Disentangling and Generating Modalities for Recommendation in Missing Modality Scenarios

Jiwan Kim
kim.jiwan@kaist.ac.kr
KAIST
Daejeon, Republic of Korea

Hongseok Kang
ghdtjr0311@kaist.ac.kr
KAIST
Daejeon, Republic of Korea

Sein Kim
rlatpdlsgns@kaist.ac.kr
KAIST
Daejeon, Republic of Korea

Kibum Kim
kb.kim@kaist.ac.kr
KAIST
Daejeon, Republic of Korea

Chanyoung Park*
cy.park@kaist.ac.kr
KAIST
Daejeon, Republic of Korea

Abstract

Multi-modal recommender systems (MRSs) have achieved notable success in improving personalization by leveraging diverse modalities such as images, text, and audio. However, two key challenges remain insufficiently addressed: (1) Insufficient consideration of missing modality scenarios and (2) the overlooking of unique characteristics of modality features. These challenges result in significant performance degradation in realistic situations where modalities are missing. To address these issues, we propose **D**isentangling and **G**enerating **M**odality **Rec**ommender (DGMRec), a novel framework tailored for missing modality scenarios. DGMRec disentangles modality features into general and specific modality features from an information-based perspective, enabling richer representations for recommendation. Building on this, it generates missing modality features by integrating aligned features from other modalities and leveraging user modality preferences. Extensive experiments show that DGMRec consistently outperforms state-of-the-art MRSs in challenging scenarios, including missing modalities and new item settings as well as diverse missing ratios and varying levels of missing modalities. Moreover, DGMRec's generation-based approach enables cross-modal retrieval, a task inapplicable for existing MRSs, highlighting its adaptability and potential for real-world applications. Our code is available at https://github.com/ptkjw1997/DGMRec.

CCS Concepts

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

Keywords

Multi-modal Recommender Systems; Missing Modalities; Collaborative Filtering; Feature Disentanglement

ACM Reference Format:

Jiwan Kim, Hongseok Kang, Sein Kim, Kibum Kim, and Chanyoung Park. 2025. Disentangling and Generating Modalities for Recommendation in Missing Modality Scenarios. 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/10.1145/3726302.3729953

*Corresponding author

SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3729953

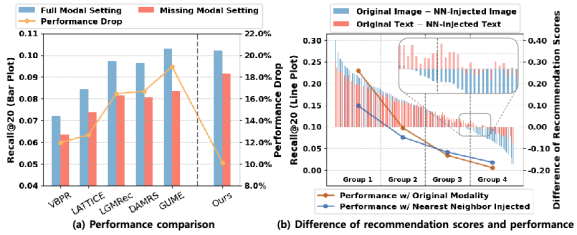

Figure 1: (a) Performance drop of recent MRSs when missing modality exists. (b) Difference between the model's recommendation scores under two conditions: when the modality exists (Original Image/Text) and when the modality is missing (NN-Injected Image/Text). Line plots indicate the performance of LGMRec [6].

1 Introduction

In recent years, e-commerce platforms such as Amazon and Alibaba, as well as social media services like YouTube and TikTok, have become an integral part of everyday life. Their recommendation systems play a pivotal role in shaping user behavior and decision-making. In particular, traditional methods using Collaborative Filtering (CF), such as Matrix Factorization (MF) [18] or Graph Neural Networks (GNNs) [9, 22], have successfully delivered personalized recommendations, making these systems increasingly important for both customers and businesses.

While collaborative filtering (CF) models have proven successful, they are still constrained by the inherent sparsity of user feedback [36]. To address this, the integration of diverse modalities, such as images, text, and audio, has emerged as a promising solution, complementing the lack of feedback and enabling a deeper understanding of user preferences compared to traditional CF approaches. This has driven the development of multi-modal recommender systems (MRSs), which leverage multi-modal content to derive rich semantic representations and uncover relationships between items that CF models alone cannot achieve.

Although MRSs have demonstrated their effectiveness, several practical challenges remain insufficiently addressed.

C1. Missing modality scenarios are not sufficiently addressed. Prior MRSs generally assume that all modality features of an item are always fully available. However, in the real-world industry, some or all modality features of an item may be missing [3]. Specifically, existing studies in MRSs often address the missing

modality scenarios by either 1) simply dropping out the items with missing-modality features from the training dataset [28, 36] or 2) injecting synthetic features for missing modalities (e.g., using the global mean of a modality [14] or the nearest neighbor mean (NN) [13]). Although injection methods are shown to be effective [13], they still encounter a significant performance drop under missing modality scenarios. In Figure 1(a), we compared the performance of several MRSs that assume all modality features are available, including early models (i.e., VBPR [8] and LATTICE [34]) as well as state-of-the-art (SOTA) models (i.e., LGMRec [6], DAMRS [29], and GUME [11]), on the Amazon Baby dataset under two scenarios: one where all modalities are available and the other where some modalities are randomly missing[1]. We found that such a naive injection approach still experiences significant performance drop in the presence of missing modalities.

To investigate why the NN-injection method fails to prevent the performance drop under missing modality, we analyze its impact on the model prediction of a recent MRS, LGMRec [6]. In Figure 1(b), we calculate the difference in recommendation scores of a positive user–item pair between scenarios where the item includes all modality features and where a specific modality feature (either image or text) is missing[1], and sort them in descending order. We also divided the positive user–item pairs into four groups based on their sorted differences, and evaluated the recommendation performance for each group (in the line plot). We observe that the drop in the recommendation performance under missing modality (i.e., red line - blue line) is more severe when the difference of the recommendation scores is larger. Since a difference of the recommendation scores indicates how well the NN-injected feature captures the original modality (i.e., the smaller the better), this result implies that the poor recommendation performance is originated from the NN-injected feature failing to successfully substitute the original modality. The performance degradation under missing modality exacerbates in recent SOTA MRSs that heavily rely on item modality features.

C2. Unique characteristics of modalities are overlooked. Existing MRSs [6, 11] generally aim to directly align among different modalities of an item, assuming that the features of various modalities for the same item inherently share semantics. However, this assumption does not always hold. Specifically, the image modality tends to capture visual attributes such as color and style, emphasizing the tangible and aesthetic aspects of an item. In contrast, the text modality conveys descriptive and contextual information, highlighting functional attributes or background details. This indicates that each modality contains unique, modality-specific information that cannot be fully captured by other modalities. For this reason, directly aligning different modalities of an item as in prior studies [6, 11] fails to account for these distinct characteristics, hindering the development of high-quality item representations.

In fact, an existing MRS, LGMRec [6], failing to capture the modality-specific information can be observed by closely examining the difference of the recommendation scores of Group 3 and 4 in Figure 1(b), where the sign of the red and the blue bars are the opposite (See the zoomed part). We argue that the items with the opposite sign in the red and blue bars are those that LGMRec failed to account for the unique characteristics of the text and image modalities. More precisely, as the injected features can be considered as the features commonly shared by all items, the difference in the recommendation scores can be interpreted as the distinct modality-specific information that remains for an item after accounting for the commonly shared features. However, as the injected features cannot be perfectly representative of the globally shared common features of a certain modality, the difference of the recommendation scores would contain not only the modality-specific information but also generally shared information within a modality of an item. From this perspective, the positive difference observed for text (red bars) indicates that the remaining information in text is informative to the model's predictions, whereas the negative difference in image (blue bars) suggests the information retained in images less helpful or may even hinder the predictions. This contrast that is present for an item arises because the modality-specific information provided by text and image modalities differs significantly. These observations highlight that each modality contains unique information that is neither shared nor aligned with other modalities. Hence, in these cases, forcing alignment between modalities can obscure their unique contributions and adversely affect recommendation performance.

While there has been some research on missing-modality aware recommender systems (MMA-RSs) [1, 5, 12, 21] to address the realistic challenge of missing modalities, these approaches primarily focus on the robustness of the models when handling items with missing modalities. However, because they use the missing modalities in their incomplete state without addressing them, these methods fail to leverage the unique characteristics inherent to each modality. Additionally, since they mainly rely on content-based approaches, they tend to underperform conventional collaborative filtering (CF) methods in general performance, limiting their practical applicability despite their effort to handle real-world challenges.

