# Multi-Modal Diffusion Model for Cross-Domain Recommendation

Anonymous

***Abstract*—Cross-domain recommendation (CDR) aims to utilize information from multiple domains to alleviate data sparsity and user cold-start problems in recommendation systems. However, current CDR methods primarily rely on ID-based features, which fail to capture the detailed characteristics of items. This limitation restricts their effectiveness, particularly in addressing the item cold-start problem, where new or rarely interacted items are difficult to recommend. To address this issue, we propose a Multi-Modal Diffusion model for Cross-Domain Recommendation (MMDCDR). Instead of using traditional ID-based representations, we extract item multi-modal (e.g., text and image) representations from pre-trained models, which effectively disentangles fine-grained item features and alleviates the cold-start problem for items. To enhance CDR with multi-modal data, we use a diffusion model to transfer information across domains. By conditioning on the user's representation in the source domain, our model progressively denoises and generates the user's representation in the target domain. Then, we design a modality-aware contrastive augmentation strategy to enhance the consistency between multi-modal representations. Extensive experiments on large real-world datasets demonstrate the effectiveness and superiority of MMDCDR in cold-start scenarios. The code has been available at https://anonymous.4open.science/r/MMDCDR-Oner.**

***Index Terms*—cross-domain recommendation, multi-modal, diffusion model, cold-start problem.**

## I. Introduction

In the era of exponential information growth, recommendation systems have become essential tools for assisting users in efficiently identifying relevant content within extensive datasets [1], [2]. However, traditional recommendation systems heavily rely on abundant historical user data. When encountering new users, known as cold-start users, limited behavioral data often leads to poor recommendation performance [3], [4].

To address this challenge, cross-domain recommendation (CDR) [5] has garnered significant attention as a method that transfers knowledge from a source domain to enhance recommendation performance in a target domain [6]. The core task of CDR is to bridge user's preferences in the source domain and the target domain, also called preference transfer [7]. CDR have shown great success in alleviating the serious data sparsity and user cold-start problems [8], [9]. Generally, most existing CDR methods [7], [10]–[12] adopt the concept of embedding and mapping. Specifically, these methods usually use collaborative filtering models to separately obtain ID representations in the source and target domains, and then train a mapper using a number of users existing in both domains (i.e., overlapping users) to map user representations from the source domain to the target domain.

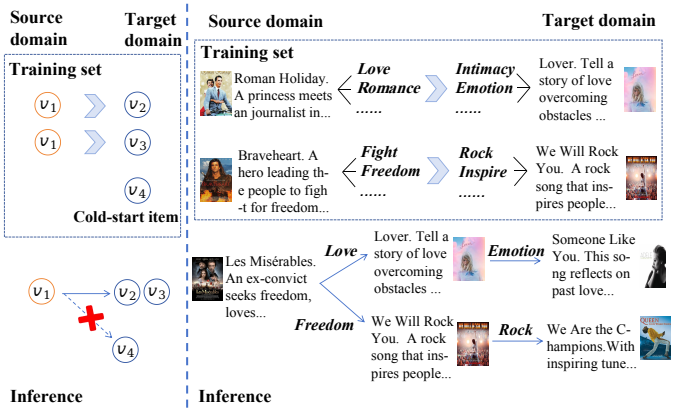


Fig. 1. A comparison of cross-domain knowledge transfer based on ID information (left) and multi-modal information (right).

Other studies [13]–[15] use meta-learning methods to capture specific user preferences in CDR.

However, most existing work relies on ID information to learn knowledge transfer across domains. But item IDs are merely symbolic identifiers and fail to capture detailed item features [16], [17]. From a statistical perspective, they capture only the co-occurrence patterns of preferences across domains. Moreover, the absence of intrinsic item features in IDs limits the model's generalization to new or rarely interacted items, which is called item cold-start problem. As illustrated in the left part of Figure 1, the co-occurrence pattern indicates that if a user interacts with item $v_1$ in the source domain and also with items $v_2$ or $v_3$ in the target domain, recommending $v_2$ and $v_3$ to a cold-start user in the target domain who interacted with $v_1$ is a reasonable strategy. However, for cold-start items like $v_4$, which have not been interacted with by any users, ID-based methods lack sufficient historical interaction data to recommend them directly, thereby highlighting the significant challenge posed by the item cold-start problem.

To address these limitations, we replace ID information with multi-modal features in CDR. Multi-modal features, such as text descriptions and images, can disentangle fine-grained characteristics of items. As shown in the right part of Figure 1, multi-modal representations can learn knowledge transfer patterns between items' fine-grained features (noted in italics in the figure). For a cold-start user in the target domain who has interacted with *Les Misérables* in the source domain, we can learn that the fine-grained features of *Les Misérables* include ***Love***, ***Freedom***, etc., which also appear in

*Roman Holiday* and *Braveheart*. Therefore, recommending the corresponding songs *Lover* and *We Will Rock You* is a wise strategy. Moreover, based on the items' fine-grained features, the model can recommend similar songs such as *Someone Like You* and *We Are the Champions* , even if they are cold-start items. **In conclusion, unlike ID information that can only capture cross-domain co-occurrence behaviors, multi-modal information can learn transfer patterns between the fine-grained features of cross-domain items, thereby improving the model's ability to address the item cold-start problem.**

Although introducing multi-modal features is a promising direction, applying them to CDR remains a challenge. Existing methods mainly rely on mapping techniques that align ID-based representations from the source domain to the target domain. However, these methods do not achieve better results on multi-modal features because they have difficulty disentangling fine-grained item features, making it challenging to capture the complex and subtle differences between domains. Additionally, these methods focus on individual data samples rather than capturing broader distribution patterns, limiting their ability to generalize to unseen data. In contrast, diffusion models progressively refine noisy inputs and naturally model complex data distributions [18], making them well-suited for estimating target embedding distributions in CDR.

To handle the challenges mentioned above, we propose a novel framework named Multi-Modal Diffusion model for Cross-Domain Recommendation (MMDCDR). Our method leverages rich multi-modal representations to replace the ID-based representations in traditional methods. Specifically, we first obtain the initial textual and visual representations of items using pre-trained models. Inspired by the success of Diffusion Models (DMs) in image generation tasks [19], [20], we then employ conditional diffusion models to transfer multi-modal representations across domains. In detail, we progressively add noise to the multi-modal representations of overlapping users in the target domain. Then, we use the source domain's multi-modal representations as conditions to guide the model in denoising and progressively generating the user's representation in the target domain. Furthermore, to improve the consistency of representations between different modalities, we propose a modality-aware contrastive enhancement module. By applying diffusion models to multi-modal feature CDR, we not only enhance the fine-grained transfer of item features across domains, but also improve the model's ability to generalize to cold-start items. Furthermore, by leveraging the advantages of the diffusion process, the model can learn to transfer patterns between multi-modal data distributions, providing a powerful research direction for cross-domain transfer techniques. The contributions of this paper are concluded as follows:

1) We deeply investigate the differences between ID-based information and multi-modal information in CDR. By utilizing multi-modal information, we model the fine-grained features of items, which enhances the transfer of user preferences across domains and effectively addresses the item cold-start problem.
2) We propose a multi-modal diffusion model for CDR, called MMDCDR. We employ a conditional diffusion model to transfer multi-modal features from the source domain to the target domain, and then design a contrastive learning framework to enhance the multi-modal representations. To the best of our knowledge, this is the first work to apply diffusion models based on multi-modal information in the field of CDR.
3) We conduct comprehensive experiments on the Amazon review dataset to validate the effectiveness and adaptability of the MMDCDR model in cold-start CDR scenarios. The experimental results demonstrate that our model significantly outperforms existing methods.

