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

Anonymous Authors[1]

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 have achieved remarkable progress in personalized ranking tasks by integrating heterogeneous modality information such as visual and textual content. However, since each modality characterizes items from distinct semantic dimensions, representational discrepancies across modalities are inevitable, giving rise to *modality inconsistency*. Notably, this inconsistency exhibits a pronounced dual nature: when it conveys complementary discriminative information across modalities, it can serve as additional decision-making evidence for ranking, termed **informative inconsistency**; conversely, when it originates from noise perturbations, it may introduce misleading signals, termed **noisy inconsistency**. This observation gives rise to a central challenge in multimodal recommendation:

How can we effectively identify and exploit informative modality inconsistency to enhance recommendation performance, while simultaneously suppressing the adverse effects of noisy inconsistency?

To address this challenge, some studies adopt modality-independent modeling, learning representations for each modality through independent encoders and late fusion. While this approach partially prevents the propagation of modality-specific noise, it restricts cross-modal interaction, making it difficult to mitigate the adverse effects of noisy inconsistency on recommendation performance and to fully capture and exploit informative inconsistency across modalities. More recently, several approaches have introduced self-supervised objectives such as contrastive learning and diffusion models to explicitly enforce cross-modal consistency. These methods uniformly treat all modality inconsistency as noise; although they effectively suppress the negative impact of misleading signals on recommendations, they also sacrifice modality-differentiated characteristics. Consequently, when a single modality is insufficient to fully capture user preferences, such strategies struggle to effectively mine and exploit informative cross-modal inconsistency, thereby limiting the potential for further improvements in recommendation performance.

Through systematic empirical analysis of mainstream multimodal recommendation methods, we identify two noteworthy phenomena: **(1) Under-exploited informative inconsistency.** In numerous mis-ranking cases, negative samples exhibit high similarity to the target item in one modality while displaying significant differences in another modality.

[1]Anonymous 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.

This indicates that existing multimodal recommendation systems fail to fully leverage informative modality inconsistency. **(2) Noisy inconsistency undermines fusion.** We observe cases where unimodal representations successfully retrieve the target item, whereas multimodal fused representations fail. This suggests that noisy modality inconsistency may weaken the advantages of multimodal fusion.

To address the aforementioned challenges, this paper proposes **BR-MRS**, a multimodal recommendation framework that achieves effective discrimination and exploitation of modality inconsistency by reformulating the classical Bayesian Personalized Ranking loss (BPR Loss). Specifically, we first design a *Cross-modal Hard Negative Sampling* (CHNS) strategy, which constructs discriminatively challenging negatives across different modalities to explicitly guide the model toward attending to cross-modal differential features that are valuable for user preference ranking, thereby more effectively capturing and exploiting informative inconsistency across modalities. Furthermore, to mitigate the degradation problem that arises during cross-modal fusion, we propose a *Synergy-aware Bayesian Personalized Ranking Loss*, which constrains the fused multimodal representation to significantly outperform each unimodal representation in distinguishing positive 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. As illustrated in Fig. **??**, BR-MRS first follows the mainstream multimodal recommendation paradigm, employing graph neural networks to learn user and item representations. Building upon this foundation, BR-MRS introduces two core components: (i) *Cross-modal Hard Negative Sampling* (CHNS), which mines confusable samples from one modality to drive the other modality to provide discriminative evidence, thereby exploiting informative inconsistency; and (ii) *Synergy-aware BPR Loss*, which constrains the preference margin of the fused representation to significantly exceed that of any unimodal branch, thereby suppressing fusion degeneration caused by noisy inconsistency. Algorithm **??** 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 λ_h, λ_s, θ, λ
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 $\mathcal{L}_{\text{syn}}$
9:     Compute $\mathcal{L}_{\text{chns}}$ and $\mathcal{L}_{\text{syn}}$
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

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

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 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 $m \in \{t, v, f\}$.

