expands each user's neighborhood to provide more behavior-consistent and clustered evidence. It then performs dual-graph propagation on the user–item bipartite graph and the co-occurrence graph, and fuses the resulting representations with an adaptive gate to balance global collaboration and local consistency. On top of the enhanced representations, BCMIRec learns multiple user intent vectors through temperature-controlled soft routing over the expanded neighborhood, and produces ranking scores via a LogSumExp-based mixture prediction with a diversity regularization to reduce interest collapse. Experiments on three Amazon benchmarks demonstrate that BCMIRec consistently outperforms strong multimodal baselines across Recall and NDCG, with clear gains on long-tail items.**

***Index Terms*—Multimodal Recommendation, Graph Neural Networks, Multi-Interest Modeling, Behavior Co-Occurrence, Implicit Feedback, Multimodal Content Alignment**

I. Introduction

Multimodal recommendation leverages auxiliary item contents (e.g., image and text) to alleviate the sparsity of implicit feedback and improve ranking quality. Despite notable progress, three challenges still hinder effective multimodal recommendation in real-world platforms. First, implicit interactions are sparse and noisy: a user may click or purchase items for diverse and transient reasons, making the observed user–item edges an imperfect signal of true preference. Second, multimodal semantics may not align well with collaborative patterns. Visual/textual similarity does not always imply similar preference, and overly relying on content can introduce spurious correlations. Third, user preference is often multi-faceted and entangled: a single user may simultaneously exhibit multiple interests (e.g., sportswear and outdoor gear), and learning a single representation may mix heterogeneous intents and hurt personalization.

A line of graph-based multimodal recommenders models collaborative relations with user–item bipartite graphs or item–item graphs, and propagates embeddings to enhance representations. However, directly propagating on the bipartite graph can be fragile when user histories are short; meanwhile, item–item graphs built from content similarity may amplify modality bias. Moreover, many multi-interest approaches depend on sequence encoders or hard clustering, which may be sensitive to noise and difficult to optimize end-to-end under implicit feedback.

In this paper, we propose BCMIRec, a Behavior Co-occurrence enhanced Multi-Interest dual-graph learning framework for multimodal recommendation. Our key intuition is that behavior co-occurrence provides a robust and complementary signal to expand user neighborhoods and guide interest extraction. Specifically, we construct an item co-occurrence graph from co-selected behaviors, and use it to expand candidate neighborhoods for each user. Then, we perform dual-path propagation on (i) the co-occurrence graph to capture local consistency among co-consumed items, and (ii) the user–item bipartite graph to preserve global collaborative structure. An adaptive gate fuses the two propagated signals to balance local and global evidence. Finally, BCMIRec learns differentiable multi-interest user representations via temperature-controlled soft routing, and predicts with a LogSumExp aggregator to model the mixture of interests while remaining fully trainable.

Our main contributions are:

- We introduce a behavior co-occurrence based neighbor expansion strategy for multimodal recommendation, which provides more consistent preference evidence under sparse implicit feedback.
- We propose a dual-graph propagation architecture with an adaptive gated fusion mechanism, integrating local co-occurrence consistency and global collaboration.
- We design a differentiable multi-interest learning module with soft routing and LogSumExp prediction, enabling robust interest disentanglement and effective ranking.
- Experiments on three Amazon benchmarks demonstrate that BCMIRec consistently outperforms strong multimodal baselines, with clear gains on long-tail items.

II. Related Work

A. Multimodal Recommendation

Multimodal recommendation incorporates item contents (e.g., images and texts) to enrich representations under sparse implicit feedback. Early methods such as VBPR fuse visual features with ID embeddings [1]. More recent work further considers adaptive fusion with gating mechanisms to balance different modality signals, such as MGCN [2]. However, multimodal features can be noisy and may be misaligned with collaborative signals, which can lead to unstable preference estimation.

B. Graph-based Multimodal Recommendation

GNN-based recommenders capture high-order signals on the user–item bipartite graph. MMGCN performs modality-aware propagation and fusion on graphs [3]. To complement bipartite modeling, DualGNN exploits user co-occurrence, while LATTICE, MICRO, and FREEDOM leverage or learn item semantic graphs [4]–[7]; LGMRec further explores higher-order structures [8]. Yet, graph propagation may amplify modality noise, and semantic graphs can be biased when content is misaligned with collaborative signals.

