BEYOND SUPERFICIAL CUES: MITIGATING POPULARITY-AMPLIFIED PRESENTATION BIAS IN MULTIMODAL RECOMMENDATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal recommendation systems significantly enhance user experience by leveraging rich multimodal features. However, their effectiveness is often undermined by presentation bias, a prevalent issue where user interactions are disproportionately influenced by superficial presentation cues (e.g., flashy thumbnails, sensational titles) rather than intrinsic content. Such cues are often highly correlated with item popularity, leading to a popularity-amplified presentation bias that compromises recommendation fairness. To address this fundamental challenge, this paper introduces a novel Intrinsic **CO**ntent **RE**presentation (**CORE**) framework. We formalize presentation bias within a Structural Causal Model (SCM), revealing that conditioning on fully coupled multimodal features acts as a collider, inducing spurious correlations between intrinsic content and presentation manner. CORE mitigates this collider bias by disentangling an item's multimodal representation into two independent components: a content-aware representation and a bias-aware representation. This is achieved through a carefully designed automated contrastive learning scheme and enforced by an independence constraint. By performing recommendations based solely on the purified intrinsic content representation, our model effectively performs a causal intervention to mitigate the spurious effects. Extensive experiments on three real-world multimodal datasets demonstrate that CORE achieves significant and consistent improvements in both recommendation accuracy and fairness.

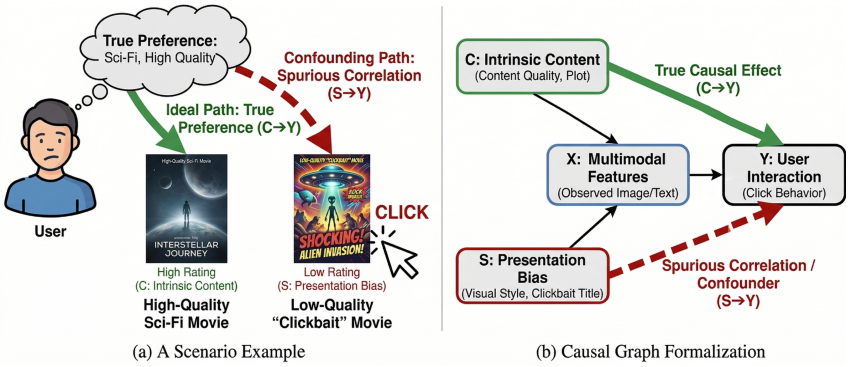
1 INTRODUCTION

With the explosive growth of internet content, multimodal recommendation systems have become a critical bridge connecting users and information, widely deployed in e-commerce platforms like Amazon, streaming services like MovieLens, and short-video applications like TikTok Xu et al. (2025a). By integrating diverse modal information such as images, videos, and text, these systems construct richer user and item profiles to deliver more precise personalized recommendations Wei et al. (2023).

However, contemporary multimodal recommendation models, in their pursuit of optimizing click-through rates, often fall into a critical pitfall: Presentation Bias. As illustrated in Figure 1, a user's click decision is a complex process influenced by both the item's Intrinsic Content (C) (e.g., a movie's plot, a product's quality) and its Presentation Manner (S) (e.g., a movie's promotional poster, a product's professionally edited image, a short video's background music). An item with mediocre content but a highly attractive cover may receive far more clicks than a high-quality item with a plain presentation. In real-world systems, these superficial presentation patterns are often tightly coupled with resource allocation, making them highly correlated with an item's cross-category popularity. Existing multimodal models typically learn from observed features (X) where intrinsic content (C) and presentation manner (S) are entangled. They fail to distinguish whether a user's click (Y) stems from genuine interest or the allure of popularity-driven presentation cues.

This entanglement leads to two severe problems: (1) Inaccurate Recommendations: The model learns spurious correlations Wei et al. (2021); Xu et al. (2026) instead of users' true preferences, diminishing the long-term user experience Wang et al. (2023b). (2) Unfair Recommendations:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) A Scenario Example (b) Causal Graph Formalization

Figure 1: Causal graph of presentation bias in multimodal recommendation. User click behavior (Y) is influenced by both intrinsic content (C) and presentation manner (S). C and S learned from the coupled multimodal feature (X) conditions on a collider. This opens the path $C \rightarrow X \leftarrow S$, creating a spurious correlation $C \leftrightarrow S \rightarrow Y$ that contaminates the model's learning of the true causal effect $C \rightarrow Y$.

High-quality but plainly presented "long-tail" content is denied fair exposure opportunities Li et al. (2023c); Said & Bellogín (2014), while over-packaged "popular" content dominates, harming the diversity of the content ecosystem Abdollahpouri et al. (2019); Wen et al. (2024).

To fundamentally address this long-standing challenge, we advocate for modeling it from a rigorous causal inference perspective Pearl (2009); Robins (1986). Instead of treating presentation bias as a completely new phenomenon, we are the first to formulate it mathematically within a causal framework. We posit that the unified multimodal feature (X) acts as a collider between the intrinsic content (C) and presentation manner (S). According to causal theory Pearl (2009), while C and S are inherently independent, conditioning on the observed X (as standard models do) opens the collider path $C \rightarrow X \leftarrow S$, creating a spurious correlation ($C \leftrightarrow S \rightarrow Y$). This entanglement contaminates the model's learning process, preventing it from capturing the invariant causal effect of content on user preference. Our goal is to break this spurious dependency to achieve debiased recommendations Li & Sui (2025); Xiao et al. (2023).

To this end, we propose the Intrinsic Content Representation (CORE) framework, a novel approach to causal disentanglement. The core idea is to disentangle an item's unified multimodal representation into two orthogonal latent variables: an intrinsic content representation that captures its core value, and a presentation bias representation that captures its superficial appeal. It is crucial to note that CORE does not rely on a rigid assumption that certain modalities (e.g., text) exclusively represent content while others (e.g., image) represent bias. As we will detail, modalities are first deeply fused, and the disentanglement operates on the unified representation space, utilizing cross-modal similarities merely as a weak supervision heuristic. During prediction, we leverage only the purified content representation, thereby performing a structural causal intervention (i.e., Pearl's do-calculus) to eliminate presentation bias Shang et al. (2024); Qiu et al. (2021). Our main contributions are summarized as follows: 1) We systematically define and analyze the "Presentation Bias" problem in multimodal recommendation from a causal inference perspective. 2) We propose the CORE framework, featuring a novel causal disentanglement module that effectively separates an item's intrinsic content from its presentation bias via an automated, scalable contrastive learning strategy. 3) We conduct extensive experiments on three real-world multimodal datasets, demonstrating that CORE significantly outperforms state-of-the-art baselines in both recommendation accuracy and fairness metrics.

