

Full Length Article

Consistency-guided evidential debiasing for multimodal recommendation

Lin Pan, Zhiqiang Pan*, Honghui Chen

National Key Laboratory of Information Systems Engineering, National University of Defense Technology, Changsha, 410000, China

ARTICLE INFO

Keywords:

Recommender systems
Multimodal recommendation
Adaptive debiasing
Evidential fusion

ABSTRACT

Multimodal recommendation integrates auxiliary modalities, such as images and text, to enrich item representations and improve personalized ranking. However, recommendation accuracy can be affected by cross-modal bias, where single-modality signals may dominate multimodal representations and lead to unreliable recommendations. Handling this bias introduces two challenges: (1) Conventional multimodal fusion strategies may dilute misleading single-modality signals during aggregation, resulting in insufficient debiasing penalties. (2) Debiasing without evaluating cross-modal semantic consistency may misidentify semantically consistent items and impose unnecessary penalties. To address these issues, we propose a Consistency-guided Evidential Debiasing framework for multimodal Recommendation (CEDRec). Specifically, CEDRec utilizes an evidential bias network that adopts a simple Dempster-Shafer formulation to represent deceptive evidence and uncertainty for each modality. Subsequently, the network performs evidential fusion over these modality-specific estimates to obtain a reliable cross-modal bias score while alleviating the dilution of strong unimodal signals. Moreover, to prevent misidentifying semantically consistent items, CEDRec employs a semantic bottleneck network based on the variational information bottleneck principle. This network distills shared cross-modal semantics to generate a dynamic gating signal, thus protecting semantically consistent items from unnecessary penalties. Finally, during the joint optimization, gradient projection is employed to resolve conflicts between collaborative preference learning and semantic bottleneck compression, ensuring that shared semantics are distilled without sacrificing ranking performance. Extensive experiments on multiple real-world datasets demonstrate that CEDRec effectively mitigates cross-modal bias and achieves a more balanced item exposure distribution while maintaining competitive ranking performance.

1. Introduction

Multimodal Recommender Systems (MRSs) serve as the core information filtering engines for modern online platforms, such as short-video applications (He et al., 2025a; Lee et al., 2025), e-commerce websites (Giahi et al., 2025; Xu et al., 2025a), and social media networks (Lin et al., 2025b; Xu et al., 2025d). By extracting semantic signals from diverse item modalities (e.g., visual covers and textual titles), MRSs incorporate auxiliary knowledge to alleviate the data sparsity and cold-start problems of traditional collaborative filtering (Lu et al., 2026; Pan et al., 2026). However, the effectiveness of MRSs depends heavily on whether the learned multimodal representations can reliably capture item semantics. Therefore, developing robust representation learning paradigms is essential for constructing effective and reliable MRSs.

Early MRSs (He & McAuley, 2016; Liu et al., 2017) integrate multimodal signals into matrix factorization architectures by concatenating item ID embeddings with visual or textual features for representation learning. Moreover, to capture higher-order user-item relationships be-

yond these early designs, subsequent studies (Wei et al., 2019; Zhang et al., 2021) introduce Graph Convolutional Networks (GCNs). In detail, these models directly embed multimodal features into the message-passing mechanism over the user-item bipartite graphs, utilizing iterative neighborhood aggregation to capture collaborative topological semantics. More recently, self-supervised learning paradigms have been widely adopted to further improve multimodal representation quality (Shi et al., 2026; Tao et al., 2022; Zhou et al., 2023). Specifically, by constructing auxiliary tasks such as cross-modal contrastive learning or feature reconstruction from unlabeled multimodal inputs, these methods enforce semantic alignment to capture underlying correlations across diverse modalities.

Despite their promising performance, existing methods generally overlook cross-modal bias in real-world data. Specifically, they rely on the assumption of cross-modal semantic alignment, which may not hold for all items in practice. A strong signal in one modality may not be supported by the other modality, yet it can still dominate the learned item representation and affect the recommendation score. Since ground-truth

* Corresponding author.

E-mail addresses: panlin@nudt.edu.cn (L. Pan), panzhiqiang@nudt.edu.cn (Z. Pan), chenhonghui@nudt.edu.cn (H. Chen).

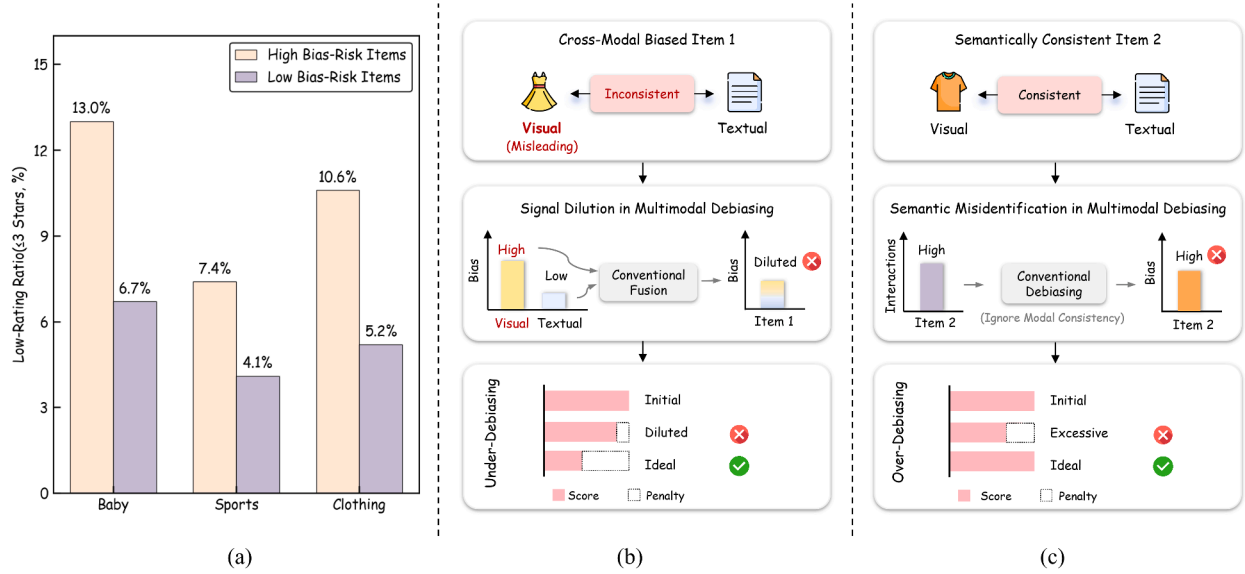


Fig. 1. Overview of cross-modal bias and challenges in multimodal recommendation. (a) Popularity-controlled comparison of item-level low-rating ratios (≤ 3 stars) between high bias-risk items and matched low bias-risk items across three Amazon datasets. The two groups are identified by the proposed cross-modal bias-risk score, while ratings serve as an external behavioral indicator. (b) Signal dilution in multimodal debiasing leads to insufficient penalties for deceptive items. (c) Semantic misidentification in multimodal debiasing leads to unnecessary penalties for semantically consistent items.

item-level bias labels are unavailable, we construct a data-driven bias-risk proxy based on cross-modal disagreement and item popularity, and examine its association with user feedback. We use this proxy to identify candidate high- and low-risk items and match them by training-set interaction counts to control for popularity. As shown in Fig. 1(a), after controlling for item popularity, high bias-risk items consistently show higher ratios of ratings of three stars or below than matched low bias-risk items across three datasets. This result provides independent behavioral evidence that the estimated bias risk is associated with less favorable user feedback, and suggests that this association is not driven by item popularity alone. However, effectively reducing this bias involves two key challenges:

(1) **Signal dilution in multimodal debiasing.** Cross-modal bias often arises when misleading signals are concentrated in a single modality. In conventional multimodal debiasing, directly aggregating modality-specific signals can smooth out the discrepancy between a high-risk modality and a less-biased one, thus diluting the strong bias evidence from the high-risk modality. As illustrated in Fig. 1(b), this signal dilution can cause the overall bias to be underestimated. As a result, the debiasing module applies an insufficient penalty, failing to fully correct the recommendation score and leaving residual bias in the final ranking.

(2) **Semantic misidentification in multimodal debiasing.** Cross-modal debiasing also faces the risk of semantic misidentification, which can lead to over-debiasing. As shown in Fig. 1(c), high-quality items may exhibit strong and consistent multimodal signals that naturally attract frequent user interactions. However, debiasing frameworks may fail to determine whether frequent interactions arise from consistent multimodal signals or misleading signals in a single modality. Consequently, semantically consistent high-quality items may be incorrectly identified as biased and receive unnecessary penalties, reducing their exposure and degrading recommendation performance.

To address these challenges, this paper proposes a **C**onsistency-guided **E**vidential **D**ebiasing framework for multimodal **R**ecommendation (CEDRec). Specifically, to mitigate the signal dilution problem in multimodal debiasing, CEDRec incorporates an Evidential Bias Network (EBN) that represents modality-specific deceptive evidence and uncertainty using a simple Dempster-Shafer formulation. These estimates are then combined to obtain a cross-modal bias score, thus alleviating the dilution of strong modality-specific signals. Moreover, to prevent se-

mantically consistent high-quality items from being misidentified as biased, CEDRec employs a Semantic Bottleneck Network (SBN) based on the variational information bottleneck principle. In detail, by distilling shared semantics from visual and textual features, SBN assesses the semantic consistency between modalities and generates a dynamic gating signal to modulate the debiasing penalty, thus protecting naturally aligned items from unnecessary penalties. Finally, during the joint optimization, gradient projection is applied to dynamically mitigate conflicts between the primary collaborative recommendation task and the semantic bottleneck compression task. Comprehensive experiments across multiple real-world datasets validate the effectiveness of CEDRec in mitigating cross-modal bias while preserving competitive recommendation performance.

The main contributions of this paper can be summarized as follows:

- We propose CEDRec, an adaptive debiasing framework for multimodal recommendation that mitigates cross-modal bias by reducing the influence of potentially deceptive modality-specific signals on collaborative preference learning.
- We introduce an evidential bias network to separately estimate modality-specific deceptive evidence and uncertainty. A simple Dempster-Shafer formulation is used to combine these estimates across modalities, preserving strong deceptive evidence under cross-modal uncertainty and alleviating signal dilution.
- We design a semantic bottleneck network based on the variational information bottleneck principle to distill shared semantics and dynamically gate the debiasing penalty, reducing unnecessary penalties on semantically consistent items.
- Extensive experiments demonstrate that CEDRec achieves a favorable balance between debiasing effectiveness and recommendation quality.

2. Related work

In this section, we review related work in two areas: multimodal recommendation and debiasing in recommender systems. The former focuses on modeling and integrating multimodal features, while the latter addresses various biases in observational data.

2.1. Multimodal recommendation

Traditional collaborative filtering methods rely on historical user-item interactions to learn latent ID representations, but they suffer from data sparsity and representation bottlenecks (Chen et al., 2025a; He et al., 2025b). To overcome these constraints, multimodal recommendation incorporates auxiliary modalities (e.g., visual and textual features) to enrich item semantics. In a pioneering study, He and McAuley (2016) extend traditional matrix factorization by directly fusing item IDs with visual features to improve representation learning. Subsequently, to capture higher-order structural correlations, GCNs have been employed. For example, Wei et al. (2019) utilize modality-specific GCNs to aggregate collaborative node features from topological neighborhoods, while Zhang et al. (2021) enhance item representation learning by mining latent item-item relationships directly from multimodal content. Recent research has increasingly focused on learning robust multimodal item representations. To alleviate data sparsity and semantic misalignment, self-supervised learning paradigms have been widely adopted. For example, Tao et al. (2022) and Zhou et al. (2023) leverage contrastive tasks to align internal semantic representations across different modalities. Xu et al. (2025c) introduce ID-guided multi-level alignment to reduce cross-modal discrepancies while retaining collaborative information. Moreover, to mitigate the impact of interaction noise, Zhou and Shen (2023) implement degree-sensitive edge pruning on the user-item bipartite graph to filter out unreliable historical interactions. To address the inherent multimodal noise and the interference introduced during contrastive learning, Song et al. (2025) employ diffusion models to generate robust contrastive views of the modality-specific user-item graphs. Ong and Khong (2025) further suppress modality-specific noise through adaptive spectral filtering, while Xu et al. (2025b) couple dual-stage modality fusion with heterogeneous and homogeneous graph modeling to reduce irrelevant multimodal information. In addition, Guo et al. (2024) and Hu et al. (2025) focus on extracting stable topological signals via global hypergraphs or modality-independent Transformers to further enhance representation stability. Recent studies further refine multimodal representations from complementary structural perspectives. Yu et al. (2025) preserve individual preference information by extracting principal local structures, while Chen et al. (2025b) employ multi-component hypercomplex representations and nonlinear cross-modal interactions to improve representation diversity. Shi et al. (2026) jointly capture global behavioral patterns and local multimodal structures through Transformer and hypergraph learning. Ma et al. (2026) reconstruct item-item and user-item graphs to reduce noisy structural relations and enrich sparse collaborative signals. Furthermore, Shi et al. (2025) construct heterogeneous modality environments with cross-modal alignment to improve robustness to incomplete multimodal information.

However, existing multimodal recommendation methods generally assume cross-modal semantic alignment and thus overlook cross-modal bias in real-world data. Although these methods improve multimodal representations through fusion, alignment, or denoising, they do not explicitly assess whether strong modality-specific signals are semantically supported across modalities. Consequently, misleading signals concentrated in a single modality may still exert disproportionate influence on the learned item representation and recommendation score.

