

---

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

---

Anonymous Authors<sup>1</sup>

## 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

Multimodal recommendation systems integrate heterogeneous item modalities (e.g., visual and textual content) and have been shown to substantially improve personalized ranking performance (Zhou et al., 2023b;a; Guo et al., 2024).

<sup>1</sup>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.

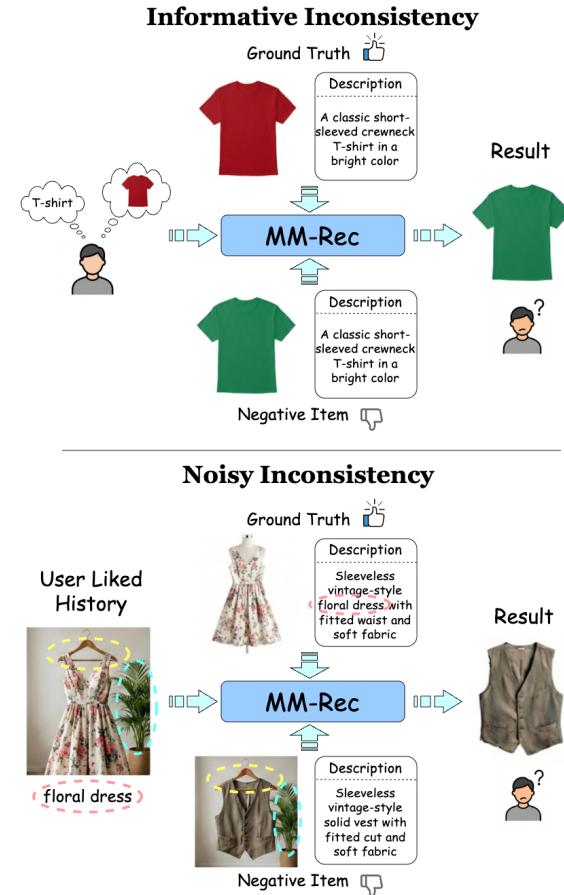


Figure 1. Illustration of two phenomena associated with modality inconsistency. (1) **Under-exploited informative inconsistency.** A user prefers a red T-shirt; two candidates are indistinguishable in text but differ in image color, yet the model misses this cross-modal cue and mis-ranks a hard negative. (2) **Noisy inconsistency harms fusion.** A user prefers a floral dress; text alone separates items correctly, but noisy visual resemblance of the negative misleads multimodal fusion, causing fusion degeneration.

Different modalities capture distinct aspects of the same item, which naturally leads to *modality inconsistency*. Such inconsistency can be *beneficial* when modality-specific differences provide complementary evidence for recommendation, yet it can also be *harmful* when it originates from noise and introduces misleading signals. Based on this observa-

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

tion, we distinguish two types of inconsistency: **informative inconsistency**, which provides discriminative cues for ranking, and **noisy inconsistency**, which impairs ranking performance. Therefore, a central challenge in multimodal recommendation is:

When a single modality is insufficient, how can we exploit informative inconsistency while suppressing noisy inconsistency?

Existing solutions tackle modality inconsistency from different perspectives. Some studies adopt *modality-independent modeling* with independent encoders and late fusion (Zhou et al., 2023b; Guo et al., 2024), which can partially prevent modality-specific noise from propagating, but it restricts cross-modal interaction and makes it difficult to both avoid the harm of noisy inconsistency and fully utilize informative inconsistency. More recent approaches introduce self-supervised objectives such as contrastive learning (e.g., InfoNCE (Oord et al., 2018)) and diffusion models to *explicitly enforce cross-modal consistency* (Zhou et al., 2023a; Zhang et al., 2024; Jiang et al., 2024; Xu et al., 2025; Yi et al., 2022). These methods largely treat all inconsistency as noise: while effective at suppressing misleading signals, they also discard modality-differentiated characteristics. As a result, when one modality alone cannot fully capture user preference, purely alignment-driven strategies may fail to excavate and utilize informative cross-modal differences, limiting further ranking gains.

Our empirical analysis further reveals two noteworthy phenomena that expose these limitations. **(1) Informative inconsistency is under-exploited.** We observe many misranked items (i.e., *hard negatives*) that appear highly similar to the target in *one* modality, but exhibit clear differences in *another* modality. This indicates that *informative inconsistency* can provide discriminative cues that could resolve confusion, yet alignment-dominant designs tend to wash them out, leading to ranking mistakes. **(2) Noisy inconsistency harms fusion.** In some cases, a unimodal representation can successfully retrieve the target item, while the fused multimodal representation fails. This indicates that fusion without explicitly separating *noisy inconsistency* from informative signals may propagate spurious cross-modal conflicts, thereby missing complementarity and even weakening unimodal advantages.

To address these challenges, we propose **BR-MRS**, a multimodal recommendation framework that reconstructs the classic Bayesian Personalized Ranking objective (BPR loss) to explicitly separate and utilize modality inconsistency. Specifically, we design *Cross-modal Hard Negative Sampling* (CHNS) to construct challenging negatives across modalities, explicitly guiding the model to attend to modality-specific differences that are valuable for rank-

ing, thereby better capturing and exploiting *informative inconsistency*. Furthermore, to mitigate fusion degeneration, we propose a *Synergy-aware BPR Loss* that constrains the fused multimodal representation to be *significantly better* than each unimodal branch at separating positives from negatives, thereby suppressing the adverse effect of *noisy inconsistency* and improving multimodal fusion.