2 Related Work

2.1 Multimodal Recommendation

Multimodal recommendation systems seek to enhance recommendation performance by integrating diverse information sources, such as text, visuals, and audio. Pioneering works, such as VBPR He &

McAuley (2016), integrated visual features extracted from Convolutional Neural Networks (CNNs) into the classic matrix factorization framework, demonstrating the significance of visual information for capturing user preferences. With the advancement of deep learning, more sophisticated fusion models have emerged. Models based on Graph Neural Networks (GNNs) Wei et al. (2019); Tan et al. (2024) have become a dominant paradigm. For instance, MMGCN Wei et al. (2019) constructs separate graphs for different modalities and learns unified user and item representations through a carefully designed cross-modal message-passing mechanism. To capture more complex inter-item relationships, HMRec Tan et al. (2024) further introduced a heterogeneous hypergraph structure, thereby enabling the modeling of high-order relations beyond pairwise interactions. Beyond GNNs, the Transformer architecture Zhou et al. (2023) has also been widely adopted for its powerful sequence modeling and feature interaction capabilities. For example, some works leverage self-attention mechanisms to dynamically weigh the contributions of different modalities to a user's final decision Sun et al. (2018); Avas et al. (2024), while others employ cross-modal attention to learn deeply fused inter-modal representations Zhang et al. (2021), and enhance the consistency of the representation through contrastive learning Xia et al. (2022). Recently, researchers have started exploring causal inference specifically for multimodal scenarios Chen et al. (2025); Li et al. (2023b); Xu et al. (2025b). However, these works largely focus on modality missingness or visual popularity. These pioneering works highlight an emerging trend, yet the specific problem of disentangling intrinsic content from superficial presentation style remains under-explored.

2.2 Debiasing in Recommendation

Bias in recommendation systems is a widely studied topic, with major areas including selection bias Hernán et al. (2004) (where users tend to rate items they like), exposure bias Joachims et al. (2017) (where users can only interact with recommended items), and popularity bias Chen et al. (2023) (where popular items are over-recommended). Existing debiasing methods can be broadly categorized into three groups. The first category comprises methods based on Inverse Propensity Scoring (IPS) Schnabel et al. (2016). These approaches correct for biases arising from selective data observation by first estimating the probability of an item being observed (i.e., the propensity score) and then using its inverse to weight the loss of each training sample. However, IPS methods are highly sensitive to the accuracy of propensity score estimation, as inaccuracies can amplify errors rather than mitigate them. The second category includes methods based on data augmentation or adversarial learning. For instance, some works generate counterfactual samples to augment the training data, aiming to balance the distribution of items across different groups Wang et al. (2024). Adversarial learning, in contrast, introduces a discriminator to compel the learned user or item representations to be indistinguishable with respect to sensitive attributes (e.g., popularity), thereby achieving debiasing He et al. (2018). The third category is grounded in causal inference Zhao et al. (2024); Zhang et al. (2024); Li et al. (2023a). These methods leverage causal graphs to explicitly model the generative process of bias, treating bias as a confounder. By employing causal tools such as the backdoor criterion, instrumental variables, or counterfactual reasoning, these approaches aim to estimate the unbiased causal effect from observational data. For example, GCDF Zhao et al. (2024) treats noisy interactions as confounders and performs causal denoising. Other works Zhang et al. (2024); Wang et al. (2025a) attempt to answer counterfactual questions such as, "Would the user still have clicked this item if it were not popular?" Our work is deeply inspired by the principles of causal inference. However, we apply them to a novel and distinct problem specific to the multimodal context: disentangling the spurious correlations that arises from an item's intrinsic content versus its external presentation. This marks a significant departure from previous work, which has predominantly focused on popularity or selection bias.

3 Preliminaries

3.1 Problem Formulation

Given a set of users $\mathcal{U}$ and a set of items $\mathcal{I}$, we have a user-item interaction matrix $R \in \{0,1\}^{|\mathcal{U}|\times|\mathcal{I}|}$, where $R_{ui} = 1$ indicates that user u has interacted with item i. Each item i is associated with multimodal features, such as visual features v_i and textual features t_i. Our goal is to learn a function $f(\cdot)$ that predicts a user u's preference score $\hat{y}_{ui} = f(u, i, v_i, t_i)$ for an unobserved item i, and to generate a debiased recommendation list based on these scores.

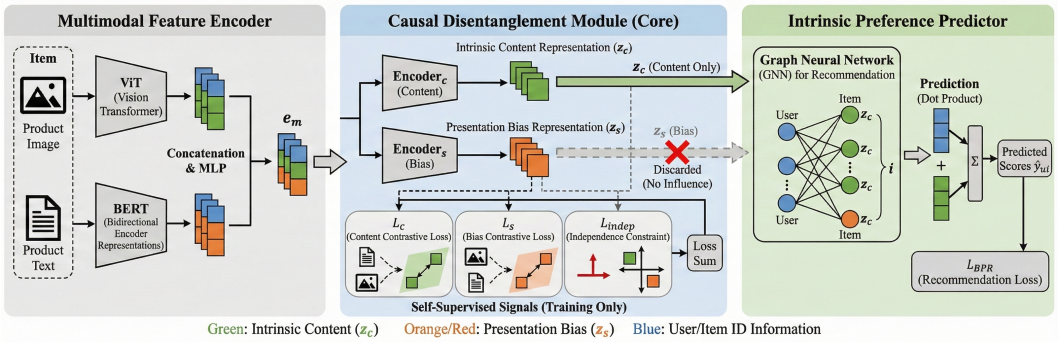


Figure 2: The overview of the proposed CORE framework. On the left side, the Multimodal Feature Encoder extracts and fuses visual and textual modalities into an initial coupled representation. In the middle of the diagram lies the core Causal Disentanglement Module. On the right side, the Intrinsic Preference Predictor performs a causal intervention by exclusively utilizing the purified content representation within a GNN, actively discarding the confounding bias signal.

3.2 Causal Formulation of Presentation Bias

We formalize the problem using a SCM Pearl (2009) based on the causal graph in Figure 1. Let C, S, X, and Y denote the latent intrinsic content, latent presentation manner, observed multimodal features, and user click, respectively. The data generation process is described by the following structural equations:

- $C \sim P(C), S \sim P(S)$ (Content and presentation are independent exogenous factors)
- $X = f_X(C, S, \epsilon_X)$ (Observed multimodal features are generated from both C and S)
- $Y = f_Y(C, S, \epsilon_Y)$ (User clicks are caused by C and S)

where f_X and f_Y are unknown functions and ϵ represents stochastic noise. In standard recommendation models, we learn the conditional probability $P(Y|X)$. However, since X is a collider ($C \rightarrow X \leftarrow S$), observing X renders the exogenous variables C and S statistically dependent ($C \not\perp S|X$). Because presentation cues strongly drive interactions ($S \rightarrow Y$), this induces a spurious correlation path $C \leftrightarrow S \rightarrow Y$. Consequently, the learned association $P(Y|X)$ is a biased estimation of the invariant causal mechanism $C \rightarrow Y$.

To mitigate this collider bias, our strategy is to recover the independent latent mechanisms. We aim to learn disentangled representations, z_c and z_s, acting as proxies for C and S. Crucially, by enforcing statistical independence between z_c and z_s (via HSIC constraints during training), we structurally break the spurious path $C \leftrightarrow S$. While strictly identifying $P(Y|do(C))$ requires strong parametric assumptions (which are often intractable in deep multimodal scenarios), utilizing only the purified z_c for prediction serves as a robust proxy-based causal disentanglement. It effectively filters out the presentation-induced variations, approximating an invariant predictor that relies solely on intrinsic content. Further discussions on causal identifiability limits are provided in Appendix A.

4 The Proposed Method

This section details the proposed CORE framework. We begin with an overview of the framework. We then provide a causal analysis of the problem, followed by a detailed description of the causal disentanglement module and the intrinsic preference predictor. Finally, we present the model's optimization objective.

4.1 Framework Overview

The overall architecture of the CORE framework is illustrated in Figure 2. It consists of three core components: (1) Multimodal Feature Encoder: Extracts initial features from raw multimodal data. (2) Causal Disentanglement Module: Disentangles the fused multimodal representation into an

intrinsic content representation $\mathbf{z}_c$ and a presentation bias representation $\mathbf{z}_s$. (3) Intrinsic Preference Predictor: Utilizes only the disentangled content representation $\mathbf{z}_c$ for recommendation, effectively blocking the bias pathway.