To overcome the inherent limitations of MRSs and MMA-RSs, we propose a novel model called **D**isentangling and **G**enerating **M**odality **Rec**ommender (DGMRec), which effectively addresses the challenges of handling missing modality features and extracting common and unique characteristics of modalities.

C1: To handle missing modality features, DGMRec employs an autoencoder architecture to reconstruct an item's modality features to closely resemble the actual ones, ensuring the preservation of the item's distinct attributes. For items with missing modalities, DGMRec generates modality features by leveraging two sources of knowledge: aligned features from other available modalities and interacted users' modality preferences. Using generated features, DGMRec enhances the item-item graph, achieving more robust and richer semantic relations between items, which existing models have struggled to capture when missing modality exists.

C2: To extract common and unique modality features, DGMRec derives general and specific modality features using separate encoders. Additionally, it employs two information-based loss functions to disentangle the general and specific features within a single modality while simultaneously learning shared traits across different modalities.

These two challenges are closely interconnected, as disentangling modality attributes directly impacts the quality of feature generation for accurately representing an item.

Our contributions are summarized as follows:

[1]The missing modality feature is injected based on the NN-injection approach [13].

- We identify and analyze the significant performance degradation of current MRSs in missing modality scenarios, highlighting the inadequacy of naive injection methods as the core limitation.
- We propose a robust generation-based approach that reconstructs missing modalities by leveraging an item's distinct characteristics, enabling DGMRec to achieve superior performance across diverse real-world scenarios.
- The proposed fine-grained missing modality feature generation enables DGMRec to perform additional tasks, such as cross-modal retrieval, to assist user behavior in missing scenarios, which remains unattainable for existing recommenders.
- We introduce a novel approach from the perspective of mutual information to separate and learn general and specific modality features.

2 Related Works

2.1 Multi-modal Recommender Systems

Multi-modal Recommenders (MRSs) leverage multi-modal content in various ways. **(1) Feature-based** approaches that directly utilize features have been extensively studied. For instance, VBPR [8] integrates visual features directly with ID embeddings. Similarly, methods such as SLMRec [19], GRCN [24], and MMGCN [25] use graph convolution networks (GCNs) to integrate modality knowledge with CF knowledge. BM3 [39] employs contrastive views of modalities for self-supervised learning, while LGMRec [6] adopts a hybrid approach by utilizing both hypergraphs and local graphs to balance the learning of global and local knowledge in modality feature extraction. On the other hand, **(2) Graph-based** approaches focus on identifying relationships between items based on modality features rather than directly utilizing these features. Notable examples include LATTICE [34], MICRO [35], FREEDOM [38], and DAMRS [29]. Recently, **(3) Hybrid** approaches combining both strategies have emerged. For instance, GUME [11] and MGCN [31] simultaneously leverage modality features and item-item graphs.

These MRSs have benefited significantly from multi-modal alignment. However, existing methods either completely disregard the relationships between modalities [19, 39] or rely solely on direct alignment to extract shared information between modalities [6, 11]. As a result, the unique information inherent to different modalities is often overlooked and remains unlearned. Some recent works have highlighted the importance of unique modality features for the qualitative representation of items and have adopted orthogonal learning methods from other domains [11, 31, 32]. However, their simplistic approach—taking the mean of modality features as shared commonality and substituting it with the modality features for the unique features—lacks sufficient logical justification from an informational perspective.

2.2 Missing Modality-Aware Recommender Systems

Most existing MRSs assume that all modalities are available and complete, which is rarely the case in real-world applications. Missing modality-aware recommender systems (MMA-RSs) address these challenges by considering two common scenarios: (1) **Incomplete Modality**, where some feature values are missing in a modality [21, 27, 40], and (2) **Missing Modality**, where an entire modality is unavailable [1, 5, 12, 13]. MILK [1] and SIBRAR [5] tackle missing modalities by leveraging invariant learning and single-branch networks to make robust recommendations without directly addressing the missing modality. Alternatively, [13] investigates which features can serve as effective substitutes for missing modalities while leaving the model architecture largely unchanged. However, most MMA-RS methods not only fail to effectively capture CF knowledge due to their content-based architecture but also rely on naive approaches that lack sufficient consideration of the unique characteristics of each modality, resulting in suboptimal overall performance. In this regard, the work most aligned with our motivation is CI2MG [12], which utilizes hypergraphs and optimal transport (OT) to generate missing modalities. However, CI2MG suffers from significant computational overhead in calculating OT and lacks integration between the OT process and other recommendation modules.

3 Methodology

To accurately generate missing modalities that reflect the distinct characteristics of items, it is essential to consider both the general (shared) and specific (unique) features of modalities. To achieve this, we propose a **Disentangling Modality Feature** module (Sec 3.2), which uses separate encoders to extract general and specific features. Additionally, we introduce information-based losses to disentangle these features and ensure the alignment of general features across modalities.

Subsequently, to extract meaningful modality representations and construct item-item graphs in the presence of missing modalities, we propose a **Missing Modality Generation** module (Section 3.3), which generates general and specific features that capture the distinct characteristics of items in a fine-grained manner and adaptively refines the item-item graph using these generated features.

Finally, while modality features are effective in capturing semantic content, they lack the collaborative knowledge critical for recommendation tasks [10, 33]. To mitigate this, we introduce two additional **alignment methods** (Section 3.4) to connect user representations with item representations, thereby bridging collaborative filtering with modality features.

3.1 Preliminaries

Let $\mathcal{U}$ and $\mathcal{I}$ denote the sets of users and items, with $|\mathcal{U}|$ and $|\mathcal{I}|$ representing the number of users and items, respectively. The User-Item Interaction matrix is defined as $\mathcal{R} = \mathbb{R}^{|\mathcal{U}|\times|\mathcal{I}|}$. Each item's modality feature is extracted using pre-trained models (i.e., image with a CNN model [7] and text with a SBERT [17]) and represented as $X_m \in \mathbb{R}^{|\mathcal{I}|\times d_m}$, where missing modality features are initialized with their mean values. Additionally, the ID embeddings for users and items are defined as $E_{id} \in \mathbb{R}^{(|\mathcal{U}|+|\mathcal{I}|)\times d}$.[2] $\mathcal{N}_i$ denotes the set of users that interacted with item i, and $\mathcal{N}_u$ denotes the set of items that user u interacted with.

3.2 Disentangling Modality Feature Module

Existing MRSs directly align modality features [6, 11] or combine adjacency views across modalities [34, 38], obscuring unique characteristics of modalities. To address this, we extract two types of features—general and specific—from each modality using encoder

[2]Matrices are denoted by uppercase letters and vectors by lowercase letters for clarity and consistency (e.g., $e_{i,id}$ represents the ID embedding vector for the i-th item, and $e_{u,id}$ for the u-th user, where i, j index items and u, v index users).

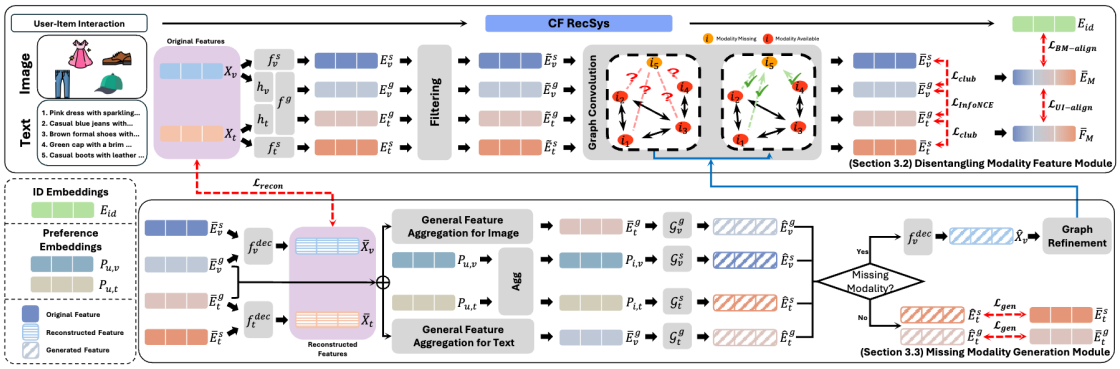

Figure 2: Overview of DGMRec framework. It consists of the Disentangling Modality Feature module and the Missing Modality Generation module. In the Missing Modality Generation module, we illustrate the case where an item is associated with the text modality while the image modality is missing.

functions (Sec 3.2.1). These features are further refined through a GCN-based item-item graph and information-driven disentanglement (Sec 3.2.2).

