
BR-MRS: Synergy-Aware Multimodal Recommendation with Cross-Modal Hard Negatives

Anonymous Authors¹

Abstract

Multimodal recommendation systems integrate visual and textual features to enhance personalized ranking. However, existing methods that directly transfer components from general multimodal learning—such as InfoNCE-style alignment and orthogonality-based decorrelation—fail to explicitly capture *modality-unique* and *synergistic* information under the user-conditioned ranking objective. Through systematic empirical analysis, we reveal that stronger orthogonality regularization does not yield richer unique information but instead shifts learning toward redundant components, while contrastive alignment provides little incentive for synergistic signals that emerge only through fusion. To address these limitations, we propose **BR-MRS**, a multimodal recommendation framework with two core designs: (i) *Cross-modal Hard Negative Sampling* (CHNS), which assigns each unimodal branch the task of resolving confusable cases identified by the other modality, thereby explicitly activating modality-specific evidence; and (ii) a *Synergy-aware BPR Loss* that enforces a larger preference margin for the fused representation than any single-modality branch, explicitly encouraging synergistic learning. Extensive experiments on three benchmark datasets demonstrate that BR-MRS significantly outperforms state-of-the-art methods, achieving up to 23.1% improvement in NDCG@10.

1. Introduction

In recent years, multimodal recommendation systems (MMRS) have achieved remarkable progress in personalized ranking tasks by integrating heterogeneous modality information such as visual and textual content (Liu et al., 2024a).

¹Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

However, since different modalities often describe an item from distinct semantic dimensions, discrepancies across modality-specific representations are inevitable, resulting in *modality inconsistency* (Liu et al., 2024b). Notably, this inconsistency exhibits a pronounced dual nature: when it reflects complementary discriminative information across modalities, it can provide additional evidence for ranking decisions, termed **informative inconsistency** (Williams & Beer, 2010; Yang et al., 2025); conversely, when it stems from modality noise or unreliable modality cues, it may introduce misleading signals, termed **noisy inconsistency** (Yu et al., 2023; Xu et al., 2025b). This observation gives rise to a central challenge in multimodal recommendation:

How can we effectively leverage informative modality inconsistency while mitigating the negative impact of noisy inconsistency in multimodal recommendation?

To address this challenge, some studies adopt modality-independent modeling, where each modality is encoded by an independent encoder and the predictions are combined via late fusion (Xu et al., 2025a). While this design can partially reduce the propagation of modality-specific noise into collaborative signals, it also limits cross-modal interactions, making it harder to explicitly resolve noisy inconsistency and to fully exploit informative inconsistency across modalities (Yu et al., 2023; Xu et al., 2025a). More recently, several approaches introduce self-supervised objectives such as contrastive learning and diffusion models to explicitly enforce cross-modal consistency (Zhou et al., 2023; Wei et al., 2023; Jiang et al., 2024). Although these methods effectively suppress the negative impact of misleading signals on recommendations, they often treat all modality inconsistency as noise, sacrificing modality-differentiated characteristics. Consequently, when a single modality is insufficient to fully capture user preferences, such strategies struggle to effectively exploit informative cross-modal inconsistency, thereby limiting further improvements in recommendation performance (Liu et al., 2024b; Xu et al., 2025b).

Through systematic empirical analysis of representative multimodal recommenders (Guo et al., 2024; Jiang et al., 2024; Zhou et al., 2023; Xu et al., 2025b), we identify two noteworthy phenomena. (1) **Under-exploited informative in-**

055
056
057
058
059
060
061
062
063
064
065
066
067
068
069
070
071
072
073
074
075
076
077
078
079
080
081
082
083
084
085
086
087
088
089
090
091
092
093
094
095
096
097
098
099
100
101
102
103
104
105
106
107
108
109

consistency. In many mis-ranking cases, negative samples are highly similar to the target item in one modality while differ significantly in another modality. This indicates that existing MMRS fail to fully leverage informative modality inconsistency. **(2) Noisy inconsistency undermines fusion.** We also observe that unimodal representations can successfully retrieve the target item, whereas multimodal fused representations fail. This suggests that noisy modality inconsistency may diminish the advantages of multimodal fusion.

