RESEARCH-ARTICLE

MELON: Learning Multi-Aspect Modality Preferences for Accurate Multimedia Recommendation

DONGHO JEONG, Hanyang University, Seoul, South Korea

TAERI KIM, Hanyang University, Seoul, South Korea

DONGHYEON CHO, Hanyang University, Seoul, South Korea

SANG-WOOK KIM, Hanyang University, Seoul, South Korea

Open Access Support provided by:

Hanyang University

Published: 13 July 2025

SIGIR '25: The 48th International ACM SIGIR Conference on Research and Development in Information Retrieval
July 13 - 18, 2025
Padua, Italy

Conference Sponsors:
SIGIR

https://doi.org/10.1145/3726302.3730031
ISBN: 9798400715921

MELON: Learning Multi-Aspect Modality Preferences for Accurate Multimedia Recommendation

Dongho Jeong
Hanyang University
Seoul, Korea
mars9954@hanyang.ac.kr

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

Donghyun Cho
Hanyang University
Seoul, Korea
doncho@hanyang.ac.kr

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

Abstract

Existing multimedia recommender systems have made the best efforts to predict user preferences for items by utilizing behavioral similarities between users and the modality features of items a user has interacted with. However, we identify two key limitations in existing methods regarding preferences for modality features: (L1) although preferences for modality features is an important aspect of users' preferences, existing methods *only* leverage neighbors with similar interactions and *do not consider the neighbors* who may have *similar preferences for modality features* while having *different interactions*; (L2) although modality features of a user and an item may have a *complex geometric relationship* in the latent space, existing methods *overlook and face challenges* in *precisely* capturing this relationship. To address these two limitations, we propose a novel multimedia recommendation framework, named **MELON**, which is based on two core ideas: (Idea 1) **M**odality-c**E**ntered embedding extraction; (Idea 2) re**L**ati**O**nship-ce**N**tered embedding extraction. We validate the effectiveness and validity of MELON through extensive experiments with four real-world datasets, showing 10.51% higher accuracy compared to the best competitor in terms of recall@10. The code and dataset of MELON is available at https://github.com/Bigdasgit/MELON.

CCS Concepts

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

Keywords

multimedia recommendation, modality-based neighbor, multi-aspect modality preferences

ACM Reference Format:
Dongho Jeong, Taeri Kim, Donghyun Cho, and Sang-Wook Kim. 2025. MELON: Learning Multi-Aspect Modality Preferences for Accurate Multimedia Recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3726302.3730031

*Corresponding author.

SIGIR '25, July 13–18, 2025, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730031

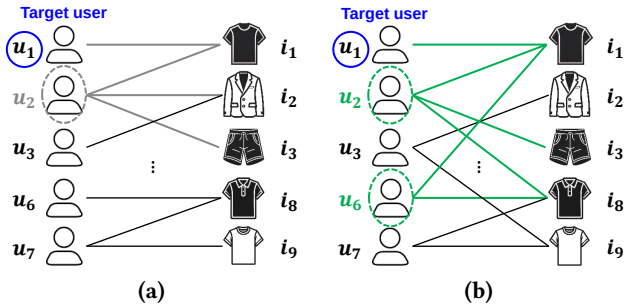

Figure 1: Examples of bipartite graphs between users and items: (a) A *interaction-centered* user–item bipartite graph; (b) A *modality-centered* user–item bipartite graph w.r.t. visual modality. These two bipartite graphs help identify different sets of neighbors (dotted-line ovals) for each user.

1 Introduction

Collaborative Filtering (CF), a popular technique in recommender systems, suggests such items that a user is likely to prefer by leveraging her *user–item interactions* (*e.g.*, clicks). CF methods identify a group of users (*i.e.*, *neighbors*) who have *similar user–item interactions* with a target user and recommends to her those items that her neighbors have interacted with but she has not seen yet [1, 6, 7, 13, 15, 19, 23]. However, due to the *sparsity* of interactions, CF methods face limitations in providing accurate recommendations [1, 10, 11, 24, 27, 28].

Multimedia recommender systems alleviate the *sparsity issue* by utilizing not only interactions but also *multimodal features* (*e.g.*, visual and textual modality features) of items [2, 4, 10, 11, 26, 29]; multimodal features are extracted from different types of item data (*e.g.*, images and descriptions) by using pre-trained deep-learning models such as convolutional neural networks [14] or recurrent neural networks [8].[1] Since the modality features of the items that a user has interacted with reveal her inherent *preference for modality features* [11], we refer to these modality features as '*modality features of user*' for convenience. To summarize, multimedia recommender systems provide more accurately the items that a user is likely to prefer by considering both important aspects of preferences: *interactions* and *modality features*.

Most existing methods perform the following steps in common [2, 4, 10, 11, 24, 26, 29]. **(Step 1)** They construct *a user–item interaction bipartite graph* using *the interaction data*. **(Step 2)** They *randomly* set the initial embedding of a *user* and set the initial embedding of an item as the compressed representation of its

[1] In general, multimedia recommender systems do not focus on how to extract the multimodal features of items [2, 4, 5, 10, 11, 24, 25, 29–32].

(given) modality features, obtained by using a multi-layer perceptron (MLP) [22]. **(Step 3)** They obtain the final user (resp. item) embedding by aggregating the embeddings of items (resp. users) *the user (resp. item) has interacted with* by applying graph convolutional networks (GCN) [12] to the user–item interaction bipartite graph. This process helps improve accuracy by explicitly exploiting a user's relationships with her neighbors inherent in the *high-order connectivity* of the interaction graph, also incorporating the modality features into the final user and item embeddings [11]. **(Step 4)** They predict a *user*'s preference for an *item* by calculating the *dot-product* of their final embeddings. In summary, most existing methods use an **interaction-centered** user–item bipartite graph (the interaction-centered graph, in short) to obtain the final embeddings of users and items. We refer to these embeddings as *interaction-centered user and item embeddings*.

In this paper, we discover two important challenges about preferences in terms of modality features, which was overlooked in earlier work; then propose two novel ideas to address them.

(Idea 1) Although most existing methods have identified and utilized neighbors with similar interactions, which is one of the key aspects of user preferences, they have *not* discovered and leveraged neighbors with *similar preferences for modality features*. However, based on a user's preference for modality features, we can identify a *new set of neighbors* whose preferences for modality features are similar, which may not have appeared in the interaction-centered graph. For instance, in Fig. 1-(a), u_1 and u_6 share the same preferences for visual modality features (*i.e.*, dark), even though they share no interacted items. However, as shown in Fig. 1-(b), u_1 can have a *new neighbor relationship* w.r.t. modality features with u_6. If we mine such a new set of neighbors, we can leverage the relationships with the neighbors for recommendations. To this end, we propose a new method of constructing and leveraging a **modality-centered** user–item bipartite graph (the modality-centered graph, in short), where two disjoint node sets represent users and items, with edges indicating high modality feature similarities between user and item nodes. In Section 3, we validate whether this new set of neighbors mined from modality features include indeed useful neighbors that have not been revealed in the interaction-centered graph. Then, we obtain user and item embeddings from the modality-centered graph and use them to perform recommendations alongside the interaction-centered embeddings. We refer to these new user and item embeddings as *modality-centered user and item embeddings*.

(Idea 2) Most existing methods [4, 5, 10, 11, 24–26] obtain a user's preference for an item by simply computing the *dot-product* of their interaction-centered embeddings in the latent space. However, in general CF, the limitation of using a simple dot-product to estimate user–item interactions was discussed in [7]. Inspired by their findings, we claim that a user's embedding representing her preference for modality features and an item's embedding for modality features may have a *complex geometric relationship* in the latent space, and that such a relationship is difficult to capture just with a simple dot-product operation. To address this issue, we aim to accurately uncover the complex relationships between the modality features of users and items, precisely capturing the preference hidden in this complex relationship *from her standpoint* for accurate recommendations. To this end, we propose a new idea of precisely extracting the complex relationship between a user and an item from their modality features by using neural networks, and learning the user's preference for an item from this relationship. We refer to the embedding precisely representing the complex relationship between a user and an item and the embedding representing the characteristics of her standpoint as *relational embedding* and *user-standpoint embedding*, respectively.