Our contributions are summarized as follows:

- **New Findings.** We explicitly distinguish *informative* versus *noisy* modality inconsistency, and empirically show that prior multimodal recommenders often under-exploit the former and are vulnerable to the latter.
- **Novel Method.** We propose **BR-MRS**, which combines cross-modal hard negative sampling to activate informative inconsistency with a synergy-aware ranking objective that constrains the fused representation to outperform unimodal branches.
- **Impressive Performance.** Extensive experiments on multiple benchmarks demonstrate that **BR-MRS** consistently outperforms state-of-the-art methods, with ablations and case studies validating the effectiveness of each component.

## 2. Related Work

**Multimodal Recommendation.** Multimodal recommendation has attracted extensive attention in recent years, with a growing body of research addressing various aspects of the field. Leveraging graph neural networks (GNNs), multimodal graph neural networks such as LGMRec(Guo et al., 2024) and FREEDOM(Zhou et al., 2023b) have been developed to model users' multimodal preferences. Integrating multimodal alignment algorithms, including multimodal contrastive learning (e.g., BM3 (Zhou et al., 2023a), FET-TLE(Zhang et al., 2024), AlignRec(Liu et al., 2024)) and multimodal diffusion models (Jiang et al., 2024) (e.g., Dif, MCDRec), researchers have successfully captured cross-modal representation consistency for items and multimodal relationships between users and items. Following these developments, approaches like MENTOR(Xu et al., 2025) and MMGCL(Yi et al., 2022) have employed random graph augmentation strategies to alleviate biases inherent in interaction data. However, these methods largely overlook biases arising from modality confounding and fail to address interaction biases from the perspective of modality representations.

**Partial Information Decomposition in Multimodal Learning.** Partial Information Decomposition (PID) provides a principled framework for decomposing multimodal information into unique, redundant, and synergistic components (Williams & Beer, 2010). In general multimodal

learning, PID-inspired methods have been proposed to disentangle modality-specific and shared representations. For instance, orthogonality constraints are commonly used to suppress redundancy (Liu et al., 2021), while contrastive objectives like InfoNCE (Oord et al., 2018) enforce cross-modal consistency. Despite their success in tasks such as cross-modal retrieval, these components are designed for agreement-based objectives rather than user-conditioned ranking. Our work reveals that directly transferring these designs to recommendation leads to suboptimal utilization of unique and synergistic information.

**Hard Negative Mining.** Hard negative sampling has proven effective in representation learning by focusing on difficult-to-distinguish samples (Robinson et al., 2021; Kalantidis et al., 2020). In recommendation, hard negatives have been leveraged to improve the discriminability of learned representations (Zhang et al., 2023). However, existing approaches typically mine hard negatives within a single representation space without considering cross-modal interactions. Our proposed Cross-modal Hard Negative Sampling (CHNS) differs fundamentally by using one modality to identify confusable cases and assigning the other modality to resolve them, thereby explicitly activating modality-specific evidence for personalized ranking.

### 3. Preliminaries

#### 3.1. Problem Formulation

**Implicit-feedback recommendation.** We consider an implicit-feedback setting with a user set  $\mathcal{U}$  and an item set  $\mathcal{I}$ . Observed interactions are denoted by  $\mathcal{O} \subseteq \mathcal{U} \times \mathcal{I}$ , where  $(u, i) \in \mathcal{O}$  indicates that user  $u$  has interacted with item  $i$ . For each user  $u$ , we write  $\mathcal{O}_u = \{i \in \mathcal{I} : (u, i) \in \mathcal{O}\}$ .

**Graph construction.** Following prior work (Zhou et al., 2023b; Guo et al., 2024), we construct three types of graphs to capture different relational structures: (i) a *user-item bipartite graph*  $\mathcal{G}_{ui} = (\mathcal{U} \cup \mathcal{I}, \mathcal{O})$  encoding interaction signals; (ii) a *user-user homogeneous graph*  $\mathcal{G}_{uu}$  where edges connect users with similar interaction patterns; and (iii) an *item-item homogeneous graph*  $\mathcal{G}_{ii}^{(m)}$  for each modality  $m$ , where edges connect items with similar content features. These graphs are processed by graph neural networks to obtain refined user and item representations.

**Scoring functions.** For each user  $u \in \mathcal{U}$ , we learn an embedding  $\mathbf{p}_u \in \mathbb{R}^d$ . For each item  $i \in \mathcal{I}$ , we construct modality-specific representations  $\mathbf{q}_i^{(t)}, \mathbf{q}_i^{(v)} \in \mathbb{R}^d$  from text and visual features respectively, and a fused representation  $\mathbf{q}_i^{(f)} \in \mathbb{R}^d$  via a fusion function  $\phi(\cdot)$ :

$$\mathbf{q}_i^{(f)} = \phi\left(\mathbf{q}_i^{(t)}, \mathbf{q}_i^{(v)}\right). \quad (1)$$

We define unimodal and fused ranking scores by dot product:

$$\begin{aligned} s_t(u, i) &= \langle \mathbf{p}_u, \mathbf{q}_i^{(t)} \rangle, \\ s_v(u, i) &= \langle \mathbf{p}_u, \mathbf{q}_i^{(v)} \rangle, \\ s_f(u, i) &= \langle \mathbf{p}_u, \mathbf{q}_i^{(f)} \rangle. \end{aligned} \quad (2)$$

The fused score  $s_f$  is used for final ranking.