The rest of our paper is organized as follows. We briefly review the related work in Section II. In Section III, the proposed MMDCDR is described in detail. The experimental results are reported in Section IV. At last, we draw the conclusion in Section V.

## II. Related work

### A. Cross-Domain Recommendation

In recent years, numerous CDR methods have been proposed. Since our work focuses specifically on the cold-start recommendation problem, we only focus on CDR approaches which recommend items to cold-start users. Early CDR methods, such as Collective Matrix Factorization (CMF) [5], integrate knowledge across domains by combining rating matrices and sharing user factors. Recently, the embedding and mapping paradigm (EMCDR) [10] are introduced to transfer representations across domain. SSCDR [11] proposes a semi-supervised mapping function for cross-domain transfer. LACDR [12] further refines this approach by aligning user representations in a low-dimensional space. Another line of research applies meta-learning to CDR. PTUPCDR [15] and TMCDR [13] replace the mapping function with a meta-network to enhance transferability. DREAM [21] enhances CDR by decoupling preferences, leveraging contrastive learning and focal loss for better performance. CDRNP [22] leverages neural processes to capture both user-specific preferences and correlations among users. CSNBR [23] addresses the negative transfer problem in CDR through graph reconstruction. However, existing methods rely solely on ID-based numerical information for transfer, which offers limited effectiveness in capturing item features and user behavioral preferences.

### B. Diffusion Models in Recommendation

Diffusion Models, known for their effectiveness in uncertainty injection and data augmentation in image synthesis [24], have recently been explored in recommendation systems. DiffuRec [25] utilize the DM for the sequential recommendation. DiffRec [26] generates globally similar but personalized collaborative signals during the denoising process. DiffKG [27] applies knowledge graphs to DMs. Recent methods employ conditional diffusion models for recommendation tasks. By introducing conditional variables, DMs guide and constrain the

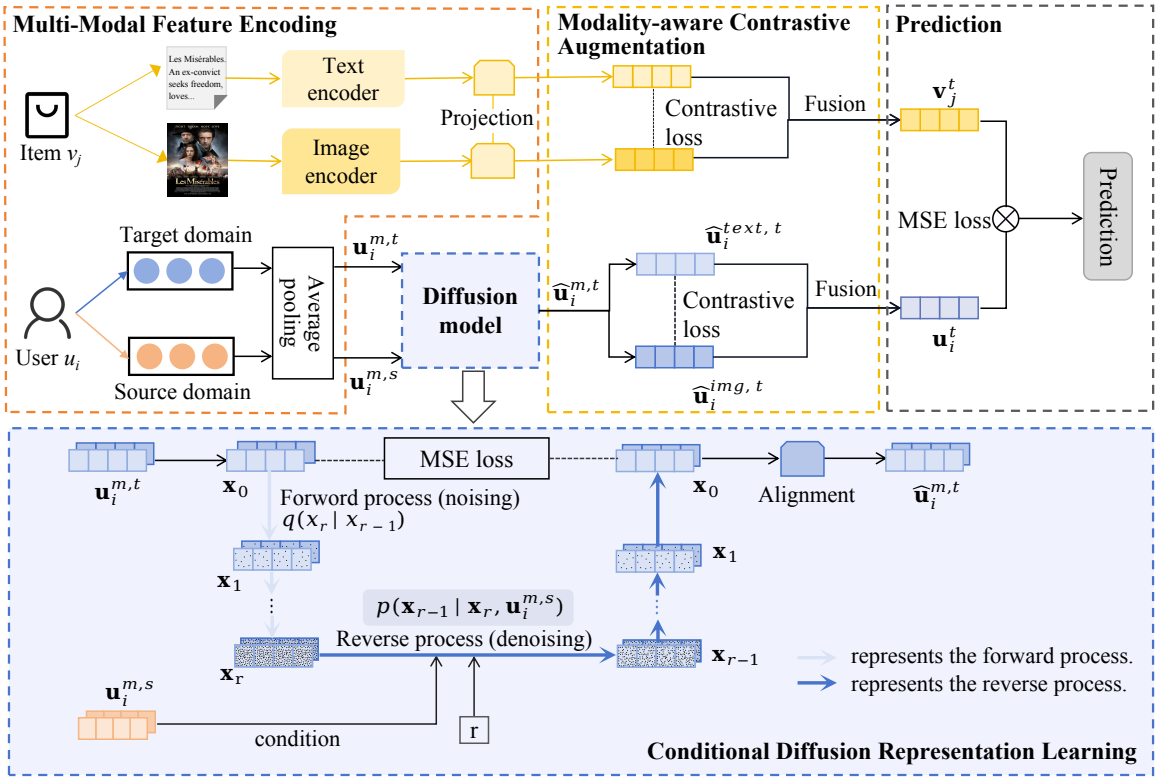


Fig. 2. The overall framework of the proposed MMDCDR.

generative process, enabling the model to produce target data that aligns with the specified conditions. MCDRec [28] integrates multi-modal data into the collaborative representation of items by using it as a conditioning factor. DiffuASR [29] utilizes item sequences as conditions to perform data augmentation. Inspired by conditional diffusion models, we introduce this approach into CDR. We hypothesize that using user representations from the source domain as conditions to guide the generation of target domain representations is a promising direction.

### C. Multi-modal Recommendation

The multi-modal recommendation aims to integrate the multi-modal features of items into representation learning, addressing the challenge of data sparsity in recommendation systems [28]. Previously, attention-based models such as ACF [30] and VECF [31] utilize multi-modal content to capture complex user preferences. Inspired by Graph Neural Networks (GNNs), MMGCN [32] employs graph convolutional networks for feature aggregation to enhance user and item representations. MMSR [33] adopts a dual-attention mechanism to distinguish between homogeneous and heterogeneous neighbor nodes. DRAGON [34] builds separate homogeneous and heterogeneous graphs to obtain dual representations for users and items. However, there is limited research on incorporating multi-modal information into CDR. We argue that leveraging multi-modal information for cross-domain knowledge transfer is more effective than relying on traditional collaborative information.

## III. Method

### A. Problem Formulation

In the context of CDR, we consider both a source domain and a target domain. Each domain consists of a set of users, $\mathcal{U} = \{u_1, u_2, \ldots\}$, a set of items $\mathcal{I} = \{v_1, v_2, \ldots\}$ and a rating matrix $\mathcal{R}$. Each element $r_{ij} \in \mathcal{R}$ represents the rating between user $u_i$ and item $v_j$. To differentiate between the two domains, we denote the user, item sets, and the rating matrix of the source domain as $\mathcal{U}^s, \mathcal{I}^s, \mathcal{R}^s$, while $\mathcal{U}^t, \mathcal{I}^t, \mathcal{R}^t$ for the target domain. The set of overlapping users is defined as $\mathcal{U}^o = \mathcal{U}^s \cap \mathcal{U}^t$. In contrast, $\mathcal{I}^s$ and $\mathcal{I}^t$ are disjoint, implying that there are no shared items between the two domains.

### B. Framework

Figure 2 illustrates the proposed MMDCDR. First, we obtain initial multi-modal representations of items and users using pre-trained models. To facilitate knowledge transfer between domains, we introduce a diffusion model to generate user representations in the target domain. Specifically, we progressively corrupt the original user representations in the target domain. By conditioning on the user's representation in the source domain, we guide the model to iteratively recover the user's representation in the target domain through a denosing process. Finally, we introduce modality-aware contrastive learning to enhance the multi-modal representations of users

and items. In this way, the model learns modality-invariant features and effectively aligns the multi-modal representations.

