Formalizing Multimedia Recommendation through Multimodal Deep Learning

DANIELE MALITESTA*, Université Paris-Saclay, CentraleSupélec, Inria, France
GIANDOMENICO CORNACCHIA†, IBM Research Europe, Ireland
CLAUDIO POMO, Dipartimento di Ingegneria Elettrica e dell'Informazione, Politecnico di Bari, Italy
FELICE ANTONIO MERRA‡, Amazon Science, Germany
TOMMASO DI NOIA, Dipartimento di Ingegneria Elettrica e dell'Informazione, Politecnico di Bari, Italy
EUGENIO DI SCIASCIO, Dipartimento di Ingegneria Elettrica e dell'Informazione, Politecnico di Bari, Italy

Recommender systems (RSs) provide customers with a personalized navigation experience within the vast catalogs of products and services offered on popular online platforms. Despite the substantial success of traditional RSs, recommendation remains a highly challenging task, especially in specific scenarios and domains. For example, human affinity for items described through multimedia content (e.g., images, audio, and text), such as fashion products, movies, and music, is multi-faceted and primarily driven by their diverse characteristics. Therefore, by leveraging all available signals in such scenarios, multimodality enables us to tap into richer information sources and construct more refined user/item profiles for recommendations. Despite the growing number of multimodal techniques proposed for multimedia recommendation, the existing literature lacks a shared and universal schema for modeling and solving the recommendation problem through the lens of multimodality. Given the recent advances in multimodal deep learning for other tasks and scenarios where precise theoretical and applicative procedures exist, we also consider it imperative to formalize a general multimodal schema for multimedia recommendation. In this work, we first provide a comprehensive literature review of multimodal approaches for multimedia recommendation from the last eight years. Second, we outline the theoretical foundations of a multimodal pipeline for multimedia recommendation by identifying and formally organizing recurring solutions/patterns; at the same time, we demonstrate its rationale by conceptually applying it to selected state-of-the-art approaches in multimedia recommendation. Third, we conduct a benchmarking analysis of recent algorithms for multimedia recommendation within Elliot, a rigorous framework for evaluating recommender systems, where we re-implement such multimedia recommendation approaches. Finally, we highlight the significant unresolved challenges in multimodal deep learning for multimedia recommendation and suggest possible avenues for addressing them. The primary aim of this work is to provide guidelines for designing and implementing the next generation of multimodal approaches in multimedia recommendation.

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

Additional Key Words and Phrases: Multimodal Deep Learning, Multimedia Recommender Systems, Benchmarking

ACM Reference Format:
Daniele Malitesta, Giandomenico Cornacchia, Claudio Pomo, Felice Antonio Merra, Tommaso Di Noia, and Eugenio Di Sciascio. 2024. Formalizing Multimedia Recommendation through Multimodal Deep Learning. *ACM Trans. Recomm. Syst.* 1, 1, Article 1 (January 2024), 32 pages. https://doi.org/10.1145/3662738

*Work done while at Politecnico di Bari as a PhD student.
†Work done while at Politecnico di Bari before joining IBM.
‡Work done while at Politecnico di Bari before joining Amazon.

Authors' addresses: Daniele Malitesta, Université Paris-Saclay, CentraleSupélec, Inria, Gif-sur-Yvette, France, daniele.malitesta@centralesupelec.fr; Giandomenico Cornacchia, IBM Research Europe, Dublin, Ireland, giandomenico.cornacchia1@ibm.com; Claudio Pomo, Dipartimento di Ingegneria Elettrica e dell'Informazione, Politecnico di Bari, Bari, Italy, claudio.pomo@poliba.it; Felice Antonio Merra, Amazon Science, Berlin, Germany, felmerra@amazon.de; Tommaso Di Noia, Dipartimento di Ingegneria Elettrica e dell'Informazione, Politecnico di Bari, Bari, Italy, tommaso.dinoia@poliba.it; Eugenio Di Sciascio, Dipartimento di Ingegneria Elettrica e dell'Informazione, Politecnico di Bari, Bari, Italy, eugenio.disciascio@poliba.it.

1 INTRODUCTION

Over the last few decades, companies have increasingly been developing online platforms to reach their customers and offer them a more comprehensive selection of personalized products and services in various domains, from food and fashion to e-commerce and tourism. Recommender systems (RSs) are among the prominent technologies that work behind the scenes of such platforms to unveil the implicit preference patterns within the intricate set of users and items and curate the presentation of a list of products that customers may enjoy. Such technologies are an essential element of all major Internet businesses, driving up to 35% of Amazon's sales [74] and more than 80% of Netflix's catalog [24].

There exist recommendation scenarios, such as multimedia recommendation, where items naturally come with additional side information that may complement the knowledge conveyed by the historical user/item interaction matrix. Multimedia recommendation [30, 31] is the task of recommending products or services either described through multimedia content (e.g., a fashion item with a product image and description) or multimedia content themselves (e.g., a movie with its visuals, soundtrack, and subtitles). In such a context, any recorded user/item interaction may hide multiple possible reasons why that interaction occurred. A user could be interested in buying a fashion item due to the description on the item page and could enjoy a movie because of its soundtrack. Understanding these patterns means modeling users' and items' profiles through the *multi-faceted* aspects of their interactions.

Our experience of daily life is intrinsically *multimodal*. We interact with objects surrounding us through our five senses. For instance, watching a movie can involve three senses (i.e., modalities): we watch it (*visual* modality) while listening to the dialogues (*audio* modality) and possibly reading its subtitles (*textual* modality). Multimodal learning has been one of the hot topics in deep learning for some years now, addressing applicative domains such as medical imaging [10, 36, 43, 103], autonomous driving [12, 50, 121, 142], speech/emotion recognition [59, 70, 87, 88], multimedia retrieval [14, 45, 46, 60], and, only recently, multimodal large language modelling [127]. Given the success and popularity it has encountered, some works have tried to outline, categorize, and formalize the core concepts behind multimodality in deep learning [7, 8, 84]. Remarkably, the literature recognizes five steps and challenges when designing a multimodal deep learning pipeline [7]: *representation*, *translation*, *alignment*, *fusion*, and *co-learning*.

Similarly to the cited domains and applications, approaches in multimedia recommendation have been shown to effectively apply multimodal deep learning techniques to the recommendation task. The idea is to model users' and items' profiles through the different modalities and suitably capture the multi-faceted nature of their interconnections. Recent works in the literature have brought multimodality to multimedia recommendation [67, 92, 106, 143] tackling (just to mention a few) micro-video recommendation [13, 21, 118], food recommendation [58, 82, 112], outfit fashion compatibility [19, 124, 132], and artist/song recommendation [23, 86, 109]. However, and differently from the other outlined domains and scenarios, recommendation lacks a *shared* theoretical and applicative formalization to align the multimedia recommendation problem with the same formal pipeline proposed in multimodal deep learning [7, 8, 84].

For these reasons, in this work, we first review the most popular and recent state-of-the-art approaches in multimedia recommendation. Indeed, it emerges that three main design choices are involved when proposing novel multimedia recommender systems leveraging multimodality: (i) ***Which*** modalities to suitably describe the user/item input data; (ii) ***How*** to extract and process meaningful multimodal representations; (iii) ***When*** to integrate and inject multimodal data into the training/inference steps. While observing that many multimedia recommendation approaches are rarely aligned on the techniques to adopt for (i), (ii), and (iii), we maintain this could limit the future development of novel solutions in the field. This is true since each work claims to advance with respect to the state-of-the-art but it becomes cumbersome to distinguish which conceptual and implementation *strategies* are contributing the most [75].

Thus, inspired by the multimodal pipeline formalized in multimodal deep learning [7, 8, 84], we try to align the same schema with the three design choices recognized above. Our objective is to define a conceptual and theoretical schema that uses multimodality to encompass and summarize the most diffused solutions/patterns in the multimedia recommendation literature. To the best of our knowledge, this represents the first attempt that, differently from similar works in the literature [67, 143], ***formalizes*** multimedia recommendation through the core concepts theorized in multimodal deep learning [7, 8, 84].

To sum up, we aim to answer the following research questions (RQs):

RQ1. *Which are the main solutions in the related literature?* We review existing works in multimedia recommendation adopting multimodal learning techniques, highlighting common and different architectural choices; in this respect, we categorize the reviewed papers according to the type of multimodal input (i.e., ***Which***), the technique for features processing (i.e., ***How***), and the moment to integrate modalities (i.e., ***When***).

RQ2. *Can we summarize the observed strategies into a shared formal schema, which is also conceptually applicable to existing recommendation scenarios?* On such basis, and following the related literature on multimodal deep learning, we revisit the multimedia recommendation task under the lens of multimodal deep learning; by mapping the multimodal pipeline outlined in [7, 8, 84] to the threefold categorization from RQ1, we provide the general formulations for a formal schema involving three steps: multimodal input data, multimodal feature processing, and multimodal feature fusion. To conceptually validate the rationale of the introduced theoretical formulations, we also apply them to four selected recommendation systems spanning various tasks in multimedia recommendation.

RQ3. *Can we integrate the multimodal schema into existing recommendation frameworks?* We use Elliot [2], a rigorous framework for the reproducibility of recommender systems, and integrate our multimodal schema into it to benchmark six state-of-the-art multimodal techniques for multimedia recommendation (i.e., VBPR [41], MMGCN [118], GRCN [117], LATTICE [134], BM3 [145], and FREEDOM [144]) and test them against four popular recommendation solutions which do not exploit multimodal features (i.e., BPR [90], NGCF [113], LightGCN [42], and SGL [120]); the measured recommendation metrics, accounting also for beyond-accuracy evaluation, open to future research directions.

RQ4. *Which are the next challenges in multimodal learning for recommendation?* Driven by the previous findings, we outline technical and conceptual challenges aimed to provide guidelines for future research in the field.

The rest of this paper is organized as follows. Section 2 provides a comprehensive overview of the most recent works exploiting multimodality for multimedia recommendation, and highlights the main differences between this work and similar works in the literature. Then, in Section 3, we present our formal schema which tries to embody and generalize the different solutions reviewed in the literature, and show how to apply it to a valuable sample of state-of-the-art approaches. Under the same light, in Section 4, we present the implementation of our formal schema, which we adopt to benchmark selected models from the literature. Furthermore, in Section 5, we take advantage of the lesson learned from the literature and the benchmarking study to outline existing technical challenges that may be addressed in future directions (i.e., Section 6). Finally, in Section 7, we sum up the main contributions of this paper. To foster the reproducibility of the current work, we release a GitHub repository with all the reviewed papers, along with the benchmarking framework and results at the following link: https://github.com/sisinflab/Formal-MultiMod-Rec.

2 LITERATURE REVIEW (RQ1)

In this section, we present a literature review on recent multimodal applications for the task of multimedia recommendation. Table 1 reports 43 papers collected from the proceedings of top-tier conferences and journals over the last eight years. A careful review and analysis aimed at outlining recurrent schematic and observed patterns suggests categorizing the retrieved papers according to three key questions:

- ***Which*** modalities to choose for the input data?
- ***How*** to process multimodal features in terms of feature extraction and representation?
- ***When*** to fuse the different modalities to integrate them into the final recommendation framework?

To collect all reviewed papers, we also include a public GitHub repository[1] to access their direct DOIs. We intend to update this repository with the most recent works leveraging multimodality for multimedia recommendation.

2.1 *Which* modalities?

In multimedia recommendation scenarios, input data generally comes in at least two of the three most common modalities in literature, namely *Visual*, *Textual*, and *Audio* modalities. As evident from the collected papers, the vast majority of works consider the visual and textual modalities, which mainly refer to product images and descriptions (e.g., [26, 32, 38, 64, 132, 134]), respectively, while fewer examples leverage such modalities to describe video frames and captions (e.g., [49, 93, 112]) or users' social media interactions through uploaded photographs together with texts (e.g., [18, 123, 136]). Another emerging trend from the literature is that audio is by far the most underrepresented modality, and it is usually coupled with the textual one to describe music in the form of audio signals and songs' descriptions (e.g., [86, 109]). Conversely, the related literature shows that the audio modality is frequently exploited for video input data (e.g., [104, 111, 117, 118, 126]) which is also the unique scenario involving all modalities.

The observed disparity in data modalities is not only linked to the specific task the various approaches address (e.g., product, song, or micro-video recommendation) but it is also found in each modality's different availability. In this respect, for example, datasets collecting user-item interactions on e-commerce platforms (e.g., the Amazon reviews dataset or IQON300) are more easily accessible than the ones involving social media videos. For instance, one may consider that a version of the TikTok dataset (introduced in [118]) has been made available with pre-trained multimodal features involving visual, audio, and textual modalities only recently [115]. This modality *misalignment* is among the most discussed challenges in the community, so we decide to dedicate a section to it later (refer to Section 5.1).

2.2 *How* to process modalities?

Once modalities have been selected for data inputs, two primary operations usually get involved in processing the multimodal data to be fed into the recommender system. First, high-level features are extracted from each of the available modalities. Interestingly, early approaches adopt handcrafted feature extraction (HFE) strategies (e.g., color histograms) as described in [18, 49, 61, 85]. However, with the outbreak and the increasing popularity encountered by deep learning and deep neural models for image and text classification, object detection, and speech recognition, trainable feature extractors (TFE) soon became the de facto standard in the learning of latent features from the input data. In this respect, the literature [27] indicates that the common approach is to use the activation of one of the final hidden layers of deep neural networks. For instance, the authors of [68, 69, 82, 97, 110, 115, 130] exploit features extracted from deep networks. Furthermore, we categorize TFE strategies based on the use of *Pretrained* deep networks and

[1] https://github.com/sisinflab/Formal-MultiMod-Rec.

Table 1. Overview of the core questions which arise when modelling a multimedia recommender system based upon multimodality, as observed in the most updated literature. HFE: Handcrafted Feature Extraction, TFE: Trainable Feature Extraction, MMR: Multimodal Representation.

