MARIO: Modality-Aware Attention and Modality-Preserving Decoders for Multimedia Recommendation

Taeri Kim*
Hanyang University
Seoul, Korea
taerik@hanyang.ac.kr

Yeon-Chang Lee*
Hanyang University
Seoul, Korea
lyc0324@hanyang.ac.kr

Kijung Shin
KAIST
Seoul, Korea
kijungs@kaist.ac.kr

Sang-Wook Kim†
Hanyang University
Seoul, Korea
wook@hanyang.ac.kr

ABSTRACT

We address the multimedia recommendation problem, which utilizes items' multimodal features, such as visual and textual modalities, in addition to interaction information. While a number of existing multimedia recommender systems have been developed for this problem, we point out that none of these methods individually capture the influence of each modality at the interaction level. More importantly, we experimentally observe that the learning procedures of existing works fail to preserve the intrinsic modality-specific properties of items. To address above limitations, we propose an accurate multimedia recommendation framework, named **MARIO**, based on modality-aware attention and modality-preserving decoders. MARIO predicts users' preferences by considering the individual influence of each modality on each interaction while obtaining item embeddings that preserve the intrinsic modality-specific properties. The experiments on four real-life datasets demonstrate that MARIO consistently and significantly outperforms seven competitors in terms of the recommendation accuracy: MARIO yields up to 14.61% higher accuracy, compared to the best competitor.

CCS CONCEPTS

• **Information systems** → **Recommender systems**.

KEYWORDS

multimedia recommendation; modality-specific properties

ACM Reference Format:
Taeri Kim, Yeon-Chang Lee, Kijung Shin, and Sang-Wook Kim. 2022. MARIO: Modality-Aware Attention and Modality-Preserving Decoders for Multimedia Recommendation. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22), October 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3511808.3557387

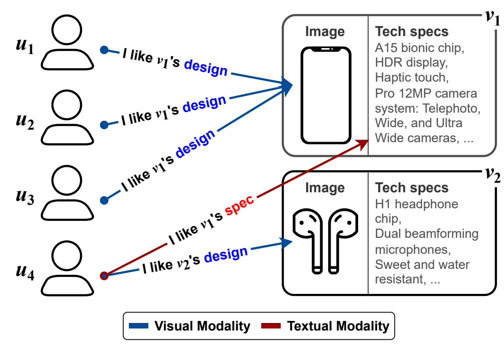


Figure 1: Toy example for the influence of each modality on each interaction. Note that the influence of each modality may differ across interactions.

1 INTRODUCTION

With the development of the Web and storage systems, the amount of available information is rapidly increasing. In addressing the problem of information overload, recommender systems could be a good solution in various domains, such as e-commerce and social media. *Collaborative filtering* (CF), an approach popularly used in recommender systems, exploits users' past interactions such as ratings and click logs for items [2–4, 8, 12, 13, 15, 17, 21, 22]. However, since interaction data are very sparse, CF methods face a difficulty in accurately capturing the preferences of users and the properties of items when they are involved with only a few or no interactions. In order to mitigate this difficulty, many prior works have exploited not only the interaction information but also additional information about users and/or items, *e.g.*, social networks between users [27, 30] and items' multimodal features [1, 5, 7, 18, 25, 26, 31, 32].

In this paper, we focus on multimedia recommender systems [5, 25, 26, 31, 32], which utilize items' *multimodal features* (*e.g.*, visual and textual modalities) along with interaction information. For instance, in the fashion domain, images (*i.e.*, visual modality) or user reviews (*i.e.*, textual modality) for clothes can be regarded as items' multimodal features. Based on such multimodal features, we can better understand items' properties not revealed from interaction information and thus capture users' preferences more accurately.

Most multimedia recommender systems [5, 7, 18, 25, 26, 31] first obtain the pre-trained embeddings of items per modality via deep learning techniques, *e.g.*, convolutional neural networks [16] for visual modality and long short-term memory [9] for textual modality. Then, they learn the embedding of each user and the embedding of each item, obtained by aggregating the pre-trained embeddings, using the interaction information. Lastly, they predict each user's preference for each item based on their embeddings.

Based on the above procedure, recent studies designed an *attention mechanism* to capture the influence of each modality on the

*Two first authors have contributed equally to this work.
†Corresponding author.

CIKM '22, October 17–21, 2022, Atlanta, GA, USA

ACM ISBN 978-1-4503-9236-5/22/10…$15.00
https://doi.org/10.1145/3511808.3557387

interactions between users and items. Specifically, GRCN [25] identifies the *user-level influence* of each modality on *each user* when she/he interacts with items, whereas LATTICE [31] identifies the *global influence* of each modality on *all users* when they interact with items. However, none of these methods capture the *individual influence* of each modality *at the interaction level.*

Furthermore, we point out a non-trivial but overlooked problem of existing multimedia recommender systems [7, 18, 25, 26, 31] in utilizing the multimodal features of items. Note that the pre-trained embeddings of items from each modality reflect their *intrinsic modality-specific properties* that cannot be captured by interaction information. However, we observe that the learning procedures of existing works fail to preserve such modality-specific properties in the final item embeddings. Specifically, they learn the final item embeddings to be similar to the embeddings of users who interact with the items, by aggregating the pre-trained embeddings for modalities based on the interaction information. Accordingly, *the final item embeddings significantly lose the intrinsic modality-specific properties*; this observation will be elaborated more in Section 3. Thus, we claim that the existing works cannot effectively utilize the multimodal features of items as side information.

In this work, we aim to accurately predict users' preferences by considering the influence of each modality on each interaction, as shown in Figure 1, while obtaining item embeddings that preserve the intrinsic modality-specific properties. Towards this goal, we propose **MARIO**, an accurate multimedia recommendation framework based on Modality-aware Attention and modality-pReservIng decOders. It consists of three components: (C1) Encoders based on interaction and multimodal information; (C2) A predictor based on attention network; (C3) Decoders for modality preservation.

In (C1), MARIO first obtains the visual/textual-modality embeddings of each item by encoding its pre-trained embeddings from visual and textual modalities into embeddings of the same dimensionality. Additionally, MARIO obtains the interaction-modality embedding of each item by employing **any** CF method (*e.g.*, BPRMF [21], NGCF [24], and LightGCN [8]) that exploits only interaction information; note that we regard *the interaction information as another modality* in this work. During this process, MARIO obtains the embedding of each user as well. Here, we highlight that this design choice makes any CF methods be easily applied to MARIO to utilize items' multimodal features. In (C2), MARIO predicts each user's preference on each item by considering the influences of the visual, textual, and interaction modalities together. To this end, we design a novel modality-aware attention mechanism that individually identifies the influence of each modality at the interaction level. In (C3), MARIO reconstructs the pre-trained embeddings by decoding the modality embeddings obtained in (C1) for modality preservation.

Finally, MARIO learns the embeddings of users and items to jointly optimize two objectives: the Bayesian personalized ranking (BPR) loss [21] and the modality preservation (MP) loss. The BPR loss is for making each user's predicted preference on rated items higher than that on unrated items while the MP loss is for preserving the intrinsic modality-specific properties in the corresponding modality embeddings of items. By minimizing the above two losses simultaneously, MARIO can fully utilize rich modality-specific semantics in addition to user-item interaction information.

Our contributions are summarized as follows:

- **Observation**: We point out two limitations that existing multimedia recommender systems overlook.
  (1) Existing works do not take into account the individual influence of each modality at the interaction level.
  (2) Learning procedures of existing works fail to preserve the intrinsic modality-specific properties of items, thereby adversely affecting the accuracy. To the best of our knowledge, this work is the first to point out this limitation in the multimedia recommendation problem.
- **General Framework**: We propose MARIO, a novel multimedia recommendation framework based on modality-aware attention and modality-preserving decoders. MARIO is easily equipped with any CF methods based on interaction information.
- **Extensive Evaluation**: We validate the effectiveness of MARIO through extensive experiments using four real-life datasets. MARIO outperforms MAML [18], MMGCN [26], GRCN [25], and LATTICE [31] significantly by up to 14.61%, 94.58%, 33.41%, and 17.21%, respectively, in terms of normalized discounted cumulative gain (NDCG) at top-10 recommendation.