### C. Multi-Modal Feature Encoding

To effectively capture the multi-modal characteristics of items and users, we employ pre-trained models to extract rich representations. For item encoding, we utilize a pre-trained RoBERTa [35] model to obtain textual representations of the item's descriptive content, and a pre-trained Vision Transformer (ViT) [36] model to derive visual representations from the item's image.

To simplify notation, we denote the multi-modal representation of item $v_j \in \mathcal{I}$ as $\mathbf{e}_j^m$, where $m \in \{\text{text}, \text{img}\}$ represents the modality (i.e., text or image). Since the item features obtained from different modality encoders reside in different feature spaces, we apply a linear projection function to map each feature vector $\mathbf{e}_j^m$ of item $v_j$ into a shared low-dimensional latent space.

$$\mathbf{v}_j^m = \mathbf{e}_j^m \mathbf{W}_m + \mathbf{b}_m, \tag{1}$$

where $\mathbf{W}_m \in \mathbb{R}^{d_m \times d}$, $\mathbf{b}_m \in \mathbb{R}^d$ denote the linear transformation matrix and bias in the linear projection function. In this way, the representations of each modality $\mathbf{v}_j^m$ can be aligned into the same latent space.

For user representations, we aggregate the multi-modal embeddings of all items the user has interacted with. Specifically, we apply average pooling to the textual and visual embeddings of the items that user $u_i \in \mathcal{U}^o$ has interacted with to generate the user's multi-modal representations $\mathbf{u}_i^{\text{text}}$ and $\mathbf{u}_i^{\text{img}}$. For simplicity in subsequent representations, we denote the multi-modal representation of user $u_i$ as $\mathbf{u}_i^m$. To explicitly distinguish between the source and target domains, we denote the source domain and target domain multi-modal representations of user $u_i$ as $\mathbf{u}_i^{m,s} \in \mathbb{R}^d$and $\mathbf{u}_i^{m,t} \in \mathbb{R}^d$, respectively.

Through this method, we obtain the initial multi-modal representations of users and items, enabling the extraction of more detailed item content information and user behavioral preferences. Even for items with little or no interactions, leveraging multi-modal information allows for effective recommendations.

### D. Conditional Diffusion Representation Learning

In this section, we introduce diffusion model into CDR. Similar to the classical DMs, our method comprises two processes: a forward process and a reverse process. The forward process involves gradually adding Gaussian noise to perturb the original data distribution. Conversely, the reverse process progressively restores the perturbed representations from a disordered state back to the representation space. The difference is that we incorporate the user's source domain representation as a condition into the diffusion process, aiming to guide the generation of the user's target domain representation. Unlike traditional mapping methods, DM estimates the user embedding distribution in the target domain while accounting for specific user preferences in the source domain. This enables more effective cross-domain transfer of multi-modal features, making the diffusion model a promising approach for knowledge transfer in CDR.

*1) Forward Process:* In the forward process of the DM, we progressively add noise to the multi-modal feature representation $\mathbf{u}_i^{m,t}$ of user $u_i \in \mathcal{U}^o$ in the target domain, where $m \in \{\text{text}, \text{img}\}$. We initialize the diffusion process with $\mathbf{x}_0 = \mathbf{u}_i^{m,t}$. The forward process is defined as:

$$q(\mathbf{x}_r \mid \mathbf{x}_{r-1}) := \mathcal{N}(\mathbf{x}_r; \sqrt{1-\beta_r}\mathbf{x}_{r-1}, \beta_r \mathbf{I}), \tag{2}$$

where $r \in \{1, 2, \ldots, R\}$ represents the diffusion timestep, $R$ is the upper limit of the diffusion step, $\beta_r \in (0, 1)$ denotes the added Gaussian noise scale at step $r$, and $\mathcal{N}$ denotes the Gaussian distribution. As $R \to \infty$, $\mathbf{x}_R$ converges to a Gaussian distribution.

By applying the reparameterization trick and leveraging the additive property of independent Gaussian noise, $\mathbf{x}_r$ can be derived directly from $\mathbf{x}_0$, as:

$$q(\mathbf{x}_r \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_r; \sqrt{\bar{\alpha}_r}\mathbf{x}_0, (1-\bar{\alpha}_r)\mathbf{I}), \tag{3}$$

where $\alpha_r = 1 - \beta_r$ and $\bar{\alpha}_r = \prod_{r'=1}^{r} \alpha_{r'}$. This formulation directly samples $\mathbf{x}_r$ at any timestep $r$ from the clean multi-modal feature $\mathbf{x}_0$, without needing to iterate over all previous steps.

*2) Reverse Process:* The reverse process is the core phase of the DM, where we iteratively denoise the noisy representation $\mathbf{x}_R$ over $R$ time steps to reconstruct the original target domain representation $\mathbf{x}_0$. To enhance the generation process specific to the target domain, we leverage the source domain user representations $\mathbf{u}_i^{m,s}$ as conditional guidance. At each reverse step $r$, the reverse process unfolds as follows:

$$p_\theta(\mathbf{x}_{r-1} \mid \mathbf{x}_r, \mathbf{u}_i^{m,s}) = \mathcal{N}(\mathbf{x}_{r-1}; \mu_\theta(\mathbf{x}_r, \mathbf{u}_i^{m,s}, r), \Sigma_\theta(\mathbf{x}_r, r)), \tag{4}$$

where $\mu_\theta$ and $\Sigma_\theta$ are parameterized functions that predict the mean and variance of the denoised representation, respectively. $\theta$ represents the learnable parameters of the model. Based on the Gaussian distribution parameterization, we can set $\Sigma_\theta = \sigma_r^2 \mathbf{I}$ as constants, where $\sigma_r^2 = \frac{1-\bar{\alpha}_{r-1}}{1-\bar{\alpha}_r}\beta_r$. As for the mean of the distribution, it can be expressed as:

$$\mu_\theta(\mathbf{x}_r, \mathbf{u}_i^{m,s}, r) = \frac{1}{\sqrt{\alpha_r}}\left(\mathbf{x}_r - \frac{\beta_r}{\sqrt{1-\bar{\alpha}_r}}\epsilon_\theta(\mathbf{x}_r, \mathbf{u}_i^{m,s}, r)\right), \tag{5}$$

where $\epsilon_\theta(\mathbf{x}_r, \mathbf{u}_i^{m,s}, r)$ is the noise estimation network, which estimates the noise present in $\mathbf{x}_r$ at step $r$ under the guidance of $\mathbf{u}_i^{m,s}$.

In image generation tasks, the noise estimation network $\epsilon_\theta$ is commonly implemented using U-Net architectures [37]. However, in our task, the data consists of low-dimensional, non-image representations, making the use of U-Nets less appropriate. We construct the noise prediction model $\epsilon_\theta$ using $L$ layers of perceptrons, which already achieve strong performance in our experiments.

*3) Classifier-Free Guidance:* To further enhance the flexibility of the conditional generation, we adopt a classifier-free guidance strategy [20]. This technique allows us to modulate the influence of the conditional information directly within the

diffusion model, without the need for an auxiliary classifier. By interpolating between conditional and unconditional noise estimates, we gain finer control over the generation process, leading to improved sample quality and better alignment with the conditioning signals. The core idea behind classifier-free guidance is to use a single noise estimation function $\epsilon_\theta$ that is capable of predicting both conditional and unconditional noise.