2.2. Debiasing in recommender systems

Recommender systems inevitably suffer from various biases in observational user behavior data, such as popularity bias (Lin et al., 2025a; Wu et al., 2025), exposure bias (Mansoury et al., 2024), and conformity bias (Zheng et al., 2026). To mitigate these issues, various strategies have been explored. From a risk discrepancy perspective, Chen et al. (2021) leverage meta-learning on a small set of uniform data to optimize debiasing parameters. Moreover, Bai et al. (2025) disentangle invariant and variant user preferences via invariant learning, and employ

distance-aware knowledge distillation to adaptively fuse them for unbiased recommendations without requiring uniform data. More recently, causal inference and counterfactual reasoning have emerged as powerful tools for debiasing (Chen et al., 2023; Gao et al., 2024). Specifically, Zhang et al. (2023) introduce proximal causal inference with proxy variables to learn identifiable latent confounders. Furthermore, to mitigate popularity bias, Ning et al. (2024) design a counterfactual framework to jointly estimate and control the direct effects of personal and global popularity. In addition, Liu et al. (2022) and Shang et al. (2024) utilize counterfactual inference to adaptively eliminate the modality bias caused by uneven multimodal content distribution, thus improving item-side fairness.

However, existing debiasing methods do not explicitly account for cross-modal semantic consistency when estimating and mitigating modality-related bias. Our work addresses this limitation by preserving strong modality-specific bias evidence to alleviate signal dilution and incorporating semantic consistency to reduce unnecessary penalties for semantically consistent items.

3. Methodology

In this section, we detail the proposed CEDRec framework, as illustrated in Fig. 2. Specifically, Section 3.1 formulates the multimodal recommendation problem and defines the essential notations. Section 3.2 introduces the base graph network to learn multimodal collaborative representations. Section 3.3 describes the evidential bias network, which quantifies the modality-specific deceptive bias via evidential fusion. Section 3.4 presents the semantic bottleneck network to distill cross-modal shared semantics for dynamic penalty gating. Section 3.5 details the adaptive debiasing and joint optimization strategy, in which conflicting gradients are projected to mitigate gradient conflicts.

3.1. Problem formulation

Let $\mathcal{U} = \{u_1, u_2, \dots, u_M\}$ and $\mathcal{I} = \{i_1, i_2, \dots, i_N\}$ represent the sets of users and items, respectively. A binary matrix $\mathbf{R} \in \{0, 1\}^{M \times N}$ is constructed to record the observed interactions, where $r_{ui} = 1$ indicates an existing interaction between user u and item i , and $r_{ui} = 0$ otherwise. For the item-side inputs, let $\mathbf{E}_{id} \in \mathbb{R}^{N \times d}$, $\mathbf{E}_v \in \mathbb{R}^{N \times d_v}$, and $\mathbf{E}_t \in \mathbb{R}^{N \times d_t}$ denote the item ID embeddings and the visual and textual feature matrices, respectively. The corresponding feature dimensions are d , d_v , and d_t .

Given the historical interactions $\mathbf{R}$ and the item-side representations $\{\mathbf{E}_{id}, \mathbf{E}_v, \mathbf{E}_t\}$, multimodal recommendation aims to learn a scoring function $f_\theta(\cdot)$ that predicts the preference score $\hat{y}_{ui}$ for each user-item pair (u, i) , which can be formulated as follows:

$$[\mathbf{R}; \mathbf{E}_{id}; \mathbf{E}_v; \mathbf{E}_t] \xrightarrow{f_\theta(\cdot)} \hat{y}_{ui}. \quad (1)$$

The learnable function $f_\theta(\cdot)$ derives user and item representations $(\mathbf{h}_u, \mathbf{h}_i)$ from these inputs and uses them to estimate $\hat{y}_{ui}$. The resulting scores are then used to rank candidate items for each user. The main notations used in this paper are summarized in Table 1.

3.2. Base graph network

Following prior studies (Hu et al., 2025; Xu et al., 2025b), we construct separate user-item bipartite graphs for each modality $m \in \{id, v, t\}$, as illustrated in Fig. 2(b). Specifically, the modality-specific user preference embeddings $\mathbf{P}_m \in \mathbb{R}^{M \times d}$ are initialized as trainable parameter matrices via the Xavier normal distribution to capture independent latent preferences. Meanwhile, to align the multimodal features with the collaborative latent space, the raw features $\mathbf{E}_m$ are projected through a two-layer Multi-Layer Perceptron (MLP) with LeakyReLU activation to obtain the aligned item embeddings $\tilde{\mathbf{E}}_m \in \mathbb{R}^{N \times d}$. The initial node embedding matrix $\mathbf{H}_m^{(0)} \in \mathbb{R}^{(M+N) \times d}$ for each graph is formed by concatenating these refined embeddings (i.e., $[\mathbf{P}_m; \tilde{\mathbf{E}}_m]$), serving as the

Table 1
Main notations used in this paper.

Notation	Description
$\mathcal{U}, \mathcal{I}$	The sets of users and items
$\mathbf{R}, r_{ui}$	The user-item interaction matrix and its binary implicit feedback entry
$\mathbf{P}_m$	The initial user preference embedding matrix for modality $m \in \{id, v, t\}$
$\mathbf{E}_m, \tilde{\mathbf{E}}_m$	The raw and MLP-aligned item embedding matrices for modality m
$\mathbf{H}_m^{(l)}$	The representation at the l -th GCN layer for modality m
$\mathbf{h}_u, \mathbf{h}_i$	The final refined representations of user u and item i
$\tilde{y}_{u,i}, \hat{y}_{u,i}$	The base collaborative matching score and the final debiased prediction score
β_m, η_m	The modality-specific deceptive evidence score and evidential uncertainty
S	The fused cross-modal bias score
$\mu_m, \log(\sigma_m^2)$	The mean and log-variance matrices of the modality-specific latent distribution
$\mathbf{Z}_m$	The L_2 -normalized shared semantic latent representation
G	The dynamic gating signal indicating cross-modal semantic consistency
$\mathcal{L}_{main}, \mathcal{L}_s$	The main BPR recommendation loss and the semantic bottleneck loss
$\mathbf{g}_m, \mathbf{g}_s$	The gradient vectors for the main recommendation task and the SBN task
γ, τ	The debiasing penalty strength and the InfoNCE temperature
λ_e, λ_s	The hyperparameters balancing the EBN and SBN objectives

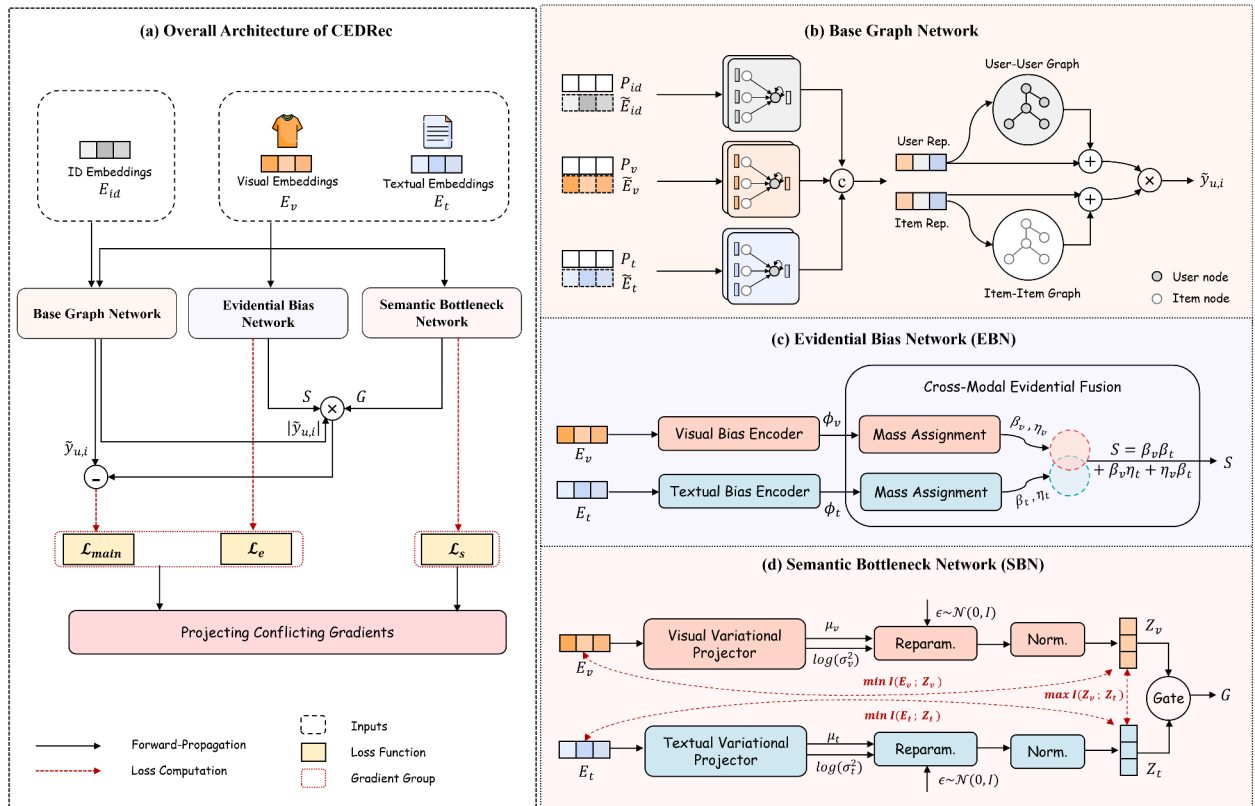


Fig. 2. The framework of CEDRec. (a) The overall architecture illustrating the forward propagation and joint optimization process; (b) Base Graph Network learns multimodal collaborative representations; (c) Evidential Bias Network quantifies modality-specific deceptive bias via evidential fusion; (d) Semantic Bottleneck Network distills cross-modal shared semantics for dynamic penalty gating.

base state to capture high-order collaborative signals. During message passing, we dynamically scale the aggregated neighborhood features based on their similarity with the initial ego-embeddings $\mathbf{H}_m^{(0)}$ to alleviate over-smoothing. Specifically, the representation at the l -th layer, $\mathbf{H}_m^{(l)} \in \mathbb{R}^{(M+N) \times d}$, is updated as:

$$\mathbf{H}_m^{(l)} = \cos(\mathbf{A}_m \mathbf{H}_m^{(l-1)}, \mathbf{H}_m^{(0)}) \odot (\mathbf{A}_m \mathbf{H}_m^{(l-1)}), \quad (2)$$

where $\mathbf{A}_m \in \mathbb{R}^{(M+N) \times (M+N)}$ is the symmetric normalized adjacency matrix, $\cos(\cdot, \cdot)$ denotes the row-wise cosine similarity function, and $\odot$ is the Hadamard product. Moreover, to capture multi-hop collaborative signals, we aggregate the representations across all $L + 1$ layers, forming the final modality-specific representation $\mathbf{H}_m \in \mathbb{R}^{(M+N) \times d}$, which can be

formulated as follows:

$$\mathbf{H}_m = \sum_{l=0}^L \mathbf{H}_m^{(l)}. \quad (3)$$

The refined outputs $\mathbf{H}_{id}, \mathbf{H}_v, \mathbf{H}_t \in \mathbb{R}^{(M+N) \times d}$ are then concatenated into a joint multimodal matrix $\mathbf{Z} \in \mathbb{R}^{(M+N) \times 3d}$. Next, to construct user-user and item-item graphs for capturing homogeneous structural signals, we partition the joint matrix $\mathbf{Z}$ into user representations $\mathbf{Z}^{(U)} \in \mathbb{R}^{M \times 3d}$ and item representations $\mathbf{Z}^{(I)} \in \mathbb{R}^{N \times 3d}$. Specifically, to capture the item-item connections, the semantic correlation between any two items i and j is first calculated via cosine similarity of their multimodal representations $\mathbf{z}_i, \mathbf{z}_j \in \mathbf{Z}^{(I)}$. Subsequently, to filter out noisy connections and formulate the final item-item adjacency matrix $\mathcal{A}^{(I)} \in \mathbb{R}^{N \times N}$, we sparsify

these dense correlations via a top- k strategy as follows:

$$\mathcal{A}_{i,j}^{(I)} = \begin{cases} \frac{\mathbf{z}_i^\top \mathbf{z}_j}{\|\mathbf{z}_i\| \|\mathbf{z}_j\|} & \text{if } j \in \mathcal{N}_k(i) \\ 0 & \text{otherwise} \end{cases}, \quad (4)$$

where $\mathcal{N}_k(i)$ denotes the set of top- k neighbors for item i . Similarly, the user-user adjacency matrix $\mathcal{A}^{(U)} \in \mathbb{R}^{M \times M}$ is constructed using $\mathbf{Z}^{(U)}$ under the same similarity-based sparsification strategy. After that, we employ residual connections during message passing to preserve previously learned heterogeneous signals. Specifically, for any given node v (either a user or an item), its final representation $\mathbf{h}_v \in \mathbb{R}^{3d}$ is updated via element-wise addition as follows:

$$\mathbf{h}_v = \mathbf{z}_v + \sum_{j \in \mathcal{N}_k(v)} \mathcal{A}_{v,j} \mathbf{z}_j. \quad (5)$$

Finally, the base collaborative matching score is computed via an inner product as $\tilde{y}_{u,i} = \mathbf{h}_u^\top \mathbf{h}_i$, which serves as the primary predictive signal prior to the subsequent debiasing correction.

3.3. Evidential bias network

To estimate cross-modal bias while preserving strong modality-specific signals, we propose an Evidential Bias Network (EBN). Specifically, EBN first adopts a simple Dempster-Shafer formulation to represent deceptive evidence and uncertainty for each modality (Denoeux, 2000; Han et al., 2022). Then, it combines these modality-specific estimates using Dempster's combination rule to obtain a cross-modal bias score, thus alleviating the dilution of strong unimodal bias signals. Finally, in the absence of item-level bias labels, EBN employs a topology-guided weak prior to regularize the fused cross-modal bias estimation.