C. Cross-modal Alignment and Fusion

Recent work improves robustness via alignment objectives and adaptive fusion. BM3 aligns ID and modality representations with contrastive learning [9], while SLMRec and MENTOR design self-supervised tasks in multimodal graph learning [10], [11]. Other methods incorporate noise suppression and adaptive fusion modules (e.g., MambaRec) [12]. Nevertheless, many approaches still rely on feature-space alignment or fixed fusion and may lack behavior-consistent structural evidence under sparsity.

D. Positioning of Our Work

BCMIRec tackles sparse/noisy implicit feedback, modality mismatch, and entangled user intents. It builds an item co-occurrence graph from behaviors and expands each user's neighborhood to provide more clustered evidence. It then performs dual-graph propagation on the co-occurrence graph and the user–item graph with an adaptive gated fusion to balance local consistency and global collaboration. Finally, BCMIRec learns multiple user intent vectors via differentiable soft routing over the expanded neighborhood and predicts with a LogSumExp-based mixture, together with regularization to reduce interest collapse, improving ranking robustness.

III. Methodology

A. Problem Definition

Let $\mathcal{U}$ and $\mathcal{I}$ denote the sets of users and items. We observe implicit feedback interactions $\mathcal{E} = \{(u, i)\}$ and define a binary interaction matrix $\mathbf{C} \in \{0,1\}^{|\mathcal{U}|\times|\mathcal{I}|}$, where $C_{ui} = 1$ indicates an observed interaction. We denote the interacted item set of user u as $\mathcal{N}(u) = \{i \mid C_{ui} = 1\}$ and the interacted user set of item i as $\mathcal{N}(i) = \{u \mid C_{ui} = 1\}$. Each item i is associated with multimodal features, including a visual feature vector $\mathbf{v}_i$ and a textual feature vector $\mathbf{t}_i$ extracted by pretrained encoders. Given a user u, our goal is to learn a scoring function $\hat{y}_{ui}$ to rank candidate items and recommend a top-K list.

B. Behavior Co-occurrence Graph and Neighbor Expansion

Co-occurrence graph. We build an item–item co-occurrence graph $G_{II} = (\mathcal{I}, \mathcal{E}_{II})$ from behavior co-selection, where two items are connected if they are frequently consumed by the same users. For an item pair (i, j), we define the co-occurrence strength

$$S_{ij} = |\mathcal{U}(i) \cap \mathcal{U}(j)|, \tag{1}$$

where $\mathcal{U}(i)$ is the set of users who interacted with item i. Unlike modality-similarity graphs, S_{ij} provides a behavior-grounded signal and is less sensitive to modality noise or encoder bias. To keep the graph sparse and scalable, we retain the top-K neighbors for each item according to S_{ij} and symmetrize the adjacency.

User neighborhood expansion. Given a user u, we expand her neighborhood by co-occurrence neighbors:

$$\mathcal{N}^+(u) = \mathcal{N}(u) \ \cup \bigcup_{i\in\mathcal{N}(u)} \mathrm{TopK}_{G_{II}}(i). \tag{2}$$

This expansion yields a more "clustered" candidate set, which is especially beneficial for users with short histories. In our framework, $\mathcal{N}^+(u)$ serves as the evidence pool for interest-aware aggregation, while the original $\mathcal{N}(u)$ is still used for modeling explicit user–item collaboration.

C. Multimodal Initialization

We project multimodal features into the latent space and combine them with ID embeddings. Let $\mathbf{e}_i^{(0)}$ be the learnable ID embedding of item i, and $\mathbf{W}_v, \mathbf{W}_t$ be projection matrices:

$$\tilde{\mathbf{v}}_i = \mathbf{W}_v \mathbf{v}_i, \quad \tilde{\mathbf{t}}_i = \mathbf{W}_t \mathbf{t}_i. \tag{3}$$

Since different modalities may contribute unevenly across items, we adopt a lightweight gated fusion to adaptively weight visual and textual signals:

$$\mathbf{m}_i = \sigma(\mathbf{W}_g[\tilde{\mathbf{v}}_i \| \tilde{\mathbf{t}}_i]) \odot \tilde{\mathbf{v}}_i + (1 - \sigma(\cdot)) \odot \tilde{\mathbf{t}}_i. \tag{4}$$

The resulting multimodal embedding is added to the ID embedding to preserve collaborative identity while injecting content semantics:

$$\mathbf{x}_i^{(0)} = \mathbf{e}_i^{(0)} + \mathbf{m}_i. \tag{5}$$

User embeddings $\mathbf{x}_u^{(0)}$ are learnable parameters.