Papers	**Year**	**Modalities (*Which?*)**			**Feature Processing (*How?*)**					**Fusion (*When?*)**	
		Visual	*Textual*	*Audio*	HFE	TFE		MMR		*Early*	*Late*
						Pretrained	*End-to-End*	*Joint*	*Coordinate*		
Ferracani et al. [33]		✓	✓			✓		✓			
Jia et al. [49]	2015	✓	✓		✓			✓			
Li et al. [61]		✓		✓	✓			✓			
Nie et al. [85]	2016	✓	✓		✓				✓	✓	
Chen et al. [18]		✓	✓		✓			✓			
Han et al. [38]		✓	✓				✓		✓	✓	
Oramas et al. [86]	2017		✓	✓			✓		✓	✓	
Zhang et al. [136]		✓	✓				✓		✓	✓	
Ying et al. [128]	2018	✓	✓			✓		✓			
Wang et al. [110]		✓	✓			✓		✓			
Liu et al. [64]		✓	✓			✓		✓			
Chen et al. [20]		✓	✓			✓					
Wei et al. [118]		✓	✓	✓		✓			✓	✓	
Cheng et al. [22]	2019	✓	✓			✓					
Dong et al. [32]		✓	✓			✓			✓	✓	
Chen et al. [19]		✓	✓			✓		✓			
Yu et al. [130]		✓	✓			✓	✓	✓	✓	✓	
Cui et al. [26]		✓	✓			✓			✓		✓
Wei et al. [117]		✓	✓	✓		✓			✓	✓	
Sun et al. [97]		✓	✓			✓		✓			
Chen and Li [16]		✓	✓			✓		✓			
Min et al. [82]	2020	✓	✓	✓		✓		✓			
Shen et al. [95]		✓	✓		✓	✓			✓	✓	
Yang et al. [124]		✓	✓			✓			✓		✓
Tao et al. [104]		✓	✓	✓		✓			✓	✓	
Yang et al. [123]		✓	✓			✓			✓	✓	
Sang et al. [93]		✓	✓			✓			✓	✓	
Liu et al. [68]		✓	✓	✓		✓			✓	✓	
Zhang et al. [134]	2021	✓	✓			✓			✓	✓	
Vaswani et al. [109]			✓	✓		✓			✓	✓	
Lei et al. [58]		✓	✓	✓		✓	✓				
Wang et al. [112]		✓	✓			✓		✓			
Zhan et al. [132]		✓	✓			✓		✓			
Wu et al. [119]		✓	✓			✓			✓	✓	
Yi and Chen [125]		✓	✓			✓			✓		✓
Yi et al. [126]	2022	✓	✓	✓		✓			✓		✓
Liu et al. [69]		✓	✓			✓			✓	✓	
Mu et al. [83]		✓	✓			✓			✓	✓	
Chen et al. [15]		✓	✓	✓		✓			✓	✓	
Zhou and Shen [144]		✓	✓			✓			✓		✓
Wang et al. [111]	2023	✓	✓	✓		✓			✓	✓	
Wei et al. [115]		✓	✓	✓		✓			✓		✓
Zhou et al. [145]		✓	✓			✓			✓		✓

End-to-End learned models. The former refer to the possibility of transferring the learned knowledge of already-trained deep networks to different domains, tasks, and datasets (e.g., see [102]), whereas the latter usually leverage custom deep neural networks trained in the downstream recommendation task. As evident from the collected papers, the pre-trained solution (e.g., [16, 20, 32, 33, 93, 95, 117, 128]) widely surpasses the end-to-end one (e.g., [38, 86, 136]) in terms of popularity, as the adoption of ready-to-use embedded features obtained from state-of-the-art deep learning models represents a more efficient and convenient approach than performing computationally-expensive and data-eager trained feature extractions. Nevertheless, an argument might be made that using features extracted through models already trained on different datasets and tasks could limit their expressiveness regarding the actual multimedia recommendation task. For this reason, we deepen into the issue in Section 5.2, trying to propose viable solutions in Section 6.1.

The second operation involved in the feature processing phase regards the implementation of a multimodal representation (MMR) solution to establish relations among the extracted modalities. We recognize two main approaches, namely, either combining all modalities so that they belong to a unique representation (*Joint*) or keeping them separated to leverage the different influence they may have on recommendation (*Coordinate*). From the collected papers, it follows that both the former (e.g., [15, 19, 33, 83, 97, 128, 132]), and the latter (e.g., [85, 118, 124, 125, 134]) are almost equally preferred; however, the coordinate multimodal representation is slightly more popular as learning different representations for each involved modality may help unveil the specific contribution it brings to the final personalized recommendation. Indeed, this could support *explainability*, which is among the hottest topics in the community [137, 138], and especially in multimedia recommendation scenarios, where user-item interactions may, by nature, be driven by non-evident and sometimes contrasting users' preferences and tastes [17, 20, 117]. Finally, the authors from [20, 22] do not integrate any multimodal representation approach since they exploit multimodality only for the optimization of the loss but not to predict user-item preferences.

2.3 *When* to fuse modalities?

The last stage in the multimodal pipeline deals with the fusion of the different processed modalities so that they can be eventually integrated into the recommendation outcome as a single representation of multiple *coordinated* modalities. This process may take place *before* or *after* the prediction of the user-item preference score. On this basis, the former and the latter approaches are usually known as *Early* (e.g., [109, 118]) and *Late* (e.g., [26, 125]) fusion, respectively. It is worth pointing out that some solutions recognize a third strategy (i.e., *Hybrid* fusion) that combines the two versions mentioned above, but for the sake of simplicity, we decide to categorize the works performing this kind of multimodal fusion as a particular case of late fusion. Additionally, we recognize that several approaches from the literature do not provide a precise differentiation between joint multimodal representation and early fusion. To better clarify this technical aspect, we propose to consider *fusion* as an optional operation that takes place after the feature processing phase only in the case of *coordinate* multimodal representation (you may refer to Section 5.3). Indeed, as evident from the table, *Joint* multimodal representation and *Early*/*Late* fusion never occur in the same approach. What is more, we observe that early fusion, employed, for instance, in [15, 32, 69, 83, 85, 93, 95, 119, 134, 136], is more popular than late fusion, used in [26, 115, 124–126, 144, 145]. Motivating this tendency is an unanswered research question that we leave as a possible open issue to impact the design of recommender systems leveraging multimodality. In this respect, you may refer to Section 5.4 for our discussion on the current challenges about modality fusion, and to Section 6.3 where we sketch possible future research directions.

2.4 Similar works to this paper

For the sake of completeness, we review the current literature works that provide similar contributions to ours to outline the main differences. As already mentioned, pioneer works such as [7, 8, 84] introduce and formalize (for the first time) the core concepts and ideas behind the field of multimodal deep learning. After that, the recent years have seen a growing interest in systematically reviewing and schematizing techniques for multimodal fusion [34], spanning different application domains such as medicine [105], conversational artificial intelligence [100], and visual content syntesis [139], up to addressing complex and novel machine learning strategies including meta-learning [73]. Although the cited works share similar rationales to ours, their focus is more general (e.g., deep learning) or heterogeneous (e.g., medicine) with respect to the multimedia recommendation task.

In the recommendation domain, the study presented in [82] is among the closest and most influential works to our proposal in the intention of introducing a unified framework for food recommendation which leverages the concept of multimodality; however, the work is different from ours in that: (i) it only addresses the task of food recommendation, and (ii) it does not provide either mathematical formalizations or benchmarking analyses of the proposed multimodal pipeline. Furthermore, it is worth recalling two surveys regarding the topic of multimodal recommender systems [67, 143] on arXiv at the moment of this submission. Among the two, the work presented in [143] shows the major similarities to our paper, especially when recognizing a multimodal pipeline for multimedia recommendation and providing an extensive benchmarking study on numerous multimodal approaches in the literature. Nevertheless, our work stands out for the following novel contributions: (i) we systematically follow the multimodal pipeline outlined in [7, 8, 84] in the attempt to adapt it to the three main questions arising in the multimedia recommendation literature, namely, ***Which?***, ***How?***, and ***When?***; (ii) we provide mathematical formalizations for each step of the proposed multimodal pipeline to sketch a formal schema for the next generation of multimodal approaches addressing multimedia recommendation; (iii) we identify a wider set of challenges regarding each step in the multimodal pipeline, and try to provide solutions to tackle them all; (iv) given the recent urge to evaluate the performance of recommender systems under other objectives apart from accuracy [1, 6, 11, 48, 94, 107, 108], and specifically when investigating the impact of multimodality on novelty and diversity [75] or popularity bias [76], our benchmarking analysis (exploiting the evaluation pipeline from Elliot [2]) represents the first effort to rigorously assess such large-scale performance measures in multimedia recommendation, opening to further research questions.

3 A FORMAL MULTIMODAL SCHEMA FOR MULTIMEDIA RECOMMENDATION (RQ2)

As previously outlined, the literature shows recurrent schematic patterns in adopting multimodal techniques for the task of multimedia recommendation. However, when considering the latest solutions in the field (Section 2) it appears evident that, differently from what happens for other applicative domains in machine learning, such approaches do not seem to follow any shared and officially recognized formal schema aligned with the principles of multimodal deep learning [7, 8, 84].

To sort things out, in this section, we propose to formally revisit multimedia recommendation under the lens of multimodal deep learning (Figure 1). First, we formalize the standard recommendation task. Then, we theoretically give answers to the core questions previously outlined, namely: ***Which*** multimodal input data to adopt, ***How*** to extract multimodal features and set relationships among them, and ***When*** to fuse modalities. Finally, we specify the multimedia recommendation task through multimodality.

Along with the formalization of the unified framework, we show how to apply the introduced theoretical notions to four selected state-of-the-art multimedia recommendation models (refer to the **Applications** paragraph). Specifically, we choose the proposed examples spanning a wide range of tasks, namely micro-video recommendation [118], food recommendation [112], outfit fashion compatibility [38], and artist/song recommendation [86]:

(1) **Micro-video recommendation.** Wei et al. [118] build a bipartite user-item graph for micro-video personalized recommendation. The idea behind the approach is to exploit high-order users-items relations leveraging the

multimodal nature of recommended items (i.e., micro-videos), to which users may experience different attitudes. The authors adopt a graph convolutional network [54] to refine user and item embeddings (conditioned on the graph topology).

(2) **Food recommendation.** Wang et al. [112] introduce a tripartite framework for food recommendation whose pipeline involves the retrieval of recipes according to user-generated videos, the profiling of users based upon their social media interactions, and the final health-aware food recommendation of recipes.

(3) **Outfit fashion compatibility.** Han et al. [38] propose a framework to recommend the next fashion item that matches a set of already chosen ones to produce a visually appealing outfit. The authors address the task by considering the items composing a fashionable outfit as a temporal sequence, so they leverage a bidirectional LSTM [37].

(4) **Artist and song recommendation.** The approach introduced by Oramas et al. [86] deals with the task of music recommendation. Specifically, the authors propose to divide the problem into artist and song recommendations by learning their separate embeddings and leveraging the textual artist biography and audio spectrogram as inputs.

Note that, in the following, we use the **bold** notation only when we explicitly define a vector for which we indicate its elements (i.e., scalars or other vectors).

3.1 Classical recommendation task

We consider users, items, and user-item interactions as the inputs to the recommender system. We denote with $u \in \mathcal{U}$, $i \in \mathcal{I}$, and $r \in \mathcal{R}$ a user, an item, and a user-item interaction, respectively. To ease the notation, we say $x \in \mathcal{X}$ is a general input to the system, with $\mathcal{X} = \mathcal{U} \cup \mathcal{I} \cup \mathcal{R}$. Given a set of input data $\mathcal{X}$, and defined $\rho(\cdot)$ as the preference score prediction function, a recommender system aims to build a top@k list of items maximizing the following posterior probability ($prob$):

$$\hat{\Theta}_\rho = \arg\max_{\mathbf{\Theta}_\rho} prob(\mathbf{\Theta}_\rho \mid \mathcal{X}), \tag{1}$$

where $\mathbf{\Theta}_\rho = [\theta_\rho^{(0)}, \theta_\rho^{(1)}, \ldots, \theta_\rho^{(|\mathcal{W}_\rho|-1)}]$ is the vector collecting all weights for the inference function $\rho(\cdot)$, $\mathcal{W}_\rho = \{\theta_\rho^{(0)}, \theta_\rho^{(1)}, \ldots, \theta_\rho^{(|\mathcal{W}_\rho|-1)}\}$ is the set of such weights, and $|\mathcal{W}_\rho|$ its cardinality.

For instance, in the case of latent factor models (e.g., matrix factorization [55]), which are the most popular ones in the literature, the set of trainable weights $\mathcal{W}_\rho$ involves the user and item embeddings: $\mathcal{W}_\rho = \{\theta_\rho^{(u_0)}, \theta_\rho^{(u_1)}, \ldots, \theta_\rho^{(u_{|\mathcal{U}|-1})}, \theta_\rho^{(i_0)}, \theta_\rho^{(i_1)}, \ldots, \theta_\rho^{(i_{|\mathcal{I}|-1})}\}$. To ease the notation, from now on, we will indicate with $\theta_\rho^{\mathcal{U}}$ and $\theta_\rho^{\mathcal{I}}$ as the embeddings of any user and item in the recommendation system, respectively. Thus, in the case of classical recommender system leveraging latent factor models, the inference function may be indicated as:

$$\hat{y} = \rho(\theta_\rho^{\mathcal{U}}, \theta_\rho^{\mathcal{I}}, \mathcal{X}), \tag{2}$$

where $\hat{y}$ is the predicted value through the inference function $\rho(\cdot)$, such as a rating score for any user-item pair.

3.2 Multimodal input data

As shown in Figure 1, the first step of our multimodal schema is to identify input modalities. A common list of modalities for each input data (i.e., user, item, user-item interaction) in multimedia scenarios may be defined as follows:

- visual ($\mathbf{v}$), e.g., images, video frames;
- textual ($\mathbf{t}$), e.g., image captions, video subtitles, song lyrics, reviews;

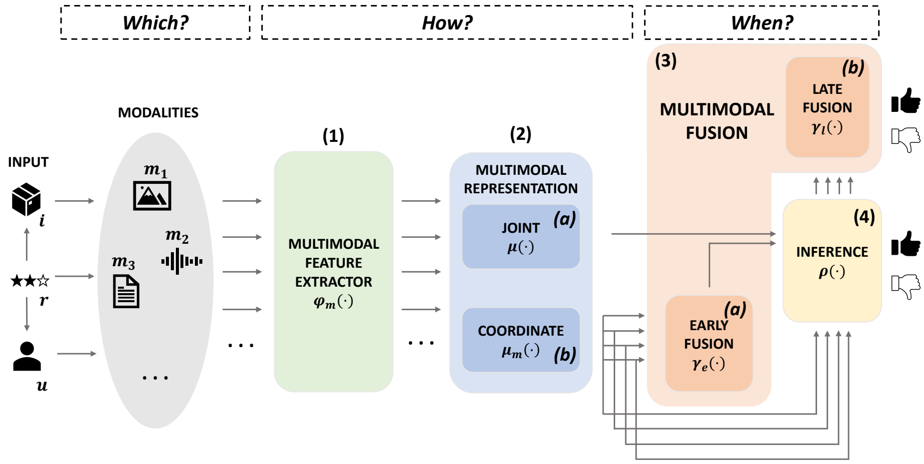

Fig. 1. Our multimodal schema for multimedia recommendation. After (1) a modality-aware feature extraction, the extracted features may be either directly represented into a unique latent space (2*a*) or projected into a different latent space for each modality (2*b*). While in the former case, the multimodal representation is used to produce a prediction (4), in the latter case, all modalities must undergo a fusion phase (3). In the early fusion (3*a*), we produce a final representation that is used for prediction (4). Otherwise, we first produce a different prediction for each modality (4), and then we fuse them (late fusion) into a single predicted value (3*b*).