In this paper, we aim to recommend items accurately by capturing user's multi-aspect modality preferences via incorporating both (Idea 1) and (Idea 2). To this end, we propose a novel multimedia recommendation framework called, **MELON** (**M**odality-c**E**ntered embedding extraction and re**L**ati**O**nship-ce**N**tered embedding extraction), consisting of an existing multimedia recommendation *backbone* for interaction-centered user and item embeddings, along with two our core modules: the Modality-Centered Embedding extraction (MCE) module explicitly captures the relationships with a new set of neighbors in terms of modality features by using a newly proposed graph, based on (Idea 1), and the Relationship-Centered Embedding extraction (RCE) module provides auxiliary clues for the user's preference by mining the complex relationship between the modality features of a user and an item, based on (Idea 2).

Our contributions are summarized as follows:

- **Important Discovery**: We discover that existing methods (1) *focus only on* utilizing neighbors with similar interactions, *overlooking* those with similar preferences for modality features and (2) *have difficulty in precisely capturing* the complex relationships between the modality features of users and items.
- **Novel Framework**: To address these limitations, we propose MELON, a multimedia recommendation framework that captures *multi-aspect* modality preferences of users.
  - Modality-centered embedding extraction (MCE), which captures the user preferences revealed from the relationships with a new set of neighbors in terms of modality features.
  - Relationship-centered embedding extraction (RCE), which considers carefully the complex relationship between modality features of a user and an item, capturing a user's preference in such a relationship *from her standpoint*.
- **Extensive Evaluation**: We validate the two proposed ideas through comprehensive experiments on four real-world datasets.
  - Most importantly, MELON *consistently* and *significantly* outperforms *all* nine state-of-the-art competitors, with up to a 10.51% improvement in recall@10 over the *best competitor*.
  - Even when a single module (either MCE or RCE) is employed in the backbone, it still outperforms the *best competitor*.

2 Related Work

In this section, we briefly review recent multimedia recommendation methods and point out their limitations. Inspired by general CF methods based on GCN that explicitly capture the collaborative signals, which are similar interactions between users (resp. items), recent multimedia recommendation methods employing GCN have emerged to incorporate both collaborative signals and modality features [2, 4, 11, 18, 24–26, 29–33]. Such methods mainly leverage the following two types of graphs for accurate recommendations: (1) an interaction-centered graph, constructed using the interaction data; (2) a semantic graph, mined from latent semantic relationships using modality features. They either utilize only

the interaction-centered graph [2, 11, 24–26] or leverage both the interaction-centered graph and the semantic graph [4, 18, 29–33].

Methods Exploiting Only an Interaction-centered Graph. Basically, they are typical methods which generate interaction-centered user and item embeddings using the original interaction-centered graph as mentioned in Section 1. However, some of these methods refine the interaction-centered graph based on users' modality preferences [24, 25] to obtain better user and item embeddings. MMGCN [26] employs GCN to infer a user's preference for items by incorporating collaborative signals and the modality preferences of her neighbors who have similar interactions. MONET [11] proposes a novel GCN that effectively incorporate *both modality features and collaborative signals* in the final user and item embeddings. GRCN [25] refines an interaction-centered graph by pruning noisy edges by referring to each user's modality-specific preferences. MMSSL [24] considers the influence of each user's modality-specific preferences on their interaction preferences and captures the inter-dependencies among users' modality-specific preferences.

Methods Exploiting Both Interaction-centered and Semantic Graphs. Additionally, some GCN-based multimedia recommendation methods have emerged, aiming to uncover and utilize a semantic graph, which is an item-item (or user-user) graph representing semantic relationships between nodes (*i.e.*, user or items) [4, 18, 29–33], for supplementing collaborative signals. LATTICE [30] captures semantic relationships between items derived from their modality features to supplement the collaborative signals of general CF models. LGMRec [4] leverages relationships between items (resp. users) based on similarities in latent attributes from their modality features to model global user interest.

Discussions. The graphs utilized in aforementioned methods have a common limitation in considering the *users with similar modality preferences*, making it *difficult to directly consider* the items that these new neighbors in terms of modality preferences are likely to prefer. In addition, all existing methods infer a user's preference for an item by simply computing the *dot-product* of their final embeddings. Thus, they have a limitation in *accurately modeling* the complex relationship between the modality features of her and an item, missing a clue for her preference. These two limitations can lead to an insufficient understanding of a user's preference, causing low accuracy.

3 Motivation

In this section, we validate (Idea 1) via preliminary experiments answering the following preliminary questions (PQs):

- **PQ1**: Do modality-based neighbors differ from interaction-based neighbors?
- **PQ2**: Do the preferences of modality-based neighbors for modality features appear more similarly compared to those of interaction-based neighbors?

We validate (Idea 2) in Section 5 by showing the framework *with* (Idea 2) outperforms significantly the framework *without* (Idea 2), since it is difficult to validate (Idea 2) via preliminary experiments.

Experimental Settings. First, we summarize the key notations used for the following sections: $\mathcal{U}$ and $\mathcal{I}$ denote the sets of whole users and whole items, respectively; Note that '*neighbors*' refer to a group of users with similar preferences to a user in either modality or interaction. Nei_u^m and $Nei_u^i \subset \mathcal{U}$ indicate *modality-based* and *interaction-based* neighbors, respectively, for each user u.

For these preliminary experiments, we use four categories of the real-world Amazon dataset [17] (*i.e.*, Women Clothing, Men Clothing, Toys & Games, and Sports), which are widely used in multimedia recommendation studies [3–5, 10, 11, 16, 24, 30], containing both visual and textual modality features of each item. Due to space limitations, we only report results on Women Clothing in this section.[2] Following [4, 11, 24, 26, 30], we pre-compute the modality features $\tilde{\mathbf{e}}_i^m$ for each item i by using a pre-trained deep-learning model, specifically employing ImageNet [14] for visual modality v and Sentence-BERT [20] for textual modality t, where m denotes a modality indicator.

PQ1: Analysis of Difference Between Nei_u^m and Nei_u^i. To validate (Idea 1), we first construct two sets of neighbors (*i.e.*, Nei_u^m and Nei_u^i) and analyze whether those two sets of neighbors are truly different by comparing the degree of overlap between them.

Modality-based Neighbor Construction. Recall that a user's preference for modality features is inferred from the modality features of the items she has interacted with, as we mentioned in Section 1. Thus, we generate user u's modality feature $\tilde{\mathbf{e}}_u$ to represent her modality preference by averaging the modality features $\tilde{\mathbf{e}}_i^m$ of the items she has interacted with for each modality m, following [4, 24], and then concatenating them over $\forall m \in \{v, t\}$. To construct modality-based neighbors, Nei_u^m, for each user u, we calculate the cosine similarity between the features (*i.e.*, $\tilde{\mathbf{e}}_u$) of users. Then, for each u, we select the top-k users with the highest similarity to hers as her modality-based neighbors, Nei_u^m.

Interaction-based Neighbor Construction. For interaction-based neighbors, we use two methods to construct them and compare the degree of their overlap with Nei_u^m in each case. For each u, the first method represents her interactions as a *multi-hot vector* with dimensionality $|\mathcal{I}|$; it selects the top-k other users with the most similar multi-hot vectors to hers by using the Jaccard similarity [9] as her interaction-based neighbors, Nei_u^i.[3]

The second method identifies neighbors by using the interaction-centered user embeddings, $\mathbf{e}_u^{ICE}$, *learned through an existing multimedia recommendation method* which utilizes only the interaction-centered graph, we mentioned in Section 2, with the modality features used as initial embeddings. We employ this method to show that, even though modality features are incorporated, a model based on the interaction-centered graph has limitations in *effectively* capturing modality-based neighbors. To this end, we obtain $\mathbf{e}_u^{ICE}$ for each u by employing MONET [11], a state-of-the-art multimedia recommendation method. Then, we construct Nei_u^{ICE} by selecting the top-k users the most similar to each u in terms of the cosine similarity between the embeddings obtained from MONET.