Unimodal evidence quantification. As depicted in Fig. 2(c), the raw visual and textual features $\mathbf{E}_m$ for $m \in \{v, t\}$ are fed into modality-specific encoders to extract latent representations that characterize potentially deceptive signals within each modality. Specifically, each modality-specific encoder progressively transforms its input through two hidden layers, with $\mathbf{E}_m^{(0)} = \mathbf{E}_m$, yielding the hidden representations $\mathbf{E}_m^{(j)} \in \mathbb{R}^{N \times d_m^{(j)}}$ ($j \in \{1, 2\}$) as follows:

$$\mathbf{E}_m^{(j)} = \text{LReLU}(\text{LN}(\mathbf{E}_m^{(j-1)} \mathbf{W}_m^{(j)} + \mathbf{b}_m^{(j)})), \quad (6)$$

where $\mathbf{W}_m^{(j)} \in \mathbb{R}^{d_m^{(j-1)} \times d_m^{(j)}}$ and $\mathbf{b}_m^{(j)} \in \mathbb{R}^{d_m^{(j)}}$ denote the learnable network weights and bias terms, respectively, LN denotes layer normalization, and LReLU denotes the LeakyReLU activation function. Moreover, to reduce overfitting in the textual branch, a dropout layer is applied to the first textual representation $\mathbf{E}_t^{(1)}$. Specifically, $\mathbf{E}_t^{(1)}$ is updated as $\mathbf{E}_t^{(1)} \leftarrow \frac{1}{1-p_d} (\mathbf{E}_t^{(1)} \odot \mathbf{B})$, where the elements of $\mathbf{B}$ are independently sampled from a Bernoulli distribution with success probability $1 - p_d$, and p_d denotes the dropout rate. Finally, the third layer of the encoder linearly projects the hidden representation $\mathbf{E}_m^{(2)}$ into a logit vector $\boldsymbol{\phi}_m \in \mathbb{R}^N$ via $\boldsymbol{\phi}_m = \mathbf{E}_m^{(2)} \mathbf{W}_m^{(3)} + \mathbf{b}_m^{(3)}$, where $\mathbf{W}_m^{(3)} \in \mathbb{R}^{d_m^{(2)} \times 1}$ and $\mathbf{b}_m^{(3)} \in \mathbb{R}$ are the trainable projection weights and bias terms, respectively.

Then, to quantify the deceptive evidence and uncertainty, we use a non-negative evidence mapping commonly adopted in evidential learning (Sensoy et al., 2018). Specifically, the learned logit vector $\boldsymbol{\phi}_m$ is transformed into a non-negative evidence vector $\boldsymbol{\psi}_m$ via the element-wise softplus function as follows:

$$\boldsymbol{\psi}_m = \ln(1 + \exp(\boldsymbol{\phi}_m)). \quad (7)$$

Since our formulation models only evidence associated with deceptive risk, while insufficient evidence is represented as uncertainty, we define the modality-specific evidence strength as $\boldsymbol{\alpha}_m = \boldsymbol{\psi}_m + 1$. Based on this strength, the deceptive evidence score vector $\boldsymbol{\beta}_m$ and uncertainty vector $\boldsymbol{\eta}_m$ are defined as follows:

$$\boldsymbol{\beta}_m = \boldsymbol{\psi}_m \odot \boldsymbol{\alpha}_m, \quad \boldsymbol{\eta}_m = 1 \odot \boldsymbol{\alpha}_m, \quad (8)$$

where $\odot$ denotes the element-wise division. For each item i , $\boldsymbol{\beta}_m + \boldsymbol{\eta}_m = 1$. A larger $\boldsymbol{\beta}_m$ indicates stronger deceptive evidence and thus a higher potential bias risk in modality m , while a larger $\boldsymbol{\eta}_m$ indicates higher uncertainty due to insufficient evidence.

Cross-modal evidential fusion. Next, we combine the modality-specific deceptive evidence to obtain an item-level cross-modal bias score under the Dempster-Shafer framework. The fusion is performed independently for each item, and the item index is omitted below for clarity. We define the set of possible states as $\Omega = \{\omega, \neg\omega\}$, where ω denotes the presence of a deceptive signal and $\neg\omega$ denotes its absence. In the Dempster-Shafer framework, a mass represents the degree of evidence assigned to a state or a set of states. Accordingly, mass assigned to $\{\omega\}$ represents support for deceptive risk, while mass assigned to Ω represents uncertainty due to insufficient evidence. In our setting, we adopt a positive-evidence construction, where each modality assigns mass only to the deceptive state and the uncertainty state, without explicitly assigning mass to $\neg\omega$. Accordingly, for each item, the modality-specific mass function is defined as $\mathcal{M}_m(\{\omega\}) = \boldsymbol{\beta}_m$ and $\mathcal{M}_m(\Omega) = \boldsymbol{\eta}_m$, with $\mathcal{M}_m(\{\neg\omega\}) = 0$.

Then, we combine the visual and textual mass functions using Dempster's orthogonal sum operator $\oplus$. Since the focal elements of both modalities are restricted to $\{\omega\}$ and Ω , their intersections are never empty. Therefore, the conflict coefficient is zero and the normalization term in Dempster's rule equals one. Thus, the combination rule reduces to:

$$[\mathcal{M}_v \oplus \mathcal{M}_t](A) = \sum_{X \cap Y = A} \mathcal{M}_v(X) \mathcal{M}_t(Y), \quad \forall A \subseteq \Omega. \quad (9)$$

Moreover, to compute the cross-modal bias score S , which represents the fused evidential mass assigned to the deceptive state ω , we aggregate all mass products satisfying $X \cap Y = \{\omega\}$. Given that each modality has only two focal elements, $\{\{\omega\}, \Omega\}$, there are three combinations whose intersections yield $\{\omega\}$: $\{\omega\} \cap \{\omega\}$ (both modalities support the deceptive state), $\{\omega\} \cap \Omega$ (the visual modality supports the deceptive state while the textual modality is uncertain), and $\Omega \cap \{\omega\}$ (the textual modality supports the deceptive state while the visual modality is uncertain). Therefore, the bias score S expands into three terms as follows:

$$S = [\mathcal{M}_v \oplus \mathcal{M}_t](\{\omega\}) = \boldsymbol{\beta}_v \odot \boldsymbol{\beta}_t + \boldsymbol{\beta}_v \odot \boldsymbol{\eta}_t + \boldsymbol{\eta}_v \odot \boldsymbol{\beta}_t. \quad (10)$$

Correspondingly, the remaining fused mass is assigned to the uncertainty state Ω , i.e., $[\mathcal{M}_v \oplus \mathcal{M}_t](\Omega) = \boldsymbol{\eta}_v \odot \boldsymbol{\eta}_t$. Since $\boldsymbol{\eta}_m = 1 - \boldsymbol{\beta}_m$, Eq. (10) can be equivalently simplified to:

$$S = \boldsymbol{\beta}_v + \boldsymbol{\beta}_t - \boldsymbol{\beta}_v \odot \boldsymbol{\beta}_t. \quad (11)$$

Under the simple Dempster-Shafer formulation adopted in this work, we use a positive-evidence setting where only evidence supporting deceptive risk is modeled, while insufficient evidence is assigned to uncertainty. In this setting, Dempster's combination rule reduces algebraically to the noisy-OR formulation above. The role of the Dempster-Shafer framework is therefore not to define a fusion operator different from noisy-OR, but to provide an evidential representation in which $\boldsymbol{\beta}_m$ denotes the mass supporting deceptive risk and $\boldsymbol{\eta}_m$ denotes the remaining mass assigned to uncertainty before fusion. For $\boldsymbol{\beta}_v, \boldsymbol{\beta}_t \in [0, 1]$, the resulting score satisfies $S \geq \boldsymbol{\beta}_v$ and $S \geq \boldsymbol{\beta}_t$. Therefore, strong deceptive evidence from either modality is preserved during fusion, while additional evidence from the other modality can be accumulated to obtain the final cross-modal bias score.

Topology-guided weak bias regularization. In addition, to optimize EBN in the absence of ground-truth labels, we design a data-driven proxy to provide a coarse item-level weak risk prior for fused cross-modal bias estimation. Specifically, we construct the prior from two observable signals: cross-modal neighborhood disagreement indicates potential relational inconsistency between modalities, while interaction frequency serves as a proxy for the potential influence of such inconsistency on collaborative recommendation.

For each item i , we construct sparse top- k neighborhood probability distributions $q_{v,i}$ and $q_{t,i}$ in the initial visual and textual spaces, respec-

tively, by applying softmax normalization to the cosine similarities of its top- k neighbors. Then, we compute the normalized Jensen-Shannon (JS) divergence as $div_i = D_{JS}(q_{v,i} \parallel q_{t,i}) / \ln 2 \in [0, 1]$ to quantify their topological disagreement. Subsequently, to formulate an item-level risk prior $w_i \in [0, 1]$, we combine div_i with the item's normalized popularity pop_i . Here, $pop_i = (c_i - \min(c)) / (\max(c) - \min(c))$, where c_i is the number of observed interactions for item i in the training set, and c is the set of training interaction counts across all items. After that, to identify items that are both highly popular and topologically mismatched, we employ the harmonic mean as follows:

$$w_i = \frac{2 \cdot pop_i \cdot div_i}{pop_i + div_i + \delta}, \quad (12)$$

where δ is a small constant for numerical stability. The prior is computed once before optimization and remains fixed throughout training.

Finally, since w_i provides a coarse item-level prior for cross-modal bias risk, we use it to regularize the fused bias score S_i . Specifically, we minimize the Mean Squared Error (MSE) over each sampled mini-batch $B \subset \mathcal{I}$ as follows:

$$\mathcal{L}_e = \frac{1}{|B|} \sum_{i \in B} (S_i - w_i)^2. \quad (13)$$

This prior-guided regularization anchors the fused bias estimation to observable risk patterns while preserving the flexibility of modality-specific evidence learning. In particular, w_i only provides a coarse and static item-level risk anchor derived from raw topological disagreement and popularity, without prescribing how the risk should be distributed between $\beta_{v,i}$ and $\beta_{t,i}$ or explicitly determining cross-modal semantic consistency. The latter is dynamically evaluated by the semantic bottleneck network during end-to-end optimization.

3.4. Semantic bottleneck network

EBN estimates cross-modal bias risk but does not explicitly assess cross-modal semantic consistency. Therefore, relying solely on the fused score S to determine the penalty may cause semantically consistent high-quality items to be misidentified as biased and receive unnecessary penalties. To provide this complementary semantic signal, we introduce a Semantic Bottleneck Network (SBN) grounded in the variational information bottleneck principle. Specifically, SBN first maps multimodal features into a probabilistic latent space via variational encoding. Then, it distills shared semantics to evaluate cross-modal consistency, and finally generates a dynamic gating signal to adjust the debiasing penalty.

Variational latent encoding. As illustrated in Fig. 2(d), SBN maps the multimodal features $\mathbf{E}_m$ ($m \in \{v, t\}$) into a shared probabilistic latent space to distill shared cross-modal semantics. Specifically, $\mathbf{E}_m$ is first projected through a non-linear layer with the LeakyReLU activation to learn compressed semantics. From this refined representation, two independent linear heads are used to parameterize the Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m^2)$, producing the mean matrix $\boldsymbol{\mu}_m$ and the log-variance matrix $\log(\boldsymbol{\sigma}_m^2)$, both in $\mathbb{R}^{N \times d_z}$. Moreover, to preserve differentiability during stochastic sampling, the latent representation $\tilde{\mathbf{Z}}_m \in \mathbb{R}^{N \times d_z}$ is computed via the reparameterization trick as follows:

$$\tilde{\mathbf{Z}}_m = \boldsymbol{\mu}_m + \boldsymbol{\sigma}_m \odot \boldsymbol{\epsilon}, \quad (14)$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ denotes an independent standard Gaussian noise matrix. Finally, $\tilde{\mathbf{Z}}_m$ is L_2 -normalized along the feature dimension to produce the representation $\mathbf{Z}_m \in \mathbb{R}^{N \times d_z}$.

Cross-modal semantic alignment. Guided by the variational information bottleneck principle, we formulate a multi-view bottleneck objective designed to filter out modality-specific noise while preserving the shared semantics. This requires minimizing the input-latent mutual information ($\min I(\mathbf{E}_m; \mathbf{Z}_m)$) alongside maximizing the cross-view mutual information ($\max I(\mathbf{Z}_v; \mathbf{Z}_t)$). Specifically, the compression term $\min I(\mathbf{E}_m; \mathbf{Z}_m)$ is optimized via its variational upper bound, implemented as the Kullback-Leibler (KL) divergence between the learned latent distribution $\mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m^2)$ and a standard Gaussian prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$. In practice,

this is approximated over each sampled mini-batch $B \subset \mathcal{I}$ as follows:

$$\mathcal{L}_{KL} = -\frac{1}{2|B|} \sum_{i \in B} \sum_{m \in \{v, t\}} \sum_{k=1}^{d_z} \left(1 + \log(\sigma_{m,i,k}^2) - \mu_{m,i,k}^2 - \sigma_{m,i,k}^2 \right). \quad (15)$$

Meanwhile, to optimize the preservation term $\max I(\mathbf{Z}_v; \mathbf{Z}_t)$, we employ the InfoNCE (van den et al., 2018) contrastive loss, which maximizes a lower bound on cross-modal mutual information. In detail, by contrasting the aligned visual-textual pairs against the batch-wise negatives, this objective encourages cross-modal semantic alignment. Specifically, the loss is defined as follows:

$$\mathcal{L}_{NCE} = -\frac{1}{2|B|} \sum_{i \in B} \left(\log \frac{\exp(\cos(\mathbf{Z}_{v,i}, \mathbf{Z}_{t,i})/\tau)}{\sum_{j \in B} \exp(\cos(\mathbf{Z}_{v,i}, \mathbf{Z}_{t,j})/\tau)} + \log \frac{\exp(\cos(\mathbf{Z}_{t,i}, \mathbf{Z}_{v,i})/\tau)}{\sum_{j \in B} \exp(\cos(\mathbf{Z}_{t,i}, \mathbf{Z}_{v,j})/\tau)} \right), \quad (16)$$

where τ is the temperature hyperparameter. InfoNCE promotes cross-modal alignment, while the information bottleneck suppresses modality-specific information. Thus, the joint objective $\mathcal{L}_s = \mathcal{L}_{NCE} + \mathcal{L}_{KL}$ encourages the latent representations to preserve shared semantics, such that their similarity can serve as an indicator of cross-modal semantic consistency.