D. Dual-Graph Propagation with Gated Fusion

We propagate embeddings on two graphs: the user–item bipartite graph G_{UI} and the co-occurrence graph G_{II}. The motivation is that G_{UI} captures global collaborative relations, while G_{II} provides item-side local consistency from co-consumption, and the two signals are complementary.

(1) Bipartite propagation. Following LightGCN-style message passing, at layer ℓ:

$$\mathbf{x}_u^{(\ell+1)} = \sum_{i\in\mathcal{N}(u)} \frac{1}{\sqrt{|\mathcal{N}(u)|\,|\mathcal{N}(i)|}} \mathbf{x}_i^{(\ell)}, \tag{6}$$

$$\mathbf{x}_i^{(\ell+1)} = \sum_{u\in\mathcal{N}(i)} \frac{1}{\sqrt{|\mathcal{N}(u)|\,|\mathcal{N}(i)|}} \mathbf{x}_u^{(\ell)}. \tag{7}$$

This normalization stabilizes training and prevents popular users/items from dominating propagation.

(2) Co-occurrence propagation. On G_{II}, item embeddings are updated by co-occurrence neighbors:

$$\mathbf{z}_i^{(\ell+1)} = \sum_{j\in\mathcal{N}_{II}(i)} \alpha_{ij} \mathbf{z}_j^{(\ell)}, \quad \alpha_{ij} = \frac{S_{ij}}{\sum_{k\in\mathcal{N}_{II}(i)} S_{ik}}. \tag{8}$$