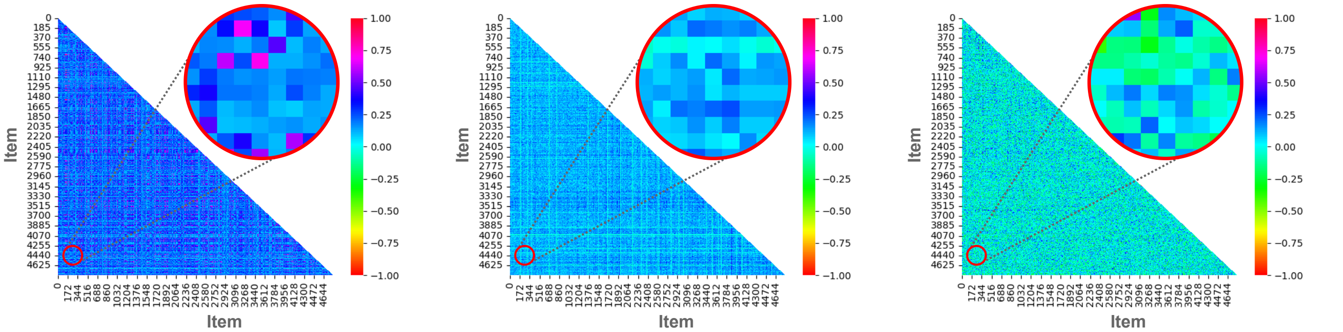
2 RELATED WORK

In this section, we briefly review existing multimedia recommender systems. The multimedia recommender systems exploit extra information of items (*e.g.*, visual and textual modalities) in addition to interaction information between users and items.

Early works have focused on exploiting a single modality of items [1, 6, 7, 11, 19, 23, 28, 29]. CTRank [28] employs a topic model using latent Dirichlet allocation (LDA) to extract textual features of items and combines it with matrix factorization. VBPR [7] employs convolutional neural networks (CNNs) to understand visual features of items and combines it with Bayesian personalized ranking (BPR). Many follow-up studies [1, 6, 11, 19, 23, 29] develop deep-learning-based methods to better capture the intrinsic properties with respect to textual or visual modalities. However, these methods cannot simultaneously exploit items' multimodal features.

Recent multimedia recommender systems [18, 25, 26, 31, 32] utilize items' multimodal features *simultaneously* in addition to interaction information. JRL [32] utilizes each item's visual and textual modalities and interactions to learn the embedding of the corresponding item and the embeddings of users who interact with the item via deep learning techniques. MAML [18] predicts each user's preference on each item by aggregating the user embedding and the item embeddings for the visual and textual modalities, in the Euclidean space. MMGCN [26] builds a modality-aware interaction graph based on the pre-trained item embeddings for each modality and captures each user's modality-specific preference on each item via graph convolution networks (GCNs) [14] on each modality-aware interaction graph. Then, it predicts each user's preference on each item by aggregating all the modality-specific preferences. GRCN [25] prunes noisy edges in the user-item interaction graph based on each user's modality-specific preferences and employs GCNs on the refined graph. Lastly, LATTICE [31] enriches the item embeddings by learning latent item-item relations captured through modality-specific similarities between items. Then, it considers them as the item embeddings of downstream CF methods.

However, the aforementioned approaches capture the influence on each modality only at the user level [25] or global level [31],



(a) Visual Modality (b) Textual Modality (c) Interaction Modality

Figure 2: Similarity of item pairs in their (a) visual modality, (b) textual modality, and (c) interaction modality. The magnified part in each subfigure shows the similarity between the same pairs of items. The results show that, even for the same item pair, the similarities in their visual modality, textual modality, and interaction modality vary significantly.

instead of capturing the influence finely for each *individual interaction*. Furthermore, their learning procedures fail to preserve intrinsic modality-specific properties in the final item embeddings.

3 MOTIVATION

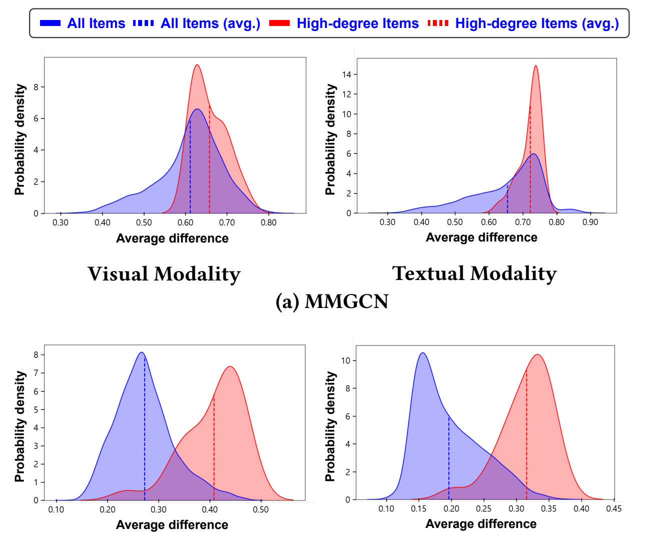
In this section, we demonstrate the limitation of existing multimedia recommender systems through experiments using the Amazon-Men Clothing dataset, which is widely used in multimedia recommendation researches [5, 7, 18, 31, 32].

First, we analyze the intrinsic modality-specific properties of items in the Amazon-Men Clothing dataset. To this end, we obtain the embedding of each item from each modality pre-trained by deep learning techniques; we employ a deep CNN [10] for visual modality, sentence-transformers [20] for textual modality, and LightGCN [8] for interaction modality.[1] Then, for every pair of items, we calculate the cosine similarity of their item embeddings for each modality, *i.e.*, visual modality, textual modality, or interaction modality.

Figures 2-(a), -(b), and -(c) show the results from visual modality, textual modality, and interaction modality, respectively. In every subfigure, the i-th row and the j-th column of colormaps correspond to the i-th item and j-the item, and the color of each (i, j)-th entry indicates the cosine similarity of the embeddings of the i-th and j-th items. The entries are in neon sky blue if the corresponding similarity is zero, and the enlarged part in a red circle in every subfigure shows the entries in the same positions, *i.e.*, similarities of randomly-sampled item pairs. As shown in Figures 2-(a), -(b), and -(c), the similarities vary across item pairs; even for the same item pair, the similarities vary across modalities. In order to statistically validate this claim, we measured the Pearson correlation coefficient between the similarities obtained by (1) visual and textual modalities, (2) visual and interaction modalities, and (3) textual and interaction modalities. As a result, the coefficient values are surprisingly low, ranging from 0.0195 to 0.2191. The results indicate that there exist the intrinsic modality-specific properties of items, which cannot be captured by interaction modality only.

Now, we carefully examine whether the item embeddings learned by existing multimedia recommender systems preserve the modality-specific properties. To this end, we first obtain the final item embeddings from the two state-of-the-art multimedia recommender systems, *spec.*, MMGCN [26] and LATTICE [31]. Then, for each item, we calculate the cosine similarity with every other item by using their final embeddings obtained by each method, which we call *f-similarities*. After that, for each modality, we compare the f-similarities with the similarities of pre-trained item embeddings, which we call *p-similarities* (see Figures 2-(a) and (b) for p-similarities). Lastly, for each item, we compute the average difference between f-similarity and p-similarity with every other item. The average differences indicate how much the final item embeddings of each method lose the intrinsic modality-specific properties in the pre-trained embeddings. That is, the smaller the differences are, the better the modality-specific properties are preserved.

[1] When employing other methods for this purpose, we observed similar tendencies.



Visual Modality Textual Modality
(a) MMGCN

Visual Modality Textual Modality
(b) LATTICE

Figure 3: Density function for the differences between the similarities of the pre-trained item embeddings (obtained from each modality) and the similarities of the final item embeddings (obtained by MMGCN and LATTICE). The modality-specific properties in the pre-trained embeddings are not accurately preserved in the final embeddings.

