On Popularity Bias of Multimodal-aware Recommender Systems: a Modalities-driven Analysis

Daniele Malitesta*
Politecnico di Bari, Italy
daniele.malitesta@poliba.it

Giandomenico Cornacchia*
Politecnico di Bari, Italy
giandomenico.cornacchia@poliba.it

Claudio Pomo
Politecnico di Bari, Italy
claudio.pomo@poliba.it

Tommaso Di Noia
Politecnico di Bari, Italy
tommaso.dinoia@poliba.it

ABSTRACT

Multimodal-aware recommender systems (MRSs) exploit multimodal content (e.g., product images or descriptions) as items' side information to improve recommendation accuracy. While most of such methods rely on factorization models (e.g., MFBPR) as base architecture, it has been shown that MFBPR may be affected by popularity bias, meaning that it inherently tends to boost the recommendation of popular (i.e., short-head) items at the detriment of niche (i.e., long-tail) items from the catalog. Motivated by this assumption, in this work, we provide one of the first analyses on how multimodality in recommendation could further amplify popularity bias. Concretely, we evaluate the performance of four state-of-the-art MRSs algorithms (i.e., VBPR, MMGCN, GRCN, LATTICE) on three datasets from Amazon by assessing, along with recommendation accuracy metrics, performance measures accounting for the diversity of recommended items and the portion of retrieved niche items. To better investigate this aspect, we decide to study the separate influence of each modality (i.e., visual and textual) on popularity bias in different evaluation dimensions. Results, which demonstrate how the single modality may augment the negative effect of popularity bias, shed light on the importance to provide a more rigorous analysis of the performance of such models.

CCS CONCEPTS

• **Information systems** → **Multimedia and multimodal retrieval**; **Personalization**.

KEYWORDS

Multimodal Recommendation, Popularity Bias

ACM Reference Format:
Daniele Malitesta, Giandomenico Cornacchia, Claudio Pomo, and Tommaso Di Noia. 2023. On Popularity Bias of Multimodal-aware Recommender Systems: a Modalities-driven Analysis. In *Proceedings of the 1st International Workshop on Deep Multimodal Learning for Information Retrieval (MMIR '23), November 2, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3606040.3617441

*Corresponding authors: Daniele Malitesta (daniele.malitesta@poliba.it) and Giandomenico Cornacchia (giandomenico.cornacchia@poliba.it).

This work is licensed under a Creative Commons Attribution International 4.0 License.

MMIR '23, November 2, 2023, Ottawa, ON, Canada
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0271-6/23/11.
https://doi.org/10.1145/3606040.3617441

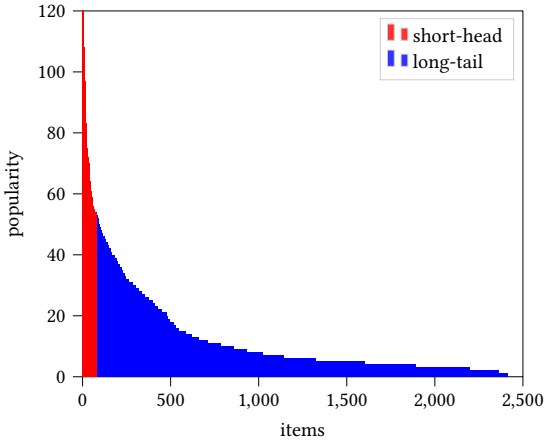

Figure 1: Short-head and long-tail items from the *Office* dataset in the Amazon catalog.

1 INTRODUCTION

The massive availability of digital data (e.g., images, texts, audio tracks) on the Internet has recently favored the raising of a novel family of recommender systems (RSs), known as multimodal-aware recommender systems (MRSs). With the integration of multimodal features (extracted through pre-trained deep learning models [26, 30, 51]) as items' side information, MRSs can generate more accurate recommendations than traditional collaborative filtering [17, 61, 66] (CF) algorithms by providing a countermeasure to common issues such as the sparsity of the user-item matrix and the cold-start scenario [27, 49, 56], or the inexplicability of users' preferences in the implicit feedback setting [15, 22, 23, 38, 39].

The vast majority of MRSs are generally based upon the famous matrix factorization with bayesian personalized ranking (MFBPR) recommendation model. On the one hand, matrix factorization [34] (MF) is a latent-factor approach that maps users and items in the recommendation system to embeddings in the latent space and is trained to reconstruct the user-item interaction matrix via the dot product of the respective factors. On the other hand, bayesian personalized ranking [52] (BPR) is an optimization schema that drives

from the assumption that, for each user, the predicted score of positive (i.e., interacted) and negative (i.e., non-interacted) items should diverge. Given its simple implementation and efficacy, MFBPR has long constituted the backbone of recommendation algorithms in CF [28, 29, 44], not only for multimodal recommendation.

Nevertheless, recommender systems (such as MFBPR) may be affected by popularity bias [2, 6, 10, 31] (Figure 1), as they tend to boost the recommendation of the items from the *short-head* (i.e., the popular ones) at the expense of the items from the *long-tail* (i.e., the niche ones). Tackling popularity bias in recommendation has primarily followed four directions [14]: (i) regularization techniques [2, 19, 32], (ii) adversarial learning [36], (iii) causal graphs [59, 67, 68], and (iv) other item re-ranking approaches [1, 3].

Despite the growing interest in popularity bias [5, 21] and potential solutions to address it, to date, very limited effort has been put into investigating ***how multimodal side information in MRSs could amplify the negative effects of popularity bias***. To the best of our knowledge, three recent works discussed the concept of bias in multimodal-aware recommendation. First, Liu et al. [40] take into account the bias towards a single modality in multimodal recommendation, and propose a solution based upon causal inference and counterfactual reasoning; however, the definition they provide about bias is conceptually different from the one of popularity bias. Then, Kowald and Lacic [35] consider popularity bias in the case of multimedia recommendation datasets (e.g., MovieLens); however, they do not support their findings by testing recommender systems leveraging multimodal features as items' side information. Last, Malitesta et al. [43] investigate how novelty and diversity metrics are influenced in multimodal recommendation, but without a finer-grained analysis on the impact of each single modality.

Driven from the assumptions above, and differently from the related literature, we propose one of the first analyses on how multimodal-aware recommender systems may amplify popularity bias in the produced recommendation lists. To this aim, we select four established and recent multimodal-aware recommender systems from the literature (i.e., VBPR [27], MMGCN [63], GRCN [62], and LATTICE [66]) and train them on three categories of the Amazon recommendation dataset [46] (i.e., *Office*, *Toys*, and *Clothing*). Then, we evaluate the performance of the models by assessing metrics accounting for recommendation accuracy and popularity bias (the latter is measured through the diversity of recommendation lists and the percentage of retrieved items from the long-tail). Finally, to tailor our investigation, we focus on the separate impact of each multimodal side information (i.e., visual or textual) on popularity bias. To conduct this further study, we train the selected recommender systems when integrating either the visual or the textual modality as items' side information, and study the performance on single metrics and across pairs of metrics.

We seek to answer: **RQ1.** How do multimodal-aware recommendation models behave in terms of accuracy, diversity, and popularity bias? **RQ2.** What is the influence of each modality (i.e., visual, textual, multimodal) on such performance measures? Results widely show that the integration of a single modality (with respect to the multimodal setting) is capable of amplifying the negative effects of popularity bias, paving the way to additional, more formal investigations on multimodal recommendation. We release the code at: https://github.com/sisinflab/MultiMod-Popularity-Bias.