Overlap Ratio Computation. Finally, we calculate the overlap ratio between two sets Nei_u^m and Nei_u^i and that between Nei_u^m and Nei_u^{ICE} for each u by using the Jaccard similarity. Subsequently, we compute the average of the overlap ratio across *all* users to determine the overall degree of overlap between the two sets of

[2] The results for other categories in the Amazon dataset showed in https://sites.google.com/hanyang.ac.kr/melon-sigir2025, showing similar tendencies to those observed for Women Clothing.

[3] Since every user in Nei_u^i shares a different number of interacted items with user u, which may result in $|Nei_u^i|$ being smaller than k in certain users.

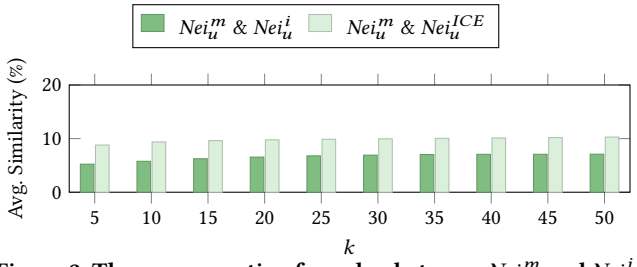

Figure 2: The average ratio of overlap between Nei_u^m and Nei_u^i (resp. Nei_u^{ICE}) according to the number of neighbors, k.

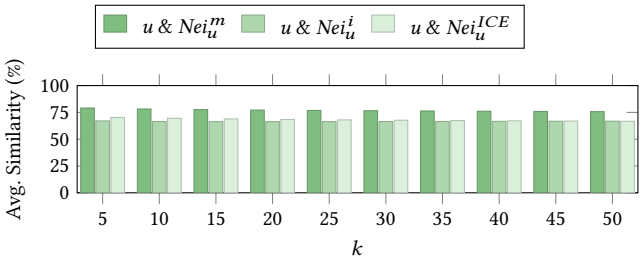

Figure 3: The average similarity of the preferences for modality features between a user and each of her Nei_u^m, Nei_u^i, and Nei_u^{ICE} according to the number of neighbors, k.

neighbors. This process is repeated, increasing the number of neighbors, k, in step of 5 until it reaches 50.

Results and Discussions. Fig. 2 shows the average ratio of overlap between Nei_u^m and Nei_u^i (resp. Nei_u^{ICE}). The x-axis indicates k, the number of neighbors per u and the y-axis does the average overlap ratio. The dark and light green bars show the results of comparing Nei_u^m with Nei_u^i and Nei_u^{ICE}, respectively. Based on these results, we make the following observations.

i) First, we observe that the overlap ratio between Nei_u^m and Nei_u^i is generally *quite low*. This demonstrates that user preferences for modality features quite differ from those for interactions.
ii) The overlap ratio between Nei_u^m and Nei_u^{ICE} is quite low as well; the overlap ratio is slightly higher than that between Nei_u^m and Nei_u^i, which is consistently observed for all k.

From the observations, we have confirmed that modality-based neighbors are *fairly different* from interaction-based neighbors.

PQ2: Analysis of Each Neighbor Group's Preferences for Modality Features. We further aim to demonstrate that a user's modality-based neighbors are more similar to her *in terms of preferences for modality features* than her neighbors in the other two groups, emphasizing the benefit of the modality-based neighbors for recommendation. To this end, we calculate and compare the average similarity of the modality features of user which represent the preference of modality features between a user and those of her neighbors in each of the three groups, Nei_u^m, Nei_u^i, Nei_u^{ICE}, as follows. First, we compute the cosine similarity between the modality features $\tilde{\mathbf{e}}_u$ of a user and those of her neighbors, and then take the average of these similarities. Then, we further calculate the overall similarity of modality features by averaging the similarities across all users. This process is repeated, increasing the number of neighbors, k, in step of 5, until 50.

Results and Discussions. Fig. 3 shows the results. The x-axis represents the number of neighbors k, while the y-axis does the average similarity of modality features. The dark, medium, and light green bars represent the modality feature similarity between a user and her neighbors in Nei_u^m, Nei_u^i, and Nei_u^{ICE}, respectively. Based on these results, we make the following observations.

i) First, the average similarity of modality features between a user and her neighbors in Nei_u^m is consistently higher than that between a user and her neighbors in the other two neighbor groups, Nei_u^i and Nei_u^{ICE}, specifically, up to 12% and 9%, respectively.
ii) The average similarity between a user and her neighbors in Nei_u^i and Nei_u^{ICE} is generally the lowest and second lowest.

Based on these observations, we have confirmed that a user's modality-based neighbors are most similar to her in terms of preferences for modality features and thus being expected useful for recommendations.

Summary. We thus claim that leveraging additional modality-based neighbors could be *beneficial* in improving the accuracy.

Table 1: Key notations used in this paper

Notation	Description
v, t	Visual and textual modalities
$\tilde{\mathbf{e}}_u^m, \tilde{\mathbf{e}}_i^m$	Modality m's features of user u and item i
$\mathbf{e}_u, \mathbf{e}_i$	Interaction-centered embeddings of user u and item i
$\bar{\mathbf{e}}_u, \bar{\mathbf{e}}_i$	Modality-centered embeddings of user u and item i
$\hat{\mathbf{e}}_{u,i}$	Relational embedding between user u and item i
$\hat{\mathbf{e}}_u$	User-standpoint embedding of user u
Nei_u^m	Modality-based neighbors of user u for modality m
$\hat{y}_{u,c}^{ICE}$	Matching score between interaction-centered embeddings of user u and target item c
$\hat{y}_{u,c}^{MCE}$	Matching score between modality-centered embeddings of user u and target item c
$\hat{y}_{u,c}^{RCE}$	Matching score between relational embedding $\hat{\mathbf{e}}_{u,c}$ and user-standpoint embedding $\hat{\mathbf{e}}_u$
$\hat{y}_{u,c}$	Preference of user u for target item c
α, β	The hyperparameters which control the weights of $\hat{y}_{u,c}^{MCE}$ and $\hat{y}_{u,c}^{RCE}$ in $\hat{y}_{u,c}$, respectively

4 MELON: Proposed Framework

In this section, we define the multimedia recommendation problem and explain the proposed framework, MELON, which is composed of three key encoder modules.

4.1 Problem Definition

Let $\mathcal{U}$ and $\mathcal{I}$ denote the complete sets of users and items, respectively. Also, $\mathcal{N}_u \subset \mathcal{I}$ represents the set of items interacted with by a user $u \in \mathcal{U}$, while $\mathcal{N}_i \subset \mathcal{U}$ denotes the set of users who have interacted with a item $i \in \mathcal{I}$. In this paper, we assume that each item i has modality features $\tilde{\mathbf{e}}_i^m \in \mathbb{R}^{d_m}$ pre-computed by using pre-trained deep-learning models for visual and textual modalities v and t as described in [4, 10, 11, 29–31], where $m \in \mathcal{M} = \{v, t\}$ and $\mathcal{M}$ is the set of modalities, and d_m denotes dimensionality of features w.r.t. modality m. The goal of multimedia recommendation is to identify the top-N items that a user u is most likely to prefer among her non-interacted items (*i.e.*, $c \in \mathcal{I} \backslash \mathcal{N}_u$) by utilizing not only user–item interactions but also the multimodal features of items. Table 1 summarizes the key notations used in this paper.