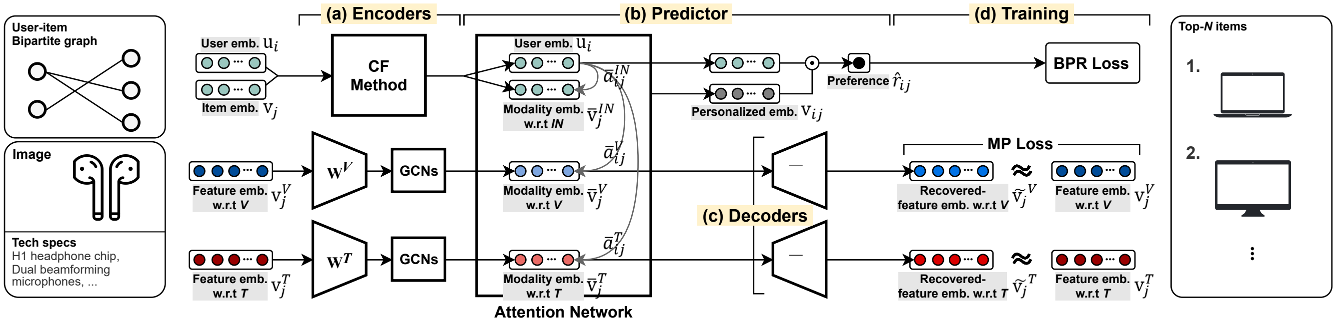


Figure 4: Overview of MARIO, which consists of three components: (C1) Encoders based on interaction and multimodal information; (C2) A predictor based on attention network; (C3) Decoders for modality preservation.

Figure 3 shows the average differences for visual and textual modalities. The x-axis indicates the average differences, and the y-axis indicates the probability density of each difference value. Also, the blue and red colors indicate the results for all items and those for the high-degree items (*spec.*, hub items with more than 38 interactions), respectively; the dotted blue and red lines indicate the average difference for all items and that for the high-degree items, respectively. As shown in Figure 3-(a), by MMGCN, the modality-specific properties of most items are not accurately preserved, regardless of the number of their interactions. However, as shown in Figure 3-(b), when LATTICE is used, the information loss is aggravated for those items with a high number of interactions. Note that the embeddings of such items need to become similar to the embeddings of many users who interacted with them. For these items, the average difference measured for the visual (textual) modality is 0.66 (0.72) and 0.41 (0.32), when MMGCN and LATTICE are used, respectively.

Therefore, we conclude that (1) the final item embeddings of existing multimedia recommender systems significantly lose the intrinsic modality-specific properties, and (2) the information loss is especially severe for the items with many interactions.

4 MARIO: PROPOSED FRAMEWORK

In this section, we propose a novel multimedia recommendation framework, named MARIO, based on modality-aware attention and modality-preserving decoders. In Section 4.1, we first define the multimedia recommendation problem and present the overall procedure of MARIO. In Sections 4.2 and 4.3, we describe the key components and learning methods of MARIO in detail, respectively.

4.1 Overview

The multimedia recommendation problem is defined as follows: Let $u_i \in \mathcal{U}$ and $v_j \in \mathcal{I}$ denote a user and an item, respectively, where $\mathcal{U}$ and $\mathcal{I}$ denote the sets of all users and all items, respectively; $\mathcal{N}_i$ denotes a set of items rated by user u_i. We denote a user u_i's embedding as $\mathbf{u}_i \in \mathbb{R}^d$, where d is the dimensionality of the embedding. We denote an item v_j's (pre-trained) feature embedding with respect to each modality $m \in \mathcal{M}$ as $\mathbf{v}_j^m \in \mathbb{R}^{d_m}$, where d_m denotes the dimensionality of the feature embedding and $\mathcal{M}$ is the set of modalities. In this paper, we use visual, textual, and interaction modalities of each item v_j, *i.e.*, $\mathcal{M} = \{V, T, IN\}$. For each user u_i, the goal is to recommend the top-N items that are most likely to be preferred by u_i among her unrated items, *i.e.*, $\mathcal{I} \setminus \mathcal{N}_i$. Note that, while we focus on three modalities in this paper, if available, additional modalities (*e.g.*, audio modality) can easily be incorporated.

Table 1: Key notations used in this paper

Notation	Description
$\mathcal{U}$, $\mathcal{I}$	Set of users u_i and the set of items v_j
$\mathcal{N}_i$, $\mathcal{M}$	Set of items rated by u_i and the set of item modalities
V, T, IN	Visual, textual, and interaction modalities
$\mathbf{u}_i$, $\mathbf{v}_{ij}$	Embedding of u_i and personalized embedding of v_j w.r.t u_i
$\mathbf{v}_j^m, \bar{\mathbf{v}}_j^m, \tilde{\mathbf{v}}_j^m$	Feature, modality, and recovered feature embeddings of v_j w.r.t modality m
d	Dimensionality of u_i's embedding and v_j's modality and personalized embeddings
d_m	Dimensionality of v_j's feature and recovered feature embeddings w.r.t modality m
$\bar{a}_{ij}^m$	Influence of modality m on each interaction between u_i and v_j
$\hat{r}_{ij}$	Preference of u_i on v_j

We present the overall procedure of MARIO (see Figure 4). First, MARIO obtains each u_i's embedding $\mathbf{u}_i \in \mathbb{R}^d$ and each v_j's multiple modality embeddings $\bar{\mathbf{v}}_j^V, \bar{\mathbf{v}}_j^T, \bar{\mathbf{v}}_j^{IN} \in \mathbb{R}^d$ with respect to visual, textual, and interaction modalities (Figure 4-(a)). Given $\mathbf{u}_i$, $\bar{\mathbf{v}}_j^V$, $\bar{\mathbf{v}}_j^T$, and $\bar{\mathbf{v}}_j^{IN}$, MARIO uses an attention network to infer the influence $\bar{a}_{ij}^m$ of each modality m on each interaction between u_i and v_j. Then, MARIO obtains v_j's personalized embedding with respect to u_i, which we denote by $\mathbf{v}_{ij} \in \mathbb{R}^d$, based on the modality-specific influences (Figure 4-(b)). Based on $\mathbf{u}_i$ and $\mathbf{v}_{ij}$, MARIO predicts each user u_i's preference $\hat{r}_{ij}$ on each item v_j. At the same time, MARIO uses decoders to preserve each v_j's modality-specific properties in its personalized embedding $\mathbf{v}_{ij}$ (Figure 4-(c)).

Finally, MARIO updates $\mathbf{u}_i$, $\bar{\mathbf{v}}_j^V$, $\bar{\mathbf{v}}_j^T$, and $\bar{\mathbf{v}}_j^{IN}$ aiming to jointly minimize two losses (Figure 4-(d)): (1) the Bayesian personalized ranking (BPR) loss for preserving the interaction information of u_i and v_j and (2) the modality preservation (MP) loss for preserving v_j's modality-specific properties with respect to visual and textual modalities. Table 1 lists the key notations used in this paper.

4.2 Key Components

In this subsection, we describe the three key components (*i.e.*, encoders, a predictor, and decoders) of MARIO in detail.

Encoders. MARIO obtains d-dimensional embeddings $\mathbf{u}_i$ and $\mathbf{v}_j$ of user u_i and item v_j by performing a CF method based on the interaction information. Here, we consider $\mathbf{v}_j$ as an interaction-modality embedding $\bar{\mathbf{v}}_j^{IN}$ that represents v_j's interaction-specific property. Note that MARIO can obtain $\bar{\mathbf{v}}_j^{IN}$ by employing *any* CF methods that learn the interaction information; thus, this design choice makes any CF methods be easily applied to MARIO to utilize items' multimodal features. In Section 5, we demonstrate equipping MARIO with each of three popular CF methods (*spec.*, BPRMF [21], NGCF [24], and LightGCN [8]) dramatically improves their recommendation accuracy.

Next, MARIO obtains v_j's visual- and textual-modality embeddings $\bar{\mathbf{v}}_j^V$ and $\bar{\mathbf{v}}_j^T$ that represent its properties with respect to visual and textual modalities, respectively. To this end, MARIO first encodes v_j's (pre-trained) feature embedding $\mathbf{v}_j^m$ for each modality m into a d-dimensional compressed modality embedding $\bar{\mathbf{v}}_j^m$ via a compression layer as follows: $\forall m \in \mathcal{M}\backslash IN$,

$$\bar{\mathbf{v}}_j^m = \mathbf{v}_j^m \mathbf{W}_m, \tag{1}$$

where $\mathbf{W}_m \in \mathbb{R}^{d_m \times d}$ represents a learnable weight matrix of the compression layer for each modality m.

