Multimodal Conditioned Diffusion Model for Recommendation

Haokai Ma[#]
School of Software, Shandong University
Jinan, China
mahaokai@mail.sdu.edu.cn

Yimeng Yang[#]
School of Software, Shandong University
Jinan, China
y_yimeng@mail.sdu.edu.cn

Lei Meng[*]
Shandong Research Institute of Industrial Technology;
School of Software, Shandong University
Jinan, China
lmeng@sdu.edu.cn

Ruobing Xie
WeChat, Tencent
Beijing, China
ruobingxie@tencent.com

Xiangxu Meng
School of Software, Shandong University
Jinan, China
mxx@sdu.edu.cn

ABSTRACT

Multimodal recommendation aims at to modeling the feature distributions of items by using their multi-modal information. Prior efforts typically focus on the denoising of the user-item graph with a degree-sensitive strategy, which may not well-handle the users' consistent preference across modalities. More importantly, it has been observed that existing methods may learn ill-posed item embeddings due to their focus on a specific auxiliary optimization task for multimodal representations rather than explicitly modeling them. This paper therefore presents a solution that takes the advantages of the explicit uncertainty injection ability of Diffusion Model (DM) for the modeling and fusion of multi-modal information. Specifically, we propose a novel Multimodal Conditioned Diffusion Model for Recommendation (MCDRec), which tailors DM with two technical modules to model the high-order multimodal knowledge. The first module is multimodal-conditioned representation diffusion (MRD), which integrates pre-extracted multimodal knowledge into the item representation modeling via a tailored DM. This smoothly bridges the insurmountable gap between the multimodal content features and the collaborative signals. Secondly, with the diffusion-guided graph denoising (DGD) module, MCDRec may effectively denoise the user-item graph by filtering the occasional interactions in user historical behaviors. This is achieved with the power of DM in aligning the users' collaborative preferences with their shared items' content information. Extensive experiments compared to several SOTA baselines on two real-word datasets demonstrate the effectiveness of MCDRec. The specific visualization also reveals the potential of MRD to precisely handling the high-order representation correlations among the user embeddings and the multi-modal heterogeneous representations of items.

CCS CONCEPTS

• **Information systems → Recommender systems**.

KEYWORDS

Recommender System, Multimodal Recommendation, Diffusion Model, Graph Structure Learning

ACM Reference Format:
Haokai Ma[#], Yimeng Yang[#], Lei Meng[*], Ruobing Xie, and Xiangxu Meng. 2024. Multimodal Conditioned Diffusion Model for Recommendation. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3589335.3651956

1 INTRODUCTION

Recommender system (RS) aims to predict the appropriate items by modeling users' historical behaviors [14, 18]. However, the ubiquity of the extensive corpus and the Matthew effect inevitably engenders the sparsity issue in real-world RSs [17]. As a straightforward method, multimodal recommendation is proposed to use the multi-modal information to enhance the item representation modeling. Its focal issue lies in how to mitigate the bias between the pre-extracted multi-modal features and collaborative signals to smoothly facilitate its adaptation to recommendation tasks. [15, 19]

Inspired by the notable achievement of graph representation learning, recent studies integrate the multi-modal information into user-item interaction graph to enhance recommendation [25, 29]. SLMRec [25] and BM3 [35] introduce self-supervised multi-modal signals to capture the content consistency among multiple modalities. LATTICE [33] constructs the item-item correlation graph for each modality and dynamically updates it to capture the high-quality item representation. FREEDOM [34] further freezes the item-item graph and designs a simple degree-sensitive denoising strategy for efficient recommendation. Recently, LD4MRec [31] attempts to leverage the notable generative capacity of diffusion models (DM) to generate interaction probabilities with the guidance

[*] indicates corresponding author.
[#] indicates equal contribution.

WWW '24 Companion, May 13–17, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0172-6/24/05
https://doi.org/10.1145/3589335.3651956

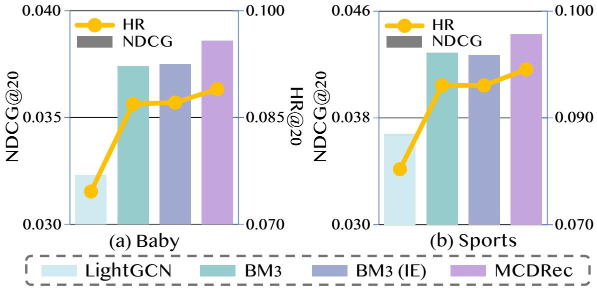

Figure 1: BM3 (IE) (replacing multi-modal knowledge with the pre-trained item embedding in LightGCN) achieves the comparable performance with the original BM3 on Baby and Sports, which verifies that the multi-modal information in BM3 is underutilized. In contrast, our MCDRec explicitly integrates the multi-modal representations in the continuous space through a tailored DM based on BM3, thereby achieving more effective utilization of multi-modal information.

of multi-modal knowledge, which follows the prevailing DM-based recommenders [16, 28] to conduct DM on discrete item indices.

To investigate whether existing multimodal recommenders fully leverage the multi-modal knowledge, we leverage the pre-trained item embeddings from LightGCN [6] to replace the multi-modal representation in BM3 [35] to define the BM3 (IE). The performance comparisons between LightGCN, BM3 and BM3 (IE) on Baby and Sports are illustrated in Fig. 1. This demonstrates that BM3 without multi-modal knowledge is able to yield a comparable performance with the original BM3, that is, the multi-modal information embedded in BM3 [35] remains underutilized. Moreover, we conduct the empirical analysis that the discrete DM in LD4MRec [31] encounters challenges in attaining optimal multimodal recommendation performance, which is primarily due to the insurmountable gap between continuous multimodal representations and discrete item indices. The informational processing misalignment between the multimodal recommendation tasks and the pure index-driven recommendation tasks (e.g., SR) also brings suboptimal performance when indiscriminate implementing existing DM-based recommenders to handle multi-modal representations. These discoveries prompts us to some questions: (1) Is there concrete evidence that multi-modal information indeed plays a role in multimodal recommendation tasks? (2) Could advanced diffusion models explicitly integrate multi-modal knowledge into the collaborative signals to support multimodal recommendation tasks?