4.2 Multimodal Feature Encoder

We employ pre-trained models to extract initial modal features for items. For the visual modality, we use a Vision Transformer (ViT) Dosovitskiy (2020) to encode item images, yielding a visual embedding $\mathbf{e}_v \in \mathbb{R}^d$. For the textual modality, we use BERT (Bidirectional Encoder Representations from Transformers) Devlin et al. (2019) to encode item titles or descriptions, obtaining a textual embedding $\mathbf{e}_t \in \mathbb{R}^d$. These embeddings are then concatenated and fused through a multi-layer perceptron (MLP) to produce the initial multimodal representation $\mathbf{e}_m$:

$$\mathbf{e}_m = \text{MLP}(\text{Concat}(\mathbf{e}_v, \mathbf{e}_t)) \tag{1}$$

4.3 Causal Disentanglement Module

This module is the core of the CORE framework, designed to project the fused multimodal representation $\mathbf{e}_m$ into two independent latent spaces. We use two separate encoders (implemented as MLPs) to generate the content and bias representations:

$$\mathbf{z}_c = \text{Encoder}_c(\mathbf{e}_m), \quad \mathbf{z}_s = \text{Encoder}_s(\mathbf{e}_m) \tag{2}$$

To achieve effective disentanglement, we design two key self-supervised tasks:

4.3.1 Contrastive Disentanglement Learning

We shape the semantics of the content and bias representation spaces using contrastive learning with carefully constructed positive and negative pairs. The core challenge is providing supervision to shape the semantics of $\mathbf{z}_c$ and $\mathbf{z}_s$. We devise a fully automated, scalable strategy to construct positive pairs, utilizing pre-trained embedding similarities as a weak supervision heuristic. The detailed procedure is outlined in Algorithm 1 in **Appendix D**.

Content Contrastive Learning: The objective is to pull together items with similar content but different presentations, while pushing apart content-dissimilar items. For an anchor item i, we construct its positive sample i^+ by preserving its textual description (representing content) but replacing its image with one from another item of the same category that has a distinct visual style (e.g., poster design, color palette). Thus, items i and i^+ share similar intrinsic content but differ in presentation. The content contrastive loss $\mathcal{L}_c$ is defined as the InfoNCE loss:

$$\mathcal{L}_c = -\sum_{i \in \mathcal{B}} \log \frac{\exp(\text{sim}(\mathbf{z}_c^i, \mathbf{z}_c^{i+})/\tau)}{\sum_{j \in \mathcal{B}} \exp(\text{sim}(\mathbf{z}_c^i, \mathbf{z}_c^j)/\tau)} \tag{3}$$

where $\mathcal{B}$ is a batch of items, $\text{sim}(\cdot, \cdot)$ is the cosine similarity, and τ is a temperature hyperparameter.

Bias Contrastive Learning: The objective here is to group items with similar presentation styles but different content. For an anchor item i, we select a positive sample $i^{+\prime}$ that shares a similar popularity or comes from the same highly popular creator but belongs to a different content category. This strategy compels the bias representation $\mathbf{z}_s$ to capture content-agnostic signals like popularity and creator effects. The bias contrastive loss $\mathcal{L}_s$ is formulated similarly to $\mathcal{L}_c$.

4.3.2 Independence Constraint

To ensure that $\mathbf{z}_c$ and $\mathbf{z}_s$ capture mutually independent information, we employ the Hilbert-Schmidt Independence Criterion (HSIC) Gretton et al. (2007) as a regularization term. This minimizes the statistical dependence between the set of content representations Z_c and the set of bias representations Z_s within a batch:

$$\mathcal{L}_{\text{indep}} = \text{HSIC}(Z_c, Z_s) \tag{4}$$

A lower HSIC value indicates greater independence between the two sets of variables.

4.4 Intrinsic Preference Predictor

After obtaining the disentangled item content representation $\mathbf{z}_c$, we use it to make unbiased recommendations. The user representation $\mathbf{u}$ is learned based on the intrinsic content of items in their interaction history. We adapt the classic LightGCN He et al. (2020) architecture, replacing the traditional ID embeddings with our intrinsic content representations $\mathbf{z}_c$ as the objects of message passing. After L layers of graph convolution, we obtain the final user and item intrinsic content representations, $\mathbf{u}^{(L)}$ and $\mathbf{z}_c^{(L)}$. The preference score of user u for item i is defined as:

$$\hat{y}_{ui} = (\mathbf{u}^{(L)})^T \mathbf{z}_{c,i}^{(L)} \tag{5}$$

Crucially, the entire recommendation prediction process is independent of the bias representation $\mathbf{z}_s$, thereby structurally blocking the influence of presentation bias.

4.5 Model Optimization

The total optimization objective consists of three parts: the main recommendation task loss, the contrastive disentanglement loss, and the independence constraint loss. We adopt the Bayesian Personalized Ranking (BPR) loss Rendle et al. (2012) as the main recommendation loss, $\mathcal{L}_{\text{BPR}}$:

$$\mathcal{L}_{\text{BPR}} = \sum_{(u,i,j)\in D_s} -\log \sigma(\hat{y}_{ui} - \hat{y}_{uj}) \tag{6}$$

where D_s is the training set of triplets containing a user, a positive item, and a negative item.

The final loss function is:

$$\mathcal{L} = \mathcal{L}_{\text{BPR}} + \alpha(\mathcal{L}_c + \mathcal{L}_s) + \beta\mathcal{L}_{\text{indep}} \tag{7}$$

where α and β are hyperparameters that control the strength of the disentanglement and independence constraints.

5 Experiments

In this section, we conduct a series of experiments to answer the following research questions:

RQ1 How does CORE perform in terms of accuracy and fairness compared to SOTA baselines?

RQ2 What are the contributions of the key components within CORE?

RQ3 How sensitive is CORE to the hyperparameters controlling the disentanglement?

RQ4 Can CORE effectively learn disentangled representations for intrinsic content and bias?

5.1 Experimental Setup

5.1.1 Datasets

We conduct our experiments on three widely used public multimodal datasets, Amazon-Beauty Ni et al. (2019), Amazon-Fashion Ni et al. (2019), and MovieLens-1M (ML-1M) Harper & Konstan (2015), all of which contain rich user interaction records and multimodal features. Since the original ML-1M dataset lacks visual modalities, we augmented it with movie posters; please refer to **Appendix C.2** for details on image gathering and preprocessing. The detailed statistics of the preprocessed datasets are summarized in Table 3 in **Appendix C**. Following standard procedures from previous works Wei et al. (2019); Dosovitskiy (2020), we convert ratings (or implicit interactions) into binary labels (ratings 4-5 are positive, others are negative). For all datasets, we apply a 5-core setting (i.e., removing users and items with fewer than 5 interactions) and split the data into training, validation, and test sets in an 8:1:1 ratio based on timestamps.

5.1.2 Baselines

For a comprehensive performance evaluation, we compare CORE against three categories of representative baselines. (1) ID-based methods, including NeuMF He et al. (2017) and LightGCN He

Table 1: Overall performance comparison on three datasets. The best results are highlighted in **bold**, and the second-best are underlined. ↑ Indicates higher is better, while ↓ indicates lower is fairer.