3.2.1 Extracting Modality Features. To extract general E_m^g and specific features E_m^s for each modality m, we employ two separate encoders composed of a fully-connected layer: the general encoder f^g applicable across all modalities and the specific encoder f_m^s tailored to each modality m.

$$E_m^g = f^g(h_m(X_m)), \quad E_m^s = f_m^s(X_m) \tag{1}$$

where the general encoder $f^g : \mathbb{R}^d \to \mathbb{R}^d$ shares parameters across modalities to extract common attributes, while the specific encoder $f_m^s : \mathbb{R}^{d_m} \to \mathbb{R}^d$ employs independent parameters to capture the unique characteristics of each modality m. Moreover, since X_m has different dimensions across modalities, we use another fully-connected layer $h_m : \mathbb{R}^{d_m} \to \mathbb{R}^d$ to project each modality into a unified dimension.

In addition, we introduce a **Modality Preference Embedding** matrix for users, denoted as $P_{u,m} \in \mathbb{R}^{|\mathcal{U}| \times d}$, to more effectively capture users' modality preferences, which is then used to compute the item's modality preference matrix $P_{i,m} \in \mathbb{R}^{|\mathcal{I}| \times d}$ as follows:

$$p_{i,m} = \frac{1}{|\mathcal{N}_i|} \sum_{u \in \mathcal{N}_i} p_{u,m} \tag{2}$$

where $p_{u,m} \in \mathbb{R}^d$ and $p_{i,m} \in \mathbb{R}^d$ are the modality preference embedding of user u and item i, respectively. As this item preference embedding matrix $P_{i,m}$ contains the modality preference of interacted users, we use it to align the modality feature with the user preference as follows: [3]

$$\tilde{E}_m = E_m \odot \sigma(P_{i,m}) \tag{3}$$

where $\odot$ is the element-wise product and σ is the sigmoid function. We expect the denoised features $\tilde{E}_m$ to retain information relevant to the user preference, leading to improved recommendations.

The denoised modality features are then enhanced through GCNs using an item-item graph. Specifically, we construct an adjacency matrix S^m with top-k similar items based on similarity scores of items' modality features X_m where S^m is computed as follows:

$$S_{i,j}^m = \frac{(x_{i,m})^\top x_{j,m}}{\|x_{i,m}\| \|x_{j,m}\|} \tag{4}$$

As the backbone GCNs, we utilize LightGCN [9] for its computational simplicity and widespread adoption as follows:

$$\bar{E}_m^{(l)} = S^m \cdot \bar{E}_m^{(l-1)}, \text{where } \bar{E}_m^{(0)} = \tilde{E}_m \text{ and } \bar{E}_m = \bar{E}_m^{(L)} \tag{5}$$

where $\bar{E}_m^{(l)} \in \mathbb{R}^{|\mathcal{I}| \times d}$ denotes the modality feature at the l-th layer of graph convolution, L is the number of layers. Note that we use the last L-th layer representation as the item's modality feature. Finally, given the item's final modality feature matrix $\bar{E}_m^{(L)}$, we compute the user modality feature matrix $\bar{F}_m \in \mathbb{R}^{|\mathcal{U}| \times d}$ by aggregating the modality features of items the user has interacted with as follows:

$$\bar{f}_{u,m} = \frac{1}{|\mathcal{N}_u|} \sum_{i \in \mathcal{N}_u} \bar{e}_{i,m} \tag{6}$$

3.2.2 Disentangling Modality Features. Separating encoders for general and specific features alone is insufficient to achieve effective disentanglement. Therefore, we introduce information-based approaches leveraging two contrastive losses: one reduces mutual information between general and specific features within a single modality, while the other enhances mutual information between general features across multiple modalities.

To minimize the mutual information between general and specific features within the same modality, we employ a sample-based approach using the Contrastive Log-ratio Upper Bound (CLUB) [2] with variational distribution $q_\phi(\cdot|\cdot)$ with parameter ϕ to estimate conditional distribution $p(\cdot|\cdot)$. q_ϕ consisting of 2-layer MLPs.

$$\mathcal{L}_{club} = \sum_{i \in \mathcal{I}} \left[\log q_\phi(\bar{e}_{i,m}^g | \bar{e}_{i,m}^s) - \frac{1}{|\mathcal{I}|} \sum_{j \in \mathcal{I}} \log q_\phi(\bar{e}_{j,m}^g | \bar{e}_{i,m}^s) \right] \tag{7}$$

Following [2], $\mathcal{L}_{club}$ approximates the upper bound of mutual information between $\bar{E}_m^g$ and $\bar{E}_m^s$. We iteratively minimize $\mathcal{L}_{club}$ alongside other model parameters, encouraging modality's general and specific features to have complementary information.

To maximize the mutual information between general features of different modalities, we adopt the InfoNCE loss [15]

[3] For notational convenience, the superscripts g and s, denoting general and specific features, are omitted in this equation and subsequent equations where the context makes their meaning clear.

to approximate a negative lower bound of mutual information.

$$\mathcal{L}_{InfoNCE} = \sum_{i \in \mathcal{I}} -\log \frac{\exp(\bar{e}^g_{i,m} \cdot \bar{e}^g_{i,m'})}{\sum_{j \in \mathcal{I}} \exp(\bar{e}^g_{i,m} \cdot \bar{e}^g_{j,m'})} \tag{8}$$

InfoNCE effectively aligns the general features of different modalities (i.e., $\bar{E}^g_m$ and $\bar{E}^g_{m'}$ for different modalities m, m'), by mapping them into the same latent space. Thus, minimizing InfoNCE loss maximizes the lower bound of mutual information, ensuring better alignment among the general features of multiple modalities.

The final loss for disentanglement is shown below:

$$\mathcal{L}_{disentangle} = \mathcal{L}_{club} + \mathcal{L}_{InfoNCE} \tag{9}$$

In summary, we expect $\mathcal{L}_{disentangle}$, to effectively disentangle general and specific features within a modality while aligning general features across different modalities. These losses enable the utilization of cross-modality information to generate missing modality representations, maintaining well-aligned modalities and preserving their unique characteristics

3.3 Missing Modality Generation Module

As discussed in Section 3.2, modality features are disentangled into general and specific components using an information-based approach. Building on these well-separated features, we train the decoder to accurately reconstruct raw modality features (Sec 3.3.1). Then, DGMRec generates general and specific features for missing modalities in a tailored manner (Sec 3.3.2). These generated features are then utilized to generate raw modality features and refine the item-item graph, addressing the instability caused by missing modalities (Sec 3.3.3).

3.3.1 Modality Feature Reconstruction. We employ an additional decoder $f^{dec}_m : \mathbb{R}^d \rightarrow \mathbb{R}^{d_m}$ for each modality m that reconstructs a modality's raw feature X_m using general features $\bar{E}^g_m$ and specific features $\bar{E}^s_m$ as follows:

$$\bar{X}_m = f^{dec}_m(\bar{E}^g_m \oplus \bar{E}^s_m) \tag{10}$$

where $\oplus$ denotes the concatenation operation. The reconstruction process is guided by a reconstruction loss $\mathcal{L}_{recon}$ between raw modality features X_m and reconstructed modality features $\bar{X}_m$. It ensures that the reconstructed features closely resemble raw feature, allowing DGMRec to accurately generate features for missing modalities by leveraging the learned $\bar{E}^g_m$ and $\bar{E}^s_m$.

$$\mathcal{L}_{recon} = \sum_{i \in \mathcal{I}} MSE(x_{i,m}, \bar{x}_{i,m}) \tag{11}$$

where *MSE* is the mean squared error loss. That is, the disentangled features $\bar{E}^g_m$ and $\bar{E}^s_m$ retain meaningful modality information to accurately reconstruct $\bar{X}_m$, rather than being merely meaningless representations of disentanglement.

