Curriculum Conditioned Diffusion for Multimodal Recommendation

Yimeng Yang[1], Haokai Ma[1], Lei Meng[1, 2*], Shuo Xu[1], Ruobing Xie[3], Xiangxu Meng[1]

[1]School of Software, Shandong University, Jinan, China
[2]Shandong Research Institute of Industrial Technology, Jinan, China
[3]Tencent, China
{y_yimeng, mahaokai, shuo.xu}@mail.sdu.edu.cn, {lmeng, mxx}@sdu.edu.cn, xrbsnowing@163.com

Abstract

Multimodal recommendation (MMRec) aims to integrate multimodal information of items to address the inherent data sparsity issue in collaborative-based recommendation. Traditional MMRec methods typically capture the structure-level item representations from the observed user behaviors within the multimodal graph, overlooking the potential impact of negative instances for personalized preference understanding. In light of the outstanding generative ability and step-by-step inference characteristic of Diffusion Models (DMs), we propose a Curriculum Conditioned Diffusion framework for Multimodal Recommendation (CCDRec), which precisely excavates the modality-aware distribution-level correlation among multi-modalities and elegantly integrates the reverse phase of DMs into negative sampling to highlight the most suitable instances in a curricular manner. Specifically, CCDRec proposes the Diffusion-controlled Multimodal Aligning module (DMA) to align multimodal knowledge with collaborative signals by capturing the fine-grained relationships among multi-modalities in the probabilistic distribution space. Furthermore, CCDRec designs the Negative-sensitive Diffusive Inferring module (NDI) to progressively synthesize the negative sample pool with diverse hardness to support the following knowledge-aware negative sampling. To gradually ramp up the training complexity, CCDRec further introduces a Curricular Negative Sampler (CNS) to tally the curriculum learning paradigm with the reverse phase of DMA, thereby adaptively sampling the gold-standard negative instances to enhance optimization. Extensive experiments on three datasets with four diverse backbones demonstrate the effectiveness and robustness of our CCDRec. The visualization analyses also clarify the underlying mechanism of our DMA in multimodal representation alignment and CNS in curricular negative discovery. The code and the corresponding dataset will be uploaded in the Appendix.

Introduction

Multimodal recommendation is crucial in the information society as it combines different data types (Ma et al. 2023a; Wang et al. 2023b), including text, images, and audio, to fully capture user preferences and deliver more personalized and relevant recommendations.

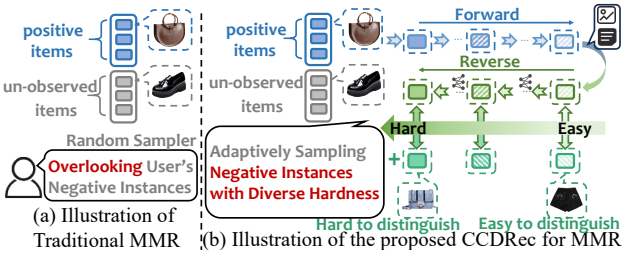

Figure 1: Illustration of the difference between the traditional MMR methods and CCDRec, which integrates the reverse process of DM with negative sampling to sample negative instances with diverse hardness adaptively.

Several research directions have emerged to incorporate multimodal content into recommender systems. Inspired by the success of graph neural networks (GNNs), several studies have frequently utilized GNNs to derive representations from different views. For instance, these methods typically generate modality preference representations under the user preference view (Zhou, Zhou, and Shen 2023) and obtain item representations within the multimodal view (Zhang et al. 2021; Zhou and Shen 2023). Additionally, some researchers leverage self-supervised multimodal signals (Yang et al.) to ensure content coherence across diverse modalities (Zhou et al. 2023; Tao et al. 2022). Recently, the effectiveness of generative models (Li et al. 2024; Sun et al. 2024) such as Variational Autoencoders and Diffusion Models (DM) have been explored in this specific recommendation task (Bai et al. 2023; Ma et al. 2024c). Existing algorithms primarily leverage the observed user interactions to improve the representation learning of items from multimodal knowledge and collaborative information. Nevertheless, these methodologies overlook the impact of negative behaviors, thereby losing some necessary knowledge that effectively understands user preferences.

In recommender systems, negative sampling is critical for capturing negative instances in a sparse user-item interaction matrix to enhance model performance. Traditional strategies typically define the fixed probabilities in recommendation (e.g., uniform sampling (Guo et al. 2017) and popularity-based sampling (Mikolov et al. 2013)), lacking flexibility in capturing users' dynamical preferences. To address this, re-

*Corresponding Author

cent studies have introduced a technique called hard negative sampling (HNS). It dynamically selects more informative negative instances, such as variance-based sampling functions (Ding et al. 2020) and hop-mixing strategies (Huang et al. 2021). In multimodal recommendation, existing negative sampling strategies typically apply methods proven effective in collaborative filtering to multimodal representations, lacking flexibility and adaptability. Benefiting from its outstanding generative ability and step-by-step inference characteristic, DM achieves notable success in image generation and speech synthesis. The multi-step Markov reverse process of DM enables both flexible access at denoising stages and adaptive difficulty control, facilitating step-wise negative sampling. To achieve this, we must address the following two challenges: (1) How to leverage DM for integrating multimodal information with collaborative knowledge, learning more accurate item-aligned representations? (2) How can we develop a negative sampling strategy tailored for multimodal recommendation that seamlessly combines the reverse process of DM with the negative sampling process in a dynamic manner?

To tackle these challenges, we propose a Curricular Conditioned Diffusion for Multimodal Recommendation (CCDRec), which utilizes a multimodal conditional diffusion model to align the multimodal representations with the collaborative signals and guides the adaptive negative sampling process. Specifically, we first propose the Diffusion-controlled Multimodal Aligning module (DMA), which leverages the DM to capture fine-grained relationships among modalities while aligning multimodal features with collaborative features, generating accurate item-aligned representations. Furthermore, we develop the Negative-sensitive Diffusive Inferring (NDI), which constructs sample pools with diverse hardness by progressively synthesizing information-rich features. This allows for a flexible selection of negative instances with varying difficulty levels. To mitigate the challenges associated with over-challenging negative samples early in training, we designed a Curricular Negative Sampler (CNS) to dynamically model user preferences, allowing the model to train from simple to complex.

Extensive experiments on three real-world datasets show that CCDRec achieves significant improvements across all datasets. Additionally, we conducted ablation studies to verify the effectiveness of all components in CCDRec and performed visualization analyses to elucidate the underlying mechanisms of DMA in multimodal integration. The main contributions of this paper are summarized as follows:

- We introduce a multimodal recommendation framework CCDRec, which skillfully combines the reverse phase of conditioned DMs into the negative sampling to pinpoint the optimal instances. To our knowledge, we are the first to explore negative sampling strategies for multimodal recommendations using diffusion models.
- We propose three model-agnostic modules, DMA, NDI, and CNS, which can be integrated at different stages and work well with multimodal recommendation systems.
- We conduct extensive experiments on three datasets with two multimodal recommendation backbones to demonstrate the effectiveness and universality of CCDRec.