Semantic consistency gating. To dynamically adjust the debiasing penalty, we derive the item-wise gating vector $G \in (0, 1)^N$ from the latent cross-modal similarity as follows:

$$G = \mathbf{1} - \sigma \left(\frac{\text{diag}(\mathbf{Z}_v \mathbf{Z}_t^\top)}{\tau_g} \right), \quad (17)$$

where $\text{diag}(\cdot)$ extracts the diagonal elements as an N -dimensional vector, $\mathbf{1} \in \mathbb{R}^N$ is the all-ones vector, τ_g is a scaling factor, and $\sigma(\cdot)$ is the element-wise sigmoid function. As $\mathbf{Z}_v$ and $\mathbf{Z}_t$ are L_2 -normalized, the diagonal entries represent item-wise cosine similarities. A higher semantic similarity produces a smaller gating value and reduces the bias penalty. In contrast, a lower semantic similarity produces a larger gating value and strengthens the penalty.

3.5. Adaptive debiasing and joint optimization

To mitigate the influence of cross-modal bias on collaborative preference prediction, we integrate the above modules into an adaptive debiasing framework. Specifically, EBN aggregates modality-specific deceptive evidence into a fused cross-modal bias score S_i , while SBN evaluates the cross-modal semantic consistency and generates a gating signal G_i . The final debiased prediction score $\hat{y}_{u,i}$ is obtained by subtracting a bias penalty proportional to the magnitude of the base collaborative score as follows:

$$\hat{y}_{u,i} = \bar{y}_{u,i} - \gamma \cdot |\bar{y}_{u,i}| \cdot G_i \cdot S_i, \quad (18)$$

where γ controls the overall debiasing strength, and $|\bar{y}_{u,i}|$ scales the penalty according to the magnitude of the base collaborative score. To optimize ranking performance, we employ the standard Bayesian Personalized Ranking (BPR) loss (Rendle et al., 2012), which encourages the debiased prediction score of an observed positive item i to exceed that of a sampled unobserved item j . The objective over the set of training triplets $\mathcal{O}$ is formulated as follows:

$$\mathcal{L}_{main} = - \sum_{(u,i,j) \in \mathcal{O}} \ln \sigma(\hat{y}_{u,i} - \hat{y}_{u,j}). \quad (19)$$

Finally, the overall training objective unifies the main recommendation task, the evidential bias supervision, and the semantic bottleneck regularization, which can be formulated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{main} + \lambda_e \mathcal{L}_e + \lambda_s \mathcal{L}_s, \quad (20)$$

where λ_e and λ_s are the balancing hyperparameters. Furthermore, we mitigate gradient interference between preference learning and bottleneck compression by using the Projecting Conflicting Gradients (PC-Grad) method (Yu et al., 2020) to decouple the optimization. Specifically, the overall objective is divided into two primary tasks: the main

Table 2

Statistics of the experimental datasets. Sparsity is calculated as $1 - \frac{|R|}{|U| \times |I|}$, with $|U|$, $|I|$, and $|R|$ representing the total number of users, items, and interactions, respectively; d_v and d_t denote the dimensions of visual and textual modalities.

Dataset	#Users	#Items	#Interactions	Sparsity	d_v	d_t
Baby	19,445	7,050	160,792	99.88%	4096	384
Sports	35,598	18,357	296,337	99.95%	4096	384
Clothing	39,387	23,033	278,677	99.97%	4096	384

task with bias penalties ($\mathcal{L}_{main}^* = \mathcal{L}_{main} + \lambda_e \mathcal{L}_e$) and the SBN task ($\mathcal{L}_s^* = \lambda_s \mathcal{L}_s$). Let $\mathbf{g}_m = \nabla_{\theta} \mathcal{L}_{main}^*$ and $\mathbf{g}_s = \nabla_{\theta} \mathcal{L}_s^*$ denote their respective gradient vectors for the shared parameters θ . When a conflict arises ($\mathbf{g}_s \cdot \mathbf{g}_m < 0$), PCGrad dynamically projects the SBN gradient onto the normal plane of the main task gradient as follows:

$$\tilde{\mathbf{g}}_s = \mathbf{g}_s - \frac{\mathbf{g}_s \cdot \mathbf{g}_m}{\|\mathbf{g}_m\|_2^2} \mathbf{g}_m, \quad (21)$$

where $\|\cdot\|_2^2$ denotes the squared L_2 -norm. Otherwise, when $\mathbf{g}_s \cdot \mathbf{g}_m \geq 0$, the original gradient is preserved as $\tilde{\mathbf{g}}_s = \mathbf{g}_s$. The final parameter update $\theta \leftarrow \theta - \eta_t(\mathbf{g}_m + \tilde{\mathbf{g}}_s)$ mitigates gradient interference while preserving effective SBN updates for shared-semantic learning.

4. Experiments

To comprehensively evaluate CEDRec, we conduct extensive experiments to answer the following questions:

- **RQ1:** How does CEDRec perform compared with state-of-the-art methods?
- **RQ2:** How do the key components and design choices affect CEDRec?
- **RQ3:** How sensitive is CEDRec to different hyperparameter settings?
- **RQ4:** What are the computational efficiency and resource overhead of CEDRec?
- **RQ5:** How well does CEDRec generalize to multimodal recommendation scenarios beyond e-commerce?

4.1. Experimental setup

4.1.1. Datasets

We mainly evaluate CEDRec on three subsets of the Amazon Review dataset¹: (a) Baby, (b) Sports and Outdoors (abbreviated as Sports), and (c) Clothing, Shoes and Jewelry (abbreviated as Clothing). Following prior research (Zhou & Shen, 2023; Zhou et al., 2023), we apply a 5-core filter to ensure data quality, which filters out users and items with fewer than five interactions. The historical interactions are then randomly divided into training, validation, and testing sets according to a ratio of 8:1:1. Table 2 details the statistics of the processed datasets.

In addition, we represent the visual and textual content of each item using pre-extracted embeddings (Zhou, 2023). Specifically, we utilize 384-dimensional text embeddings from Sentence-Transformers (Reimers & Gurevych, 2019) and 4096-dimensional visual embeddings from pre-trained VGG16 (Simonyan & Zisserman, 2014).

4.1.2. Evaluation metrics

To evaluate top- k recommendation performance, we employ two widely adopted metrics: Recall@ K and Normalized Discounted Cumulative Gain (NDCG@ K), denoted as R@ K and N@ K , respectively (Zhou et al., 2023). Specifically, R@ K measures the fraction of ground-truth items successfully retrieved in the recommendation list. A higher R@ K indicates broader coverage of the user's actual interests. In contrast, N@ K evaluates the ranking quality by assigning more weight to relevant items that appear higher in the recommendation list. A higher N@ K signifies that relevant items are ranked closer to the top.

Beyond recommendation accuracy, we follow prior studies (Ge et al., 2021; Shang et al., 2024) and adopt the Gini index (denoted as $Gini@K$) as a complementary debiasing metric. It captures the concentration of item exposure in the recommendation results. Cross-modal deceptive signals may increase the recommendation scores of certain items, leading to disproportionate exposure in the top- k recommendation lists. Therefore, we use $Gini@K$ to examine whether CEDRec alleviates such exposure concentration while preserving recommendation accuracy. Given the aggregated top- k recommendation results of all users, denoted by C_K , we sort all items in ascending order according to their recommendation frequency. Let i_x denote the x -th least recommended item and $p(i_x | C_K)$ denote its proportion in C_K . The Gini index is formulated as follows:

$$Gini@K = \frac{1}{|I| - 1} \sum_{x=1}^{|I|} (2x - |I| - 1) p(i_x | C_K), \quad (22)$$

where $|I|$ denotes the total number of items. A lower $Gini@K$ indicates a more uniform item exposure distribution, meaning that the recommendation results are less dominated by a small group of frequently recommended items. In the experiments, we report the results for $K \in \{10, 20\}$.

4.1.3. Baselines

For a comprehensive performance comparison, we evaluate fifteen representative baselines, including two traditional recommendation methods, eleven multimodal recommendation methods, and two debiasing methods:

Traditional recommendation methods: These methods rely on historical user-item interactions to capture collaborative signals, serving as fundamental baselines without utilizing multimodal side features.

- **MF-BPR** (Rendle et al., 2012) learns user and item representations through matrix factorization and optimizes them using the BPR pairwise ranking objective.
- **LightGCN** (He et al., 2020) simplifies GCN-based collaborative filtering by removing redundant operations, making it better suited for recommendation tasks.

Multimodal recommendation methods: These methods incorporate auxiliary modalities alongside interaction data to enrich item representations and improve recommendation accuracy.

- **BM3** (Zhou et al., 2023) simplifies self-supervised recommendation via latent embedding dropout while aligning inter- and intra-modality features without auxiliary graphs.
- **FREEDOM** (Zhou & Shen, 2023) achieves efficient multimodal recommendation by freezing the item-item graph and denoising the user-item interaction graph through degree-sensitive edge pruning.
- **LGMRec** (Guo et al., 2024) jointly models local topological interactions and global hypergraph relations to extract decoupled user interests for robust multimodal recommendation.
- **HPMRec** (Chen et al., 2025b) employs hypercomplex multi-component embeddings and nonlinear cross-modal multiplication with prompt-aware compensation to enhance representation diversity and alleviate over-smoothing.
- **COHESION** (Xu et al., 2025b) combines dual-stage modality fusion with composite graph convolution to model user-item, user-user, and item-item relations while reducing irrelevant multimodal information.
- **SMORE** (Ong & Khong, 2025) performs spectrum-based modality fusion with adaptive filtering to suppress modality-specific noise while jointly modeling unimodal and fusion preferences.
- **MENTOR** (Xu et al., 2025c) performs ID-guided multi-level cross-modal alignment over heterogeneous and homogeneous graphs to reduce modality gaps while preserving interaction information.
- **HEARec** (Shi et al., 2025) constructs Dirichlet-based heterogeneous multimodal environments and aligns modality embeddings in a shared hypergraph space to improve robustness against missing modalities.

¹ <http://jmcauley.ucsd.edu/data/amazon/links.html>

- **PGL** (Yu et al., 2025) extracts principal subgraphs with global- and local-aware operators to capture local structural features and preserve individual preference information.
- **SRGFormer** (Shi et al., 2026) combines multi-head self-attention with self-supervised hypergraph learning to jointly capture global user behavior patterns and local user-item structural relations.
- **DuGRec** (Ma et al., 2026) reconstructs item-item and user-item graphs with multi-view structural priors and employs contrastive learning to denoise graph relations and mitigate popularity bias.

Debiasing methods: These methods explicitly identify and remove modality-related bias from recommendation predictions.

- **EliMRec** (Liu et al., 2022) uses causal inference to capture unimodal bias and removes its direct effect from the prediction score through counterfactual reasoning.
- **M-Debias** (Shang et al., 2024) introduces unimodal prediction branches to estimate modality bias and adaptively removes the bias through fairness-aware counterfactual inference.

4.1.4. Implementation details

All models are implemented in PyTorch and evaluated on an NVIDIA RTX 5070 Ti GPU with 16 GB of memory. The trainable embedding matrices and network weight matrices are initialized using the Xavier initializer (Glorot & Bengio, 2010), and all models are optimized using the Adam optimizer (Kingma & Ba, 2014). The latent dimension d for the item ID embeddings and the modality-specific user preference embeddings P_m is set to 64. For graph construction and prior bias quantification, the number of neighbors k is fixed at 10 for the top- k sparsified homogeneous graphs (user-user and item-item) and the initial unimodal neighborhood distributions. The maximum number of training epochs is set to 1000, with early stopping using a patience of 50 epochs based on Recall@20 on the validation set.

The hyperparameters of CEDRec are configured as follows. The semantic bottleneck weight λ_s is set to 1.0 for all datasets. The bias supervision weight λ_e is set to 0.2 for Baby, 0.1 for Sports, and 0.4 for Clothing, respectively. The temperature τ in the InfoNCE objective is tuned to 0.1 for Baby and Sports, and 0.05 for Clothing. The learning rate η_r is set to $1e^{-3}$ for Baby and Clothing, and $1e^{-4}$ for Sports. The debiasing penalty strength γ is set to 0.2 for all datasets. Moreover, the sensitivity of some hyperparameters is investigated in Section 4.4.

4.2. Performance comparison (RQ1)

We evaluate CEDRec in two comparison settings. First, we compare it with representative traditional and multimodal recommendation methods, evaluating both ranking accuracy and item exposure concentration. Second, we compare it with existing debiasing methods, with all methods using the same base graph network to control for differences in backbone architectures. All results are reported as the mean $\pm$ standard deviation over five runs.

(1) **Comparison with recommendation methods.** Table 3 presents the overall results across three datasets. Regarding ranking accuracy, all multimodal methods outperform the two traditional recommendation methods in terms of Recall and NDCG, suggesting the benefits of incorporating visual and textual information into item representation learning. Among these multimodal methods, DuGRec achieves the highest Recall and NDCG results on all three datasets, making it the strongest accuracy-oriented baseline for comparison with CEDRec.