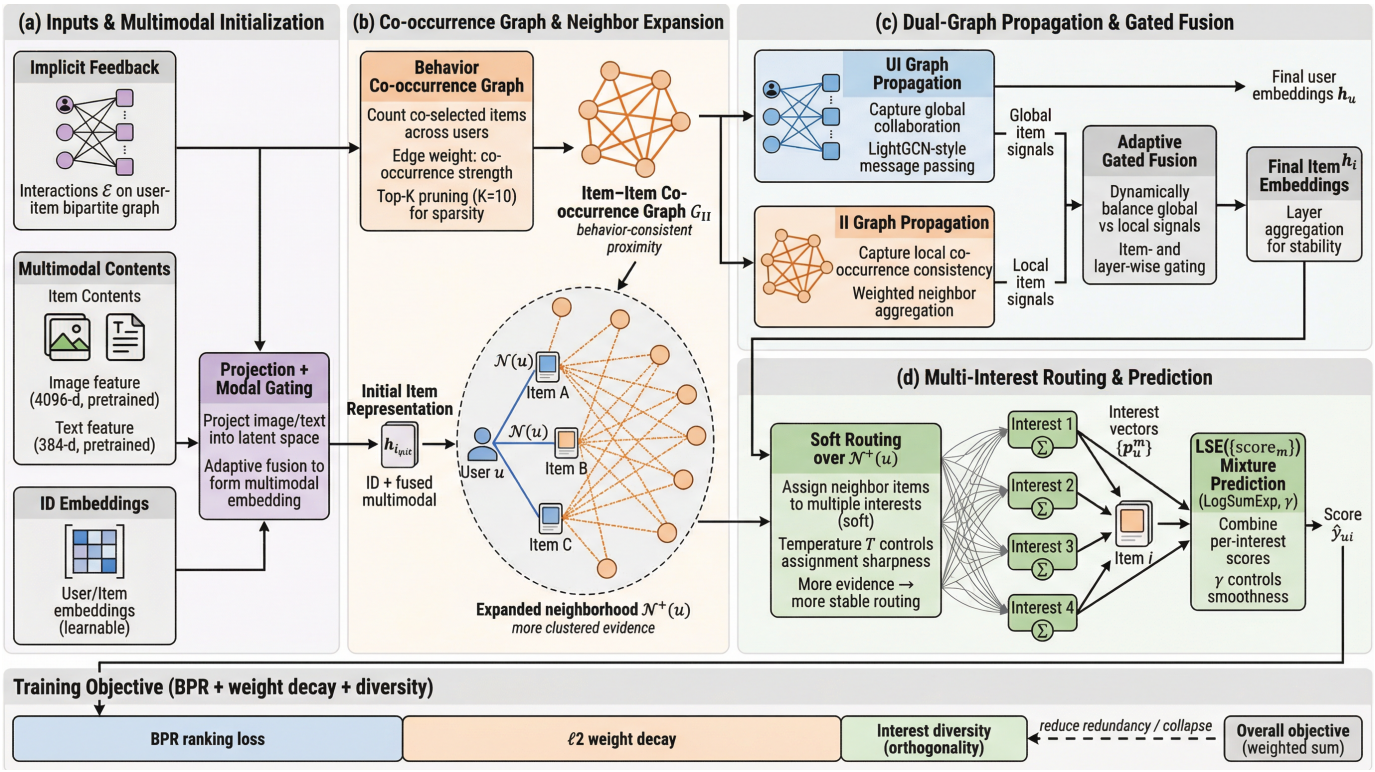


Fig. 1: Overall framework of BCMIRec: (a) multimodal initialization, (b) behavior co-occurrence graph construction and neighborhood expansion, (c) dual-graph propagation with adaptive gated fusion, and (d) differentiable multi-interest routing and mixture prediction.

Here $\mathcal{N}_{II}(i)$ denotes the top-K co-occurrence neighbors of i. The weight α_{ij} emphasizes frequently co-consumed neighbors, encouraging behavior-consistent smoothness.

Gated fusion. The two propagations capture complementary signals: G_{UI} preserves global collaboration while G_{II} provides local co-occurrence consistency. We fuse the two item representations with an adaptive gate:

$$\mathbf{h}_i^{(\ell)} = \beta_i^{(\ell)} \odot \mathbf{x}_i^{(\ell)} + (1-\beta_i^{(\ell)}) \odot \mathbf{z}_i^{(\ell)}, \quad \beta_i^{(\ell)} = \sigma(\mathbf{W}_\beta [\mathbf{x}_i^{(\ell)} \| \mathbf{z}_i^{(\ell)}]). \tag{9}$$

The gate $\beta_i^{(\ell)}$ is item- and layer-specific, allowing the model to rely more on collaborative signals for well-connected items and more on co-occurrence consistency when bipartite evidence is weak. We then use $\mathbf{h}_i^{(\ell)}$ as the item embedding for interest-aware aggregation. For stability, the final representation is the layer-wise sum:

$$\mathbf{h}_i = \sum_{\ell=0}^{L} \mathbf{h}_i^{(\ell)}, \quad \mathbf{h}_u = \sum_{\ell=0}^{L} \mathbf{x}_u^{(\ell)}. \tag{10}$$

E. Multi-Interest Soft Routing and Prediction

User behavior under implicit feedback is often diverse: the same user may interact with items from different categories or aspects, and a single user embedding can easily entangle these signals. We therefore represent each user with multiple interest vectors and learn them from the expanded neighborhood $\mathcal{N}^+(u)$, which provides a richer and more behavior-consistent evidence pool than the original history $\mathcal{N}(u)$.

Interest queries and soft routing. We use $M = 4$ interests by default. For each user u, we introduce M learnable interest queries (prototypes) $\{\mathbf{q}_u^{(m)}\}_{m=1}^M$. Given the fused item representations $\mathbf{h}_i$ (after dual-graph propagation and gated fusion), we compute routing weights that softly assign each neighbor item $i \in \mathcal{N}^+(u)$ to different interests:

$$r_{uim} = \frac{\exp(\langle \mathbf{h}_i, \mathbf{q}_u^{(m)} \rangle / T)}{\sum_{m'=1}^{M} \exp(\langle \mathbf{h}_i, \mathbf{q}_u^{(m')} \rangle / T)}. \tag{11}$$

The routing weight r_{uim} reflects how much item i contributes to the m-th interest of user u. We then obtain each interest representation by aggregating neighbor items with the routing weights:

$$\mathbf{p}_u^{(m)} = \sum_{i \in \mathcal{N}^+(u)} r_{uim} \mathbf{h}_i. \tag{12}$$

In this way, $\{\mathbf{p}_u^{(m)}\}$ are computed on-the-fly from behavior evidence, while $\{\mathbf{q}_u^{(m)}\}$ provide stable prototypes to guide routing. This design allows different interests to attend to different subsets of $\mathcal{N}^+(u)$, which helps separate heterogeneous intents and reduces mutual interference among behaviors.