**BPR objective.** We adopt Bayesian Personalized Ranking (BPR) (Rendle et al., 2009) as the canonical pairwise supervision. For each observed pair  $(u, i^+) \in \mathcal{O}$ , we sample a negative item  $i^- \notin \mathcal{O}_u$  and form a triple  $(u, i^+, i^-)$ . Let  $\Delta_f = s_f(u, i^+) - s_f(u, i^-)$  be the fused preference margin. The BPR loss is

$$\mathcal{L}_{\text{BPR}} = - \sum_{(u, i^+, i^-)} \log \sigma(\Delta_f), \quad (3)$$

where  $\sigma(\cdot)$  is the sigmoid function.

#### 3.2. User Subset Partitioning

Given a scoring function  $s(\cdot, \cdot)$ , we define the rank of the target item  $i^+$  for user  $u$  as

$$r_s(u, i^+) = 1 + |\{j \in \mathcal{I} \setminus \{i^+\} : s(u, j) > s(u, i^+)\}|. \quad (4)$$

Under leave-one-out evaluation, each user has a single held-out target item  $i_u^+$ , and all ranks and user subsets are defined with respect to this item. A recall at cutoff  $K$  is successful if  $r_s(u, i^+) \leq K$ . Using  $r_t(\cdot, \cdot)$ ,  $r_v(\cdot, \cdot)$  and  $r_f(\cdot, \cdot)$  induced by  $s_t$ ,  $s_v$  and  $s_f$ , we partition users into four sets:

$$\mathcal{U}_t = \left\{ u : r_t(u, i^+) \leq K, \begin{array}{l} r_v(u, i^+) > K, r_f(u, i^+) > K \end{array} \right\}, \quad (5)$$

$$\mathcal{U}_v = \left\{ u : r_v(u, i^+) \leq K, \begin{array}{l} r_t(u, i^+) > K, r_f(u, i^+) > K \end{array} \right\}, \quad (6)$$

$$\mathcal{U}_{tv} = \left\{ u : r_f(u, i^+) \leq K, \begin{array}{l} r_t(u, i^+) > K, r_v(u, i^+) > K \end{array} \right\}, \quad (7)$$

$$\mathcal{U}_r = \left\{ u : r_t(u, i^+) \leq K, \begin{array}{l} r_v(u, i^+) \leq K, r_f(u, i^+) > K \end{array} \right\}. \quad (8)$$

Intuitively,  $\mathcal{U}_t$  and  $\mathcal{U}_v$  correspond to cases where a single modality suffices,  $\mathcal{U}_{tv}$  captures synergy where only fusion succeeds, and  $\mathcal{U}_r$  indicates a degradation regime where fusion fails despite both unimodal branches succeeding.

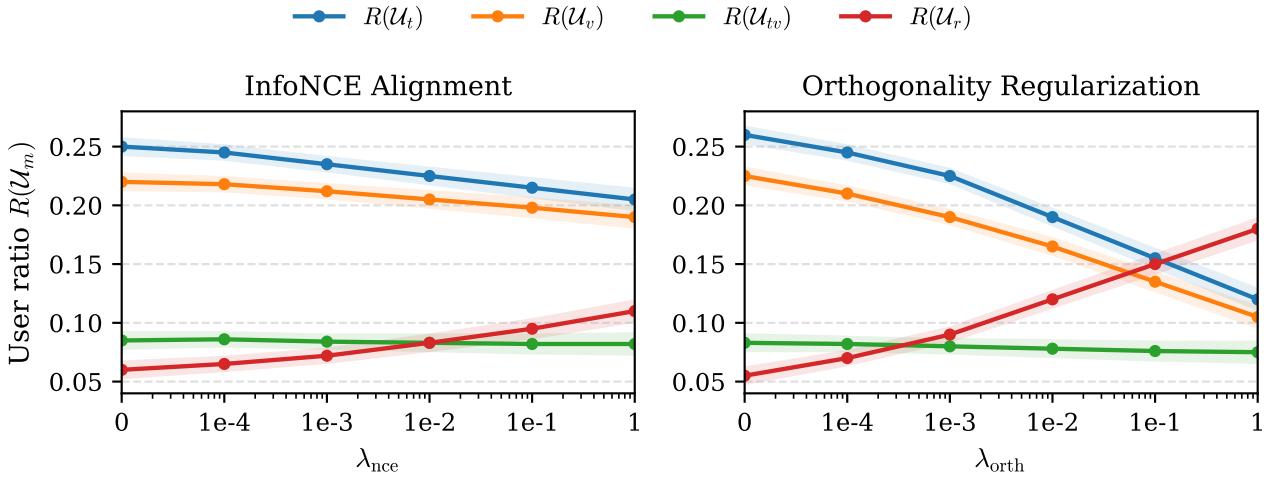


Figure 2. Effects of alignment and orthogonality on user subsets. User ratios  $R(\mathcal{U}_m)$  versus regularization strength on the Baby dataset. InfoNCE alignment leaves the synergy subset nearly unchanged, while stronger orthogonality reduces modality-unique subsets and enlarges the degradation regime.