To address these challenges, we propose **BR-MRS**, a multimodal recommendation framework that reformulates the classical Bayesian Personalized Ranking objective (Rendle et al., 2009). Specifically, we design a *Cross-modal Hard Negative Sampling* (CHNS) strategy, which constructs discriminatively challenging negatives across modalities to explicitly guide the model toward attending to cross-modal differential features that are valuable for user preference ranking, thereby more effectively exploiting informative inconsistency across modalities. Furthermore, to mitigate fusion degradation, we propose a *Synergy-aware Bayesian Personalized Ranking Loss*, which enforces the fused representation to significantly outperform each unimodal representation in distinguishing positives from negative samples, thereby suppressing the adverse effects introduced by noisy inconsistency and enhancing the recommendation performance of multimodal fusion.

2. The Proposed Method

This section presents **BR-MRS**, a multimodal recommendation framework that explicitly models modality inconsistency by reformulating the classical BPR loss (Rendle et al., 2009). As illustrated in Fig. 1, BR-MRS follows the mainstream multimodal recommendation paradigm, employing graph neural networks to learn user and item representations (Guo et al., 2024). Building upon this backbone, BR-MRS introduces two core components: (i) *Cross-modal Hard Negative Sampling* (CHNS), which mines confusable negatives from one modality to elicit discriminative evidence from another, thereby exploiting informative inconsistency; and (ii) *Synergy-aware BPR Loss*, which enforces the preference margin of the fused representation to significantly exceed that of any unimodal branch, thereby suppressing fusion degeneration caused by noisy inconsistency. Algorithm 1 summarizes the overall training procedure.

2.1. Graph-based Representation Learning

Let \mathcal{U} and \mathcal{I} denote the user and item sets, respectively, with observed interactions $\mathcal{O} \subseteq \mathcal{U} \times \mathcal{I}$. For each item $i \in \mathcal{I}$, the modality feature is denoted as $\mathbf{x}_i^{(m)}$, where $m \in \{t, v\}$.

Algorithm 1 BR-MRS training procedure

Require: Interactions \mathcal{O} , item features $\{\mathbf{x}_i^{(t)}, \mathbf{x}_i^{(v)}\}$, hyper-parameters $\lambda_h, \lambda_s, \theta, \lambda$
Ensure: Trained parameters Θ

- 1: Initialize model parameters Θ
- 2: **for** each epoch **do**
- 3: **for** each $(u, i^+) \in \mathcal{O}$ **do**
- 4: Sample candidate negatives $\mathcal{N}(u) \subseteq \mathcal{I} \setminus \mathcal{O}_u$
- 5: Compute unimodal and fused scores $s_t(u, \cdot)$, $s_v(u, \cdot)$, $s_f(u, \cdot)$
- 6: $i_v^- \leftarrow \arg \max_{j \in \mathcal{N}(u)} s_v(u, j)$
- 7: $i_t^- \leftarrow \arg \max_{j \in \mathcal{N}(u)} s_t(u, j)$
- 8: Sample a negative $i^- \in \mathcal{N}(u)$ for \mathcal{L}_{syn}
- 9: Compute $\mathcal{L}_{\text{chns}}$ and \mathcal{L}_{syn}
- 10: Update Θ by minimizing $\mathcal{L} = \lambda_h \mathcal{L}_{\text{chns}} + \lambda_s \mathcal{L}_{\text{syn}} + \lambda \|\Theta\|_2^2$
- 11: **end for**
- 12: **end for**