To answer these questions, we began to explore the feasibility of incorporating the uncertainty injection [7] and the multi-modal alignment guidance [1] of DM into the multimodal recommendation. We propose an effective **M**ultimodal **C**onditioned **D**iffusion Model for **Rec**ommendation (MCDRec), which serves as the model-agnostic framework to jointly model the multi-modal guidance and the diffused guidance in elevating the existing multimodal recommenders. Specifically, MCDRec first designs the Multimodal-conditioned Representation Diffusion (MRD) module to gradually inject the modality-aware uncertainties into item representations via a tailored DM structure, thus explicitly integrating the multi-modal information and smoothing the significant deviation between modal-aware features and collaborative signals to improve the item representation learning. In order to filter the inherent noise in user behaviors and preserve the user's modality-aware preferences, MCDRec additionally proposes a Diffusion-guided Graph Denoising (DGD) module, which fully leverages the diffusion-aware item representations from MRD to accurately denoise the user-item graph. With such two effective modules, MCDRec can seamlessly combine the tailored DM with existing multimodal recommenders, thereby ensuring the generalization capacity and high-dimensional distribution fitting capacity of DM are comprehensively exploited in multimodal recommendation from start to finish.

We have conducted extensive experiments on two real-world datasets to demonstrate the effectiveness and universality of MCDRec. Through the ablation study and visualization analysis, we also verify the validity of each module in our MCDRec. To summarise, MCDRec's contribution can be concluded as follows:

- We propose a multimodal recommendation framework, MCDRec, which jointly models the multi-modal guidance and the diffused guidance to enhance multimodal recommenders. To the best of our knowledge, this is a pioneering solution that models the continuous representation via DM in multimodal recommendation.
- We develop two effective and model-agnostic MRD and DGD modules to incorporate multimodal-guided diffusion knowledge at each phase of multimodal recommendation, enabling the seamless integration of the DM with multimodal recommendation.
- We conduct visualization analysis to uncover the superiority of MCDRec in precisely handling the correlations among heterogeneous representations of users and items, which enables a comprehensive understanding of MCDRec.

2 RELATED WORK

2.1 Multimodal Recommendation

Multimodal recommendation aims to incorporate the multi-modal information of items into the representation learning, thereby alleviating the data sparsity issue in recommendation. Early studies [5] typically inject the pre-extracted visual features of each item from Convolution Neural Network (CNN) into the original index embedding to capture the visually-aware item representations. Inspired by the success of Graph Convolutional Networks (GNNs) in recommendation, recent researchers start to leverage the graph structure to handle the multi-modal information. MMGCN [29] first incorporates it to build the user-item bipartite graph for specific modalities. BM3 [35] designs a multi-modal contrastive task to bootstrap latent representations towards to overcome the computational cost and noisy supervisory signaling issues. To precisely capture the implicit item representation, LATTICE [33] designs the item-item relation graphs for each modality to learn modality-aware structure knowledge. FREEDOM [34] freezes the item-item graphs and designs a structure denoising module on the basis of LATTICE [33] for efficient recommendation. Recently, LD4MRec [31] conducts DM on discrete item indices to generate user behaviors with the guidance of multi-modal representation and collaborative signals.

However, the challenges in existing multimodal recommendation algorithms are two-fold: Firstly, the experiments in Table. 1 have verified that certain multimodal recommenders fail to fully utilize the multi-modal information, which is the fundamental basis of multimodal recommendation. Secondly, although LD4MRec [31] has investigated the efficacy of DM in multimodal recommendation, it solely leverages the continuous multi-modal representations for predicting discrete interaction probabilities. The inherent bias between these two aspects makes it challenging to achieve optimal performance under the standard training setting (cf. Table 2 in [31]).

2.2 Diffusion Models in Recommendation

Motivated by the uncertainty injection and data augmentation ability of Diffusion Models (DM) in image synthesis [22], text generation [10], and machine translation [32], some studies have explored the effectiveness of DM in recommendation. For instance, DiffRec [28] gradually generates globally similar but personalized collaborative information via DM in the denoising process. PDRec [16] creatively proposes three plug-in modules to fully leverage the diffusion-based preferences on all items to improve SR models. LD4MRec [31] employs DM on discrete item indices with the guidance of continuous multi-modal representations. Different from these above studies that conduct DM on the discrete item indices, several DM-based recommenders employ that on the continuous item embedding space to enhance representation learning in SR. DiffuRec [11] regards the various aspects of items and the multiple user intentions as distributions to fully exploit the inherent distribution generation capability of DM. DreamRec [30] utilizes DM to explore the underlying distribution of item space and generate the oracle items with the guidance of users' sequential behaviors.

Although these DM-based recommenders have demonstrated promising performance on discrete item indices and continuous item representations, they are not explicitly designed to handle multi-modal information. Therefore, mechanically applying them to multimodal recommendation tasks may result in a decline in its performance. Furthermore, with the dependency of multimodal recommendation tasks on continuous multi-modal representations, we argue that employing continuous multi-modal representations as conditions to guide discrete DM is not the optimal choice. Instead, unifying them into the continuous feature space is deemed more suitable for the problem setting of multimodal recommendation.

3 METHODS

3.1 Problem Formulation

The objective of multimodal recommendation resides in utilizing the additional multi-modal information of items to obtain more accurate item representations in recommendation. In this work, we first define $\boldsymbol{e}_u^u \in \mathbb{R}^d$ and $\boldsymbol{e}_i^i \in \mathbb{R}^d$ as the user embedding and item embedding of the user $u \in \mathcal{U}$ and the item $i \in \mathcal{I}$ respectively. Here, d denotes the embedding dimension, $\mathcal{U}$ and $\mathcal{I}$ denote the set of users and items. To incorporate the multi-modal knowledge, we leverage v and t to represent the visual and textual modalities. With this specific modality $m \in \{v, t\}$, we represent the corresponding modality feature as $\boldsymbol{e}^m \in \mathbb{R}^{d_m}$ where d_m denotes its dimension. Given these three types of latent representation $\boldsymbol{e}_i^i$, $\boldsymbol{e}_i^v$ and $\boldsymbol{e}_i^t$ of the same item i [1], we can improve the item representation learning in multimodal recommendation to predict the preference scores of a specific user on each item.