Mixture prediction. Given a target item i, we compute an interest-specific matching score $s_{uim} = \langle \mathbf{p}_u^{(m)}, \mathbf{h}_i \rangle$. Instead of

selecting a single interest, we combine multiple interests with a smooth mixture:

$$\hat{y}_{ui} = \gamma \log \sum_{m=1}^{M} \exp(s_{uim}/\gamma). \tag{13}$$

This LogSumExp aggregation softly emphasizes the most relevant interest while still allowing other compatible interests to contribute, leading to stable gradients and end-to-end optimization with implicit ranking losses.

F. Objective Function

BPR loss. We optimize BCMIRec with a pairwise Bayesian Personalized Ranking (BPR) objective:

$$\mathcal{L}_{\text{BPR}} = -\sum_{(u,i^+,i^-)} \log \sigma(\hat{y}_{ui^+} - \hat{y}_{ui^-}), \tag{14}$$

where (u, i^+, i^-) is a training triplet, i^+ is an observed item of user u, and i^- is a sampled unobserved item.

Interest diversity regularization. In practice, the learned interest vectors may become highly similar, which makes different interests redundant and weakens the disentanglement effect. To avoid this, we add an auxiliary diversity term that encourages different interests to be less correlated (ideally close to orthogonal). For each user u, we normalize the M interest vectors and stack them as $\mathbf{P}_u = [\bar{\mathbf{p}}_u^{(1)}, \ldots, \bar{\mathbf{p}}_u^{(M)}] \in \mathbb{R}^{d \times M}$, where $\bar{\mathbf{p}}_u^{(m)} = \mathbf{p}_u^{(m)}/(\|\mathbf{p}_u^{(m)}\|_2 + \epsilon)$. We compute the interest correlation matrix $\mathbf{P}_u^\top \mathbf{P}_u$ and penalize its off-diagonal correlations:

$$\mathcal{L}_{\text{div}} = \sum_{u \in \mathcal{U}} \left\| \mathbf{P}_u^\top \mathbf{P}_u - \mathbf{I} \right\|_F^2, \tag{15}$$

where $\mathbf{I}$ is the $M \times M$ identity matrix and $\|\cdot\|_F$ is the Frobenius norm.

Minimizing $\mathcal{L}_{\text{div}}$ reduces overlap among interests, helps the soft routing produce more distinct assignments, and usually stabilizes training since the final prediction is less likely to be dominated by several nearly identical interests.

Overall objective. We also apply ℓ_2 regularization on model parameters Θ. The final loss is:

$$\mathcal{L} = \mathcal{L}_{\text{BPR}} + \lambda \|\Theta\|_2^2 + \mu \mathcal{L}_{\text{div}}, \tag{16}$$

where λ and μ control the strengths of weight decay and interest diversity regularization, respectively.

IV. Experiments

In this section, we conduct extensive experiments to verify the effectiveness of BCMIRec and answer the following research questions (RQs):

- **RQ1:** How does BCMIRec compare with strong multimodal and CF baselines?
- **RQ2:** What is the contribution of each key component in BCMIRec?
- **RQ3:** Can multi-interest modeling alleviate interest entanglement and bring benefits on long-tail preference modeling and recommendation diversity?
- **RQ4:** How robust is BCMIRec to different hyperparameter settings?

TABLE I: Statistics of the experimental datasets.

Dataset	User	Item	Interaction	Density
Baby	19445	7050	160792	0.117%
Sports	35598	18357	2296337	0.045%
Clothing	39387	23033	278677	0.031%

A. Experimental Setup

Datasets. We test BCMIRec on three commonly used subsets of the Amazon review data [13], namely *Baby*, *Sports and Outdoors* (*Sports*), and *Clothing, Shoes and Jewelry* (*Clothing*). Each subset comes with both item images and textual descriptions. We follow the standard preprocessing pipeline used in multimodal recommendation [14]: users and items are filtered with a 5-core rule, and the remaining interactions are treated as implicit positive feedback. For item content, we directly adopt the provided 4096-d visual features and encode the text into 384-d vectors using a pretrained Sentence-Transformer [15]; both modalities are then mapped into the same embedding space for model learning. Detailed statistics of the processed datasets are reported in Table I.

Baselines. We compare BCMIRec with interaction-only recommenders MF-BPR [16] and LightGCN [17]. We further include representative multimodal recommenders, including VBPR [1], MMGCN [3], DualGNN [4], LATTICE [5], SLMRec [10], BM3 [9], MGCN [2], LGMRec [8], FEW [18], HPMRec [19], and JCSMRec [20]. We use official implementations whenever available and tune hyperparameters on the validation set for a fair comparison.