During the reverse process, we adjust the noise prediction to balance between unconditional and conditional estimates. The adjusted noise prediction function $\hat{\epsilon}_\theta$ is defined as:

$$\hat{\epsilon}_\theta(\mathbf{x}_r, \mathbf{u}_i^{m,s}, r) = \epsilon_\theta(\mathbf{x}_r, r) + g \cdot \left(\epsilon_\theta(\mathbf{x}_r, \mathbf{u}_i^{m,s}, r) - \epsilon_\theta(\mathbf{x}_r, r)\right), \tag{6}$$

where $\mathbf{x}_r$ is the noisy data at time step $r$. $\mathbf{u}_i^{m,s}$ is the modality-specific conditional information. $\epsilon_\theta(\mathbf{x}_r, \mathbf{u}_i^{m,s}, r)$ is the conditional noise estimation function, predicting the noise given $\mathbf{x}_r$ and conditioning $\mathbf{u}_i^{m,s}$. $\epsilon_\theta(\mathbf{x}_r, r)$ is the unconditional noise estimation function, predicting the noise given $\mathbf{x}_r$ alone. Specifically, with a certain probability, we replace the conditioning vector $\mathbf{u}_i^{m,s}$ with a zero vector during training. $g \geq 0$ is the guidance scale that controls the strength of the conditioning influence.

*4) Training Objective:* The training objective of the diffusion model is to minimize the Mean-Squared Error (MSE) loss between the true noise $\epsilon$ and the predicted noise $\hat{\epsilon}_\theta$:

$$\mathcal{L}_{\mathrm{DM}_\theta} = \mathbb{E}_{r\in[0,R],\mathbf{x}_0\sim q(\mathbf{u}_i^{m,t}),\epsilon\sim\mathcal{N}(0,\mathbf{I})}\left[\|\epsilon - \hat{\epsilon}_\theta(\mathbf{x}_r, \mathbf{u}_i^{m,s}, r)\|^2\right]. \tag{7}$$

*5) Inference of Conditional Diffusion Model:* During the inference phase, we start with Gaussian noise $\epsilon'$ and use the user's representation in the source domain $\mathbf{u}_i^{m,s}$ as guidance to iteratively perform reverse denoising through the denoiser. Specifically, we set $\hat{\mathbf{x}}_R = \epsilon'$ to execute reverse denoising $\hat{\mathbf{x}}_R \rightarrow \hat{\mathbf{x}}_{R-1} \rightarrow \ldots \rightarrow \hat{\mathbf{x}}_0$ for $R$ steps. Finally, we take $\hat{\mathbf{x}}_0$ as the user's representation in the target domain, denoted as $\hat{\mathbf{u}}_i^{m,t}$.

Following previous work [18], we introduce an alignment module to reduce the randomness inherent in the DM. Specifically, we implement this alignment using a Multilayer Perceptron (MLP):

$$\hat{\mathbf{u}}_i^{m,t} = \mathrm{MLP}(\hat{\mathbf{u}}_i^{m,t}). \tag{8}$$

In this way, we can adjusts the predicted representation $\hat{\mathbf{u}}_i^{m,t} \in \mathbb{R}^d$ to better match the ground-truth user representation in the target domain.

### *E. Modality-aware Contrastive Augmentation*

In multi-modal recommendation scenarios, users interact with these multi-modal items in ways that exhibit consistent patterns across different modalities. For instance, a user interested in action movies may demonstrate a preference for both the visual effects and the exciting plot descriptions of such movies. Capturing these cross-modal consistencies is crucial for understanding user preferences and enhancing recommendation performance.

Building upon the DM introduced in section III-D, we obtain the user's multi-modal representations in the target domain, denoted as $\hat{\mathbf{u}}_i^{\text{text},t}$ and $\hat{\mathbf{u}}_i^{\text{img},t}$ for the textual and visual modalities, respectively. To effectively fuse these multi-modal representations and enhance the cross-modal consistencies, we propose a multi-modal contrastive enhancement module.

For each user $u_i \in \mathcal{U}^o$, we consider the representations from different modalities as positive pairs. The positive sample for user $u_i$ is the pair $(\hat{\mathbf{u}}_i^{\text{text},t}, \hat{\mathbf{u}}_i^{\text{img},t})$, representing the same user's embeddings in the textual and image modalities in target domain. Negative samples are formed by pairing the representation of user $u_i$ in one modality with the representations of other users $u_z \neq u_i$ in $\mathcal{U}^o$ in the alternative modality. The contrastive loss for user representations is defined as:

$$\mathcal{L}_{\mathrm{cl}}^{\mathrm{user}} = -\sum_{u_i\in\mathcal{U}^o} \log \frac{\exp\left(\mathrm{sim}\left(\hat{\mathbf{u}}_i^{\text{text},t}, \hat{\mathbf{u}}_i^{\text{img},t}\right)/\tau\right)}{\sum_{u_z\in\mathcal{U}^o}\exp\left(\mathrm{sim}\left(\hat{\mathbf{u}}_i^{\text{text},t}, \hat{\mathbf{u}}_z^{\text{img},t}\right)/\tau\right)}, \tag{9}$$

where $\mathrm{sim}(\cdot)$ denotes a cosine similarity function, $\tau > 0$ is a temperature hyperparameter that controls the concentration of the distribution.

Due to computational constraints, we compute the loss over mini-batches instead of the entire user set. In each training iteration, we sample a mini-batch of users, using other users within the same batch as negative samples.

Similarly, we calculate the contrastive learning loss for the item side as $\mathcal{L}_{\mathrm{cl}}^{\mathrm{item}}$ in a similar way.

By integrating the contrastive learning losses from both users and items, we define the overall multi-modal contrastive loss:

$$\mathcal{L}_{\mathrm{cl}} = \mathcal{L}_{\mathrm{cl}}^{\mathrm{user}} + \mathcal{L}_{\mathrm{cl}}^{\mathrm{item}}. \tag{10}$$

This combined loss function guides the model to align multi-modal representations consistently for both users and items, effectively fusing information across modalities.

### *F. Prediction*

To construct comprehensive embeddings that integrate multi-modal information, we concatenate the representations from different modalities and apply linear transformations for both users and items. For each user $u_i$, the fused user representation $\mathbf{u}_i^t$ in the target domain is obtained by:

$$\mathbf{u}_i^t = \mathbf{W}_u \mathrm{concat}\left(\hat{\mathbf{u}}_i^{\text{text},t}, \hat{\mathbf{u}}_i^{\text{img},t}\right) + \mathbf{b}_u, \tag{11}$$

where $\hat{\mathbf{u}}_i^{\text{text},t} \in \mathbb{R}^d$ and $\hat{\mathbf{u}}_i^{\text{img},t} \in \mathbb{R}^d$ are the user's representations in the textual and visual modalities. $\mathbf{W}_u \in \mathbb{R}^{2d\times d}$ is the weight matrix for fusion. $\mathbf{b}_u \in \mathbb{R}^d$ is the bias vector.

Similarly, we can obtain the fused representation of the item $v_j$ in the target domain, denoted as $\mathbf{v}_j^t$. With the fused user and item representations, we compute the predicted rating $\hat{r}_{ij}$ as the dot product of the fused user and item representations:

$$\hat{r}_{ij} = \mathbf{u}_i^{t\top}\mathbf{v}_j^t. \tag{12}$$

We compare the predicted rating $\hat{r}_{ij}$ with the ground truth rating $r_{ij}$ using the MSE loss:

$$\mathcal{L}_{\mathrm{task}} = \frac{1}{|\mathcal{D}|}\sum_{(u_i,v_j)\in\mathcal{D}}\left(\hat{r}_{ij} - r_{ij}\right)^2, \tag{13}$$

TABLE I
STATISTICS OF THREE CDR TASKS. (OVERLAP DENOTES THE NUMBER OF OVERLAPPING USERS.)

| Scenarios | #Items | | #Users | | | #Rating | |
|---|---|---|---|---|---|---|---|
| | Source | Target | Overlap | Source | Target | Source | Target |
| Movie2Music | 11846 | 28591 | 24513 | 260777 | 107998 | 1005241 | 616257 |
| Sport2Phone | 61647 | 39497 | 32243 | 327849 | 157150 | 1946125 | 953738 |
| Electronic2Phone | 125025 | 39497 | 83206 | 728433 | 157150 | 5783144 | 953738 |

where $\mathcal{D}$ is the set of observed user-item interactions in the training data.