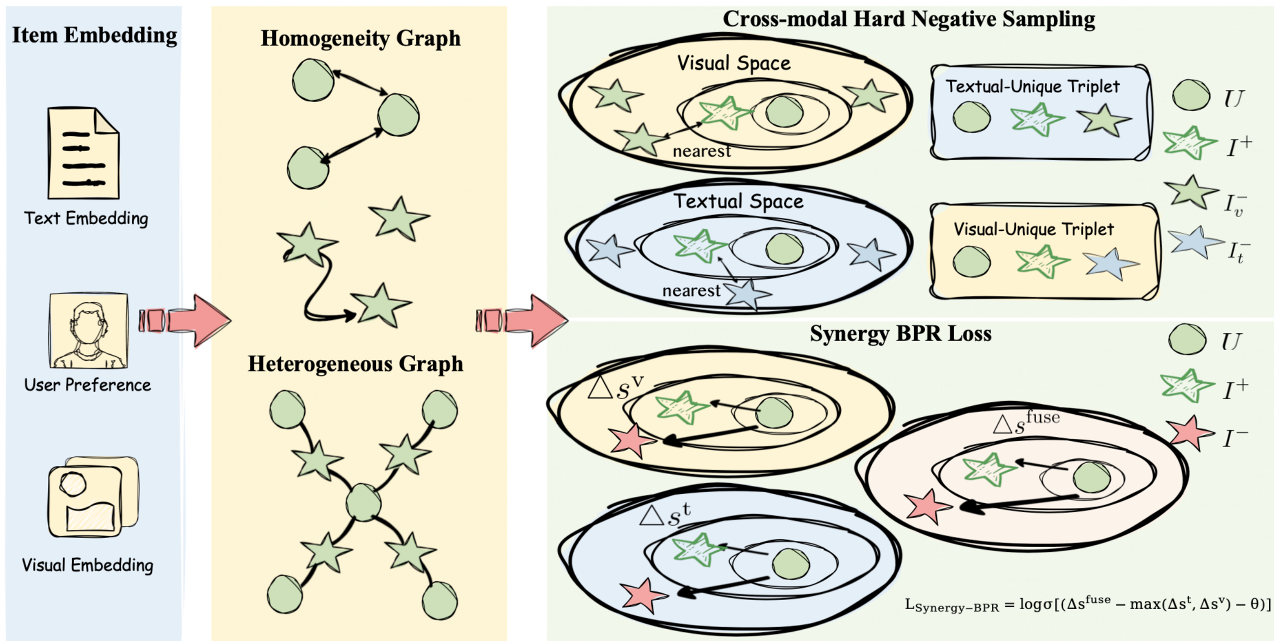

Figure 1. **Overview of BR-MRS.**

2.2. Cross-modal Hard Negative Sampling

Having obtained unimodal representations and the corresponding preference scores, we further investigate how to explicitly exploit informative inconsistency across modalities to enhance user preference characterization. Through empirical analysis, we observe that informative inconsistency often manifests as *cross-modal confusion*: for a given positive pair (u, i^+), certain negatives score highly under one modality and are thus difficult to distinguish from the positive, yet exhibit significantly lower scores under the other modality and are therefore easily separable. This phenomenon indicates that confusable samples in one modality can effectively expose the discriminative advantage of the other.

Inspired by the classical BPR negative sampling strategy, we propose *Cross-modal Hard Negative Sampling* (CHNS), which leverages confusable negatives from one modality to drive the other modality to explicitly contribute discriminative evidence. Specifically, for each positive pair (u, i^+), we first sample a candidate negative pool $\mathcal{N}(u) \subseteq \mathcal{I} \setminus \mathcal{O}_u$, 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:

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

Through this cross-modal supervision mechanism, CHNS explicitly activates the discriminative advantage of each modality, transforming informative inconsistency into effective supervisory signals to more precisely exploit modality-complementary information.

2.3. Synergy-aware BPR Loss

Although CHNS effectively exploits informative inconsistency across modalities, the presence of noisy inconsistency may lead to multimodal fusion degeneration, where the pairwise ranking capability of the fused representation becomes inferior to that of a unimodal branch. To address this, we propose the *Synergy-aware BPR Loss* based on the BPR framework, which explicitly constrains the preference margin of the fused representation to exceed that of any unimodal branch, thereby ensuring the robustness of the fusion mechanism.

Concretely, for a training triple (u, i^+, i^-) where i^- is uniformly sampled from non-interacted items, we define the preference margins for the fused and unimodal branches as:

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

Prior research has demonstrated that the magnitude of preference margins directly reflects the model's ranking confidence and discriminative capacity; larger margins correspond to more reliable pairwise ranking discrimination and exhibit strong consistency with Top-K ranking objectives in recommendation. 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:

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

By explicitly enforcing $\Delta_f > \max(\Delta_t, \Delta_v) + \theta$, the synergy-aware loss effectively suppresses the interference of noisy modalities, ensuring that the fused representation consistently maintains its advantage over any unimodal branch, thereby robustly 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:

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

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, DualGNN, GRCN, LATTICE, BM3, SLMRec, MICRO, MGCN, 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 **??** 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 **??**, 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 ✔ indicates the component is enabled, while ○ indicates it is disabled.

As shown in Table **??**, 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.

Table 1. Results on Benchmark Datasets

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	N@20
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	↑ 11.7%	↑ 6.5%	↑ 20.5%	↑ 10.8%	↑ 5.7%	↑ 4.0%	↑ 9.4%	↑ 7.9%	↑ 5.9%	↑ 8.0%	↑ 10.2%	↑ 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

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

	Baby				**Sports**				**Clothing**			
Method	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%	

Table 3. Ablation study on two benchmark datasets. We report 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**

NDCG@10. Ablation studies confirm the complementary contributions of both proposed components. Our work provides new insights into how multimodal information should be leveraged for personalized ranking and offers a principled approach for future multimodal recommendation research.

Impact Statement

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

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*

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

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.*