2 RELATED WORK

This section outlines the related literature about multimodal learning and popularity bias in recommendation. First, we provide an overview of the most popular and recent advances in multimodal-aware recommendation, from which we select four representative approaches to analyze. Then, we summarize the concept of popularity bias, underlining how our work provides one of the first comprehensive investigations on popularity bias in multimodal recommendation at the granularity of modalities.

Multimodal-aware recommendation. In various domains such as fashion [16, 17, 25], music [20, 49, 55], food [37, 47, 58], and micro-video [13, 18, 63] recommendation, the multimodal content associated with items (e.g., product images and descriptions, or audio tracks) has demonstrated to greatly enhance the representational power of recommender systems.

Following the latest advances in multimodal learning [8, 9, 48], multimodal-aware recommender systems (MRSs) aim to tackle some long-term open challenges in personalized recommendation such as data sparsity and cold-start [27, 49, 56]. Moreover, leveraging multimodal content can help reveal underlying user-item interactions and intents through attention mechanisms, contributing to the explainability of recommendations [15, 17, 38, 39, 54].

With the recent outbreak of graph neural networks in recommendation [28, 45, 50], several techniques have started integrating multimodality into the user-item bipartite graphs and knowledge graphs [11, 28, 53, 57, 60], refining the multimodal representations of users and items through different approaches implementing the message-passing schema. While some early attempts involve simply injecting multimodal item features into the graph-based pipeline [65], more advanced techniques learn separate graph representations for each modality and disentangle users' preferences at the modality level [33, 54, 62]. Recent approaches focus on uncovering multimodal structural differences among items in the catalog [41, 42, 66], in some cases by leveraging self-supervised [61, 69] and contrastive [64] learning.

In this work, we select four popular and recent approaches in multimodal recommendation, namely, VBPR [27], MMGCN [63], GRCN [62], and LATTICE [66], and test their performance to assess the impact of (multi)modalities on popularity bias.

Popularity bias in recommendation. In recommendation, popularity bias refers to the system's tendency to favor popular items (i.e., *short-head*) at the expense of less popular ones (i.e., *long-tail*) [2, 6, 10, 12, 31]. For instance, Jannach et al. [31] conduct a comprehensive algorithmic comparison across multiple datasets; their findings indicate that existing recommendation methods tend to concentrate mainly on a small fraction of the available item spectrum. More recently, Abdollahpouri et al. [3] delve into this issue using the well-known MovieLens 1M dataset and reveal that over 80% of all ratings are attributed to popular items; their main focus lies in finding ways to strike a balance between ranking accuracy and the coverage of long-tail items.

On such basis, the literature currently recognizes four main research directions [14] to address popularity bias in recommendation, namely: (i) regularization techniques [2, 19, 32], (ii) adversarial learning [36], (iii) causal graphs [59, 67, 68], and (iv) other approaches such as item re-ranking [1, 3].

In multimodal recommendation, only a few recent works discuss popularity bias, but with specific definitions [40] and neglecting the impact of multimodal features [35], or on other evaluation metrics [43]. Conversely, our analysis assesses how prone multimodal-aware recommender systems are to push items belonging to the short-head and how the different modalities affect the tendency to amplify the popularity bias.

3 BACKGROUND

This section provides useful background notions for our proposed experimental analysis. To begin with, we introduce the preliminaries about the personalized recommendation scenario. Then, we focus on factorization-based approaches for recommendation (such as MFBPR) and present their building formulations. Finally, we extend the formalism to multimodal-aware recommendation, by considering the four selected approaches (i.e., VBPR, MMGCN, GRCN, and LATTICE) and their rationales.

3.1 Preliminaries

Let $\mathcal{U}$ and $\mathcal{I}$ be the set of users and items in the recommendation system, respectively, where their cardinalities are indicated as $|\mathcal{U}|$ and $|\mathcal{I}|$. Then, let $\mathbf{X} \in \mathbb{R}^{|\mathcal{U}| \times |\mathcal{I}|}$ be the user-item interaction matrix, where $x_{ui} = 1$ if user u interacted with item i, 0 otherwise. On such basis, we also introduce $\mathcal{R} = \{(u, i) \mid x_{ui} = 1\}$ as the set of recorded user-item interactions ($|\mathcal{R}|$ is its cardinality).

3.2 Factorization-based approaches

Currently, the majority of state-of-the-art recommender systems in collaborative filtering follow the matrix factorization [34] (MF) rationale. Despite the different building solutions they propose, the core idea is to map users' and items' IDs to embeddings in the latent space. Specifically, we indicate with $\mathbf{e}_u \in \mathbb{R}^d$ and $\mathbf{e}_i \in \mathbb{R}^d$ the embeddings for user u and item i, respectively, with $d << |\mathcal{U}|, |\mathcal{I}|$. Then, given a pair of user and item (u, i), the predicted interaction score is:

$$\hat{x}_{ui} = \mathbf{e}_u^\top \mathbf{e}_i. \tag{1}$$

To learn such embeddings, MF-based approaches are usually coupled with bayesian personalized ranking [52] (BPR). This optimization method assumes that the predicted interaction score for users and their positive (i.e., interacted) items should be higher than the predicted interaction score for users and their negative (i.e., non-interacted) items. Concretely, let $\mathcal{T} = \{(u, i, j) \mid x_{ui} = 1 \wedge x_{uj} = 0\}$ be the set of triples, where each triple consists of a user, a positive, and a negative item. Bayesian personalized ranking seeks to optimize the following objective function:

$$\arg\max_{\Theta} \sum_{(u,i,j) \in \mathcal{T}} \ln \sigma(\hat{x}_{ui} - \hat{x}_{uj}), \tag{2}$$

where Θ is the vector containing all model's parameters (e.g., in the case of MF, $\mathbf{e}_u$ and $\mathbf{e}_i$), while $\sigma(\cdot)$ is the sigmoid function.

3.3 Factorization-based approaches leveraging multimodal side information

We present the formulations of four state-of-the-art multimodal-aware recommender systems (MRSs): VBPR [27], MMGCN [63], GRCN [62], and LATTICE [66]. Before diving into their approaches, we introduce some additional formalism.

Besides $\mathbf{e}_u$ and $\mathbf{e}_i$, hereafter referred to as *collaborative* user and item embeddings, we also introduce $\mathbf{f}_u$ and $\mathbf{f}_i$ as the *multimodal* embeddings for user u and item i. Moreover, we indicate $\mathcal{M}$ as the set of available modalities (e.g., visual, textual, audio), and we use m as embedding's apex to denote that the embedding refers to the $m \in \mathcal{M}$ modality (e.g., $\mathbf{f}_i^m$ stands for the m-th multimodal embedding of item i).

VBPR. Visual-bayesian personalized ranking [27] (dubbed as VBPR) adopts visual features extracted from product images as items' side information in MFBPR. The authors introduce, along with user and item *collaborative* embeddings, additional *visual* user and item embeddings, where the latter is obtained as the activation of the penultimate layer from a pre-trained convolutional neural network. Then, the collaborative and visual embeddings are used to measure a collaborative- and visual-aware prediction for the interaction score and are eventually summed to obtain the final prediction score. In this work, we follow [66] and adapt VBPR to multimodality by concatenating the visual and textual item features to generate a unique multimodal representation of the item:

$$\hat{x}_{ui} = \mathbf{e}_u^\top \mathbf{e}_i + \mathbf{f}_u^\top t(\mathbf{f}_i) \quad \text{with} \quad \mathbf{f}_i = \Big\Vert_{m \in \mathcal{M}} \mathbf{f}_i^m, \tag{3}$$

where t is a projection function such that the latent dimensions of the multimodal user and item embeddings match.