To effectively train the model components, we employ a two-stage training procedure that separates the optimization of the diffusion model from the joint optimization of the fusion and prediction tasks.

In the first stage, we train the diffusion model to generate accurate representations for users by minimizing the diffusion loss $\mathcal{L}_{\text{DM}}$ defined in Equation 7.

In the second stage, we jointly optimize the task loss and the contrastive loss defined in Equation 10.

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda \mathcal{L}_{\text{cl}}, \tag{14}$$

where $\lambda \geq 0$ is a hyperparameter controlling the balance between the task loss and the contrastive loss.

## IV. EXPERIMENT

Extensive experiments are conducted to evaluate the effectiveness of the proposed model. The experiments aim to answer the following research questions:

RQ1: Does the proposed MMDCDR outperform the state-of-the-art CDR methods?

RQ2: How do different components contribute to MMDCDR?

RQ3: How do different choices of hyper-parameters affect the performance of MMDCDR?

### A. Experimental Setting

*1) Datasets:* Following existing methods [6], [15] on CDR, we evaluate MMDCDR on a real-world recommendation dataset, namely the Amazon review dataset [1] [38]. We use the 5-core subsets, in which all users and items have at least 5 reviews. Specifically, we select five types of datasets: Movies and TV (Movie), CDs and Vinyl (Music), Sports and Outdoors (Sport), Cell Phones and Accessories (Phone), and Electronics (Electronic). These datasets can be divided into three different cross-domain tasks: Movie2Music (Task 1), Sport2Phone (Task 2), and Electronic2Phone (Task 3). Statistics of these datasets are listed in Table I.

[1]https://cseweb.ucsd.edu/ jmcauley/datasets/amazon_v2/

*2) Evaluation Metrics:* Amazon review dataset contains rating data (0 - 5 score). Following previous work [10], [15], we adopt Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) as evaluation metrics to assess the recommendation performance.

MAE measures the average absolute difference between the predicted ratings $\hat{r}_{ij}$ and the ground truth ratings $r_{ij}$, and is defined as:

$$\text{MAE} = \frac{1}{|\mathcal{T}|} \sum_{(i,j)\in\mathcal{T}} |\hat{r}_{ij} - r_{ij}|, \tag{15}$$

where $\mathcal{T}$ denotes the set of user-item pairs in the test set.

RMSE, on the other hand, penalizes larger errors more heavily by computing the square root of the average squared differences between predicted and actual ratings:

$$\text{RMSE} = \sqrt{\frac{1}{|\mathcal{T}|} \sum_{(i,j)\in\mathcal{T}} (\hat{r}_{ij} - r_{ij})^2}. \tag{16}$$

Both metrics provide complementary insights into prediction accuracy: while MAE offers a more robust measure against outliers, RMSE emphasizes larger deviations and is more sensitive to them. Lower values of MAE and RMSE indicate better predictive performance.

*3) Baselines:* For a fair comparison, we have selected the following CDR methods as baselines.

1) **CMF** [5] shares same embeddings for overlapping users belonging to different domains.
2) **EMCDR** [10] follows the embedding and mapping paradigm, learning a mapping function to transfer user representations.
3) **SSCDR** [11] improves upon EMCDR by introducing a semi-supervised strategy for learning the cross-domain mapping function.
4) **LACDR** [12] uses an encoder-decoder structure to construct a non-linear and more expressive mapping function.
5) **PTUPCDR** [15] introduces a personalized transfer mechanism by designing a meta-network that generates user-specific bridge functions.
6) **P2M2-CDR** [39] proposes a privacy-preserving multi-modal CDR framework that incorporates both textual and visual modalities of items.
7) **DiffCDR** [18] introduces diffusion model to transfer user representations.

*4) Implementation and Hyperparameter Setting:* In this paper, we focus on the scenario of cold-start users in the target domain. Specifically, we only consider the overlapping users during the training, validation, and testing phases. The proposed method learns to transfer the behavioral preferences of overlapping users from the source domain to the target domain. To evaluate the effectiveness of the model, the overlapping user set $\mathcal{U}^o$ is divided into training, validation, and testing subsets. The rating interactions of users in both the source and target domains within the training set are used to train the CDR model. Meanwhile, the rating interactions of users in the

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT BASELINE METHODS.THE BEST AND THE SECOND-BEST PERFORMANCE IS BOLD AND UNDERLINED RESPECTIVELY.

| Task | $\alpha$ | Metric | CMF | EMCDR | SSCDR | LACDR | PTUPCDR | P2M2-CDR | DiffCDR | MMDCDR | Improve |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Movie2Music | 80% | MAE | 1.6044 | 1.5089 | 1.0340 | 1.2501 | 1.0395 | 1.0249 | 1.0238 | **0.9531** | 6.91% |
| | | RMSE | 2.1247 | 1.8292 | 1.3551 | 1.6208 | 1.3859 | 1.2998 | 1.3310 | **1.2563** | 3.35% |
| | 50% | MAE | 1.8033 | 1.9113 | 1.2738 | 1.4877 | 1.1913 | 1.0386 | 1.1870 | **0.9752** | 6.10% |
| | | RMSE | 2.3700 | 2.2242 | 1.5843 | 1.8988 | 1.6050 | 1.3334 | 1.5719 | **1.2671** | 4.97% |
| | 20% | MAE | 2.3634 | 2.2003 | 1.7012 | 1.7060 | 1.3805 | 1.0761 | 1.3764 | **1.0041** | 6.69% |
| | | RMSE | 3.0780 | 2.5253 | 2.0550 | 2.1876 | 1.8817 | 1.3360 | 1.8686 | **1.2891** | 3.51% |
| Sport2Phone | 80% | MAE | 1.4962 | 1.4924 | 1.4027 | 1.3294 | 1.0555 | 1.0709 | 1.0733 | **0.9020** | 14.54% |
| | | RMSE | 1.9450 | 1.7231 | 1.5634 | 1.6012 | 1.3928 | 1.3884 | 1.4213 | **1.1749** | 15.37% |
| | 50% | MAE | 1.8313 | 1.6107 | 1.4742 | 1.4397 | 1.1226 | 1.0915 | 1.1214 | **0.9531** | 12.68% |
| | | RMSE | 2.3405 | 1.8794 | 1.6976 | 1.7507 | 1.4823 | 1.4015 | 1.5053 | **1.3175** | 5.99% |
| | 20% | MAE | 2.8392 | 1.9606 | 1.8515 | 1.7915 | 1.3165 | 1.1201 | 1.3478 | **1.1043** | 1.41% |
| | | RMSE | 3.4482 | 2.2481 | 2.1116 | 2.1290 | 1.7480 | 1.4741 | 1.7740 | **1.4131** | 4.14% |
| Electric2Phone | 80% | MAE | 1.4595 | 1.3762 | 1.2977 | 1.2192 | 1.0440 | 1.1201 | 1.1202 | **0.9020** | 13.60% |
| | | RMSE | 1.9507 | 1.6057 | 1.5179 | 1.4864 | 1.3796 | 1.4253 | 1.4888 | **1.2267** | 11.08% |
| | 50% | MAE | 1.4679 | 1.8947 | 1.7423 | 1.3882 | 1.2414 | 1.1427 | 1.2847 | **0.9195** | 19.53% |
| | | RMSE | 1.9726 | 2.1433 | 1.9981 | 1.7442 | 1.6611 | 1.4434 | 1.7076 | **1.2612** | 12.62% |
| | 20% | MAE | 1.8073 | 2.7567 | 2.5630 | 2.2415 | 1.7956 | 1.1516 | 1.7618 | **0.9418** | 18.22% |
| | | RMSE | 2.4277 | 3.0372 | 2.8753 | 2.6691 | 2.4006 | 1.4292 | 2.3365 | **1.2687** | 11.24% |

target domain within the validation and testing sets serve as ground truth for evaluating model performance.