- audio ($\mathbf{a}$), e.g., songs, podcasts, movie soundtracks.

Formally, we define $m \in \mathcal{M}$ as an admissible modality for the system (i.e., $\mathcal{M} = \{\mathbf{v}, \mathbf{t}, \mathbf{a}\}$). We should mention that data may come with all such modalities or just a subset. For instance, videos from video streaming platforms (such as Netflix or Amazon Prime Video) have frames ($\mathbf{v}$), subtitles and/or descriptions ($\mathbf{t}$), and an audio track and/or soundtrack ($\mathbf{a}$). Similarly, e-commerce platforms (such as Amazon or Zalando) sell products that may come with photographs ($\mathbf{v}$) and reviews which stand for the textual feedback users express towards those products ($\mathbf{t}$).

Let $x \in \mathcal{X}$ be an input to the recommender system, whose set of available modalities is indicated as $\mathcal{M}_x \subseteq \mathcal{M}$. We represent the *content* data of input x in modality m as $c_x^{(m)}$, with $m \in \mathcal{M}_x$, and the vector of content data for input x in all modalities as $\mathbf{c}_x$. Concerning the examples from above, a video item x may be described through three modalities (i.e., $\mathcal{M}_x = \{\mathbf{v}, \mathbf{a}, \mathbf{t}\}$) and, for example, its visual content data (a frame) is an RGB image indicated as $c_x^{(\mathbf{v})}$. Similarly, a fashion item x may be described through two modalities (i.e., $\mathcal{M}_x = \{\mathbf{v}, \mathbf{t}\}$) and, for example, its textual content data (the description) is a set of words indicated as $c_x^{(\mathbf{t})}$.

Applications

(1) **Micro-video recommendation [118].** Micro-videos (the items) are described via three modalities, namely: *visual* (i.e., frames), *textual* (i.e., user-generated captions and descriptions) and *audio* (i.e., the audio track, that is not always available). It is worth pointing out that also users are described through three embeddings representing how each item modality might influence them differently. Nevertheless, they cannot be formally considered as multimodal input data (we do not report any information about the multimodal input data and feature extraction columns in the table).

(2) **Food recommendation [112].** On the user side, it should be noticed that the input data does not properly follow the above definition we provide about multimodality, as users are profiled only according to the textual description

of their generated tweets. However, we maintain the importance of this example since it represents one of the few approaches in the literature that proposes to model users through a multimodality-like solution. On the other side, items' description is multimodal because it integrates frames of user-generated videos to retrieve recipes from (i.e., *visual* modality) and descriptions of recipe ingredients (i.e., *textual* modality).

(3) **Outfit fashion compatibility [38].** Recommendation is multimodal because the authors adopt both product images (i.e., *visual* modality) and text descriptions of the fashion items extracted from the product details (i.e., *textual* modality).

(4) **Artist and song recommendation [86].** Multimodality is to be found in the item's description, which is based upon artist biography (i.e., *textual* modality) and audio spectrogram derived from songs (i.e., *audio* modality).

3.3 Multimodal feature processing

As in Figure 1, multimodal inputs are processed to be transferred into a low-dimensional representation. This step runs through a multimodal feature extractor and a component that constructs a multimodal feature representation.

3.3.1 *Feature extraction.*

Content input data is generally not exploitable as it is in a recommender model (e.g., the matrix of pixels from an image is not meant to be directly integrated into a recommender). Hence, our schema introduces a *Feature Extractor* (FE) component to extract features, which should follow two principles, being (i) *high-level* (i.e., meaningful for the recommender system) and (ii) *functional* to the final recommendation task. Indeed, choosing the most suitable feature extractor for each modality may affect the performance.

Let $c_x^{(m)}$ be the content data for input x in modality $m \in \mathcal{M}_x$. Then, let $\varphi_m(\cdot)$ be the feature extractor function for the modality m. We define the feature extraction process in the modality m as:

$$\bar{c}_x^{(m)} = \varphi_m(c_x^{(m)}) \quad \forall m \in \mathcal{M}_x, \tag{3}$$

where $\bar{c}_x^{(m)}$ is the extracted feature for input x in modality m. We use the notation $\bar{\mathbf{c}}_x = [\bar{c}_x^{(0)}, \bar{c}_x^{(1)}, \ldots, \bar{c}_x^{(|\mathcal{M}_x|-1)}]$ to refer to the vector of extracted features for input x in all modalities. Generally speaking, $\varphi_m(\cdot)$ may refer either to a handcrafted extractor, HFE (e.g., SIFT and color histogram for visual, and MFCCs for audio), or to a trainable extractor, TFE (e.g., deep learning-based models such as CNNs for visual, audio, and textual). In the latter case, $\varphi_m(\cdot)$ can be either pre-trained or trained end-to-end along with the recommender system. Regarding the extraction of multimodal features for recommendation, the authors in [5, 77] propose Ducho, which aims to standardize and unify the whole process.

APPLICATIONS

(1) **Micro-video recommendation [118].** Visual features are extracted through a pretrained ResNet50 [39] only from key video frames. Textual features are derived from Sentence2Vector [4]. Audio features are extracted using a pre-trained VGGish [44].

(2) **Food recommendation [112].** Users' tweets are the input to what the authors define as a word-class interaction-based recurrent convolutional network (WIRCNN), which involves a recurrent neural network (RNN) and a convolutional neural network (CNN) to classify user tags. As for items, sampled video frames are encoded through a pre-trained VGGNet-19 [96], while textual recipe ingredients are processed via TextCNN [53].

(3) **Outfit fashion compatibility [38].** The visual features of fashion items are extracted from a GoogleNet InceptionV3 [101] pretrained network (the TFE), whose dimensionality is 2048. As for the textual description, each word is a one-hot-encoded vector.

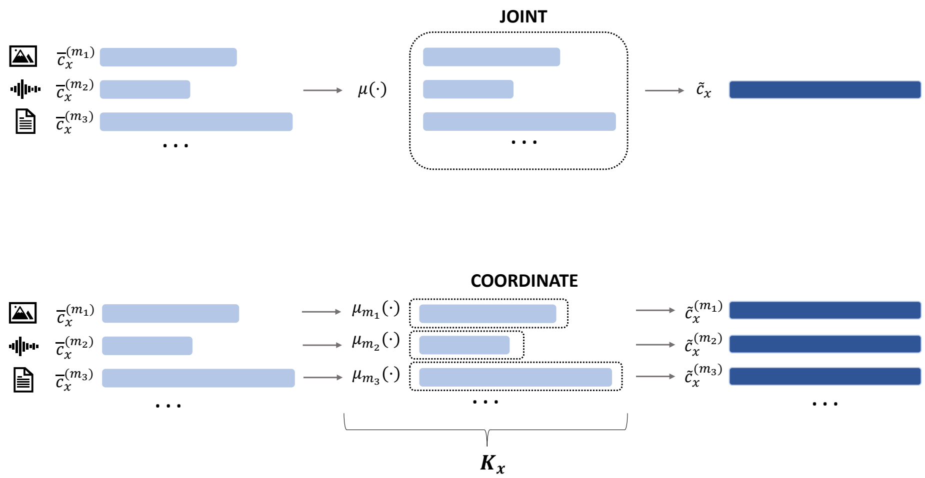

Fig. 2. A visual representation of *Joint* and *Coordinate* multimodal representation (above and below, respectively).

(4) **Artist and song recommendation [86].** Artist biographies are processed through the state-of-the-art CNN, which is re-trained using word2vec word embeddings pre-trained on the Google News dataset [81]. As for the song latent factors, a custom CNN with 256, 512, and 1024 convolutional filters is trained on the time axis, having as output the 4096-dense layer.

3.3.2 Multimodal representation. Once high-level features have been extracted from each modality of the input data, the next step is to design a *Representation* strategy to handle the relationships among modalities and eventually inject such data into the recommender system. As shown in Section 2, the literature follows two main approaches: *Joint* and *Coordinate* (Figure 2). The former relies on projecting multimodal features into a shared latent space to produce a unique final representation (e.g., concatenation is usually the simplest approach). Conversely, the latter involves adopting a different latent space for each modality, with the possibility of setting specific constraints among modalities that are expressed, for instance, through similarity metrics. In the following, we mathematically formalize the two strategies.

Joint representation. Let $\bar{\mathbf{c}}_x$ be the vector of extracted features for input x in all modalities. In the case of *Joint* representation, we assume $\mu(\cdot)$ is the function to produce the multimodal representation of the extracted features. Thus:

$$\tilde{c}_x = \mu(\bar{\mathbf{c}}_x), \tag{4}$$

where $\tilde{c}_x$ is the multimodal representation for input x.

Coordinate representation. Let $\bar{c}_x^{(m)}$ be the extracted feature for input x in modality $m \in \mathcal{M}_x$. In the case of *Coordinate* representation, we assume $\mu_m(\cdot)$ is the multimodal representation function for modality m, and let $\mathcal{K}_x$ be a set of constraints on multimodal representations of input x. Thus, we say:

$$\tilde{c}_x^{(m)} = \mu_m(\bar{c}_x^{(m)}) \textbf{ subject to } \mathcal{K}_x, \textbf{ with } |\mathcal{K}_x| \geq 0, \tag{5}$$

where $\tilde{c}_x^{(m)}$ is the coordinate multimodal representation for input x in modality m. Note that with Equation (5) we are referring to $\tilde{\mathbf{c}}_x = [\tilde{c}_x^{(0)}, \tilde{c}_x^{(1)}, \dots, \tilde{c}_x^{(|\mathcal{M}_x|-1)}]$ as the vector of coordinate multimodal representations for input x in all modalities.

Applications

(1) **Micro-video recommendation [118].** The framework leverages three versions of the same bipartite user-item graph (i.e., one for each micro-video modality). The graph convolutional layer first aggregates the neighborhood features of the ego node and then combines the result of such aggregation with the collaborative embedding and the multimodal representation from the previous iteration. Given the formalism introduced above, we might say this approach goes under the definition of *Coordinate* representation. The model adopts a linear projection for each modality to map the input into a modality-specific latent space, both in the aggregation and combination steps. No explicit constraints are introduced.
(2) **Food recommendation [112].** Given that the user profile is not multimodal *per se*, we do not recognize any multimodal representation stage. On the item side, the multimodal representation is *Coordinate*, but no particular operation is performed on the extracted multimodal features.
(3) **Outfit fashion compatibility [38].** Visual and textual extracted features are projected into a unique latent space, whose dimensionality is 512. According to the earlier formalism, this approach follows a *Joint* multimodal representation. On the one hand, 2048-dimensional visual features are compressed into a 512-dimensional embedding through a fully connected neural network layer, which is trained end-to-end with the recommendation model. On the other hand, the textual features for each description are first projected into the 512-dimensional latent space through linear mapping (i.e., adopting a projection matrix, which is also trained end-to-end). Then, the authors adopt bag-of-words to obtain a unique representation for the description of each fashion item.
(4) **Artist and song recommendation [86].** Before further processing the extracted multimodal features, both textual and audio features are normalized. Afterward, they optionally go through two separate MLPs (i.e., *Coordinate* representation).

3.4 Multimodal feature fusion

As an optional third step, when *Coordinate* representation is used, our multimodal schema allows an additional *Fusion* step to combine all produced multimodal representations. In the following, we describe the inference step in the two cases of *Early* and *Late* fusion.

Early fusion. Let $\tilde{\mathbf{c}}_x$ be the vector of coordinate multimodal representations for input x in all modalities. Then, let $\gamma_e(\cdot)$ be the function for *Early* fusion. We generate the multimodal representation for input x as:

$$\tilde{c}_x = \gamma_e(\tilde{\mathbf{c}}_x). \tag{6}$$

Note that after applying Equation (6), everything we describe in the following also applies to *Joint* representation. We obtain the predicted output $\hat{y}$ for input x as:

$$\hat{y} = \rho(\tilde{c}_x). \tag{7}$$

Late fusion. Let $\tilde{c}_x^{(m)}$ be the coordinate multimodal representation for input x in modality $m \in \mathcal{M}_x$. We first predict the different output values for each modality as:

$$\hat{y}^{(m)} = \rho(\tilde{c}_x^{(m)}) \quad \forall m \in \mathcal{M}_x. \tag{8}$$

Let $\hat{\mathbf{y}}$ be the vector of multimodal predicted outputs in all modalities. If we denote $\gamma_l(\cdot)$ as the function for *Late* fusion, we finally aggregate (fuse) all modality-aware predictions:

$$\hat{y} = \gamma_l(\hat{\mathbf{y}}). \tag{9}$$

Whatever the type of *Fusion*, the literature shows that various works perform this operation differently, from more straightforward solutions such as concatenation and element-wise addition, multiplication, or average to more refined techniques (i.e., neural-based ones, like attention mechanisms). Note that, in this work, we consider *Late* fusion also when multimodal representations are exploited for some specific components of the loss function; indeed, in such settings, multimodal fusion does not occur even until the very last stage of the recommendation pipeline (i.e., the calculation and optimization of the loss function).

Applications

(1) **Micro-video recommendation [118].** The adoption of a multimodal coordinate representation requires a modality fusion phase. This is performed through element-wise addition among modalities for users and items. As this occurs before feeding them into the inference function, we categorize it as *Early* fusion.
(2) **Food recommendation [112].** No modality fusion is run over user profiles. Regarding items, authors adopt an *Early* modality fusion by concatenating the visual and textual features.
(3) **Outfit fashion compatibility [38].** As the method adopts *Joint* multimodal representation (see above) no fusion takes place.
(4) **Artist and song recommendation [86].** The authors explore two possibilities: if no MLP processing occurred in the multimodal representation stage, then normalized features are concatenated and fed into a one-layer MLP; otherwise, multimodal representations are connected to the one-layer MLP. They adopt an *Early* multimodal fusion in both cases.