MMGCN. One of the first approaches leveraging the representational power of graph convolutional networks (GCNs) with multimodal content is multimodal graph convolution network for recommendation [63] (dubbed as MMGCN). By designing one GCN for each modality, the model learns the different preferences users have towards each representation of the items. Finally, to fuse all multimodal representations into one for both users and items embeddings, the authors adopt the element-wise addition, and the predicted interaction score is calculated via the dot product:

$$\hat{x}_{ui} = \mathbf{f}_u^\top \mathbf{f}_i \quad \text{with} \quad \mathbf{f}_u = \sum_{m \in \mathcal{M}} c(\mathbf{e}_u, g(\mathbf{f}_u^m), t(\mathbf{f}_u^m, \mathbf{e}_u)), \tag{4}$$

where c and g are a combination and GCN-based functions. We report only the user-side formulation for the sake of space.

GRCN. Similarly to MMGCN, graph-refined convolutional network for multimedia recommendation [62] (dubbed as GRCN) utilizes a GCN-architecture to update user and item embeddings. Specifically, the adjacency matrix entries are refined by pruning the noisy user-item interactions according to the preference of users toward each item's modality. Collaborative and multimodal versions of the user and item embeddings are eventually combined through concatenation to estimate the interaction score via their dot product:

$$\hat{x}_{ui} = \mathbf{f}_u^\top \mathbf{f}_i \quad \text{with} \quad \mathbf{f}_u = g(\mathbf{e}_u, \mathbf{f}_u^m, \forall m \in \mathcal{M}) \,||\, \left(\Big\Vert_{m \in \mathcal{M}} t(\mathbf{f}_u^m) \right). \tag{5}$$

Again, we report only the user-wise formulation for lack of space.

LATTICE. Latent structure mining method for multimodal recommendation [66] (dubbed as LATTICE) performs graph structure learning on multiple modality-aware item-item graphs (one for each modality). The obtained adjacency matrices are aggregated

Table 1: Statistics of the tested datasets.

Datasets	$\lvert\mathcal{U}\rvert$	$\lvert\mathcal{I}\rvert$	$\lvert\mathcal{R}\rvert$	Sparsity (%)
Office	4,905	2,420	53,258	99.5513
Toys	19,412	11,924	167,597	99.9276
Clothing	39,387	23,033	278,677	99.9693

through weighted element-wise addition, and the final adjacency matrix is exploited to perform graph convolution to update the representation of the collaborative item embeddings. Then, this updated version is added to the initial collaborative item embedding. Finally, the dot product between the collaborative user and (updated) item embeddings predicts the interaction score:

$$\hat{x}_{ui} = \mathbf{e}_u^\top \mathbf{f}_i \quad \text{with} \quad \mathbf{f}_i = \mathbf{e}_i + \frac{g(\mathbf{e}_i, \mathbf{f}_i^m, \forall m \in \mathcal{M})}{||g(\mathbf{e}_i, \mathbf{f}_i^m, \forall m \in \mathcal{M})||_2}, \tag{6}$$

where g is a LightGCN [28] architecture performing graph structure learning as stated above.

4 PROPOSED ANALYSIS

In this section, we present the details to conduct our analysis. Initially, we report on the used datasets, describing the methodologies employed for extracting multimodal features. Subsequently, we introduce and formally define the evaluation metrics employed, encompassing accuracy, diversity, and popularity bias. Finally, we provide a thorough summary of the reproducibility information for our study, detailing the methods used for dataset splitting and filtering as well as the strategy for hyperparameter search.

4.1 Datasets

he multimodal recommender systems have been tested on three popular [17, 33, 66, 69] datasets from the Amazon catalog [46]: Office Products (*Office*), (b) Toys & Games (*Toys*), and (c) Clothing, Shoes & Jewelry (*Clothing*). The multimodal datasets provide both images and descriptions for each available item. Specifically, we utilize the pre-extracted 4,096-dimensional visual features [24] which are made publicly available[1]. For the textual modality, we follow the existing literature [66], which aggregates the item's title, descriptions, categories, and brand, thereby generating textual embeddings by leveraging sentence transformers [51]. The generated features are 1,024-dimensional embeddings. Additional dataset information can be found in Table 1.

4.2 Evaluation metrics

In the proposed study, we refer to various metrics that may bring out additional insights which have not been investigated yet in multimodal recommendation. Indeed, we do not solely rely on accuracy metrics (i.e., Recall and nDCG) but also on diversity (i.e., item coverage) and popularity bias (i.e., APLT) metrics. The metrics listed hereinafter are calculated on top-k recommendation lists.

Recall. The Recall assesses the system's capacity to retrieve relevant items from the recommendation list, highlighting the need for thorough coverage to the list of user interactions [7]:

$$\text{Recall@}k = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{|\text{Rel}_u \, @k|}{|\text{Rel}_u|}, \tag{7}$$

where Rel_u indicates the set of relevant items for user u, while $\text{Rel}_u \, @k$ is the set of relevant recommended items in the top-k list.

Normalized discount cumulative gain. The normalized discount cumulative gain (nDCG) considers the relevance and the ranking position of recommended products, taking into account the varied degrees of relevance:

$$\text{nDCG@}k = \frac{1}{|\mathcal{U}|} \sum_{u} \frac{\text{DCG}_u@k}{\text{IDCG}_u@k}, \tag{8}$$

where $\text{DCG@}k = \sum_{i=1}^{k} \frac{2^{rel_{u,i}}-1}{\log_2(i+1)}$ quantifies the cumulative gain of relevance scores through the recommended list, with $rel_{u,i} \in \text{Rel}_u$, and IDCG represents the cumulative gain of relevance scores for a perfect (ideal) recommender system.

Item coverage. The item coverage (abbreviated "iCov" in the following) gives information on the coverage (item-side) measured in recommendation lists. A higher item coverage suggests that a larger fraction of the item space is being scrutinized and recommended to consumers, implying a more comprehensive coverage of user preferences and potentially a more comprehensive recommendation experience. In particular, we have:

$$\text{iCov@}k = \frac{|\bigcup_u \hat{\mathcal{I}}_u@k|}{|\mathcal{I}_{train}|}, \tag{9}$$

where $\hat{\mathcal{I}}_u@k$ is the list of top-k recommended items for a user u.

Average percentage of long-tail items. The average percentage of long-tail items (APLT) is a measure used to assess the presence of popularity bias in recommendation systems [2]. Popularity bias refers to the tendency of recommendation algorithms to prioritize popular or mainstream items over less well-known or niche items. This bias can lead to limited exposure of users to diverse and personalized recommendations. The metric measure the percentage of items belonging to the medium/long-tail distribution in the recommendation lists averaged over all users:

$$\text{APLT@}k = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{|\{i \mid i \in (\hat{\mathcal{I}}_u@k \cap \sim \Phi)\}|}{k}, \tag{10}$$

where Φ is the set of items belonging to the short-tail distribution while $\sim \Phi$ is the set of items from the medium/long-tail distribution. Note that we decide to integrate the evaluation of the APLT along with the iCov (introduced above) because the latter may be functional to provide a complete interpretation of the former. Indeed, following their definitions and formulations, the two metrics are conceptually related.

Metrics value interpretation An ideal recommender system should increase all the metrics listed above according to the principle "higher is better" to boost accuracy and diversity while reducing the popularity bias of the produced recommendations. ***Nevertheless, with the current work, we try to unveil whether and why multimodal-aware recommender systems are affected by popularity bias. Thus, in the following, we will take into account those settings in which accuracy is high, while diversity and popularity bias are low (according to the metrics definitions).***

[1] https://cseweb.ucsd.edu/~jmcauley/datasets/amazon/links.html.