We implement our framework using PyTorch. We extract the title, brand, and category as the textual description of the item. In Equation 1, we map both the text and image representations to a 256-dimensional vector. For the denoiser, we employ a 3-layer perceptron with 512 hidden units. The dropout probability for the conditional input is set to 10%. In Equation 14, $\lambda$ is set to 0.2 to balance the task loss and the contrastive loss. The temperature hyperparameter of contrastive loss is set to 0.02. Additionally, the inference process of DM typically involves numerous reverse steps, making direct sampling computationally expensive and time-consuming. To improve inference efficiency, we draw inspiration from existing studies [40], [41] and adopt the DPM solver, a well-established fast sampling solver, to accelerate the sampling process. The number of function evaluations used in the DPM solver is 50. The layers of perceptrons $L$ in the noise estimation network is selected from $\{2, 3, 4, 5\}$, and the diffusion step $R$ is selected from$\{1000, 3000, 5000, 7000\}$. We perform all the experiments on NVIDIA GeForce RTX 3090 GPUs.

All baselines are conducted based on their GitHub source code. All models are trained for 30 epochs to achieve convergence. We employ the same fully connected layer to facilitate comparison for the cross-domain bridge functions of EMCDR, SSCDR, and PTUPCDR. To ensure a fair comparison with P2M2-CDR, we employ the same item images and textual descriptions as in our method, processing them using the same pre-trained model. For each task and method, the initial learning rate for the Adam [42] optimizer is tuned by grid searches within $\{0.0001, 0.0005, 0.001, 0.005, 0.01, 0.1\}$. In our experiments, we vary the proportion $\alpha$ of overlapping users included in the training dataset, setting it to 80%, 50%, and 20% of the total overlapping user base. The remaining overlapping users are divided equally between the validation and test sets in a 1:1 ratio.

### *B. Performance Comparisons (RQ1)*

We compare the performance of MMDCDR with several baseline methods under cold-start scenarios, and the experimental results are presented in Table II. From these results, it can be observed that MMDCDR consistently outperforms all baseline models across various tasks. Specifically, under the experimental setting of $\alpha$ value at 80%, MMDCDR outperforms the existing SOTA in MAE by 6.91%, 14.54%, and 13.60% in three CDR tasks.

Compared to ID-based methods (e.g., DiffCDR, PTUPCDR, LACDR, SSCDR, EMCDR, and CMF), MMDCDR demonstrates a significant performance advantage, particularly in scenarios with limited training data. According to the experimental results, DiffCDR and PTUPCDR exhibit noticeable advantages only when the training data volume is highest, i.e., when proportion $\alpha$ = 80%, and their advantage quickly diminishes as the training data decreases. This trend highlights the intrinsic limitation of ID-based paradigms: they fail to

TABLE III
MODEL PERFORMANCE OF ABLATION STUDY. THE BEST RESULTS ARE BOLD.

| Methods | Movie2Music | | | | | | Sport2Phone | | | | | | Electric2Phone | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 80% | | 50% | | 20% | | 80% | | 50% | | 20% | | 80% | | 50% | | 20% | |
| | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE |
| w/o diff | 1.0399 | 1.3269 | 1.0904 | 1.3827 | 1.1189 | 1.3909 | 1.0872 | 1.3582 | 1.1590 | 1.4501 | 1.3021 | 1.6819 | 1.0278 | 1.2992 | 1.1870 | 1.4990 | 1.2526 | 1.5946 |
| w/o mm | 1.0885 | 1.5236 | 1.2234 | 1.7748 | 1.5610 | 2.2541 | 1.3887 | 1.8933 | 1.5742 | 2.0928 | 1.7295 | 2.3425 | 1.2179 | 1.6409 | 1.5311 | 1.9209 | 1.8628 | 2.5099 |
| w/o image | 1.0651 | 1.3528 | 1.1042 | 1.4128 | 1.1297 | 1.4712 | 1.2436 | 1.5718 | 1.3457 | 1.6746 | 1.4475 | 1.8486 | 1.1826 | 1.5386 | 1.3750 | 1.7211 | 1.3963 | 1.7070 |
| w/o text | 1.0559 | 1.3545 | 1.1891 | 1.5099 | 1.1567 | 1.4866 | 1.2846 | 1.6141 | 1.2904 | 1.6469 | 1.5161 | 1.8838 | 1.0823 | 1.2816 | 1.1092 | 1.3820 | 1.1709 | 1.5161 |
| w/o cl | 0.9667 | 1.2769 | 1.0010 | 1.2827 | 1.1119 | 1.3679 | 0.9643 | 1.2526 | 1.0070 | 1.3341 | 1.1645 | 1.4363 | 1.0611 | 1.3732 | 1.0847 | 1.4247 | 1.1269 | 1.4524 |
| MMDCDR | **0.9531** | **1.2563** | **0.9752** | **1.2671** | **1.0041** | **1.2891** | **0.9020** | **1.1749** | **0.9531** | **1.3175** | **1.1043** | **1.4131** | **0.9020** | **1.2267** | **0.9195** | **1.2612** | **0.9418** | **1.2687** |

generalize well to unseen or infrequent entities due to the lack of sufficient interaction signals. In contrast, multi-modal information provides fine-grained semantic representations of items, which are independent of interaction sparsity. These rich content features enable the model to better infer user preferences for novel or rarely interacted items, thereby alleviating the cold-start challenge. MMDCDR benefits from this modality-rich representation by incorporating both textual and visual features into the diffusion framework, enabling robust learning even in severely underrepresented regions of the data.

Moreover, while both MMDCDR and DiffCDR adopt a diffusion model framework, MMDCDR significantly outperforms DiffCDR across all evaluated scenarios. This performance gap underscores the value of integrating multi-modal information within the diffusion process. Rather than relying solely on user and item ID embeddings, which are often brittle under cold-start conditions, MMDCDR leverages a broader and more informative feature space, leading to improved transferability and generalization.

In addition, our approach consistently surpasses another multi-modal-based model, P2M2-CDR, which highlights the importance of the underlying architecture in exploiting multi-modal signals. The superior performance of MMDCDR suggests that the diffusion-based modeling of latent user-item interaction patterns enables more expressive and structured knowledge transfer across domains. These results collectively underscore the effectiveness of combining multi-modal representations with a diffusion modeling framework, positioning MMDCDR as a robust and scalable solution for cold-start cross-domain recommendation tasks.

### C. Ablation Study (RQ2)

In this section, we conduct an ablation study to evaluate the effectiveness of various components in MMDCDR. Specifically, we design the following five model variants.