### 3.3. Empirical Observations

We empirically examine how two widely adopted components—InfoNCE-style alignment and orthogonality-based decorrelation—shape *different types of multimodal interaction* under personalized ranking. Based on the user subset partitioning defined above, we partition users into  $\{\mathcal{U}_t, \mathcal{U}_v, \mathcal{U}_{tv}, \mathcal{U}_r\}$  and quantify each type by its ratio

$$R(\mathcal{U}_m) = \frac{|\mathcal{U}_m|}{|\mathcal{U}|}, \quad \mathcal{U}_m \in \{\mathcal{U}_t, \mathcal{U}_v, \mathcal{U}_{tv}, \mathcal{U}_r\}. \quad (9)$$

Larger  $R(\mathcal{U}_t)$  or  $R(\mathcal{U}_v)$  indicates that *unique* modality evidence is necessary; larger  $R(\mathcal{U}_{tv})$  indicates *synergy* where only fusion succeeds; and larger  $R(\mathcal{U}_r)$  indicates a *degradation* regime where fusion fails despite both unimodal branches succeeding.

We sweep the regularization strengths  $\lambda_{\text{orth}}$  and  $\lambda_{\text{nec}}$  on the Baby dataset. Two consistent patterns emerge (Fig. 2).

**Orthogonality suppresses uniqueness.** As  $\lambda_{\text{orth}}$  increases,  $R(\mathcal{U}_t)$  and  $R(\mathcal{U}_v)$  decrease steadily, while  $R(\mathcal{U}_r)$  increases. That is, stronger decorrelation does not yield more users that benefit from modality-specific evidence; instead, it enlarges the regime where fusion degrades.

**Alignment does not induce synergy.** As  $\lambda_{\text{nec}}$  increases,  $R(\mathcal{U}_{tv})$  remains nearly unchanged. In other words, enforcing cross-modal consistency alone provides little incentive to form synergistic signals that become useful *only* through fusion.

These observations reveal a mismatch between generic geometric regularization and personalized ranking: orthog-

onality tends to erode unique utility, and InfoNCE-style alignment fails to create synergy.

## 4. The Proposed Method

In this section, we present **BR-MRS**, a multimodal recommendation framework designed to exploit informative inconsistency while suppressing noisy inconsistency under personalized ranking. An overview is illustrated in Fig. 3. BR-MRS comprises two core components: (i) *Cross-modal Hard Negative Sampling* (CHNS), which exploits informative inconsistency by constructing hard negatives across modalities and explicitly driving the model to leverage modality-specific differences that are discriminative for ranking; and (ii) *Synergy-aware BPR Loss*, which suppresses noisy inconsistency by constraining the fused representation to outperform each unimodal branch in distinguishing positive and negative items, thereby alleviating fusion degeneration. Algorithm 1 summarizes the overall training procedure.

### 4.1. Cross-modal Hard Negative Sampling

To effectively exploit informative inconsistency for better user preference characterization, we need a mechanism that identifies and leverages the discriminative evidence from each modality. By explicitly mining such cases, we find that informative inconsistency manifests when one modality is confused by certain negatives while the other modality can provide discriminative evidence.

Motivated by this observation, we revisit the negative sampling strategy in recommender systems (Zhang et al., 2023). Prior methods mine hard negatives based on the fused repre-

220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230  
231  
232  
233  
234  
235  
236  
237  
238  
239  
240