Turning to item exposure concentration, we further evaluate the distribution of recommended items using the Gini index. In this respect, CEDRec achieves the lowest Gini values on all three datasets, while PGL performs best among the competing baselines and therefore serves as the strongest exposure-oriented baseline. Specifically, compared with PGL, CEDRec achieves relative reductions of 2.64%, 1.84%, and 3.76%

in Gini@20 on Baby, Sports, and Clothing, respectively. Similar reductions are also observed for Gini@10, with the relative reduction reaching 4.33% on Clothing. Overall, these results indicate that CEDRec effectively reduces exposure concentration while maintaining competitive ranking accuracy, thus achieving a favorable balance between debiasing effectiveness and recommendation quality.

(2) **Comparison with debiasing methods.** Table 4 compares CEDRec with two debiasing methods under a controlled setting in which all methods use the same base graph network. Under this setting, CEDRec achieves the best result for every metric on each of the three datasets. Specifically, relative to the best competing result for each metric, CEDRec improves R@20 by 2.31%, 1.24%, and 3.19% on Baby, Sports, and Clothing, respectively, while the corresponding improvements in N@20 are 2.43%, 1.98%, and 4.75%. In terms of item exposure concentration, CEDRec further achieves relative reductions of 1.65%, 1.32%, and 2.28% in Gini@20 on the three datasets, respectively. Overall, these results indicate that CEDRec's advantages persist after controlling for backbone architecture, supporting the effectiveness of its adaptive debiasing design.

4.3. Ablation study (RQ2)

We conduct a series of ablation studies to examine the contributions of the key components and design choices in CEDRec. Specifically, we analyze the effects of the main model components and optimization strategy, cross-modal fusion strategy, prior proxy construction, and semantic consistency modeling. The results are shown in Figs. 3–6 and Tables 5 and 6.

4.3.1. Effect of model components and optimization strategy

To evaluate the contributions of EBN and SBN, the score scaling design, and the PCGrad optimization strategy, we compare CEDRec with the following variants:

- **CEDRec (w/o EBN & SBN)** removes both debiasing modules and retains only the base graph network.
- **CEDRec (w/o SBN)** removes SBN and sets G to an all-ones vector, leaving the penalty determined by S .
- **CEDRec (w/o EBN)** removes EBN and sets S to an all-ones vector, leaving the penalty determined by G .
- **CEDRec (w/o Scaling)** removes the score scaling factor $|\tilde{y}_{u,i}|$, leaving the penalty modulated by G and S .
- **CEDRec (w/o PCG)** removes the PCGrad strategy and directly performs joint optimization.

(1) **Component ablation analysis.** The results are shown in Fig. 3. Overall, the full CEDRec achieves the best results across all three datasets. Compared with CEDRec (w/o EBN & SBN), it improves NDCG@20 by 3.58%, 3.82%, and 5.50% on Baby, Sports, and Clothing, respectively, while reducing Gini@20 by 2.83%, 2.51%, and 3.84%. For the individual modules, CEDRec (w/o SBN) consistently outperforms the base model, suggesting that EBN provides useful item-specific signals for adaptive debiasing. Its lower performance than the full model further supports the benefit of SBN in adjusting the penalty based on semantic consistency. In contrast, CEDRec (w/o EBN) performs close to the base model, slightly worse on Baby and Sports but better on Clothing. This suggests that semantic consistency alone is insufficient to determine the penalty strength reliably. Moreover, CEDRec (w/o Scaling) consistently performs worse than the full model, confirming the benefit of scaling the debiasing penalty by the base matching score. Without this scaling, candidates with different score magnitudes receive penalties on the same absolute scale, potentially causing disproportionate corrections and poorer rankings. Finally, CEDRec (w/o PCG) performs worse than the full model on both metrics, confirming that resolving gradient conflicts benefits the joint optimization of recommendation and semantic bottleneck learning.

Table 3

Performance comparison (%) across three datasets. **Bold** and underlined denote the best and second-best results.

Dataset	Model	$R@10(\uparrow)$	$R@20(\uparrow)$	$N@10(\uparrow)$	$N@20(\uparrow)$	$Gini@10(\downarrow)$	$Gini@20(\downarrow)$
Baby	MF-BPR	3.57 $\pm$ 0.05	5.75 $\pm$ 0.10	1.92 $\pm$ 0.07	2.49 $\pm$ 0.08	89.40 $\pm$ 0.04	89.07 $\pm$ 0.08
	LightGCN	4.79 $\pm$ 0.09	7.54 $\pm$ 0.12	2.57 $\pm$ 0.09	3.28 $\pm$ 0.10	89.81 $\pm$ 0.06	89.72 $\pm$ 0.06
	BM3	5.64 $\pm$ 0.04	8.83 $\pm$ 0.07	3.01 $\pm$ 0.10	3.83 $\pm$ 0.07	88.93 $\pm$ 0.04	88.47 $\pm$ 0.08
	FREEDOM	6.27 $\pm$ 0.07	9.92 $\pm$ 0.02	3.30 $\pm$ 0.03	4.24 $\pm$ 0.02	87.39 $\pm$ 0.09	86.80 $\pm$ 0.04
	LGMRec	6.39 $\pm$ 0.02	9.89 $\pm$ 0.01	3.37 $\pm$ 0.08	4.30 $\pm$ 0.01	88.22 $\pm$ 0.03	87.69 $\pm$ 0.05
	HPMRec	6.67 $\pm$ 0.04	10.35 $\pm$ 0.10	3.56 $\pm$ 0.07	4.51 $\pm$ 0.03	87.56 $\pm$ 0.03	86.94 $\pm$ 0.04
	COHESION	6.72 $\pm$ 0.08	10.43 $\pm$ 0.02	3.44 $\pm$ 0.08	4.48 $\pm$ 0.09	88.23 $\pm$ 0.05	87.70 $\pm$ 0.03
	SMORE	6.71 $\pm$ 0.06	10.21 $\pm$ 0.07	3.59 $\pm$ 0.11	4.49 $\pm$ 0.04	84.81 $\pm$ 0.08	84.16 $\pm$ 0.13
	MENTOR	6.68 $\pm$ 0.06	10.29 $\pm$ 0.05	3.53 $\pm$ 0.04	4.53 $\pm$ 0.07	87.80 $\pm$ 0.06	87.35 $\pm$ 0.04
	HEARec	6.69 $\pm$ 0.05	10.21 $\pm$ 0.04	3.63 $\pm$ 0.09	4.54 $\pm$ 0.02	87.07 $\pm$ 0.03	87.03 $\pm$ 0.05
	PGL	6.77 $\pm$ 0.01	10.17 $\pm$ 0.01	3.54 $\pm$ 0.02	4.47 $\pm$ 0.05	<u>79.73 $\pm$ 0.05</u>	<u>78.91 $\pm$ 0.06</u>
	SRGFormer	6.74 $\pm$ 0.03	10.32 $\pm$ 0.05	3.64 $\pm$ 0.05	4.58 $\pm$ 0.09	88.81 $\pm$ 0.02	88.42 $\pm$ 0.05
	DuGRec	7.18 $\pm$ 0.02	10.84 $\pm$ 0.04	3.94 $\pm$ 0.06	4.88 $\pm$ 0.06	81.10 $\pm$ 0.05	80.06 $\pm$ 0.03
	CEDRec	<u>6.94 $\pm$ 0.05</u>	<u>10.62 $\pm$ 0.02</u>	<u>3.68 $\pm$ 0.08</u>	<u>4.63 $\pm$ 0.03</u>	77.27 $\pm$ 0.02	76.83 $\pm$ 0.06
Sports	MF-BPR	4.32 $\pm$ 0.03	6.53 $\pm$ 0.04	2.41 $\pm$ 0.08	2.98 $\pm$ 0.03	89.09 $\pm$ 0.07	88.41 $\pm$ 0.02
	LightGCN	5.69 $\pm$ 0.05	8.62 $\pm$ 0.03	3.11 $\pm$ 0.02	3.87 $\pm$ 0.04	89.58 $\pm$ 0.08	88.95 $\pm$ 0.05
	BM3	6.56 $\pm$ 0.08	9.80 $\pm$ 0.02	3.55 $\pm$ 0.06	4.38 $\pm$ 0.05	88.03 $\pm$ 0.08	87.44 $\pm$ 0.03
	FREEDOM	7.17 $\pm$ 0.03	10.89 $\pm$ 0.05	3.85 $\pm$ 0.07	4.81 $\pm$ 0.07	86.92 $\pm$ 0.06	86.30 $\pm$ 0.07
	LGMRec	7.19 $\pm$ 0.03	10.68 $\pm$ 0.03	3.87 $\pm$ 0.01	4.77 $\pm$ 0.05	87.39 $\pm$ 0.09	86.71 $\pm$ 0.03
	HPMRec	7.54 $\pm$ 0.03	11.27 $\pm$ 0.06	4.02 $\pm$ 0.04	4.94 $\pm$ 0.05	87.07 $\pm$ 0.05	86.46 $\pm$ 0.08
	COHESION	7.51 $\pm$ 0.09	11.30 $\pm$ 0.02	3.94 $\pm$ 0.04	4.97 $\pm$ 0.08	87.11 $\pm$ 0.05	86.59 $\pm$ 0.02
	SMORE	7.48 $\pm$ 0.06	11.22 $\pm$ 0.06	4.05 $\pm$ 0.05	4.97 $\pm$ 0.04	84.35 $\pm$ 0.03	83.61 $\pm$ 0.03
	MENTOR	7.51 $\pm$ 0.07	11.34 $\pm$ 0.02	4.06 $\pm$ 0.01	5.03 $\pm$ 0.09	86.68 $\pm$ 0.04	86.02 $\pm$ 0.09
	HEARec	7.66 $\pm$ 0.03	11.48 $\pm$ 0.05	4.15 $\pm$ 0.04	5.20 $\pm$ 0.08	86.87 $\pm$ 0.02	86.40 $\pm$ 0.02
	PGL	<u>7.82 $\pm$ 0.09</u>	<u>11.68 $\pm$ 0.09</u>	<u>4.26 $\pm$ 0.07</u>	<u>5.29 $\pm$ 0.08</u>	<u>80.07 $\pm$ 0.02</u>	<u>79.28 $\pm$ 0.06</u>
	SRGFormer	7.59 $\pm$ 0.03	11.44 $\pm$ 0.03	4.10 $\pm$ 0.04	5.11 $\pm$ 0.02	86.92 $\pm$ 0.02	86.27 $\pm$ 0.07
	DuGRec	8.24 $\pm$ 0.06	12.13 $\pm$ 0.08	4.56 $\pm$ 0.03	5.58 $\pm$ 0.05	81.04 $\pm$ 0.03	80.49 $\pm$ 0.06
	CEDRec	7.68 $\pm$ 0.07	11.47 $\pm$ 0.08	4.18 $\pm$ 0.06	5.16 $\pm$ 0.07	78.65 $\pm$ 0.04	77.82 $\pm$ 0.08
Clothing	MF-BPR	1.99 $\pm$ 0.05	2.98 $\pm$ 0.05	1.05 $\pm$ 0.09	1.34 $\pm$ 0.05	82.97 $\pm$ 0.07	82.41 $\pm$ 0.02
	LightGCN	2.98 $\pm$ 0.09	4.49 $\pm$ 0.04	1.67 $\pm$ 0.03	2.05 $\pm$ 0.06	83.20 $\pm$ 0.09	82.73 $\pm$ 0.05
	BM3	4.22 $\pm$ 0.09	6.21 $\pm$ 0.04	2.31 $\pm$ 0.04	2.81 $\pm$ 0.09	82.43 $\pm$ 0.06	81.88 $\pm$ 0.05
	FREEDOM	6.28 $\pm$ 0.08	9.41 $\pm$ 0.04	3.41 $\pm$ 0.02	4.20 $\pm$ 0.04	80.64 $\pm$ 0.03	79.81 $\pm$ 0.07
	LGMRec	5.55 $\pm$ 0.05	8.28 $\pm$ 0.07	3.02 $\pm$ 0.03	3.71 $\pm$ 0.04	81.72 $\pm$ 0.05	81.49 $\pm$ 0.08
	HPMRec	6.46 $\pm$ 0.03	9.61 $\pm$ 0.09	3.48 $\pm$ 0.01	4.24 $\pm$ 0.03	80.84 $\pm$ 0.04	80.05 $\pm$ 0.07
	COHESION	6.54 $\pm$ 0.02	9.79 $\pm$ 0.07	3.56 $\pm$ 0.04	4.38 $\pm$ 0.02	82.03 $\pm$ 0.03	81.08 $\pm$ 0.03
	SMORE	6.52 $\pm$ 0.02	9.69 $\pm$ 0.04	3.53 $\pm$ 0.02	4.36 $\pm$ 0.06	77.62 $\pm$ 0.07	76.77 $\pm$ 0.09
	MENTOR	6.61 $\pm$ 0.04	9.85 $\pm$ 0.03	3.60 $\pm$ 0.07	4.45 $\pm$ 0.08	81.27 $\pm$ 0.03	80.20 $\pm$ 0.06
	HEARec	5.98 $\pm$ 0.05	8.87 $\pm$ 0.06	3.35 $\pm$ 0.04	4.04 $\pm$ 0.08	80.46 $\pm$ 0.04	79.37 $\pm$ 0.07
	PGL	<u>7.04 $\pm$ 0.02</u>	<u>10.18 $\pm$ 0.02</u>	<u>3.79 $\pm$ 0.03</u>	<u>4.59 $\pm$ 0.05</u>	<u>71.64 $\pm$ 0.07</u>	<u>70.69 $\pm$ 0.05</u>
	SRGFormer	6.02 $\pm$ 0.05	8.78 $\pm$ 0.08	3.31 $\pm$ 0.02	4.01 $\pm$ 0.03	82.86 $\pm$ 0.08	81.72 $\pm$ 0.03
	DuGRec	7.22 $\pm$ 0.04	10.42 $\pm$ 0.07	3.94 $\pm$ 0.03	4.73 $\pm$ 0.02	74.30 $\pm$ 0.04	72.94 $\pm$ 0.07
	CEDRec	6.81 $\pm$ 0.03	10.04 $\pm$ 0.06	3.78 $\pm$ 0.05	4.41 $\pm$ 0.09	68.54 $\pm$ 0.03	68.03 $\pm$ 0.03

Table 4

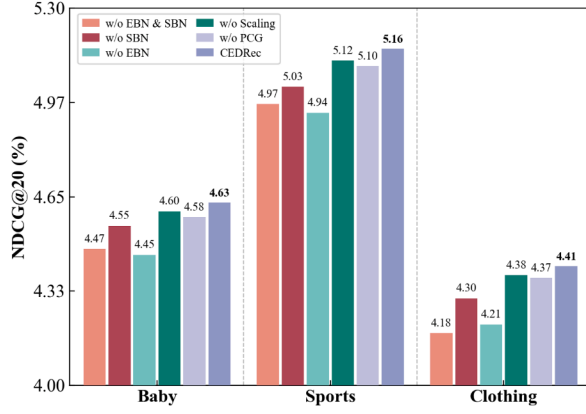
Performance comparison (%) with debiasing methods across three datasets using the same base graph network. **Bold** denotes the best result.