4.3 Reproducibility

We investigate the models' behavior in three different settings: (i) *visual* modality, in which we employ only visual features, (ii) *textual* modality, in which we employ only textual features, and (iii) *multimodal*, where both modalities are considered and combined.

In the first step, we evaluate the models in the multimodal setting which is the same setting as the original one for each tested approach. Then, we focused on quantifying the singular modality influence on the multimodal scenario in terms of accuracy, diversity, and popularity bias. Furthermore, to ensure the reproducibility of our work, in the following, we provide comprehensive details regarding the preprocessing and splitting of the datasets, as well as the tuning and evaluation of the models.

The datasets are filtered using the p-core strategy, where we set p to 5. Subsequently, we employ an 80%/20% train-test hold-out strategy to split the dataset. During the hyper-parameter tuning phase, we further divide the test set by removing 50% of its instances for the validation, specifically evaluating the results using the Recall@20 metric (as in the original work). In terms of models' training, we set the maximum number of epochs to 200 and select the model weights based on the epoch that yields the best performance on the validation set.

The code is implemented in Elliot [4]. Note that the explored hyper-parameter values are not entirely aligned with the ones in the original papers and codes. Indeed, we want to tune the selected baselines ***on an extensive, shared set of hyper-parameter values across all models for the sake of fair comparison.***

5 RESULTS AND DISCUSSION

In this section, we answer the following research questions (RQs):

RQ1. *How do the selected multimodal-aware recommendation models behave in terms of accuracy, diversity, and popularity bias?* Section 5.1 investigates the recommendation performance in terms of accuracy (i.e., Recall, nDCG), diversity (i.e., iCov), and popularity bias (i.e., APLT). Note that, for the sake of completeness, we also report the performance of a recommender system generating recommendations in a random manner (i.e., Random) or based upon the most popular items in the catalog (i.e., MostPop); then, we train and evaluate MFBPR, that is the building model of the other multimodal baselines. We regard the performance of Random, MostPop, and MFBPR as a reference for the other multimodal-aware recommender systems we want to analyze.

RQ2. *What is the influence of each modality setting (i.e., visual, textual, multimodal) on such performance measures?* Section 5.2 takes a step further by analyzing how each modality (i.e., visual, textual, and multimodal) influences accuracy, diversity, and popularity bias; the evaluation is conducted both on the single metric and across pairs of metrics.

5.1 Recommendation accuracy, diversity, and popularity bias (RQ1)

The results of the accuracy, diversity, and popularity bias metrics are reported in Table 2. The measured values refer to top@10, top@20, and top@50 recommendation lists. In the following, we discuss the obtained results considering the three metrics families separately.

Accuracy. Overall, LATTICE is the top-performing model, in alignment with the recent literature [66]. Indeed, its ability to learn more refined items' embeddings based upon the multimodal item-item similarities may positively impact the accuracy performance. Conversely, VBPR's outstanding performance with respect to the other multimodal approaches comes as quite a surprise, considering that more complex and recent models leveraging graph neural networks (such as MMGCN and GRCN) do not outperform it.

Considering the performance on a dataset level, the most significant variation in metrics between LATTICE and VBPR is observed on *Toys* and *Clothing*, while the difference is reduced on *Office*. Notably, *Toys* and *Clothing* store three and four times more interactions than *Office*, respectively, but they are much sparser. This emphasizes LATTICE's ability to recommend more accurate items despite the higher dataset sparsity. Assessing the other models' performance, MMGCN works exceptionally well on *Toys* but shows the lowest performance as the number of interactions and sparsity increase. GRCN, in contrast, excels with highly sparse data, exhibiting an opposite trend to MMGCN.

From a metric-wise analysis, LATTICE outperforms VBPR in correctly predicting relevant items (high Recall) that are more likely to appear at the top of the recommendation lists (nDCG). However, the same trend is not as evident on the Recall, partly due to its normalization w.r.t. the k recommended items, which can lead to a smaller difference between LATTICE and VBPR as k increases.

Diversity. As far as recommendation diversity (i.e., iCov) is concerned, the worst-performing model is MMGCN, since its iCov is, in any case, negatively out of scale compared to the other models. For instance, when taking into account *Office*, MMGCN's iCov is slightly better than MostPop (whose item diversity is, by construction, the lowest) demonstrating a restricted ability to engage diverse items in the recommendation lists. Unexpectedly, the second-worst model is LATTICE, even if its performance is still more balanced to the other approaches than MMGCN's one. Indeed, we observe that while MMGCN is affected by poor accuracy due to the lack of item diversity, LATTICE can deal with both accuracy and diversity.

As an opposite (but noteworthy) trend, we underline that VBPR and GRCN are generally capable of recommending a wider portion of items than MMGCN and LATTICE, independently on the selected top-k. Overall, their iCov values are quite comparable to the ones of Random, which should provide (by definition) the highest item coverage from the catalog. We intend to further investigate (and justify) this aspect by assessing the effects of popularity bias.

Popularity bias. In terms of popularity bias (i.e., APLT), the worst and second-worst models are once again MMGCN and LATTICE (the former on *Office* and *Clothing*, while the latter on *Toys*). As already discussed in Section 4.2, it makes sense to conceptually bind iCov and APLT. When assessing MMGCN's performance on *Office*, it becomes clear how the model is recommending only a few items (see again the iCov) while achieving good results in terms of accuracy; this demonstrates how the user-item interactions from *Office* may likely be biased towards popular items, and the phenomenon is even amplified due to the dataset small size. The same does not hold on *Clothing* where MMGCN, usually prone to popularity bias, gets also really low performance in terms of accuracy. Conversely, LATTICE can recommend popular items thus pushing its accuracy performance without amplifying the

Table 2: Results in terms of recommendation accuracy (Recall, nDCG), diversity (iCov) and popularity bias (APLT). For accuracy metrics, ↑ means better performance, while ↓ means less diversity and more popularity bias. We remind that, while iCov and APLT metrics would generally adhere to the principle of "higher is better" (↑) for an ideal recommender system, in this work we consider the opposite as we want to emphasize which models are performing worst in terms of diversity and popularity bias.