Related Work

Multimodal Recommendation The goal of multimodal recommendation is to enhance item representations by incorporating supplementary multimodal content alongside historical interactions. Early studies (He and McAuley 2016; Chen et al. 2019) used pre-extracted visual features to enrich item representations. Based on the success of Graph Neural Networks (GNN) in recommendation, some methods (Wang et al. 2021; Zhang et al. 2021; Zhou and Shen 2023) incorporate it to extract user-specific modal preferences and high-order relationships between items. Additionally, self-supervised learning (SSL) techniques (Tao et al. 2022; Zhou et al. 2023) have been introduced to improve latent item representations. Recently, generative methods have also been applied in multimodal recommendations (Bai et al. 2023; Yu et al. 2023; Ma et al. 2024c). For instance, MCDRec (Ma et al. 2024c) uses DM to model multimodal and collaborative data in a continuous space. However, existing methods have overlooked the modeling of negative behaviors.

Diffusion Models in Recommendation Motivated by the uncertainty injection and data augmentation capabilities of Diffusion Models (DM) in computer vision (Rombach et al. 2022; Zheng et al. 2024; Yu et al. 2024), some studies have explored the effectiveness of DM in recommendation. For instance, DiffRec (Wang et al. 2023a) gradually generates global yet personalized collaborative information through a denoising process. PDRec (Ma et al. 2024a) introduces an approach with three plug-in modules to fully utilize diffusion-based preferences across items. Some methods (Li, Sun, and Li 2023; Yang et al. 2024) explore the underlying distribution of item spaces using DM to enhance insights into item dynamics guided by users' sequential behaviors. MCDRec (Ma et al. 2024c) injects modality-aware uncertainty into item representations to mitigate biases between multimodal and collaborative features.

Negative Sampling in Recommendation Recommendation systems commonly use Bayesian Personalized Ranking (BPR) (Rendle et al. 2012) and static negative sampling based on fixed probability distributions to optimize models (Guo et al. 2017; Mikolov et al. 2013). However, uniformly selected negative items may result in smaller gradients and less contribution to convergence. To overcome this issue, researchers proposed hard negative sampling (HNS) methods such as DNS (Zhang et al. 2013) to oversample high-score negatives, obtaining more information. For instance, SRNS (Ding et al. 2020) employs a variance-based function to detect high-information negative samples, and MixGCF (Huang et al. 2021) generates synthetic negative samples by combining negatives from multiple layers. Nonetheless, the primary reliance of the mentioned methods on collaborative filtering and graph representation learning limits their suitability for the multimodal recommender.

Methods

Task Formulation and Overall Framework

The goal of multimodal recommendation is to leverage the additional multimodal information on items to obtain more

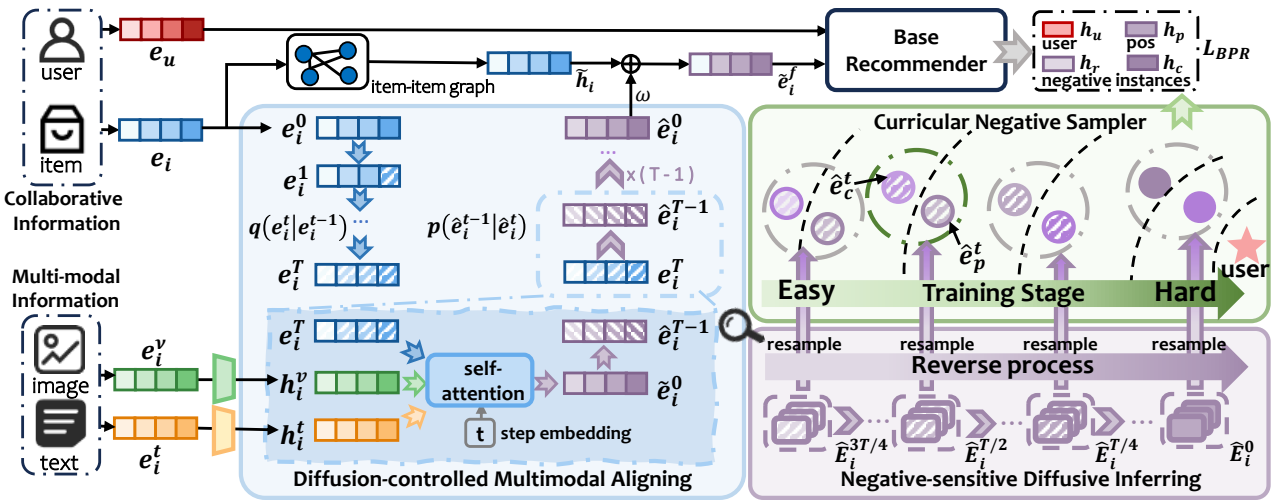

Figure 2: The overall structure of CCDRec. DMA explicitly aligns multimodal knowledge with collaborative signals via DM while NDI and CNS elegantly integrate the reverse phase of DM into negative sampling to highlight the most suitable instances.

precise item representations for recommendations. We define the embeddings $\boldsymbol{e}_u$ and $\boldsymbol{e}_i$ for user $u \in \mathcal{U}$ and item $i \in \mathcal{I}$, where $\mathcal{U}$ and $\mathcal{I}$ denote the set of users and items. For each item, we have its visual feature $\boldsymbol{e}_i^v$ and textual feature $\boldsymbol{e}_i^t$ as additional information.

We thoroughly introduce the proposed CCDRec, which uses DM to enhance the multimodal fusion (Chen et al. 2023b; Wang et al. 2022) of items and applies diffusion-generated knowledge for adaptive negative sampling, selecting different negative samples at various stages. The main architecture of CCDRec is shown in Figure.2. Specifically, CCDRec introduces a Diffusion-controlled Multimodal Aligning module (DMA), which utilizes DM to capture the fine-grained correlations across different modalities to generate the aligned item representations. Subsequently, CCDRec introduces a Negative-sensitive Diffusive Inferring module (NDI), which forms sample pools using item-aligned features generated at various diffusion steps for negative sampling. To improve generalization and convergence, CCDRec has designed a Curricular Negative Sampler (CNS) that selects progressively harder negative samples throughout training.