Datasets	Metrics	ID-based Methods		Multimodal Methods			Debiasing Methods				CORE (Ours)	Improve
		NeuMF	LightGCN	VBPR	MMGCN	HMRec	IPS-MMGCN	GCDF	PPAC	CDRSB		
Amazon Beauty	R@20 ↑	0.0630	0.0651	0.0684	0.0752	0.0812	0.0730	0.0795	0.0805	0.0825	**0.0870**	+5.45%
	N@20 ↑	0.0341	0.0353	0.0381	0.0423	0.0455	0.0409	0.0441	0.0450	0.0463	**0.0487**	+5.18%
	C@20 ↑	0.0502	0.0521	0.0498	0.0617	0.0695	0.0753	0.0788	0.0815	0.0831	**0.0854**	+2.77%
	Gini ↓	0.468	0.453	0.461	0.442	0.425	0.391	0.380	0.375	0.369	**0.351**	-4.88%
ML-1M	R@20 ↑	0.1805	0.1835	0.1891	0.2014	0.2088	0.1995	0.2053	0.2071	0.2113	**0.2198**	+4.02%
	N@20 ↑	0.1431	0.1452	0.1508	0.1625	0.1693	0.1601	0.1664	0.1680	0.1712	**0.1783**	+4.15%
	C@20 ↑	0.1015	0.1033	0.0989	0.1156	0.1240	0.1211	0.1235	0.1298	0.1325	**0.1371**	+3.47%
	Gini ↓	0.625	0.612	0.623	0.589	0.551	0.535	0.519	0.505	0.498	**0.481**	-3.41%
Amazon Fashion	R@20 ↑	0.0715	0.0742	0.0788	0.0865	0.0921	0.0844	0.0903	0.0915	0.0938	**0.0982**	+4.69%
	N@20 ↑	0.0388	0.0401	0.0425	0.0471	0.0508	0.0460	0.0495	0.0503	0.0517	**0.0543**	+5.03%
	C@20 ↑	0.0553	0.0581	0.0560	0.0688	0.0754	0.0811	0.0845	0.0886	0.0902	**0.0935**	+3.66%
	Gini ↓	0.482	0.470	0.479	0.451	0.433	0.402	0.389	0.381	0.370	**0.352**	-4.86%

et al. (2020). (2) Multimodal methods, including VBPR He & McAuley (2016), MMGCN Wei et al. (2019) and HMRec Tan et al. (2024). (3) Debiasing methods, including IPS-MMGCN Schnabel et al. (2016), GCDF Zhao et al. (2024), PPAC Ning et al. (2024) and CDRSB Wang et al. (2025b). The detailed descriptions of all baselines are provided in **Appendix B**.

5.1.3 Evaluation Metrics

We employ two categories of metrics for a comprehensive evaluation. We use the widely accepted Recall@20 and NDCG@20 to measure the accuracy of the top-K recommendation lists. We use Coverage@20 (the percentage of unique items in recommendation lists relative to the total item pool; higher is better) and the Gini Index (measuring the inequality of recommended item popularity distribution; lower is fairer) to evaluate the model's debiasing effectiveness and diversity.

5.1.4 Implementation Details

All models were implemented using the PyTorch framework and the RecBole library Zhao et al. (2022). For multimodal features, we used a pre-trained ViT-B/32 to extract visual features and BERT-base for textual features. The embedding dimension was set to 64 for all models. We used the Adam optimizer with a learning rate searched in 1e-3, 5e-4, 1e-4. The batch size was fixed at 2048. For CORE, the number of graph convolution layers L was set to 2, and the temperature τ was 0.1. Hyperparameters α and β were tuned on the validation set via a grid search over [0.01, 0.1, 1.0]. All reported results are the average of three independent runs. Additional implementation details are elaborated in **Appendix C**.

5.2 Overall Performance Comparison (RQ1)

Table 1 presents the overall performance comparison of CORE against all baseline models on the three datasets. CORE achieves the best performance across all metrics, providing strong evidence for the effectiveness of our proposed causal disentanglement framework. By explicitly separating intrinsic content from presentation bias, CORE not only enhances recommendation accuracy but also significantly improves fairness. The results lead to the following key observations:

In Terms of Accuracy. As shown in Table 1, CORE consistently achieves the best performance in Recall@20 and NDCG@20 across all three datasets. This demonstrates that by isolating and removing the interference from presentation bias, our model more accurately captures users intrinsic preferences. Notably, CORE outperforms not only standard multimodal models like HMRec but also sophisticated debiasing models. For instance, compared to the strongest debiasing baseline CDRSB, CORE achieves a relative NDCG@20 improvement of 5.18% on Amazon-Beauty, 4.15% on ML-1M, and 5.03% on Amazon-Fashion. This suggests that our specific formulation of presentation bias and the targeted causal disentanglement strategy are more effective for this problem than general-purpose debiasing frameworks.

In Terms of Fairness. CORE's advantages in fairness are particularly pronounced. It consistently achieves the highest Cov@20 and the lowest Gini Index, indicating more diverse and fair recom-

mendations. On the ML-1M dataset, for example, it reduces the Gini Index by 3.41% compared to the best debiasing baseline CDRSB, while simultaneously increasing Coverage. This indicates that CORE successfully provides exposure to a wider range of long-tail items, effectively mitigating the popularity bias exacerbated by presentation cues.

Improvement over Debiasing Baselines. CORE demonstrates clear advantages over all dedicated debiasing methods. While models like IPS and PPAC address popularity bias from statistical and adversarial perspectives, respectively, they do not explicitly model the confounding nature of presentation in multimodal contexts. Even compared to CDRSB, which also employs causal disentanglement, CORE's superiority highlights the effectiveness of our novel contrastive learning strategy tailored for separating intrinsic content from presentation style. By proactively removing the source of bias at the representation level through a principled causal intervention, our method is more targeted and fundamental in addressing the specific challenge of "Presentation Bias."

5.3 Ablation Study (RQ2)

To investigate the specific contribution of each proposed component within CORE and validate our causal disentanglement strategy (RQ2), we conduct a comprehensive ablation study. We compare the full CORE model against several variants on the Amazon-Beauty dataset, each created by removing or simplifying a key module. The variants are defined as follows:

- **w/o-Dis:** Removes the entire Causal Disentanglement Module.
- **w/o-CL:** Removes both the content and bias contrastive learning losses ($\mathcal{L}_c$ and $\mathcal{L}_s$).
- **w/o-HSIC:** Removes the HSIC independence constraint ($\mathcal{L}_{indep}$).

The results in Table 2 clearly demonstrate the contribution of each component. Removing the entire disentanglement module (CORE-w/o-Dis) leads to a significant performance drop, especially on fairness metrics, with the highest Gini Index and lowest Coverage. This confirms the necessity of causal disentanglement. Both the contrastive learning loss (CL) and the independence constraint (HSIC) are crucial for achieving effective representation separation. The removal of contrastive learning (CORE-w/o-CL) results in the largest performance degradation, underscoring that our designed contrastive tasks are the core driving force for shaping the semantics of the content and bias spaces. The independence constraint (CORE-w/o-HSIC) plays a key regularization role, ensuring the orthogonality of the two latent spaces and preventing information leakage. Together, these components work synergistically to ensure CORE's performance.

Table 2: Ablation study results on the Amazon-Beauty dataset.