Datasets	Models	top@10				top@20				top@50			
		Recall↑	**nDCG↑**	**iCov↓**	**APLT↓**	**Recall↑**	**nDCG↑**	**iCov↓**	**APLT↓**	**Recall↑**	**nDCG↑**	**iCov↓**	**APLT↓**
Office	Random	0.0034	0.0020	2,414	0.5950	0.0079	0.0034	2,414	0.5948	0.0220	0.0068	2,414	0.5924
	MostPop	0.0302	0.0208	20	0.0000	0.0533	0.0282	32	0.0000	0.1143	0.0439	66	0.0000
	MFBPR	0.0602	0.0389	2,268	0.2294	0.0955	0.0500	2,357	0.2379	0.1657	0.0677	2,398	0.2513
	VBPR	0.0652	0.0419	2,265	0.2321	0.1025	0.0533	2,354	0.2375	0.1774	0.0721	2,404	0.2469
	MMGCN	0.0455	0.0300	**74**	**0.0016**	0.0798	0.0405	**112**	**0.0078**	0.1575	0.0598	**247**	**0.0205**
	GRCN	0.0393	0.0253	2,390	0.3438	0.0667	0.0339	2,409	0.3469	0.1250	0.0488	2,414	0.3548
	LATTICE	**0.0664**	**0.0449**	2,121	0.1752	**0.1029**	**0.0566**	2,315	0.2039	**0.1780**	**0.0751**	2,397	0.2413
Toys	Random	0.0011	0.0006	11,879	0.4894	0.0021	0.0008	11,879	0.4896	0.0051	0.0015	11,879	0.4902
	MostPop	0.0130	0.0075	13	0.0000	0.0229	0.0104	24	0.0000	0.0451	0.0156	56	0.0000
	MFBPR	0.0641	0.0403	10,016	0.1167	0.0903	0.0481	10,944	0.1268	0.1394	0.0596	11,544	0.1460
	VBPR	0.0710	0.0458	10,085	0.1064	0.1006	0.0545	11,026	0.1180	0.1523	0.0667	11,624	0.1400
	MMGCN	0.0256	0.0150	**4,499**	0.0961	0.0426	0.0200	**6,238**	0.1058	0.0785	0.0285	**8,657**	0.1263
	GRCN	0.0554	0.0354	11,007	0.2368	0.0831	0.0436	11,609	0.2482	0.1355	0.0559	11,847	0.2679
	LATTICE	**0.0805**	**0.0512**	8,767	**0.0546**	**0.1165**	**0.0617**	10,285	**0.0684**	**0.1771**	**0.0759**	11,397	**0.0950**
Clothing	Random	0.0004	0.0002	23,016	0.4487	0.0010	0.0003	23,016	0.4478	0.0024	0.0006	23,016	0.4482
	MostPop	0.0089	0.0046	13	0.0000	0.0157	0.0063	24	0.0000	0.0322	0.0095	56	0.0000
	MFBPR	0.0303	0.0156	18,414	0.0729	0.0459	0.0195	20,582	0.0824	0.0734	0.0249	22,171	0.1017
	VBPR	0.0339	0.0181	19,195	0.0809	0.0529	0.0229	21,251	0.0915	0.0847	0.0292	22,555	0.1112
	MMGCN	0.0227	0.0119	**1,744**	**0.0044**	0.0348	0.0150	**2,864**	**0.0066**	0.0609	0.0201	**5,373**	**0.0121**
	GRCN	0.0319	0.0164	21,490	0.2358	0.0496	0.0209	22,503	0.2459	0.0858	0.0281	22,954	0.2631
	LATTICE	**0.0502**	**0.0275**	13,463	0.0134	**0.0744**	**0.0336**	17,538	0.0207	**0.1186**	**0.0425**	21,458	0.0385

popularity bias phenomenon as much as MMGCN does. Indeed, even if LATTICE's iCov is the second-worst across all the datasets, the metric is always close to the best models in terms of diversity.

Finally, VBPR and GRCN confirm their ability (already observed on the diversity measure) to tackle also popularity bias in all experimental settings. Particularly, while we recognize that VBPR performance is slightly increased with respect to MFBPR in terms of iCov and APLT (the two approaches are almost similar), GRCN results are quite remarkable. It might be the case that its graph edges pruning technique (driven by multimodal signals) is reducing the influence of noisy user-item interactions (i.e., redundant edges which might involve popular items), thus helping to diversify the recommendations by considering also several long-tail items.

Summary. *In a standard multimodal setting, LATTICE stands out for its accuracy performance and ability to handle dataset sparsity, but at the detriment of amplifying popularity bias; MMGCN struggles with diversity, exhibits strong popularity bias, and sacrifices accuracy in certain scenarios; VBPR and GRCN, in different manners, better manage all the metrics by finding the right compromise among them.*

5.2 Modalities influence on recommendation performance (RQ2)

While the previous section has answered how multimodal recommender systems perform in terms of accuracy, diversity, and popularity bias when leveraging the ***full*** modalities, in the following, we discuss the influence of each ***single*** modality on the performance. We consider two evaluation dimensions where modalities influence is assessed (i) on accuracy, diversity, and popularity bias separately, and (ii) on pairs of metrics to investigate their joint variations.

Modalities influence on the single metric. Figure 2 displays the influence of each modality calculated as percentage variation with respect to the multimodal baseline, on the top@20 recommendation lists. We select the Recall (Figure 2a), iCov (Figure 2b), and APLT (Figure 2c) for accuracy, diversity, and popularity bias, respectively.

As regards the accuracy performance (Figure 2a), we notice how the trend is not consistent across all the datasets and models. Particularly, when considering *Office*, we observe that only VBPR and LATTICE fully exploit multimodality (indeed, their performance decreases when the modalities are injected separately); on an opposite level, on MMGCN, the visual modality slightly improves the multimodal setting, while the textual modality even worsens it; then, GRCN achieves better performance on both the visual and textual modalities, suggesting that this approach may not take advantage of the multimodal configuration. On the *Toys* dataset, the only textual setting generally improves the performance, bringing important information to the model learning interaction. The model benefiting from the single modality the most is MMGCN, which has an improvement of at least 20% on both visual and textual. For the remaining models, the trend is quite stable with the textual and visual modalities improving and reducing the performance, respectively. Finally, we observe that *Clothing* is the only dataset