Base Multimodal Recommender

We utilize FREEDOM(Zhou and Shen 2023) as our base multimodal recommender, constructing modality-aware item-item graphs with raw features $\boldsymbol{e}_i^v$ and $\boldsymbol{e}_i^t$ and simplifying them using KNN sparsification to form normalized adjacency matrices. By merging these matrices, we create a unified latent item-item graph S and apply graph convolutions for feature aggregation and information propagation to obtain $\hat{\boldsymbol{h}}_i$. In the user-item graph $\widehat{A}$, we carry out multiple convolutional operations using the default settings of LightGCN to derive the ID embeddings of users and items, designated as $\widetilde{\boldsymbol{h}}_i$ and $\widetilde{\boldsymbol{h}}_u$. In the end, the representations for users and items are $\boldsymbol{h}_u = \widetilde{\boldsymbol{h}}_u$ and $\boldsymbol{h}_i = \hat{\boldsymbol{h}}_i + \widetilde{\boldsymbol{h}}_i$. Additionally, we utilize Multilayer Perceptrons (MLPs) to project features of each modality as $\boldsymbol{h}_i^m = \boldsymbol{e}_i^m \mathbf{W}_m + \boldsymbol{b}_m$.

Diffusion-controlled Multimodal Aligning

Diffusion models (Po et al. 2024; Li et al. 2023; Hoogeboom, Heek, and Salimans 2023) fundamentally transform data progressively into noise, subsequently generating reconstructed samples via a parameterized denoising trajectory that mirrors the original data's distribution (Wallace et al. 2024; Prabhudesai et al. 2023; Giannone et al. 2023). Motivated by this principle, we introduce the Diffusion-controlled Multimodal Aligning (DMA) module, derived from Denoising Diffusion Probabilistic Models (DDPM). This module is designed to accurately capture the probabilistic correlations between multiple modalities and align multimodal information with collaborative signals. By generating aligned multimodal fused features, it aims to address the inconsistencies between collaborative and multimodal information while better-capturing users' deeper preferences.

Learning Phase of DMA Given an item ID embedding $\boldsymbol{e}_i$, we initially denote it as $\boldsymbol{e}_i^0$. During the **forward process**, we gradually introduce Gaussian noise into $\boldsymbol{e}_0^i$, transforming it into an uncertain distribution after t steps, where $t \sim \text{Uniform}\{1, 2, \ldots, T\}$:

$$q\left(\boldsymbol{e}_i^t \mid \boldsymbol{e}_i^0\right) = \mathcal{N}\left(\boldsymbol{e}_i^t, \sqrt{\overline{\alpha}_t}\boldsymbol{e}_i^0, \left(1-\overline{\alpha}_t\right)\mathrm{I}\right). \quad (1)$$

Through reparameterization, we can obtain the item representation $\boldsymbol{e}_i^t = \sqrt{\overline{\alpha}_t}\boldsymbol{e}_i^0 + \sqrt{1-\overline{\alpha}_t}\epsilon$. The **reverse process** is essential in DM to iteratively eliminate the noise introduced during the forward phase. To generate the subsequent denoised representation at each step t, the reverse process is employed:

$$p_\theta\left(\boldsymbol{e}_i^{t-1} \mid \boldsymbol{e}_i^t\right) = \mathcal{N}\left(\boldsymbol{e}_i^t; f_\theta\left(\boldsymbol{e}_i^t, t, \boldsymbol{h}_i^v, \boldsymbol{h}_i^t\right), \Sigma_\theta\left(\boldsymbol{e}_i^t, t\right)\right), \quad (2)$$

where $\Sigma_\theta\left(\boldsymbol{e}_i^t, t\right) = \sigma_t^2\mathrm{I} = \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\beta_t\mathrm{I}$ denotes the variance. Typically, traditional diffusion models achieve inference by

predicting the noise at each step. However, in the recommendation domain, it is common to train suitable estimators directly to approximate $\boldsymbol{e}_0^i$ for performing the reverse steps. Here, the mean $\mu_\theta\left(\boldsymbol{e}_i^t, t, \boldsymbol{h}_i^v, \boldsymbol{h}_i^t\right)$ can be calculated by:

$$\mu_\theta\left(\boldsymbol{e}_i^t, t, \boldsymbol{h}_i^v, \boldsymbol{h}_i^t\right) = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{e}_i^t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} f_\theta\left(\boldsymbol{e}_i^t, t, \boldsymbol{h}_i^v, \boldsymbol{h}_i^t\right)\right), \quad (3)$$

where $f_\theta(.)$ is the tailored conditional estimator.

The core of generative tasks is to optimize the underlying data generation distribution, typically done by optimizing the variational bound(VLB) of the negative log-likelihood. To learn a high-quality conditional estimator, we follow the DDPM (Ho, Jain, and Abbeel 2020) setup and minimize the KL divergence between the two distributions $q\left(\boldsymbol{e}_i^t \mid \boldsymbol{e}_i^{t-1}, \boldsymbol{e}_0^i\right)$ and $p_\theta\left(\boldsymbol{e}_i^{t-1} \mid \boldsymbol{e}_i^t\right)$:

$$\mathcal{L}_{vlb} = D_{KL}\left(q\left(\boldsymbol{e}_i^t \mid \boldsymbol{e}_i^{t-1}, \boldsymbol{e}_i^0\right) \| p_\theta\left(\boldsymbol{e}_i^{t-1} \mid \boldsymbol{e}_i^t\right)\right). \quad (4)$$

Then, we transform it into a simpler Mean-Squared Error (MSE) loss function:

$$\mathcal{L}_{dm} = E_{\boldsymbol{e}_i^0, \boldsymbol{e}_i^t}\left[\left\|\boldsymbol{e}_i^0 - f_\theta\left(\boldsymbol{e}_i^t, t, \boldsymbol{h}_i^v, \boldsymbol{h}_i^t\right)\right\|^2\right]. \quad (5)$$

In this context, $\boldsymbol{e}_i^0$ signifies the initial item embedding, while f_θ is the conditional estimator, responsible for generating the estimated item representation $\widetilde{\boldsymbol{e}}_i^0$. Next, we merge the higher-order item representation $\hat{\boldsymbol{h}}_i$ from the base recommender with the item-aligned representation to form the fused representation $\hat{\boldsymbol{e}}_i^f = (1-\mu)\,\hat{\boldsymbol{h}}_i + \mu \cdot \widetilde{\boldsymbol{e}}_i^0$, where μ is an adjustable parameter that controls the diffused weight.

Conditional Estimator Following (Li, Sun, and Li 2023; Wang et al. 2024), we adopt the Transformer architecture as the conditional estimator $f_\theta(.)$ in the reverse process to generate $\widetilde{\boldsymbol{e}}_i^0$. We integrate various modal features, including noised item representation $\boldsymbol{e}_i^t$, textual feature $\boldsymbol{h}_i^t$, visual feature $\boldsymbol{h}_i^v$, and alongside a sinusoidal time step embedding t_i to form the input feature matrix $\mathbf{F} \in \mathbb{R}^{B \times M \times d}$, where B is the batch size, M is the number of modalities, and d is the feature dimension. The self-attention mechanism selectively focuses on different parts of the input data, allowing the model to capture complex dependencies and relationships between the various modalities. The aggregated attention output $\widetilde{\boldsymbol{e}}_i^0$ is obtained by averaging across modalities, which ensures the sophisticated incorporation of multimodal data precisely condition the estimation of $\hat{\boldsymbol{e}}_i^0$.