3.5 Multimodal recommendation task

Let $\mathcal{W}_\varphi$, $\mathcal{W}_\mu$, and $\mathcal{W}_\gamma$ be the sets of the additional model trainable weights from (i) feature extraction, (ii) multimodal representation, and (iii) multimodal fusion, respectively. Note that they could be empty, as the correspondent functions may be non-trainable. Then, given the set of multimodal input data $\mathcal{X}$, we extend Equation (1):

$$\hat{\Theta} = \arg\max_{\Theta} prob(\Theta|\mathcal{X}), \tag{10}$$

where $\Theta = [\Theta_\rho, \Theta_\varphi, \Theta_\mu, \Theta_\gamma]$, with

$$\Theta_\varphi = [\theta_\varphi^{(0)}, \theta_\varphi^{(1)}, \ldots, \theta_\varphi^{(|\mathcal{W}_\varphi|-1)}], \quad \Theta_\mu = [\theta_\mu^{(0)}, \theta_\mu^{(1)}, \ldots, \theta_\mu^{(|\mathcal{W}_\mu|-1)}], \quad \Theta_\gamma = [\theta_\gamma^{(0)}, \theta_\gamma^{(1)}, \ldots, \theta_\gamma^{(|\mathcal{W}_\gamma|-1)}], \tag{11}$$

as the vectors of the model's feature extractor weights, multimodal representation weights, and multimodal fusion weights, respectively. For instance, let us consider the case of simple (but popular) latent factor models leveraging multimodal features (e.g., visual Bayesian personalized ranking [41] in the multimodal version proposed in [134]). Specifically, we have a set of trainable weights for: (i) the inference function comprising the users' and items' embeddings $\mathcal{W}_\rho = \{\theta_\rho^{\mathcal{U}}, \theta_\rho^{\mathcal{I}}\}$ (as previously defined, see again Section 3.1); (ii) the multimodal *Coordinate* representation function comprising the projection matrices, each translating the items' multimodal input features into the same latent space as users' and items' embeddings $\mathcal{W}_\mu = \{\theta_\mu^{\mathcal{M}}\}$, with $\theta_\mu^{\mathcal{M}}$ representing such projection matrices for all modalities. Thus, in the case of multimodal recommender systems leveraging latent factor models, the inference function may be indicated:

$$\hat{y} = \rho(\theta_\rho^{\mathcal{U}}, \theta_\rho^{\mathcal{I}}, \theta_\mu^{\mathcal{M}}, \mathcal{X}). \tag{12}$$

Generally, we solve Equation (10) by optimizing the loss L:

$$L = L_{rec}(\Theta, \hat{y}, y) + \alpha L_{reg}(\Theta), \tag{13}$$

where y is the ground-truth value corresponding to the predicted output $\hat{y}$, and α is a model hyper-parameter to weight the *regularization* component of the loss function (i.e., L_{reg}).

APPLICATIONS

(1) **Micro-video recommendation [118].** As in several collaborative filtering approaches, the inference is performed through the inner product between the multimodal representations of users and items. Regarding the loss function, the authors use the broadly-adopted BPR [90] optimization framework, maximizing the distance between predicted ratings for positive items (i.e., the ones interacted by users) and negative items (i.e., the ones not already interacted by users).
(2) **Food recommendation [112].** User embeddings are learned for tag prediction, with a one-layer MLP used to predict scores and cross-entropy as a loss function. The final user embeddings are eventually exploited for the main task of food recommendation. Contrarily, item embeddings are directly adopted for the score prediction, run with a one-layer MLP trained on a binary cross-entropy loss.
(3) **Outfit fashion compatibility [38].** Only the visual modality is adopted as input for the recommendation inference. However, the textual modality is exploited jointly with the visual input to minimize the contrastive component of the loss function, for whom the cosine similarity measures the distance between visual and textual modalities in the shared latent space.
(4) **Artist and song recommendation [86].** Inference is run through the inner product between user and item final embeddings, while cosine distance is the chosen loss function as the learned latent embeddings are l2-normalized.

To conclude, Algorithm 1 provides a general overview of the overall multimodal schema we presented, while Table 2 summarizes the main contributions of the four selected recommender systems to conceptually validate our theoretical framework.

4 IMPLEMENTATION AND BENCHMARKING (RQ3)

In this section, we show how we integrate the proposed multimodal schema for multimedia recommendation into Elliot [2], a framework for the rigorous reproducibility and evaluation of recommender systems. Specifically, we use this implementation to benchmark six state-of-the-art multimedia recommendation approaches (i.e., VBPR [41], MMGCN [118], GRCN [117], LATTICE [134], BM3 [145], and FREEDOM [144]). To complement our benchmarking analysis, we also select four classical recommender systems which do not leverage multimodal features, namely: BPR [90], NGCF [113], LightGCN [42], and SGL [120]; this is useful to assess the performance improvement obtained through the adoption of multimodal features. Additionally, Elliot provides evaluation metrics accounting for recommendation accuracy and beyond-accuracy measures, we maintain the importance of assessing the performance of such models especially on the latter category of metrics. Indeed, the related literature about multimedia recommendation has mainly focused on the performance evaluation through standard accuracy metrics, often disregarding the potentially huge relevance that other evaluation dimensions could have by measuring the novelty and diversity of the produced recommendations [75], or the portion of popular items recommended to users [76].

Algorithm 1: Multimodal schema for multimedia recommendation

Input: Set of available modalities $\mathcal{M}$; set of multimodal input data $\mathcal{X}$ and admissible modalities $\mathcal{M}_x$, $\forall x \in \mathcal{X}$.
Output: Trained model's weights $\hat{\Theta}$.
Initialize all model's trainable weights $\mathbf{\Theta}$.
repeat
  extract features according to Equation (3)
  if *Joint representation* **then**
    get joint representation according to Equation (4)
    get model's prediction according to Equation (7)
  else if *Coordinate representation* **then**
    get coordinate multimodal representations according to Equation (5)
    if *Early fusion* **then**
      get multimodal representation according to Equation (6)
      get model's prediction according to Equation (7)
    else if *Late fusion* **then**
      get predictions for each modality according to Equation (8)
      get model's prediction according to Equation (9)
  for $\hat{\theta} \in \hat{\Theta}$ **do**
    update $\hat{\theta}$ according to Equation (10), by optimizing the loss function L in Equation (13)
  end
until *convergence*;
Return $\hat{\Theta}$.

Table 2. Four literature examples of multimodal frameworks for multimedia recommendation. For each work, we report the performed task, the considered modalities for each input to the system (e.g., user and item), the feature extraction and multimodal representation strategies, the multimodal fusion, and the adopted inference/loss functions.

	Paper	Input	Modalities	Multimodal Input Data	Feature Extraction	Multimodal Representation	Multimodal Fusion	Inference and Loss
Micro-video recommendation	Wei et al. [118]	User	Visual			Coordinate, aggregation of node's neighborhood and combination with the ego node. Projection in both aggregation and combination.	Early, modalities are combined through addition.	Inference via inner-product between final user and item embeddings. BPR is the optimization function.
			Textual					
			Audio					
		Item	Visual	Video frames	ResNet50			
			Textual	Captions and descriptions	Sentence2Vector			
			Audio	Audio track	VGGish			
Food recommendation	Wang et al. [112]	User	Textual	Users' tweets	Bi-RNN & CNN			Score prediction with MLP for user tags. Loss is cross-entropy. User embedding is later adopted for recommendation.
		Item	Visual	Video Frames	VGGNet-19	Coordinate, no particular operation performed.	Early, modalities are combined via concatenation.	User-item score prediction with MLP and cross-entropy.
			Textual	Recipe ingredients	TextCNN			
Outfit fashion compatibility	Han et al. [38]	Item	Visual	Product images	InceptionV3	Joint, features are projected into a shared embedding space.		The visual modality is used for the inference.
			Textual	Descriptions	One-hot-encode			The textual modality is used for the contrastive component of the loss function.
Artist and song recommendation	Oramas et al. [86]	Item	Textual	Artist biography	Custom CNN	Coordinate, textual and audio features normalized and optionally fed into two separate MLPs.	Early, either normalized features are concatenated and fed into a one-layer MLP, or multimodal representations are connected to the one-layer MLP.	Inference via inner-product between user and final item embeddings. Cosine distance is the loss function.
			Audio	Audio spectrogram	Custom CNN			

In the following, we report on the datasets we used, the technical aspects of the (multimedia) recommender systems involved, the evaluation metrics we adopted (spanning both accuracy and beyond-accuracy recommendation metrics), the reproducibility details of our framework, and the results of the benchmarking analysis.

4.1 Datasets

For the benchmarking, we use five popular [20, 52, 134, 145] datasets which collect the purchase history from five product categories of the Amazon catalog [40, 80], namely, Office Products (i.e., Office), Toys & Games (i.e., Toys), All Beauty (i.e., Beauty), Sports & Outdoors (i.e., Sports), and Clothing Shoes & Jewelry (i.e., Clothing). Besides containing the records of user-product interactions with timestamps and other metadata, such datasets come with the visual features extracted from the product images, stored as 4,096-dimensional embeddings which are publicly available at the same URL of the datasets[2]. Conversely, in terms of textual modality, we adopt the same procedure indicated in [134], and concatenate the item's title, descriptions, categories, and brand, to extract the 1,024-dimensional textual embeddings through sentence transformers [89]. Overall dataset information can be found in Table 3.

4.2 Recommender systems baselines

We decide to benchmark the results of six state-of-the-art multimedia recommender systems, namely, VBPR [41], MMGCN [118], GRCN [117], LATTICE [134], BM3 [145], and FREEDOM [144]. Such approaches represent a group of techniques that are widely recognized in the related literature as strong baselines in multimedia recommendation exploiting multimodality, as well as recently proposed solutions at top-tier conferences (Table 4). Moreover, we also present the four selected recommendation solutions which do not leverage multimodal features: BPR [90], NGCF [113], LightGCN [42], and SGL [120]. In the following, we describe the main technical details for each of them.

Classical recommendation

4.2.1 BPR. Bayesian-personalized ranking [90], also known as BPR, is among the pioneer optimization algorithms for recommendation. Specifically, it works by maximizing the distance of predicted ratings for positive and negative items for the same user. Coupled with the matrix factorization [55] (MF) strategy, which decouples the user-item interaction matrix into latent factors (i.e., embeddings) representing users and items, it has been exploited as building block for a large plethora of recommender systems from the last decades.

4.2.2 NGCF. Neural graph collaborative filtering [113], indicated as NGCF, is among the first techniques in recommendation leveraging the representational power of graph neural networks (GNNs). Its message-passing formulation works by aggregating the information coming from the neighborhood nodes into the ego nodes at different distance hops, and it also leverages the inter-dependencies among the ego and the neighborhood nodes.

4.2.3 LightGCN. Light graph convolutional network [42], always referred to as LightGCN, propose a light-weight formulation of the graph convolutional layer proposed by Kipf and Welling [54], suggesting (and empirically demonstrating) that such a variation could lead to superior performance in personalized recommendation. To be specific, the authors design the model such that by dropping feature transformations and non-linearities from the message-passing.

4.2.4 SGL. Self-supervised graph learning [120], presented as SGL, proposes to apply self-supervised [47] and contrastive [51] learning to GNN-based recommendation. The model is based upon a LightGCN backbone, but is designed

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

Table 3. Statistics of the tested datasets.

Datasets	$\|\mathcal{U}\|$	$\|\mathcal{I}\|$	$\|\mathcal{R}\|$	**Sparsity (%)**
Office	4,905	2,420	53,258	99.55%
Toys	19,412	11,924	167,597	99.93%
Beauty	22,363	12,101	198,502	99.93%
Sports	35,598	18,357	296,337	99.95%
Clothing	39,387	23,033	278,677	99.97%

to learn different views of nodes by performing node/edge dropout and random walk operations on the underlying user-item graph structure. At this stage, a self-supervised contrastive loss component is added to promote the consistency and divergence among same and different views of the same and different nodes, respectively.

Multimedia recommendation

4.2.5 VBPR. Visual-bayesian personalized ranking [41], abbreviated as VBPR, represents one of the pioneering efforts to integrate visual-aware content such as high-level visual features derived from product images, into the BPR-MF recommendation algorithm. In addition to the collaborative embeddings for users and items, the authors introduce a pair of visual embeddings for users and items. The latter captures high-level features extracted from product images using a pre-trained convolutional neural network. The prediction score is determined by summing the inner products of the collaborative and visual embeddings. While VBPR was initially conceived as a recommendation system focused on a single modality, we have followed the approach outlined in [134] to introduce additional user and item embeddings for each additional modality, mirroring the methodology applied to the visual modality.

4.2.6 MMGCN. Multimodal graph convolution network [118], referred to as MMGCN, introduces the concept of multimodality in graph-based recommendation systems. In this approach, the authors suggest training distinct graph convolutional networks for each modality under consideration. This results in three sets of user and item representations, accommodating the various perspectives that users may have toward each modality. Ultimately, all the embeddings related to different modalities, both for users and items, are merged through element-wise addition, and the preference prediction score is computed using an inner product.

4.2.7 GRCN. Graph-refined convolutional network [117], denoted as GRCN leverages the information from various modalities to refine the values of the adjacency matrix. This is particularly important due to the implicit nature of user-item interactions in the bipartite graph. The primary objective is to identify and potentially remove edges that do not accurately reflect each user's actual preferences. The multimodal representations of users and items that are learned are then fused through concatenation to produce a comprehensive representation for predicting preference scores.

4.2.8 LATTICE. Latent structure mining method for multimodal recommendation [134], dubbed LATTICE, creates a similarity graph between items for each modality and enhances this structure using graph structure learning techniques. These improved adjacency matrices are then merged through a weighted sum, assigning varying importance weights to each modality. The resulting adjacency matrix is employed to refine item embeddings using a graph convolutional network. Ultimately, the obtained item embeddings can serve as the building blocks for various user and item latent factor-based models, such as BPR-MF.

Table 4. An overview on the selected multimedia recommender systems, along with: (i) their year and publication venue; (ii) whether they represent users and/or items through multimodal embeddings; (iii) whether they use GNNs and (in case) on which type of graph (U-U and I-I stand for user-user and item-item, respectively); (iv) a non-exhaustive set of papers where they are used as baselines.

Models	**Year**	**Venue**	**Multimodal embeddings**		**GNN**	**Multimodal graphs**		**Baseline in**
			Users	**Items**		**U-I**	**I-I**	
VBPR [41]	2016	AAAI	✓	✓	✗			[20, 22, 64, 66, 104, 134]
MMGCN [118]	2019	MM	✓	✓	✓	✓	✗	[97, 111, 115, 116, 132, 145]
GRCN [117]	2020	MM	✗	✓	✓	✓	✗	[57, 65, 66, 115, 134, 145]
LATTICE [134]	2021	MM	✗	✓	✓	✗	✓	[52, 83, 115, 135, 145]
BM3 [145]	2023	WWW	✗	✓	✓	✓	✗	[63, 129, 143]
FREEDOM [144]	2023	MM	✗	✓	✓	✗	✓	[143]

4.2.9 BM3. Bootstrapped multimodal model [145], indicated as BM3, proposes a self-supervised multimodal technique for recommendation. Different from previous approaches using computationally expensive augmentations, BM3 leverages dropout as a simple operation for generating contrastive views of the same embeddings. In detail, the loss function consists of three components, where a reconstruction loss minimizes the similarity between the contrastive views of user and item embeddings, while an inter- and intra-modality alignment loss works to minimize the distance between the contrastive views generated for the same or different modalities.