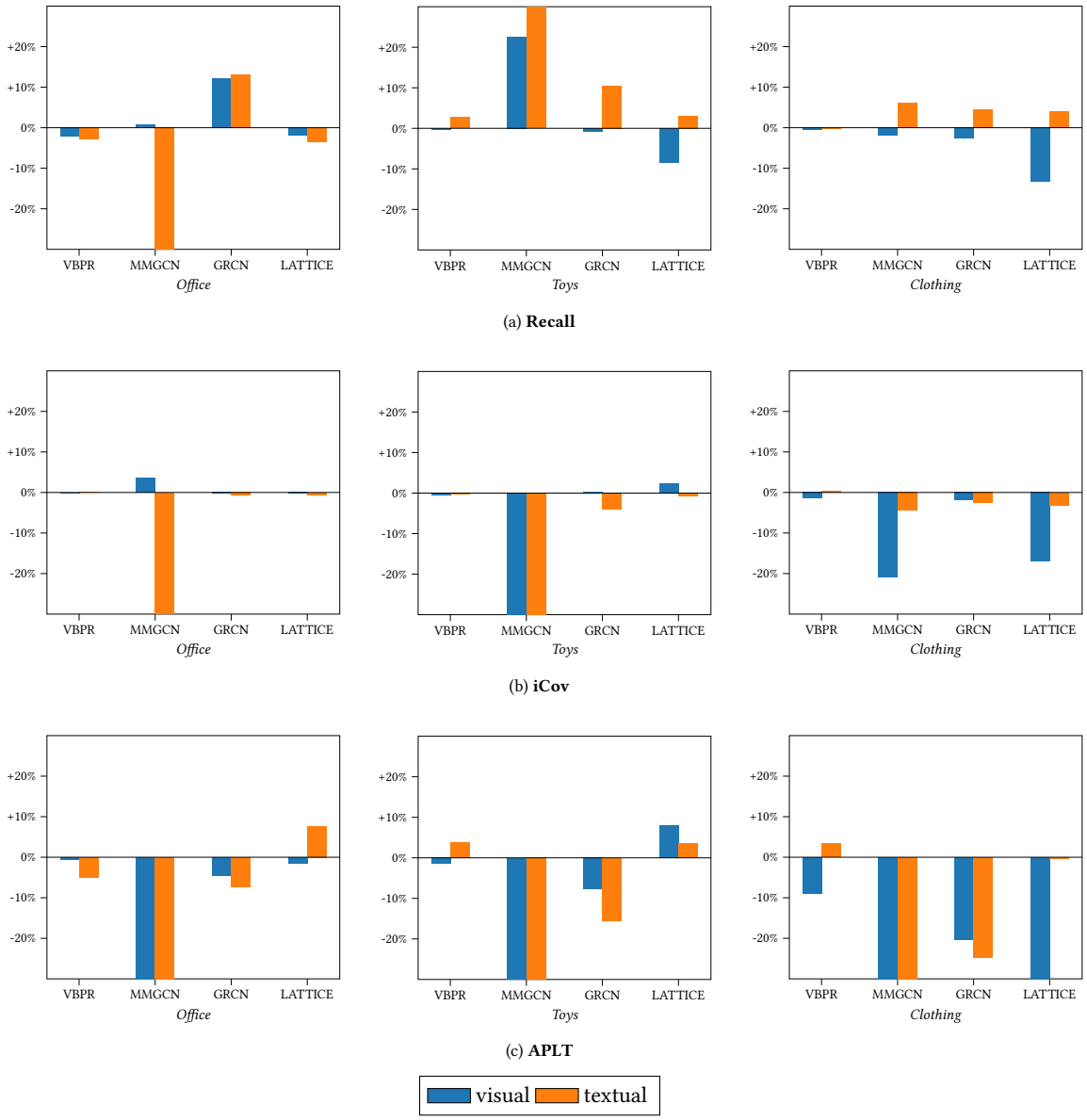

Figure 2: Percentage variation on the (a) Recall, (b) iCov, and (c) APLT when training the multimodal recommender systems with either visual or textual modalities. The 0% line stands for the reference performance provided by the multimodal version of the model. All results refer to the top@20 recommendation lists.

showing consistent trends. Indeed, the visual modality reduces the Recall while the textual increases it (with the only exception of VBPR whose percentage variation is negligible).

Differently from the accuracy analysis, we recognize a quasi-stable trend in the performance variation measured for the diversity metric (Figure 2b). Considering the *Office* dataset, each modality's contribution is generally irrelevant except for MMGCN, for which the visual modality slightly improves the coverage across the whole recommendation list, while the textual one worsens the performance by a large margin. Assessing the trend on *Toys*, both the modalities decrease the coverage performance of the model when injected separately in the recommendation pipeline; remarkably, MMGCN is once again the model affected by the single modality presence the most, but this time the coverage performance widely deteriorates because of both the visual and textual modalities. Finally, on *Clothing*, both modalities lower the model's item coverage, with specific reference to the visual modality.

As the last part of our analysis, we take into account each modality's contribution to the popularity bias dimension (Figure 2c). Starting from *Office*, we notice how both modalities are prone to enforce popularity bias if injected singularly, with the only exception of LATTICE whose textual modality limits the popularity bias (the

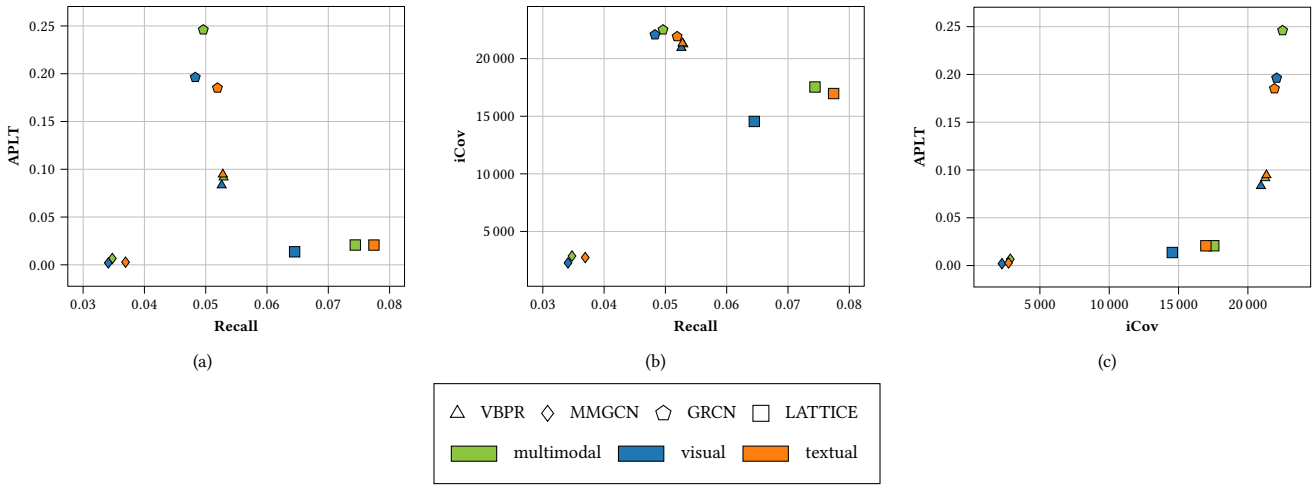

Figure 3: Performance analysis on *Clothing* when comparing (a) Recall *vs.* APLT, (b) Recall *vs.* iCov, and (c) iCov *vs.* APLT for different modality settings involving the multimodal, visual, and textual modalities. Metrics are on top@20.

APLT increases); this is interesting as we remind that LATTICE is the second-worst model in terms of popularity bias, but using only the textual modality reduces its accuracy performance and the influence of popular items in the recommendation list. When it comes to the *Toys* dataset, every single modality enforces the popularity bias of MMGCN and GRCN; for VBPR, the visual and textual modalities amplify and reduce the bias, respectively, while for LATTICE both the visual and textual modalities limit the popularity bias. Finally, on *Clothing*, both the modalities show to increase the popularity bias of the model (but the textual one on VBPR and LATTICE).

Modalities cross-influence on metrics pairs. To conclude, we discuss the cross-influence of each modality setting (i.e., visual, textual, and multimodal) on pairs of metrics. In this respect, we decide to display (Figure 3) the joint trend of (a) accuracy and popularity bias (i.e., Recall *vs.* APLT), (b) accuracy and diversity (i.e., Recall *vs.* iCov), and (c) diversity and popularity bias (i.e., iCov *vs.* APLT). We only report the results on *Clothing* for top@20 recommendations.

In detail, VBPR and MMGCN are the models being affected by each specific modality the least, since the performance measures assessed on visual and textual are generally aligned with the multimodal reference. Regarding LATTICE, we notice that the textual modality has a major accuracy influence with respect to popularity bias and diversity. Indeed, the textual modality improves the Recall without having a relevant effect in terms of iCov and APLT; conversely, the visual modality reduces the accuracy by jointly worsening the diversity and the popularity bias. Finally, when considering GRCN, we observe that the multimodal setting reduces the popularity bias without affecting the accuracy and diversity.

Summary. *In a single modality setting, the textual one improves the accuracy, while both modalities negatively affect the diversity and reinforce the popularity bias. When evaluating the modalities' influence across metrics pairs, the textual modality has a significant influence on accuracy but minimal effects on diversity and popularity bias; conversely, the visual modality reduces accuracy and jointly worsens the popularity bias and diversity.*

6 CONCLUSION AND FUTURE WORK

Motivated by the assumption that factorization models in recommendation (such as MFBPR) are affected by popularity bias, in this work, we provided one of the first systematic analyses on how multimodal-aware recommender systems (largely built upon MFBPR) further amplify the recommendation of popular items. After having selected four state-of-the-art multimodal recommender systems, namely, VBPR, MMGCN, GRCN, and LATTICE, we proposed an exhaustive experimental study involving three datasets from the Amazon catalog, four metrics spanning three evaluation dimensions (i.e., accuracy, diversity, and popularity bias), and three modalities settings (i.e., multimodal, only visual, and only textual). Results demonstrated that, in a standard multimodal setting, VBPR and GRCN can strike a better compromise between all evaluated metrics than MMGCN and LATTICE; furthermore, the separate injection of the visual and textual modalities can improve the accuracy but negatively impact the diversity and popularity bias. Conclusively, a complementary investigation regarding the modalities' influence on metrics pairs outlined that the textual modality has a considerable impact on accuracy but little effect on diversity and popularity bias, whereas the visual modality reduces accuracy while exacerbating popularity bias and limiting the diversity. Such findings pave the way to a more complete study on the performance of these models and other more recent approaches in multimodal recommendation.