3.2 Overall Framework

In this section, we detail our proposed Multimodal Conditioned Diffusion Model for Recommendation (MCDRec), which incorporates multi-modal information into the item representation modeling process through a tailored DM and leverages the diffusion-aware knowledge into the user-item interaction graph for accurate denoising. As shown in Fig. 2, MCDRec consists of two main components, including multimodal-conditioned representation diffusion (MRD) and diffusion-guided graph denoising (DGD). Specifically, MCDRec first proposes MRD to guide the more precise item representations learning with the condition as multi-modal features, aiming to incorporate the modality-specific uncertainty. In order to achieve accurate denoising on the interaction graph, MCDRec designs a DGD strategy that leverage the diffusion-aware knowledge from MRD to identify real noise edges and smoothly prune them to consistently maintain the relative noise-free interaction graph.

3.3 Base multimodal recommender

Inspired by the success in graph representation learning in recommendation, we adopt BM3 [35] as the base multimodal recommender in this work. Given the modality-specific pre-extracted feature $\boldsymbol{e}_i^m \in \mathbb{R}^{d_m}$ of item i, we first conduct a Multilayer Perceptron (MLP) to get the hidden modality-aware representation $\boldsymbol{h}_m$ for the modality m as following:

$$\boldsymbol{h}_i^m = \mathrm{MLP}^m(\boldsymbol{e}_i^m) = \boldsymbol{e}_i^m \mathbf{W}_m + \boldsymbol{b}_m \tag{1}$$

where $\mathbf{W}_m \in \mathbb{R}^{d_m \times d}$, $\boldsymbol{b}_m \in \mathbb{R}^d$ denote the weight matrix and the bias vector respectively.

Then, we leverage LightGCN [6] to encode the user-item interaction graph as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \mathcal{U} \cup \mathcal{I}$ and $\mathcal{E}$ denote the nodes and edges of this graph. Besides, $\mathrm{A} = \begin{pmatrix} 0 & \mathrm{R} \\ \mathrm{R}^\top & 0 \end{pmatrix} \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$ and D are leveraged to represent the adjacency matrix and the diagonal degree matrix of $\mathcal{G}$, where R is the user-item interaction matrix. We also denote the initial index embeddings of all the nodes in graph $\mathcal{G}$ as $\mathrm{H}^0 = [\boldsymbol{e}_1^u, \boldsymbol{e}_2^u, \cdots, \boldsymbol{e}_{|\mathcal{U}|}^u, \boldsymbol{e}_1^i, \boldsymbol{e}_2^i, \cdots, \boldsymbol{e}_{|\mathcal{I}|}^i]$. With the feed-forward propagation, we can obtain the hidden index embeddings H^{l+1} of the $(l+1)$-th layer from the the hidden index embeddings H^l of the l-th layer as following:

$$\mathrm{H}^{l+1} = \hat{\mathrm{A}}\mathrm{H}^l = \left(\mathrm{D}^{-1/2}\mathrm{A}\mathrm{D}^{-1/2}\right)\mathrm{H}^l \tag{2}$$

To generate the ultimate node representations of users and items for the consequent recommendation phase, we also utilize a readout function to aggregate all representations from all the hidden layers. Following [33, 35], we also conduct a residual connection to incorporate the initial item embeddings H_i^0 into the final representation for items $\mathrm{H_i}$. This process can be succinctly expressed as follows:

$$\begin{aligned} \mathrm{H_u} &= \mathrm{READOUT}\left(\mathrm{H}_u^0, \mathrm{H}_u^1, \mathrm{H}_u^2, \ldots, \mathrm{H}_u^L\right) \\ \mathrm{H_i} &= \mathrm{READOUT}\left(\mathrm{H}_i^0, \mathrm{H}_i^1, \mathrm{H}_i^2, \ldots, \mathrm{H}_i^L\right) + \mathrm{H}_i^0 \end{aligned} \tag{3}$$

where L denotes the number of convolutional layers.

[1] For brevity, we omit the subscript i in $\boldsymbol{e}_i^i$, $\boldsymbol{e}_i^v$ and $\boldsymbol{e}_i^t$ and the subscript u in $\boldsymbol{e}_u^u$ in the following sections.

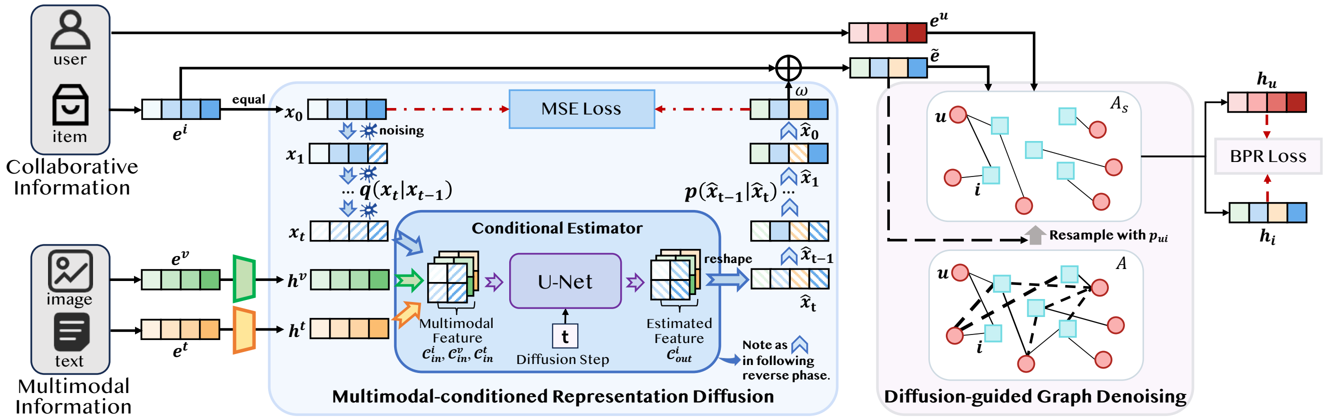

Figure 2: The overall structure of MCDRec. MRD explicitly incorporates the multi-modal knowledge into item representation modeling via the tailored DM while DGD strategy leverages the diffusion-aware user preference consistency on multiple modalities for better user behaviors denoising.