3.3.2 Missing Modality Generation. To address missing modalities, it is crucial to effectively generate both general and specific features separately, which are used to generate raw modality feature through decoders. Although these two features (general and specific) originate from the same modality, they possess fundamentally different characteristics and thus require distinct approaches for generation.

To **generate general features** $\hat{E}^g_m$, we leverage the original general features $\bar{E}^g_{m'}$ from other modalities m', aligned by $\mathcal{L}_{InfoNCE}$ in Eq 8. As directly using these features could be unstable, we introduce a general feature generator $\mathcal{G}^g_m$, which consists of 2-layer MLPs, for each modality m to generate general features as follows:

$$\hat{E}^g_m = \mathcal{G}^g_m(\bigoplus_{m'} \bar{E}^g_{m'}) \tag{12}$$

where $\bigoplus$ denotes the concatenation operation applied across all available modalities m'. Note that when the input modality m' is also missing (i.e., when two or more modalities are missing), we use the mean of the modality features to ensure that the concatenated vector maintains a consistent dimension regardless of the number of missing modalities.

To **generate specific features** $\hat{E}^s_m$, information from other modalities cannot be utilized. Instead, we leverage user modality preferences, which are aligned with the items' modality features as shown in Eq 3. Since the modality-specific knowledge of an item is implicitly captured within the modality preferences of all users associated with that item, a specific feature generator $\mathcal{G}^s_m$ for each modality, which consists of 2-layer MLPs, uses this information to generate the specific features for each item as follows:

$$\hat{E}^s_m = \mathcal{G}^s_m(P_{i,m}) \tag{13}$$

To guarantee that the generated features preserve the modality's original information, we introduce the generation loss $\mathcal{L}_{gen}$, which encourages the original features ($\bar{E}^g_m$, $\bar{E}^s_m$) and the generated features ($\hat{E}^g_m$, $\hat{E}^s_m$) to be as similar as possible as follows:

$$\mathcal{L}_{gen} = \sum_{i \in \mathcal{I}} \left(MSE(\bar{e}^g_{i,m}, \hat{e}^g_{i,m}) + MSE(\bar{e}^s_{i,m}, \hat{e}^s_{i,m}) \right) \tag{14}$$

Importantly, these two generation-related losses ($\mathcal{L}_{recon}$ and $\mathcal{L}_{gen}$) are computed only for items with available modalities, preventing missing modality items from hindering the training process.

3.3.3 Refining Item-Item Graph via Generated Features. At regular intervals determined by a hyperparameter (i.e., every 5 epochs), DGMRec generates features for items with missing modalities. The generated $\hat{E}^g_m$ and $\hat{E}^s_m$ are then used to generate $\hat{X}_m$, approximating the raw feature X_m using the decoder as follows:

$$\hat{X}_m = f^{dec}_m(\hat{E}^g_m \oplus \hat{E}^s_m) \tag{15}$$

However, simply substituting raw modality features with the generated features is insufficient for addressing the issues caused by missing modalities. As observed in Figure 1, SOTA MRSs that heavily rely on modality features using item-item graphs suffer severe performance degradation. This occurs because inappropriate edges that are formed when NN features are injected disrupt stable training and hinder the propagation of true semantic relationships. Hence, we construct a new adjacency matrix $\hat{S}^m$ that accurately reflects the semantic relationships among items, based on the generated raw features $\hat{X}_m$, following the same process shown in Eq 4.

Moreover, to prevent the instability in the model training caused by abrupt graph changes, we propose an adaptive update strategy to smoothly integrate new connections using the adjustable hyperparameter α as follows:

$$S^m = \alpha S^m + (1 - \alpha)\hat{S}^m \tag{16}$$

This is expected to enhance the graph structure and capture richer semantic relationships between items.

It is important to note that we only update the edges associated with items with missing modalities, and construct directed edges from items containing a modality to those with missing modalities.

The purpose of this design is twofold: It prevents contamination of original modality features by under-trained generated features, thereby avoiding instability during training. Additionally, by computing similarity scores solely for pairs involving items with missing modalities, the process reduces computational overhead, ensuring both stability and efficiency in graph refinement. This approach enables the item-item graph to capture meaningful semantic relationships while maintaining robustness during training.

3.4 Alignment for Recommendation Task

Modality features effectively capture semantic content but lack the collaborative knowledge essential for recommendation tasks [10, 33]. To address this limitation, we propose two alignment methods that bridge collaborative filtering with modality features. By aligning modality features extracted from pre-trained models with ID embeddings, we seamlessly integrate modality knowledge and collaborative knowledge.

3.4.1 Fusing ID embedding and Modality Features. The final modality representation is computed by first obtaining the general features of all modalities through mean pooling. Subsequently, these combined general features are mean pooled with the specific features of the modalities to produce the final modality representation.

$$\begin{aligned} \bar{E}_M &= MeanPool_{m\in\mathcal{M}}(\bar{E}^s_m, MeanPool_{m'\in\mathcal{M}}(\bar{E}^g_{m'})) \\ \bar{F}_M &= MeanPool_{m\in\mathcal{M}}(\bar{F}^s_m, MeanPool_{m'\in\mathcal{M}}(\bar{F}^g_{m'})) \end{aligned} \tag{17}$$

where $\mathcal{M}$ is a set of all modalities. Incorporating modality features, the final recommendation representation for the user and item is derived by combining the ID embedding from the CF model with the respective modality representations as follows:

$$\bar{E}_u = E_{u,id} + \bar{F}_M, \quad \bar{E}_i = E_{i,id} + \bar{E}_M \tag{18}$$

The final recommendation score $y_{u,i}$ is computed by the inner product between the user and item representations as follows:

$$y_{u,i} = \bar{e}_u^\top \bar{e}_i \tag{19}$$

3.4.2 Behavior-Modality Alignment. While modality features are important for item representations, the lack of collaborative knowledge makes it difficult to capture user-item relationships [10, 33]. To alleviate this challenge, we incorporate collaborative knowledge by aligning user behavior with modality features through a contrastive loss.

$$\begin{aligned} \mathcal{L}_{BM\text{-}align} = &\sum_{u\in\mathcal{U}} -\log \frac{\exp(e_{u,id}\cdot \bar{f}_{u,M})}{\sum_{v\in\mathcal{U}}\exp(e_{u,id}\cdot \bar{f}_{v,M})} \\ &+ \sum_{i\in\mathcal{I}} -\log \frac{\exp(e_{i,id}\cdot \bar{e}_{i,M})}{\sum_{j\in\mathcal{I}}\exp(e_{i,id}\cdot \bar{e}_{j,M})} \end{aligned} \tag{20}$$

3.4.3 User-Item Alignment. We further refine modality features among users and items to ensure coherence within each modality as follows:

$$\mathcal{L}_{UI\text{-}align} = \sum_{m\in\mathcal{M}}\sum_{(u,i)\in O} -\log \frac{\exp(\bar{f}_{u,m}\cdot \bar{e}_{i,m})}{\sum_{j\in\mathcal{I}}\exp(\bar{f}_{u,m}\cdot \bar{e}_{j,m})} \tag{21}$$

Dataset	# Users	# Items	# Interactions	Sparsity	Modalities		
					Image	Text	Audio
Baby	19,445	7,050	160,792	99.88%	✓	✓	✗
Sports	35,598	18,357	296,337	99.95%	✓	✓	✗
Clothing	39,387	23,033	278,677	99.97%	✓	✓	✗
TikTok	9,308	6,710	68,722	99.89%	✓	✓	✓

Table 1: Statistics of Datasets

where O is a set of positive observation in interaction matrix $\mathcal{R}$ (i.e., (u,i) is contained in O when $\mathcal{R}_{u,i}=1$).