241  
242  
243

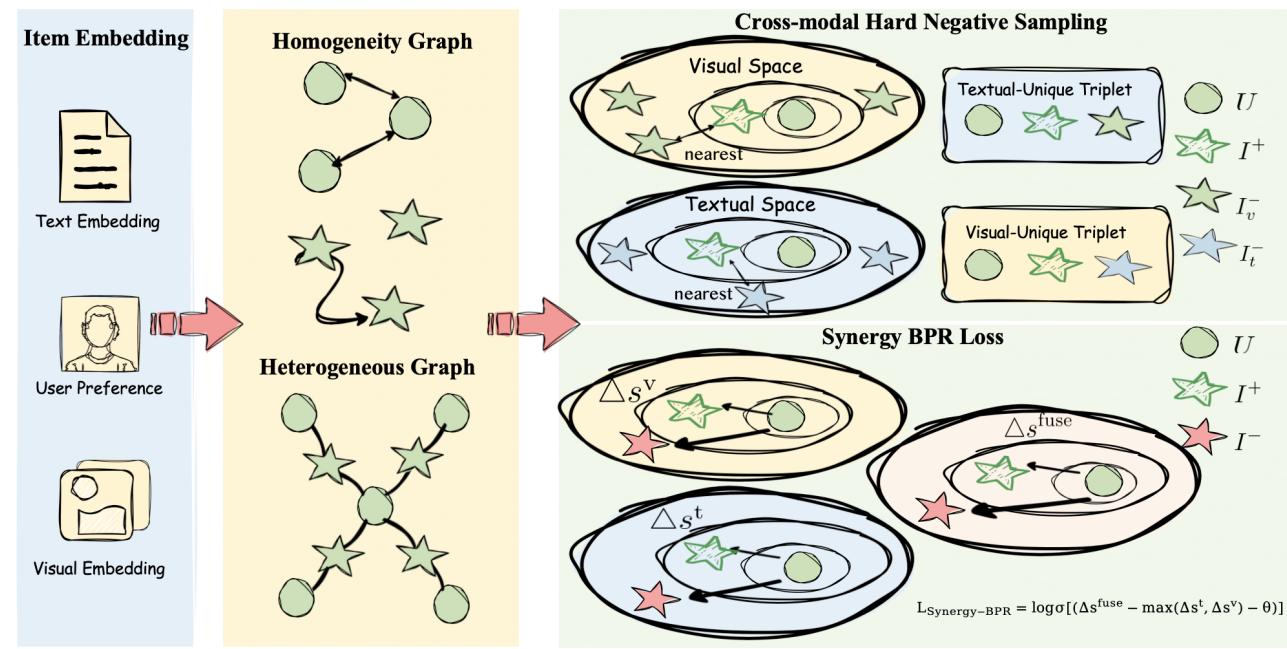


Figure 3. Overview of BR-MRS.

**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  $\Theta$

- 1: Initialize model parameters  $\Theta$
- 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  $\Theta$  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**

sentation, while ranking errors may originate from specific modalities where user preferences and informative inconsistency have not been sufficiently captured. To this end, we propose *Cross-modal Hard Negative Sampling* (CHNS), which constructs hard negatives across modalities: each modality is trained to resolve confusable cases identified by the other modality, thereby explicitly activating modality-specific discriminative capacity to exploit informative inconsistency. Specifically, for each positive pair  $(u, i^+) \in \mathcal{O}$ , we first sample a candidate negative pool  $\mathcal{N}(u) \subseteq \mathcal{I} \setminus \mathcal{O}_u$ ,

and then identify the most confusable negative under each modality:

$$\begin{aligned} i_v^-(u, i^+) &= \arg \max_{j \in \mathcal{N}(u)} s_v(u, j), \\ i_t^-(u, i^+) &= \arg \max_{j \in \mathcal{N}(u)} s_t(u, j). \end{aligned} \quad (10)$$

Here,  $j_v$  denotes the negative that is hardest to distinguish from the positive under the visual modality, while  $j_t$  denotes the most confusable negative under the textual modality. We then *swap* these confusable negatives to train the *other* modality branch, explicitly requiring each modality to contribute its unique information as discriminative evidence:

$$\begin{aligned} \mathcal{L}_{\text{chns}}^v &= - \sum_{(u, i^+) \in \mathcal{O}} \log \sigma(s_v(u, i^+) - s_v(u, i_t^-(u, i^+))) \\ \mathcal{L}_{\text{chns}}^t &= - \sum_{(u, i^+) \in \mathcal{O}} \log \sigma(s_t(u, i^+) - s_t(u, i_v^-(u, i^+))) \\ \mathcal{L}_{\text{chns}} &= \mathcal{L}_{\text{chns}}^v + \mathcal{L}_{\text{chns}}^t \end{aligned} \quad (11)$$

$\mathcal{L}_{\text{chns}}^v$  requires the visual modality to discriminate negatives that confuse the textual modality;  $\mathcal{L}_{\text{chns}}^t$  requires the textual modality to resolve visually confusable cases. Through this cross-modal supervision design, CHNS explicitly exploits informative inconsistency by forcing each modality to leverage its unique discriminative strengths for personalized ranking.

---

275    **4.2. Synergy-aware BPR Loss**

276 CHNS targets *informative* inconsistency via cross-modal  
 277 hard negatives; however, multimodal fusion can still *de-*  
 278 *generate* when noisy inconsistency misleads the fusion  
 279 process. We therefore propose a *Synergy-aware BPR Loss* that  
 280 enforces a strict preference-margin constraint: the fused  
 281 branch must achieve a margin larger than the strongest uni-  
 282 modal branch by a positive gap  $\theta$ . For a training triple  
 283  $(u, i^+, i^-)$  where  $i^-$  is sampled from  $\mathcal{N}(u)$  uniformly, we  
 284 define the preference margins:  
 285

$$\begin{aligned}\Delta_f &= s_f(u, i^+) - s_f(u, i^-), \\ \Delta_t &= \text{sg}[s_t(u, i^+) - s_t(u, i^-)], \\ \Delta_v &= \text{sg}[s_v(u, i^+) - s_v(u, i^-)].\end{aligned}\quad (12)$$

291 Where sg means stop gradient. We impose a margin  $\theta > 0$   
 292 and define the synergy-aware term:

$$\mathcal{L}_{\text{syn}} = - \sum_{(u, i^+, i^-)} \log \sigma(\Delta_f - \max(\Delta_t, \Delta_v) - \theta). \quad (13)$$

297 By constraining  $\Delta_f > \max(\Delta_t, \Delta_v) + \theta$ , this loss directly  
 298 operationalizes the principle that *fusion should be stronger*  
 299 *than any single modality* in pairwise ranking discrimina-  
 300 tion. Through this explicit constraint, the synergy-aware  
 301 loss suppresses noisy inconsistency by ensuring that multi-  
 302 modal fusion surpasses any unimodal branch in personalized  
 303 ranking, preventing fusion degeneration.