Dataset	Debiasing Method	$R@10(\uparrow)$	$R@20(\uparrow)$	$N@10(\uparrow)$	$N@20(\uparrow)$	$Gini@10(\downarrow)$	$Gini@20(\downarrow)$
Baby	EliMRec	6.77 $\pm$ 0.02	10.38 $\pm$ 0.05	3.51 $\pm$ 0.07	4.43 $\pm$ 0.12	80.35 $\pm$ 0.05	79.71 $\pm$ 0.06
	M-Debias	6.76 $\pm$ 0.03	10.30 $\pm$ 0.05	3.59 $\pm$ 0.02	4.52 $\pm$ 0.09	78.82 $\pm$ 0.05	78.12 $\pm$ 0.08
	CEDRec	6.94 $\pm$ 0.05	10.62 $\pm$ 0.02	3.68 $\pm$ 0.08	4.63 $\pm$ 0.03	77.27 $\pm$ 0.02	76.83 $\pm$ 0.06
Sports	EliMRec	7.54 $\pm$ 0.05	11.33 $\pm$ 0.06	4.02 $\pm$ 0.13	5.02 $\pm$ 0.02	80.74 $\pm$ 0.04	80.03 $\pm$ 0.02
	M-Debias	7.51 $\pm$ 0.03	11.27 $\pm$ 0.04	4.07 $\pm$ 0.04	5.06 $\pm$ 0.09	79.62 $\pm$ 0.07	78.86 $\pm$ 0.08
	CEDRec	7.68 $\pm$ 0.07	11.47 $\pm$ 0.08	4.18 $\pm$ 0.06	5.16 $\pm$ 0.07	78.65 $\pm$ 0.04	77.82 $\pm$ 0.08
Clothing	EliMRec	6.54 $\pm$ 0.02	9.73 $\pm$ 0.06	3.59 $\pm$ 0.07	4.21 $\pm$ 0.09	72.08 $\pm$ 0.02	71.35 $\pm$ 0.03
	M-Debias	6.31 $\pm$ 0.05	9.46 $\pm$ 0.09	3.51 $\pm$ 0.09	4.10 $\pm$ 0.04	70.31 $\pm$ 0.06	69.62 $\pm$ 0.10
	CEDRec	6.81 $\pm$ 0.03	10.04 $\pm$ 0.06	3.78 $\pm$ 0.05	4.41 $\pm$ 0.09	68.54 $\pm$ 0.03	68.03 $\pm$ 0.03

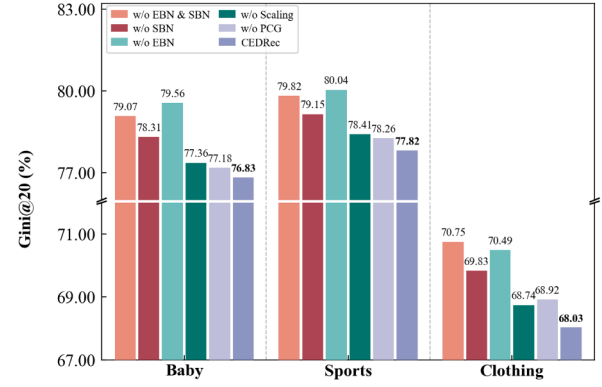
(2) **Dependency analysis between bias score S and gating signal G .** We examine the statistical dependence between S and G using Spearman correlation ρ and Normalized Mutual Information (NMI). Spearman ρ measures their monotonic rank association, while NMI captures more general dependence. For NMI computation, the continuous item-level outputs are discretized into 20 equal-frequency bins. The mutual information is then normalized by the average entropy. Table 5 reports Spear-

man correlations from 0.29 to 0.49 and NMI values from 0.10 to 0.23. These results indicate that the two signals are related but not redundant. EBN estimates the item-specific bias intensity. SBN adjusts the penalty according to semantic consistency. Therefore, their joint use provides complementary signals for more reliable adaptive debiasing.

(3) **Gradient conflict analysis with PCGrad.** To examine whether PCGrad persistently suppresses SBN, we track the cosine similarity be-



(a) NDCG@20



(b) Gini@20

Fig. 3. Ablation results of the key modules and optimization strategy across three datasets.

Table 5

Statistical dependence between the EBN output S and the SBN output G . Spearman ρ measures monotonic association, while NMI measures general statistical dependence. Larger $|\rho|$ and NMI indicate stronger dependence.

Dataset	Spearman $\rho(S, G)$	NMI(S, G)
Baby	0.398	0.174
Sports	0.286	0.103
Clothing	0.488	0.229

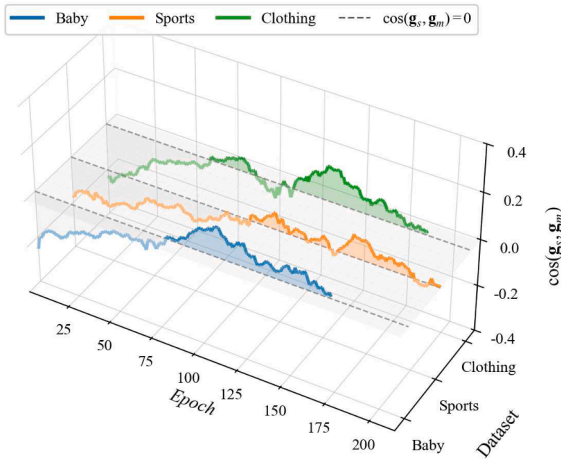


Fig. 4. Cosine similarity between the main-task gradient $\mathbf{g}_m$ and the SBN gradient $\mathbf{g}_s$ during training across three datasets. The dashed line indicates $\cos(\mathbf{g}_m, \mathbf{g}_s) = 0$.

tween the main-task gradient $\mathbf{g}_m$ and the SBN gradient $\mathbf{g}_s$ during training. For each epoch, we compute the similarity for each batch and report its average value over all batches. As shown in Fig. 4, the average gradient similarities vary across epochs on all three datasets. Specifically, positive values occur in many epochs, indicating aligned update directions, while negative values appear at some stages, indicating occasional gradient conflicts. Overall, PCGrad only adjusts gradients when conflicts arise and does not continuously suppress SBN updates.

4.3.2. Effect of cross-modal fusion strategies

To examine the effect of different cross-modal fusion strategies, we replace the evidential fusion in EBN while keeping the remaining components unchanged. Specifically, we compare CEDRec with the following variants:

- **CEDRec (w/ Max)** fuses the visual and textual deceptive evidence scores using their maximum.
- **CEDRec (w/ Avg.)** fuses the visual and textual deceptive evidence scores using their arithmetic mean.
- **CEDRec (w/ WS)** fuses the visual and textual deceptive evidence scores using their weighted sum.

As shown in Fig. 5, the proposed evidential fusion achieves the highest NDCG@20 and the lowest Gini@20 across all three datasets. CEDRec (w/ Avg.) and CEDRec (w/ WS) consistently yield weaker results, indicating that linear fusion may dilute strong modality-specific deceptive evidence when the two modalities provide imbalanced signals. CEDRec (w/ Max) avoids such dilution by preserving the stronger unimodal evidence, but cannot incorporate complementary evidence from the other modality. In contrast, the proposed fusion preserves strong deceptive evidence from either modality while accumulating complementary evidence when both modalities indicate potential bias. This design provides a more effective bias signal for determining the debiasing penalty, thus improving ranking accuracy and reducing item exposure concentration.

4.3.3. Effect of prior proxy construction

To investigate the effect of the prior proxy for bias supervision, we vary its supervision and construction strategies while keeping the remaining model settings unchanged. Specifically, we compare CEDRec with the following variants:

- **CEDRec (w/o Prior)** removes the prior-guided regularization on the fused cross-modal bias score.
- **CEDRec (w/o JS)** removes the JS divergence and uses only item popularity as the prior proxy.
- **CEDRec (w/o Pop)** removes item popularity and uses only the JS divergence as the prior proxy.
- **CEDRec (w/ AM)** replaces the harmonic mean with the arithmetic mean to combine popularity and JS divergence.
- **CEDRec (w/ GM)** replaces the harmonic mean with the geometric mean to combine popularity and JS divergence.

As shown in Fig. 6, CEDRec (w/o Prior) consistently performs worse than CEDRec on all three datasets, demonstrating the importance of prior-guided regularization for constraining the fused cross-modal bias estimation. Furthermore, CEDRec (w/o JS) and CEDRec (w/o Pop) both yield weaker results, indicating that topological disagreement and popularity provide complementary bias-risk information. Regarding the aggregation strategy, CEDRec (w/ AM) and CEDRec (w/ GM) show comparable performance to CEDRec, with our harmonic mean achieving a small but consistent advantage. Overall, the aggregation strategy has only a limited impact, and the harmonic mean provides a simple way to

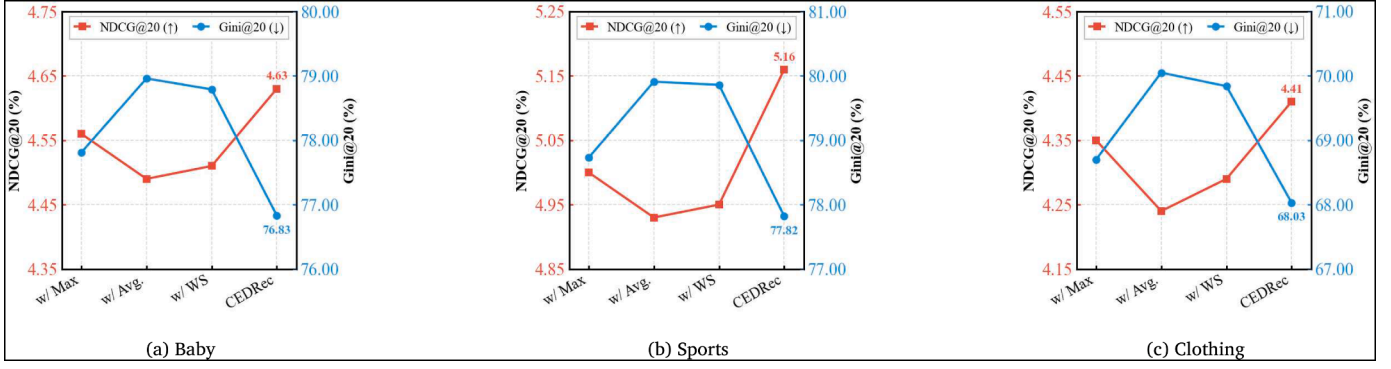


Fig. 5. Comparison of different cross-modal fusion strategies across three datasets.

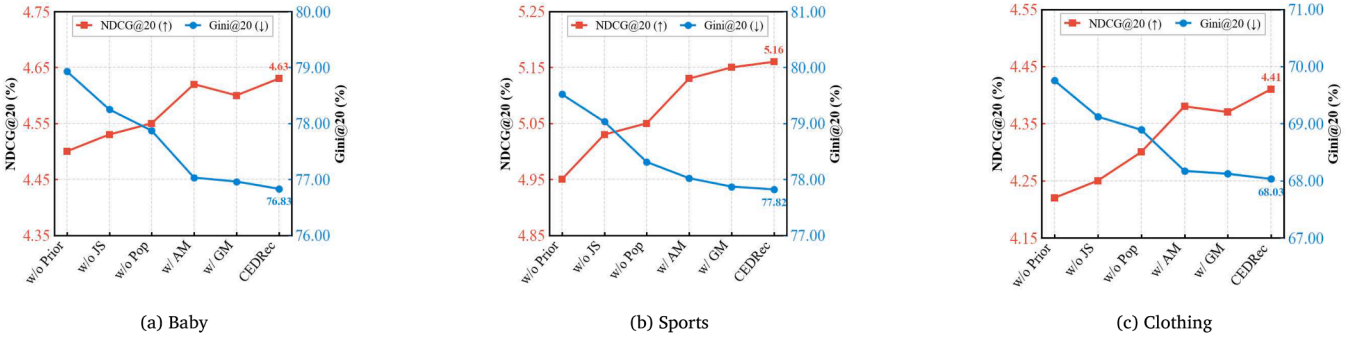


Fig. 6. Comparison of different prior proxy construction strategies across three datasets.

emphasize items with jointly high popularity and topological disagreement.

