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

Multimodal recommendation systems integrate heterogeneous item modalities (e.g., visual and textual content) and have been shown to substantially improve personalized ranking performance (**???**). 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 observation, 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:

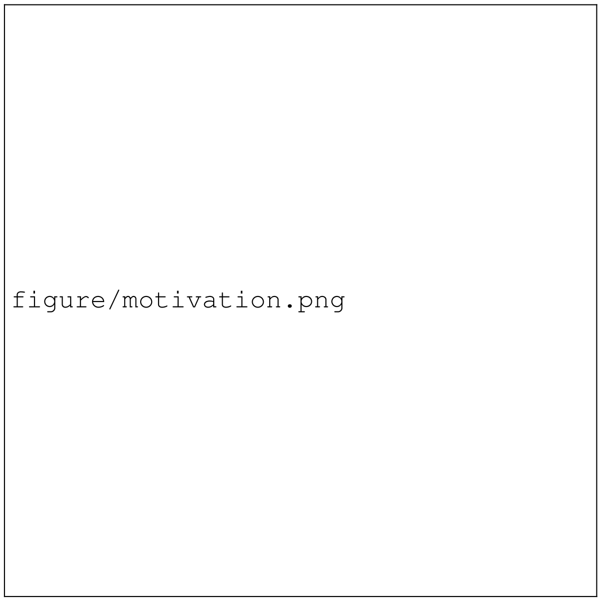

Figure 1. Illustration of two phenomena associated with modality inconsistency. **(1) Under-exploited informative inconsistency.** A user prefers a red T-shirt; two candidates are indistinguishable in text but differ in image color, yet the model misses this cross-modal cue and mis-ranks a hard negative. **(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.

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

Existing solutions tackle modality inconsistency from differ-

[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.

ent perspectives. Some studies adopt *modality-independent modeling* with independent encoders and late fusion (**??**), 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 (**?**)) and diffusion models to *explicitly enforce cross-modal consistency* (**?????**). 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 ranking, 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 underexploit 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. 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 $m \in \{t, v\}$.

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

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

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.

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

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

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. 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.

3.4. 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.

3.5. 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.

3.6. 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.

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

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%

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	–	–
✔	✔	**0.0819**	**0.0452**	–	–

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 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.

figure/OverView_1.png

Figure 2. **Overview of BR-MRS.**

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.*

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

Block-orthogonal parametrization. Let the embedding space decompose into three orthogonal subspaces $\mathbb{R}^d = \mathcal{S}_t \oplus \mathcal{S}_v \oplus \mathcal{S}_p$ 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 factor $\mathbf{u}_i \in \mathbb{R}^{d_p}$. We realize modality embeddings as

$$\mathbf{h}_t^i = \begin{bmatrix} \mathbf{c}_i \\ \mathbf{0} \\ \mathbf{u}_i \end{bmatrix}, \qquad \mathbf{h}_v^i = \begin{bmatrix} \mathbf{0} \\ \mathbf{c}_i \\ \mathbf{0} \end{bmatrix}. \tag{7}$$

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

$$\tilde{\mathbf{h}}_t^i = \mathbf{P}_t \mathbf{h}_t^i = \mathbf{c}_i, \qquad \tilde{\mathbf{h}}_v^i = \mathbf{P}_v \mathbf{h}_v^i = \mathbf{c}_i. \tag{8}$$

Orthogonality term is exactly minimized. Stacking item embeddings yields

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

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

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

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

InfoNCE can be made arbitrarily small. By (**??**), the contrastive pair for item i is $(\mathbf{c}_i, \mathbf{c}_i)$. Choose $\{\mathbf{c}_i\}_{i \in \mathcal{I}}$ to be (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 $f(\mathbf{c}_i, \mathbf{c}_j) \approx 0$ for $j \neq i$, implying the InfoNCE denominator is dominated by the positive term. As d_c increases (or equivalently by increasing separation among $\{\mathbf{c}_i\}$), $\mathcal{L}_{\text{InfoNCE}}$ can be driven below any prescribed $\varepsilon > 0$.

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

$$\mathbf{h}_f^i = \phi_f(\mathbf{h}_t^i, \mathbf{h}_v^i) = \begin{bmatrix} \mathbf{c}_i \\ \mathbf{c}_i \\ \mathbf{0} \end{bmatrix}. \tag{9}$$

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 $s_f(u, i) = 2\langle \mathbf{w}_u, \mathbf{c}_i \rangle$. Hence BPR training reduces to learning $(\mathbf{w}_u, \mathbf{c}_i)$ to separate positives from sampled negatives in the shared factor space.

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 Lemma **??**, the contribution of constraints on S to the sampled BPR objective is at most $pM \leq \rho_A M$. Therefore, by 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 $\mathcal{L}_{\text{BPR}} \leq \varepsilon + \rho_A M$.

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

This completes the proof. □