4.2.10 FREEDOM. The authors from [144] demonstrate that freezing the item-item multimodal similarity graph (derived from LATTICE) and denoising the user-item graph can lead to improved recommendation performance (the proposed model is named FREEDOM). As for the denoising operation of the user-item graph, the authors propose a degree-sensitive edge pruning to remove noisy edges from the user-item adjacency matrix. Moreover, and differently from LATTICE, the model optimizes a double BPR-like loss function, where the first component of the loss integrates a multimodal-enhanced representation of the item embedding, while the second component explicitly leverages the item projected multimodal features.

4.3 Evaluation metrics

To conduct the benchmarking analysis, we measure the recommendation performance through accuracy and beyond-accuracy metrics. While the former are commonly adopted measures in the related literature on multimedia recommendation, the latter have been brought to the attention of the community only in recent works, especially to assess the novelty and diversity of the recommended items [75], or how multimedia recommender systems are biased towards suggesting popular items [76]. For the recommendation accuracy, we consider the Recall@k and the nDCG@k; for the novelty [108] and diversity [98], we measure the EFD@k and the Gini@k, respectively; for the popularity bias [1], we calculate the APTL@k; finally, as a general index of how recommendations cover the entire catalog of products, we adopt the iCov@k. In the following, we present each of these metrics along with their mathematical formulations.

4.3.1 Recall. The recall (Recall) evaluates to what extent the recommender systems can retrieve relevant items from the recommendation list:

$$\text{Recall@}k = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{|\text{Rel}_u \text{ @}k|}{|\text{Rel}_u|}, \tag{14}$$

with Rel_u as the set of relevant items for user u, and Rel_u @k as the set of relevant recommended items in the top@k.

4.3.2 Normalized discount cumulative gain. The normalized discount cumulative gain (nDCG) measures both the relevance and the ranking position of items from the recommendation lists, considering the various relevance degrees:

$$\text{nDCG@}k = \frac{1}{|\mathcal{U}|} \sum_{u} \frac{\text{DCG}_u\text{@}k}{\text{IDCG}_u\text{@}k}, \tag{15}$$

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

4.3.3 Expected free discovery. The expected free discovery (EFD), as introduced in [108], is a metric accounting for novelty in recommendation that utilizes the *inverse collection frequency*. Specifically, it provides a quantification on how a recommender system can suggest relevant *long-tail* items (i.e., niche products):

$$\text{EFD@}k = C \sum_{i_k \in R} \text{disc}(k) P\left(\text{Rel}_u\text{ @}k \mid i_k, u\right) \cdot \left(-\log_2 p(i \mid \text{ seen}, \theta)\right). \tag{16}$$

4.3.4 Gini index. The Gini index (Gini) indicates the disparity in items' popularity when considering the recommendation lists for each user. In detail, it provides a measurement of how certain items are consistently favored by a large portion of users with respect to other items. For the sake of this analysis, we use a normalized version of Gini according to which high values stand for a wide set of items being suggested to users, and low values indicate that a few popular items are generally recommended, leading to a less diverse recommendation experience [98]. Its (normalized) formulation is:

$$\text{Gini@}k = 1 - \left(\frac{\sum_{i=1}^{|\mathcal{I}|} (2i - |\mathcal{I}| - 1) P|_{@k}(i)}{|\mathcal{I}| \sum_{i=1}^{|\mathcal{I}|} P|_{@k}(i)} \right), \tag{17}$$

where $P|_{@k}(i)$ represents the popularity of item i, in the top@k recommendation lists, sorted in non-decreasing order (i.e., $P|_{@k}(i) \leq P|_{@k}(i+1)$).

4.3.5 Average percentage of long-tail items. The average percentage of long-tail items (APLT) assesses the presence of popularity bias in recommendation [1]. When referring to popularity bias, we indicate the tendency of recommender systems to boost the recommendation of popular/mainstream items (i.e., *short-head*) at the detriment of unpopular/niche products (i.e., *long-tail*), thus limiting the exposure of certain item categories to users. The APLT calculates the percentage of *long-tail* items belonging to the recommendation lists (averaged over all users):

$$\text{APLT@}k = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{|\{i \mid i \in (\hat{\mathcal{I}}_u\text{@}k \cap \sim \Phi)\}|}{k}, \tag{18}$$

where $\hat{\mathcal{I}}_u\text{@}k$ is the list of top@k recommended items for user u, and Φ is the set of items belonging to the *short-tail* distribution while $\sim \Phi$ stands for the remaining *long-tail* items.

4.3.6 Item coverage. The item coverage (iCov) is a percentage estimating how the recommended items span the entire catalog of products in the recommendation system:

$$\text{iCov@}k = \frac{|\bigcup_u \hat{\mathcal{I}}_u\text{@}k|}{|\mathcal{I}_{train}|}\%, \tag{19}$$

where $\mathcal{I}_{train}$ is the set of items in the training set.

4.4 Reproducibility

First, we pre-process the datasets following the 5-core filtering on users and items to remove cold-start users and items as done in [134]. Second, we split them according to the 80:20 hold-out strategy for the training and test sets, where the former and the latter contain the 80% and the 20% of interactions recorded for each user, respectively. Then, we decide to train the recommendation models so that the number of epochs (i.e., 200) and the batch size (i.e., 1024) are the same for all of them to ensure fair comparison. As for the other models' hyper-parameters, we empirically find that, even if we consider (partially) different datasets with respect to the ones from the original works, exploring only the learning rate and regularization coefficient while fixing the remaining (model-specific) hyper-parameters to the indicated best values is enough to generally obtain performance trends comparable with the original works. For this reason, we explore only the learning rate (i.e., [0.0001, 0.0005, 0.001, 0.005, 0.01]) and the regularization coefficient (i.e., [10^{-5}, 10^{-2}]) for a total of 10 explorations. Note that the search spaces for the learning rate and the regularization coefficient overlap across all baselines because this (i) is what usually happens in the original works, and (ii) ensures (once again) fair comparison. Finally, to select the best configuration for each model and dataset, we remove the 50% of the test set for the validation set (following again [134]), and select the hyper-parameter setting providing the highest Recall@20 value on the validation data measured for a specific epoch (maximum 200 epochs). To foster the reproducibility of the proposed benchmarks, we provide the codes, datasets, and configuration files to replicate our results at: https://github.com/sisinflab/Formal-MultiMod-Rec, where we integrated the selected multimedia recommender systems into Elliot [2]. Explored hyper-parameter values are also reported in Table 5.

4.5 Benchmarking results

We organize the proposed benchmarking analysis as follows. First, we present the results regarding the overall performance of both classical and multimedia recommender systems in their best configurations. Then, we provide a finer-grained investigation on the impact of multimodal features under different settings.

4.5.1 Overall performance. Table 6 reports on the results of the extensive benchmarking analysis we conduct on the selected datasets and state-of-the-art classical and multimedia recommendation systems (in the Table categorized under the 'Classical' and 'Multimedia' fields, respectively). The calculated metrics involve both accuracy (i.e., Recall and nDCG) and beyond-accuracy (i.e., EFD, Gini, APLT, and iCov) measures when considering top@10 and top@20 recommendation lists. Based on how we defined all the recommendation metrics, higher values indicate better performance.

In terms of **accuracy** performance, we observe that one of LATTICE, BM3, and FREEDOM is steadily among the two best recommendation models, outperforming the best classical recommendation approaches by a great margin; indeed, this is something that also emerges from the related literature. This holds across all datasets and top@k under analysis, with the only exception of Office. Nevertheless, we notice an interesting trend when comparing VBPR to the other multimedia recommendation techniques. Particularly, we see how this model is almost always among the top-3 recommendation techniques despite being one of the shallowest approaches compared to other more recent and complex models. As already stated in recent works [75, 76], this finding demonstrates how even a not-so-deep, but still careful hyper-parameter exploration (such as the one we performed) may help uncover unexpected results with respect to what described in the literature.

However, the most surprising behaviors involves the analysis of the **beyond-accuracy** performance. First, we notice that, differently from what observed in the accuracy analysis, the classical recommendation approaches almost always settle as second-to-best solutions under beyond-accuracy metrics, surpassing most of the multimedia counterparts.

Table 5. Hyper-parameter values as explored in our benchmarking. Note that values reported within "[. . .]" are explored in different settings, while those with "(. . .)" refer to different values for the same configuration (e.g., multimodal factors for each modality).

Families	Models	Hyper-parameters
Classical	BPR	`learning_rate`: [0.0001, 0.0005, 0.001, 0.005, 0.01], `factors`: 64, `regularization`: [1e-2, 1e-5]
	NGCF	`learning_rate`: [0.0001, 0.0005, 0.001, 0.005, 0.01], `factors`: 64, `regularization`: [1e-2, 1e-5], `n_layers`: 3, `weight_size`: 64, `node_dropout`: 0.1, `message_dropout`: 0.1, `adj_normalization`: True
	LightGCN	`learning_rate`: [0.0001, 0.0005, 0.001, 0.005, 0.01], `factors`: 64, `regularization`: [1e-2, 1e-5], `n_layers`: 3, `adj_normalization`: True
	SGL	`learning_rate`: [0.0001, 0.0005, 0.001, 0.005, 0.01], `factors`: 64, `regularization`: [1e-2, 1e-5], `n_layers`: 3, `node_dropout`: 0.1, `ssl_temp`: 0.2, `ssl_reg`: 0.1, `ssl_ratio`: 0.1, `sampling`: edge_dropout
Multimedia	VBPR	`learning_rate`: [0.0001, 0.0005, 0.001, 0.005, 0.01], `factors`: 64, `regularization`: [1e-2, 1e-5], `comb_mod`: concat
	MMGCN	`learning_rate`: [0.00001, 0.00003, 0.0001, 0.001, 0.01], `factors`: 64, `regularization`: [1e-2, 1e-5], `n_layers`: 3, `factors_mm`: (256, None), `aggregation`: mean, `concatenation`: False, `has_id`: True
	GRCN	`learning_rate`: [0.0001, 0.001, 0.01, 0.1, 1], `factors`: 64, `regularization`: [1e-2, 1e-5], `n_layers`: 2, `n_routings`: 3, `factors_mm`: 128, `aggregation`: add, `weight_mode`: confid, `pruning`: True, `has_act`: False, `fusion_mode`: concat
	LATTICE	`learning_rate`: [0.0001, 0.0005, 0.001, 0.005, 0.01], `factors`: 64, `regularization`: [1e-2, 1e-5], `n_layers`: 1, `n_ui_layers`: 2 `top_k`: 20, `l_m`: 0.7, `factors_mm`: 64
	BM3	`learning_rate`: [0.0001, 0.0005, 0.001, 0.005, 0.01], `factors`: 64, `regularization`: [1e-1, 1e-2], `n_layers`: 2, `cl_weight`: 2.0, `dropout`: 0.3, `lr_sched`: (1.0,50), `factors_mm`: 64
	FREEDOM	`learning_rate`: [0.0001, 0.0005, 0.001, 0.005, 0.01], `factors`: 64, `regularization`: [1e-2, 1e-5], `n_layers`: 1, `n_ui_layers`: 2 `top_k`: 10, `factors_mm`: 64, `mw`: (0.1,0.9), `dropout`: 0.8, `lr_sched`: (1.0,50)

Additionally, when considering the multimedia recommendation scenario only, models dominating the accuracy performance such as LATTICE, BM3, and FREEDOM cannot provide similar results on the other metrics accounting for novelty, diversity, and popularity bias. Indeed, the only observable trend in this setting is that GRCN steadily settle as best-performing algorithms. Particularly, it is worth pointing out how this approach can strike a sufficient trade-off between accuracy and beyond-accuracy measures. Once again, such observations corroborate what has recently been pointed out in similar works [75, 76], by extending the analysis to additional datasets and recommendation systems.

4.5.2 Impact of different multimodal feature settings. To assess the impact of injecting multimodal features into the recommendation pipeline, we select BPR [90] as recommender system, and calculate the performance variation in three different settings: (i) **visual** modality only (i.e., VBPR [41] in its original formulation), (ii) **textual** modality only, and (iii) **visual + textual** modalities (the one reported in the overall performance from above). Overall (Table 7), results show an interesting trend, which complements the observation from the previous analysis. While on the accuracy metric (i.e., Recall@20) the integration of (multi)modal features generally leads to improved recommendation performance, the same is not true on the other two beyond-accuracy measures (i.e., APLT@20 and iCov@20). Indeed, in the latter setting, it is evident how even the base configuration (i.e., without multimodal features) can reach best or second-to-best performance. The only exception to the outlined observations is on the Clothing dataset, where (as normally acknowledged in the literature) the textual and visual+textual (i.e., multimodal) settings are the best ones.

Table 6. Benchmarking results on selected datasets and multimedia recommenders for accuracy and beyond-accuracy recommendation metrics, on top@10 and top@20 lists. For each metric-dataset pair, **boldface** and underlined indicate best and second-to-best values.