ACKNOWLEDGMENTS

This work was partially supported by the following projects: Secure Safe Apulia, MISE CUP: I14E20000020001 CTEMT - Casa delle Tecnologie Emergenti Comune di Matera, CT_FINCONS_III, OVS Fashion Retail Reloaded, LUTECH DIGITALE 4.0, KOINÈ.

REFERENCES

[1] Himan Abdollahpouri. 2019. Popularity Bias in Ranking and Recommendation. In *AIES*. ACM, 529–530.
[2] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2017. Controlling Popularity Bias in Learning-to-Rank Recommendation. In *RecSys*. ACM, 42–46.

[3] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2019. Managing Popularity Bias in Recommender Systems with Personalized Re-Ranking. In *FLAIRS*. AAAI Press, 413–418.

[4] Vito Walter Anelli, Alejandro Bellogín, Antonio Ferrara, Daniele Malitesta, Felice Antonio Merra, Claudio Pomo, Francesco Maria Donini, and Tommaso Di Noia. 2021. Elliot: A Comprehensive and Rigorous Framework for Reproducible Recommender Systems Evaluation. In *SIGIR*. ACM, 2405–2414.

[5] Vito Walter Anelli, Yashar Deldjoo, Tommaso Di Noia, Daniele Malitesta, Vincenzo Paparella, and Claudio Pomo. 2023. Auditing Consumer- and Producer-Fairness in Graph Collaborative Filtering. In *ECIR (1) (Lecture Notes in Computer Science, Vol. 13980)*. Springer, 33–48.

[6] Ricardo Baeza-Yates. 2020. Bias in Search and Recommender Systems. In *RecSys*. ACM, 2.

[7] Ricardo Baeza-Yates and Berthier A. Ribeiro-Neto. 2011. *Modern Information Retrieval - the concepts and technology behind search, Second edition*. Pearson Education Ltd., Harlow, England.

[8] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Challenges and applications in multimodal machine learning. In *The Handbook of Multimodal-Multisensor Interfaces, Volume 2 (2)*. Association for Computing Machinery, 17–48.

[9] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis Philippe Morency. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 2 (2019), 423–443.

[10] Ludovico Boratto, Gianni Fenu, and Mirko Marras. 2021. Connecting user and item perspectives in popularity debiasing for collaborative recommendation. *Inf. Process. Manag.* 58, 1 (2021), 102387.

[11] Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Velickovic. 2021. Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. *CoRR* abs/2104.13478 (2021).

[12] Erik Brynjolfsson, Yu Jeffrey Hu, and Michael D Smith. 2006. From niches to riches: Anatomy of the long tail. *Sloan management review* 47, 4 (2006), 67–71.

[13] Desheng Cai, Shengsheng Qian, Quan Fang, and Changsheng Xu. 2022. Heterogeneous Hierarchical Feature Aggregation Network for Personalized Micro-Video Recommendation. *IEEE Trans. Multim.* 24 (2022), 805–818.

[14] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023. Bias and Debias in Recommender System: A Survey and Future Directions. *ACM Trans. Inf. Syst.* 41, 3 (2023), 67:1–67:39.

[15] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive Collaborative Filtering: Multimedia Recommendation with Item- and Component-Level Attention. In *SIGIR*. ACM, 335–344.

[16] Wen Chen, Pipei Huang, Jiaming Xu, Xin Guo, Cheng Guo, Fei Sun, Chao Li, Andreas Pfadler, Huan Zhao, and Binqiang Zhao. 2019. POG: Personalized Outfit Generation for Fashion Recommendation at Alibaba iFashion. In *KDD*. ACM.

[17] Xu Chen, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2019. Personalized Fashion Recommendation with Visual Explanations based on Multimodal Attention Network: Towards Visually Explainable Recommendation. In *SIGIR*. ACM, 765–774.

[18] Xusong Chen, Dong Liu, Zhiwei Xiong, and Zheng-Jun Zha. 2021. Learning and Fusing Multiple User Interest Representations for Micro-Video and Movie Recommendations. *IEEE Trans. Multim.* 23 (2021), 484–496.

[19] Zhihong Chen, Rong Xiao, Chenliang Li, Gangfeng Ye, Haochuan Sun, and Hongbo Deng. 2020. ESAM: Discriminative Domain Adaptation with Non-Displayed Items to Improve Long-Tail Performance. In *SIGIR*. ACM, 579–588.

[20] Zhiyong Cheng, Jialie Shen, and Steven C. H. Hoi. 2016. On Effective Personalized Music Retrieval by Exploring Online User Behaviors. In *SIGIR*. ACM, 125–134.

[21] Giandomenico Cornacchia, Vito Walter Anelli, Giovanni Maria Biancofiore, Fedelucio Narducci, Claudio Pomo, Azzurra Ragone, and Eugenio Di Sciascio. 2023. Auditing fairness under unawareness through counterfactual reasoning. *Inf. Process. Manag.* 60, 2 (2023), 103224.

[22] Giandomenico Cornacchia, Francesco M. Donini, Fedelucio Narducci, Claudio Pomo, and Azzurra Ragone. 2021. Explanation in Multi-Stakeholder Recommendation for Enterprise Decision Support Systems. In *CAiSE Workshops (Lecture Notes in Business Information Processing, Vol. 423)*. Springer, 39–47.

[23] Giandomenico Cornacchia, Fedelucio Narducci, and Azzurra Ragone. 2021. A General Model for Fair and Explainable Recommendation in the Loan Domain (Short paper). In *KaRS/ComplexRec@RecSys (CEUR Workshop Proceedings, Vol. 2960)*. CEUR-WS.org.

[24] Yashar Deldjoo, Tommaso Di Noia, Daniele Malitesta, and Felice Antonio Merra. 2021. A Study on the Relative Importance of Convolutional Neural Networks in Visually-Aware Recommender Systems. In *CVPR Workshops*. Computer Vision Foundation / IEEE, 3961–3967.

[25] Yashar Deldjoo, Tommaso Di Noia, Daniele Malitesta, and Felice Antonio Merra. 2022. Leveraging Content-Style Item Representation for Visual Recommendation. In *ECIR (2) (Lecture Notes in Computer Science, Vol. 13186)*. Springer, 84–92.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. IEEE Computer Society, 770–778.

[27] Ruining He and Julian J. McAuley. 2016. VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback. In *AAAI*. AAAI Press, 144–150.

[28] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yong-Dong Zhang, and Meng Wang. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *SIGIR*. ACM, 639–648.

[29] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *WWW*. ACM, 173–182.

[30] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin W. Wilson. 2017. CNN architectures for large-scale audio classification. In *ICASSP*. IEEE, 131–135.

[31] Dietmar Jannach, Lukas Lerche, Iman Kamehkhosh, and Michael Jugovac. 2015. What recommenders recommend: an analysis of recommendation biases and possible countermeasures. *User Model. User Adapt. Interact.* 25, 5 (2015), 427–491.

[32] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2014. Correcting Popularity Bias by Enhancing Recommendation Neutrality. In *RecSys Posters (CEUR Workshop Proceedings, Vol. 1247)*. CEUR-WS.org.

[33] Taeri Kim, Yeon-Chang Lee, Kijung Shin, and Sang-Wook Kim. 2022. MARIO: Modality-Aware Attention and Modality-Preserving Decoders for Multimedia Recommendation. In *CIKM*. ACM, 993–1002.