Then, MARIO enriches $\bar{\mathbf{v}}_j^m$ based on the similarities between v_j and other items v_p with respect to each modality m, which cannot be captured by interaction modality (see Figure 2). To this end, we build two types of k-nearest-neighbor (kNN) item graphs. Specifically, we calculate the cosine similarities s_{jp}^m between v_j's feature embedding $\mathbf{v}_j^m$ and v_p's feature embedding $\mathbf{v}_p^m$, respectively. Based on s_{jp}^m, we construct a set $\mathcal{N}_j^m$ that consists of the most similar k items with v_j, *i.e.*, neighborhood of v_j w.r.t m. In the same manner, we construct a set $\bar{\mathcal{N}}_j^m$ by using v_j's modality embedding $\bar{\mathbf{v}}_j^m$ and v_p's modality embedding $\bar{\mathbf{v}}_p^m$. We repeat these processes for all items and then build the following two kNN item graphs as in [31]: $\forall m \in \mathcal{M}\backslash IN$,

$$\mathbf{A}^m = \begin{cases} s_{jp}^m, & v_p \in \mathcal{N}_j^m, \\ 0, & \text{otherwise}, \end{cases} \quad \bar{\mathbf{A}}^m = \begin{cases} \bar{s}_{jp}^m, & v_p \in \bar{\mathcal{N}}_j^m, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

where $\mathbf{A}^m \in \mathbb{R}^{|I|\times|I|}$ and $\bar{\mathbf{A}}^m \in \mathbb{R}^{|I|\times|I|}$ represent the adjacency matrices of the kNN item graphs based on feature and modality embeddings with respect to each modality m, respectively. Then, we combine $\mathbf{A}^m$ and $\bar{\mathbf{A}}^m$ for each modality m, constructing the final kNN item graph $\mathbf{G}^m$ as follows: $\forall m \in \mathcal{M}\backslash IN$,

$$\mathbf{G}^m = \lambda \mathbf{A}^m + (1-\lambda)\bar{\mathbf{A}}^m, \tag{3}$$

where $\lambda \in (0,1)$ indicates a hyperparameter that controls the weights of $\mathbf{A}^m$ and $\bar{\mathbf{A}}^m$.

Now, MARIO enrichs $\bar{\mathbf{v}}_j^m$ by applying graph convolutional networks (GCNs) to $\mathbf{G}^m$ for each modality m: $\forall m \in \mathcal{M}\backslash IN$,

$$\begin{aligned} (\bar{\mathbf{v}}_j^m)^l &= \sum\nolimits_{v_p \in \mathcal{N}_j^m \cup \bar{\mathcal{N}}_j^m} \bar{g}_{jp}^m (\bar{\mathbf{v}}_p^m)^{l-1}, \\ \bar{\mathbf{G}}^m &= (\mathbf{D}^m)^{-\frac{1}{2}} \mathbf{G}^m (\mathbf{D}^m)^{-\frac{1}{2}}, \end{aligned} \tag{4}$$

where $l \in \{1, \cdots, L\}$ represents the l-th GCNs layer; we set $(\bar{\mathbf{v}}_j^m)^0$ as v_j's modality embedding $\bar{\mathbf{v}}_j^m$ with respect to each modality m.[2]

[2]We also used v_j's interaction-modality embedding $\bar{\mathbf{v}}_j^{IN}$ for this setting, but observed no significant improvement of the recommendation accuracy.

Also, $\mathbf{D}^m \in \mathbb{R}^{|I|\times|I|}$ represents a diagnoal degree matrix of $\mathbf{G}^m$. Finally, MARIO considers v_j's embedding $(\bar{\mathbf{v}}_j^m)^L$ obtained from the last L-th GCNs layer as v_j's final modality embedding $\bar{\mathbf{v}}_j^m$.

Predictor. MARIO predicts each user u_i's preference $\hat{r}_{ij}$ on each item v_j based on $\mathbf{u}_i$, $\bar{\mathbf{v}}_j^V$,$\bar{\mathbf{v}}_j^T$, and $\bar{\mathbf{v}}_j^{IN}$. Here, as mentioned in Section 1, we note that the influence of each modality may vary across interactions. As in the example shown in Figure 1, suppose that four users u_1, u_2, u_3, and u_4 purchased an iPhone 13 (*i.e.*, v_1). In this example, a user u_4 purchased the item v_1 because she/he prefers its spec (*i.e.*, textual modality) while the remaining users u_1, u_2, and u_3 purchased the item v_1 because they prefer its design (*i.e.*, visual modality). Furthermore, u_4 purchased an AirPods 3 (*i.e.*, v_2) because she/he prefers its design rather than its spec.

For this reason, we design a modality-aware attention mechanism to identify the influence of each modality m on each interaction between u_i and v_j. Using $\mathbf{u}_i$ as a query, and $\bar{\mathbf{v}}_j^V$,$\bar{\mathbf{v}}_j^T$, and $\bar{\mathbf{v}}_j^{IN}$ as keys and values, MARIO calculates the influence $\bar{a}_{ij}^m$ of each modality m on each interaction between u_i and v_j as follows: $\forall m \in \mathcal{M}$,

$$\bar{a}_{ij}^m = \frac{\exp(a_{ij}^m)}{\sum_{m \in \mathcal{M}} \exp(a_{ij}^m)}, \quad \text{where} \quad a_{ij}^m = \frac{\mathbf{u}_i \odot \bar{\mathbf{v}}_j^m}{\sqrt{d}}, \tag{5}$$

and $\odot$ and $\sqrt{d}$ represent the dot product and the scaling factor, respectively.

Then, MARIO obtains v_j's *personalized embedding* $\mathbf{v}_{ij}$ with respect to u_i by fusing v_j's modality embeddings $\bar{\mathbf{v}}_j^V$,$\bar{\mathbf{v}}_j^T$, and $\bar{\mathbf{v}}_j^{IN}$ based on their attentions $\bar{a}_{ij}^V$, $\bar{a}_{ij}^T$, and $\bar{a}_{ij}^{IN}$:

$$\mathbf{v}_{ij} = \sum\nolimits_{m \in \mathcal{M}} \bar{a}_{ij}^m \bar{\mathbf{v}}_j^m. \tag{6}$$

Here, we highlight that v_j's personalized embedding $\mathbf{v}_{ij}$ with respect to each user u_i enables to identify which modality m has the most influence on the interaction between u_i and v_j. We further clarify that, unlike MMGCN [26], MARIO does not reflect the interaction information in representing v_j's modality embeddings $\bar{\mathbf{v}}_j^V$ and $\bar{\mathbf{v}}_j^T$ except for $\bar{\mathbf{v}}_j^{IN}$; instead, the interaction information is reflected only when MARIO aggregates all the modality embeddings $\bar{\mathbf{v}}_j^V$, $\bar{\mathbf{v}}_j^T$, and $\bar{\mathbf{v}}_j^{IN}$, which are called *late fusion*. We believe that this design choice contributes to preserving v_j's modality-specific properties in $\bar{\mathbf{v}}_j^V$ and $\bar{\mathbf{v}}_j^T$. As a demonstration, in Section 5, we validate that this late fusion is more effective than early fusion (*i.e.*, the aggregation strategy of MMGCN) in terms of accuracy.

Finally, MARIO predicts u_i's preference on v_j, as follows, and recommends the most preferred top-N items to u_i:

$$\hat{r}_{ij} = \mathbf{u}_i \odot \mathbf{v}_{ij}. \tag{7}$$

Decoders for Modality Preservation. Lastly, we design a decoder layer to effectively preserve v_j's modality-specific properties in its modality embeddings. Recall that MARIO compresses v_j's feature embedding $\mathbf{v}_j^m \in \mathbb{R}^{d_m}$ with respect to each modality $m \in \mathcal{M}\backslash IN$ into the d-dimensional embedding via a compression layer, *i.e.*, Eq. (1). For each modality m, MARIO decodes v_j's compressed modality embedding $\bar{\mathbf{v}}_j^m \in \mathbb{R}^d$ to restore the d_m-dimension of its feature embedding $\mathbf{v}_j^m$ via a decoder layer as follows: $\forall m \in \mathcal{M}\backslash IN$,

$$\tilde{\mathbf{v}}_j^m = \bar{\mathbf{v}}_j^m \bar{\mathbf{W}}^m, \tag{8}$$

Algorithm 1 MARIO

Input: (1) users $\mathcal{U}$; (2) items $\mathcal{V}$; (3) items $\mathcal{N}_i$ rated by each user u_i, for $\forall u_i$; (4) modalities $\mathcal{M}$; (5) feature embeddings $\mathbf{v}_j^m$ of each item v_j, for $\forall v_j, \forall m \in \mathcal{M}$; (6) the number of neighbors k used for kNN graphs; (7) the maximum number of epochs h

Output: $\mathbf{u}_i$, $\bar{\mathbf{v}}_j^m$, for $\forall u_i$, $\forall v_j$, $\forall m$