Metrics and evaluation. Following prior work [14], we randomly split each user's interaction history into training/validation/test sets with an $8:1:1$ ratio. The best checkpoint is selected according to the highest Recall@20 on the validation set. We report the average performance over all test users using Recall@10, Recall@20, NDCG@10, and NDCG@20.

Implementation details. We implement BCMIRec and all baselines in MMRec [14] with PyTorch. We set the embedding size to $d = 64$ and use Xavier initialization; all models are optimized with Adam [21]. For each baseline and BCMIRec, we conduct grid search on the validation set. Specifically, weight decay and the diversity weight μ are searched in $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$; the routing temperature T in $\{0.2, 0.3, 0.5, 0.7\}$ and LogSumExp sharpness γ in $\{0.1, 0.2, 0.5, 1.0\}$. We keep top-K co-occurrence neighbors with $K = 10$ by default, and set $L \in \{1, 2, 3, 4\}$ and $L_{II} \in \{1, 2, 3\}$ for propagation on the user–item and item–item graphs, respectively. We select the final model by the best Recall@20 on the validation set, and report Recall@10/20 and NDCG@10/20 on the test set. All experiments are run on a single NVIDIA RTX 4070Ti GPU.

B. Overall Performance (RQ1)

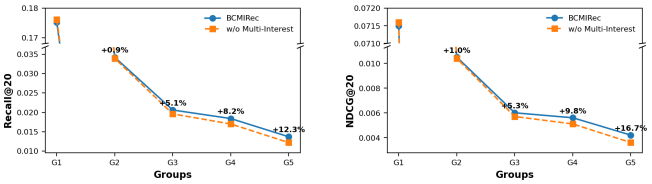
Table II reports the overall results on Baby, Sports, and Clothing. BCMIRec achieves the best performance across all datasets and metrics. In particular, BCMIRec improves over

TABLE II: Performance comparison on three datasets.

Datasets	Baby				Sports				Clothing			
Model	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20
MF-BPR (UAI'09)	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 (SIGIR'20)	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 (AAAI'16)	0.0423	0.0633	0.0233	0.0284	0.0558	0.0856	0.0307	0.0384	0.0281	0.0415	0.0158	0.0192
MMGCN (MM'19)	0.0378	0.0615	0.0200	0.0261	0.0370	0.0605	0.0193	0.0254	0.0218	0.0345	0.0110	0.0142
DualGNN (TMM'21)	0.0448	0.0716	0.0240	0.0309	0.0568	0.0859	0.0310	0.0385	0.0454	0.0683	0.0241	0.0299
LATTICE (MM'21)	0.0547	0.0850	0.0292	0.0370	0.0620	0.0953	0.0335	0.0421	0.0492	0.0733	0.0268	0.0330
SLMRec (TMM'22)	0.0540	0.0810	0.0296	0.0361	0.0676	0.1007	0.0374	0.0462	0.0452	0.0675	0.0247	0.0303
BM3 (WWW'23)	0.0564	0.0883	0.0301	0.0383	0.0656	0.0980	0.0355	0.0438	0.0422	0.0621	0.0231	0.0281
MGCN (MM'23)	0.0620	0.0964	0.0339	0.0427	0.0729	0.1106	0.0379	0.0496	0.0641	0.0945	0.0347	0.0428
FEW (IJCNN'24)	0.0633	0.0993	0.0350	0.0442	0.0674	0.1003	0.0375	0.0460	0.0490	0.0723	0.0272	0.0331
LGMRec (AAAI'24)	0.0644	0.1002	0.0349	0.0440	0.0720	0.1068	0.0390	0.0480	0.0555	0.0828	0.0302	0.0371
HPMRec (CIKM'25)	0.0667	0.1033	0.0357	0.0451	0.0751	0.1129	0.0410	0.0507	0.0658	0.0963	0.0351	0.0429
JCSMRec (IJCNN'25)	0.0648	0.0993	0.0349	0.0433	0.0754	0.1103	0.0387	0.0478	0.0636	0.0958	0.0345	0.0431
BCMIRec	**0.0689**	**0.1068**	**0.0372**	**0.0467**	**0.0780**	**0.1180**	**0.0424**	**0.0524**	**0.0685**	**0.1010**	**0.0370**	**0.0453**
Improve.	3.3%	3.4%	4.2%	3.5%	3.4%	4.5%	3.4%	3.6%	4.1%	4.9%	5.4%	5.1%

TABLE III: Ablation results of BCMIRec.