Inference Phase of DMA After each training epoch, the inference phase of the diffusion model is executed. First, similar to the training operations, given a complete diffusion step $t = T$, we add noise to the item embedding $\boldsymbol{e}_i^0$ to obtain $\hat{\boldsymbol{e}}_i^t$. After that, we perform a step-by-step reverse denoising operation $\hat{\boldsymbol{e}}_i^T \rightarrow \hat{\boldsymbol{e}}_i^{T-1} \rightarrow \cdots \rightarrow \hat{\boldsymbol{e}}_i^0$ to achieve the final $\hat{\boldsymbol{e}}_i^0$. Through this process, we can generate an item-aligned representation that conforms to the collaborative representation distribution. Aligned with the training task, we generate $\hat{\boldsymbol{e}}_i^f = (1-\mu)\,\hat{\boldsymbol{h}}_i + \mu \cdot \hat{\boldsymbol{e}}_i^0$ as the item-fused representation.

Negative-sensitive Diffusive Inferring

Hard negative instances (Liu et al. 2023; Li et al. 2021) typically refer to items highly relevant to a user's interests, providing rich information that enhances the model's ability to discern and adapt to user interest preferences(Chen et al. 2023a). However, existing research (Ma et al. 2023b; Qi et al. 2022) has shown that optimizing with the hardest negatives in the early stages can lead to local minima, resulting in sub-optimal performance. Therefore, synthesizing negative sample pools with varying difficulty levels to devise appropriate negative sampling strategies poses a challenge. To address this, we propose the Negative-sensitive Diffusive Inferring module (NDI). Integrated with the DMI inference process, NDI leverages features from different steps to create a knowledge-aware negative sample pool, allowing the selection of negatives with varying difficulty.

As described in the previous section, we update DM during training to incorporate multimodal information into the item collaborate feature. Subsequently, the trained DM incrementally generates fused item representations across multiple steps, enabling natural access at different stages. Notably, this process is executed once at the start of each epoch, ensuring minimal computational cost. Specifically, the reverse process starts from noise and gradually generates the final representation over T steps. These samples progressively approach the item-fused representations with the richest information. To this end, we establish a fixed step interval and extract item representations after $T/4$, $T/2$, $3T/4$, and T steps. These extracted representations form the four sample candidate pools, which can be expressed as:

$$\hat{E}_{|\mathcal{I}|}^t = \left[\hat{\boldsymbol{e}}_0^t, \hat{\boldsymbol{e}}_1^t, \cdots, \hat{\boldsymbol{e}}_{|\mathcal{I}|}^t\right] \in \mathbb{R}^{|\mathcal{I}| \times d}, t \in \left\{\tfrac{3T}{4}, \tfrac{T}{2}, \tfrac{T}{4}, 0\right\}. \quad (6)$$

Then, we can flexibly use different sample pools to identify indices of negative instances with varying difficulty levels.

Curricular Negative Sampler

Including all difficult negative samples (Ma et al. 2024b) early in training can hinder model performance and slow convergence. To address this, we use Curriculum Learning (CL) (Chen et al. 2021) to gradually increase sample difficulty, improving generalization and speeding up convergence. Therefore, we developed an adaptive Curricular Negative Sampler (CNS) to enhance the learning process of multimodal recommendion. During training, we progressively introduce harder negative samples at different stages, warming up the model and mitigating the impact of difficult negatives. At a specified training epoch n, the sample pool $\hat{E}_{|\mathcal{I}|}^t$ utilized is determinable by the formula:

$$t = (T/4) \times \left(3 - \min\left(3, \lfloor n/\Delta\tau \rfloor\right)\right). \quad (7)$$

Here, $\Delta\tau$ controls the interval between epochs in CL, and τ_{end} marks the ending of the CL strategy within the training process. After epoch $n > \tau_{\text{end}}$, we use the final item representations as the sample pool for negative sampling, considered the hardest samples.

Once the sample pools are selected, we randomly sample 10% of the items as candidates. As mentioned in Section 3.3, the final item representations are denoted as $H_{|\mathcal{I}|} = \left[\boldsymbol{h}_0, \boldsymbol{h}_1, \cdots, \boldsymbol{h}_{|\mathcal{I}|}\right] \in \mathbb{R}^{I \times d}$. Given a positive item $\boldsymbol{h}_p$, we retrieve its representation $\hat{\boldsymbol{e}}_p^t$ at the specific diffusion step from the sample pool and compute the similarity with all

Dataset	#Users	#Items	#Interactions	Sparsity
Baby	19,445	7,050	160,792	99.88%
Sports	35,598	18,357	296,337	99.95%
Clothing	39,387	23,033	237,488	99.97%

Table 1: Statistics of three real-world multimodal datasets.

candidate items. An item is randomly sampled from the top k candidates with the highest similarity scores. Formally, let $S\left(\hat{\boldsymbol{e}}_p^t, \hat{\boldsymbol{e}}_j^t\right)$ denote the similarity score between the positive item representation $\hat{\boldsymbol{e}}_p^t$ and the candidate items at step t. The top k candidates $\mathcal{C}_k$ are selected based on the highest similarity scores, where k represents the proportion of the top-rated items. Next, we resample an item $\hat{\boldsymbol{e}}_c^t$ where $c \in \mathcal{C}_k$, designating c as the curricular negative item index. Once the item index c is selected, it is mapped back to its final representation $\boldsymbol{h}_c$ as the final negative instance. To enhance the model's generalization capability and stability, we consider the inclusion of easy negative instances during the training process is necessary. Therefore, we incorporate randomly selected $\boldsymbol{h}_r$ and the chosen negative sample $\boldsymbol{h}_c$ for joint training.

Optimization Objectives

Building on the methodology of traditional recommendation algorithms, we implement the Bayesian personalized ranking (BPR) loss (Rendle et al. 2012) for each user u and positive item $\boldsymbol{h}_p$. To incorporate two types of negative samples $\boldsymbol{h}_r$ and $\boldsymbol{h}_c$, we introduce two separate loss functions, which can be written as follows:

$$\begin{cases} \mathcal{L}_{bpr}^r = \sum\limits_{(u,p,r)\in\mathcal{R}} \left(-\log\sigma\left(\boldsymbol{h}_u^\top \boldsymbol{h}_p - \boldsymbol{h}_u^\top \boldsymbol{h}_r\right)\right), \\ \mathcal{L}_{bpr}^c = \sum\limits_{(u,p,c)\in\mathcal{R}} \left(-\log\sigma\left(\boldsymbol{h}_u^\top \boldsymbol{h}_p - \boldsymbol{h}_u^\top \boldsymbol{h}_c\right)\right). \end{cases} \tag{8}$$

The overall objective function $\mathcal{L}$ can be formulated as:

$$\mathcal{L} = (1-\omega)\cdot\mathcal{L}_{bpr}^r + \omega\cdot\mathcal{L}_{bpr}^c + \lambda\cdot\mathcal{L}_{dm}, \tag{9}$$

where λ and ω denote the weight of different losses.