Variant	R@20 ↑	N@20 ↑	C@20 ↑	Gini ↓
w/o-Dis	0.0765 (-12.1%)	0.0428 (-12.1%)	0.0583 (-31.7%)	0.435 (+23.9%)
w/o-CL	0.0791 (-9.1%)	0.0440 (-9.7%)	0.0652 (-23.7%)	0.410 (+16.8%)
w/o-HSIC	0.0845 (-2.9%)	0.0471 (-3.3%)	0.0793 (-7.1%)	0.368 (+4.8%)
CORE	**0.0870**	**0.0487**	**0.0854**	**0.351**

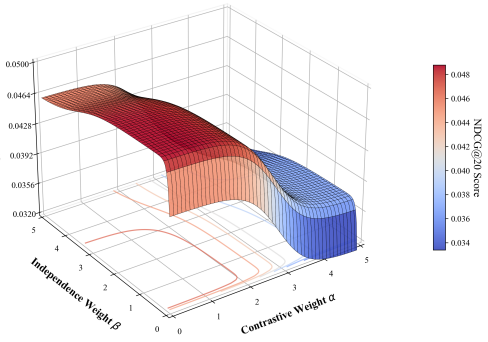
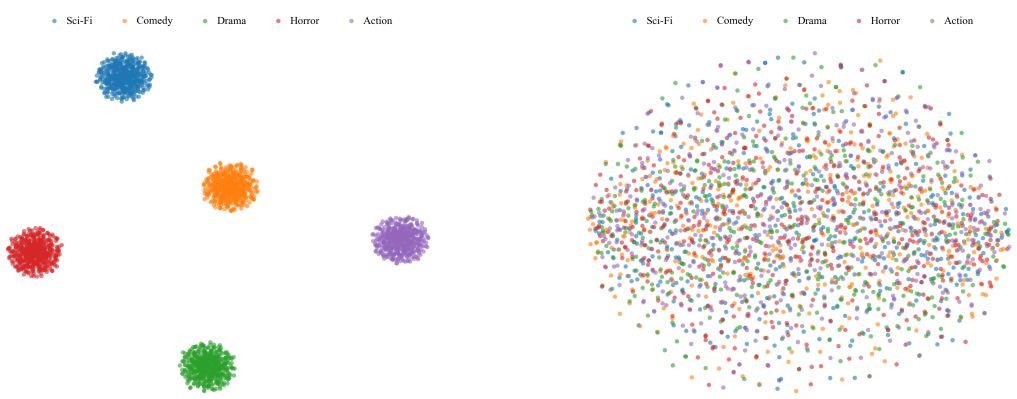


Figure 3: Hyperparameter sensitivity of α, β.

5.4 Hyperparameter Sensitivity Analysis (RQ3)

We conduct a sensitivity analysis on the Amazon-Beauty dataset for the two key hyperparameters: α (contrastive loss weight) and β (HSIC independence constraint weight). We vary both parameters in the range of $[0.01, 5.0]$. The resulting NDCG@20 performance is plotted as a 3D surface in Figure 3.

As illustrated in Figure 3, CORE demonstrates high stability. Within a reasonably wide operational range (e.g., $\alpha \in [0.05, 0.5]$ and $\beta \in [0.1, 1.0]$), the performance remains at a high plateau. Extreme performance degradation only occurs at absolute boundaries. For instance, when $\alpha > 2.0$, the



(a) Intrinsic Content ($\mathbf{z}_c$) (b) Presentation Bias ($\mathbf{z}_s$)

Figure 4: The t-SNE visualization of the learned (a) intrinsic content representations and (b) presentation bias representations on ML-1M dataset. Colors denote different movie genres.

contrastive objective overshadows the primary BPR recommendation loss. Conversely, when $\beta \to 0.01$, the independence constraint is too weak, causing the representations to entangle. This analysis validates that CORE is robust and does not rely on narrowly tuned settings.

5.5 Visualization Analysis (RQ4)

We specifically select the ML-1M dataset for visualization analysis because, unlike e-commerce items, movies possess highly unambiguous and well-defined content labels (i.e., Genres). This provides a clear, objective ground truth to verify whether our disentangled representation $\mathbf{z}_c$ has truly captured the intrinsic semantic content. To visually inspect the disentanglement effect, we used t-SNE to project the learned item intrinsic content representations $\mathbf{z}_c$ and presentation bias representations $\mathbf{z}_s$ from the MovieLens dataset into a 2D space, as shown in Figure 4.

In the intrinsic content space (a), movies of the same genre form distinct clusters. This indicates that $\mathbf{z}_c$ successfully captures content-related semantic information. In the presentation bias space (b), movies of different genres are intermingled, showing no clear structure based on content. Instead, we observe that items with similar popularity levels (not explicitly labeled) tend to be closer in this space. This suggests that $\mathbf{z}_s$ has captured content-agnostic bias signals. This visualization provides compelling evidence that our CORE framework successfully disentangles intrinsic content and superficial biases into separate representation spaces.

6 Conclusion

In this paper, we addressed the critical challenge of presentation bias contaminating multimodal representations in recommendation, arising from the collider bias and spurious correlation between intrinsic content and superficial presentation manner. We introduced CORE, a novel framework leveraging the principles of causal inference through a structural disentanglement strategy. By decomposing multimodal representations into independent content-aware and bias-aware components via an automated contrastive learning scheme and an independence constraint, CORE effectively blocks the spurious dependency path. By performing recommendation as a causal intervention using only the purified content representation, our framework achieves more accurate and fair user preference modeling. Extensive experiments on three benchmark datasets demonstrate the substantial superiority of our approach over state-of-the-art debiasing and disentanglement methods. Future work could involve extending this framework to sequential recommendation scenarios and exploring more advanced disentanglement techniques and counterfactual reasoning methods to further enhance the robustness and fairness of recommendation systems.

AI USE STATEMENT

In this work, we used generative AI tools solely for text polishing and grammatical correction to improve the readability of the manuscript. We have not used generative AI tools for data generation, writing code, designing experiments, or generating scientific ideas, and the rest of the required disclosure tasks are not applicable to this work. We have carefully reviewed all AI-assisted language revisions to ensure they accurately preserve our original scientific intent, methodologies, and findings. We take full responsibility for the final content of this work, including text, claims, and all artifacts presented in this paper.

ETHICS STATEMENT

This paper proposes the CORE framework to mitigate popularity-amplified presentation bias in multimodal recommendation systems. The core objective of our research explicitly aligns with ethical AI principles, as it aims to improve recommendation fairness and diversity. By disentangling intrinsic content from superficial presentation cues, our model ensures equitable exposure for high-quality "long-tail" items and prevents over-packaged popular content from dominating the ecosystem. All experiments were conducted using publicly available, anonymized datasets (Amazon-Beauty, Amazon-Fashion, and MovieLens-1M) that do not involve sensitive personally identifiable information (PII) or direct human subject research. We foresee no potentially harmful insights, methodologies, or negative societal impacts arising from our findings. There are no conflicts of interest to declare.

REPRODUCIBILITY STATEMENT

We have made concerted efforts to ensure the full reproducibility of the proposed CORE framework and our experimental results. In Section 5.1.4 of the main text, we provide a clear description of the evaluation metrics, baseline setups, and key hyperparameter configurations. To supply further necessary details without cluttering the main paper, Appendix C reports the comprehensive statistics of the preprocessed datasets along with additional implementation configurations. Furthermore, Appendix D details the precise step-by-step procedures for our automated contrastive pair construction and provides a rigorous complexity analysis. To support hands-on verification, the complete source code has been included in the supplementary materials and will be made publicly available upon publication.