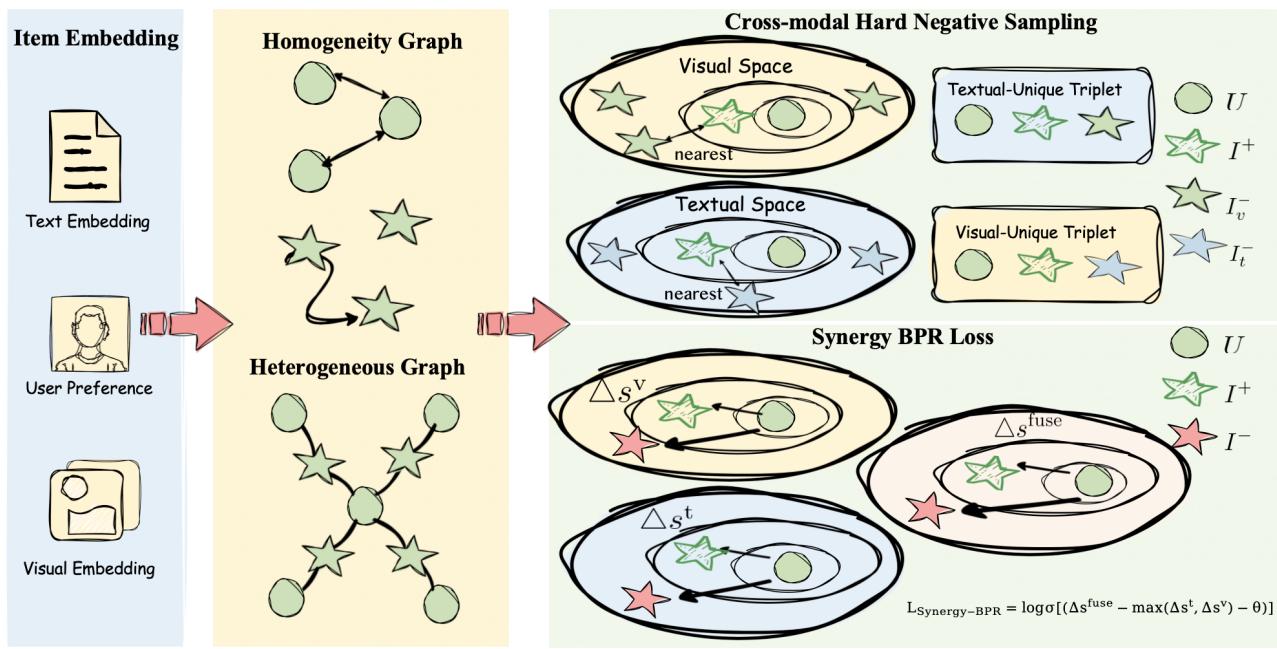
Homogeneous graph construction and propagation. To capture latent associations among entities of the same type, we construct the item homogeneous graph \mathcal{G}_{ii} , whose edge weights integrate both interaction co-occurrence and modality semantics. Specifically, the edge weight between item pair (i, j) is defined as

$$e_{ij} = \alpha \cdot \text{overlap}(\mathcal{N}_i^u, \mathcal{N}_j^u) + (1 - \alpha) \sum_m \beta_m \cos(\mathbf{x}_i^{(m)}, \mathbf{x}_j^{(m)}), \quad (1)$$

where \mathcal{N}_i^u denotes the set of users who have interacted with item i , and α, β_m are balancing coefficients. Top- k sparsification is applied to retain salient connections. The user homogeneous graph \mathcal{G}_{uu} is constructed symmetrically. After graph convolution propagation, node representations encode behavioral patterns and semantic affinities among entities of the same type.

Heterogeneous graph propagation. On the user-item bipartite graph $\mathcal{G}_{ui} = (\mathcal{U} \cup \mathcal{I}, \mathcal{O})$, we employ LightGCN (He et al., 2020) to perform L layers of neighborhood aggregation. Each modality feature propagates through independent channels, and the final representations are obtained via mean pooling across layers. Specifically, for item i , we obtain unimodal representations $\mathbf{q}_i^{(t)}, \mathbf{q}_i^{(v)} \in \mathbb{R}^d$ and fused representation $\mathbf{q}_i^{(f)} = \phi(\mathbf{q}_i^{(t)}, \mathbf{q}_i^{(v)})$; for user u , we symmetrically obtain unimodal preference representations $\mathbf{p}_u^{(t)}, \mathbf{p}_u^{(v)} \in \mathbb{R}^d$ and fused representation $\mathbf{p}_u^{(f)} = \phi(\mathbf{p}_u^{(t)}, \mathbf{p}_u^{(v)})$. The user-item preference score is defined as $s_m(u, i) = \langle \mathbf{p}_u^{(m)}, \mathbf{q}_i^{(m)} \rangle$, where $m \in \{t, v, f\}$.

110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130



131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164

Figure 1. Overview of BR-MRS.

2.2. Cross-modal Hard Negative Sampling

Having obtained unimodal representations and preference scores, we next consider how to explicitly exploit informative inconsistency across modalities. Empirically, we find that it often appears as *cross-modal confusion*: for a positive pair (u, i^+) , some negatives are highly confusable with i^+ under one modality (high scores) but are clearly separable under the other (low scores). This indicates that the other modality can offer a larger preference margin when one modality fails to separate positive and negative items.