Experiments

In this section, we conduct comprehensive experiments to answer the following research questions:

- **RQ1:** How does CCDRec perform against the CF methods and the SOTA multimodal recommendation methods?
- **RQ2:** How do different components in our CCDRec impact its recommendation performance?
- **RQ3:** Is CCDRec still effective with diverse multimodal recommendation backbones?
- **RQ4:** Does CCDRec still maintain its superiority compared to other negative sampling algorithms?
- **RQ5:** How does DMA affect the distribution of user representation and item representations?
- **RQ6:** What is the underlying mechanism of CNS in curricular negative discovery?

Experimental Settings

Datasets Following previous works (Zhou 2023), we perform experiments on the $Baby$, $Sports$, and $Clothing$ datasets from the Amazon platform. We pre-process the data with a 5-core setting on items and users, as used in (He and McAuley 2016), and present the results in Table 1. Visual features are directly used as pre-extracted with a dimension of 4096 (Zhou and Shen 2023), while textual features are obtained using sentence-transformers (Reimers and Gurevych 2019) with 384-dimensional embeddings.

Baselines To demonstrate the effectiveness of our method, we compare it with the following baseline models. First, we select two general CF-based recommenders **BPR** (Rendle et al. 2012) and **LightGCN** (He et al. 2020). Additionally, we further compare it with seven multimodal recommenders: **MMGCN** (Wei et al. 2019), **SLMRec** (Tao et al. 2022), **LATTICE** (Zhang et al. 2021), **BM3** (Zhou et al. 2023), **FREEDOM** (Zhou and Shen 2023), **MG** (Zhong et al. 2024) and **MCDRec** (Ma et al. 2024c).

Parameter Settings Following the classical works (Zhang et al. 2021; Zhou et al. 2023; Zhou and Shen 2023), we set the embedding size of both users and items to 64 for all models. To ensure a fair comparison, we present the results of other methods using two random negative samples. 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. We set the loss weight λ at $\{0.5, 1, 2\}$ and ω at $\{0.5, 0.7, 0.8, 0.9\}$. As for the diffusion process, the step t is tuned in $\{5, 10, 20, 40, 100\}$. Respectively, the diffused weight μ is chosen from $\{0.3, 0.5, 0.8\}$. $\Delta\tau$ is searched in the set $\{3, 5, 10, 15, 20\}$, and τ_{end} is searched in $\{30, 50, 75, 100\}$. Following (Zhou and Shen 2023), we opt for the early stopping strategy.

Performance Comparison (RQ1)

We conduct experiments on three real-world datasets with two standard evaluation metrics: NDCG@k (N@k) and Recall@k (R@k), where k is in $5, 10$. We accentuate the best results of the same backbone with bold font. As shown in Table 2, we can observe the following insights:

(1) CCDRec significantly outperforms all baselines across all metrics in three datasets. This indirectly demonstrates that the combination of multimodal diffusion-enhanced item fusion and diffusion knowledge-guided negative sampling strategies can effectively leverage multimodal information, enabling the model to learn users' fine-grained multimodal preferences. Additionally, multimodal recommendation methods generally outperform traditional CF-based recommendations, and CCDRec achieves further improvements over state-of-the-art multimodal recommenders. This further demonstrates the superiority of incorporating multimodal information in recommendation systems.

(2) Comparing the performance of CCDRec on various base models, we observe that CCDRec delivers the most notable improvement on LATTICE and achieves peak results across all datasets when combined with FREEDOM. Significant improvements are achieved on LATTICE, FREEDOM, and MG (without any multimodal diffusion strategy), which further underscores the ability of CCDRec to model item multimodal fusion representations with the tailored DMs.

Versions	Algorithms	Baby				Sports				Clothing			
		R@5	R@10	N@5	N@10	R@5	R@10	N@5	N@10	R@5	R@10	N@5	N@10
CF-based recommenders	BPR-MF	0.0208	0.0344	0.0138	0.0183	0.0257	0.0410	0.0177	0.0228	0.0118	0.0191	0.0079	0.0102
	LightGCN	0.0307	0.0488	0.0204	0.0263	0.0354	0.0554	0.0242	0.0308	0.0219	0.0355	0.0145	0.0189
Multimodal recommenders	MMGCN	0.0251	0.0410	0.0164	0.0217	0.0236	0.0388	0.0154	0.0204	0.0128	0.0210	0.0085	0.0111
	SLMRec	0.0320	0.0486	0.0216	0.0271	0.0420	0.0650	0.0285	0.0361	0.0290	0.0440	0.0192	0.0240
	BM3	0.0326	0.0535	0.0219	0.0288	0.0401	0.0627	0.0269	0.0343	0.0273	0.0417	0.0180	0.0226
	LATTICE	0.0352	0.0545	0.0228	0.0291	0.0395	0.0625	0.0263	0.0338	0.0330	0.0499	0.0217	0.0272
	CCDRec(LATTICE)	**0.0371**	**0.0596**	**0.0251**	**0.0325**	**0.0470**	**0.0715**	**0.0316**	**0.0397**	**0.0393**	**0.0613**	**0.0259**	**0.0330**
	Improvement	5.40%	9.36%	10.09%	11.68%	18.99%	14.40%	20.15%	17.46%	19.09%	22.85%	19.35%	21.32%
	FREEDOM	0.0389	0.0626	0.0250	0.0328	0.0455	0.0713	0.0299	0.0384	0.0403	0.0623	0.0265	0.0337
	CCDRec(FREEDOM)	**0.0426**	**0.0679**	**0.0274**	**0.0356**	**0.0481**	**0.0760**	**0.0315**	**0.0406**	**0.0433**	**0.0677**	**0.0288**	**0.0368**
	Improvement	9.51%	8.47%	9.60%	8.54%	5.71%	6.59%	5.35%	5.73%	7.44%	8.67%	8.68%	9.20%
	MCDRec	0.0381	0.0651	0.0255	0.0343	0.0463	0.0709	0.0305	0.0386	0.0415	0.0653	0.0276	0.0353
	CCDRec(MCDRec)	**0.0409**	**0.0667**	**0.0269**	**0.0354**	**0.0478**	**0.0740**	**0.0315**	**0.0400**	**0.0434**	**0.0670**	**0.0288**	**0.0364**
	Improvement	7.35%	2.46%	5.49%	3.21%	3.24%	4.37%	3.28%	3.63%	4.58%	2.60%	4.35%	3.12%
	MG	0.0390	0.0624	0.0253	0.0330	0.0460	0.0714	0.0302	0.0385	0.0400	0.0622	0.0264	0.0336
	CCDRec(MG)	**0.0399**	**0.0651**	**0.0262**	**0.0344**	**0.0489**	**0.0746**	**0.0319**	**0.0404**	**0.0428**	**0.0664**	**0.0284**	**0.0361**
	Improvement	2.31%	4.33%	3.56%	4.24%	6.30%	4.48%	5.63%	4.94%	7.00%	6.75%	7.58%	7.44%

Table 2: Performance comparison on three datasets. *Improvement* stands for the relative improvement over its backbone.