REFERENCES

Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. The unfairness of popularity bias in recommendation. *arXiv preprint arXiv:1907.13286*, 2019.

Ignacio Avas, Liesbeth Allein, Katrien Laenen, and Marie-Francine Moens. Align macridvae: multimodal alignment for disentangled recommendations. In *European Conference on Information Retrieval*, pp. 73–89. Springer, 2024.

Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems*, 41(3):1–39, 2023.

Weixin Chen, Li Chen, Yongxin Ni, and Yuhan Zhao. Causality-inspired fair representation learning for multimodal recommendation. *ACM Transactions on Information Systems*, 43(6):1–29, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. *Advances in neural information processing systems*, 20, 2007.

F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

Ruining He and Julian McAuley. Vbpr: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pp. 173–182, 2017.

Xiangnan He, Zhankui He, Xiaoyu Du, and Tat-Seng Chua. Adversarial personalized ranking for recommendation. In *The 41st International ACM SIGIR conference on research & development in information retrieval*, pp. 355–364, 2018.

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pp. 639–648, 2020.

Miguel A Hernán, Sonia Hernández-Díaz, and James M Robins. A structural approach to selection bias. *Epidemiology*, 15(5):615–625, 2004.

Euiju Jeong, Xinzhe Li, Angela Kwon, Seonu Park, Qinglong Li, and Jaekyeong Kim. A multimodal recommender system using deep learning techniques combining review texts and images. *Applied Sciences*, 14(20):9206, 2024.

Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. Unbiased learning-to-rank with biased feedback. In *Proceedings of the tenth ACM international conference on web search and data mining*, pp. 781–789, 2017.

Meng Li and Haochen Sui. Causal recommendation via machine unlearning with a few unbiased data. In *AAAI 2025 Workshop on Artificial Intelligence with Causal Techniques*, volume 2, 2025.

Qian Li, Xiangmeng Wang, Zhichao Wang, and Guandong Xu. Be causal: De-biasing social network confounding in recommendation. *ACM Transactions on Knowledge Discovery from Data*, 17(1): 1–23, 2023a.

Shuaiyang Li, Dan Guo, Kang Liu, Richang Hong, and Feng Xue. Multimodal counterfactual learning network for multimedia-based recommendation. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pp. 1539–1548, 2023b.

Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, Juntao Tan, Shuchang Liu, and Yongfeng Zhang. Fairness in recommendation: Foundations, methods, and applications. *ACM Transactions on Intelligent Systems and Technology*, 14(5):1–48, 2023c.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 188–197, 2019.

Wentao Ning, Reynold Cheng, Xiao Yan, Ben Kao, Nan Huo, Nur Al Hasan Haldar, and Bo Tang. Debiasing recommendation with personal popularity. In *Proceedings of the ACM web conference 2024*, pp. 3400–3409, 2024.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Ruihong Qiu, Sen Wang, Zhi Chen, Hongzhi Yin, and Zi Huang. Causalrec: Causal inference for visual debiasing in visually-aware recommendation. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 3844–3852, 2021.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*, 2012.

James Robins. A new approach to causal inference in mortality studies with a sustained exposure periodapplication to control of the healthy worker survivor effect. *Mathematical modelling*, 7 (9-12):1393–1512, 1986.

Alan Said and Alejandro Bellogín. Comparative recommender system evaluation: benchmarking recommendation frameworks. In *Proceedings of the 8th ACM Conference on Recommender systems*, pp. 129–136, 2014.

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *international conference on machine learning*, pp. 1670–1679. PMLR, 2016.

Yu Shang, Chen Gao, Jiansheng Chen, Depeng Jin, and Yong Li. Improving item-side fairness of multimodal recommendation via modality debiasing. In *Proceedings of the ACM Web Conference 2024*, pp. 4697–4705, 2024.

Peijie Sun, Le Wu, and Meng Wang. Attentive recurrent social recommendation. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pp. 185–194, 2018.

Yanchao Tan, Zhenghong Lin, Sujie Pan, Siying Xu, Weiming Liu, Guofang Ma, and Shiping Wang. Heterogeneous hypergraph structure learning for multimedia recommendation. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2024.

Dongjing Wang, Haojiang Yao, Dongjin Yu, Shiyu Song, He Weng, Guandong Xu, and Shuiguang Deng. Graph intention embedding neural network for tag-aware recommendation. *Neural Networks*, 184:107062, 2025a.

Li Wang, Min Xu, Quangui Zhang, Yunxiao Shi, and Qiang Wu. Causal disentanglement for regulating social influence bias in social recommendation. *Neurocomputing*, 618:129133, 2025b.

Siyu Wang, Xiaocong Chen, Quan Z Sheng, Yihong Zhang, and Lina Yao. Causal disentangled variational auto-encoder for preference understanding in recommendation. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pp. 1874–1878, 2023a.

Xiangmeng Wang, Qian Li, Dianer Yu, Qing Li, and Guandong Xu. Reinforced path reasoning for counterfactual explainable recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3443–3459, 2024.

Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. A survey on the fairness of recommender systems. *ACM Transactions on Information Systems*, 41(3):1–43, 2023b.

Tianxin Wei, Fuli Feng, Jiawei Chen, Ziwei Wu, Jinfeng Yi, and Xiangnan He. Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 1791–1800, 2021.

Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM international conference on multimedia*, pp. 1437–1445, 2019.

Yinwei Wei, Wenqi Liu, Fan Liu, Xiang Wang, Liqiang Nie, and Tat-Seng Chua. Lightgt: A light graph transformer for multimedia recommendation. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pp. 1508–1517, 2023.

Xin Wen, Weizhi Nie, Jing Liu, Yuting Su, Yongdong Zhang, and An-An Liu. Cdcm: Chatgpt-aided diversity-aware causal model for interactive recommendation. *IEEE Transactions on Multimedia*, 26:6488–6500, 2024.

Lianghao Xia, Chao Huang, Yong Xu, Jiashu Zhao, Dawei Yin, and Jimmy Huang. Hypergraph contrastive collaborative filtering. In *Proceedings of the 45th International ACM SIGIR conference on research and development in information retrieval*, pp. 70–79, 2022.

Teng Xiao, Zhengyu Chen, and Suhang Wang. Reconsidering learning objectives in unbiased recommendation: A distribution shift perspective. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2764–2775, 2023.

Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, Wei Wang, Xiping Hu, Steven Hoi, and Edith Ngai. A survey on multimodal recommender systems: Recent advances and future directions. *arXiv preprint arXiv:2502.15711*, 2025a.

Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, Wei Wang, Xiping Hu, Steven Hoi, and Edith Ngai. A survey on multimodal recommender systems: Recent advances and future directions. *IEEE Transactions on Multimedia*, 2026.

Shuyuan Xu, Jianchao Ji, Yunqi Li, Yingqiang Ge, Juntao Tan, and Yongfeng Zhang. Causal inference for recommendation: Foundations, methods, and applications. *ACM Transactions on Intelligent Systems and Technology*, 16(3):1–51, 2025b.

Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. Mining latent structures for multimedia recommendation. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 3872–3880, 2021.

Qing Zhang, Xiaoying Zhang, Yang Liu, Hongning Wang, Min Gao, Jiheng Zhang, and Ruocheng Guo. Debiasing recommendation by learning identifiable latent confounders. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 3353–3363, 2023.

Yuyao Zhang, Ke Guo, and Xiao Zhou. Causally aware generative adversarial networks for light pollution control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 22529–22537, 2024.