By aligning the modality features of an item with those of users who interacted with it, DGMRec captures recommendation-relevant knowledge, enhancing the consistency of the modality features.

The final alignment loss is defined as:

$$\mathcal{L}_{align} = \mathcal{L}_{BM\text{-}align} + \mathcal{L}_{UI\text{-}align} \tag{22}$$

By combining $\mathcal{L}_{BM\text{-}align}$ and $\mathcal{L}_{UI\text{-}align}$, our approach effectively integrates modality features with collaborative filtering knowledge, resulting in a robust and comprehensive recommendation system.

To optimize user and item representations for the recommendation task, we employ the Bayesian Personalized Ranking (BPR) loss [18],

$$\mathcal{L}_{bpr} = \sum_{(u,i^+,i^-)\in\mathcal{D}} \left(-\sigma(y_{u,i^+} - y_{u,i^-})\right) \tag{23}$$

where $\mathcal{D} = \{(u,i^+,i^-)|(u,i^+)\in O, (u,i^-)\notin O\}$ represents the dataset of triplets, and $\sigma()$ denotes the sigmoid function.

The final objective function of DGMRec is give by:

$$\mathcal{L} = \mathcal{L}_{bpr} + \mathcal{L}_{recon} + \mathcal{L}_{gen} + \lambda_1\mathcal{L}_{disentangle} + \lambda_2\mathcal{L}_{align} \tag{24}$$

where λ_1, λ_2 are hyper-parameters.

4 Experiment

4.1 Experimental Settings

Datasets. We use datasets with diverse modalities, including the Amazon Baby, Sports, and Clothing datasets, as well as the TikTok dataset with 5-core setting following previous works [11, 23, 29, 34]. The Amazon datasets contain image and text modalities. Modality features in Amazon datasets are extracted using the same pre-trained models following [37] (i.e., image with a CNN model [7] and text with a SBERT [17]). The TikTok dataset, published by TikTok[4], includes image, text, and audio modalities. However, the raw features and pre-trained models used for feature extraction are not publicly available. The statistics of datasets are summarized in Table 1.

Missing Modality Setting. Similar to other MMA-RSs [1, 5, 12], we introduce settings where missing modalities are present. For datasets with two modalities, items were evenly divided such that 1/3 of items have 0, 1, or 2 missing modalities. For three modalities, 1/4 of items have 0, 1, 2, or 3 missing modalities. The specific modality chosen as missing was randomly selected for each item.

New Items Setting. Following the setup of [1], we select 20% of the items as new items that appeared only in the test set, ensuring these items were unseen during the training and validation phases. This setup evaluates the model's ability to generalize to previously unseen items, testing its performance in realistic scenarios.

Compared Methods. To ensure fair comparisons, we evaluated DGMRec against a wide range of models, including 5 traditional CF models, 10 multi-modal RSs, and 3 missing modality-aware RSs.

[4]https://www.tiktok.com/

Table 2: Performance Comparison. The best and runner-ups are marked in bold and underlined, respectively.

Missing Modality Setting																	
Dataset		Baby				Sports				Clothing				TikTok			
Metric		R@20	R@50	N@20	N@50	R@20	R@50	N@20	N@50	R@20	R@50	N@20	N@50	R@20	R@50	N@20	N@50
CF	MF	0.0611	0.1091	0.0273	0.0370	0.0707	0.1112	0.0327	0.0416	0.0346	0.0533	0.0164	0.0201	0.0558	0.0909	0.0220	0.0289
	NGCF	0.0602	0.1137	0.0258	0.0366	0.0701	0.1215	0.0304	0.0408	0.0422	0.0728	0.0176	0.0237	0.0722	0.1284	0.0284	0.0394
	LightGCN	0.0733	0.1323	0.0320	0.0440	0.0829	0.1369	0.0379	0.0488	0.0514	0.0818	0.0227	0.0288	0.0916	0.1576	0.0406	0.0536
	SGL	0.0804	0.1422	0.0348	0.0473	0.0917	0.1492	0.0414	0.0531	0.0600	0.0936	0.0271	0.0338	0.0939	0.1490	0.0403	0.0521
	SimGCL	0.0809	0.1409	0.0349	0.0471	0.0910	0.1465	0.0410	0.0523	0.0542	0.0833	0.0252	0.0310	0.0952	0.1451	0.0401	0.0509
Multi-Modal Recommenders	VBPR	0.0514	0.0937	0.0213	0.0299	0.0741	0.1229	0.0328	0.0427	0.0462	0.0737	0.0207	0.0226	0.0410	0.0699	0.0172	0.0229
	MMGCN	0.0519	0.0991	0.0215	0.0310	0.0509	0.0913	0.0215	0.0297	0.0289	0.0530	0.0120	0.0168	0.0883	0.1431	0.0372	0.0484
	GRCN	0.0644	0.1151	0.0274	0.0377	0.0681	0.1157	0.0300	0.0397	0.0381	0.0644	0.0161	0.0214	0.0716	0.1257	0.0283	0.0389
	SLMRec	0.0753	0.1254	0.0340	0.0422	0.0914	0.1462	0.0415	0.0526	0.0624	0.0979	0.0281	0.0351	0.0932	0.1523	0.0364	0.0480
	BM3	0.0683	0.1235	0.0296	0.0408	0.0908	0.1466	0.0400	0.0513	0.0591	0.0920	0.0268	0.0334	0.0768	0.1215	0.0322	0.0409
	LATTICE	0.0738	0.1297	0.0319	0.0432	0.0867	0.1401	0.0306	0.0384	0.0581	0.0929	0.0262	0.0332	0.0824	0.1353	0.0372	0.0477
	MGCN	0.0833	0.1389	0.0366	0.0481	0.0941	0.1525	0.0425	0.0544	0.0665	0.1052	0.0300	0.0377	0.0870	0.1395	0.0356	0.0460
	LGMRec	0.0813	0.1410	0.0352	0.0471	0.0906	0.1496	0.0403	0.0522	0.0624	0.1015	0.0277	0.0355	0.0791	0.1376	0.0335	0.0450
	DAMRS	0.0804	0.1390	0.0355	0.0474	0.0941	0.1526	0.0416	0.0534	0.0670	0.1066	0.0301	0.0380	0.1044	0.1638	0.0452	0.0569
	GUME	0.0835	0.1429	0.0369	0.0489	0.0947	0.1554	0.0424	0.0546	0.0639	0.1016	0.0291	0.0366	0.0968	0.1645	0.0389	0.0524
MMA RSs	CI2MG	0.0720	0.1285	0.0305	0.0420	0.0717	0.1179	0.0331	0.0425	0.0523	0.0845	0.0237	0.0301	0.0772	0.1284	0.0327	0.0429
	MILK	0.0427	0.0763	0.0182	0.0250	0.0362	0.0626	0.0155	0.0209	0.0226	0.0376	0.0094	0.0124	0.0404	0.0640	0.0184	0.0230
	SIBRAR	0.0480	0.0888	0.0207	0.0289	0.0434	0.0758	0.0190	0.0255	0.0264	0.0453	0.0110	0.0148	0.0548	0.0854	0.0220	0.0280
	DGMRec	**0.0897**	**0.1531**	**0.0404**	**0.0528**	**0.1024**	**0.1625**	**0.0462**	**0.0584**	**0.0725**	**0.1134**	**0.0324**	**0.0406**	**0.1093**	**0.1773**	**0.0476**	**0.0611**
	Improv.	7.43%	7.14%	9.49%	7.98%	8.13%	4.57%	8.71%	6.96%	8.21%	6.00%	7.64%	6.84%	4.69%	7.78%	5.31%	7.38%