304    **4.3. Overall Objective**

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

$$\mathcal{L} = \lambda_h \mathcal{L}_{\text{chns}} + \lambda_s \mathcal{L}_{\text{syn}} + \lambda \|\Theta\|_2^2, \quad (14)$$

312 where  $\lambda_h$  and  $\lambda_s$  control the contributions of CHNS and  
 313 the synergy-aware loss respectively,  $\lambda$  is the regularization  
 314 coefficient, and  $\Theta$  denotes all trainable parameters.

316 This unified design ensures that informative inconsistency  
 317 is effectively exploited through cross-modal hard negative  
 318 mining, while noisy inconsistency is suppressed through  
 319 the synergy-aware constraint, enabling BR-MRS to achieve  
 320 superior personalized ranking performance.

321    **5. Experiment**

322    **5.1. Experimental Setup**

325 We evaluate BR-MRS on three public multimodal recom-  
 326 mendation benchmarks, namely Baby, Sports, and Clothing,  
 327 where each item is associated with both visual and textual  
 328 content features. We follow standard preprocessing and

329 splitting protocols used in prior multimodal recommenda-  
 330 tion work to ensure fair comparison. We adopt leave-one-out  
 331 evaluation and report Recall@K and NDCG@K with  
 332  $K \in \{10, 20\}$ . Baselines cover classical CF models (e.g.,  
 333 BPR, LightGCN, ApeGNN, MGDN) and a broad range of  
 334 multimodal recommenders (e.g., VBPR, MMGCN, Dual-  
 335 GNN, GRCN, LATTICE, BM3, SLMRec, MICRO, MGCN,  
 336 FREEDOM, LGMRec, DRAGON, MIG-GT, REARM). For  
 337 all methods, hyperparameters are tuned on validation sets,  
 338 and we use the same multimodal features and evaluation  
 339 pipeline for a fair comparison.

340    **5.2. Overall Performance**

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

350    **5.3. Ablation Study**

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

357 As shown in Table 2, disabling either component leads to  
 358 performance degradation. When only CHNS is enabled, the  
 359 model can mine modality-specific discriminative evidence  
 360 but lacks explicit synergy constraints. When only Syn is  
 361 enabled, the model enforces fusion superiority but misses the  
 362 cross-modal hard negative mining. The full model with both  
 363 components achieves the best performance, demonstrating  
 364 their complementary contributions.

365    **5.4. Hyper-parameter and Robustness Analysis**

366 We analyze the sensitivity of BR-MRS to key hyperparam-  
 367 eters, including  $\lambda_h$  (weight of CHNS),  $\lambda_s$  (weight of synergy-  
 368 aware loss), and  $\theta$  (synergy margin). Performance remains  
 369 stable across a broad range of values, with moderate  $\theta$  yield-  
 370 ing the best trade-off between unimodal stability and fusion  
 371 gains. We also observe that BR-MRS maintains consistent  
 372 improvements under different evaluation cutoffs, indicating  
 373 robustness to the choice of ranking metric. Detailed  
 374 curves and additional robustness results are deferred to the  
 375 Appendix.

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
BPR	0.0357	0.0575	0.0192	0.0249	0.0432	0.0653	0.0241	0.0298	0.0206	0.0303	0.0114	0.0138
LightGCN	0.0479	0.0754	0.0257	0.0328	0.0569	0.0864	0.0311	0.0387	0.0361	0.0544	0.0197	0.0243
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	0.0547	0.0850	0.0292	0.0370	0.0620	0.0953	0.0335	0.0419	0.0492	0.0733	0.0268	0.0330
BM3	0.0564	0.0883	0.0301	0.0383	0.0656	0.0980	0.0355	0.0438	0.0422	0.0621	0.0231	0.0281
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	0.0620	0.0964	0.0339	0.0427	0.0729	0.1106	0.0397	0.0496	0.0641	0.0945	0.0347	0.0428
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	0.0662	0.1021	0.0345	0.0435	0.0752	0.1139	0.0413	0.0512	0.0671	0.0979	0.0365	0.0443
MIG-GT	0.0665	0.1021	0.0361	0.0452	0.0753	0.1130	0.0414	0.0511	0.0636	0.0934	0.0347	0.0422
REARM	0.0705	0.1105	0.0377	0.0479	0.0836	0.1231	0.0455	0.0553	0.0700	0.0998	0.0377	0.0454
SSR	0.0728	0.1103	0.0395	0.0491	0.0825	0.1203	0.0449	0.0547	0.0708	0.1032	0.0386	0.0466
<b>BR-MRS</b>	<b>0.0819</b>	<b>0.1215</b>	<b>0.0452</b>	<b>0.0554</b>	—	—	—	—	—	—	—	—
Improve	↑ 16.1%	↑ 9.9%	↑ 23.1%	↑ 15.4%	↑ -%	↑ -%	↑ -%	↑ -%	↑ -%	↑ -%	↑ -%	↑ -%