Motivated by the negative sampling procedure in classical BPR training (Rendle et al., 2009) and the general benefit of hard negatives in contrastive learning (Kalantidis et al., 2020; Robinson et al., 2021), we propose *Cross-modal Hard Negative Sampling* (CHNS). CHNS leverages confusable negatives selected by one modality to supervise the other modality, encouraging it to provide explicit discriminative evidence. Concretely, for each observed positive pair (u, i^+) , we first sample a candidate negative pool $\mathcal{N}(u) \subseteq \mathcal{I} \setminus \mathcal{O}_u$, and then select the highest-scoring negative under each modality: $i_v^- = \arg \max_{j \in \mathcal{N}(u)} s_v(u, j)$ and $i_t^- = \arg \max_{j \in \mathcal{N}(u)} s_t(u, j)$. These modality-specific negatives are then *cross-assigned* to train the opposite modality branch, yielding the cross-modal BPR loss:

$$\begin{aligned}\mathcal{L}_{\text{chns}}^v &= - \sum_{(u, i^+) \in \mathcal{O}} \log \sigma(s_v(u, i^+) - s_v(u, i_t^-)), \\ \mathcal{L}_{\text{chns}}^t &= - \sum_{(u, i^+) \in \mathcal{O}} \log \sigma(s_t(u, i^+) - s_t(u, i_v^-)), \\ \mathcal{L}_{\text{chns}} &= \mathcal{L}_{\text{chns}}^v + \mathcal{L}_{\text{chns}}^t.\end{aligned}\quad (2)$$

Through this cross-modal supervision, CHNS activates the discriminative strengths of each modality and turns informative inconsistency into effective training signals, exploiting more precise modality-complementary information.

2.3. Synergy-aware BPR Loss

Although CHNS helps exploit informative inconsistency across modalities, noisy inconsistency may still cause *fusion degradation*, where the fused representation becomes less discriminative than a unimodal branch (Yu et al., 2023; Ong & Khong, 2025; Dong et al., 2025). To address this issue, we propose a *Synergy-aware BPR Loss* built upon the BPR framework (Rendle et al., 2009), which explicitly encourages the preference margin of the fused representation to exceed that of any unimodal branch, thereby improving the robustness of multimodal fusion.

Concretely, for a training triple (u, i^+, i^-) where i^- is uniformly sampled from non-interacted items, we define the

165 preference margins for the fused and unimodal branches as:

$$\begin{aligned}\Delta_f &= s_f(u, i^+) - s_f(u, i^-), \\ \Delta_t &= s_t(u, i^+) - s_t(u, i^-), \\ \Delta_v &= s_v(u, i^+) - s_v(u, i^-).\end{aligned}\quad (3)$$

In pairwise learning, the margin Δ is commonly used as a proxy for ranking confidence: a larger margin indicates stronger separation between positive and negative items under the corresponding scoring function (Rendle et al., 2009; Dong et al., 2025). Accordingly, we take the best-performing unimodal branch as an adaptive reference and introduce a strict positive margin constraint $\theta > 0$, yielding the synergy-aware BPR loss:

$$\mathcal{L}_{\text{syn}} = - \sum_{(u, i^+, i^-)} \log \sigma(\Delta_f - \max(\Delta_t, \Delta_v) - \theta). \quad (4)$$

By enforcing $\Delta_f > \max(\Delta_t, \Delta_v) + \theta$, the synergy-aware loss mitigates the adverse effect of noisy modality information during fusion, ensuring that the fused representation consistently maintains an advantage over any unimodal branch, and thereby enhancing the reliability and effectiveness of multimodal fusion.

2.4. Overall Objective

We integrate the cross-modal hard negative sampling loss and the synergy-aware loss into a unified training objective for BR-MRS:

$$\mathcal{L} = \lambda_h \mathcal{L}_{\text{chns}} + \lambda_s \mathcal{L}_{\text{syn}} + \lambda \|\Theta\|_2^2, \quad (5)$$

where λ_h and λ_s control the contributions of CHNS and the synergy-aware loss respectively, λ is the regularization coefficient, and Θ denotes all trainable parameters.