Variant	Baby		Sports		Clothing	
	R@20	N@20	R@20	N@20	R@20	N@20
BCMIRec	0.1068	0.0467	0.1180	0.0524	0.1010	0.0453
w/o Co-Graph	0.1036	0.0452	0.1143	0.0505	0.0981	0.0441
w/o Expansion	0.1029	0.0447	0.1135	0.0502	0.0975	0.0437
w/o Gate	0.1052	0.0456	0.1154	0.0512	0.0989	0.0444
w/o Multi-Interest	0.1009	0.0437	0.1108	0.0492	0.0943	0.0417



(a) Recall@20 (b) NDCG@20

Fig. 2: Long-tail performance across item popularity groups on Sports.

the strongest baseline by 3.4% / 4.5% / 4.9% in Recall@20 on Baby / Sports / Clothing, respectively, showing consistent gains under implicit feedback. The improvements indicate that behavior co-occurrence expansion and dual-graph fusion provide more reliable preference evidence than using the bipartite graph or content similarity alone. Moreover, the multi-interest module further improves personalization by disentangling heterogeneous intents from sparse implicit feedback.

C. Ablation Study (RQ2)

To quantify the contribution of each component, we construct the following ablated variants and report results in Table III: (1) **w/o Co-Graph**: remove the item–item co-occurrence graph (local path) and propagate only on the user–item graph; (2) **w/o Expansion**: disable neighbor expansion and perform routing/aggregation only on the first-order interaction set $\mathcal{N}(u)$; (3) **w/o Gate**: replace gated fusion with a fixed fusion (e.g., simple average) for the local/global item representations; (4) **w/o Multi-Interest**: remove multi-interest modeling and degenerate to a single user representation. Each component leads to a consistent performance drop across datasets, and the full model performs best.

D. Multi-Interest Modeling on Long-tail Items (RQ3)

We study whether multi-interest modeling brings larger gains on long-tail items, where feedback is sparse and a single user embedding may entangle heterogeneous intents. On the *Sports* dataset, we rank items by training-set popularity and split them into five equal-sized groups (G1: head, G5: tail), then report group-wise Recall@20/NDCG@20.

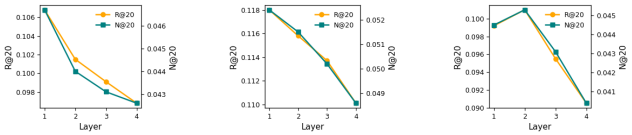
We compare the full model ($M_{\text{int}} = 4$) with a single-interest variant ($M_{\text{int}} = 1$) while keeping other components unchanged. As shown in Fig. 2, multi-interest yields larger improvements on low-popularity groups (especially G4–G5). This is because tail items often reflect minority or occasional preferences with weak, fragmented evidence; multi-interest routing can allocate limited signals into different intent components and reduce interference from dominant head behaviors. In contrast, a single representation is more likely to be dominated by frequent interactions and under-represent tail intents.

Moreover, behavior co-occurrence expansion provides a more clustered neighborhood for routing, which stabilizes interest assignment under sparsity and leads to more consistent tail gains.

E. Hyperparameter Sensitivity (RQ4)

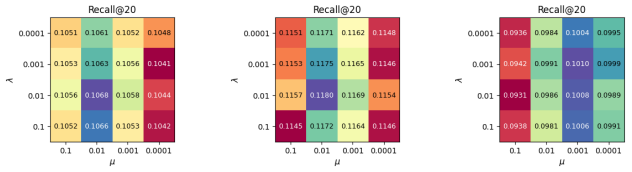
Finally, we study the sensitivity of BCMIRec to key hyperparameters, while keeping other settings fixed to the best configuration on the validation set. We report Recall@20 on all three datasets (the trend is consistent on NDCG@20).

Propagation depth. We vary the propagation depth $L \in \{1, 2, 3, 4\}$ on the user–item interaction bipartite graph. In general, a small depth (1–2 layers) works best: deeper propagation brings limited gains and may introduce over-smoothing, which makes user/item representations less discriminative.



(a) Baby (b) Sports (c) Clothing

Fig. 3: Sensitivity to Propagation Depth L.