Table 2. 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	—	—
○	✓	0.0767	0.0411	—	—
✓	✓	<b>0.0819</b>	<b>0.0452</b>	—	—

## 5.5. 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.

## 5.6. 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.

## 6. 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 NDCG@10. Ablation studies confirm the complementary

385 contributions of both proposed components. Our work pro-  
386 vides new insights into how multimodal information should  
387 be leveraged for personalized ranking and offers a principled  
388 approach for future multimodal recommendation research.  
389

## 390 Impact Statement

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

## 396 References

- 399 Guo, Z. et al. Lgmrec: Local and global graph learning for  
400 multimodal recommendation. In *Proceedings of the AAAI*  
401 Conference on Artificial Intelligence, 2024.
- 402 Jiang, Y. et al. Diffmm: Multi-modal diffusion model for  
403 recommendation. In *Proceedings of the ACM Interna-*  
404 *tional Conference on Multimedia*, 2024.
- 405 Kalantidis, Y. et al. Hard negative mixing for contrastive  
406 learning. In *Advances in Neural Information Processing*  
407 Systems, 2020.
- 408 Liu, W. et al. Orthogonal projection loss. In *Proceedings*  
409 *of the IEEE/CVF International Conference on Computer*  
410 *Vision*, 2021.
- 411 Liu, Y. et al. Alignrec: Aligning and training in multimodal  
412 recommendations, 2024.
- 413 Oord, A. v. d., Li, Y., and Vinyals, O. Representation  
414 learning with contrastive predictive coding. In *Advances*  
415 *in Neural Information Processing Systems*, 2018.
- 416 Rendle, S., Freudenthaler, C., Gantner, Z., and Schmidt-  
417 Thieme, L. Bpr: Bayesian personalized ranking from im-  
418 plicit feedback. In *Proceedings of the Twenty-Fifth Con-*  
419 *ference on Uncertainty in Artificial Intelligence*, 2009.
- 420 Robinson, J. et al. Contrastive learning with hard nega-  
421 tive samples. In *International Conference on Learning*  
422 *Representations*, 2021.
- 423 Williams, P. L. and Beer, R. D. Nonnegative decom-  
424 position of multivariate information. *arXiv preprint*  
425 *arXiv:1004.2515*, 2010.
- 426 Xu, J. et al. Mentor: Multi-level self-supervised learning  
427 for multimodal recommendation. In *Proceedings of the*  
428 *Web Conference*, 2025.
- 429 Yi, Z. et al. Multi-modal graph contrastive learning for  
430 micro-video recommendation. In *Proceedings of the ACM*  
431 *SIGIR Conference*, 2022.
- 432

Zhang, A. et al. Empowering collaborative filtering with  
principled adversarial contrastive loss. In *Advances in  
Neural Information Processing Systems*, 2023.

Zhang, W. et al. Fettle: Feature-enhanced text-to-text learn-  
ing for recommendation. In *Proceedings of the ACM  
SIGIR Conference*, 2024.

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

Zhou, X. et al. A tale of two graphs: Freezing and denoising  
graph structures for multimodal recommendation. In  
*Proceedings of the ACM International Conference on  
Multimedia*, 2023b.

---

## 440 A. appendix

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

$$445 \quad \mathcal{L}_{\text{InfoNCE}}(\Theta_\varepsilon) \leq \varepsilon, \quad \mathcal{L}_{\text{orth}}(\Theta_\varepsilon) = 0, \quad \mathcal{L}_{\text{BPR}}(\Theta_\varepsilon) \leq \varepsilon + \rho_A M, \quad (15)$$

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

449 *Proof.* Fix an arbitrary  $\varepsilon > 0$ . We construct a family of representations that attains low loss while provably lacking  
 450 modality-unique discriminative evidence.

451 **Block-orthogonal parametrization.** Let the embedding space decompose into three orthogonal subspaces  $\mathbb{R}^d = \mathcal{S}_t \oplus \mathcal{S}_v \oplus$   
 452  $\mathcal{S}_p$  with dimensions  $d = d_c + d_c + d_p$ . For each item  $i$ , define a shared factor  $\mathbf{c}_i \in \mathbb{R}^{d_c}$  and an (optional) modality-private  
 453 factor  $\mathbf{u}_i \in \mathbb{R}^{d_p}$ . We realize modality embeddings as

$$454 \quad \mathbf{h}_t^i = \begin{bmatrix} \mathbf{c}_i \\ \mathbf{0} \\ \mathbf{u}_i \end{bmatrix}, \quad \mathbf{h}_v^i = \begin{bmatrix} \mathbf{0} \\ \mathbf{c}_i \\ \mathbf{0} \end{bmatrix}. \quad (16)$$