3.4 Multimodal-conditioned Representation Diffusion

In this section, we will detail our Multimodal-conditioned Representation Diffusion (MRD) based on the Denoising Diffusion Probabilistic Models (DDPM) framework [7]. Inspired by the promising success of DM in Computer Vision (CV) and Natural Language Processing (NLP), some researchers have started to explore DM in recommendation. However, these DM-based recommendation models typically employ the tailored diffusion model on the direct item indices [16, 28] or the continuous item embeddings [11, 30], overlooking the multi-modal knowledge modeling. Different from these prior algorithms, we attempt to incorporate multi-modal information as the conditions into the diffusion process, with the aim of guiding the generation of item representations. Similar to the classical DM-based recommendation algorithms, the proposed MRD consists of two processes: the forward process and the reverse process. The forward process involves the gradual addition of the Gaussian noise to perturb the original data distribution. In contrast, the reverse process gradually recovers the perturbed representation from the disorder-state to the representation space.

3.4.1 Forward Process. Without loss of generality, we firstly denote x_0 as the initial index embedding e^i for item $i \in \mathcal{I}$ in MRD. The forward process constitutes a Markov Chain that Gaussian noise is gradually incorporated into x_0 as follows:

$$\begin{aligned} q\left(x_{1:T} \mid x_0\right) &:= \prod_{t=1}^{T} q\left(x_t \mid x_{t-1}\right) \\ q\left(x_t \mid x_{t-1}\right) &:= \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I\right) \end{aligned} \tag{4}$$

where t denotes the diffusion steps and $\beta_t \in (0,1)$ denotes the added Gaussian noise scale. Utilizing the reparameterization trick [13], we can aggregate the noising process at each step. Consequently, the forward process can be formulated as:

$$q\left(x_t \mid x_0\right) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}x_t, \left(1-\bar{\alpha}_t\right)\mathbf{I}\right) \tag{5}$$

Let $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\alpha_t = 1-\beta_t$ and $\bar{\alpha}_t = \prod_{t'=1}^{t} \alpha_{t'}$, then we can acquire that $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$.

3.4.2 Reverse Process. As the core phase of the diffusion model, the reverse process aims to iteratively denoise x_t for t steps to ultimately approximate the initial item representation x_0. Notably, we incorporate the multi-modal knowledge of items as the additional conditions to guide the conditional generation of item representations via a tailored conditional estimator. The conditional generation process at each step t can be formulated as:

$$\begin{aligned} p_\theta\left(x_{0:T}\right) &= p\left(x_T\right) \prod_{t=1}^{T} p_\theta\left(x_{t-1} \mid x_t\right) \\ p_\theta\left(x_{t-1} \mid x_t\right) &= \mathcal{N}\left(x_t; \mu_\theta\left(x_t, t, h_i^v, h_i^t\right), \Sigma_\theta\left(x_t, t\right)\right) \end{aligned} \tag{6}$$

where $\Sigma_\theta\left(\mathrm{x}_t, t\right) = \sigma_t^2 \mathrm{I} = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \mathrm{I}$ denotes the variance and $h_i^v, h_i^t \in \mathbb{R}^d$ are visual and textual latent representation of the same item i respectively. Here, the mean $\mu_\theta\left(x_t, t, h_i^v, h_i^t\right)$ can be calculated by:

$$\mu_\theta\left(x_t, t, h_i^v, h_i^t\right) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} f_\theta\left(x_t, t, h_i^v, h_i^t\right)\right) \tag{7}$$

The conditional estimator $f_\theta(.)$ is typically constructed by many deep neural networks (e.g., MLP [28], Transformer [11] or U-Net [12]) to generate the estimated item representation with the multimodal informative knowledge. After t steps, we could obtain the predicted $\hat{x}_0$ as x_p. To preserve the personalized features of items while introducing diffused multi-modal information, we denote $\widetilde{e} = x_0 + \omega \cdot x_p$ as the final item representation for the subsequent recommendation task, where ω is the pre-defined diffused weight.

3.4.3 Conditional Estimator. In this section, we will introduce our designed conditional estimator, a tailored model architecture on U-Net, to adapt to multimodal-guided diffusion. U-Net [24] is widely utilized in diffusion models to enhance various CV and NLP tasks, including image synthesis [7], image super-resolution [9] and semantic segmentation [2]. It is principally ascribed to the heightened proficiency exhibited by U-Net in adeptly handling high-dimensional distributions. Therefore, directly incorporating the raw structure of U-Net into a 1-dimensional vector may result in sub-optimal performance.

To tackle this issue, we regard the estimated item representation x_p, the visual representation h_i^v, and the textual representation

$\boldsymbol{h}_i^t \in \mathbb{R}^d$ as the separate channels in U-Net to capture informative knowledge from diverse item modalities, as channels are independently modeled with the convolutional network. To fully leverage the advantage of receptive-field in convolutional networks, we reshape these representations as channels matrices C_{in}^i, C_{in}^v and $\mathrm{C}_{in}^t \in \mathbb{R}^{\sqrt{d}\times\sqrt{d}}$ and subsequently fed them into U-Net. Moreover, we also add the step representation $\boldsymbol{t}_i$ in conjunction with above channels matrices to each convolutional neural network block, thus facilitating the fusion of step information and item representation. Here, $\boldsymbol{t}_i$ is generated from the scalar diffusion step t by sinusoidal embedding technique. With the forward propagation of U-Net, we can obtain the estimated channel matrices C_{out}^i, C_{out}^v and $\mathrm{C}_{out}^t \in \mathbb{R}^{\sqrt{d}\times\sqrt{d}}$ and recover the C_{out}^i back to the estimated representation $\hat{x}_0 \in \mathbb{R}^d$. After such operations, the diffusion step t and multi-modal embeddings $\boldsymbol{h}_i^v$ and $\boldsymbol{h}_i^t$ can simultaneously used to precisely condition the estimation of $\hat{x}_0$.

3.4.4 Optimization. Similar to the optimization of the underlying data generation distribution in other generation tasks, DM also compels the posterior distribution $q\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0\right)$ closer to the prior distribution $p_\theta\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right)$ during the reverse process. This optimization function is expressed in the form of KL divergence:

$$\mathcal{L}_{vlb} = D_{KL}\left(q\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0\right) \| p_\theta\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right)\right) \tag{8}$$

Thanks to the DDPM framework [7], it can be easily simplified to a Mean-Squared Error (MSE) loss as follows:

$$\mathcal{L}_{dm} = E_{\boldsymbol{x}_0,\boldsymbol{x}_t}\left[\left\|x_0 - f_\theta\left(x_t, t, \boldsymbol{h}_i^v, \boldsymbol{h}_i^t\right)\right\|^2\right] \tag{9}$$

Here, x_0 denotes the initial item embedding, and f_θ refers to the above conditional estimator to generate the estimated item representation $\hat{x_0}$.