1) w/o diff: it replaces the diffusion model in the original framework with an MLP.
2) w/o mm: it replaces multi-modal information with ID collaborative information.
3) w/o text: it removes item text features when learning item representations.
4) w/o image: it removes item image features when learning item representations.
5) w/o cl: it disables the modality-aware contrastive augmentation.

The results of the ablation experiments are shown in Table III. The results indicate that each component of the MMDCDR contributes significantly to its overall performance. Specifically, disabling the DM and replacing it with a simple MLP leads to a notable decline in performance. This finding underscores the importance of the diffusion mechanism in capturing the complex, dynamic transfer of knowledge across domains. Furthermore, when the multi-modal features are disabled, the model's performance experiences a substantial decrease, reinforcing the idea that multi-modal information plays a crucial role in extracting fine-grained item characteristics and capturing user behavioral preferences. When the model is restricted to using only text or image information, performance reductions are observed, but they remain moderate. This indicates that both text and image modalities offer complementary information to the recommendation process. Additionally, removing the modality-aware contrastive augmentation module results in further performance drops. This finding highlights the critical role of contrastive learning in improving the alignment and fusion of multi-modal data. By emphasizing consistent features across modalities, this module enhances the model's ability to effectively integrate textual and visual information, leading to more accurate and coherent item representations. Overall, the ablation studies provide strong evidence that each component of MMDCDR contributes synergistically to its superior performance in cross-domain recommendation tasks.

### D. Parameter Sensitivity Analysis (RQ3)

In this section, we investigate the impact of two key hyperparameters on model performance: (1) the number of perceptron layers $L$ in the noise estimation network, and (2) the number of diffusion steps $R$ in the forward process. The analysis is conducted under three different overlapping user ratios ($\alpha = 80\%, 50\%, 20\%$) to evaluate model robustness in varying cold-start scenarios. Specifically, Figures 3 and 5

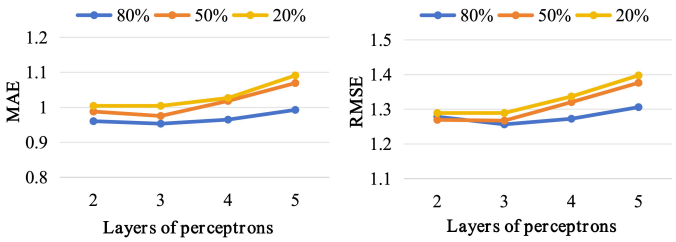


Fig. 3. Performance with different numbers of perceptron layers on Movie2Music.

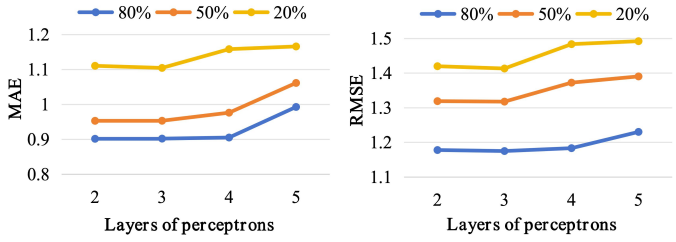


Fig. 4. Performance with different numbers of perceptron layers on Sport2Phone.

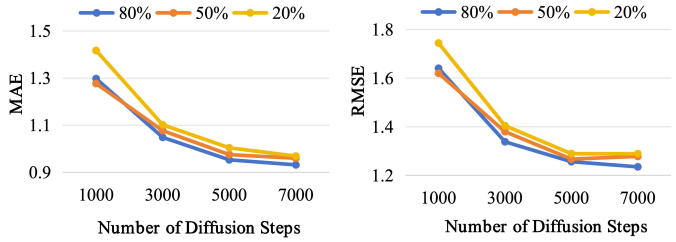


Fig. 5. Performance with different number of diffusion steps on Movie2Music.

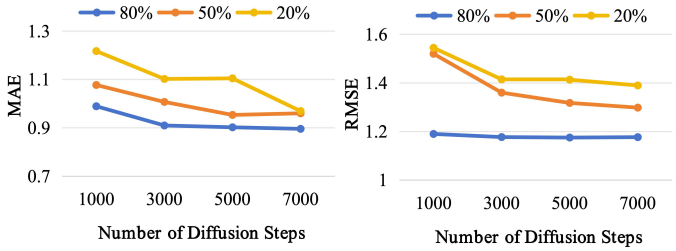


Fig. 6. Performance with different number of diffusion steps on Sport2Phone.

report results on the Movie2Music task, while Figures 4 and 6 present results on the Sport2Phone task.

*1) Impact of the Number of Perceptron Layers:* Figures 3 and 4 show the model's MAE and RMSE performance with varying numbers of perceptron layers $L$ in the denoising network (2, 3, 4, 5). The results across both datasets indicate that increasing the number of layers generally leads to a degradation in performance, especially under lower $\alpha$ values. This is likely due to overfitting caused by increased model complexity in data-sparse scenarios.

On Movie2Music, the optimal performance is achieved with 3 layers, beyond which both MAE and RMSE begin to rise. A similar trend is observed on Sport2Phone, where deeper networks (4 or 5 layers) show degraded performance. These observations suggest that a relatively shallow denoising model (e.g., 3 layers) is sufficient and more stable across different data scales.

*2) Impact of the Number of Diffusion Steps:* Figures 5 and 6 depict the model's performance with different numbers of diffusion steps $R$ (1000, 3000, 5000, 7000) during the forward process. In both datasets, we observe a clear trend: increasing the number of diffusion steps consistently improves model accuracy.

On Movie2Music, both MAE and RMSE decrease substantially as the number of steps increases from 1000 to 5000, with marginal improvements beyond 5000. This indicates that increasing the number of steps helps the model better capture the underlying data distribution up to a point, after which additional steps bring diminishing returns. On Sport2Phone, the performance shows a steep improvement when the number of diffusion steps increases from 1000 to 3000, especially under lower $\alpha$ values. However, beyond 3000 steps, the performance gains become more gradual, and the curves begin to plateau. This suggests that on this dataset, 3000 steps are sufficient to achieve most of the benefit of the diffusion process. These results demonstrate that while increasing the number of diffusion steps generally enhances the model's generative and predictive capabilities, the optimal number of steps may vary across datasets.

## V. Conclusion

In this paper, we focus on addressing the cold-start problem through cross-domain recommendation. Existing methods predominantly rely on ID-based collaborative information for CDR, which limits their effectiveness when faced with new or rarely-interacted items. To overcome these limitations, we propose a multi-modal diffusion model for CDR. By extracting the multi-modal representations of items, we can disentangle item fine-grained features, thereby gaining a deeper understanding of users' behavioral preference. The use of multi-modal features not only enriches item embeddings but also enhances the generalization ability of the model in cold-start settings. Furthermore, to enable effective knowledge transfer between domains, we employ a conditional diffusion model as the backbone of our architecture. The diffusion framework is capable of modeling complex and high-dimensional data distributions, which allows our model to capture intricate relations across source and target domains. In addition, we introduce a modality-aware contrastive augmentation strategy to strengthen the consistency and alignment between multi-modal item representations. Extensive experiments conducted on multiple benchmark datasets confirm the superiority of MMDCDR over existing state-of-the-art CDR methods.

## REFERENCES


[1] Y. Wang, X. Wang, X. Huang, Y. Yu, H. Li, M. Zhang, Z. Guo, and W. Wu, "Intent-aware recommendation via disentangled graph contrastive learning," in *IJCAI*, 2023, pp. 2343–2351.