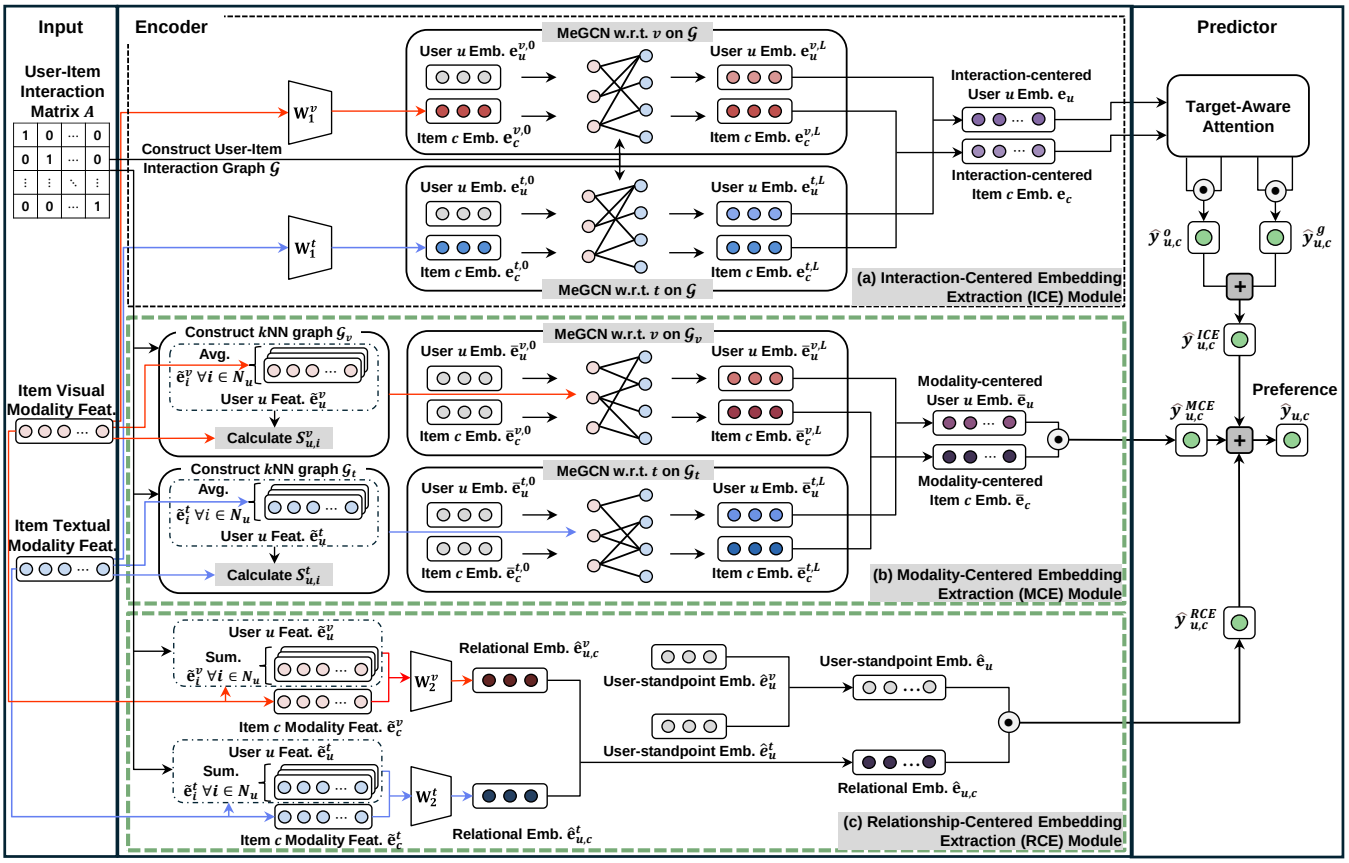

Figure 4: Overview of MELON composed of two key components: An encoder based on (a) the ICE module, (b) the MCE module, (c) the RCE module and a predictor utilizing all the preferences across multi-aspects.

4.2 Key Components

Fig. 4 shows the schematic overview of MELON. Note that MELON consists of the two key components: an *encoder* and a *predictor*. In the encoder, MELON generates the embeddings carrying a user's multi-aspect modality preferences: (i) the interaction-centered embedding extraction (ICE) module aims to capture a user's interaction-centered preference (Fig. 4-(a)); (ii) the modality-centered embedding extraction (MCE) module obtains a user's modality-centered preference by leveraging the relationships with her modality-based neighbors (Fig. 4-(b)); (iii) the relationship-centered embedding extraction (RCE) module considers the complex geometric relationship between a user and an item from her standpoint (Fig. 4-(c)). In the predictor, MELON computes a user's final preference for the target item by using all embeddings obtained from the encoder (Fig. 4-predictor). Lastly, MELON trains all learnable parameters by using the Bayesian Personalized Ranking (BPR) loss [4, 11, 21, 24, 29–31].

Encoder. The ICE, MCE, and RCE modules utilize multimodal features in *different ways* to accurately capture the user's preferences.

Interaction-Centered Embedding Extraction (ICE) Module. Multimedia recommendation methods exploiting only interaction-centered graph have provided accurate recommendations by leveraging interaction-centered embeddings which consider both user u's Nei_u^i and users' modality preferences. Therefore, we obtain and utilize interaction-centered embeddings for recommendations by employing MeGCN of MONET [11] on the ICE module, which is the state-of-the-art multimedia recommendation method. Using a user–item interaction matrix $A \in \mathbb{R}^{|\mathcal{U}|\times|\mathcal{I}|}$, MELON constructs a user–item interaction bipartite graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = (\mathcal{U} \cup \mathcal{I})$ and $\mathcal{E} = \{(u, i) \in \mathcal{U}, i \in \mathcal{N}_u\}$ denote a set of nodes and edges, respectively. Given a user–item interaction bipartite graph $\mathcal{G}$, MELON *randomly* initializes interaction-centered user embedding $\mathbf{e}_u^{m,0} \in \mathbb{R}^d$ differently w.r.t. modality m. Also, following [4, 11, 24, 26], MELON initializes interaction-centered item embedding $\mathbf{e}_i^{m,0} \in \mathbb{R}^d$ by compressing item i's modality feature $\tilde{\mathbf{e}}_i^m$ into a low-dimensional embedding via MLP as follows: $\forall m \in \{v, t\}$,

$$\mathbf{e}_i^{m,0} = \mathbf{W}_1^m \tilde{\mathbf{e}}_i^m + \mathbf{b}_1^m, \tag{1}$$

where $\mathbf{W}_1^m \in \mathbb{R}^{d \times d_m}$ and $\mathbf{b}_1^m \in \mathbb{R}^d$ denote a trainable weight matrix and a bias vector, respectively.

Then, MELON independently applies MeGCN to $\mathcal{G}$ for each modality. Formally, interaction-centered embeddings $\mathbf{e}_u^{m,l}$ and $\mathbf{e}_i^{m,l}$ for modality m via the l-th MeGCN layer ($1 \le l \le L$, where L is the number of layers) can be expressed as follows: $\forall m \in \{v, t\}$,

$$\mathbf{e}_u^{m,l} = \sum_{i \in \mathcal{N}_u} \frac{\mathbf{e}_i^{m,l-1}}{\sqrt{|\mathcal{N}_u||\mathcal{N}_i|}} + \mathbf{e}_u^{m,l-1}, \mathbf{e}_i^{m,l} = \sum_{u \in \mathcal{N}_i} \frac{\mathbf{e}_u^{m,l-1}}{\sqrt{|\mathcal{N}_u||\mathcal{N}_i|}} + \mathbf{e}_i^{m,l-1}. \tag{2}$$

Lastly, MELON obtains interaction-centered user (resp. item) embedding $\mathbf{e}_u$ (resp. $\mathbf{e}_i$) by concatenating user (resp. item) embeddings $\mathbf{e}_u^{m,L}$ (resp. $\mathbf{e}_i^{m,L}$) from the L-th MeGCN layer for $\forall m \in \{v, t\}$ [11].

Modality-Centered Embedding Extraction (MCE) Module. MELON constructs a modality-centered user–item bipartite graph $\mathcal{G}_m$ and captures the relationships with Nei_u^m from $\mathcal{G}_m$ in the MCE module. By leveraging the modality-centered graph, MELON can recommend to her the items with modality features that a user's Nei_u^m prefer. Note that, since a user's preferences for different modalities may differ [10], her modality-based neighbors can also differ across modalities. Therefore we choose to construct modality-centered graph $\mathcal{G}_m$ for each modality m to capture the relationships with various neighbors in Nei_u^m over $\forall m \in \{v, t\}$. The effectiveness of the MCE module will be empirically validated in Section 5.