4.3.4. Effect of semantic consistency modeling

To examine the effect of semantic consistency modeling strategies in SBN, we replace the proposed bottleneck-based representation learning with two simpler alternatives while keeping the remaining model settings unchanged:

- **CEDRec (w/ CosAlign)** directly optimizes the alignment between visual and textual representations using cosine similarity without variational compression.
- **CEDRec (w/ CL)** employs standard contrastive learning to capture cross-modal consistency without the information bottleneck constraint.

As reported in Table 6, CEDRec (w/ CosAlign) yields the weakest overall performance, indicating that direct pairwise alignment alone is insufficient to distinguish shared semantics from modality-specific information. In comparison, CEDRec (w/ CL) performs better than CEDRec (w/ CosAlign), showing that contrastive learning with negative samples provides a stronger signal for cross-modal semantic alignment. Moreover, CEDRec consistently outperforms CEDRec (w/ CL), demonstrating the benefit of introducing the information bottleneck constraint. By suppressing modality-specific information while preserving shared cross-modal semantics, the bottleneck produces a more reliable consistency signal for adjusting the debiasing penalty.

4.4. Parameter sensitivity study (RQ3)

We evaluate the impact of core hyperparameters on CEDRec's recommendation performance. Specifically, we analyze the model's sensitivity

to the weight coefficients (i.e., λ_e and λ_s), the debiasing penalty strength γ , and the learning rate η_r across three datasets. The results are shown in Figs. 7–9.

(1) **Weight coefficients λ_e and λ_s .** As illustrated in Fig. 7, we evaluate the model's sensitivity to the bias supervision weight λ_e and the semantic bottleneck weight λ_s , which are varied over $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ and $\{0.7, 0.8, 0.9, 1.0, 1.1\}$, respectively. The optimal λ_s remains at 1.0 across all three datasets, indicating that a stable bottleneck constraint is essential for distilling shared cross-modal semantics. In contrast, the optimal λ_e varies across datasets, achieving the peak Recall@20 at 0.2, 0.1, and 0.4 on the Baby, Sports, and Clothing datasets, respectively. These results imply that categories like Clothing may contain a higher proportion of cross-modal bias-risk items, thus requiring a larger λ_e to effectively penalize the modality-specific biases.

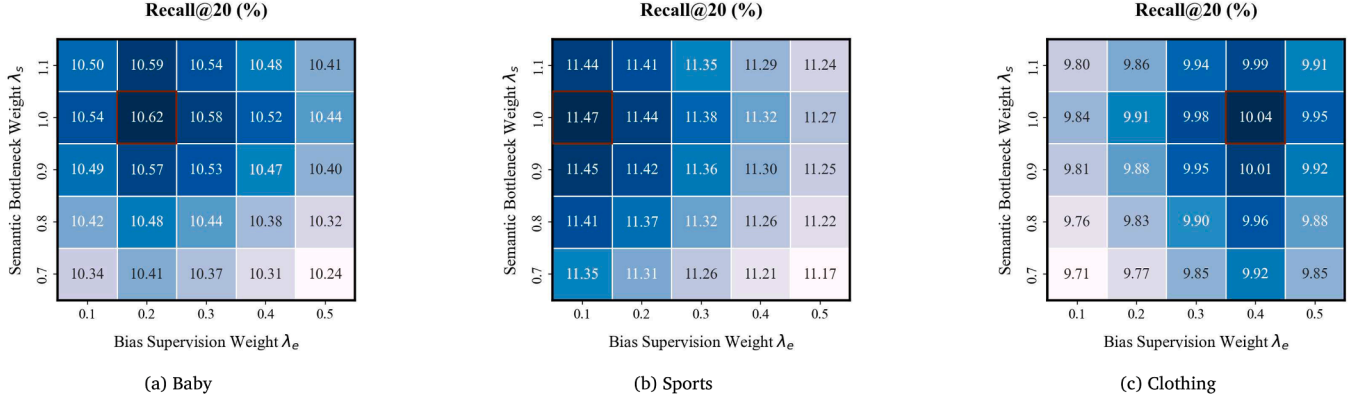
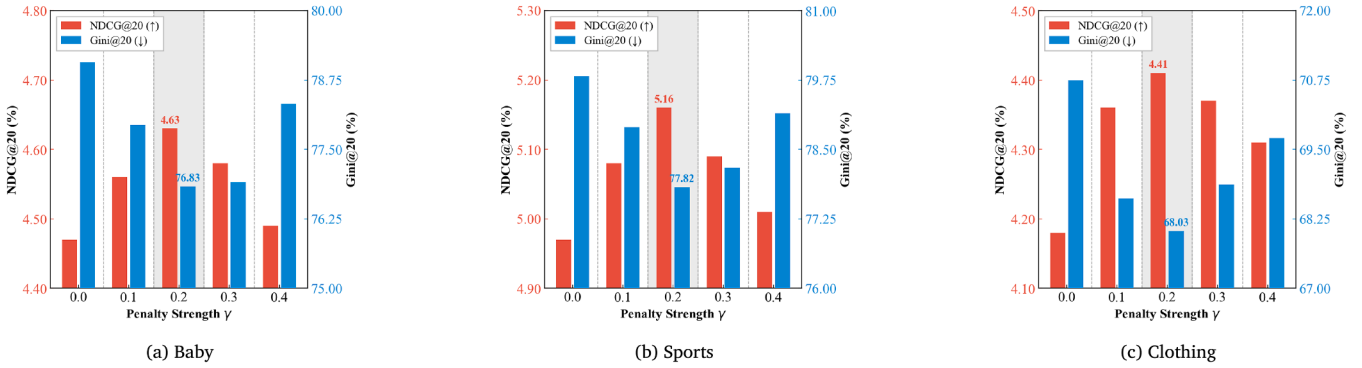
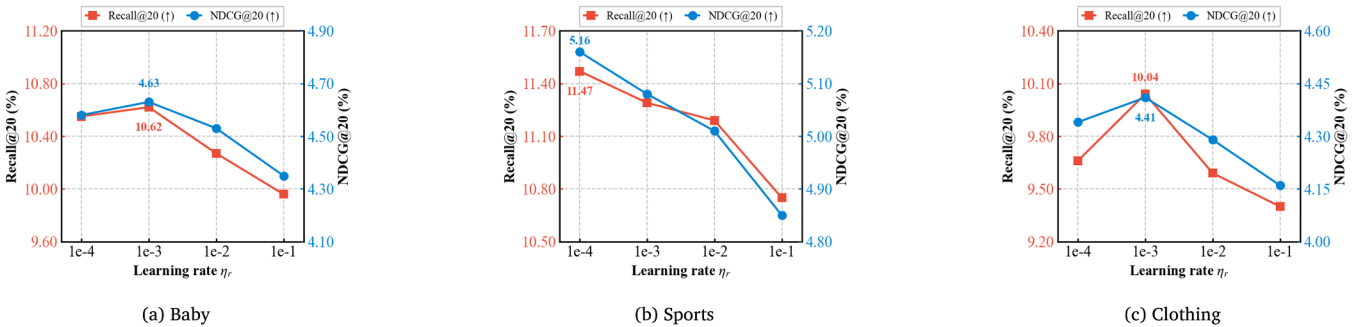
(2) **Debiasing penalty strength γ .** We investigate the impact of γ across $\{0.0, 0.1, 0.2, 0.3, 0.4\}$, with results plotted in Fig. 8, from which we can observe that CEDRec consistently achieves the highest NDCG@20 and the lowest Gini@20 at $\gamma = 0.2$ on all three datasets. At lower settings such as 0.0 and 0.1, insufficient debiasing allows misleading items with exaggerated features to remain under-penalized, reducing ranking quality and leading to more concentrated item exposure. Moreover, further increasing γ beyond 0.2 yields a continuous decrease in NDCG@20 and an increase in Gini@20. This indicates that an excessively large penalty can over-correct the recommendation scores and suppress useful collaborative preference signals.

(3) **Learning rate η_r .** We investigate the impact of the learning rate η_r by varying it within the range of $\{1e^{-4}, 1e^{-3}, 1e^{-2}, 1e^{-1}\}$, with results presented in Fig. 9. It can be observed that CEDRec achieves optimal recommendation performance at smaller learning rates, specifically $1e^{-4}$ on the Sports dataset, and $1e^{-3}$ on the Baby and Clothing datasets. Increasing η_r to $1e^{-2}$ or $1e^{-1}$ causes a sharp decline in terms of both Recall@20 and NDCG@20. This indicates that a large learning rate disrupts the stable joint optimization of the main collaborative ranking task and the

Table 6

Comparison of different semantic consistency modeling strategies across three datasets.

Dataset	Baby		Sports		Clothing	
Metrics	$N@20(\uparrow)$	$Gini@20(\downarrow)$	$N@20(\uparrow)$	$Gini@20(\downarrow)$	$N@20(\uparrow)$	$Gini@20(\downarrow)$
CEDRec (w/ CosAlign)	4.58	77.51	5.07	78.47	4.34	68.97
CEDRec (w/ CL)	4.60	77.04	5.11	78.15	4.37	68.32
CEDRec	4.63	76.83	5.16	77.82	4.41	68.03

**Fig. 7.** Parameter sensitivity of the bias supervision weight λ_e and the semantic bottleneck weight λ_s across three datasets.**Fig. 8.** Parameter sensitivity of the debiasing penalty strength γ .**Fig. 9.** Parameter sensitivity of the learning rate η_r .

semantic bottleneck objective, thus degrading the final recommendation accuracy.

4.5. Efficiency analysis (RQ4)

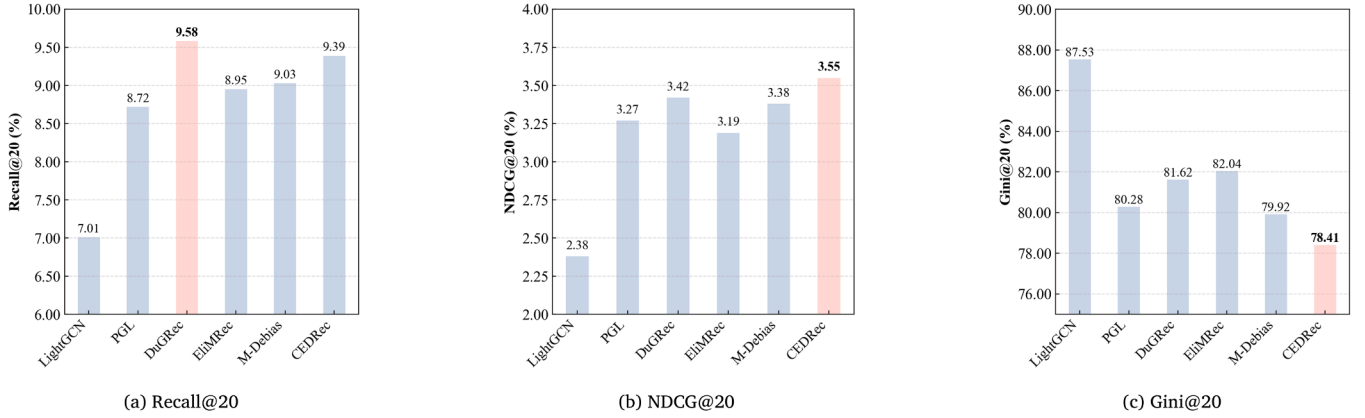
To examine the computational cost of CEDRec, we first analyze the complexity of its main components. Here, M , N , and $|R|$ denote the numbers of users, items, and observed interactions, respectively. The symbols d , L , k , B , and d_z denote the embedding dimension, the

number of graph layers, the number of retained neighbors, the batch size, and the SBN latent dimension, respectively. In the base graph network, message passing over the three user-item graphs has a complexity of $O(L|R|d)$. The subsequent construction of user-user and item-item graphs requires $O((M^2 + N^2)d)$ for pairwise similarity computation, followed by $O(k(M + N)d)$ for aggregation on the sparse top- k graphs. EBN mainly performs item-wise MLP transformations and element-wise evidential fusion, so its cost grows linearly with the number of items. For SBN, the main computational cost comes from the batch-wise In-

Table 7

Efficiency comparison of CEDRec with state-of-the-art baselines. (Training Time: s/epoch; Inference Time: ms/batch; Memory: GB; Parameter: M).

Dataset	Baby				Sports				Clothing			
Metrics	Training	Inference	Memory	Parameter	Training	Inference	Memory	Parameter	Training	Inference	Memory	Parameter
SMORE	6.62	10.58	3.31	33.6	10.37	18.61	5.02	86.0	13.55	22.43	6.89	107.5
MENTOR	7.31	14.21	7.12	44.7	22.28	32.94	8.44	103.3	28.13	38.02	12.99	126.1
HEARec	12.46	30.49	2.42	39.7	43.15	58.75	4.50	97.4	41.35	67.84	4.65	120.8
PGL	5.75	7.96	2.17	34.4	19.41	15.20	3.43	87.1	17.93	17.52	4.37	108.6
SRGFormer	7.82	19.33	2.39	39.7	29.28	55.63	4.63	97.4	26.04	54.68	4.77	120.8
DuGRec	2.04	1.20	0.75	37.7	7.39	2.73	1.48	91.5	6.80	3.27	1.81	113.1
CEDRec	5.72	7.59	1.38	39.3	19.76	17.22	2.88	94.8	20.57	12.18	3.39	116.8

**Fig. 10.** Performance comparison (%) of different methods on the Netflix dataset.

foNCE objective, with a complexity of $O(B^2 d_z)$. In addition, CEDRec constructs a static bias prior before training through top- k search in the visual and textual feature spaces, incurring a one-time complexity of $O(N^2(d_v + d_t))$.