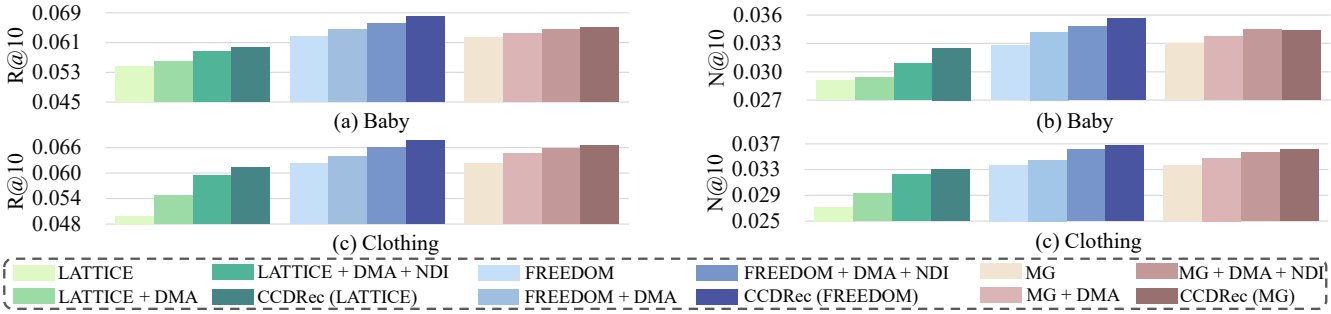

Figure 3: Results on ablation study of CCDRec on LATTICE, FREEDOM and MG. All components are effective.

(3) Moreover, our CCDRec shows consistent improvements on MCDRec which incorporates diffusion-guided item modeling, suggesting the limitations of U-Net for item feature reconstruction in MCDRec, like the loss of crucial modal information. Additionally, this result indicates that CCDRec offers valuable guidance for model training through diffusion-guided negative sampling strategies.

Ablation Study (RQ2 & RQ3)

We conduct ablation studies to examine the efficacy of different components in CCDRec. Specifically, we compare CCDRec with its ablation versions to verify the effectiveness of DMA, NDI, and CNS respectively. Here, "FREEDOM+DMA+NDI" refers to randomly selecting samples from the last sample pool $\hat{E}^0_{|\mathcal{I}|}$ in NDI as the negative instances for model optimization. Notably, CCDRec (FREEDOM) is equivalent to FREEDOM+DMA+NDI+CNS. As shown in Figure 3, we observe that:

(1) FREEDOM+DMA consistently outperforms FREEDOM across two datasets, demonstrating that DMA better captures users' fine-grained modal preferences and generates more accurate item-aligned representations.

(2) FREEDOM+DMA+NDI exhibits significant improvement over FREEDOM+DMA, indicating that the NDI module effectively uncovers latent negatives using item-aligned representations. CCDRec (FREEDOM) further boosts its performance, underscoring the value of a curriculum-based dynamic negative sampling strategy that adapts difficulty based on inference steps for optimal training.

(3) We also perform a series of progressive ablation experiments on different variants of various base models, consistently finding that CCDRec outperforms all other variants. This indicates that the different components we propose are effective and generalizable across various multimodal recommendation models.

Performance against Other Hard Negative Sampling Methods (RQ4)

To further verify the performance of CCDRec, we compared it with three hard negative sampling methods: DNS (Zhang et al. 2013), MixGCF (Huang et al. 2021), and RealHNS (Ma et al. 2023b). To ensure a fair comparison, we integrate them into three base models and conduct a comprehensive evaluation under consistent experimental settings. The re-

Versions	Baby		Clothing	
	Recall@10	NDCG@10	Recall@10	NDCG@10
LATTICE	0.0545	0.0291	0.0499	0.0272
+DNS	0.0572	0.0311	0.0580	0.0322
+MixGCF	0.0582	0.0316	0.0582	0.0321
+RealHNS	0.0586	0.0313	0.0586	0.0322
+CCDRec	**0.0596**	**0.0356**	**0.0613**	**0.0330**
FREEDOM	0.0626	0.0328	0.0623	0.0337
+DNS	0.0637	0.0339	0.0650	0.0354
+MixGCF	0.0654	0.0348	0.0644	0.0350
+RealHNS	0.0659	0.0351	0.0641	0.0351
+CCDRec	**0.0679**	**0.0356**	**0.0677**	**0.0368**
MG	0.0624	0.0330	0.0622	0.0336
+DNS	0.0635	0.0336	0.0639	0.0346
+MixGCF	0.0643	0.0342	0.0648	0.0349
+RealHNS	0.0644	0.0338	0.0651	0.0352
+CCDRec	**0.0651**	**0.0344**	**0.0664**	**0.0361**

Table 3: Comparative results of CCDRec with other hard negative sampling methods on two datasets.

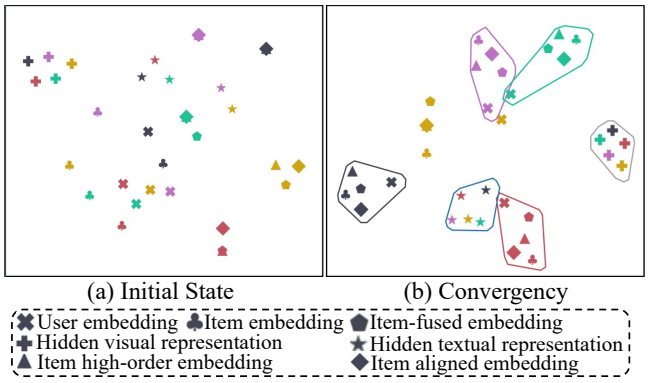

Figure 4: Visualization of the multimodal representation distribution of CCDRec on different training stages from the perspective of different users.

sults are shown in Table 3, where we can observe that: (1) CCDRec consistently outperforms other negative sampling methods across different base models (LATTICE, FREEDOM and MG) on two datasets, which further illustrates the superiority of our method in negative sampling. (2) Using different HNS methods with various backbones consistently improves performance. This substantiates the significant latent potential of negative information in the multimodal recommendation, showing that the judicious selection of negative sampling strategies can enhance the model's ability to model users' multimodal preferences consistently.

In-depth Analyses of CCDRec

In this section, we employ t-SNE (Van der Maaten and Hinton 2008) to visualize how the proposed DMA impacts the distribution of multimodal item representations as shown in Fig. 4, and also clarify the underlying mechanism of the proposed CNS in negative selection in Fig.5.