(a) Baby (b) Sports (c) Clothing

Fig. 4: Performance Landscape over (λ, μ).

Joint effect of λ and μ. We investigate the combined impact of the weight decay coefficient λ and the diversity loss weight μ, where λ controls overall model complexity and μ balances the strength of interest diversification. When λ is too small, the model may overfit noisy implicit interactions; when it is too large, the embeddings are over-shrunk and the ranking accuracy degrades. Similarly, a moderate μ helps prevent interest collapse and encourages complementary interests, while an overly large μ can dominate optimization and over-constrain interest vectors, hurting the main ranking objective.

V. Conclusion

We proposed BCMIRec, a behavior co-occurrence enhanced multi-interest dual-graph framework for multimodal recommendation. By constructing a co-occurrence graph for neighborhood expansion, performing dual-graph propagation with adaptive gated fusion, and learning differentiable multi-interest representations with soft routing and LogSumExp prediction, BCMIRec achieves consistent improvements on three Amazon benchmarks. Future work includes incorporating richer user behavior types and exploring more adaptive graph construction for dynamic multimodal scenarios.

References

[1] R. He and J. McAuley, "Vbpr: visual bayesian personalized ranking from implicit feedback," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.

[2] P. Yu, Z. Tan, G. Lu, and B.-K. Bao, "Multi-view graph convolutional network for multimedia recommendation," in *Proceedings of the 31st ACM international conference on multimedia*, 2023, pp. 6576–6585.

[3] Y. Wei, X. Wang, L. Nie, X. He, R. Hong, and T.-S. Chua, "Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 1437–1445.

[4] Q. Wang, Y. Wei, J. Yin, J. Wu, X. Song, and L. Nie, "Dualgnn: Dual graph neural network for multimedia recommendation," *IEEE Transactions on Multimedia*, vol. 25, pp. 1074–1084, 2021.

[5] J. Zhang, Y. Zhu, Q. Liu, S. Wu, S. Wang, and L. Wang, "Mining latent structures for multimedia recommendation," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 3872–3880.

[6] J. Zhang, Y. Zhu, Q. Liu, M. Zhang, S. Wu, and L. Wang, "Latent structure mining with contrastive modality fusion for multimedia recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 9, pp. 9154–9167, 2022.

[7] X. Zhou and Z. Shen, "A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation," in *Proceedings of the 31st ACM international conference on multimedia*, 2023, pp. 935–943.

[8] Z. Guo, J. Li, G. Li, C. Wang, S. Shi, and B. Ruan, "Lgmrec: Local and global graph learning for multimodal recommendation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 8, 2024, pp. 8454–8462.

[9] 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 *Proceedings of the ACM web conference 2023*, 2023, pp. 845–854.

[10] Z. Tao, X. Liu, Y. Xia, X. Wang, L. Yang, X. Huang, and T.-S. Chua, "Self-supervised learning for multimedia recommendation," *IEEE Transactions on Multimedia*, vol. 25, pp. 5107–5116, 2022.

[11] J. Xu, Z. Chen, S. Yang, J. Li, H. Wang, and E. C. Ngai, "Mentor: multi-level self-supervised learning for multimodal recommendation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 12, 2025, pp. 12 908–12 917.

[12] K. Ren, C.-Y. Ju, and D.-H. Lee, "Modality alignment with multi-scale bilateral attention for multimodal recommendation," in *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, 2025, pp. 2493–2502.

[13] J. McAuley, C. Targett, Q. Shi, and A. Van Den Hengel, "Image-based recommendations on styles and substitutes," in *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, 2015, pp. 43–52.

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

[15] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[16] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "Bpr: Bayesian personalized ranking from implicit feedback," *arXiv preprint arXiv:1205.2618*, 2012.

[17] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, "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*, 2020, pp. 639–648.

[18] Q. Ye, L. Qiao, Z. Ou, K. Yang, and F. Yang, "Few: Multi-modal recommendation for cold-start," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–9.

[19] Z. Chen, J. Xu, H. Wang, S. Yang, Z. Wan, and H. Hu, "Hypercomplex prompt-aware multimodal recommendation," in *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, 2025, pp. 403–414.

[20] X. Zhang, Y. He, J. Li, J. Chang, K. Zhu, and G. Li, "Joint content semantic relation learning with mamba for multimodal recommendation," in *2025 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2025, pp. 1–8.

[21] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.