3.5 Diffusion-guided Graph Denoising

Previous graph-based recommendation studies have verified that the occasional interactions in user behavioral sequence introduce an inescapable semblance of noise into the user-item interaction graph [3]. Graph structures are confronted with the issue of nodes with higher popularity being more susceptible to over-smoothing. Most existing graph denoising techniques randomly discard edges on the graph at a certain percentage throughout the training process, which are stochastic and potentially destroy the graph correlations. Intuitively, the incorporation of multi-modal knowledge in recommendation contributes substantively to user preference modeling and item representation modeling. This assertion forms the foundational underpinning of multimodal recommendation tasks. This motivates us to investigate how to harness the diffused item representations obtained in MRD to refine the extant graph denoising strategies. To the end, we introduce a Diffusion-guided Graph Denoising (DGD) strategy, which identifies the authentic noised edges and smoothly prunes them by incorporating of the diffusion-aware knowledge from MRD.

Specifically, given the user-item interaction graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, and the specific edge e_{ui} in $\mathcal{G}$, we can obtain the connected nodes u and i, and their corresponding representations as $\boldsymbol{e}_u^u$ and $\boldsymbol{e}_i^i$. With the MRD module mentioned in Sec. 3.4, we inject the multi-modal knowledge into the original item representation $\boldsymbol{e}_i^i$ to obtain the multimodal-enhanced diffused representation $\widetilde{\boldsymbol{e}}_i$ of item i. Then we compute the score $s_{ui} = \boldsymbol{e}_u^\top \widetilde{\boldsymbol{e}}_i$ between the user embedding $\boldsymbol{e}_u$ and the multimodal-enhanced diffused representation $\widetilde{\boldsymbol{e}}_i$ as the diffusion-aware interaction probability. For each edge e_{ui} in $\mathcal{G}$, we can update its weight as $(1 + \tau \cdot s_{ui})$ where τ denotes the pre-defined score weight and then re-calculate the degrees of the connected nodes u and i as d_u and d_i. Finally, we sample a denoised sub-graph $\widetilde{\mathcal{G}}_s$ from the original user-item interaction graph $\mathcal{G}$ with the probability $p_{ui} = \frac{1}{\sqrt{d_u}\sqrt{d_i}}$ of each edge e_{ui} in $\mathcal{G}$ until the number of edges in $\widetilde{\mathcal{G}}_s$ reach $|\mathcal{E}|(1-\rho)$, where ρ denotes the dropout proportion. Thus, we can generate the new symmetric adjacency matrix A_s based on $\widetilde{\mathcal{G}}_s$ to achieve the graph denoising at the beginning of each training epoch, and normalize it as: $\widehat{\mathrm{A}}_s = \mathrm{D}^{-1/2}\mathrm{A}_s\mathrm{D}^{-1/2}$. Following the classical graph denoising method [23], we only conduct DGD in the training phase while utilizing the original normalized matrix A as the adjacency matrix in the inference phase to implement the graph representation modeling. This further enables the low computational complexity of MCDRec when serving online.

Table 1: Statistics of two real-world multimodal datasets

Dataset	#Users	#Items	#Interactions	Sparsity
Baby	19,445	7,050	160,792	99.88%
Sports	35,598	18,357	296,337	99.95%

3.6 Optimization Objectives

With the final item representation $\widetilde{x}_i$ obtained in MRD module and the user embedding $\boldsymbol{e}_u$, we feed them into the DGD module to generate the $\boldsymbol{h}_u$ and $\boldsymbol{h}_i$ as the final representation. Following the classical multimodal recommendation algorithms [33, 34], we opt the Bayesian personalized ranking (BPR) loss [21] for each user-item triplet (u, i, j) in training set $\mathcal{R}$ to force the score between user u and positive item i is higher than that of negative item j:

$$\mathcal{L}_{bpr} = \sum_{(u,i,j)\in\mathcal{R}} \left(-\log\sigma\left(\boldsymbol{h}_u^\top \boldsymbol{h}_i - \boldsymbol{h}_u^\top \boldsymbol{h}_j\right)\right) \tag{10}$$

where $\sigma(\cdot)$ is the sigmoid function. The overall objective function $\mathcal{L}$ can be formulated as a linear combination of $\mathcal{L}_{bpr}$ and $\mathcal{L}_{dm}$:

$$\mathcal{L} = \mathcal{L}_{bpr} + \lambda \cdot \mathcal{L}_{dm} \tag{11}$$

where λ is used to control the weight of $\mathcal{L}_{dm}$.

4 EXPERIMENTS

In this section, we perform extensive experiments to answer the following research questions:

- **RQ1:** How does MCDRec perform against the general CF methods and the SOTA multimodal recommendation methods?
- **RQ2:** How do different components of MCDRec benefit its recommendation performance?
- **RQ3:** How does MRD affect the distribution of user embeddings and multi-modal item representations?

4.1 Datasets

Following previous works [34, 35], we conduct comprehensive experiments on *Baby* and *Sports* from the Amazon platform. The dataset includes both visual and textual features, and each review rating is considered a record of a positive user-item interaction. The raw data of each dataset are pre-processed with a 5-core setting on both items and users, which have been widely used in [5, 33, 35], and the results are presented in Table. 1. Referring to [34], we directly utilize the pre-extracted visual features with the dimension as 4096. For textual features, we use the sentence-transformers [20] to obtain 384-dimensional sentence embeddings. Moreover, the metric of data sparsity is computed through the division of the total interactions by the product of the number of items and users.

4.2 Baselines

To demonstrate the effectiveness of the proposed method, we compare MCDRec with the following baseline models.

The first category consists of two general CF-based recommenders which recommend personalized items to users based only on their interactions with the items:

- **BPR** [21] enhances the latent representations of users and items within the matrix factorization (MF) framework, employing a BPR loss for optimization.
- **LightGCN** [6] simplifies GCN by dismissing the feature transformation and nonlinear activation and using the hidden layer embeddings for prediction.