Estimation of Multimodal Alignment in DMA (RQ5) First, we randomly select five users to extract their embedding and multimodal item representations of their interacted items on Baby at the initial state and the convergence state. The same color represents representations belonging to the same user. We find that at the initial stage, the related representations of the same user are scattered in the whole low-dimensional space. In contrast, these representations of the same user exhibit significant clustering distributions in the convergence stage, with the item-fused representation being the closest to the user. This may be attributed to the effectiveness of DMA in precisely capturing the fine-grained relationships among multi-modalities of the same item.

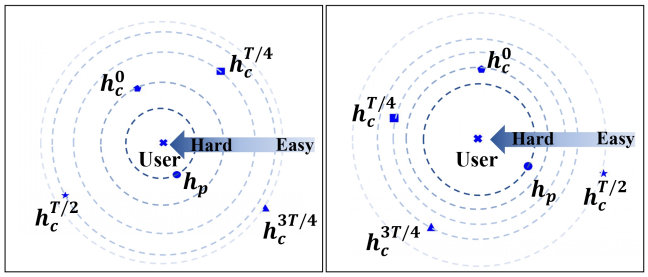

Figure 5: Visualization of the representation distribution of negative instances with diverse hardness (from step $3T/4$ to step 0 of the inference phase of DMA) in training.

Estimation of Negative Inference in CNS (RQ6) To investigate the effectiveness of sampled negative instances in CNS, we randomly select two users with the positive samples they have interacted with and select the negative instances with diverse hardness which are sampled in CNS. From Fig. 5, we can observe that: Obviously, the hardness of the sampled negative instances increases with the inference phase of DMA, specifically demonstrated by these samples progressively moving closer to the representation of the corresponding user and positive samples in the low-dimensional space. That is, fewer reverse steps (e.g., 3T/4, T/2) result in noisier and simpler negative samples, whereas more steps sample more informative negative instances. This phenomenon highlights the effectiveness of our proposed CNS in boosting recommender optimization.

Conclusion

In this paper, we proposed a novel Curricular Conditioned Diffusion for Multimodal Recommendation (CCDRec) framework, which ingeniously integrates the reverse process of DM with the negative sampling process to select suitable negative instances. Initially, we propose the DMA to capture fine-grained relationships between multimodal features of items and aligning them with collaborative signals. Then, CCDRec introduces the NDI and CNS dynamically select negative samples of varying difficulty during training. The extensive evaluation on three real-world datasets and four base models verify the effectiveness of CCDRec. In the future, we will continue to explore the untapped potential of DM in negative sampling and investigate the effectiveness in other more challenging scenarios(Qi et al. 2024b,a), such as multimodal cross-domain or sequential recommendation.

Acknowledgments

This work is supported in part by the Shandong Province Excellent Young Scientists Fund Program (Overseas) (Grant no. 2022HWYQ-048), the Young Elite Scientists Sponsorship Program by CAST (2023QNRC001).

References

Bai, H.; Hou, M.; Wu, L.; Yang, Y.; Zhang, K.; Hong, R.; and Wang, M. 2023. Gorec: a generative cold-start recommendation framework. In *MM*, 1004–1012.

Chen, H.; Chen, Y.; Wang, X.; Xie, R.; Wang, R.; Xia, F.; and Zhu, W. 2021. Curriculum Disentangled Recommendation with Noisy Multi-feedback. *NIPS*, 34: 26924–26936.

Chen, X.; Chen, H.; Xu, H.; Zhang, Y.; Cao, Y.; Qin, Z.; and Zha, H. 2019. Personalized Fashion Recommendation with Visual Explanations based on Multimodal Attention Network. In *SIGIR*, 765–774.

Chen, Z.; Qi, Z.; Cao, X.; Li, X.; Meng, X.; and Meng, L. 2023a. Class-level Structural Relation Modeling and Smoothing for Visual Representation Learning. In *MM*, 2964–2972.

Chen, Z.; Qi, Z.; Li, X.; Wang, Y.; Meng, L.; and Meng, X. 2023b. Class-aware convolution and attentive aggregation for image classification. In *MM Asia*, 1–7.

Ding, J.; Quan, Y.; Yao, Q.; Li, Y.; and Jin, D. 2020. Simplify and robustify negative sampling for implicit collaborative filtering. *NIPS*, 33: 1094–1105.

Giannone, G.; Srivastava, A.; Winther, O.; and Ahmed, F. 2023. Aligning optimization trajectories with diffusion models for constrained design generation. *NIPS*, 36: 51830–51861.

Guo, H.; TANG, R.; Ye, Y.; Li, Z.; and He, X. 2017. DeepFM: A Factorization-Machine based Neural Network for CTR Prediction. In *IJCAI*.

He, R.; and McAuley, J. 2016. VBPR: visual bayesian personalized ranking from implicit feedback. In *AAAI*.

He, X.; Deng, K.; Wang, X.; Li, Y.; Zhang, Y.; and Wang, M. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *SIGIR*, 639–648.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. *NIPS*, 33: 6840–6851.

Hoogeboom, E.; Heek, J.; and Salimans, T. 2023. simple diffusion: End-to-end diffusion for high resolution images. In *ICML*, 13213–13232.

Huang, T.; Dong, Y.; Ding, M.; Yang, Z.; Feng, W.; Wang, X.; and Tang, J. 2021. Mixgcf: An improved training method for graph neural network-based recommender systems. In *KDD*, 665–674.

Li, X.; Liu, Y.; Lian, L.; Yang, H.; Dong, Z.; Kang, D.; Zhang, S.; and Keutzer, K. 2023. Q-diffusion: Quantizing diffusion models. In *ICCV*, 17535–17545.

Li, X.; Ma, H.; Meng, L.; and Meng, X. 2021. Comparative study of adversarial training methods for long-tailed classification. In *ADVM*, 1–7.

Li, X.; Meng, L.; Wu, L.; Li, M.; and Meng, X. 2024. Dreamfont3d: personalized text-to-3D artistic font generation. In *SIGGRAPH*, 1–11.

Li, Z.; Sun, A.; and Li, C. 2023. DiffuRec: A Diffusion Model for Sequential Recommendation. *ACM Transactions on Information Systems*, 42(3): 1–28.

Liu, T.; Qi, Z.; Chen, Z.; Meng, X.; and Meng, L. 2023. Cross-Training with Prototypical Distillation for improving the generalization of Federated Learning. In *ICME*, 648–653.

Ma, H.; Qi, Z.; Dong, X.; Li, X.; Zheng, Y.; Meng, X.; and Meng, L. 2023a. Cross-modal content inference and feature enrichment for cold-start recommendation. In *IJCNN*, 1–8.

Ma, H.; Xie, R.; Meng, L.; Chen, X.; Zhang, X.; Lin, L.; and Kang, Z. 2024a. Plug-in Diffusion Model for Sequential Recommendation. In *AAAI*, 8886–8894.