3. Experiment

3.1. Experimental Setup

We evaluate BR-MRS on three public multimodal recommendation benchmarks, namely Baby, Sports, and Clothing, where each item is associated with both visual and textual content features. We follow standard preprocessing and splitting protocols used in prior multimodal recommendation work to ensure fair comparison. We adopt leave-one-out evaluation and report Recall@K and NDCG@K with $K \in \{10, 20\}$. Baselines cover classical CF models (e.g., BPR, LightGCN, ApeGNN, MGDN) and a broad range of multimodal recommenders (e.g., VBPR, MMGCN, Dual-GNN, GRCN, LATTICE, BM3, SLMRec, MICRO, MGCGN, FREEDOM, LGMRec, DRAGON, MIG-GT, REARM). For all methods, hyperparameters are tuned on validation sets, and we use the same multimodal features and evaluation pipeline for a fair comparison.

3.2. Overall Performance

Table 1 summarizes the overall performance. BR-MRS consistently outperforms strong baselines across different model families, achieving state-of-the-art results on the reported metrics. On the Baby dataset, BR-MRS yields substantial improvements over the strongest baseline, with gains up to 23.1% in NDCG@10. These results validate that explicitly modeling modality-unique evidence and cross-modal synergy is more effective than applying generic alignment or fusion-only objectives.

3.3. Plug-and-Play Improvements

To further demonstrate the generalizability of our proposed components, we apply BR-MRS as a plug-and-play module to various existing multimodal recommendation methods. Specifically, we integrate the CHNS and Synergy-aware BPR loss into the training objectives of representative baselines without modifying their original architectures.

As shown in Table 2, incorporating BR-MRS components yields consistent and significant improvements across all baseline methods on all three datasets. Notably, weaker baselines (e.g., MMGCN) benefit more substantially, with up to 21.1% relative gain in NDCG@10, while stronger baselines (e.g., DRAGON, MIG-GT) still achieve 6–9% improvements. On average, integrating BR-MRS improves Recall@10 by approximately 12% and NDCG@10 by approximately 13% across all datasets. These results demonstrate that our proposed cross-modal hard negative sampling and synergy-aware loss are model-agnostic and can effectively enhance existing multimodal recommenders as plug-and-play modules.

3.4. Ablation Study

To validate the effectiveness of each proposed component, we conduct ablation studies on two benchmark datasets. We study two variants of BR-MRS: merely providing Cross-modal Hard Negative Sampling (CHNS) or Synergy-aware BPR loss (Syn). The checkmark \checkmark indicates the component is enabled, while \circ indicates it is disabled.

As shown in Table 3, disabling either component leads to performance degradation. When only CHNS is enabled, the model can mine modality-specific discriminative evidence but lacks explicit synergy constraints. When only Syn is enabled, the model enforces fusion superiority but misses the cross-modal hard negative mining. The full model with both components achieves the best performance, demonstrating their complementary contributions.