Datasets		Models	top@10						top@20					
			Accuracy		Beyond-accuracy				Accuracy		Beyond-accuracy			
			Recall	**nDCG**	**EFD**	**Gini**	**APLT**	**iCov**	**Recall**	**nDCG**	**EFD**	**Gini**	**APLT**	**iCov**
Office	Classical	BPR	0.0604	0.0390	0.1655	0.3675	0.2292	93.95%	0.0958	0.0501	0.1387	0.4047	0.2378	97.64%
		NGCF	0.0582	0.0365	0.1604	0.4428	0.2726	97.47%	0.0958	0.0478	0.1365	0.4757	0.2817	99.59%
		LightGCN	**0.0802**	**0.0520**	**0.2099**	0.1684	0.1594	76.14%	**0.1237**	**0.0655**	**0.1740**	0.2144	0.1817	86.70%
		SGL	0.0695	0.0469	0.1991	0.3479	0.2833	84.92%	0.1012	0.0566	0.1592	0.3992	0.2928	91.55%
	Multimedia	VBPR	0.0652	0.0419	0.1753	0.3634	0.2321	93.83%	0.1025	0.0533	0.1479	0.3960	0.2375	97.51%
		MMGCN	0.0455	0.0300	0.1140	0.0128	0.0016	3.07%	0.0798	0.0405	0.1027	0.0231	0.0078	4.64%
		GRCN	0.0393	0.0253	0.1215	**0.4587**	**0.3438**	**99.01%**	0.0667	0.0339	0.1051	**0.4892**	**0.3469**	**99.79%**
		LATTICE	0.0664	0.0449	0.1827	0.2128	0.1752	87.86%	0.1029	0.0566	0.1513	0.2652	0.2039	95.90%
		BM3	0.0701	0.0460	0.1837	0.1407	0.1427	77.13%	0.1081	0.0583	0.1550	0.1900	0.1715	91.55%
		FREEDOM	0.0560	0.0365	0.1493	0.1922	0.1875	79.12%	0.0884	0.0469	0.1282	0.2439	0.2080	90.64%
Toys	Classical	BPR	0.0643	0.0404	0.1712	0.2673	0.1166	84.32%	0.0906	0.0482	0.1345	0.3049	0.1268	92.12%
		NGCF	0.0622	0.0403	0.1715	0.3535	0.1609	92.36%	0.0902	0.0485	0.1365	0.3960	0.1760	97.24%
		LightGCN	0.0714	0.0455	0.1810	0.0740	0.0288	52.69%	0.1036	0.0550	0.1446	0.1037	0.0374	68.41%
		SGL	0.0789	0.0519	0.2203	0.2352	0.0977	78.27%	0.1061	0.0601	0.1681	0.2838	0.1106	88.77%
	Multimedia	VBPR	0.0710	0.0458	0.1948	0.2645	0.1064	84.90%	0.1006	0.0545	0.1527	0.3011	0.1180	92.82%
		MMGCN	0.0256	0.0150	0.0648	0.0989	0.0961	37.87%	0.0426	0.0200	0.0570	0.1450	0.1058	52.51%
		GRCN	0.0554	0.0354	0.1604	**0.3954**	**0.2368**	**92.66%**	0.0831	0.0436	0.1298	**0.4329**	**0.2482**	**97.73%**
		LATTICE	0.0805	0.0512	0.2090	0.1656	0.0546	73.80%	0.1165	0.0617	0.1665	0.2026	0.0684	86.58%
		BM3	0.0613	0.0393	0.1582	0.0776	0.0486	56.23%	0.0901	0.0478	0.1270	0.1154	0.0658	73.50%
		FREEDOM	**0.0870**	**0.0548**	**0.2284**	0.1474	0.0756	62.09%	**0.1249**	**0.0660**	**0.1820**	0.2007	0.0951	78.42%
Beauty	Classical	BPR	0.0676	0.0414	0.1869	0.2348	0.1099	83.78%	0.0993	0.0511	0.1519	0.2689	0.1195	91.51%
		NGCF	0.0661	0.0408	0.1787	0.2375	0.0929	86.47%	0.0987	0.0505	0.1473	0.2754	0.1071	94.08%
		LightGCN	0.0785	0.0493	0.2044	0.0555	0.0290	49.92%	0.1141	0.0599	0.1647	0.0794	0.0343	65.74%
		SGL	0.0810	0.0524	**0.2291**	0.1554	0.0807	66.78%	0.1144	0.0626	0.1824	0.1935	0.0870	78.36%
	Multimedia	VBPR	0.0760	0.0483	0.2119	0.2076	0.0833	83.06%	0.1102	0.0586	0.1700	0.2376	0.0915	91.41%
		MMGCN	0.0496	0.0294	0.1300	0.0252	0.0282	13.75%	0.0772	0.0379	0.1105	0.0423	0.0345	21.37%
		GRCN	0.0575	0.0370	0.1817	**0.3823**	**0.2497**	**94.59%**	0.0892	0.0466	0.1498	**0.4178**	**0.2608**	**98.56%**
		LATTICE	**0.0867**	**0.0544**	0.2272	0.1153	0.0386	65.82%	0.1259	0.0661	0.1830	0.1558	0.0511	81.60%
		BM3	0.0713	0.0443	0.1831	0.0245	0.0179	32.31%	0.1051	0.0545	0.1490	0.0414	0.0228	48.75%
		FREEDOM	0.0864	0.0539	0.2279	0.0921	0.0486	55.89%	**0.1286**	**0.0666**	**0.1868**	0.1359	0.0653	72.96%
Sports	Classical	BPR	0.0411	0.0251	0.1047	0.1713	0.0709	76.60%	0.0627	0.0314	0.0861	0.1963	0.0774	86.62%
		NGCF	0.0419	0.0260	0.1073	0.1821	0.0616	82.50%	0.0633	0.0323	0.0876	0.2093	0.0708	92.28%
		LightGCN	0.0558	0.0346	0.1324	0.0215	0.0086	33.59%	0.0839	0.0428	0.1081	0.0342	0.0113	49.49%
		SGL	0.0545	0.0347	0.1413	0.0680	0.0305	45.99%	0.0788	0.0419	0.1122	0.0888	0.0346	58.65%
	Multimedia	VBPR	0.0450	0.0281	0.1167	0.1501	0.0497	75.77%	0.0677	0.0349	0.0949	0.1722	0.0552	86.54%
		MMGCN	0.0342	0.0207	0.0791	0.0095	0.0046	5.10%	0.0551	0.0269	0.0678	0.0168	0.0065	8.39%
		GRCN	0.0330	0.0202	0.0885	**0.3087**	**0.2190**	**91.28%**	0.0523	0.0259	0.0746	**0.3386**	**0.2273**	**97.09%**
		LATTICE	**0.0610**	0.0372	0.1465	0.0573	0.0129	48.44%	0.0898	0.0456	0.1185	0.0802	0.0185	64.90%
		BM3	0.0548	0.0349	0.1372	0.0776	0.0283	59.13%	0.0825	0.0430	0.1118	0.1120	0.0385	76.75%
		FREEDOM	0.0603	**0.0375**	**0.1494**	0.0621	0.0319	48.37%	**0.0911**	**0.0465**	**0.1219**	0.0926	0.0442	65.81%
Clothing	Classical	BPR	0.0303	0.0156	0.0427	0.2260	0.0729	80.00%	0.0459	0.0195	0.0347	0.2600	0.0824	89.42%
		NGCF	0.0284	0.0146	0.0395	0.2438	0.0633	87.55%	0.0443	0.0186	0.0327	0.2788	0.0743	95.65%
		LightGCN	0.0393	0.0210	0.0534	0.0379	0.0086	41.61%	0.0602	0.0262	0.0438	0.0569	0.0111	58.59%
		SGL	0.0408	0.0220	0.0583	0.1296	0.0352	66.77%	0.0580	0.0263	0.0457	0.1699	0.0435	81.32%
	Multimedia	VBPR	0.0339	0.0181	0.0502	0.2437	0.0809	83.40%	0.0529	0.0229	0.0413	0.2791	0.0915	92.33%
		MMGCN	0.0227	0.0119	0.0292	0.0136	0.0044	7.58%	0.0348	0.0150	0.0240	0.0236	0.0066	12.44%
		GRCN	0.0319	0.0164	0.0481	**0.3990**	**0.2358**	**93.37%**	0.0496	0.0209	0.0397	**0.4368**	**0.2459**	**97.77%**
		LATTICE	0.0502	0.0275	0.0738	0.1022	0.0134	58.49%	0.0744	0.0336	0.0589	0.1384	0.0207	76.20%
		BM3	0.0418	0.0226	0.0596	0.1348	0.0319	72.88%	0.0633	0.0281	0.0486	0.1825	0.0449	88.65%
		FREEDOM	**0.0547**	**0.0294**	**0.0805**	0.1509	0.0600	65.54%	**0.0822**	**0.0363**	**0.0652**	0.2078	0.0843	81.91%

Table 7. Recommendation results calculated as Recall@20, APLT@20, and iCov@20 on all tested datasets for BPR in four configurations: (i) without multimodal features (i.e., the original BPR [90]), (ii) BPR + visual (i.e., VBPR in its original version [41]), (iii) BPR + textual, and (iv) BPR + (visual + textual). **Boldface** and underlined stand for best and second-to-best values.

Datasets	Models	Recall@20	APLT@20	iCov@20
Office	BPR	0.0958	**0.2378**	**97.64%**
	+ visual	0.1002	0.2361	97.43%
	+ textual	0.0997	0.2256	97.47%
	+ (visual + textual)	**0.1025**	0.2375	97.51%
Toys	BPR	0.0906	**0.1268**	92.12%
	+ visual	0.1002	0.1164	92.33%
	+ textual	**0.1033**	0.1227	92.50%
	+ (visual + textual)	0.1006	0.1180	**92.82%**
Beauty	BPR	0.0993	**0.1195**	91.51%
	+ visual	0.1074	0.1116	92.14%
	+ textual	0.1066	0.1174	**92.48%**
	+ (visual + textual)	**0.1102**	0.0915	91.41%
Sports	BPR	0.0627	0.0774	86.62%
	+ visual	**0.0691**	0.0702	86.14%
	+ textual	0.0678	**0.0875**	**89.11%**
	+ (visual + textual)	0.0677	0.0552	86.54%
Clothing	BPR	0.0459	0.0824	89.42%
	+ visual	0.0526	0.0833	90.99%
	+ textual	0.0528	**0.0947**	**92.71%**
	+ (visual + textual)	**0.0529**	0.0915	92.33%

We deem this different trend to be ascribed to the specific dataset characteristics of the Clothing dataset (as already suggested by other previous works [29, 78]) and not to the multimodal characteristics of such a dataset.

Summing up, the proposed benchmarking studies indicate that training and evaluating multimedia recommender systems remain an open challenge in the literature, showing (in same cases) unexpected outcomes that pave the way to more rigorous and careful analyses to be conducted in future work (see Section 6.4).

5 TECHNICAL CHALLENGES (RQ4)

This section aims to overview the main technical challenges we recognize in multimodal approaches for multimedia recommendation. Starting from the proposed schema we presented in the previous sections, we outline the evident (or even less evident) issues emerging from the literature.

5.1 Missing modalities in the input data

Describing data under the lens of multimodality may be a two-sided coin. From one perspective, multimodality helps enrich the informative content carried by the input, thus exploring data's multi-faceted nature to learn better-tailored user-item preference patterns [5, 77]. On the other side, the need to provide descriptive content for every input modality may come at the expense of some missing modalities (e.g., a video dataset could integrate videos having no textual content, for example, subtitles or descriptions may be only sometimes available). Tackling the modality misalignment in the data is a recent and widely discussed challenge in other domains [56, 71, 91, 131, 133], and requires ad-hoc techniques to provide equal representation of all involved modalities to fully exploit their informative richness [72, 99]. Nevertheless, apart from very recent preliminary attempts [79], to the best of our knowledge, the issue remains open in recommendation.

5.2 Pre-trained feature extractors

Deep learning models processing images, texts, or audio have been shown to enrich the informative content of items' profiles in several recommendation algorithms. In most solutions, such architectures are used as pre-trained blocks to extract high-level features from the input data, thus exploiting the capability of deep neural networks to transfer knowledge among different datasets and/or tasks. Despite the ease of adopting ready-to-use feature extraction networks, we seek to underline a conceptual limitation that, to the best of our knowledge, is only partially investigated in the literature. Indeed, pre-trained representations extracted through state-of-the-art deep learning models are not necessarily supposed to capture those semantic features, which will likely captivate users for their final decision-making process. As an example, the embedding feature extracted from a product image (e.g., a bag) through a pre-trained deep convolutional network (e.g., ResNet50) is carrying high-level informative content driven by the task of *image classification*, but this does not mean the same knowledge will be helpful to predict whether the product could be *recommended* to a user.

5.3 Modalities representation

The multimodal representation of the extracted input data is among the main stages in the multimodal schema we described since it establishes the relationships for the selected input modalities. Nonetheless, the literature is not generally aligned on its definition since most of the works usually refer to *Joint* representation and *Early* fusion interchangeably. We recognize this as a conceptual issue because the two stages (i.e., representation and fusion) should be considered separately. We maintain that the former stands for the initial step to set interconnections among early-extracted multimodal features, while the latter, despite dealing again with modalities relationships, involves features that have been further processed towards the task optimization (i.e., recommendation in our case), thus embodying different rationales and techniques. Furthermore, the related literature suggests two possible solutions to multimodal representation, either *Joint* or *Coordinate*, where the latter additionally requires the subsequent fusion step. However, each of the paradigms' advantages and whether they might depend on the task remain under investigation.

5.4 Multimodal-aware fusion and optimization

While multimodal representation builds on input modalities in the early stages of the schema, multimodal fusion accounts for multimodal features that have already been processed, with a specific focus on the last steps (i.e., inference and model optimization). Similarly to what was observed above, multimodal fusion may come in the form of *Early* or *Late* fusion. The significant difference between the two approaches lies in preserving or not modalities separation during the inference (i.e., *Late* and *Early*, respectively). The literature demonstrates the vast predominance of *Early* solutions, whereas several works quite often refer to *Late* fusion by mistaking it for *Early* fusion. Indeed, providing a precise definition for the two is of paramount importance because the two approaches may serve different purposes. The rationale of *Late* fusion is to keep the modalities separation explicit during the inference phase so that the contribution of each modality is observable up to the last operation. Moreover, the literature is not aligned on the operation to fuse modalities. Non-trainable fusion functions (e.g., element-wise addition) are usually the preferred direction given that it is more lightweight and easy to perform to trainable approaches (e.g., neural networks) which (on their side) may allow to better tailor user-item preference prediction.

6 FUTURE DIRECTIONS (RQ4 — EXTENSION)

The scope of this section is to outline possible future directions for the application of multimodality for multimedia recommendation. While some of the presented solutions may apply to the above-raised challenges, we also discuss different research paths we suggest to follow in future work.

6.1 Domain-specific multimodal features

Given the limitations imposed by the adoption of pre-trained multimodal features (see again Section 5.2), we wish to underline the benefits of *domain-specific* features in the multimodal schema we have outlined. Extracting such high-level features from input data would entail injecting meaningful and task-aware informative content into the recommendation system, thereby better-profiling items and users on the platform to generate more tailored recommendations. Domain-specific features should necessitate domain-specific extraction models, which may have been previously trained and optimized on similar tasks to the one we are pursuing. Regarding *fashion* recommendation, for instance, we recall the work by Ge et al. [35], a pre-trained architecture for the comprehensive visual analysis of clothing photos. Another example is the *food* recommendation system proposed by Yang et al. [122], which analyzes food-related photos.

Furthermore, in the field of *audio* and *text* understanding and classification, Choi et al. [25] construct a deep model based on convolutional recurrent neural networks for music tagging by taking into account songs' local features and temporal characteristics, whereas Barbieri et al. [9] tackle the task of sentiment analysis in user-generated tweets.

6.2 Multimodality on user-item interactions

Multimodality is the most intuitive approach to describe the multifaceted nature of items in multimedia recommendation [30, 31], but this does not hold for the users' profile.

First, from a technical point of view, profiling each user through multimodal features (e.g., her voice, her visual appearance) would require sophisticated technologies that users' digital devices could not necessarily support (e.g., smartphones). Second, from a practical point of view, it is likely that users would not be disposed to share such personal data on online platforms, primarily for privacy concerns. Despite the raised critical aspects, a few examples from the literature [104, 118] propose to model the user profile in such a way that her preferences toward each multimodal aspect of items are made explicit and learned during model's training. However, these systems rely solely on the multimodal profiles of the items, disregarding alternative information sources. *Product reviews*, which express opinions and comments about items that have been clicked, watched, or purchased, could be a valuable tool for revealing users' nuanced preferences toward each item in the recommendation system.