MELON constructs a modality-centered graph $\mathcal{G}_m$ which is the k-nearest-neighbor (kNN) graph based on similarities between the modality features of users and items. Due to the *absence of explicit modality features of user* in datasets as in [4, 24], MELON computes the modality features of a user u, $\tilde{\mathbf{e}}_u^m$ for modality m as follows:

$$\tilde{\mathbf{e}}_u^m = \sum_{i \in \mathcal{N}_u} \frac{1}{|\mathcal{N}_u|} \tilde{\mathbf{e}}_i^m. \tag{3}$$

After that, MELON calculates cosine similarities $S_{u,i}^m$ between the modality features of a user u $\tilde{\mathbf{e}}_u^m$ and an item i $\tilde{\mathbf{e}}_i^m$, $\forall u \in \mathcal{U}$, $\forall i \in \mathcal{I}$, for modality m. Based on $S_{u,i}^m$, MELON constructs modality-centered graph $\mathcal{G}_m$ as follows: $\forall m \in \{v, t\}$,

$$A^m = \begin{cases} 1, & \text{if } S_{u,i}^m \in \text{top-}k(S_u^m), \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where A^m is an adjacency matrix for a $\mathcal{G}_m$ w.r.t. modality m.[4]

Then, for each modality m, MELON applies MeGCN to $\mathcal{G}_m$ independently with the embeddings randomly initialized for users and items $\bar{\mathbf{e}}_u^{m,0}$, $\bar{\mathbf{e}}_i^{m,0} \in \mathbb{R}^d$. Finally, MELON generates modality-centered user and item embeddings, $\bar{\mathbf{e}}_u$ and $\bar{\mathbf{e}}_i$, by concatenating modality m's $\bar{\mathbf{e}}_u^{m,L}$ (resp. $\bar{\mathbf{e}}_i^{m,L}$) for $\forall m \in \{v, t\}$ as in the ICE module.

Relationship-Centered Embedding Extraction (RCE) Module. MELON aims to capture such a complex relationship between the modality features of a user and an item very carefully and exploit her preference revealed from the relationship for recommendation in the RCE module. The effectiveness and structural validity of the RCE module will be provided in Section 5.

First of all, MELON generates a relational embedding $\hat{\mathbf{e}}_{u,i}$ for each modality m which contains a complex relationship between the modality features of a user u and an item i. To this end, MELON constructs a relational modality feature $\tilde{\mathbf{e}}_{u,i}^m$ of a user u and an item i for $\forall m \in \{v, t\}$ which intuitively represents their complex relationship as mentioned in [7], by concatenating modality features of a user u $\tilde{\mathbf{e}}_u^m$ and an item i $\tilde{\mathbf{e}}_i^m$ (*i.e.*, $[\tilde{\mathbf{e}}_u^m || \tilde{\mathbf{e}}_i^m]$).

In this module, MELON computes the modality features of a user u, $\tilde{\mathbf{e}}_u^m$ for modality m, as follows [4, 24]:

$$\tilde{\mathbf{e}}_u^m = \sum_{i \in \mathcal{N}_u} \tilde{\mathbf{e}}_i^m. \tag{5}$$

[4] MELON pre-computes and stores A_m once, then loads it during the training.

To capture their complex relationship precisely from the raw information, MELON transforms $\tilde{\mathbf{e}}_{u,i}^m$ to relational embedding $\hat{\mathbf{e}}_{u,i}^m \in \mathbb{R}^d$ for modality m, which is expressed as follows: $\forall m \in \{v, t\}$,

$$\hat{\mathbf{e}}_{u,i}^m = \mathbf{W}_2^m \tilde{\mathbf{e}}_{u,i}^m + \mathbf{b}_2^m, \tag{6}$$

where $\mathbf{W}_2^m \in \mathbb{R}^{d \times 2d_m}$ and $\mathbf{b}_2^m \in \mathbb{R}^d$ denote a trainable weight matrix and a bias vector, respectively. After that, MELON obtains the final relational embeddings $\hat{\mathbf{e}}_{u,i}$ by concatenating modality m's relational embedding $\hat{\mathbf{e}}_{u,i}^m$ for $\forall m \in \{v, t\}$.

Finally, MELON concatenates randomly initialized user-standpoint embedding $\hat{\mathbf{e}}_u^m \in \mathbb{R}^d$ for $\forall m \in \{v, t\}$, to generate a user u's final user-standpoint embedding $\hat{\mathbf{e}}_u$ representing her preference on the above complex relationship. Like other learnable parameters, relational embedding $\hat{\mathbf{e}}_{u,i}$ and user-standpoint embedding $\hat{e}_u$ are also trained in a supervised learning paradigm by using BPR loss [21]; the details about learning $\hat{\mathbf{e}}_u$ and $\hat{\mathbf{e}}_{u,i}$ will be described in the next subsection.

Predictor. MELON predicts a final preference $\hat{y}_{u,c}$ for a user u on a target item $c \in \mathcal{I} \setminus \mathcal{N}_u$ by incorporating the multi-aspect preferences. First, MELON computes $\hat{y}_{u,c}^{ICE}$, which is the preference of a user u for a target item c by exploiting interaction-centered embeddings obtained from the relationships with interaction-based neighbors. Specifically, we consider both u's general interest $\hat{y}_{u,c}^g(\mathbf{e}_u{}^\top \mathbf{e}_c)$ and her target-specific interest $\hat{y}_{u,c}^o$ $(\mathbf{e}_u^c{}^\top \mathbf{e}_c)$ through target-aware attention [11] to capture two beneficial aspects of user u's interaction-centered preference as follows:

$$\hat{y}_{u,c}^{ICE} = (1 - \gamma)\hat{y}_{u,c}^g + \gamma \hat{y}_{u,c}^o, \tag{7}$$

where the target-oriented user embedding of u for a target item c $\mathbf{e}_u^c$ is computed by $\sum_{i \in \mathcal{N}_u} a_u^{c,i} \times \mathbf{e}_i$, $a_u^{c,i} = \frac{\exp(\mathbf{e}_c^\top \mathbf{e}_i)}{\sum_{j \in \mathcal{N}_u} \exp(\mathbf{e}_c^\top \mathbf{e}_j)}$, and hyperparameter $\gamma \in (0, 1)$ denotes a coefficient which controls the balance between a general interest and a target-specific interest.

Also, MELON computes $\hat{y}_{u,c}^{MCE}$, which is the preference of a user u for a target item c by utilizing modality-centered embeddings considering for preferences revealed in the relationships with modality-based neighbors as follows:

$$\hat{y}_{u,c}^{MCE} = \bar{\mathbf{e}}_u^\top \bar{\mathbf{e}}_c. \tag{8}$$

Next, MELON computes $\hat{y}_{u,c}^{RCE}$, which is the preference of a user u for a target item c by leveraging the relational embedding and the user-standpoint embedding considering her preference on the complex relationship formed between modality features of a user and an item as follows:

$$\hat{y}_{u,c}^{RCE} = \hat{\mathbf{e}}_u^\top \hat{\mathbf{e}}_{u,c}. \tag{9}$$

Finally, MELON predicts user u's preference $\hat{y}_{u,c}$ for target item c by leveraging *all* the preferences, as follows:

$$\hat{y}_{u,c} = \hat{y}_{u,c}^{ICE} + \alpha \hat{y}_{u,c}^{MCE} + \beta \hat{y}_{u,c}^{RCE}, \tag{10}$$

where $\alpha \in (0, 1)$ and $\beta \in (0, 1)$ denote the coefficients controlling the weights of $\hat{y}_{u,c}^{MCE}$ and $\hat{y}_{u,c}^{RCE}$ in $\hat{y}_{u,c}$, respectively.

To train all parameters, MELON employs the BPR loss [21], widely adopted in multimedia recommender systems [4, 5, 10, 11, 24–26, 29–31]. The BPR loss learns all parameters with the intuition that a user u's preference for an item i which she has interacted

with is likely to be higher than u's preference for a *randomly selected* item j which she has not interacted with, as follows:

$$\mathcal{L} = -\sum_{u\in\mathcal{U}}\sum_{i\in\mathcal{N}_u}\sum_{j\in\mathcal{I}\setminus\mathcal{N}_u} \ln\sigma(\hat{y}_{u,i}-\hat{y}_{u,j}) + \lambda\|\Theta\|_2^2, \quad (11)$$

where $\sigma(\cdot)$, λ, and Θ denote a sigmoid function, a regularization weight, learnable parameters of MELON, respectively.