1: Initialize $\mathbf{u}_i$, $\mathbf{v}_j$, for $\forall u_i$, $\forall v_j$
2: **for** $epoch \leftarrow 1$ to h **do**
3: $\quad$ $\mathbf{u}_i, \bar{\mathbf{v}}_j^{IN} = cf\_method(\mathbf{u}_i, \mathbf{v}_j, \mathcal{N}_i)$, for $\forall u_i$, $\forall v_j$
4: $\quad$ **for** $m \in \mathcal{M} \backslash IN$ **do**
5: $\quad\quad$ $\bar{\mathbf{v}}_j^m \leftarrow \mathbf{v}_j^m \mathbf{W}^m$, for $\forall v_j$ ▷ Compression Layer
6: $\quad\quad$ Construct kNN item graphs for m via Eq. (2) and Eq. (3)
7: $\quad\quad$ Update $\bar{\mathbf{v}}_j^m$ via Eq. (4), for $\forall v_j$ ▷ GCNs Layer
8: $\quad$ **end for**
9: $\quad$ Calculate $\bar{a}_{ij}^m$ via Eq. (5), for $\forall u_i$, $\forall v_j$, $\forall m$ ▷ Attention Network
10: $\quad$ Obtain $\mathbf{v}_{ij}$ via Eq. (6), for $\forall u_i$, $\forall v_j$
11: $\quad$ $\hat{r}_{ij} = \mathbf{u}_i \odot \mathbf{v}_{ij}$, for $\forall u_i$, $\forall v_j$
12: $\quad$ **for** $m \in \mathcal{M} \backslash IN$ **do** ▷ Decoder Layer
13: $\quad\quad$ $\tilde{\mathbf{v}}_j^m = \bar{\mathbf{v}}_j^m \bar{\mathbf{W}}^m$, for $\forall v_j$
14: $\quad$ **end for**
15: $\quad$ Update $\mathbf{u}_i$, $\mathbf{v}_j$, $\mathbf{W}^m$, via Eq. (9), for $\forall u_i$, $\forall v_j$, $\forall m \in \mathcal{M} \backslash IN$
16: ▷ BPR Loss
17: $\quad$ Update $\mathbf{W}^m$, $\bar{\mathbf{W}}^m$ via Eq. (10), for $\forall m \in \mathcal{M} \backslash IN$ ▷ MP Loss
18: **end for**

where $\bar{\mathbf{W}}^m \in \mathbb{R}^{d \times d_m}$ represents a learnable weight matrix of the decoder layer for each modality m. The v_j's recovered feature embedding $\tilde{\mathbf{v}}_j^m$ with respect to each modality m is used to preserve v_j's modality-specific properties in the training process of MARIO, as elaborated in detail in the following subsection.

4.3 Training

Finally, we learn each user u_i's embedding $\mathbf{u}_i$ and each item v_j's modality embeddings $\bar{\mathbf{v}}_j^V$, $\bar{\mathbf{v}}_j^T$, and $\bar{\mathbf{v}}_j^{IN}$ with the loss functions below.

Bayesian Personalized Ranking (BPR) Loss. Basically, MARIO employs the Bayesian personalized ranking (BPR) loss $\mathcal{L}_{BPR}$ to preserve the interaction information:

$$\mathcal{L}_{BPR} = -\sum\nolimits_{u_i \in \mathcal{U}} \sum\nolimits_{v_j \in \mathcal{N}_i} \sum\nolimits_{v_p \in \mathcal{N}_i \backslash \mathcal{I}} \ln \sigma(\hat{r}_{ij} - \hat{r}_{ip}), \quad (9)$$

where σ indicates the sigmoid function. That is, the embeddings of u_i, v_j, v_p are learned based on the intuition that u_i's preference $\hat{r}_{ij}$ on the rated item v_j is likely to be higher than u_i's preference $\hat{r}_{ip}$ on an (randomly-sampled) unrated item v_p. However, as shown in Section 3, we observed that when the parameters are learned based only on the interaction information, the final item embeddings significantly lose their intrinsic modality-specific properties.

Modality Preservation (MP) Loss. To address this problem, MARIO additionally uses a modality preservation (MP) loss to preserve v_j's modality-specific properties in its modality embeddings. Specifically, MARIO aims to minimize the difference between v_j's (pre-trained) feature embedding $\mathbf{v}_j^m$ and v_j's recovered feature embedding $\tilde{\mathbf{v}}_j^m$ (*i.e.*, Eq. (8)) with respect to each modality m as follows:

$$\mathcal{L}_{MP} = \sum\nolimits_{m \in \mathcal{M} \backslash IN} \sum\nolimits_{v_j \in \mathcal{I}} \sum_{f=1}^{d_m} |\mathbf{v}_j^m(f) - \tilde{\mathbf{v}}_j^m(f)|. \quad (10)$$

Table 2: Dataset statistics

Dataset	# User	# Item	# Interaction	Sparsity	Dim. of V / T
Baby	19,445	7,050	160,792	99.88%	4,096 / 1,024
Clothing	4,955	5,028	32,363	99.87%	4,096 / 1,024
Office	4,874	2,406	52,957	99.55%	4,096 / 1,024
Musical	1,429	900	10,261	99.20%	4,096 / 1,024

Note that the interaction modality is not included in the MP loss because MARIO does not assume items' feature embeddings obtained from interaction modality. Via the MP loss, MARIO learns the compression layer (*i.e.*, $\mathbf{W}^m$ in Eq. (1)) and the decoder layer (*i.e.*, $\bar{\mathbf{W}}^m$ in Eq. (8)), aiming to make v_j's visual- and textual-modality embeddings $\bar{\mathbf{v}}_j^V$ and $\bar{\mathbf{v}}_j^T$ preserve the intrinsic modality-specific properties in feature embeddings $\mathbf{v}_j^V$ and $\mathbf{v}_j^T$, respectively. To the best of our knowledge, the MP loss is the first attempt that effectively preserves the intrinsic modality-specific properties for accurate recommendation. We believe that our discovery suggests a promising direction for the multimedia recommendation problem.

Final Loss. The final loss function of MARIO is as follows:

$$\mathcal{L} = \mathcal{L}_{BPR} + \mu \mathcal{L}_{MP}, \quad (11)$$

where $\mu \in (0, 1]$ represents a hyperparameter that controls the balance between $\mathcal{L}_{BPR}$ and $\mathcal{L}_{MP}$. By jointly optimizing the BPR and MP losses, MARIO fully utilizes rich modality-specific semantics in addition to user-item interaction information.

Algorithm 1 sketches the overall procedure of MARIO. First, MARIO obtains user embeddings $\mathbf{u}_i$ and the interaction-modality embeddings $\bar{\mathbf{v}}_j^{IN}$ of items by employing a CF method (Lines 1-3). Also, MARIO obtains the visual- and textual-modality embeddings $\bar{\mathbf{v}}_j^V$ and $\bar{\mathbf{v}}_j^T$ of items based on their feature embeddings $\mathbf{v}_j^V$ and $\mathbf{v}_j^T$, respectively (Lines 4-8). Next, MARIO builds the personalized embeddings $\mathbf{v}_{ij}$ of the items v_j with respect to users u_i by fusing the modality embeddings $\bar{\mathbf{v}}_j^V$, $\bar{\mathbf{v}}_j^T$, and $\bar{\mathbf{v}}_j^{IN}$ based on their attentions $\bar{a}_{ij}^V$, $\bar{a}_{ij}^T$, and $\bar{a}_{ij}^{IN}$ (Lines 9-10). Then, MARIO predicts the preferences $\hat{r}_{ij}$ of users u_i on the items v_j while it decodes the compressed modality embeddings $\bar{\mathbf{v}}_j^m$ to the recovered feature embeddings $\tilde{\mathbf{v}}_j^m$ (Lines 11-14). Finally, MARIO updates user embeddings $\mathbf{u}_i$, interaction-modality embeddings of items $\bar{\mathbf{v}}_j^{IN}$, and $\mathbf{W}^m$ ($m \in \mathcal{M} \backslash IN$) based on the BPR loss (Lines 15-16). In addition, MARIO updates the parameters $\mathbf{W}^m$ and $\bar{\mathbf{W}}^m$ ($m \in \mathcal{M} \backslash IN$) of the compression layer and the decoder layer based on the MP loss (Line 17).