Method	Baby				Sports				Clothing			
	Metric	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20	R@10	R@20	N@10
BPR	0.0363	0.0582	0.0196	0.0254	0.0439	0.0661	0.0246	0.0304	0.0212	0.0310	0.0118	0.0142
LightGCN	0.0486	0.0763	0.0263	0.0334	0.0576	0.0873	0.0318	0.0394	0.0368	0.0552	0.0203	0.0249
ApeGNN	0.0501	0.0775	0.0267	0.0338	0.0608	0.0892	0.0333	0.0407	0.0378	0.0538	0.0204	0.0244
MGDN	0.0495	0.0783	0.0272	0.0346	0.0614	0.0932	0.0340	0.0422	0.0362	0.0551	0.0199	0.0247
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	0.0421	0.0660	0.0220	0.0282	0.0401	0.0636	0.0209	0.0270	0.0227	0.0361	0.0154	0.0154
DualGNN	0.0513	0.0803	0.0278	0.0352	0.0588	0.0899	0.0324	0.0404	0.0452	0.0675	0.0242	0.0298
GRCN	0.0532	0.0824	0.0282	0.0358	0.0599	0.0919	0.0330	0.0413	0.0421	0.0657	0.0224	0.0284
LATTICE	0.0555	0.0861	0.0299	0.0378	0.0628	0.0965	0.0343	0.0427	0.0501	0.0744	0.0275	0.0338
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
SLMRec	0.0521	0.0772	0.0289	0.0354	0.0663	0.0990	0.0365	0.0450	0.0442	0.0659	0.0241	0.0296
MICRO	0.0584	0.0929	0.0318	0.0407	0.0679	0.1050	0.0367	0.0463	0.0521	0.0772	0.0283	0.0347
MGCN	0.0628	0.0975	0.0346	0.0435	0.0737	0.1118	0.0405	0.0504	0.0650	0.0956	0.0355	0.0436
FREEDOM	0.0627	0.0992	0.0330	0.0424	0.0717	0.1089	0.0385	0.0481	0.0629	0.0941	0.0341	0.0420
LGMRec	0.0644	0.1002	0.0349	0.0440	0.0720	0.1068	0.0390	0.0480	0.0555	0.0828	0.0302	0.0371
DRAGON	0.0670	0.1032	0.0352	0.0443	0.0761	0.1150	0.0421	0.0520	0.0680	0.0990	0.0372	0.0451
MIG-GT	0.0673	0.1033	0.0368	0.0460	0.0762	0.1142	0.0422	0.0519	0.0645	0.0945	0.0354	0.0430
REARM	0.0733	0.1141	0.0375	0.0500	0.0820	0.1199	0.0446	0.0544	0.0693	0.0994	0.0361	0.0437
BR-MRS	0.0819	0.1215	0.0452	0.0554	0.0867	0.1247	0.0488	0.0587	0.0734	0.1074	0.0398	0.0484
Improve	$\uparrow 11.7\%$ $\uparrow 6.5\%$ $\uparrow 20.5\%$ $\uparrow 10.8\%$ $\uparrow 5.7\%$ $\uparrow 4.0\%$ $\uparrow 9.4\%$ $\uparrow 7.9\%$ $\uparrow 5.9\%$ $\uparrow 8.0\%$ $\uparrow 10.2\%$ $\uparrow 10.8\%$											

3.5. Hyper-parameter and Robustness Analysis

We analyze the sensitivity of BR-MRS to key hyperparameters, including λ_h (weight of CHNS), λ_s (weight of synergy-aware loss), and θ (synergy margin). Performance remains stable across a broad range of values, with moderate θ yielding the best trade-off between unimodal stability and fusion gains. We also observe that BR-MRS maintains consistent improvements under different evaluation cutoffs, indicating robustness to the choice of ranking metric. Detailed curves and additional robustness results are deferred to the Appendix.

3.6. Effectiveness of CHNS

We further compare CHNS with alternative negative sampling strategies, including uniform sampling and hard negatives mined within a single (fused or unimodal) representation space. CHNS consistently yields stronger gains, as it deliberately selects negatives that are confusable in one modality but separable in the other. This cross-modal contrast forces each unimodal branch to contribute discriminative cues that would otherwise be ignored, leading to larger unique subsets (\mathcal{U}_t and \mathcal{U}_v) and improved overall ranking performance.

3.7. Effectiveness of Synergy-aware Loss

To evaluate the synergy-aware loss, we analyze how fusion quality changes compared to unimodal branches. The synergy constraint reduces fusion degradation by shrinking the

degradation subset \mathcal{U}_r and expanding the synergy subset \mathcal{U}_{tv} , indicating that fused representations more frequently achieve correct ranking than either unimodal branch. In practice, this translates into more reliable fused scores and fewer cases where multimodal fusion hurts performance.

4. Conclusion

In this paper, we investigated the limitations of directly transferring general multimodal learning components—specifically InfoNCE-style alignment and orthogonality-based decorrelation—to multimodal recommendation systems. Through systematic empirical analysis, we revealed that stronger orthogonality regularization fails to enhance modality-unique information and instead enlarges the degradation regime, while contrastive alignment provides little incentive for synergistic signals.

To address these limitations, we proposed **BR-MRS**, a synergy-aware multimodal recommendation framework with two key innovations. First, Cross-modal Hard Negative Sampling (CHNS) explicitly activates modality-specific evidence by assigning each unimodal branch to resolve confusable cases identified by the other modality. Second, the Synergy-aware BPR Loss enforces that the fused representation achieves a larger preference margin than any single-modality branch, explicitly inducing synergistic learning.

Extensive experiments on three benchmark datasets demonstrate that BR-MRS significantly outperforms state-of-the-art methods, achieving up to 23.1% improvement in

275 *Table 2.* Plug-and-play improvements when applying BR-MRS components to existing methods. We report Recall@10 and NDCG@10
 276 on three benchmark datasets. Δ denotes relative improvement.