[34] Yehuda Koren, Robert M. Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *Computer* 42, 8 (2009), 30–37.

[35] Dominik Kowald and Emanuel Lacic. 2022. Popularity Bias in Collaborative Filtering-Based Multimedia Recommender Systems. In *BIAS (Communications in Computer and Information Science, Vol. 1610)*. Springer, 1–11.

[36] Adit Krishnan, Ashish Sharma, Aravind Sankar, and Hari Sundaram. 2018. An Adversarial Approach to Improve Long-Tail Performance in Neural Collaborative Filtering. In *CIKM*. ACM, 1491–1494.

[37] Zhenfeng Lei, Anwar Ul Haq, Adnan Zeb, Md Suzauddola, and Defu Zhang. 2021. Is the suggested food your desired?: Multi-modal recipe recommendation with demand-based knowledge graph. *Expert Syst. Appl.* 186 (2021), 115708.

[38] Jiao Li, Xing Xu, Wei Yu, Fumin Shen, Zuo Cao, Kai Zuo, and Heng Tao Shen. 2021. Hybrid Fusion with Intra- and Cross-Modality Attention for Image-Recipe Retrieval. In *SIGIR*. ACM, 244–254.

[39] Fan Liu, Zhiyong Cheng, Changchang Sun, Yinglong Wang, Liqiang Nie, and Mohan S. Kankanhalli. 2019. User Diverse Preference Modeling by Multimodal Attentive Metric Learning. In *ACM Multimedia*. ACM, 1526–1534.

[40] Xiaohao Liu, Zhulin Tao, Jiahong Shao, Lifang Yang, and Xianglin Huang. 2022. EliMRec: Eliminating Single-modal Bias in Multimedia Recommendation. In *ACM Multimedia*. ACM, 687–695.

[41] Yong Liu, Susen Yang, Chenyi Lei, Guoxin Wang, Haihong Tang, Juyong Zhang, Aixin Sun, and Chunyan Miao. 2021. Pre-training Graph Transformer with Multimodal Side Information for Recommendation. In *ACM Multimedia*. ACM, 2853–2861.

[42] Zhuang Liu, Yunpu Ma, Matthias Schubert, Yuanxin Ouyang, and Zhang Xiong. 2022. Multi-Modal Contrastive Pre-training for Recommendation. In *ICMR*. ACM, 99–108.

[43] Daniele Malitesta, Giandomenico Cornacchia, Claudio Pomo, and Tommaso Di Noia. 2023. Disentangling the Performance Puzzle of Multimodal-aware Recommender Systems. In *EvalRS@KDD (CEUR Workshop Proceedings, Vol. 3450)*. CEUR-WS.org.

[44] Kelong Mao, Jieming Zhu, Jinpeng Wang, Quanyu Dai, Zhenhua Dong, Xi Xiao, and Xiuqiang He. 2021. SimpleX: A Simple and Strong Baseline for Collaborative Filtering. In *CIKM*. ACM, 1243–1252.

[45] Kelong Mao, Jieming Zhu, Xi Xiao, Biao Lu, Zhaowei Wang, and Xiuqiang He. 2021. UltraGCN: Ultra Simplification of Graph Convolutional Networks for Recommendation. In *CIKM*. ACM, 1253–1262.

[46] Julian J. McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-Based Recommendations on Styles and Substitutes. In *SIGIR*. ACM, 43–52.

[47] Weiqing Min, Shuqiang Jiang, and Ramesh C. Jain. 2020. Food Recommendation: Framework, Existing Solutions, and Challenges. *IEEE Trans. Multim.* 22, 10 (2020), 2659–2671.

[48] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. Multimodal Deep Learning. In *ICML*. Omnipress, 689–696.

[49] Sergio Oramas, Oriol Nieto, Mohamed Sordo, and Xavier Serra. 2017. A Deep Multimodal Approach for Cold-start Music Recommendation. In *DLRS@RecSys*. ACM, 32–37.

[50] Shaowen Peng, Kazunari Sugiyama, and Tsunenori Mine. 2022. Less is More: Reweighting Important Spectral Graph Features for Recommendation. In *SIGIR*. ACM, 1273–1282.

[51] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *EMNLP/IJCNLP (1)*. Association for Computational Linguistics, 3980–3990.

[52] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *UAI*.

[53] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2009. The Graph Neural Network Model. *IEEE Trans. Neural Networks* 20, 1 (2009), 61–80.

[54] Zhulin Tao, Yinwei Wei, Xiang Wang, Xiangnan He, Xianglin Huang, and Tat-Seng Chua. 2020. MGAT: Multimodal Graph Attention Network for Recommendation. *Inf. Process. Manag.* 57, 5 (2020), 102277.
[55] Kunal Vaswani, Yudhik Agrawal, and Vinoo Alluri. 2021. Multimodal Fusion Based Attentive Networks for Sequential Music Recommendation. In *BigMM*. IEEE, 25–32.
[56] Dhruv Verma, Kshitij Gulati, Vasu Goel, and Rajiv Ratn Shah. 2020. Fashionist: Personalising Outfit Recommendation for Cold-Start Scenarios. In *ACM Multimedia*. ACM, 4527–4529.
[57] Qifan Wang, Yinwei Wei, Jianhua Yin, Jianlong Wu, Xuemeng Song, and Liqiang Nie. 2023. DualGNN: Dual Graph Neural Network for Multimedia Recommendation. *IEEE Trans. Multim.* 25 (2023), 1074–1084.
[58] Wenjie Wang, Ling-Yu Duan, Hao Jiang, Peiguang Jing, Xuemeng Song, and Liqiang Nie. 2021. Market2Dish: Health-aware Food Recommendation. *ACM Trans. Multim. Comput. Commun. Appl.* 17, 1 (2021), 33:1–33:19.
[59] Wenjie Wang, Fuli Feng, Xiangnan He, Xiang Wang, and Tat-Seng Chua. 2021. Deconfounded Recommendation for Alleviating Bias Amplification. In *KDD*. ACM, 1717–1725.
[60] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural Graph Collaborative Filtering. In *SIGIR*. ACM, 165–174.
[61] Wei Wei, Chao Huang, Lianghao Xia, and Chuxu Zhang. 2023. Multi-Modal Self-Supervised Learning for Recommendation. In *WWW*. ACM, 790–800.
[62] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2020. Graph-Refined Convolutional Network for Multimedia Recommendation with Implicit Feedback. In *ACM Multimedia*. ACM, 3541–3549.
[63] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal Graph Convolution Network for Personalized Recommendation of Micro-video. In *ACM Multimedia*. ACM, 1437–1445.
[64] Zixuan Yi, Xi Wang, Iadh Ounis, and Craig MacDonald. 2022. Multi-modal Graph Contrastive Learning for Micro-video Recommendation. In *SIGIR*. ACM, 1807–1811.
[65] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L. Hamilton, and Jure Leskovec. 2018. Graph Convolutional Neural Networks for Web-Scale Recommender Systems. In *KDD*. ACM, 974–983.
[66] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. 2021. Mining Latent Structures for Multimedia Recommendation. In *ACM Multimedia*. ACM, 3872–3880.
[67] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. 2021. Causal Intervention for Leveraging Popularity Bias in Recommendation. In *SIGIR*. ACM, 11–20.
[68] Yu Zheng, Chen Gao, Xiang Li, Xiangnan He, Yong Li, and Depeng Jin. 2021. Disentangling User Interest and Conformity for Recommendation with Causal Embedding. In *WWW*. ACM / IW3C2, 2980–2991.
[69] Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. 2023. Bootstrap Latent Representations for Multi-modal Recommendation. In *WWW*. ACM, 845–854.