Missing Modality + New Items Setting																	
Dataset		Baby				Sports				Clothing				TikTok			
Metric		R@20	R@50	N@20	N@50	R@20	R@50	N@20	N@50	R@20	R@50	N@20	N@50	R@20	R@50	N@20	N@50
CF	MF	0.0349	0.0583	0.0174	0.0228	0.0376	0.0598	0.0201	0.0253	0.0196	0.0288	0.0100	0.0120	0.0286	0.0464	0.0112	0.0148
	NGCF	0.0336	0.0599	0.0160	0.0220	0.0389	0.0648	0.0196	0.0255	0.0262	0.0418	0.0124	0.0158	0.0443	0.0737	0.0186	0.0246
	LightGCN	0.0434	0.0723	0.0218	0.0285	0.0458	0.0788	0.0240	0.0302	0.0290	0.0457	0.0147	0.0184	0.0527	0.0829	0.0245	0.0306
	SGL	0.0434	0.0682	0.0228	0.0285	0.0484	0.0788	0.0251	0.0316	0.0337	0.0505	0.0173	0.0210	0.0548	0.0775	0.0247	0.0293
	SimGCL	0.0391	0.0630	0.0208	0.0262	0.0475	0.0750	0.0242	0.0313	0.0307	0.0492	0.0159	0.0200	0.0550	0.0771	0.0245	0.0289
Multi-Modal Recommenders	VBPR	0.0347	0.0640	0.0177	0.0244	0.0393	0.0641	0.0200	0.0257	0.0265	0.0414	0.0133	0.0166	0.0244	0.0417	0.0118	0.0221
	MMGCN	0.0326	0.0596	0.0157	0.0218	0.0274	0.0489	0.0133	0.0182	0.0170	0.0308	0.0079	0.0110	0.0439	0.0599	0.0186	0.0218
	GRCN	0.0347	0.0621	0.0170	0.0233	0.0368	0.0606	0.0185	0.0239	0.0226	0.0379	0.0109	0.0143	0.0378	0.0661	0.0166	0.0223
	SLMRec	0.0434	0.0702	0.0223	0.0284	0.0477	0.0755	0.0245	0.0308	0.0344	0.0526	0.0176	0.0217	0.0548	0.0775	0.0247	0.0293
	BM3	0.0407	0.0717	0.0204	0.0274	0.0496	0.0796	0.0255	0.0324	0.0317	0.0496	0.0163	0.0202	0.0588	0.0869	0.0262	0.0318
	LATTICE	0.0423	0.0730	0.0213	0.0283	0.0432	0.0713	0.0218	0.0283	0.0344	0.0539	0.0173	0.0216	0.0444	0.0742	0.0209	0.0269
	MGCN	0.0446	0.0802	0.0230	0.0302	0.0478	0.0775	0.0240	0.0316	0.0358	0.0562	0.0182	0.0228	0.0357	0.0694	0.0140	0.0208
	LGMRec	0.0450	0.0772	0.0230	0.0303	0.0462	0.0742	0.0236	0.0300	0.0353	0.0557	0.0175	0.0221	0.0388	0.0632	0.0142	0.0191
	DAMRS	0.0455	0.0779	0.0229	0.0304	0.0480	0.0784	0.0248	0.0317	0.0380	0.0583	0.0192	0.0237	0.0598	0.0872	0.0267	0.0333
	GUME	0.0447	0.0795	0.0225	0.0304	0.0476	0.0776	0.0244	0.0313	0.0357	0.0563	0.0179	0.0224	0.0567	0.0918	0.0217	0.0289
MMA RSs	CI2MG	0.0415	0.0716	0.0210	0.0279	0.0437	0.0718	0.0226	0.0290	0.0294	0.0461	0.0149	0.0186	0.0427	0.0660	0.0188	0.0235
	MILK	0.0247	0.0429	0.0120	0.0162	0.0192	0.0323	0.0093	0.0123	0.0133	0.0226	0.0064	0.0085	0.0212	0.0332	0.0105	0.0129
	SIBRAR	0.0280	0.0495	0.0138	0.0188	0.0257	0.0435	0.0128	0.0169	0.0153	0.0259	0.0070	0.0094	0.0351	0.0527	0.0154	0.0190
	DGMRec	**0.0519**	**0.0876**	**0.0257**	**0.0336**	**0.0532**	**0.0845**	**0.0276**	**0.0348**	**0.0413**	**0.0631**	**0.0211**	**0.0260**	**0.0639**	**0.0973**	**0.0285**	**0.0353**
	Improv.	14.06%	9.22%	11.73%	10.53%	7.26%	6.16%	8.67%	7.41%	8.68%	8.23%	9.89%	9.70%	6.86%	5.99%	6.74%	6.00%

- **Traditional CF Models**: Matrix factorization methods (MFBPR [18]), GNN-based methods (NGCF [22] and LightGCN [9]), and contrastive learning methods (SGL [26] and SimGCL [30]).
- **Multi-modal Recommender Systems**[5]: Feature-based methods (VBPR [8], MMGCN [25], GRCN [24], SLMRec [19], BM3 [39] and LGMRec [6]), Graph-based methods (LATTICE [34] and DAMRS [29]) and Hybrid methods (MGCN [31] and GUME [11]).
- **Missing Modality-Aware Recommender Systems**: Robust learning methods that utilize invariant learning (MILK [1]) and single-branch network (SIBRAR [5]), and generation-based methods that utilize Optimal Transport and hypergraphs (CI2MG [12]).

Evaluation Metrics. We split each dataset into training, validation, and test sets with 8:1:1 following [6, 11, 29]. For evaluation, we employ the widely used Recall@K and NDCG@K metrics with K = 20 and 50. We denote Recall and NDCG as R and N, respectively.

Implementation We implement DGMRec and other baselines in PyTorch [16]. We adopt Adam [4] as optimizer. The embedding dimension is fixed to 64. For DGMRec, the hyperparameters λ_1 and λ_2 are tuned in {1.0, 1e-1, 1e-2, 1e-3}, and τ in contrastive learning and α for adjusting graph structure are tuned from [0,1] with an interval of 0.2. The interval of modality generation is fixed to 5 epochs. For convergence consideration, the early stopping and total epochs are fixed at 30 and 1,000, respectively.

[5]The missing modality feature is injected based on the NN-injection approach [13].

4.2 Performance Comparison

For comprehensive evaluations of DGMRec, we perform evaluations under various scenarios: overall performance (Sec 4.2.1), different missing modality levels (Sec 4.2.2), varying missing ratios (Sec 4.2.3), and cross-modal retrieval (Sec 4.2.4).

4.2.1 Overall Performance. In Table 2, we present the performance of DGMRec and other models under the **Missing Modality Setting**. Additionally, to simulate a more realistic and challenging scenario, we evaluate performance under the **Missing Modality and New Items Setting**, which combines missing modalities with new items.

We have the following observations: **1)** DGMRec consistently outperforms existing MRSs and MMA-RSs across all datasets in both settings. This highlights the benefit of the proposed generation-based approach to handle missing modalities compared with existing injection-based methods. By dynamically generating missing modality features, DGMRec achieves robust and superior performance compared to baseline methods. **2)** The performance gap between the best CF model and the best MRSs narrows when comparing the Missing Modality Setting with the combined Missing Modality and New Item Setting, while the improvement achieved by

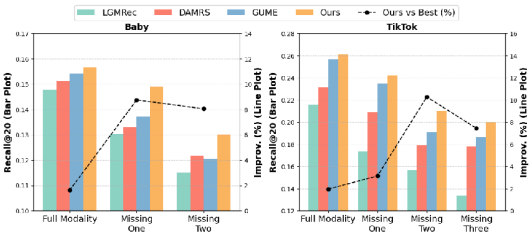

Figure 3: Performance on various missing levels on Amazon Baby and TikTok datasets.

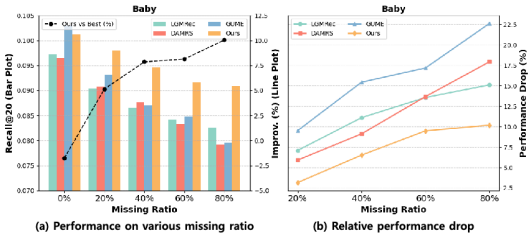

Figure 4: (a) Performance on various missing ratios, and (b) relative performance drop on Amazon Baby dataset.