Existing review-based approaches [114, 141] work by integrating reviews as the *textual modality* to represent *items*. However, we believe that a more logical and effective way to integrate reviews would be to view them as a medium to represent *user* preference *over items*, thereby providing additional and complementary preference scores in addition to numeric ratings or implicit feedback that are typically used to compute recommendations. Such reasoning may be easily generalized to include user-generated data regarding interacting things (such as images or videos of delivered products), which we can characterize as *multimodal feedback* (see Figure 3). When compared to numerical feedback, which tends to be *atomic* (single-faceted), multimodal feedback could be considered as *composite*, revealing nuance and the user's multi-faceted opinion of the products [3].

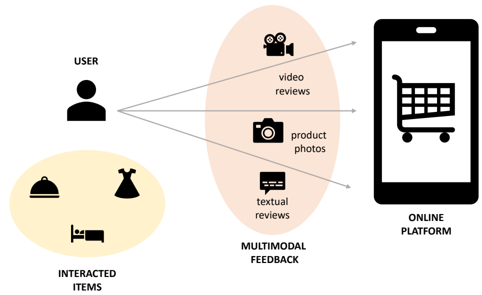

Fig. 3. An example of how users generate and upload *multimodal feedback* about interacted items (e.g., textual reviews, product photos, or even video reviews) on online platforms. Such *user-item* sources of information may be suitably exploited to better profile user' preferences [3].

6.3 Fine-grained multimodal features

Multimodality is a way to effectively profile the multi-faceted aspects of items and users' preference (e.g., I bought this smartphone because its technical *description* is quite exhaustive and its *display* amazes me; I like this song since I love the *music* and the *lyrics*). Nevertheless, analyzing and learning users' tastes at modalities' granularity might not be enough to uncover all aspects underlying every user-item interaction. In contexts where modalities bring a great source of heterogeneous information, a finer-grained feature processing could help better unveil hidden facets. For instance, when it comes to the recommendation of fashion items (e.g., dresses, shoes, jewelry), user attention may be captivated by specific item visual characteristics, such as colors, shapes, and particular patterns and motifs [28]. Similarly, a song involves several features [62] (i.e., pitch, rhythm, and dynamics), which could differently influence users' attitudes towards it. Uncovering and understanding details at this finer granularity should be one of the main directions toward the novel recommendation approaches of multimedia products and services.

6.4 An extensive and fair evaluation of multimedia recommender systems

To date, very limited effort has been put into the extensive evaluation of multimedia recommender systems. The principal reason is that, apart from some recent frameworks [92, 143] which integrate multimedia recommender systems into their pipelines, each novel multimedia recommender system introduces its own implementation of the proposed approach with different dataset pre-processing solutions, sampling strategies, and evaluation protocols. Indeed, this may undermine the fair comparison of multimedia recommender systems, which cannot benefit from shared and unified training and evaluation frameworks to run rigorous and reproducible experiments as in other recommendation domains and scenarios [2, 140]. To this end, we plan to start from the initial benchmarking analysis we proposed in this work to further assess the reproducibility of the tested baselines. On such basis, the next steps would be to evaluate the recommendation performance under more comprehensive experimental settings involving, for instance, (i) a larger plethora of pre-trained deep learning models for the extraction of multimodal features; (ii) other multimodal datasets involving all modalities (as our framework offers the possibility to inject visual, textual, and audio features); (iii) a more careful evaluation of such models under beyond-accuracy recommendation metrics [75, 76].

7 CONCLUSION

In this paper, we highlighted the importance of formalizing the multimedia recommendation task under the lens of multimodal deep learning. By recognizing how the main recommendation approaches in the related literature fall into some recurrent strategy patterns, we outlined a unified multimodal schema that, following the established multimodal deep learning pipeline, formalizes the core stages of multimedia recommendation as (i) multimodal input data, (ii) multimodal feature processing, (iii) multimodal feature fusion, and (iv) the multimodal recommendation task. By applying each of the outlined phases to four selected multimedia recommendation scenarios, we conceptually validated its rationale. Then, we integrated the same schema into Elliot to benchmark the results of six state-of-the-art multimedia recommender systems. The obtained results, which assess the recommendation performance in terms of accuracy and beyond-accuracy measures, along with the proposed formal schema, allowed highlighting technical challenges as well as possible avenues to address such challenges in future directions.

ACKNOWLEDGMENTS

This work was partially supported by the following projects: CT_FINCONS_III, OVS Fashion Retail Reloaded, LUTECH DIGITALE 4.0, VAI2C, IDENTITA, REACH-XY. We acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources and support.

REFERENCES

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

[2] 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.

[3] Vito Walter Anelli, Yashar Deldjoo, Tommaso Di Noia, Eugenio Di Sciascio, Antonio Ferrara, Daniele Malitesta, and Claudio Pomo. 2022. Reshaping Graph Recommendation with Edge Graph Collaborative Filtering and Customer Reviews. In *DL4SR@CIKM (CEUR Workshop Proceedings, Vol. 3317)*. CEUR-WS.org.

[4] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A Simple but Tough-to-Beat Baseline for Sentence Embeddings. In *ICLR (Poster)*. OpenReview.net.

[5] Matteo Attimonelli, Danilo Danese, Daniele Malitesta, Claudio Pomo, Giuseppe Gassi, and Tommaso Di Noia. 2024. Ducho 2.0: Towards a More Up-to-Date Unified Framework for the Extraction of Multimodal Features in Recommendation. *CoRR* abs/2403.04503 (2024).

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

[7] 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.

[8] 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.

[9] Francesco Barbieri, José Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In *EMNLP (Findings) (Findings of ACL, Vol. EMNLP 2020)*. Association for Computational Linguistics, 1644–1650.

[10] Gaurav Bhatnagar, Q. M. Jonathan Wu, and Zheng Liu. 2013. Directive Contrast Based Multimodal Medical Image Fusion in NSCT Domain. *IEEE Trans. Multim.* 15, 5 (2013), 1014–1024.

[11] 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.

[12] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuScenes: A Multimodal Dataset for Autonomous Driving. In *CVPR*. Computer Vision Foundation / IEEE, 11618–11628.

[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] Dapeng Chen, Min Wang, Haobin Chen, Lin Wu, Jing Qin, and Wei Peng. 2022. Cross-Modal Retrieval with Heterogeneous Graph Embedding. In *ACM Multimedia*. ACM, 3291–3300.

[15] Feiyu Chen, Junjie Wang, Yinwei Wei, Hai-Tao Zheng, and Jie Shao. 2022. Breaking Isolation: Multimodal Graph Fusion for Multimedia Recommendation by Edge-wise Modulation. In *ACM Multimedia*. ACM, 385–394.

[16] Huiyuan Chen and Jing Li. 2020. Neural Tensor Model for Learning Multi-Aspect Factors in Recommender Systems. In *IJCAI*. ijcai.org, 2449–2455.

[17] 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.

[18] Tao Chen, Xiangnan He, and Min-Yen Kan. 2016. Context-aware Image Tweet Modelling and Recommendation. In *ACM Multimedia*. ACM, 1018–1027.

[19] 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.

[20] 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.

[21] 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.

[22] Zhiyong Cheng, Xiaojun Chang, Lei Zhu, Rose Catherine Kanjirathinkal, and Mohan S. Kankanhalli. 2019. MMALFM: Explainable Recommendation by Leveraging Reviews and Images. *ACM Trans. Inf. Syst.* 37, 2 (2019), 16:1–16:28.

[23] Zhiyong Cheng, Jialie Shen, and Steven C. H. Hoi. 2016. On Effective Personalized Music Retrieval by Exploring Online User Behaviors. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, Raffaele Perego, Fabrizio Sebastiani, Javed A. Aslam, Ian Ruthven, and Justin Zobel (Eds.). ACM, 125–134. https://doi.org/10.1145/2911451.2911491

[24] Sameer Chhabra. [n. d.]. Netflix says 80 percent of watched content is based on algorithmic recommendations. https://mobilesyrup.com/2017/08/22/80-percent-netflix-shows-discovered-recommendation/. Accessed: 2021-03-13.

[25] Keunwoo Choi, György Fazekas, Mark B. Sandler, and Kyunghyun Cho. 2017. Convolutional recurrent neural networks for music classification. In *ICASSP*. IEEE, 2392–2396.

[26] Qiang Cui, Shu Wu, Qiang Liu, Wen Zhong, and Liang Wang. 2020. MV-RNN: A Multi-View Recurrent Neural Network for Sequential Recommendation. *IEEE Trans. Knowl. Data Eng.* 32, 2 (2020), 317–331.

[27] 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.

[28] 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.

[29] Yashar Deldjoo, Tommaso Di Noia, Eugenio Di Sciascio, and Felice Antonio Merra. 2020. How Dataset Characteristics Affect the Robustness of Collaborative Recommendation Models. In *SIGIR*. ACM, 951–960.

[30] Yashar Deldjoo, Markus Schedl, Paolo Cremonesi, and Gabriella Pasi. 2020. Recommender Systems Leveraging Multimedia Content. *ACM Comput. Surv.* 53, 5 (2020), 106:1–106:38.

[31] Yashar Deldjoo, Markus Schedl, Balasz Hidasi, Xiangnan He, and Yinwei Wei. 2022. Multimedia Recommender Systems: Algorithms and Challenges. In *Recommender Systems Handbook*. Springer US.

[32] Xue Dong, Xuemeng Song, Fuli Feng, Peiguang Jing, Xin-Shun Xu, and Liqiang Nie. 2019. Personalized Capsule Wardrobe Creation with Garment and User Modeling. In *ACM Multimedia*. ACM, 302–310.

[33] Andrea Ferracani, Daniele Pezzatini, Marco Bertini, Saverio Meucci, and Alberto Del Bimbo. 2015. A System for Video Recommendation using Visual Saliency, Crowdsourced and Automatic Annotations. In *ACM Multimedia*. ACM.

[34] Jing Gao, Peng Li, Zhikui Chen, and Jianing Zhang. 2020. A Survey on Deep Learning for Multimodal Data Fusion. *Neural Comput.* 32, 5 (2020), 829–864.

[35] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. 2019. DeepFashion2: A Versatile Benchmark for Detection, Pose Estimation, Segmentation and Re-Identification of Clothing Images. In *CVPR*. Computer Vision Foundation / IEEE, 5337–5345.

[36] Mariana-Iuliana Georgescu, Radu Tudor Ionescu, Andreea-Iuliana Miron, Olivian Savencu, Nicolae-Catalin Ristea, Nicolae Verga, and Fahad Shahbaz Khan. 2023. Multimodal Multi-Head Convolutional Attention with Various Kernel Sizes for Medical Image Super-Resolution. In *WACV*. IEEE, 2194–2204.

[37] Alex Graves. 2012. *Supervised Sequence Labelling with Recurrent Neural Networks*. Studies in Computational Intelligence, Vol. 385. Springer.

[38] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S. Davis. 2017. Learning Fashion Compatibility with Bidirectional LSTMs. In *ACM Multimedia*. ACM.

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

[40] Ruining He and Julian J. McAuley. 2016. Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. In *WWW*. ACM, 507–517.

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

[42] 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.

[43] Haithem Hermessi, Olfa Mourali, and Ezzeddine Zagrouba. 2021. Multimodal medical image fusion review: Theoretical background and recent advances. *Signal Process.* 183 (2021), 108036.

[44] 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.
[45] Peng Hu, Liangli Zhen, Dezhong Peng, and Pei Liu. 2019. Scalable Deep Multimodal Learning for Cross-Modal Retrieval. In *SIGIR*. ACM, 635–644.
[46] Xuming Hu, Zhijiang Guo, Zhiyang Teng, Irwin King, and Philip S. Yu. 2023. Multimodal Relation Extraction with Cross-Modal Retrieval and Synthesis. In *ACL (2)*. Association for Computational Linguistics, 303–311.
[47] Chao Huang, Lianghao Xia, Xiang Wang, Xiangnan He, and Dawei Yin. 2022. Self-Supervised Learning for Recommendation. In *CIKM*. ACM, 5136–5139.
[48] 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.
[49] Xiaowei Jia, Aosen Wang, Xiaoyi Li, Guangxu Xun, Wenyao Xu, and Aidong Zhang. 2015. Multi-modal learning for video recommendation based on mobile application usage. In *IEEE BigData*. IEEE Computer Society, 837–842.
[50] Atsushi Kawasaki and Akihito Seki. 2021. Multimodal Trajectory Predictions for Autonomous Driving without a Detailed Prior Map. In *WACV*. IEEE, 3722–3731.
[51] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning. In *NeurIPS*.
[52] 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.
[53] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *EMNLP*. ACL, 1746–1751.
[54] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR (Poster)*. OpenReview.net.
[55] Yehuda Koren, Robert M. Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *Computer* 42, 8 (2009), 30–37.
[56] Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. 2023. Multimodal Prompting with Missing Modalities for Visual Recognition. In *CVPR*. IEEE, 14943–14952.
[57] Fei Lei, Zhongqi Cao, Yuning Yang, Yibo Ding, and Cong Zhang. 2023. Learning the User's Deeper Preferences for Multi-Modal Recommendation Systems. *ACM Trans. Multimedia Comput. Commun. Appl.* 19, 3s, Article 138 (feb 2023), 18 pages. https://doi.org/10.1145/3573010
[58] 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.
[59] Jiang Li, Xiaoping Wang, Guoqing Lv, and Zhigang Zeng. 2023. GraphMFT: A graph network based multimodal fusion technique for emotion recognition in conversation. *Neurocomputing* 550 (2023), 126427.
[60] 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.
[61] Zhan Li, Jinye Peng, Guohua Geng, Xiaojiang Chen, and Pan-Pan Zheng. 2015. Video recommendation based on multi-modal information and multiple kernel. *Multim. Tools Appl.* 74, 13 (2015), 4599–4616.
[62] Hongru Liang, Wenqiang Lei, Paul Yaozhu Chan, Zhenglu Yang, Maosong Sun, and Tat-Seng Chua. 2020. PiRhDy: Learning Pitch-, Rhythm-, and Dynamics-aware Embeddings for Symbolic Music. In *ACM Multimedia*. ACM, 574–582.
[63] Fan Liu, Huilin Chen, Zhiyong Cheng, Liqiang Nie, and Mohan S. Kankanhalli. 2023. Semantic-Guided Feature Distillation for Multimodal Recommendation. In *ACM Multimedia*. ACM, 6567–6575.
[64] 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.
[65] Kang Liu, Feng Xue, Dan Guo, Peijie Sun, Shengsheng Qian, and Richang Hong. 2023. Multimodal Graph Contrastive Learning for Multimedia-Based Recommendation. *IEEE Trans. Multim.* 25 (2023), 9343–9355.
[66] Kang Liu, Feng Xue, Dan Guo, Le Wu, Shujie Li, and Richang Hong. 2023. MEGCF: Multimodal Entity Graph Collaborative Filtering for Personalized Recommendation. *ACM Trans. Inf. Syst.* 41, 2 (2023), 30:1–30:27.
[67] Qidong Liu, Jiaxi Hu, Yutian Xiao, Jingtong Gao, and Xiangyu Zhao. 2023. Multimodal Recommender Systems: A Survey. *CoRR* abs/2302.03883 (2023).
[68] 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.
[69] Zhuang Liu, Yunpu Ma, Matthias Schubert, Yuanxin Ouyang, and Zhang Xiong. 2022. Multi-Modal Contrastive Pre-training for Recommendation. In *ICMR*. ACM, 99–108.
[70] Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. 2021. Progressive Modality Reinforcement for Human Multimodal Emotion Recognition From Unaligned Multimodal Sequences. In *CVPR*. Computer Vision Foundation / IEEE, 2554–2562.
[71] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. 2022. Are Multimodal Transformers Robust to Missing Modality?. In *CVPR*. IEEE, 18156–18165.
[72] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. 2021. SMIL: Multimodal Learning with Severely Missing Modality. In *AAAI*. AAAI Press, 2302–2310.
[73] Yao Ma, Shilin Zhao, Weixiao Wang, Yaoman Li, and Irwin King. 2022. Multimodality in meta-learning: A comprehensive survey. *Knowl. Based Syst.* 250 (2022), 108976.
[74] Ian MacKenzie, Chris Meyer, and Steve Noble. 2013. How retailers can keep up with consumers. *McKinsey & Company* 18 (2013).