Method	Baby				Sports				Clothing			
	Recall@10	Δ	NDCG@10	Δ	Recall@10	Δ	NDCG@10	Δ	Recall@10	Δ	NDCG@10	Δ
MMGCN	0.0421	–	0.0220	–	0.0401	–	0.0209	–	0.0227	–	0.0154	–
+BR-MRS	0.0505	+20.0%	0.0265	+20.5%	0.0481	+19.9%	0.0253	+21.1%	0.0273	+20.3%	0.0186	+20.8%
DualGNN	0.0513	–	0.0278	–	0.0588	–	0.0324	–	0.0452	–	0.0242	–
+BR-MRS	0.0587	+14.4%	0.0320	+15.1%	0.0671	+14.1%	0.0372	+14.8%	0.0518	+14.6%	0.0279	+15.3%
GRCN	0.0532	–	0.0282	–	0.0599	–	0.0330	–	0.0421	–	0.0224	–
+BR-MRS	0.0610	+14.7%	0.0326	+15.6%	0.0686	+14.5%	0.0381	+15.5%	0.0483	+14.7%	0.0259	+15.6%
LATTICE	0.0555	–	0.0299	–	0.0628	–	0.0343	–	0.0501	–	0.0275	–
+BR-MRS	0.0622	+12.1%	0.0338	+13.0%	0.0703	+11.9%	0.0390	+13.7%	0.0561	+12.0%	0.0311	+13.1%
MGCN	0.0628	–	0.0346	–	0.0737	–	0.0405	–	0.0650	–	0.0355	–
+BR-MRS	0.0689	+9.7%	0.0382	+10.4%	0.0805	+9.2%	0.0447	+10.4%	0.0712	+9.5%	0.0391	+10.1%
DRAGON	0.0670	–	0.0352	–	0.0761	–	0.0421	–	0.0680	–	0.0372	–
+BR-MRS	0.0720	+7.5%	0.0381	+8.2%	0.0817	+7.4%	0.0458	+8.8%	0.0728	+7.1%	0.0397	+6.7%
MIG-GT	0.0673	–	0.0368	–	0.0762	–	0.0422	–	0.0645	–	0.0354	–
+BR-MRS	0.0718	+6.7%	0.0395	+7.3%	0.0814	+6.8%	0.0456	+8.1%	0.0688	+6.7%	0.0380	+7.3%
Avg. Improv.	+12.2%		+12.9%		+11.9%		+12.8%		+12.1%		+12.7%	

295 *Table 3.* Ablation study on two benchmark datasets. We report
 296 Recall@10 (R@10) and NDCG@10 (N@10).

CHNS	Syn-BPR	Baby		Sports	
		R@10	N@10	R@10	N@10
✓	○	0.0803	0.0446	0.0845	0.0472
○	✓	0.0767	0.0411	0.0812	0.0451
✓	✓	0.0819	0.0452	0.0867	0.0488

304 NDCG@10. Ablation studies confirm the complementary
 305 contributions of both proposed components. Our work pro-
 306 vides new insights into how multimodal information should
 307 be leveraged for personalized ranking and offers a principled
 308 approach for future multimodal recommendation research.
 309

311 Impact Statement

313 This paper presents work whose goal is to advance the field
 314 of Machine Learning. There are many potential societal
 315 consequences of our work, none which we feel must be
 316 specifically highlighted here.

318 References

320 Dong, X., Song, X., Zheng, N., Zhao, S., and Ding, G.
 321 Modality reliability guided multimodal recommenda-
 322 tion, 2025. URL <https://arxiv.org/abs/2504.16524>.

325 Guo, Z. et al. Lgmrec: Local and global graph learning for
 326 multimodal recommendation. In *Proceedings of the AAAI
 327 Conference on Artificial Intelligence*, 2024.

328 He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., and Wang, M.
 329

Lightgcn: Simplifying and powering graph convolution
 network for recommendation. In *Proceedings of the ACM
 SIGIR Conference*, 2020.

Jiang, Y. et al. Diffmm: Multi-modal diffusion model for
 recommendation. In *Proceedings of the ACM Interna-
 tional Conference on Multimedia*, 2024.