Via learning with the BPR loss, MELON can capture a complex relationship between the modality features of a user and an item on the relational embedding $\hat{\mathbf{e}}_{u,i}$ and a user's additional preference on the user-standpoint embedding $\hat{\mathbf{e}}_u$. User-standpoint embedding $\hat{\mathbf{e}}_u$ moves closer to the relevant relational embeddings $\hat{\mathbf{e}}_{u,i}$ in which item $\forall i \in \mathcal{N}_u$ is involved, while moving far away from irrelevant relational embeddings $\hat{\mathbf{e}}_{u,j}$ in which item $\forall j \in \mathcal{I}\setminus\mathcal{N}_u$ involved.

4.3 Analysis of Time & Space Complexity

We analyze the time and space complexities of MELON. We assume that the dimensionality of visual and textual modality feature is same as d_m, and the number of items $|\mathcal{N}_u|$ per user is the average of all user's, *i.e.*, $\frac{|\mathcal{E}|}{|\mathcal{U}|}$, for simplicity.

First, the time complexity of MELON is as follows:

- In the training phase, the computational costs for the ICE, MCE, and RCE modules are $O(|\mathcal{M}|d(3d_m+2|\mathcal{E}|(L+2)))$, $O(2|\mathcal{M}|d(|\mathcal{U}|kL+|\mathcal{E}|))$, $O(|\mathcal{M}|(3d_md+d+2|\mathcal{E}|))$, respectively. The difference between the cost of the ICE and the cost of the newly proposed two modules is expressed as $2|\mathcal{E}|d-(2|\mathcal{E}|+d)$ and is a positive value if both $|\mathcal{E}|$ and d are greater than two.
- In the inference phase, the computational cost of the ICE module is $O(|\mathcal{M}||\mathcal{U}||\mathcal{N}_u||\mathcal{I}|d+2|\mathcal{M}||\mathcal{U}||\mathcal{I}|d)$, while the combined cost of the MCE and RCE modules is $O(|\mathcal{M}||\mathcal{U}|(|\mathcal{I}|d_md+3|\mathcal{I}|d))$. The difference between the costs of these modules is expressed as $\mathcal{N}_u|\mathcal{I}|d-(|\mathcal{I}|d+|\mathcal{I}|+d_md)$, which is positive in most cases.

Next, the space complexity for the learnable parameters of each module in MELON for each modality m is as follows; (1) the ICE module has a complexity of $|\mathcal{M}|(|\mathcal{U}|+d_m)$; (2) the MCE module has a complexity of $|\mathcal{M}|(|\mathcal{U}|+|\mathcal{I}|)d$; (3) RCE module has a complexity of $|\mathcal{M}|(|\mathcal{U}|+2d_m)$.

5 Evaluation

In this section, we demonstrate the superiority of MELON by answering the following key research questions (RQs):

- **RQ1**: Is MELON more accurate in top-N recommendations than state-of-the-art recommender systems?
- **RQ2**: Are the two core ideas of MELON (the MCE and RCE modules) effective in improving accuracy?
- **RQ3**: Is the MCE module effective in enabling MELON to capture the relationships with modality-based neighbors?
- **RQ4**: How sensitive is the accuracy of MELON to the hyperparameters α and β?

5.1 Experimental Settings

Datasets. We conduct the experiments by using four real-world Amazon datasets [17], which are widely used in existing multimedia recommender systems [4, 10, 11, 24, 29, 30], as follows: Women Clothing, Men Clothing, Toys & Games, and Sports. Following [4, 10, 11, 24, 29–31], we only use such users and items that had at least five interactions, along with the item's visual and textual modality features, across all datasets. Some statistics of the four datasets are provided in Table 2.

Table 2: Dataset statistics

Dataset	# Users	# Items	# Interactions	Sparsity
Women Clothing	19,244	14,596	135,326	99.95%
Men Clothing	4,955	5,028	32,363	99.87%
Toys & Games	19,412	11,924	167,597	99.93%
Sports	35,598	18,357	256,308	99.61%

Competitors. To validate the effectiveness and validity of MELON, we compare MELON with nine state-of-the-art recommender methods, which can be divided into two groups: (1) GCN-based general CF recommender methods (LightGCN [6], LightGCL [1]) and (2) multimedia recommender methods (VBPR [5], MMGCN [26], GRCN [25], LATTICE [30], MMSSL [24], LGMRec [4], MONET [11]). We use only user–item interactions for (1) GCN-based CF methods, while utilizing both user–item interactions and each items' visual and textual features for (2) multimedia recommender methods.

Evaluation Protocols. Following existing studies, [4, 11, 24–26, 30], we randomly split user–item interactions for each dataset into training (80%), validation (10%), and test (10%) sets for evaluation. Also, we evaluate top-10 recommendation results obtained from each method, in terms of precision, recall, and normalized discounted cumulative gain (NDCG) [4, 11, 24–26, 30]. We report the average values over five independent evaluations, with the all p-values being below 0.05, indicating statistical significance.

Implementation Details. For fair comparisons, we set the dimensionality of embedding d to 64 for all methods [1, 4, 6, 10, 11, 24–26, 29–31] including MELON. Then, we fine-tune all hyperparameters of our competitors based on the values reported in their original papers by using a validation set. Lastly, we set hyperparameters of MELON as follows: learning rate = 0.0001; the number of GCN layer L = 2; the number of neighbors for kNN graph k = 5; the values of (α, β, γ) is set to (0.3, 0.6, 0.4), (1.0, 0.4, 0.5), (0.3, 0.2, 0.6), and (0.1, 0.001, 0.1) for Women Clothing, Men Clothing, Toys & Games, and Sports, respectively.

5.2 Results and Analysis

We report the results of all datasets for RQ1; however, we only present the results on Women Clothing for the remaining RQs due to space limitations.[5] For ease of comparisons, we report the values of precision@10 after multiplying by 10 for all methods and all datasets except for Toys & Games. In the following tables and figures, we use P@10 and R@10 as the abbreviations for precision@10 and recall@10, respectively. The *best* and *second-best* results in each column are highlighted with **bold** and underline, respectively.

RQ1: Comparison with Nine Competitors. We compare the accuracy of MELON and nine competitors in Table 3; our findings from the results are summarized as follows:

i) MELON *consistently* and *significantly* outperforms *all* nine competitors on *all* datasets for *all* metrics. Specifically, on Women

[5]The rest of the results generally show similar tendencies as those on Women Clothing and are provided in https://sites.google.com/hanyang.ac.kr/melon-sigir2025.

Table 3: Accuracies of nine competitors and MELON on four datasets. The improvements of MELON over the best competitors are all statistically significant with p-value ≤ 0.05.