Wayne Xin Zhao, Yupeng Hou, Xingyu Pan, Chen Yang, Zeyu Zhang, Zihan Lin, Jingsen Zhang, Shuqing Bian, Jiakai Tang, Wenqi Sun, et al. Recbole 2.0: Towards a more up-to-date recommendation library. In *Proceedings of the 31st ACM international conference on information & knowledge management*, pp. 4722–4726, 2022.

Yibowen Zhao, Yonghui Xu, Ning Liu, Yixin Zhang, Wei Guo, Xudong Lu, and Lizhen Cui. Causal denoising framework for generalizable recommendation system using graph neural network. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2024.

Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM web conference 2023*, pp. 845–854, 2023.

Xinyuan Zhu, Yang Zhang, Fuli Feng, Xun Yang, Dingxian Wang, and Xiangnan He. Mitigating hidden confounding effects for causal recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 36(9):4794–4805, 2024.

This appendix provides extended discussions, detailed baseline information, additional implementation details, and in-depth fairness evaluations to ensure the reproducibility and rigorousness of our proposed CORE framework. These sections aim to further clarify the theoretical rationale of our proxy-based causal disentanglement, explicitly compare CORE with recent advanced causal models, and rigorously validate its debiasing capabilities under unbiased protocols.

A Discussions on Causal Identifiability and Limitations

A.1 Identifiability Limits and Proxy-based Disentanglement

In rigorous causal inference, non-parametric identification of the interventional distribution $P(Y|do(C))$ from observational data necessitates a fully specified Structural Causal Model (SCM)

and strong parametric assumptions. In the context of deep multimodal recommendation, the true data-generating functions (e.g., f_X, f_Y) governing complex, high-dimensional modalities (images, text) are highly non-linear and largely intractable.

Therefore, rather than claiming a strict mathematical identification of $P(Y|do(C))$, our CORE framework adopts a pragmatic, proxy-based causal disentanglement approach. Recognizing that the coupled multimodal feature X acts as a collider ($C \rightarrow X \leftarrow S$), we acknowledge that the latent intrinsic content (C) and presentation manner (S) become spuriously correlated in the observational data. Our objective is to approximate the independent exogenous distributions. The Hilbert-Schmidt Independence Criterion (HSIC) loss acts as a structural regularization, compelling the learned proxies ($\mathbf{z}_c$ and $\mathbf{z}_s$) to be statistically orthogonal. By breaking the collider-induced spurious correlation at the representation level, this heuristic disentanglement provides a highly scalable and robust approximation of an invariant predictor. This trade-off is critical for real-world recommendation systems where strict identifiability is often unattainable.

A.2 Justification of Modality-Guided Heuristics

The automated pair construction (Algorithm 1) relies on modality-guided proxy heuristics: assuming text-similarity reflects intrinsic content, and popularity reflects presentation bias. We explicitly acknowledge this as a weak supervision trade-off. In domains like movie recommendations (ML-1M), text (plot) and visual (poster) semantics are relatively distinct, making this heuristic highly reliable. However, in domains like Fashion, visual appearance (e.g., color, cut) is intrinsically part of the content. In such cases, relying predominantly on text for content supervision risks partial information leakage, where some content signals might bleed into the bias representation $\mathbf{z}_s$.

Nevertheless, by combining this weak contrastive supervision with the HSIC independence constraint, CORE compels the network to discover the maximally independent factors of variation. The consistent empirical improvements across all three diverse datasets validate that these proxy heuristics are sufficiently effective for debiasing without requiring prohibitively expensive, fine-grained manual annotations. For future work, utilizing large foundation models (e.g., CLIP-based cross-modal semantics) to construct semantic content graphs presents a promising alternative to refine these proxies.

B Detailed Baselines and Comparisons

B.1 Detailed Baselines

From the perspective of data modeling and causal debiasing, we select representative models from general collaborative filtering, multimodal recommendation, and debiasing approaches.

- **NeuMF** He et al. (2017) is a fundamental collaborative filtering framework that replaces the inner product in matrix factorization with a neural architecture to learn non-linear user-item interactions.
- **LightGCN** He et al. (2020) simplifies the design of Graph Convolution Networks for recommendation by removing non-linear activations and feature transformations, relying purely on neighborhood aggregation of ID embeddings.
- **VBPR** He & McAuley (2016) is a pioneering multimodal method that integrates visual features extracted from deep neural networks into the classic matrix factorization framework using Bayesian Personalized Ranking.
- **MMGCN** Wei et al. (2019) constructs separate bipartite graphs for different modalities (e.g., visual, textual) and learns unified representations through modal-specific message passing and aggregation.
- **HMRec** Tan et al. (2024) employs heterogeneous hypergraphs to capture complex, high-order relations among users and multimodal items beyond simple pairwise interactions.
- **IPS-MMGCN** Schnabel et al. (2016) applies Inverse Propensity Scoring (Schnabel et al. 2016) to MMGCN by reweighting the training loss of each interaction based on the inverse of the item's observed popularity to mitigate exposure bias.

- **GCDF** Zhao et al. (2024) is a recent causal denoising framework for GNN-based recommendation that explicitly treats noisy or biased interactions as confounders and performs causal denoising to enhance generalizability.
- **PPAC** Ning et al. (2024) utilizes a popularity-penalized adversarial learning framework to compel item representations to be indistinguishable with respect to their popularity levels.
- **CDRSB** Wang et al. (2025b) is a recent causal disentangled representation learning framework designed to regulate social influence bias by separating preference from conformity.

B.2 Comparison with Related Causal Models

While not directly included in the main quantitative tables due to differing problem settings (e.g., non-multimodal, requirement of auxiliary graphs), we qualitatively compare CORE with several advanced causal models to highlight our conceptual positioning:

- **CaMRec** Jeong et al. (2024) utilizes two-stage front-door and back-door adjustments via diffusion models and causal subgraph reconstruction.
- **HCR** Zhu et al. (2024) demonstrates front-door adjustment for hidden confounders in multimodal settings.
- **iDCF** Zhang et al. (2023) employs instrumental variable (IV) strategies to provide theoretical identification under hidden confounding.
- **CaD-VAE** Wang et al. (2023a) explicitly embeds an SCM inside a generative VAE for causal disentanglement.

C Dataset Statistics and Experimental Details

C.1 Dataset Statistics

To complement the experimental setup described in Section 5.1.1 of the main paper, we provide the detailed statistical information of the three evaluated public datasets: Amazon-Beauty, Amazon-Fashion, and MovieLens-1M (ML-1M). As summarized in Table 3, these datasets encompass varying scales, interaction densities, and distinct content modalities. For all datasets, we applied a strict 5-core preprocessing step, ensuring that only users and items with at least five interactions were retained to guarantee the quality of the interaction graph. The implicit feedback is converted into binary labels, where ratings of 4 and 5 are treated as positive preferences, and all others as negative.

Table 3: Statistics of the preprocessed datasets.

Dataset	#Users	#Items	#Inters	#Density	avg.
Beauty	22,363	12,101	198,502	0.073%	8.87
Fashion	39,387	23,033	278,677	0.031%	7.07
ML-1M	6,040	3,706	1,000,209	4.469%	165.59

C.2 Additional Implementation Details

In Algorithm 1, the definition of Category varies by dataset to best reflect intrinsic content. For Amazon-Beauty and Amazon-Fashion, Category refers to the fine-grained, second-level product metadata tags provided in the raw dataset (e.g., “Skin Care", “Makeup"). For ML-1M, it refers to the primary movie genre label (e.g., “Action", “Sci-Fi"). The thresholds for positive candidate selection were determined empirically to balance sample quality and variance. Across all datasets, we set k_c and k_s to encompass the top 5% of the respective candidate item pools. Sensitivity analysis indicates that the model's performance remains stable as long as these thresholds are kept within the $[2\%, 10\%]$ range. All experiments were conducted on a Linux server equipped with four NVIDIA RTX 4070 GPUs, using Python 3.9 and PyTorch 1.13.1.