Since the prior construction described above is an additional preprocessing operation introduced by CEDRec, we separately evaluate its practical overhead in terms of preprocessing time and peak GPU memory. On the same hardware used in the main experiments, this one-time preprocessing requires 1.69 s and 0.67 GB on Baby, 7.83 s and 2.05 GB on Sports, and 11.51 s and 2.88 GB on Clothing, respectively. These results indicate that the preprocessing cost remains manageable on the evaluated datasets. However, its quadratic dependence on the number of items may limit scalability for substantially larger item catalogs. For large-scale applications, approximate nearest-neighbor methods, such as Faiss-based ANN search, can be used to reduce the cost of neighborhood construction and improve scalability.

Table 7 further compares the practical efficiency of CEDRec with state-of-the-art baselines in terms of training time (s/epoch), inference time (ms/batch), GPU memory (GB), and parameter count (M). We can see that CEDRec maintains moderate training and inference costs across the three datasets. Although DuGRec is the most lightweight baseline, CEDRec still shows competitive efficiency compared with the other multimodal methods. Its GPU memory usage remains low across all three datasets, and its parameter count is smaller than those of MENTOR, HEARec, and SRGFormer. Overall, the results show that CEDRec achieves practical training and inference efficiency while introducing the additional EBN and SBN modules.

4.6. Generalization evaluation on the Netflix dataset (RQ5)

We further examine the applicability of CEDRec in a different recommendation scenario using the Netflix dataset ² (Wei et al., 2024).

Specifically, this movie recommendation dataset contains 13,187 users, 17,366 items, and 68,933 interactions, with a sparsity of 99.97%. It provides 768-dimensional textual features and 512-dimensional visual features. For evaluation, we use the same evaluation metrics as the main experiments, including Recall@20, NDCG@20, and Gini@20.

The results are shown in Fig. 10. It can be observed that CEDRec achieves the best NDCG@20 of 3.55 and the lowest Gini@20 of 78.41. In terms of Recall@20, DuGRec obtains the highest value of 9.58, while CEDRec achieves a comparable result of 9.39. Overall, these results show that CEDRec maintains competitive recommendation accuracy while achieving a more balanced item exposure distribution, further supporting its applicability to another visually rich recommendation scenario.

5. Conclusion

In this work, we introduce CEDRec, an adaptive debiasing framework that addresses signal dilution and semantic misidentification in multimodal recommendation. Its evidential bias network preserves strong modality-specific deceptive evidence during cross-modal fusion, while its semantic bottleneck network adjusts debiasing penalties according to cross-modal semantic consistency. Moreover, PCGrad mitigates gradient conflicts between collaborative preference learning and semantic bottleneck compression. Experimental results on three Amazon datasets and an additional Netflix dataset show that CEDRec reduces item exposure concentration while maintaining competitive ranking accuracy, demonstrating a favorable balance between debiasing effectiveness and recommendation quality.

Despite these contributions, CEDRec has several limitations. First, it operates on pre-extracted and fixed multimodal features, which may affect the reliability of cross-modal bias estimation and semantic consistency modeling. In addition, both the fused bias score and the gating signal are defined at the item level and therefore do not explicitly account for user-specific responses to modality-specific signals. Accordingly, future work will explore representations learned by vision-language mod-

² <https://www.kaggle.com/datasets/netflix-inc/netflix-prize-data>

els such as CLIP and BLIP. We will also investigate user-conditioned debiasing strategies to capture heterogeneous user responses to modality-specific signals.

CRedit authorship contribution statement

Lin Pan: Writing – original draft, Visualization, Methodology, Formal analysis, Conceptualization; **Zhiqiang Pan:** Writing – review & editing, Validation, Supervision; **Honghui Chen:** Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- Bai, T., Chen, W., Yang, C., & Shi, C. (2025). Invariant debiasing learning for recommendation via biased imputation. *Information Processing & Management*, 62(3), 104028.
- Chen, J., Dong, H., Qiu, Y., He, X., Xin, X., Chen, L., Lin, G., & Yang, K. (2021). Autodebias: Learning to debias for recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval* (pp. 21–30).
- Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., & He, X. (2023). Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems*, 41(3), 1–39.
- Chen, R., Fan, J., Wu, M., & Ma, S. (2025a). Conditional diffusion model for recommender systems. *Neural Networks*, 185, 107204.
- Chen, Z., Xu, J., Wang, H., Yang, S., Wan, Z., & Hu, H. (2025b). Hypercomplex prompt-aware multimodal recommendation. In *Proceedings of the 34th ACM international conference on information and knowledge management* (pp. 403–414).
- van den, O. A., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.
- Denoeux, T. (2000). A neural network classifier based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 30(2), 131–150.
- Gao, C., Zheng, Y., Wang, W., Feng, F., He, X., & Li, Y. (2024). Causal inference in recommender systems: A survey and future directions. *ACM Transactions on Information Systems*, 42(4), 1–32.
- Ge, Y., Liu, S., Gao, R., Xian, Y., Li, Y., Zhao, X., Pei, C., Sun, F., Ge, J., Ou, W., & Zhang, Y. (2021). Towards long-term fairness in recommendation. In *Proceedings of the 14th ACM international conference on web search and data mining* (pp. 445–453).
- Giahi, R., Yao, K., Kollipara, S., Zhao, K., Mirjalili, V., Xu, J., Biswas, T., Korpeoglu, E., & Achan, K. (2025). VL-CLIP: Enhancing multimodal recommendations via visual grounding and LLM-augmented CLIP embeddings. In *Proceedings of the 19th ACM conference on recommender systems* (pp. 482–491).
- Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 249–256).
- Guo, Z., Li, J., Li, G., Wang, C., Shi, S., & Ruan, B. (2024). LGMRec: Local and global graph learning for multimodal recommendation. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 8454–8462). (Vol. 38).
- Han, Z., Zhang, C., Fu, H., & Zhou, J. T. (2022). Trusted multi-view classification with dynamic evidential fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2), 2551–2566.
- He, R., & McAuley, J. (2016). VBPR: Visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 144–150). (Vol. 30).
- He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., & Wang, M. (2020). 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).
- He, Z., Zhang, S., Ma, W., Li, J., Sun, P., Ai, Q., Liu, Y., & Zhang, M. (2025a). User immersion-aware short video recommendation. *ACM Transactions on Information Systems*, 44(1), 1–33.
- He, Z.-Y., Lin, J.-Q., Wang, C.-D., & Guizani, M. (2025b). Interaction-knowledge semantic alignment for recommendation. *Neural Networks*, 181, 106755.
- Hu, J., Hooi, B., He, B., & Wei, Y. (2025). Modality-independent graph neural networks with global transformers for multimodal recommendation. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 11790–11798). (Vol. 39).
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
- Lee, S., Park, S., & Lee, J. (2025). Exploiting fine-grained skip behaviors for micro-video recommendation. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 12004–12012). (Vol. 39).
- Lin, S., Gao, C., Chen, J., Zhou, S., Hu, B., Feng, Y., Chen, C., & Wang, C. (2025a). How do recommendation models amplify popularity bias? An analysis from the spectral perspective. In *Proceedings of the eighteenth ACM international conference on web search and data mining* (pp. 659–668).
- Lin, X., Liu, R., Cao, Y., Zou, L., Li, Q., Wu, Y., Liu, Y., Yin, D., & Xu, G. (2025b). Contrastive modality-disentangled learning for multimodal recommendation. *ACM Transactions on Information Systems*, 43(3), 1–31.
- Liu, Q., Wu, S., & Wang, L. (2017). Deepstyle: Learning user preferences for visual recommendation. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval* (pp. 841–844).
- Liu, X., Tao, Z., Shao, J., Yang, L., & Huang, X. (2022). EIMRec: Eliminating single-modal bias in multimedia recommendation. In *Proceedings of the 30th ACM international conference on multimedia* (pp. 687–695).
- Lu, Y.-H., Xi, W.-D., Xing, X.-X., Wan, W., & Wang, C.-D. (2026). ID-guided multimodal experts with contrastive diffusion for sequential recommendation. *Neural Networks*, 198, 108612.
- Ma, H., Zhang, Y., Zhou, Y., Yang, B., Yu, D., & Li, Z. (2026). Let two graphs talk: Self-supervised dual-graph reconstruction for multimodal recommendation. *Information Fusion*, (p. 104299).
- Mansoury, M., Mobasher, B., & van Hoof, H. (2024). Mitigating exposure bias in online learning to rank recommendation: A novel reward model for cascading bandits. In *Proceedings of the 33rd ACM international conference on information and knowledge management* (pp. 1638–1648).
- Ning, W., Cheng, R., Yan, X., Kao, B., Huo, N., Haldar, N. A. H., & Tang, B. (2024). Debiasing recommendation with personal popularity. In *Proceedings of the ACM web conference 2024* (pp. 3400–3409).
- Ong, R. K., & Khong, A. W. H. (2025). Spectrum-based modality representation fusion graph convolutional network for multimodal recommendation. In *Proceedings of the eighteenth ACM international conference on web search and data mining* (pp. 773–781).
- Pan, L., Pan, Z., Cai, F., & Chen, H. (2026). Multimodal recommender systems: A survey of representation, modeling, and optimization. *Information Fusion*, 128, 103991.
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese bert-networks. 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. 3982–3992).
- Rendle, S., Freudenthaler, C., Gantner, Z., & Schmidt-Thieme, L. (2012). BPR: Bayesian personalized ranking from implicit feedback. arXiv preprint arXiv:1205.2618.
- Sensory, M., Kaplan, L., & Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. In *Advances in neural information processing systems* (pp. 1–11). (Vol. 31).
- Shang, Y., Gao, C., Chen, J., Jin, D., & Li, Y. (2024). Improving item-side fairness of multimodal recommendation via modality debiasing. In *Proceedings of the ACM web conference 2024* (pp. 4697–4705).
- Shi, K., Zhang, Y., Zhang, M., Chen, L., Yi, J., Xiao, K., Hou, X., & Li, Z. (2026). Structurally refined graph transformer for multimodal recommendation. *IEEE Transactions on Multimedia*, (pp. 1–13).
- Shi, K., Zhang, Y., Zhang, M., Xiao, K., Yu, D., Zhou, Y., Huang, W., & Li, Z. (2025). Heterogeneous environment-aware multimodal recommendation with modality alignment. *Information Fusion*, (p. 103989).
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
- Song, Q., Hu, J., Xiao, L., Sun, B., Gao, X., & Li, S. (2025). DiffCL: A diffusion-based contrastive learning framework with semantic alignment for multimodal recommendations. *IEEE Transactions on Neural Networks and Learning Systems*, 36(10), 18587–18597.
- Tao, Z., Liu, X., Xia, Y., Wang, X., Yang, L., Huang, X., & Chua, T.-S. (2022). Self-supervised learning for multimedia recommendation. *IEEE Transactions on Multimedia*, 25, 5107–5116.
- Wei, W., Ren, X., Tang, J., Wang, Q., Su, L., Cheng, S., Wang, J., Yin, D., & Huang, C. (2024). LLMRec: Large language models with graph augmentation for recommendation. In *Proceedings of the 17th ACM international conference on web search and data mining* (pp. 806–815).
- Wei, Y., Wang, X., Nie, L., He, X., Hong, R., & Chua, T.-S. (2019). 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).
- Wu, J., Yang, B., Zhu, Q., Mao, R., & Li, Q. (2025). Price-aware debiased learning model for recommendation. *Neural Networks*, 192, 107872.
- Xu, C., He, Y., Wang, J., & Zhang, W. (2025a). STAIR: Manipulating collaborative and multimodal information for e-commerce recommendation. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 12899–12907). (Vol. 39).
- Xu, J., Chen, Z., Wang, W., Hu, X., Kim, S.-W., & Ngai, E. C. H. (2025b). Cohesion: Composite graph convolutional network with dual-stage fusion for multimodal recommendation. In *Proceedings of the 48th international ACM SIGIR conference on research and development in information retrieval* (pp. 1830–1839).
- Xu, J., Chen, Z., Yang, S., Li, J., Wang, H., & Ngai, E. C. H. (2025c). MENTOR: Multi-level self-supervised learning for multimodal recommendation. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 12908–12917). (Vol. 39).
- Xu, X., Zhang, Y., Zhou, F., & Song, J. (2025d). Improving multimodal social media popularity prediction via selective retrieval knowledge augmentation. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 932–940). (Vol. 39).
- Yu, P., Tan, Z., Lu, G., & Bao, B.-K. (2025). Mind individual information! principal graph learning for multimedia recommendation. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 13096–13105). (Vol. 39).
- Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., & Finn, C. (2020). Gradient surgery for multi-task learning. In *Advances in neural information processing systems* (pp. 5824–5836). (Vol. 33).
- Zhang, J., Zhu, Y., Liu, Q., Wu, S., Wang, S., & Wang, L. (2021). Mining latent structures for multimedia recommendation. In *Proceedings of the 29th ACM international conference on multimedia* (pp. 3872–3880).

- Zhang, Q., Zhang, X., Liu, Y., Wang, H., Gao, M., Zhang, J., & Guo, R. (2023). Debiasing recommendation by learning identifiable latent confounders. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 3353–3363).
- Zheng, T., Xu, X., Li, S., Pan, Q., Jiang, J., & Du, M. (2026). Mitigating influence of selection and conformity biases for user interest in recommender systems. *Engineering Applications of Artificial Intelligence*, 167, 113810.
- Zhou, X. (2023). MMRec: Simplifying multimodal recommendation. In *Proceedings of the 5th ACM international conference on multimedia in Asia workshops* (pp. 1–2).
- Zhou, X., & Shen, Z. (2023). A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation. In *Proceedings of the 31st ACM international conference on multimedia* (pp. 935–943).
- Zhou, X., Zhou, H., Liu, Y., Zeng, Z., Miao, C., Wang, P., You, Y., & Jiang, F. (2023). Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM web conference 2023* (pp. 845–854).