5 EVALUATION

We designed our experiments, aiming at answering the following key research questions (RQs):

- **RQ1**: Does MARIO provide more-accurate top-N recommendation than state-of-the-art multimedia recommender systems?
- **RQ2**: Is exploiting all the item modalities under the MARIO effective for multimedia recommendation?
- **RQ3**: Are the modality-aware attention network of MARIO effective for multimedia recommendation?
- **RQ4**: Does the MP loss of MARIO help modality preservation and accurate multimedia recommendation?
- **RQ5**: Does equipping MARIO with different CF methods consistently improve their accuracies?

Table 3: Accuracies of four multimedia recommender systems and MARIO. The symbol * denotes that p-values are below 0.05, indicating the differences are statistically significant. MARIO significantly outperforms all competitors in most cases. That is, MARIO provides most accurate multimedia recommendation.

Datasets	Baby			Clothing			Office			Musical		
Metrics	**NDCG@10**	**Recall@10**	**Pre@10**	**NDCG@10**	**Recall@10**	**Pre@10**	**NDCG@10**	**Recall@10**	**Pre@10**	**NDCG@10**	**Recall@10**	**Pre@10**
MAML	–	–	–	0.0226	0.0442	0.0044	0.0505	0.0887	**0.0112**	0.0662	0.1302	0.0134
MMGCN	0.0187	0.0370	0.0039	0.0133	0.0260	0.0026	0.0281	0.0534	0.0067	0.0517	0.0980	0.0101
GRCN	0.0262	0.0468	0.0050	0.0194	0.0355	0.0036	0.0553	0.0845	0.0106	0.0541	0.1068	0.0110
LATTICE	0.0276	0.0503	0.0053	0.0221	0.0404	0.0040	**0.0589**	0.0902	0.0109	0.0769	0.1459	0.0148
MARIO	**0.0300***	**0.0539***	**0.0056***	**0.0259***	**0.0484***	**0.0048***	0.0583	**0.0932***	0.0110	**0.0790***	**0.1513***	**0.0153***
Improvement	**8.58%**	**7.20%**	**6.81%**	**14.61%**	**9.59%**	**9.09%**	-0.89%	**3.38%**	-1.47%	**2.73%**	**3.68%**	**3.30%**

Metrics	NDCG@20	Recall@20	Pre@20	NDCG@20	Recall@20	Pre@20	NDCG@20	Recall@20	Pre@20	NDCG@20	Recall@20	Pre@20
MAML	–	–	–	0.0289	0.0694	**0.0035**	0.0628	0.1350	**0.0086**	0.0822	0.1919	0.0099
MMGCN	0.0249	0.0615	0.0033	0.0168	0.0399	0.0020	0.0375	0.0892	0.0056	0.0692	0.1682	0.0086
GRCN	0.0335	0.0743	0.0040	0.0242	0.0544	0.0027	0.0689	0.1326	0.0084	0.0707	0.1726	0.0089
LATTICE	0.0355	0.0804	0.0042	0.0279	0.0634	0.0032	0.0725	0.1360	0.0083	0.0944	0.2126	0.0109
MARIO	**0.0378***	**0.0838***	**0.0044***	**0.0314***	**0.0706***	**0.0035**	**0.0728***	**0.1418***	**0.0086**	**0.0968***	**0.2195***	**0.0113***
Improvement	**6.58%**	**4.13%**	**4.02%**	**8.50%**	**1.79%**	0.00%	**0.38%**	**4.26%**	0.00%	**2.53%**	**3.26%**	**3.21%**

–: out-of-memory

5.1 Experimental Settings

Datasets. We used four real-life Amazon datasets in different categories, which are widely used in previous studies of multimedia recommendation [1, 5, 7, 18, 31, 32]: Baby, Men Clothing (Clothing, in short), Office, and Musical Instruments (Musical, in short). They contain not only the interaction information between users and items but also the visual and textual modalities of items. Following [18], we filtered out the users and the items with less than five interactions. For visual and textual modalities, we used the 4,096- and 1,024-dimensional feature embeddings extracted by a deep CNN [10] and sentence-transformers [20] in the same way as in [31]. All the datasets are publicly available.[3] Table 2 provides some statistics of the four datasets.

Competitors. To evaluate the effectiveness of MARIO, we compare MARIO with seven competitors. Following [31], we employed three CF methods as baselines (*i.e.*, BPRMF [21], NGCF [24], and LightGCN [8]) which exploit the interaction information only. Also, we used four state-of-the-art multimedia recommender systems, *i.e.*, MAML [18], MMGCN [26], GRCN [25], and LATTICE [31]. For evaluation, we used the source code provided by the authors.

Evaluation Task. For testing, we split a dataset into training (80%), validation (10%), and test (10%) sets in the same way as in [25, 26, 31]. Then, we performed top-10/20 recommendation by using each method. To evaluate accuracy, we employed the following three popular measures: precision, recall, and normalized discounted cumulative gain (NDCG). For all of our experiments, we conducted t-tests with a 95% confidence level to verify the accuracy differences between MARIO and competitors. The results in all measures and all datasets show that all p-values are below 0.05, indicating the differences are statistically significant.

Implementation Details. Following [18, 25, 26, 31], for a fair comparison, we set the dimensionality of the embeddings for users and items to 64 in all methods including MARIO. Then, we carefully tuned the hyperparameters of competitors and MARIO. Specifically, for hyperparameters of competitors, we used the best settings found via grid search with the validation set in the following ranges: {0.0001, 0.0005, 0.001, 0.005, 0.01, 0.1, 1} for learning rate; {0, 0.00001, 0.0001, 0.001, 0.01, 0.1} for regularization weight; {128, 256, 512, 1024} for the batch size; {0.1, 0.2, ..., 2.0} for the margin in the hinge loss of MAML; {32, 64, 128} for the dimensionality of latent feature embeddings of MMGCN; {0, 0.1, ..., 0.8} for the dropout ratio of LATTICE. For MARIO, we set its hyperparameters as follows: learning rate=0.0005 for Baby and Office and 0.005 for other datasets; regularization weight=0.00001; batch size=1024; k=10; λ=0.9; μ=1.

[3] http://jmcauley.ucsd.edu/data/amazon/links.html

5.2 Results

Due to space limitations, for RQ2, RQ3, and RQ4, we omit the results of MARIO on Baby, Office, and Musical datasets in this paper. Instead, the details for the omitted results are available at https://sites.google.com/view/mario-cikm2022.

RQ1: Comparison with Four Competitors. We conducted comparative experiments on four datasets to demonstrate the superiority of MARIO over the following four state-of-the-art multimedia recommender systems: MAML [18], MMGCN [26], GRCN [25], and LATTICE [31]. For MARIO, we report the results of MARIO equipped with LightGCN as a CF method. Table 3 shows the results. The values in boldface and underlined indicate the best and 2nd best accuracies in each column (*i.e.*, each measure), respectively. Also, 'Improvement' indicates the degree of accuracy improvements by MARIO over the *best* competitors. Note that, on Baby, the accuracies of MAML could not be obtained due to its out-of-memory issue.[4] Below, we summarize the results in Table 3.

First and most importantly, MARIO significantly outperforms all competitors in almost all cases. Specifically, on Clothing, MARIO outperforms the best competitor (*i.e.*, MAML) by up to 9.09%, 9.59%, and 14.61%, in terms of precision@10, recall@10, and NDCG@10, respectively. Also, on Baby, MARIO outperforms the best competitor (*i.e.*, LATTICE) by up to 6.81%, 7.20%, and 8.58% in terms of precision@10, recall@10, and NDCG@10, respectively.

Second, among the competitors *except for* MARIO, no single method consistently outperforms all others. Best competitors change depending on datasets and measures: LATTICE in terms of all measures on Baby and Musical, and in terms of recall and NDCG on Office; MAML in terms of all measures on Clothing, and in terms of

[4] All the experiments were conducted in Ubuntu 18.04 LTS running on Nvidia v100×2ea and 16 vCPUs with 180GB RAM.

Table 4: The effects of exploiting all the item modalities. MARIO consistently achieves the best accuracy when it exploits all the item modalities.

Model	NDCG@10	Recall@10	Pre@10
MARIO$_{w/o\ \mathrm{V}}$	0.0233	0.0415	0.0042
MARIO$_{w/o\ \mathrm{T}}$	0.0243	0.0460	0.0046
MARIO	**0.0259**	**0.0484**	**0.0048**

Table 5: The effects of our attention network. The modality-aware attention mechanism is most effective.