DGMRec increases. This demonstrates the effectiveness of DGMRec's modality generation module in handling new items. **3)** The MMA-RSs based on contents, MILK and SIBRAR, exhibit lower performance compared to CF models. Additionally, CI2MG, which leverages CF knowledge by utilizing LightGCN as its backbone, performs worse than the vanilla LightGCN. This indicates that generating modality features without proper alignment (i.e., Optimal Transport) can lead to performance degradation.

4.2.2 Performance at Different Missing Modality Levels. In Figure 3, we evaluate the performance of DGMRec across different levels of missing modalities. The missing modality settings were consistent with those used in Section 4.2.1, and comparisons were made with LGMRec, DAMRS, and GUME. For better analysis, we reported the performance improvement of DGMRec relative to the best-performing baseline in a line plot. **1)** DGMRec consistently outperformed all baselines across all levels of missing modalities. This demonstrates that DGMRec not only obtains meaningful representations through disentangling modality features but also effectively generates missing modalities to enhance performance. **2)** Performance gains were greater when some modalities were available compared to cases with no modalities (i.e., missing one in Baby dataset and missing one & two in TikTok dataset). This highlights the importance of utilizing other available modalities to generate general features during the generation process. **3)** Even in scenarios where no modalities were available (i.e., missing two in Baby and missing three in TikTok), DGMRec still achieved significant performance improvements. This demonstrates that even in situations where general features cannot be generated due to the complete absence of available modalities, the specific features generated using the preference-based approach with interaction data remain effective.

4.2.3 Performance under Varying Missing Ratios. We conduct experiments with missing ratios set to 0% (No Missing Modality), 20%, 40%, 60%, and up to an extreme scenario of 80%. For each setting, items with missing modalities and the types of missing modalities were randomly selected, with selections kept consistent across experiments. For example, if an item's image modality was missing at 20%, the same item's image modality remained missing at 40%. In Figure 4(a), we present the results of DGMRec in comparison with LGMRec, DAMRS, and GUME on the Amazon Baby dataset using a bar plot. Additionally, the performance improvement of DGMRec relative to the best-performing baseline is shown in a line plot. **1)** Across all missing ratios, DGMRec consistently outperformed all baselines, except for the 0% baseline, where its performance was only slightly lower. Notably, as GUME's performance dropped significantly with increasing missing ratios, the performance gap between DGMRec and the other models widened substantially, reaching up to 10.05% at a missing ratio of 80%. This demonstrates DGMRec's robustness across a wide range of missing ratios, from low levels to extreme cases. To further confirm that DGMRec's robust performance is not solely attributed to its high performance at the 0% missing ratio (No Missing Modality), we also show the percentage performance drop at each missing ratio relative to the 0% baseline in figure 4(b). **2)** The performance drop of DGMRec was consistently smaller compared to other models , with 2.7% (3.2% vs. 5.9%) at 20% missing and 4.9% (10.2% vs. 15.1%) at a missing ratio of 80%. This highlights that DGMRec not only achieves high performance by effectively leveraging modality features through the Disentangling Modality Feature module but also demonstrates robust capability in addressing missing modalities via the Missing Modality Generation module.

Table 3: Results (Hit@10 and Hit@20) on Cross-Modal Retrieval of DGMRec. NN means the Nearest Neighbor method.

Datasets	Baby			
	Missing 1 Modality		Missing 2 Modalities	
	NN	DGMRec	NN	DGMRec
Hit@10	0.1344	0.3577	-	0.3496
Hit@20	0.1999	0.3801	-	0.3690

4.2.4 Cross-Modal Retrieval via Modality Generation. In this section, we demonstrate the potential of DGMRec's generation-based approach for real-world industry applications by demonstrating its ability to retrieve items with missing modalities—something that conventional MRSs cannot achieve. For comparisons, we employed a nearest-neighbor (NN) approach since existing MRSs are completely inapplicable for this task. Specifically, the NN approach retrieves the top-5 similar items based on available modalities while utilizing the mean of the existing modality features for retrieval.

Table 3 presents the retrieval performance for cases with one missing modality and scenarios with two missing modalities (i.e., all modalities missing). DGMRec significantly outperforms NN methods when a single modality is missing. Moreover, even in the challenging case where all modalities are missing, DGMRec achieves remarkable performance, whereas NN fails entirely.

We would like to emphasize that by leveraging generated modality features, DGMRec enables the retrieval of similar items and provides meaningful descriptions, even when modalities are missing during the item-streaming process. This highlights DGMRec's strong practical applicability in real-world and industrial scenarios, where delivering effective recommendations and descriptions despite incomplete modality data is crucial.

Table 4: Ablation studies on the components of DGMRec.

Datasets	Baby		Clothing		TikTok	
Metric	R@20	N@20	R@20	N@20	R@20	N@20
DGMRec	**0.0897**	**0.0404**	**0.0725**	**0.0324**	**0.1093**	**0.0476**
w/o Disentangle	0.0756	0.0331	0.0596	0.0268	0.0985	0.0419
w/o CLUB	0.0854	0.0373	0.0617	0.0277	0.1031	0.0452
w/o InfoNCE	0.0778	0.0347	0.0631	0.0280	0.1001	0.0429
w/o Generation	0.0848	0.0373	0.0646	0.0282	0.0988	0.0402
w/o Recon Loss	0.0872	0.0376	0.0703	0.0313	0.1034	0.0434
w/o Gen Loss	0.0862	0.0374	0.0647	0.0284	0.1041	0.0438
w/o Alignment	0.0554	0.0248	0.0392	0.0142	0.0745	0.0335
w/o UI-align	0.0789	0.0335	0.0634	0.0284	0.0903	0.0389
w/o BM-align	0.0811	0.0346	0.0576	0.0260	0.1011	0.0422

4.3 Model Analysis

4.3.1 Ablation Study. In Table 4, we conducted ablation studies to highlight the contribution of each component in DGMRec. In general, excluding any loss resulted in a performance decline across all datasets. More precisely, **1)** excluding $\mathcal{L}_{disentangle}$ led to a performance drop, demonstrating the effectiveness of disentangling modality features. This implies that modality disentanglement is effective for capturing modality representation. **2)** Removing generation-related losses such as $\mathcal{L}_{recon}$ and $\mathcal{L}_{gen}$ resulted in performance declines. These losses were introduced to effectively generate missing modalities, indicating that aligning reconstructed features with original features is crucial. **3)** The removal of $\mathcal{L}_{align}$ led to the most significant performance degradation. This emphasizes the critical role of aligning modality features with the CF knowledge. Alignment with the CF knowledge ensures that modality features contribute effectively to the recommendation, demonstrating its importance in DGMRec.

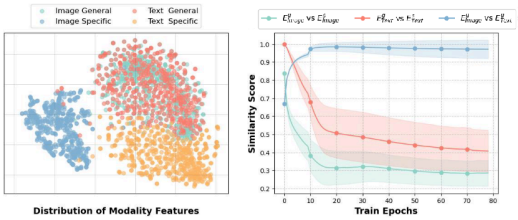

Figure 5: (a) Visualization of disentangled modality features and (b) similarity score between the features during training

4.3.2 Impact of Modality Disentanglement. To evaluate the effectiveness of the Disentangling Modality Feature module, we visualized the general and specific modality features of 500 randomly selected items from the Baby dataset using TSNE [20] and tracked their similarity scores during training to observe the disentanglement process. As shown in Figure 5(a), specific features are well-separated, while general features appear more entangled, reflecting their shared nature across modalities. Additionally, Figure 5(b) shows that the similarity scores between general and specific features within the same modality gradually decrease during training, while the similarity scores between general features across different modalities increase. These trends demonstrate the effectiveness of the disentanglement module in DGMRec.

Table 5: Time Complexity of compared methods.