[75] 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.

[76] Daniele Malitesta, Giandomenico Cornacchia, Claudio Pomo, and Tommaso Di Noia. 2023. On Popularity Bias of Multimodal-aware Recommender Systems: A Modalities-driven Analysis. In *MMIR@MM*. ACM, 59–68.

[77] Daniele Malitesta, Giuseppe Gassi, Claudio Pomo, and Tommaso Di Noia. 2023. Ducho: A Unified Framework for the Extraction of Multimodal Features in Recommendation. In *ACM Multimedia*. ACM, 9668–9671.

[78] Daniele Malitesta, Claudio Pomo, Vito Walter Anelli, Alberto Carlo Maria Mancino, Eugenio Di Sciascio, and Tommaso Di Noia. 2023. A Topology-aware Analysis of Graph Collaborative Filtering. *CoRR* abs/2308.10778 (2023).

[79] Daniele Malitesta, Emanuele Rossi, Claudio Pomo, Fragkiskos D. Malliaros, and Tommaso Di Noia. 2024. Dealing with Missing Modalities in Multimodal Recommendation: a Feature Propagation-based Approach. *CoRR* abs/2403.19841 (2024).

[80] 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.

[81] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *ICLR (Workshop Poster)*.

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

[83] Zongshen Mu, Yueting Zhuang, Jie Tan, Jun Xiao, and Siliang Tang. 2022. Learning Hybrid Behavior Patterns for Multimedia Recommendation. In *ACM Multimedia*. ACM, 376–384.

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

[85] Weizhi Nie, Anan Liu, Xiaorong Zhu, and Yuting Su. 2016. Quality models for venue recommendation in location-based social network. *Multim. Tools Appl.* 75, 20 (2016), 12521–12534.

[86] 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.

[87] Xichen Pan, Peiyu Chen, Yichen Gong, Helong Zhou, Xinbing Wang, and Zhouhan Lin. 2022. Leveraging Unimodal Self-Supervised Learning for Multimodal Audio-Visual Speech Recognition. In *ACL (1)*. Association for Computational Linguistics, 4491–4503.

[88] Georgios Paraskevopoulos, Srinivas Parthasarathy, Aparna Khare, and Shiva Sundaram. 2020. Multimodal and Multiresolution Speech Recognition with Transformers. In *ACL*. Association for Computational Linguistics, 2381–2387.

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

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

[91] Emanuele Rossi, Henry Kenlay, Maria I. Gorinova, Benjamin Paul Chamberlain, Xiaowen Dong, and Michael M. Bronstein. 2022. On the Unreasonable Effectiveness of Feature Propagation in Learning on Graphs With Missing Node Features. In *LoG (Proceedings of Machine Learning Research, Vol. 198)*. PMLR, 11.

[92] Aghiles Salah, Quoc-Tuan Truong, and Hady W. Lauw. 2020. Cornac: A Comparative Framework for Multimodal Recommender Systems. *J. Mach. Learn. Res.* 21 (2020), 95:1–95:5.

[93] Lei Sang, Min Xu, Shengsheng Qian, Matt Martin, Peter Li, and Xindong Wu. 2021. Context-Dependent Propagating-Based Video Recommendation in Multimodal Heterogeneous Information Networks. *IEEE Trans. Multim.* 23 (2021), 2019–2032.

[94] Guy Shani and Asela Gunawardana. 2011. Evaluating Recommendation Systems. In *Recommender Systems Handbook*. Springer, 257–297.

[95] Tiancheng Shen, Jia Jia, Yan Li, Hanjie Wang, and Bo Chen. 2020. Enhancing Music Recommendation with Social Media Content: an Attentive Multimodal Autoencoder Approach. In *IJCNN*. IEEE, 1–8.

[96] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*.

[97] Rui Sun, Xuezhi Cao, Yan Zhao, Junchen Wan, Kun Zhou, Fuzheng Zhang, Zhongyuan Wang, and Kai Zheng. 2020. Multi-modal Knowledge Graphs for Recommender Systems. In *CIKM*. ACM, 1405–1414.

[98] Wenlong Sun, Sami Khenissi, Olfa Nasraoui, and Patrick Shafto. 2019. Debiasing the Human-Recommender System Feedback Loop in Collaborative Filtering. In *WWW (Companion Volume)*. ACM, 645–651.

[99] Wangbin Sun, Fei Ma, Yang Li, Shao-Lun Huang, Shiguang Ni, and Lin Zhang. 2021. Semi-Supervised Multimodal Image Translation for Missing Modality Imputation. In *ICASSP*. IEEE, 4320–4324.

[100] Anirudh Sundar and Larry Heck. 2022. Multimodal Conversational AI: A Survey of Datasets and Approaches. In *ConvAI@ACL*. Association for Computational Linguistics, 131–147.

[101] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *CVPR*. IEEE Computer Society, 2818–2826.

[102] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. 2018. A Survey on Deep Transfer Learning. In *ICANN (3) (Lecture Notes in Computer Science, Vol. 11141)*. Springer, 270–279.

[103] Wei Tang, Fazhi He, Yu Liu, and Yansong Duan. 2022. MATR: Multimodal Medical Image Fusion via Multiscale Adaptive Transformer. *IEEE Trans. Image Process.* 31 (2022), 5134–5149.

[104] 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.

[105] Nahed Tawfik, Heba A. Elnemr, Mahmoud Fakhr, Moawad I. Dessouky, and Fathi E. Abd El-Samie. 2021. Survey study of multimodality medical image fusion methods. *Multim. Tools Appl.* 80, 4 (2021), 6369–6396.

[106] Quoc-Tuan Truong, Aghiles Salah, and Hady W. Lauw. 2021. Multi-Modal Recommender Systems: Hands-On Exploration. In *RecSys*. ACM, 834–837.

[107] Saúl Vargas. 2014. Novelty and diversity enhancement and evaluation in recommender systems and information retrieval. In *SIGIR*. ACM, 1281.

[108] Saul Vargas and Pablo Castells. 2011. Rank and relevance in novelty and diversity metrics for recommender systems. In *RecSys*. ACM, 109–116.

[109] Kunal Vaswani, Yudhik Agrawal, and Vinoo Alluri. 2021. Multimodal Fusion Based Attentive Networks for Sequential Music Recommendation. In *BigMM*. IEEE, 25–32.

[110] Cheng Wang, Mathias Niepert, and Hui Li. 2018. LRMM: Learning to Recommend with Missing Modalities. In *EMNLP*. Association for Computational Linguistics.

[111] 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.

[112] 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.

[113] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural Graph Collaborative Filtering. In *SIGIR*. ACM, 165–174.

[114] Xi Wang, Iadh Ounis, and Craig Macdonald. 2021. Leveraging Review Properties for Effective Recommendation. In *WWW*. ACM / IW3C2, 2209–2219.

[115] Wei Wei, Chao Huang, Lianghao Xia, and Chuxu Zhang. 2023. Multi-Modal Self-Supervised Learning for Recommendation. In *WWW*. ACM, 790–800.

[116] Yinwei Wei, Xiang Wang, Xiangnan He, Liqiang Nie, Yong Rui, and Tat-Seng Chua. 2022. Hierarchical User Intent Graph Network for Multimedia Recommendation. *IEEE Trans. Multim.* 24 (2022), 2701–2712.

[117] 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.

[118] 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.

[119] Chuhan Wu, Fangzhao Wu, Tao Qi, Chao Zhang, Yongfeng Huang, and Tong Xu. 2022. MM-Rec: Visiolinguistic Model Empowered Multimodal News Recommendation. In *SIGIR*. ACM, 2560–2564.

[120] Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. 2021. Self-supervised Graph Learning for Recommendation. In *SIGIR*. ACM, 726–735.

[121] Yi Xiao, Felipe Codevilla, Akhil Gurram, Onay Urfalioglu, and Antonio M. López. 2022. Multimodal End-to-End Autonomous Driving. *IEEE Trans. Intell. Transp. Syst.* 23, 1 (2022), 537–547.

[122] Longqi Yang, Cheng-Kang Hsieh, Hongjian Yang, John P. Pollak, Nicola Dell, Serge J. Belongie, Curtis Cole, and Deborah Estrin. 2017. Yum-Me: A Personalized Nutrient-Based Meal Recommender System. *ACM Trans. Inf. Syst.* 36, 1 (2017), 7:1–7:31.

[123] Qi Yang, Gaosheng Wu, Yuhua Li, Ruixuan Li, Xiwu Gu, Huicai Deng, and Junzhuang Wu. 2020. AMNN: Attention-Based Multimodal Neural Network Model for Hashtag Recommendation. *IEEE Trans. Comput. Soc. Syst.* 7, 3 (2020), 768–779.

[124] Xun Yang, Xiaoyu Du, and Meng Wang. 2020. Learning to Match on Graph for Fashion Compatibility Modeling. In *AAAI*. AAAI Press, 287–294.

[125] Jing Yi and Zhenzhong Chen. 2022. Multi-Modal Variational Graph Auto-Encoder for Recommendation Systems. *IEEE Trans. Multim.* 24 (2022), 1067–1079.

[126] Zixuan Yi, Xi Wang, Iadh Ounis, and Craig MacDonald. 2022. Multi-modal Graph Contrastive Learning for Micro-video Recommendation. In *SIGIR*. ACM, 1807–1811.

[127] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A Survey on Multimodal Large Language Models. *CoRR* abs/2306.13549 (2023).

[128] 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.

[129] Penghang Yu, Zhiyi Tan, Guanming Lu, and Bing-Kun Bao. 2023. Multi-View Graph Convolutional Network for Multimedia Recommendation. In *ACM Multimedia*. ACM, 6576–6585.

[130] Tong Yu, Yilin Shen, Ruiyi Zhang, Xiangyu Zeng, and Hongxia Jin. 2019. Vision-Language Recommendation via Attribute Augmented Multimodal Reinforcement Learning. In *ACM Multimedia*. ACM, 39–47.

[131] Jiandian Zeng, Tianyi Liu, and Jiantao Zhou. 2022. Tag-assisted Multimodal Sentiment Analysis under Uncertain Missing Modalities. In *SIGIR*. ACM, 1545–1554.

[132] Huijing Zhan, Jie Lin, Kenan Emir Ak, Boxin Shi, Ling-Yu Duan, and Alex C. Kot. 2022. A^3-FKG: Attentive Attribute-Aware Fashion Knowledge Graph for Outfit Preference Prediction. *IEEE Trans. Multim.* 24 (2022), 819–831.

[133] Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. 2022. M3Care: Learning with Missing Modalities in Multimodal Healthcare Data. In *KDD*. ACM, 2418–2428.

[134] 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.

[135] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Mengqi Zhang, Shu Wu, and Liang Wang. 2021. Latent Structures Mining with Contrastive Modality Fusion for Multimedia Recommendation. *CoRR* abs/2111.00678 (2021).

[136] Qi Zhang, Jiawen Wang, Haoran Huang, Xuanjing Huang, and Yeyun Gong. 2017. Hashtag Recommendation for Multimodal Microblog Using Co-Attention Network. In *IJCAI*. ijcai.org, 3420–3426.

[137] Yongfeng Zhang. 2017. Explainable Recommendation: Theory and Applications. *CoRR* abs/1708.06409 (2017).

[138] Yongfeng Zhang and Xu Chen. 2020. Explainable Recommendation: A Survey and New Perspectives. *Found. Trends Inf. Retr.* 14, 1 (2020), 1–101. https://doi.org/10.1561/1500000066

[139] Ziqi Zhang, Zeyu Li, Kun Wei, Siduo Pan, and Cheng Deng. 2022. A survey on multimodal-guided visual content synthesis. *Neurocomputing* 497 (2022), 110–128.

[140] Wayne Xin Zhao, Yupeng Hou, Xingyu Pan, Chen Yang, Zeyu Zhang, Zihan Lin, Jingsen Zhang, Shuqing Bian, Jiakai Tang, Wenqi Sun, Yushuo Chen, Lanling Xu, Gaowei Zhang, Zhen Tian, Changxin Tian, Shanlei Mu, Xinyan Fan, Xu Chen, and Ji-Rong Wen. 2022. RecBole 2.0: Towards a More Up-to-Date Recommendation Library. In *CIKM*. ACM, 4722–4726.

[141] Lei Zheng, Vahid Noroozi, and Philip S. Yu. 2017. Joint Deep Modeling of Users and Items Using Reviews for Recommendation. In *WSDM*. ACM, 425–434.

[142] Tianyue Zheng, Ang Li, Zhe Chen, Hongbo Wang, and Jun Luo. 2023. AutoFed: Heterogeneity-Aware Federated Multimodal Learning for Robust Autonomous Driving. In *MobiCom*. ACM, 15:1–15:15.

[143] Hongyu Zhou, Xin Zhou, Zhiwei Zeng, Lingzi Zhang, and Zhiqi Shen. 2023. A Comprehensive Survey on Multimodal Recommender Systems: Taxonomy, Evaluation, and Future Directions. *CoRR* abs/2302.04473 (2023).

[144] Xin Zhou and Zhiqi Shen. 2023. A Tale of Two Graphs: Freezing and Denoising Graph Structures for Multimodal Recommendation. In *ACM Multimedia*. ACM, 935–943.

[145] 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.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009