Model	NDCG@10	Recall@10	Pre@10
MARIO$_{mean}$	0.0256	0.0478	**0.0048**
MARIO$_{max}$	0.0232	0.0429	0.0043
MARIO$_{fc}$	0.0117	0.0214	0.0022
MARIO	**0.0259**	**0.0484**	**0.0048**

precision on Office. Note that our work made a direct comparison between LATTICE and MAML for the first time.

Lastly, we see that MMGCN always shows the lowest accuracy. While MMGCN uses *early fusion* (*i.e.*, the interaction information is reflected when learning the visual- and textual-modality embeddings), the remaining methods, including MARIO, use late fusion. In this context, the result supports the importance of preserving the intrinsic modality-specific properties for accurate multimedia recommendation (as examined in greater detail in RQ4).

RQ2: Effectiveness of Item Modalities. We verify whether, under the MARIO, both visual and textual modalities help capture users' preferences accurately. For RQ2, we compare MARIO with its two variants: (1) MARIO$_{w/o\ \mathrm{V}}$ does not use visual-modality embeddings of items; (2) MARIO$_{w/o\ \mathrm{T}}$ does not use textual-modality embeddings of items. As shown in Table 4, MARIO consistently and significantly outperforms the two variants with all measures. Specifically, MARIO improves the accuracies of MARIO$_{w/o\ \mathrm{V}}$ and MARIO$_{w/o\ \mathrm{T}}$, up to 11.38% and 6.51% in terms of NDCG@10, respectively. This result shows that using both visual and textual modalities is effective in MARIO.

RQ3: Effectiveness of Our Attention Network. In Section 4, we designed the attention mechanism that infers the influence of each modality differently on each interaction. For RQ3, we compare MARIO with its three variants: (1) MARIO$_{mean}$ employs the mean pooling when fusing all modality embeddings of items; (2) MARIO$_{max}$ employs the max pooling when fusing all modality embeddings of items; (3) MARIO$_{fc}$ employs a fully connected layer that uses a concatenation of each item's all modality embeddings as an input. Table 5 shows the accuracy of MARIO and its three variants. We observe that MARIO consistently outperforms MARIO$_{mean}$, MARIO$_{max}$, and MARIO$_{fc}$. The results show that it is most effective to use the proposed attention mechanism to aggregate all modality embeddings of items by considering the individual influence of each modality at the interaction level.

RQ4: Effectiveness of Our MP Loss. In Section 3, we demonstrated that the learning procedures of existing works fail to preserve the intrinsic modality-specific properties in final item embeddings. To alleviate such a limitation, we designed the MP loss for modality preservation. To verify its effectiveness, we conduct experiments to answer the following two subquestions:

Table 6: The average differences between p-similarities (which are obtained from each modality) and f-similarities (which are obtained by MMGCN, LATTICE, and MARIO). The smaller the differences are, the better the modality-specific properties are preserved. Note that MARIO is most effective in preserving the intrinsic modality-specific properties.

Modality		MMGCN	LATTICE	MARIO
All Items	Visual	0.6111	0.2721	**0.2668**
	Textual	0.6541	0.1961	**0.1842**
Items with High-degree	Visual	0.6576	0.4075	**0.3062**
	Textual	0.7221	0.3155	**0.2739**

- **RQ4-1**: Do the final item embeddings by MARIO better preserve the intrinsic modality-specific properties?
- **RQ4-2**: Does the MP loss help for the accurate multimedia recommendation?

For RQ4-1, we conducted an experiment in the same way as in Section 3. Specifically, we compared the cosine similarities (*i.e.*, p-similarities) for all item pairs based on pre-trained item embeddings, and those (*i.e.*, f-similarities) based on the final item embeddings obtained by MARIO. Then, for each item, we compared the average of the differences between p-similarities and f-similarities over all other items. Table 6 shows the differences averaged over all items and those averaged only over high-degree items (*spec.*, items with more than 38 interactions) in MMGCN, LATTICE, and MARIO. Due to space limitations, we compare MARIO only with MMGCN and LATTICE, which were analyzed in Section 3. However, other methods resulted in tendencies similar to those shown in Table 6.

First, we observe that the differences are always smallest in MARIO. That is, MARIO is most effective in preserving the intrinsic modality-specific properties. Second, the differences are significantly larger in MMGCN, which performs early fusion, than in LATTICE and MARIO, which perform late fusion. The results show that such early fusion harms accurate preservation of the intrinsic modality-specific properties.

Furthermore, we examine the results for the high-degree items in detail. Note that, for MMGCN and LATTICE, the information loss is especially severe for high-degree items, as we mentioned in Section 3. On the other hand, we observe that the MP loss of MARIO affects modality preservation more for high-degree items than for low-degree items. In particular, the differences for visual modality in MARIO are significantly lower than those in MMGCN and LATTICE. Recall that, as shown in Figure 2, on Clothing, the visual-modality-specific property is quite different from the interaction-modality-specific property, while the textual-modality-specific property is relatively similar to the interaction-modality-specific property. Thus, it means that the visual modality is informative as the additional information, compared to the textual modality; in this regard, in RQ2, we confirmed that exploiting visual modality is more effective rather than exploiting textual modality in terms of recommendation accuracy. From this observation, we expect that the MP loss of MARIO is more effective when the modality-specific property cannot be captured by interaction information.

For RQ4-2, we analyze how the accuracy of MARIO depends on the weight μ for the MP loss. In Figure 6, the x-axis represents the value of μ, and the y-axis does the accuracy. We see that the accuracy

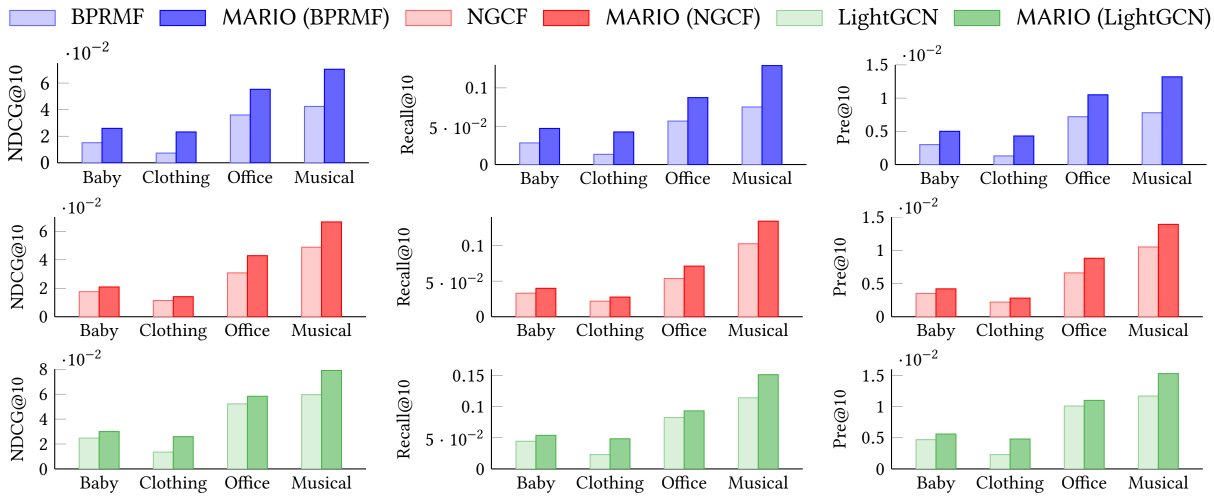


Figure 5: Accuracies of three CF methods and MARIOs equipped with them. Any CF methods can be easily applied to MARIO to utilize items' multimodal features, and by doing so, their accuracies significantly improve.

of MARIO is always the worst when μ=0. Specifically, MARIO yields up to 6.19%, 6.19%, and 2.31% higher precision@10, recall@10, and NDCG@10, respectively, when μ is 1 than when μ is 0. The results show that preserving the modality-specific properties is also effective in improving the accuracy of multimedia recommendation.

RQ5: Effectiveness of MARIOs Equipped with Three CF Methods. MARIO employs a CF method to obtain user embeddings and interaction-modality embeddings of items. We claimed that this design choice makes any CF methods be easily applied to MARIO to utilize items' multimodal features. To validate this claim, we compare the accuracies of three popular CF methods (*i.e.*, BPRMF [21], NGCF [24], and LightGCN [8]) which use only the interaction information, and those of MARIO equipped with them, denoted by MARIO (BPRMF), MARIO (NGCF), and MARIO (LightGCN).