Ma, H.; Xie, R.; Meng, L.; Chen, X.; Zhang, X.; Lin, L.; and Zhou, J. 2023b. Exploring False Hard Negative Sample in Cross-Domain Recommendation. In *Recsys*, 502–514.

Ma, H.; Xie, R.; Meng, L.; Chen, X.; Zhang, X.; Lin, L.; and Zhou, J. 2024b. Triple sequence learning for cross-domain recommendation. *ACM Transactions on Information Systems*, 42(4): 1–29.

Ma, H.; Yang, Y.; Meng, L.; Xie, R.; and Meng, X. 2024c. Multimodal Conditioned Diffusion Model for Recommendation. In *WWW*, 1733–1740.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed Representations of Words and Phrases and their Compositionality. *NIPS*, 26.

Po, R.; Yifan, W.; Golyanik, V.; Aberman, K.; Barron, J. T.; Bermano, A.; Chan, E.; Dekel, T.; Holynski, A.; Kanazawa, A.; et al. 2024. State of the art on diffusion models for visual computing. In *Computer Graphics Forum*, volume 43, e15063.

Prabhudesai, M.; Goyal, A.; Pathak, D.; and Fragkiadaki, K. 2023. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*.

Qi, Z.; He, W.; Meng, X.; and Meng, L. 2024a. Attentive modeling and distillation for out-of-distribution generalization of federated learning. In *ICME*, 1–6.

Qi, Z.; Meng, L.; He, W.; Zhang, R.; Wang, Y.; Qi, X.; and Meng, X. 2024b. Cross-Training with Multi-View Knowledge Fusion for Heterogenous Federated Learning. *arXiv*.

Qi, Z.; Wang, Y.; Chen, Z.; Wang, R.; Meng, X.; and Meng, L. 2022. Clustering-based curriculum construction for sample-balanced federated learning. In *CAAI*, 155–166.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *EMNLP-IJCNLP*.

Rendle, S.; Freudenthaler, C.; Gantner, Z.; and Schmidt-Thieme, L. 2012. BPR: Bayesian personalized ranking from implicit feedback.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*, 10684–10695.

Sun, W.; Li, M.; Li, P.; Cao, X.; Meng, X.; and Meng, L. 2024. Sequential selection and calibration of video frames for 3D outdoor scene reconstruction. *CAAI Transactions on Intelligence Technology*.

Tao, Z.; Liu, X.; Xia, Y.; Wang, X.; Yang, L.; Huang, X.; and Chua, T.-S. 2022. Self-supervised Learning for Multimedia Recommendation. *IEEE Transactions on Multimedia*, 25: 5107–5116.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11).

Wallace, B.; Dang, M.; Rafailov, R.; Zhou, L.; Lou, A.; Purushwalkam, S.; Ermon, S.; Xiong, C.; Joty, S.; and Naik, N. 2024. Diffusion model alignment using direct preference optimization. In *CVPR*, 8228–8238.

Wang, Q.; Wei, Y.; Yin, J.; Wu, J.; Song, X.; and Nie, L. 2021. DualGNN: Dual Graph Neural Network for Multimedia Recommendation. *IEEE Transactions on Multimedia*, 1074–1084.

Wang, W.; Xu, Y.; Feng, F.; Lin, X.; He, X.; and Chua, T. 2023a. Diffusion Recommender Model. In *SIGIR*, 832–841.

Wang, Y.; Li, X.; Qi, Z.; Li, J.; Li, X.; Meng, X.; and Meng, L. 2022. Meta-causal feature learning for out-of-distribution generalization. In *ECCV*, 530–545.

Wang, Y.; Meng, L.; Ma, H.; Wang, Y.; Huang, H.; and Meng, X. 2024. Modeling Event-level Causal Representation for Video Classification. In *MM*, 3936–3944.

Wang, Y.; Qi, Z.; Li, X.; Liu, J.; Meng, X.; and Meng, L. 2023b. Multi-channel attentive weighting of visual frames for multimodal video classification. In *IJCNN*, 1–8.

Wei, Y.; Wang, X.; Nie, L.; He, X.; Hong, R.; and Chua, T.-S. 2019. MMGCN: Multi-modal Graph Convolution Network for Personalized Recommendation of Micro-video. In *MM*, 1437–1445.

Yang, X.; Zeng, J.; Guo, D.; Wang, S.; Dong, J.; and Wang, M. ???? Robust Video Question Answering via Contrastive Cross-Modality Representation Learning. *SCIENCE CHINA Information Sciences*, 67: 1–16.

Yang, Z.; Wu, J.; Wang, Z.; Wang, X.; Yuan, Y.; and He, X. 2024. Generate What You Prefer: Reshaping Sequential Recommendation via Guided Diffusion. volume 36.

Yu, M.; Wu, L.; Wang, C.; Meng, L.; and Meng, X. 2024. LayoutDM: Precision Multi-Scale Diffusion for Layout-to-Image. In *ICME*, 1–6.

Yu, P.; Tan, Z.; Lu, G.; and Bao, B.-K. 2023. LD4MRec: Simplifying and Powering Diffusion Model for Multimedia Recommendation. *arXiv*.

Zhang, J.; Zhu, Y.; Liu, Q.; Wu, S.; Wang, S.; and Wang, L. 2021. Mining Latent Structures for Multimedia Recommendation. In *MM*, 3872–3880.

Zhang, W.; Chen, T.; Wang, J.; and Yu, Y. 2013. Optimizing top-n collaborative filtering via dynamic negative item sampling. In *SIGIR*, 785–788.

Zheng, Y.; Li, Z.; Li, X.; Liu, J.; Wang, Y.; Meng, X.; and Meng, L. 2024. Unifying Visual and Semantic Feature Spaces with Diffusion Models for Enhanced Cross-Modal Alignment. In *ICANN*, 110–125.

Zhong, S.; Huang, Z.; Li, D.; Wen, W.; Qin, J.; and Lin, L. 2024. Mirror Gradient: Towards Robust Multimodal Recommender Systems via Exploring Flat Local Minima. In *WWW*, 3700–3711.

Zhou, H.; Zhou, X.; and Shen, Z. 2023. Enhancing Dyadic Relations with Homogeneous Graphs for Multimodal Recommendation. 3123–3130.

Zhou, X. 2023. MMRec: Simplifying Multimodal Recommendation. *arXiv*.

Zhou, X.; and Shen, Z. 2023. A Tale of Two Graphs: Freezing and Denoising Graph Structures for Multimodal Recommendation. In *MM*, 935–943.

Zhou, X.; Zhou, H.; Liu, Y.; Zeng, Z.; Miao, C.; Wang, P.; You, Y.; and Jiang, F. 2023. Bootstrap Latent Representations for Multi-modal Recommendation. In *WWW*, 845–854.