[2] Y. Lu, C. Wang, P. Lai, and J. Lai, "PKAT: pre-training in collaborative knowledge graph attention network for recommendation," in *ICDM*, 2023, pp. 448–457.

[3] L. Wang, B. Jin, Z. Huang, H. Zhao, D. Lian, Q. Liu, and E. Chen, "Preference-adaptive meta-learning for cold-start recommendation," in *IJCAI*, 2021, pp. 1607–1614.

[4] Z. Li, J. Wang, Z. Chen, K. Wu, Y. Wei, and H. Huang, "Adaptive graph neural networks for cold-start multimedia recommendation," in *ICDM*, 2024, pp. 201–210.

[5] A. P. Singh and G. J. Gordon, "Relational learning via collective matrix factorization," in *SIGKDD*, 2008, pp. 650–658.

[6] C. Sun, J. Gu, B. Hu, X. Dong, H. Li, L. Cheng, and L. Mo, "REMIT: reinforced multi-interest transfer for cross-domain recommendation," in *AAAI*, 2023, pp. 9900–9908.

[7] C. Zhao, C. Li, R. Xiao, H. Deng, and A. Sun, "CATN: cross-domain recommendation for cold-start users via aspect transfer network," in *SIGIR*, 2020, pp. 229–238.

[8] Z. Lin, W. Huang, H. Zhang, J. Xu, W. Liu, X. Liao, F. Wang, S. Wang, and Y. Tan, "Enhancing dual-target cross-domain recommendation with federated privacy-preserving learning," in *IJCAI*, 2024, pp. 2153–2161.

[9] W. Liu, C. Chen, X. Liao, M. Hu, J. Yin, Y. Tan, and L. Zheng, "Federated probabilistic preference distribution modelling with compactness co-clustering for privacy-preserving multi-domain recommendation," in *IJCAI*, 2023, pp. 2206–2214.

[10] T. Man, H. Shen, X. Jin, and X. Cheng, "Cross-domain recommendation: An embedding and mapping approach," in *IJCAI*, 2017, pp. 2464–2470.

[11] S. Kang, J. Hwang, D. Lee, and H. Yu, "Semi-supervised learning for cross-domain recommendation to cold-start users," in *CIKM*, 2019, pp. 1563–1572.

[12] T. Wang, F. Zhuang, Z. Zhang, D. Wang, J. Zhou, and Q. He, "Low-dimensional alignment for cross-domain recommendation," in *CIKM*, 2021, pp. 3508–3512.

[13] Y. Zhu, K. Ge, F. Zhuang, R. Xie, D. Xi, X. Zhang, L. Lin, and Q. He, "Transfer-meta framework for cross-domain recommendation to cold-start users," in *SIGIR*, 2021, pp. 1813–1817.

[14] Y. Zhu, R. Xie, F. Zhuang, K. Ge, Y. Sun, X. Zhang, L. Lin, and J. Cao, "Learning to warm up cold item embeddings for cold-start recommendation with meta scaling and shifting networks," in *SIGIR*, 2021, pp. 1167–1176.

[15] Y. Zhu, Z. Tang, Y. Liu, F. Zhuang, R. Xie, X. Zhang, L. Lin, and Q. He, "Personalized transfer of user preferences for cross-domain recommendation," in *WSDM*, 2022, pp. 1507–1515.

[16] Z. Yuan, F. Yuan, Y. Song, Y. Li, J. Fu, F. Yang, Y. Pan, and Y. Ni, "Where to go next for recommender systems? ID- vs. modality-based recommender models revisited," in *SIGIR*, 2023, pp. 2639–2649.

[17] X. Zhang, B. Xu, F. Ma, C. Li, L. Yang, and H. Lin, "Beyond co-occurrence: Multi-modal session-based recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 4, pp. 1450–1462, 2024.

[18] Y. Xuan, "Diffusion cross-domain recommendation," *CoRR*, vol. abs/2402.02182, 2024.

[19] P. Dhariwal and A. Q. Nichol, "Diffusion models beat gans on image synthesis," in *NeurIPS*, 2021, pp. 8780–8794.

[20] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *CoRR*, vol. abs/2207.12598, 2022.

[21] X. Ye, Y. Li, and L. Yao, "DREAM: decoupled representation via extraction attention module and supervised contrastive learning for cross-domain sequential recommender," in *RecSys*, 2023, pp. 479–490.

[22] X. Li, J. Sheng, J. Cao, W. Zhang, Q. Li, and T. Liu, "CDRNP: cross-domain recommendation to cold-start users via neural process," in *WSDM*, 2024, pp. 378–386.

[23] L. Ma, Y. Li, Z. Mai, F. Liang, C. Wang, M. Chen, and M. Guizani, "Cross-store next-basket recommendation," in *ICDM*, 2024, pp. 301–310.

[24] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10 674–10 685.

[25] Z. Li, A. Sun, and C. Li, "Diffurec: A diffusion model for sequential recommendation," *ACM Transactions on Information Systems*, vol. 42, no. 3, pp. 1–28, 2023.

[26] W. Wang, Y. Xu, F. Feng, X. Lin, X. He, and T. Chua, "Diffusion recommender model," in *SIGIR*, 2023, pp. 832–841.

[27] Y. Jiang, Y. Yang, L. Xia, and C. Huang, "DiffKG: Knowledge graph diffusion model for recommendation," in *WSDM*, 2024, pp. 313–321.

[28] H. Ma, Y. Yang, L. Meng, R. Xie, and X. Meng, "Multimodal conditioned diffusion model for recommendation," in *WWW*, 2024, pp. 1733–1740.

[29] Q. Liu, F. Yan, X. Zhao, Z. Du, H. Guo, R. Tang, and F. Tian, "Diffusion augmentation for sequential recommendation," in *CIKM*, 2023, pp. 1576–1586.

[30] J. Chen, H. Zhang, X. He, L. Nie, W. Liu, and T. Chua, "Attentive collaborative filtering: Multimedia recommendation with item- and component-level attention," in *SIGIR*, 2017, pp. 335–344.

[31] X. Chen, H. Chen, H. Xu, Y. Zhang, Y. Cao, Z. Qin, and H. Zha, "Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation," in *SIGIR*, 2019, pp. 765–774.

[32] Y. Wei, X. Wang, L. Nie, X. He, R. Hong, and T. Chua, "MMGCN: multi-modal graph convolution network for personalized recommendation of micro-video," in *MM*, 2019, pp. 1437–1445.

[33] H. Hu, W. Guo, Y. Liu, and M. Kan, "Adaptive multi-modalities fusion in sequential recommendation systems," in *CIKM*, 2023, pp. 843–853.

[34] H. Zhou, X. Zhou, L. Zhang, and Z. Shen, "Enhancing dyadic relations with homogeneous graphs for multimodal recommendation," in *ECAI*, vol. 372, 2023, pp. 3123–3130.

[35] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019.

[36] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[37] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *MICCAI*, vol. 9351, 2015, pp. 234–241.

[38] J. Ni, J. Li, and J. McAuley, "Justifying recommendations using distantly-labeled reviews and fine-grained aspects," in *(EMNLP-IJCNLP)*, 2019, pp. 188–197.

[39] L. Wang, L. Sang, Q. Zhang, Q. Wu, and M. Xu, "A privacy-preserving framework with multi-modal data for cross-domain recommendation," *Knowl. Based Syst.*, vol. 304, p. 112529, 2024.

[40] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps," in *NeurIPS*, 2022.

[41] F. Bao, C. Li, J. Zhu, and B. Zhang, "Analytic-DPM: an analytic estimate of the optimal reverse variance in diffusion probabilistic models," in *ICLR*, 2022.

[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.