While the second class of work consists of six multi-modal recommenders that utilize multi-modal information of items for recommendation:

- **MMGCN** [29] represent each modality of items individually by GCN, and integrate these modal representations.
- **SLMRec** [25] adopts a self-supervised learning strategy, introduces innovative data augmentation techniques, and utilizes a contrastive learning loss.
- **DualGNN** [27] build the user co-occurrence graph to draw the user's attention to various modalities
- **LATTICE** [33] performs GCN on both user-item graph and item-item graph to learn latent representations.
- **BM3** [35] proposes a multi-modal contrastive task to bootstrap latent representations by designing inter-modality and intra-modality contrastive losses.
- **FREEDOM** [34] freezes the dynamic item-item graph in LATTICE and proposes a degree-sensitive denoising strategy to denoise the user-item interaction graph.

4.3 Experimental settings

We implement the above methods with PyTorch 1.12.0 and Python 3.8.10. Following the classical works [33–35], we set the embedding size of both users and items to 64 for all models. Moreover, we initialize the embedding parameters using the Xavier method [4], and optimize all models with the Adam [8] optimizer with the learning rate as 0.001. To ensure a fair comparison, we carefully adjust the parameters of each model with their published papers. We perform a comprehensive grid search to select the optimal universal hyper-parameters. To be specific, the number of GCN layers is set to 2, and we fix the hyperparameter λ at $1e-3$. As for the diffusion process, the steps t is tuned in $\{5, 10, 20, 40, 100\}$. The dropout rate ρ of diffusion-guided graph denoising is serched from $\{0.1, 0.3, 0.5, 0.8\}$. Respectively, the diffused weight ω and the score weight τ are chosen from $\{0.05, 0.1, 0.3, 0.5, 0.8, 1.0\}$. Following [33], we opt the early stopping strategy with 20 epochs and the total epochs are set to 1000, while Recall@20 is the stopping indicator.

4.4 Performance Comparison (RQ1)

To verify the effectiveness of MCDRec, we conduct our experiments on two real-world datasets to compare our MCDRec with various CF-based recommenders and multimodal recommenders. As illustrated in Table. 2, we use NDCG@k (N@k) and Recall@k (R@k) with k in $\{5, 10, 20\}$ as evaluation metrics and highlight the best results in boldface. From this table, we can observe that:

(1) MCDRec significantly outperforms all the baselines across all metrics and datasets. The improvements are larger with smaller k, which is natural that the precise utilization of multi-modal knowledge in MCDRec is more beneficial in mining users' authentic preference for the top positions.

(2) Comparing the recommendation performance between two types of recommenders, most multimodal recommenders exhibit superiority over CF-based recommenders across diverse datasets. MCDRec achieves further improvements over the state-of-the-art multimodal recommenders, which indicates that MCDRec can fully harness the uncertainty injection ability of DM to explicitly incorporate multi-modal information into the item representation modeling, thereby enhancing the representation learning in existing multimodal recommenders.

(3) Upon analyzing the improvements on various base models, MCDRec achieves the most substantial improvement over BM3 (without any graph denoise strategies). Furthermore, when integrated with the state-of-the-art baseline FREEDOM (with the degree-sensitive denoising strategy), MCDRec obtains peak results across all datasets. This highlights the ability of MCDRec to guide the graph denoising towards the direction of user consistent preference modeling on the multi-modal content level of items.

4.5 Ablation Study (RQ2)

In this section, we conduct an ablation study to explore the effectiveness of different components within MCDRec. Here, "BM3+MK" denotes merely leveraging the fused multi-modal item representation $\mathrm{MLP}(\boldsymbol{e}^v|\boldsymbol{e}^t|\boldsymbol{e}^i)$ as the final item representation $\widetilde{\boldsymbol{e}}$ in BM3. It is notable that MCDRec (BM3) is equivalent to BM3+DGD+MRD. By comparing BM3, BM3+MK, BM3+MRD, BM3+DGD and MCDRec (BM3), we can verify the benefits of our proposed MRD and DGD. From Table. 3, we observed that:

(1) In general, BM3+MK performs worse than BM3, demonstrating that the simple fusion of multi-modal information and collaborative signals may introduce additional bias, hindering accurate item representation modeling.

(2) Comparing BM3+MRD with BM3+MK on two real-world datasets, we further discover the indispensability of MRD in our MCDRec. This is mainly due to the fact that MRD leverages the multi-modal information to conditionally guide the diffusion process, introducing the information uncertainty of each modality into item representations.

(3) With DGD, MCDRec achieves the best performance across all datasets and metrics. This highlights the necessity of utilizing diffusion knowledge to guide the graph denoising process, thereby

Table 2: Performance comparison on Baby and Sports. *Improvement* stands for the relative improvement over its backbone.

Version	Algorithms	Baby						Sports					
		R@5	R@10	R@20	N@5	N@10	N@20	R@5	R@10	R@20	N@5	N@10	N@20
CF-based recommenders	BPR-MF [21]	0.0199	0.0389	0.0442	0.0145	0.0167	0.0200	0.0275	0.0376	0.0435	0.0167	0.0198	0.0208
	LightGCN [6]	0.0282	0.0478	0.0746	0.0190	0.0250	0.0323	0.0342	0.0524	0.0804	0.0236	0.0296	0.0368
Multi-modal recommenders	MMGCN [29]	0.0253	0.0403	0.0646	0.0170	0.0219	0.0281	0.0231	0.0371	0.0631	0.0150	0.0196	0.0263
	SLMRec [25]	0.0438	0.0486	0.0741	0.0216	0.0271	0.0337	0.0418	0.0650	0.0967	0.0285	0.0361	0.0443
	DualGNN [27]	0.0324	0.0506	0.0799	0.0213	0.0274	0.0350	0.0348	0.0579	0.0892	0.0238	0.0316	0.0402
	LATTICE [33]	0.0349	0.0542	0.0845	0.0228	0.0292	0.037	0.0395	0.0625	0.0958	0.0262	0.0337	0.0423
	BM3 [35]	0.0326	0.0535	0.0869	0.0219	0.0288	0.0374	0.0401	0.0627	0.0961	0.0269	0.0343	0.0429
	MCDRec (BM3)	0.0355	0.0566	0.0890	0.0242	0.0306	0.0386	0.0419	0.0654	0.0991	0.0279	0.0355	0.0443
	Improvement	8.90%	5.79%	2.42%	10.50%	6.25%	3.21%.	4.49%	4.31%	3.12%	3.72%	3.50%	3.26%
	FREEDOM [34]	0.0376	0.0624	0.0985	0.0243	0.0324	0.0416	0.0455	0.0713	0.1075	0.0299	0.0384	0.0477
	MCDRec (FREEDOM)	**0.0397**	**0.0644**	**0.1013**	**0.0263**	**0.0343**	**0.0438**	**0.0466**	**0.0737**	**0.1100**	**0.0306**	**0.0392**	**0.0488**
	Improvement	5.59%	3.21%	2.84%	8.23%	5.86%	5.29%	2.42%	3.37%	2.33%	2.34%	2.08%	2.31%

Table 3: Results on ablation study of MCDRec (BM3) on Baby and Sports. Generally, all components are effective.