Datasets	Women Clothing			Men Clothing			Toys & Games			Sports		
Measures	**P@10**	**R@10**	**NDCG@10**	**P@10**	**R@10**	**NDCG@10**	**P@10**	**R@10**	**NDCG@10**	**P@10**	**R@10**	**NDCG@10**
LightGCN (SIGIR'20)	0.0450	0.0441	0.0250	0.0360	0.0356	0.0192	0.0780	0.0748	0.0437	0.0596	0.0564	0.0318
LightGCL (ICLR'23)	0.0558	0.0550	0.0315	0.0398	0.0395	0.0221	0.0939	0.0886	0.0565	0.0691	0.0653	0.0374
VBPR (AAAI'16)	0.0345	0.0339	0.0182	0.0313	0.0312	0.0162	0.0583	0.0549	0.0314	0.0289	0.0277	0.0152
MMGCN (MM'19)	0.0270	0.0265	0.0141	0.0235	0.0233	0.0118	0.0423	0.0396	0.0219	0.0362	0.0340	0.0186
GRCN (MM'20)	0.0480	0.0474	0.0249	0.0344	0.0343	0.0177	0.0925	0.0863	0.0500	0.0607	0.0573	0.0322
LATTICE (MM'21)	0.0570	0.0564	0.0314	0.0503	0.0501	0.0274	0.0940	0.0894	0.0523	0.0611	0.0580	0.0318
MMSSL (WWW'23)	0.0610	0.0602	0.0330	0.0518	0.0514	0.0270	0.0971	0.0911	0.0523	0.0708	0.0671	0.0378
LGMRec (AAAI'24)	0.0575	0.0567	0.0310	0.0570	0.0563	0.0298	0.0993	0.0935	0.0517	0.0708	0.0674	0.0367
MONET (WSDM'24)	0.0678	0.0668	0.0363	0.0583	0.0580	0.0318	0.1116	0.1052	0.0618	0.0721	0.0688	0.0388
MELON	**0.0720**	**0.0710**	**0.0393**	**0.0643**	**0.0641**	**0.0349**	**0.1183**	**0.1114**	**0.0641**	**0.0744**	**0.0708**	**0.0399**
Improvements (%)	**6.21**	**6.26**	**8.19**	**10.29**	**10.51**	**9.76**	**5.99**	**5.94**	**3.79**	**3.18**	**2.96**	**2.98**

Clothing, Men Clothing, Toys & Games, and Sports, MELON shows the improvements by up to 6.26%, 10.51%, 5.94%, 2.96% over the *best competitors*, *i.e.*, MONET, in terms of R@10, respectively. Also, these are *considerable improvements* in the sense that MELON overall outperforms our best competitor which provided a significant improvement over its best competitors by an average of 22.85% in terms of R@20 in the original paper [11].

ii) Multimedia recommenders tend to outperform GCN-based general CF methods. Specifically, MELON outperforms GCN-based general CF state-the-of-art, LightGCL, by up to 62.28% in terms of R@10. This is the result of capturing multi-aspect preferences effectively in MELON by utilizing both interactions and modality features. Similarly, incorporating multimodal features to capture additional aspects of user preferences improves accuracy, as multimedia recommenders with LightGCN (*i.e.*, MMSSL, LATTICE) outperform that with LightGCN alone [4].

RQ2: Ablation Study. To validate the effectiveness of the MCE and RCE modules, which correspond to the two core ideas of MELON, we compare MELON with its variants.

MELON Variants w.r.t. MCE Module. (1) MELON–MCE indicates the variant that does not employ the MCE module within MELON. (2) MCE_k indicates the variant that employ $\mathcal{G}_m$ constructed with k = 5, 10, 15, 20. (3) $\text{MCE}_{|\mathcal{N}_u|}$ indicates the variant that employs modified $\mathcal{G}_m$ where a user u is connected with $|\mathcal{N}_u|$-items. (4) $\text{MCE}_{\text{i-i}}$ indicates the variant that employs a *latent* item-item graph as used in the state-of-the-art multimedia recommendation methods [30, 31], instead of our $\mathcal{G}_m$.

MELON Variants w.r.t. RCE Module. (1) MELON–RCE excludes the RCE module within MELON. (2) RCE–rel indicates the variant that obtains user-standpoint embeddings by transforming a user u's modality feature $\tilde{\mathbf{e}}_u^m$ through the MLP, without considering the complex relationship between modality features of a user and an item. (3) RCE–user indicates the variant that obtains a user u's preference on an item i by transforming $\tilde{\mathbf{e}}_{u,i}^m$ into a single scalar value through the MLP, without considering the preference for the complex relationship from her standpoint.

Table 4 shows the accuracies of MELON and its variants w.r.t. the MCE and RCE modules. From Table 4, we emphasize that

Table 4: The effects of MELON's two core ideas (MCE, RCE)

Measures	P@10	R@10	NDCG@10
MELON(MCE_5)	**0.0720**	**0.0710**	**0.0393**
MELON–MCE	0.0704	0.0695	0.0380
MCE_{10}	0.0694	0.0684	0.0379
MCE_{15}	0.0681	0.0672	0.0371
MCE_{20}	0.0671	0.0662	0.0365
$\text{MCE}_{\lvert\mathcal{N}_u\rvert}$	0.0710	0.0700	0.0385
$\text{MCE}_{\text{i-i}}$	0.0702	0.0692	0.0380
MELON–RCE	0.0689	0.0680	0.0371
RCE–rel	0.0690	0.0681	0.0373
RCE–user	0.0698	0.0690	0.0383

MELON outperforms *all* its *variants for all metrics* and discuss the effectiveness of each module as follows.

Effectiveness of the MCE Module. Our findings w.r.t. the effectiveness of the MCE module are summarized as follows:

i) MELON *outperforms* MELON–MCE. Also MELON–RCE shows *higher accuracy* than MONET (*i.e.*, the backbone). These results indicate that incorporating the relationships with modality-based neighbors is *crucial* in improving accuracy.

ii) MCE_k *outperforms* $\text{MCE}_{|\mathcal{N}_u|}$ in *all* metrics when k=5. This result indicates that constructing $\mathcal{G}_m$ by simply connecting k items for all users, although it may seem ad hoc, is more *beneficial* in improving accuracy than connecting $|\mathcal{N}_u|$ items flexibly for a user u. Furthermore, since MCE_5 achieves the highest accuracy in all datasets, we set k=5 for MELON in the following experiments.

iii) Although $\text{MCE}_{|\mathcal{N}_u|}$ shows *lower accuracy* than MELON, it still *outperforms* MELON–MCE. This comparison between the variants of MELON, demonstrates the importance of *leveraging modality-centered user–item relationships* to improve accuracy.

iv) MELON and $\text{MCE}_{|\mathcal{N}_u|}$ show *higher accuracy* than $\text{MCE}_{\text{i-i}}$. These improvements support the superiority of the modality-centered user–item bipartite graph $\mathcal{G}_m$ over the latent item-item graph, which *can not directly consider* the relationship between a user and her modality-based neighbors.

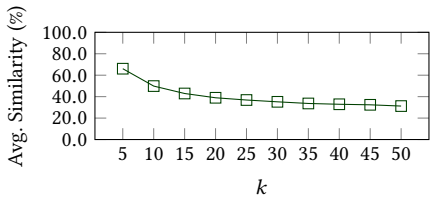

Figure 5: The average overlap ratio of modality-based neighbors captured by ICE and MCE modules.

Effectiveness of the RCE Module. Our findings w.r.t. the effectiveness of the RCE module are summarized as follows:

i) MELON *outperforms* MELON–RCE. Also, MELON–MCE shows *higher accuracy* than MONET. These results demonstrate that it is *beneficial* in improving accuracy to capture a complex relationship between the modality features of a user and an item from her standpoint.
ii) MELON *outperforms both* RCE–rel and RCE–user. This implies that taking *both* her standpoint and the complex relationship into account is *more effective* in improving accuracy than considering only one.
iii) Although RCE–rel (resp. RCE–user) only considers the user-standpoint (resp. the complex relationship), it still *outperforms* the best competitor, *i.e.*, MONET. This result indicates that considering either the complex relationship or the user-standpoint is *beneficial* for improving accuracy.

In conclusion, we have confirmed that our proposed (Idea 1) and (Idea 2) are *all effective* in improving accuracy and current design of the MCE and RCE modules is *reasonable* for realizing them. **RQ3: Effectiveness of the MCE Module in Capturing Relationships with Modality-based Neighbors**. We demonstrate that the MCE module helps the ICE module to capture effectively the relationships with modality-based neighbors via the following process. Specifically, we examine the degree of overlap between two sets of the modality-based neighbors captured by ICE and MCE modules. First, as in the PQ1, we construct a user u's different three sets of neighbors, Nei_u^m, Nei_u^{ICE}, and Nei_u^{MCE} by using u's modality feature $\tilde{\mathbf{e}}_u$, interaction-centered user embeddings $\mathbf{e}_u$, and modality-centered user embeddings $\bar{\mathbf{e}}_u$. After that, we calculate the similarity between $Nei_u^m \cap Nei_u^{ICE}$ and $Nei_u^m \cap Nei_u^{MCE}$, which indicates how well the ICE and MCE modules of MELON capture jointly a user's modality-based neighbors, by using the Jaccard similarity for each user and averaging it across all users. This process is repeated, increasing the number of neighbors, k, by 5 until it reaches 50.