As shown in Figure 5, the three versions of MARIO consistently and dramatically outperform BPRMF, NGCF, and LightGCN, respectively, on all datasets in terms of all measures. On Clothing, MARIO (BPRMF) achieves up to 219.70%, 218.94%, and 215.57% higher precision@10, recall@10, and NDCG@10, respectively, than BPRMF. MARIO (NGCF) also yields up to 26.61%, 26.85%, and 23.45% higher precision@10, recall@10, and NDCG@10, respectively, than NGCF. Lastly, MARIO (LightGCN) gives 110.53%, 110.53%, and 92.52% higher precision@10, recall@10, and NDCG@10, respectively, than LightGCN. Among the three versions, MARIO (LightGCN) achieves the best accuracy on all datasets in terms of all measures.

The results indicate that equipping MARIO with CF methods significantly improves their accuracies. In this sense, if better CF methods would be available, they thus can be employed to enhance the recommendation accuracy of MARIO.

The experimental results can be summarized as follows: (1) MARIO consistently and significantly outperforms the state-of-the-art multimedia recommender systems; (2) our attention network helps to accurately capture each user's preferences; (3) our MP loss helps to preserve the intrinsic modality-specific properties of items; (4) by being equipped with any CF methods, MARIO significantly improves their original accuracy.

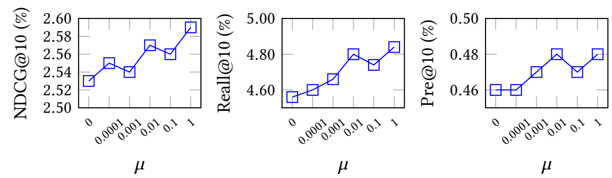


Figure 6: The effect of μ on the accuracies of MARIO. It is most effective when the value of μ is 1.

6 CONCLUSIONS

Observation. We pointed out that existing multimedia recommender systems face difficulties in (1) accurately capturing the individual influence of each modality at the interaction level and (2) effectively exploiting the intrinsic modality-specific properties of items. We demonstrated that learning procedures of existing works fail to preserve the intrinsic modality-specific properties of items.

Framework Design. We proposed an accurate multimedia recommendation framework MARIO based on modality-aware attention and modality-preserving decoders. Our framework is designed to be easily equipped with any CF methods that exploit only interaction information so that they can utilize items' multimodal features.

Experiments. We demonstrated that MARIO consistently and significantly outperforms its seven competitors on four real-life datasets. Moreover, we validated that MARIO better preserves the intrinsic properties of item modalities, compared to the state-of-the-art multimedia recommender systems.

ACKNOWLEDGMENT

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00352, No. RS-2022-00155586, and No. 2020-0-01373, Artificial Intelligence Graduate School Program (Hanyang University)). Also, we thank the NHN Cloud Corporation for their support of providing the cloud computing environment, which helped us greatly in performing this research successfully.

REFERENCES

[1] Yang Bao, Hui Fang, and Jie Zhang. 2014. Topicmf: Simultaneously exploiting ratings and reviews for recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 28. 2–8.

[2] Dong-Kyu Chae, Jin-Soo Kang, Sang-Wook Kim, and Jaeho Choi. 2019. Rating augmentation with generative adversarial networks towards accurate collaborative filtering. In *The World Wide Web Conference (WWW)*. 2616–2622.

[3] Dong-Kyu Chae, Jihoo Kim, Duen Horng Chau, and Sang-Wook Kim. 2020. AR-CF: Augmenting Virtual Users and Items in Collaborative Filtering for Addressing Cold-Start Problems. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR)*. 1251–1260.

[4] Lei Chen, Le Wu, Kun Zhang, Richang Hong, and Meng Wang. 2021. Set2setRank: Collaborative Set to Set Ranking for Implicit Feedback Based Recommendation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR)*. 585–594.

[5] 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 *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR)*. 765–774.

[6] Rami Cohen, Oren Sar Shalom, Dietmar Jannach, and Amihood Amir. 2021. A Black-Box Attack Model for Visually-Aware Recommender Systems. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*. 94–102.

[7] Ruining He and Julian McAuley. 2016. VBPR: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 30. 144–150.

[8] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR)*. 639–648.

[9] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[10] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*. 675–678.

[11] Dong Hyun Kim, Chanyoung Park, Jinoh Oh, Sungyoung Lee, and Hwanjo Yu. 2016. Convolutional Matrix Factorization for Document Context-Aware Recommendation. In *Proceedings of the ACM Conference on Recommender Systems (ACM RecSys)*. 233–240.

[12] Minseok Kim, Hwanjun Song, Doyoung Kim, Kijung Shin, and Jae-Gil Lee. 2021. PREMERE: Meta-Reweighting via Self-Ensembling for Point-of-Interest Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 35. 4164–4171.

[13] Minseok Kim, Hwanjun Song, Yooju Shin, Dongmin Park, Kijung Shin, and Jae-Gil Lee. 2022. Meta-Learning for Online Update of Recommender Systems. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 36. 4065–4074.

[14] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[15] Taeyong Kong, Taeri Kim, Jinsung Jeon, Jeongwhan Choi, Yeon-Chang Lee, Noseong Park, and Sang-Wook Kim. 2022. Linear, or Non-Linear, That is the Question!. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (ACM WSDM)*. 517–525.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the Annual Conference. on Neural Information Processing Systems (NeurIPS)*. 1106–1114.

[17] Hongjun Lim, Yeon-Chang Lee, Jin-Seo Lee, Sanggyu Han, Seunghyeon Kim, Yeongjong Jeong, Changbong Kim, Jaehun Kim, Sunghoon Han, Solbi Choi, et al. 2022. AiRS: A Large-Scale Recommender System at NAVER News. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*. 3386–3398.

[18] Fan Liu, Zhiyong Cheng, Changchang Sun, Yinglong Wang, Liqiang Nie, and Mohan Kankanhalli. 2019. User diverse preference modeling by multimodal attentive metric learning. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*. 1526–1534.

[19] Qiang Liu, Shu Wu, and Liang Wang. 2017. DeepStyle: Learning User Preferences for Visual Recommendation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR)*. 841–844.

[20] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3982–3992.

[21] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *Proceedigns of the International Conference on Uncertainty in Artificial Intelligence (UAI)*. 452–461.

[22] Yixin Su, Rui Zhang, Sarah M. Erfani, and Junhao Gan. 2021. Neural Graph Matching Based Collaborative Filtering. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR)*. 849–858.

[23] Suhang Wang, Yilin Wang, Jiliang Tang, Kai Shu, Suhas Ranganath, and Huan Liu. 2017. What Your Images Reveal: Exploiting Visual Contents for Point-of-Interest Recommendation. In *Proceedings of the International Conference on World Wide Web (WWW)*. 391–400.

[24] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural graph collaborative filtering. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR)*. 165–174.

[25] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2020. Graph-refined convolutional network for multimedia recommendation with implicit feedback. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*. 3541–3549.

[26] 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 *Proceedings of the ACM International Conference on Multimedia (ACM MM)*. 1437–1445.

[27] Liangwei Yang, Zhiwei Liu, Yingtong Dou, Jing Ma, and Philip S. Yu. 2021. ConsisRec: Enhancing GNN for Social Recommendation via Consistent Neighbor Aggregation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR)*. 2141–2145.

[28] Weilong Yao, Jing He, Hua Wang, Yanchun Zhang, and Jie Cao. 2015. Collaborative Topic Ranking: Leveraging Item Meta-Data for Sparsity Reduction. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 29. 374–380.

[29] Haochao Ying, Liang Chen, Yuwen Xiong, and Jian Wu. 2016. Collaborative Deep Ranking: A Hybrid Pair-Wise Recommendation Algorithm with Implicit Feedback. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*. 555–567.

[30] Junliang Yu, Hongzhi Yin, Jundong Li, Qinyong Wang, Nguyen Quoc Viet Hung, and Xiangliang Zhang. 2021. Self-Supervised Multi-Channel Hypergraph Convolutional Network for Social Recommendation. In *Proceedings of the International Conference on World Wide Web (WWW)*. 413–424.

[31] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. 2021. Mining Latent Structures for Multimedia Recommendation. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*. 3872–3880.

[32] Yongfeng Zhang, Qingyao Ai, Xu Chen, and W Bruce Croft. 2017. Joint representation learning for top-n recommendation with heterogeneous information sources. In *Proceedings of the ACM on Conference on Information and Knowledge Management (ACM CIKM)*. 1449–1458.