Datasets	Versions	R@5	R@10	R@20	N@5	N@10	N@20
Baby	BM3	0.0326	0.0535	0.0869	0.0219	0.0288	0.0374
	BM3+MK	0.0331	0.0541	0.0848	0.0220	0.0289	0.0368
	BM3+MRD	0.0348	0.0558	0.0886	0.0230	0.0297	0.0380
	BM3+DGD	0.0335	0.0547	0.0875	0.0226	0.0291	0.0375
	MCDRec (BM3)	**0.0355**	**0.0566**	**0.0890**	**0.0242**	**0.0306**	**0.0386**
Sports	BM3	0.0401	0.0627	0.0961	0.0269	0.0343	0.0429
	BM3+MK	0.0403	0.0620	0.0946	0.0268	0.0340	0.0424
	BM3+MRD	0.0411	0.0641	0.0988	0.0275	0.0350	0.0436
	BM3+DGD	0.0409	0.0644	0.0983	0.0276	0.0350	0.0435
	MCDRec (BM3)	**0.0419**	**0.0654**	**0.0991**	**0.0279**	**0.0355**	**0.0443**

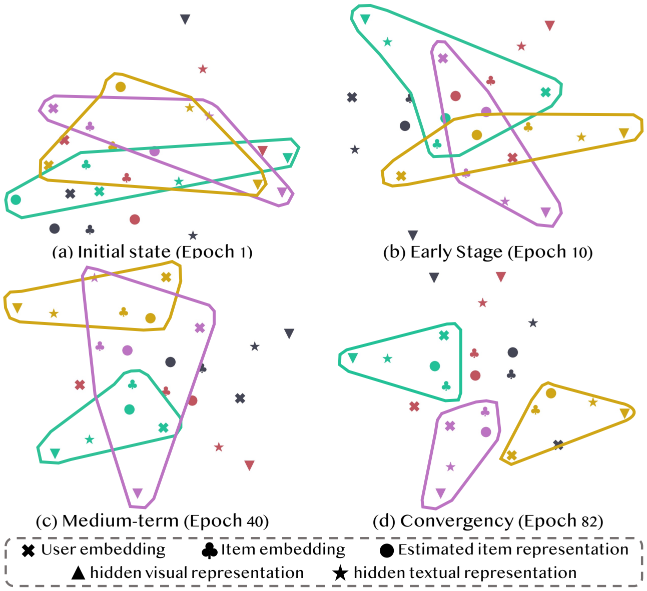
(a) Initial state (Epoch 1)
(b) Early Stage (Epoch 10)
(c) Medium-term (Epoch 40)
(d) Convergency (Epoch 82)

Figure 3: Visualization of the representation distribution of MCDRec on different training stages from the perspective of different users. We employ color differentials to distinguish different users while utilizing shapes to differentiate the user embedding e^u, item embedding e^i, hidden visual representation h^v, hidden textual representation h^t and the estimated item representation $\widetilde{e}^i$ obtained from MRD.

enabling the identification of real noisy interactions and the refinement of existing graph denoising strategies in conjunction with the diffusion-aware knowledge from MRD.

4.6 Visualization (RQ3)

To investigate how the proposed MRD affect the distribution of multi-modal item representations, we randomly select five users to extract their embedding and the multi-modal representation of their interacted items on Baby at different training stages, including (a) initial state at epoch 1, (b) early stage at epoch 10, (c) medium-term at epoch 40 and (d) convergency at epoch 82. It is noteworthy that these user-item interactions are derived from the evaluation phase specific to each given epoch, thereby remaining beyond the scope of training visibility. Then we leverage t-SNE [26] to visualize the aforementioned representations in Fig. 3.

The observations are as follows: (1) In the initial state (see Fig. 3 (a)), the intrinsic aggregation of the pre-trained representations from the same modality among diverse users poses a challenge for the subsequent recommender, hindering its ability to distinguish user multi-modal preferences effectively. (2) The clustering issues among heterogeneous modalities have been alleviated in the early stage (see Fig. 3 (b)), yet the effective differentiation of multi-modal representations from the same user remains challenging. (3) With the training of MRD, we progressively achieve consistent modeling of multi-modal preferences from the same user. However, the overlapping regions in Fig. 3 (c) indicate that our MRD still requires several iterations to better demonstrate its efficacy. (4) For the convergence of MCDRec in Fig. 3 (d), multi-modal item representations from the same user exhibit the significant clustering distributions. So far, the foundational multimodal recommenders are able to thoroughly mine users' modality-aware personalized preferences. This may be attributed to the effectiveness of our designed conditional estimator in precisely handling high-order representation correlations among multi-modal heterogeneous representations, thereby substantiating the superiority of MRD.

5 CONCLUSION

In this paper, we propose a novel Multimodal Conditioned Diffusion Model for Recommendation (MCDRec), which is able to co-model multi-modal guidance and diffusion guidance to enhance the performance of existing multi-modal recommenders. To inject modality-aware uncertainty into item representations, MCDRec first proposes the MRD module to reduce the significant deviation between modality-aware features and the collaborative information and improve the modeling of item representation. Then MCDRec presents the DGD which makes full leverage of the diffusion-aware item representations in MRD to accurately denoise the user-item interaction graph. The extensive evaluation and analyses on two real-world datasets verify the effectiveness of MCDRec and demonstrate that MCDRec can precisely capture users' modality-aware personalized preferences. In the future, we will continue to explore the fine-grained modeling of multi-modal representations in DM and to validate its effectiveness in more challenging scenarios such as multimodal sequential recommendation and cross-domain multimodal recommendation.