Kalantidis, Y. et al. Hard negative mixing for contrastive
 learning. In *Advances in Neural Information Processing
 Systems*, 2020.

Liu, Q., Hu, J., Xiao, Y., Zhao, X., Gao, J., Wang, W.,
 Li, Q., and Tang, J. Multimodal recommender systems:
 A survey, 2024a. URL <https://arxiv.org/abs/2302.03883>.

Liu, Y. et al. Alignrec: Aligning and training in multimodal
 recommendations, 2024b.

Ong, R. K. and Khong, A. W. H. Spectrum-based modality
 representation fusion graph convolutional network for
 multimodal recommendation. In *Proceedings of the Eight-
 eenth ACM International Conference on Web Search and
 Data Mining*, WSDM ’25, pp. 773–781. ACM, March
 2025. doi: 10.1145/3701551.3703561. URL <http://dx.doi.org/10.1145/3701551.3703561>.

Rendle, S., Freudenthaler, C., Gantner, Z., and Schmidt-
 Thieme, L. Bpr: Bayesian personalized ranking from im-
 plicit feedback. In *Proceedings of the Twenty-Fifth Con-
 ference on Uncertainty in Artificial Intelligence*, 2009.

Robinson, J. et al. Contrastive learning with hard nega-
 tive samples. In *International Conference on Learning
 Representations*, 2021.

330 Wei, W., Huang, C., Xia, L., and Zhang, C. Multi-modal
331 self-supervised learning for recommendation. In *Pro-*
332 *ceedings of the ACM Web Conference 2023*, WWW '23,
333 pp. 790–800. ACM, April 2023. doi: 10.1145/3543507.
334 3583206. URL <http://dx.doi.org/10.1145/3543507.3583206>.

335
336 Williams, P. L. and Beer, R. D. Nonnegative decom-
337 position of multivariate information. *arXiv preprint*
338 *arXiv:1004.2515*, 2010.

339
340 Xu, J., Chen, Z., Yang, S., Li, J., Wang, W., Hu, X., Hoi,
341 S., and Ngai, E. A survey on multimodal recommender
342 systems: Recent advances and future directions, 2025a.
343 URL <https://arxiv.org/abs/2502.15711>.

344
345 Xu, J. et al. Mentor: Multi-level self-supervised learning
346 for multimodal recommendation. In *Proceedings of the*
347 *Web Conference*, 2025b.

348
349 Yang, Z., Wang, H., and Hu, D. Efficient quantification
350 of multimodal interaction at sample level, 2025. URL
351 <https://arxiv.org/abs/2506.17248>.

352
353 Yu, P., Tan, Z., Lu, G., and Bao, B.-K. Multi-view graph
354 convolutional network for multimedia recommendation.
355 In *Proceedings of the 31st ACM International Conference*
356 *on Multimedia*, MM '23, pp. 6576–6585. ACM, October
357 2023. doi: 10.1145/3581783.3613915. URL <http://dx.doi.org/10.1145/3581783.3613915>.

358
359 Zhou, X. et al. Bootstrap latent representations for multi-
360 modal recommendation. In *Proceedings of the Web Con-*
361 *ference*, 2023.

362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384

385 A. appendix

Theorem A.1 (No guarantee to resolve unimodal indistinguishability (refined)). *Consider $\mathcal{L}_{\text{total}}$ in (??) trained with negative sampling (??). Under Assumption ??, for any $\varepsilon > 0$ there exist parameters Θ_ε (i.e., encoders ϕ_t, ϕ_v , projection heads used in $\tilde{\mathbf{h}}_t, \tilde{\mathbf{h}}_v$, fusion ϕ_f , and user embeddings $\{\mathbf{e}_u\}$) such that*

$$\mathcal{L}_{\text{InfoNCE}}(\Theta_\varepsilon) \leq \varepsilon, \quad \mathcal{L}_{\text{orth}}(\Theta_\varepsilon) = 0, \quad \mathcal{L}_{\text{BPR}}(\Theta_\varepsilon) \leq \varepsilon + \rho_A M, \quad (6)$$

for some finite constant M (as in Lemma ??), yet the learned model fails to separate modality-ambiguous negatives $\mathcal{A}_v(u, i^+)$ (Definition ??) for a non-negligible fraction of (u, i^+) . Consequently, minimizing (??) does not guarantee eliminating unimodal indistinguishability.