Dataset	Baby			Sports			Clothing		
(sec)	Train	Inference	Total	Train	Inference	Total	Train	Inference	Total
LGMRec	2.81	3.39	171.5	7.01	6.54	603.2	6.58	7.20	559.1
DAMRS	3.56	3.33	238.7	14.99	6.93	1723.6	16.53	7.23	1934.2
GUME	2.74	2.94	152.1	9.45	7.32	662.1	9.29	7.96	667.2
DGMRec	2.63	3.38	210.4	8.83	6.18	666.5	9.91	7.32	802.4

4.3.3 Time Complexity Analysis. The missing modality generation process in DGMRec can incur additional computational overhead per specific epoch. To evaluate the complexity of this process, we compared three key metrics: **1)** the average time per training epoch (Train), **2)** the total inference time (Inference), and **3)** the overall training time throughout the entire process (Total) across three Amazon datasets in Table 5.

We have the following observations: **1)** For the relatively small Baby dataset, the time required by DGMRec is nearly identical to that of other models, demonstrating that the additional processes in DGMRec do not create noticeable inefficiencies in smaller datasets. **2)** In the larger Sports and Clothing datasets, while DGMRec incurs slightly higher overhead due to the modality generation and graph refinement processes, the difference remains modest. Considering the significant performance gains achieved by DGMRec, we argue that this additional time cost is reasonable, indicating that DGMRec remains competitive even in larger-scale datasets.

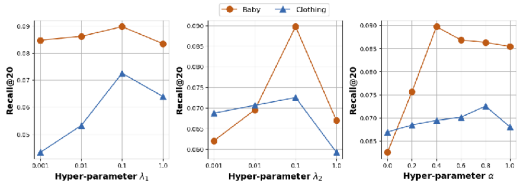

Figure 6: Performance on different hyper-parameters

4.3.4 Parameter Sensitivity. In Figure 6, we investigate the impact of *every* hyperparameter in DGMRec, i.e., λ_1 for $\mathcal{L}_{disentangle}$, λ_2 for $\mathcal{L}_{align}$, and α for graph refinement, on the Baby and Clothing datasets. **1)** The results show that both λ_1 and λ_2 achieve the best performance when set to 0.01, with significant performance drops observed when either value is too low or too high. Notably, λ_2, which directly affects performance through alignment loss, is found to be more sensitive than λ_1. **2)** For the balancing hyperparameter α, a value of 0.0—indicating no graph refinement—leads to the lowest performance, emphasizing the importance of the graph refinement process in DGMRec. Moreover, the best results are achieved with moderate α values, rather than extremes like 1.0, underscoring the need to maintain a balanced approach.

5 Conclusion

In this paper, we propose a novel model, DGMRec that addresses two key challenges: **1)** Missing modality scenarios are not sufficiently addressed, and **2)** Unique characteristics of modalities are overlooked. The core idea of DGMRec lies in effectively disentangling general and specific modality features, which are then utilized to generate missing modalities in a fine-grained manner. These two modules work synergistically in DGMRec framework. As a result, DGMRec achieves high performance in realistic settings thus demonstrating its strong potential for industrial applications, including information retrieval tasks and scenarios with extremely missing modalities.

Acknowledgment

This work was supported by Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2023-00216011, RS-2022-II220077), and National Research Foundation of Korea(NRF) funded by Ministry of Science and ICT (NRF-2022M3J6A1063021).

References

[1] Haoyue Bai, Le Wu, Min Hou, Miaomiao Cai, Zhuangzhuang He, Yuyang Zhou, Richang Hong, and Meng Wang. 2024. Multimodality invariant learning for multimedia-based new item recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 677–686.

[2] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. 2020. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*. PMLR, 1779–1788.

[3] Jae Won Cho, Dong-Jin Kim, Jinsoo Choi, Yunjae Jung, and In So Kweon. 2021. Dealing with missing modalities in the visual question answer-difference prediction task through knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1592–1601.

[4] P Kingma Diederik. 2014. Adam: A method for stochastic optimization. *(No Title)* (2014).

[5] Christian Ganhör, Marta Moscati, Anna Hausberger, Shah Nawaz, and Markus Schedl. 2024. A Multimodal Single-Branch Embedding Network for Recommendation in Cold-Start and Missing Modality Scenarios. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 380–390.

[6] 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*, Vol. 38. 8454–8462.

[7] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*. 507–517.

[8] Ruining He and Julian McAuley. 2016. VBPR: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 30.

[9] 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 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 639–648.

[10] Sein Kim, Hongseok Kang, Seungyoon Choi, Donghyun Kim, Minchul Yang, and Chanyoung Park. 2024. Large language models meet collaborative filtering: An efficient all-round llm-based recommender system. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1395–1406.

[11] Guojiao Lin, Meng Zhen, Dongjie Wang, Qingqing Long, Yuanchun Zhou, and Meng Xiao. 2024. GUME: Graphs and User Modalities Enhancement for Long-Tail Multimodal Recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 1400–1409.

[12] Zhenghong Lin, Yanchao Tan, Yunfei Zhan, Weiming Liu, Fan Wang, Chaochao Chen, Shiping Wang, and Carl Yang. 2023. Contrastive intra-and inter-modality generation for enhancing incomplete multimedia recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 6234–6242.

[13] Daniele Malitesta, Emanuele Rossi, Claudio Pomo, Tommaso Di Noia, and Fragkiskos D Malliaros. 2024. Do We Really Need to Drop Items with Missing Modalities in Multimodal Recommendation?. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 3943–3948.

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

[15] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[16] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[17] N Reimers. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint arXiv:1908.10084* (2019).

[18] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618* (2012).

[19] Zhulin Tao, Xiaohao Liu, Yewei Xia, Xiang Wang, Lifang Yang, Xianglin Huang, and Tat-Seng Chua. 2022. Self-supervised learning for multimedia recommendation. *IEEE Transactions on Multimedia* 25 (2022), 5107–5116.

[20] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[21] Cheng Wang, Mathias Niepert, and Hui Li. 2018. LRMM: Learning to recommend with missing modalities. *arXiv preprint arXiv:1808.06791* (2018).

[22] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural graph collaborative filtering. In *Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval*. 165–174.

[23] 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*. 790–800.

[24] 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 28th ACM international conference on multimedia*. 3541–3549.

[25] 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 27th ACM international conference on multimedia*. 1437–1445.

[26] 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*. 726–735.

[27] Le Wu, Yonghui Yang, Kun Zhang, Richang Hong, Yanjie Fu, and Meng Wang. 2020. Joint item recommendation and attribute inference: An adaptive graph convolutional network approach. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 679–688.

[28] Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. 2024. Deep multimodal learning with missing modality: A survey. *arXiv preprint arXiv:2409.07825* (2024).

[29] Guipeng Xv, Xinyu Li, Ruobing Xie, Chen Lin, Chong Liu, Feng Xia, Zhanhui Kang, and Leyu Lin. 2024. Improving Multi-modal Recommender Systems by Denoising and Aligning Multi-modal Content and User Feedback. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3645–3656.

[30] 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*. 1294–1303.

[31] 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*. 6576–6585.

[32] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 10790–10797.

[33] Zheng Yuan, Fajie Yuan, Yu Song, Youhua Li, Junchen Fu, Fei Yang, Yunzhu Pan, and Yongxin Ni. 2023. Where to go next for recommender systems? id-vs. modality-based recommender models revisited. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2639–2649.

[34] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. 2021. Mining latent structures for multimedia recommendation. In *Proceedings of the 29th ACM international conference on multimedia*. 3872–3880.

[35] 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* 35, 9 (2022), 9154–9167.

[36] Hongyu Zhou, Xin Zhou, Zhiwei Zeng, Lingzi Zhang, and Zhiqi Shen. 2023. A comprehensive survey on multimodal recommender systems: Taxonomy, evaluation, and future directions. *arXiv preprint arXiv:2302.04473* (2023).

[37] Xin Zhou. 2023. Mmrec: Simplifying multimodal recommendation. In *Proceedings of the 5th ACM International Conference on Multimedia in Asia Workshops*. 1–2.

[38] 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*. 935–943.

[39] Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. 2023. Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM Web Conference 2023*. 845–854.

[40] Lipeng Zhu and David J Hill. 2021. Data/model jointly driven high-quality case generation for power system dynamic stability assessment. *IEEE Transactions on Industrial Informatics* 18, 8 (2021), 5055–5066.