455 Let  $\mathbf{P}_t = [\mathbf{I}_{d_c} \ \mathbf{0} \ \mathbf{0}]$  and  $\mathbf{P}_v = [\mathbf{0} \ \mathbf{I}_{d_c} \ \mathbf{0}]$  be selection matrices. Define the contrastive embeddings by projection heads

$$456 \quad \tilde{\mathbf{h}}_t^i = \mathbf{P}_t \mathbf{h}_t^i = \mathbf{c}_i, \quad \tilde{\mathbf{h}}_v^i = \mathbf{P}_v \mathbf{h}_v^i = \mathbf{c}_i. \quad (17)$$

457 **Orthogonality term is exactly minimized.** Stacking item embeddings yields

$$458 \quad \mathbf{H}_t = \begin{bmatrix} \mathbf{C} \\ \mathbf{0} \\ \mathbf{U} \end{bmatrix}, \quad \mathbf{H}_v = \begin{bmatrix} \mathbf{0} \\ \mathbf{C} \\ \mathbf{0} \end{bmatrix},$$

459 where  $\mathbf{C} = [\mathbf{c}_1, \dots, \mathbf{c}_{|\mathcal{I}|}]$  and  $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_{|\mathcal{I}|}]$ . Therefore,

$$460 \quad \mathbf{H}_t^\top \mathbf{H}_v = \mathbf{C}^\top \mathbf{0} + \mathbf{0}^\top \mathbf{C} + \mathbf{U}^\top \mathbf{0} = \mathbf{0},$$

461 hence  $\mathcal{L}_{\text{orth}} = \|\mathbf{H}_t^\top \mathbf{H}_v\|_F^2 = 0$ .

462 **InfoNCE can be made arbitrarily small.** By (17), the contrastive pair for item  $i$  is  $(\mathbf{c}_i, \mathbf{c}_i)$ . Choose  $\{\mathbf{c}_i\}_{i \in \mathcal{I}}$  to be  
 463 (approximately) orthonormal in  $\mathbb{R}^{d_c}$  with  $d_c$  sufficiently large, and take  $f(\mathbf{a}, \mathbf{b}) = \langle \mathbf{a}, \mathbf{b} \rangle$ . Then  $f(\mathbf{c}_i, \mathbf{c}_i) = 1$  and  
 464  $f(\mathbf{c}_i, \mathbf{c}_j) \approx 0$  for  $j \neq i$ , implying the InfoNCE denominator is dominated by the positive term. As  $d_c$  increases (or  
 465 equivalently by increasing separation among  $\{\mathbf{c}_i\}$ ),  $\mathcal{L}_{\text{InfoNCE}}$  can be driven below any prescribed  $\varepsilon > 0$ .

466 **Fused BPR can be small while ignoring modality-unique evidence.** Let the fusion module ignore the private channel  
 467  $\mathcal{S}_p$ :

$$468 \quad \mathbf{h}_f^i = \phi_f(\mathbf{h}_t^i, \mathbf{h}_v^i) = \begin{bmatrix} \mathbf{c}_i \\ \mathbf{c}_i \\ \mathbf{0} \end{bmatrix}. \quad (18)$$

469 Choose user embeddings  $\mathbf{e}_u = [\mathbf{w}_u; \mathbf{w}_u; \mathbf{0}]$  so that  $s_t(u, i) = \langle \mathbf{w}_u, \mathbf{c}_i \rangle$  and  $s_v(u, i) = \langle \mathbf{w}_u, \mathbf{c}_i \rangle$ , and  $s_f(u, i) = 2\langle \mathbf{w}_u, \mathbf{c}_i \rangle$ .  
 470 Hence BPR training reduces to learning  $(\mathbf{w}_u, \mathbf{c}_i)$  to separate positives from sampled negatives in the shared factor space.

471 Now consider  $S = \mathcal{A}_v(u, i^+)$ . By Assumption ??,  $p = q(S \mid u) \leq \rho_A$  for a non-negligible fraction of  $(u, i^+)$ . Applying  
 472 Lemma ??, the contribution of constraints on  $S$  to the sampled BPR objective is at most  $pM \leq \rho_A M$ . Therefore, by  
 473 choosing  $(\mathbf{w}_u, \mathbf{c}_i)$  to yield arbitrarily small loss on the complement  $\mathcal{I} \setminus (\mathcal{O}_u \cup S)$ , we obtain  $\mathcal{L}_{\text{BPR}} \leq \varepsilon + \rho_A M$ .

---

495 **Failure on unimodal indistinguishability.** By Definition ??, negatives in  $\mathcal{A}_v(u, i^+)$  admit task-relevant modality-unique  
496 evidence that is not captured by the shared factor alone. Our construction makes the fused scorer and both unimodal scorers  
497 depend only on  $\mathbf{c}_i$  and completely ignore the private evidence in  $\mathbf{u}_i$ . Thus, for those ambiguous negatives, the model is  
498 not compelled by  $\mathcal{L}_{\text{BPR}} + \lambda_1 \mathcal{L}_{\text{InfoNCE}} + \lambda_2 \mathcal{L}_{\text{orth}}$  to learn the unique evidence needed for disambiguation, and unimodal  
499 indistinguishability can persist.

500 This completes the proof. □

501  
502  
503  
504  
505  
506  
507  
508  
509  
510  
511  
512  
513  
514  
515  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539  
540  
541  
542  
543  
544  
545  
546  
547  
548  
549