ACKNOWLEDGMENTS

This work is supported in part by the TaiShan Scholars Program (Grant no. tsqn202211289), the Excellent Youth Scholars Program of Shandong Province (Grant no. 2022HWYQ-048), the Oversea Innovation Team Project of the "20 Regulations for New Universities" funding program of Jinan (Grant no. 2021GXRC073).

REFERENCES

[1] Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. 2023. One transformer fits all distributions in multi-modal diffusion at scale. *arXiv preprint arXiv:2303.06555* (2023).

[2] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. 2021. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126* (2021).

[3] Ziwei Fan, Ke Xu, Zhang Dong, Hao Peng, Jiawei Zhang, and Philip S Yu. 2023. Graph collaborative signals denoising and augmentation for recommendation. In *In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. 2037–2041.

[4] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. *Journal of Machine Learning Research (JMLR)* (2010).

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

[6] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. (2020).

[7] Jonathan Ho, AjayN. Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. *Neural Information Processing Systems,Neural Information Processing Systems* (2020).

[8] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2014).

[9] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. 2022. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing* 479 (2022), 47–59.

[10] XiangLisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and TatsunoriB. Hashimoto. 2022. Diffusion-LM Improves Controllable Text Generation. (2022).

[11] Zihao Li, Aixin Sun, and Chenliang Li. 2023. DiffuRec: A Diffusion Model for Sequential Recommendation. (2023).

[12] Qidong Liu, Fan Yan, Xiangyu Zhao, Zhaocheng Du, Huifeng Guo, Ruiming Tang, and Feng Tian. 2023. Diffusion Augmentation for Sequential Recommendation. In *In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM)*.

[13] Romain Lopez, Pierre Boyeau, Nir Yosef, Michael I. Jordan, and Jeffrey Regier. 2020. AUTO-ENCODING VARIATIONAL BAYES.

[14] Haokai Ma, Xiangxian Li, Lei Meng, and Xiangxu Meng. 2021. Comparative study of adversarial training methods for cold-start recommendation. In *In Proceedings of the 1st International Workshop on Adversarial Learning for Multimedia (ADVM)*.

[15] Haokai Ma, Zhuang Qi, Xinxin Dong, Xiangxian Li, Yuze Zheng, and Xiangxu Mengand Lei Meng. 2023. Cross-Modal Content Inference and Feature Enrichment for Cold-Start Recommendation. *In Proceedings of International Joint Conference on Neural Networks (IJCNN)* (2023).

[16] Haokai Ma, Ruobing Xie, Lei Meng, Xin Chen, Xu Zhang, Leyu Lin, and Zhanhui Kang. 2024. Plug-in Diffusion Model for Sequential Recommendation. *arXiv preprint arXiv:2401.02913* (2024).

[17] Haokai Ma, Ruobing Xie, Lei Meng, Xin Chen, Xu Zhang, Leyu Lin, and Jie Zhou. 2023. Exploring False Hard Negative Sample in Cross-Domain Recommendation. In *In Proceedings of the ACM Conference on Recommender Systems (Recsys)*.

[18] Haokai Ma, Ruobing Xie, Lei Meng, Xin Chen, Xu Zhang, Leyu Lin, and Jie Zhou. 2023. Triple Sequence Learning for Cross-domain Recommendation. *ACM Transactions on Information Systems (TOIS)* (2023).

[19] Lei Meng, Fuli Feng, Xiangnan He, Xiaoyan Gao, and Tat-Seng Chua. 2020. Heterogeneous fusion of semantic and collaborative information for visually-aware food recommendation. In *In Proceedings of ACM Multimedia (MM)*.

[20] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

[21] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. *Uncertainty in Artificial Intelligence,Uncertainty in Artificial Intelligence* (2009).

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[23] Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang. 2019. DropEdge: Towards Deep Graph Convolutional Networks on Node Classification. *Learning,Learning* (2019).

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Lecture Notes in Computer Science,Lecture Notes in Computer Science* (2015).

[25] Zhulin Tao, Xiaohao Liu, Yewei Xia, Xiang Wang, Lifang Yang, Xianglin Huang, and Tat-Seng Chua. 2023. Self-supervised Learning for Multimedia Recommendation. *IEEE Transactions on Multimedia* (2023).

[26] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research (JMLR)* 9, 11 (2008).

[27] Qifan Wang, Yinwei Wei, Jianhua Yin, Jianlong Wu, Xuemeng Song, and Liqiang Nie. 2023. DualGNN: Dual Graph Neural Network for Multimedia Recommendation. *IEEE Transactions on Multimedia* (2023).

[28] Wenjie Wang, Yiyan Xu, Fuli Feng, Xinyu Lin, Xiangnan He, and Tat-Seng Chua. 2023. Diffusion Recommender Model. In *In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

[29] 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 *In Proceedings of the 27th ACM International Conference on Multimedia (MM)*.

[30] Zhengyi Yang, Jiancan Wu, Zhicai Wang, Xiang Wang, Yancheng Yuan, and Xiangnan He. 2023. Generate What You Prefer: Reshaping Sequential Recommendation via Guided Diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*.

[31] Penghang Yu, Zhiyi Tan, Guanming Lu, and Bing-Kun Bao. 2023. LD4MRec: Simplifying and Powering Diffusion Model for Multimedia Recommendation. *arXiv preprint arXiv:2309.15363* (2023).

[32] Ryoya Yuasa, Akihiro Tamura, Tomoyuki Kajiwara, Takashi Ninomiya, and Tsuneo Kato. 2023. Multimodal Neural Machine Translation Using Synthetic Images Transformed by Latent Diffusion Model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

[33] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. 2021. Mining Latent Structures for Multimedia Recommendation. In *In Proceedings of the 29th ACM International Conference on Multimedia (MM)*.

[34] Xin Zhou and Zhiqi Shen. 2023. A Tale of Two Graphs: Freezing and Denoising Graph Structures for Multimodal Recommendation. In *In Proceedings of the 31st ACM International Conference on Multimedia (MM)*.

[35] Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. 2023. Bootstrap Latent Representations for Multimodal Recommendation. In *In Proceedings of the ACM Web Conference 2023 (WWW)*.