We utilize the Gaussian (RBF) kernel to compute the HSIC penalty ($\mathcal{L}_{indep}$), as it effectively captures complex, non-linear statistical dependencies between the representation spaces $\mathbf{z}_c$ and $\mathbf{z}_s$. The

computational complexity of HSIC is $\mathcal{O}(B^2d)$. Given our batch size of $B = 2048$ and embedding dimension $d = 64$, this matrix operation is highly optimized on GPUs. Empirically, the HSIC constraint adds less than 5% overhead to the total training time per epoch compared to a standard LightGCN, ensuring high efficiency.

Image Gathering and Preprocessing for ML-1M. As noted in Section 5.1.1, the original MovieLens-1M dataset lacks visual modalities. To conduct multimodal evaluations, we augmented the dataset by gathering the corresponding movie posters. Specifically, we mapped the standard MovieLens item IDs to their corresponding TMDB (The Movie Database) and IMDb identifiers. Using the official TMDB API, we successfully crawled the high-resolution movie poster images for the matched items. For data preprocessing, all raw images from the three datasets (Amazon-Beauty, Amazon-Fashion, and ML-1M) were standardized before feature extraction. The images were centrally cropped, resized to 224×224 pixels, and normalized using the standard ImageNet dataset statistics (mean: $[0.485, 0.456, 0.406]$, standard deviation: $[0.229, 0.224, 0.225]$). Finally, these preprocessed images were fed into the pre-trained ViT-B/32 model to extract the initial 64-dimensional visual embeddings $\mathbf{e}_v$.

D Algorithmic Details and Complexity Analysis

D.1 Automated Contrastive Pair Construction

In Section 4.3.1 of the main text, we introduced the contrastive disentanglement learning strategy designed to shape the semantics of the intrinsic content representation $\mathbf{z}_c$ and the presentation bias representation $\mathbf{z}_s$. Here, we provide the detailed, step-by-step procedure for our fully automated positive pair construction in Algorithm 1. The core advantage of this procedure is that it relies purely on pre-trained embedding similarities and metadata (e.g., categories and popularity distributions) as weak supervision heuristics. This eliminates the need for any costly fine-grained manual annotations. As outlined in the algorithm, the content positive (i^+) is selected to maximize text similarity while minimizing visual overlap, whereas the bias positive (i'^+) targets items with similar global popularity but across entirely different content categories.

Algorithm 1 Automated Contrastive Pair Construction

Input: Anchor item i, Item set $\mathcal{I}$, Pre-trained textual and visual embeddings E_T, E_V, Top-K thresholds k_c, k_s
Output: Content positive i^+, Bias positive i'^+

// Construct Content Positive i^+
$\mathcal{C}_i \leftarrow \text{Category}(i)$
$\mathcal{I}_{same} \leftarrow \{j \in \mathcal{I} \mid \text{Category}(j) == \mathcal{C}_i, j \neq i\}$
$TopK_{text} \leftarrow$ Top-k_c items in $\mathcal{I}_{same}$ with highest cosine similarity to i based on E_T
$i^+ \leftarrow \arg\min_{j \in TopK_{text}} \text{CosineSim}(E_V(i), E_V(j))$
// Construct Bias Positive i'^+
$Pop_i \leftarrow \text{Popularity}(i)$
$\mathcal{I}_{diff} \leftarrow \{j \in \mathcal{I} \mid \text{Category}(j) \neq \mathcal{C}_i\}$
$TopK_{pop} \leftarrow$ Top-k_s items in $\mathcal{I}_{diff}$ with popularity closest to Pop_i
$i'^+ \leftarrow \text{RandomSample}(TopK_{pop})$
return i^+, i'^+

D.2 Complexity Analysis

To demonstrate that our proposed CORE framework is scalable to real-world, large-scale multimodal recommendation scenarios, we present a detailed theoretical analysis of its time complexity.

We analyze the time complexity of CORE. Let $|\mathcal{V}| = |\mathcal{U}| + |\mathcal{I}|$ be the number of nodes, $|\mathcal{E}|$ the number of interactions, d the embedding size, and B the batch size. The LightGCN predictor takes $O(|\mathcal{E}|d)$. The MLP-based disentanglement encoders take $O(|\mathcal{I}|d^2)$. The contrastive pair construction (Algorithm 1) is performed offline. During training, the InfoNCE loss takes $O(B^2d)$ and the HSIC constraint takes $O(B^2d)$. Since $B \ll |\mathcal{I}|$, the overall training complexity is approximately

$O(|\mathcal{E}|d + |\mathcal{I}|d^2)$, which is on par with standard GNN-based recommendation models, ensuring scalability for large datasets.

E Further Analysis

To rigorously validate that CORE genuinely mitigates popularity-amplified presentation bias rather than merely shifting the overall score distribution, we conduct two granular evaluations.

E.1 Fairness Breakdown by Popularity Deciles

We divide the items in the ML-1M dataset into 10 decile groups based on their interaction frequency in the training set (Decile 1 = extremely long-tail, Decile 10 = extremely popular head items). We compute the Recall@20 for the Tail (Deciles 1-3), Mid (Deciles 4-7), and Head (Deciles 8-10) groups.

As shown in Table 4, standard models like LightGCN suffer heavily from popularity bias, achieving high recall on Head items but severely neglecting Tail items. In contrast, CORE significantly boosts the recall of Tail and Mid items (achieving a +41.5% relative improvement on Tail items compared to LightGCN), with only a marginal sacrifice in Head recall. This confirms that CORE structurally reallocates exposure from over-packaged popular items to high-quality long-tail items, yielding genuine diversity benefits.

E.2 Exposure-Adjusted (Intervened) Evaluation

Table 4: Recall@20 Breakdown by Item Popularity Groups (ML-1M).

Method	**Tail (1-3)** ↑	**Mid (4-7)** ↑	**Head (8-10)** ↑
LightGCN	0.0412	0.1255	**0.2915**
CDRSB	0.0504	0.1421	0.2783
CORE (Ours)	**0.0583**	**0.1594**	0.2710
% Improve	*+15.67%*	*+12.17%*	*-7.03%*

Table 5: IPS-Reweighted Unbiased Evaluation on Amazon-Beauty Test Set.

Method	**IPS-R@20**↑	**IPS-N@20**↑
LightGCN	0.0512	0.0275
CDRSB	0.0684	0.0381
CORE (Ours)	**0.0751**	**0.0422**
% Improve	*+9.80%*	*+10.76%*

Standard observational test sets are naturally biased towards popular items. To perform an unbiased evaluation, we apply Inverse Propensity Scoring (IPS) reweighting to the test set instances to simulate a uniform exposure distribution. The propensity score for an item is estimated based on its relative frequency in the training data.

As shown in Table 5, under this intervened unbiased evaluation protocol, the performance gap between CORE and the baselines widens further. This demonstrates that CORE's purified content representation $\mathbf{z}_c$ more accurately captures users' true, unbiased preferences, remaining highly robust even when evaluated outside the biased observational distribution.