Fig. 5 illustrates the result of the average overlap ratio between modality-based neighbors captured by ICE and MCE modules across all users. The average overlap ratio is highest when k=5, and decreases as k increases further. This result suggests that the MCE module captures distinct modality-based neighbors that are not identified by the ICE module, thereby enhancing MELON's overall ability to capture a user's modality-based neighbors. Furthermore, combining this finding with the results of RQ2—specifically, the accuracy comparison between MONET and MELON–RCE employing only the MCE module on MONET—demonstrates that additionally capturing the relationships with modality-based neighbors *contributes* to improving accuracy.

RQ4: Hyperparameter Sensitivity Analysis. The hyperparameter α in Eq. (10) adjusts the weight of $\hat{y}_{u,i}^{MCE}$ in $\hat{y}_{u,i}$, determining how much the relationships with modality-based neighbors are reflected. As shown in the top row of Fig. 6, as α increases up to 0.3, all metrics show a *gradual improvement*; after that, the accuracies progressively decrease. Specifically, MELON shows improvements by up to 1.84%, 1.29%, and 1.02% (resp. 4.05%, 3.51%, and 3.40%) in terms of P/R/NDCG@10 respectively, when α = 0.3 compared to the case of α= 0 (resp. α = 1). These results indicate that MELON appears to be *insensitive* to the changes in the hyperparameter α. Furthermore, the decrease in accuracy observed at α = 1 compared to that with α = 0.3 implies that overly emphasizing the relationships with modality-based neighbors may degrade accuracy. Next, the hyperparameter β in Eq. (10) controls the weight of $\hat{y}_{u,i}^{RCE}$ in $\hat{y}_{u,i}$, determining the extent to which her preference on the complex relationship formed between the modality features of a user and an item is reflected. The bottom row of Fig. 6 shows that the accuracies increase steadily until β reaches 0.5, after which they decrease for all metrics. Specifically, MELON shows improvements by up to 8.76%, 9.03%, and 14.58% (resp. 1.29%, 1.30%, and 2.67%) in terms of P/R/NDCG@10, respectively, when β = 0.5 compared to the case of β= 0 (resp. β = 1). These results indicate that MELON does not seem to be *insensitive* to the changes in the hyperparameter β.

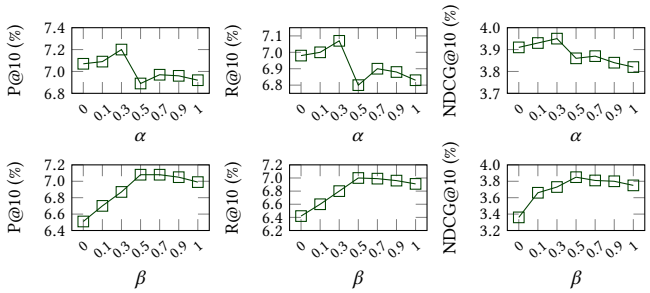

Figure 6: The effect of α and β on the accuracies of MELON.

6 Conclusion

We have explored two previously uncovered important challenges regarding the preferences related to modality features: (1) we showed that the modality-based neighbors *differ* from interaction-based neighbors and leveraging both two sets of neighbors is *beneficial* to improving accuracy; (2) we highlighted the importance of considering the *complex relationship* between the modality features of a user and an item from *her standpoint* to capture user preferences accurately. From these discoveries, we proposed MELON, a novel multimedia recommendation framework with two novel modules that capture users' *multi-aspect modality preferences*. We validated the effectiveness of MELON through extensive experiments on four real-world datasets, showing that it *consistently* and *significantly* outperforms *all* nine state-of-the-art competitors by up to 10.51% in terms of recall@10, compared to the best competitor.

Acknowledgments

This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2022-00155586, A High-Performance Big-Hypergraph Mining Platform for Real-World Downstream Tasks) and (No.RS-2020-II201373, Artificial Intelligence Graduate School Program (Hanyang University)).

References

[1] Xuheng Cai, Chao Huang, Lianghao Xia, and Xubin Ren. 2023. LightGCL: Simple Yet Effective Graph Contrastive Learning for Recommendation. In *The Eleventh International Conference on Learning Representations (ICLR)*.

[2] 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 *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*. 385–394.

[3] 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.

[4] Zhiqiang Guo, Jianjun Li, Guohui Li, Chaoyang Wang, Si Shi, and Bin Ruan. 2024. LGMRec: Local and Global Graph Learning for Multimodal Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 38. 8454–8462.

[5] Ruining He and Julian J. McAuley. 2016. VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. 144–150.

[6] 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 *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR)*. 639–648.

[7] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the ACM on Web Conference 2017 (ACM WWW)*. 173–182.

[8] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural computation* (1997), 1735–1780.

[9] Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull Soc Vaudoise Sci Nat* 37 (1901), 547–579.

[10] 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 ACM International Conference on Information and Knowledge Management (ACM CIKM)*. 993–1002.

[11] Yungi Kim, Taeri Kim, Won-Yong Shin, and Sang-Wook Kim. 2024. MONET: Modality-Embracing Graph Convolutional Network and Target-Aware Attention for Multimedia Recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (ACM WSDM)*. 332–340.

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

[13] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.

[14] 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.

[15] Zihan Lin, Changxin Tian, Yupeng Hou, and Wayne Xin Zhao. 2022. Improving graph collaborative filtering with neighborhood-enriched contrastive learning. In *Proceedings of the ACM on Web Conference 2022 (ACM WWW)*. 2320–2329.

[16] 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 *Proceedings of the ACM International Conference on Multimedia (ACM MM)*. 1526–1534.

[17] Julian J. McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-Based Recommendations on Styles and Substitutes. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR)*. 43–52.

[18] Zongshen Mu, Yueting Zhuang, Jie Tan, Jun Xiao, and Siliang Tang. 2022. Learning hybrid behavior patterns for multimedia recommendation. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*. 376–384.

[19] Rong Pan, Yunhong Zhou, Bin Cao, Nathan N Liu, Rajan Lukose, Martin Scholz, and Qiang Yang. 2008. One-class collaborative filtering. In *2008 Eighth IEEE international conference on data mining (IEEE ICDM)*. IEEE, 502–511.

[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)*. 3980–3990.

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

[22] David E Rumelhart, Geoffrey E Hinton, James L McClelland, et al. 1986. A general framework for parallel distributed processing. *Parallel distributed processing: Explorations in the microstructure of cognition* 1, 45-76 (1986), 26.

[23] 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.

[24] Wei Wei, Chao Huang, Lianghao Xia, and Chuxu Zhang. 2023. Multi-modal self-supervised learning for recommendation. In *Proceedings of the ACM Web Conference 2023 (ACM WWW)*. 790–800.

[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] Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. 2021. Self-supervised graph learning for recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval (ACM SIGIR)*. 726–735.

[28] Junliang Yu, Hongzhi Yin, Xin Xia, Tong Chen, Lizhen Cui, and Quoc Viet Hung Nguyen. 2022. Are graph augmentations necessary? simple graph contrastive learning for recommendation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval (ACM SIGIR)*. 1294–1303.

[29] Penghang Yu, Zhiyi Tan, Guanming Lu, and Bing-Kun Bao. 2023. Multi-view graph convolutional network for multimedia recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*. 6576–6585.

[30] 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.

[31] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Mengqi Zhang, Shu Wu, and Liang Wang. 2022. Latent Structure Mining with Contrastive Modality Fusion for Multimedia Recommendation. *IEEE Transactions on Knowledge and Data Engineering (IEEE TKDE)* (2022).

[32] Xin Zhou and Zhiqi Shen. 2023. A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*. 935–943.

[33] Yan Zhou, Jie Guo, Hao Sun, Bin Song, and Fei Richard Yu. 2023. Attention-guided multi-step fusion: a hierarchical fusion network for multimodal recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR)*. 1816–1820.