Multimodality Invariant Learning for Multimedia-Based New Item Recommendation

Haoyue Bai
Hefei University of Technology
Hefei, China
baihaoyue621@gmail.com

Le Wu*
Hefei University of Technology
Institute of Dataspace,
Hefei Comprehensive National Science Center
Hefei, China
lewu.ustc@gmail.com

Min Hou
Hefei University of Technology
Hefei, China
hmhoumin@gmail.com

Miaomiao Cai
Hefei University of Technology
Hefei, China
cmm.hfut@gmail.com

Zhuangzhuang He
Hefei University of Technology
Hefei, China
hyicheng223@gmail.com

Yuyang Zhou
Academy of Cyber
Beijing, China
yzhou193@alumni.jh.edu

Richang Hong
Hefei University of Technology
Institute of Dataspace,
Hefei Comprehensive National Science Center
Hefei, China
hongrc.hfut@gmail.com

Meng Wang
Hefei University of Technology
Institute of Artificial Intelligence,
Hefei Comprehensive National Science Center
Hefei, China
eric.mengwang@gmail.com

ABSTRACT

Multimedia-based recommendation provides personalized item suggestions by learning the content preferences of users. With the proliferation of digital devices and APPs, a huge number of new items are created rapidly over time. How to quickly provide recommendations for new items at the inference time is challenging. What's worse, real-world items exhibit varying degrees of modality missing (e.g., many short videos are uploaded without text descriptions). Though many efforts have been devoted to multimedia-based recommendations, they either could not deal with new multimedia items or assumed the modality completeness in the modeling process.

In this paper, we highlight the necessity of tackling the modality missing issue for new item recommendation. We argue that users' inherent content preference is stable and better kept invariant to arbitrary modality missing environments. Therefore, we approach this problem from a novel perspective of invariant learning. However, how to construct environments from finite user behavior training data to generalize any modality missing is challenging. To tackle this issue, we propose a novel **M**ultimodality **I**nvariant **L**earning re**C**ommendation (a.k.a. *MILK*) framework. Specifically, *MILK* first designs a cross-modality alignment module to keep semantic consistency from pretrained multimedia item features. After that, *MILK* designs multi-modal heterogeneous environments with cyclic mixup to augment training data, in order to mimic any modality missing for invariant user preference learning. Extensive experiments on three real datasets verify the superiority of our proposed framework. The code is available at https://github.com/HaoyueBai98/MILK.

CCS CONCEPTS

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

KEYWORDS

Multimedia-Based Recommendation, Invariant Learning, Modality Missing

ACM Reference Format:
Haoyue Bai, Le Wu, Min Hou, Miaomiao Cai, Zhuangzhuang He, Yuyang Zhou, Richang Hong, and Meng Wang. 2024. Multimodality Invariant Learning for Multimedia-Based New Item Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3626772.3658596

1 INTRODUCTION

With the proliferation of digital devices and APPs, individuals are exposed to abundant multimedia content, such as e-commerce and short-video sharing applications. Multimedia-based recommender systems have become indispensable components of these online services, aiming at learning user preferences from multimedia content

*Corresponding author.

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07…$15.00
https://doi.org/10.1145/3626772.3658596

to facilitate personalized item suggestions [2, 4, 11]. The key idea of these models lies in better model item content representations from pretrained features, and then align users' content preference with the help of users' historical records.

In real-world scenarios, multimedia-based recommender systems encounter distinctive characteristics. First, *a huge number of new items emerge rapidly over time*, particularly on the news and short-video sharing platforms. E.g., about 500 hours of video are uploaded to YouTube every minute[1] [18]. Unlike the old items encountered during training, new items lack users' behavior data and need to be quickly recommended to keep freshness. Most of previous works focused on designing sophisticated content representations from the complete multimedia content [10, 17, 24, 26]. Others proposed to leverage interchange between users and items with a fused graph from multiple content channels [33, 34, 40, 41]. E.g., researchers proposed to exploit the user-item interaction records to guide the representation learning of each modality for final recommendation fusion [34]. Most of these graph-based models show better performance for old items compared to pure content representation techniques. When faced with new items, most models need to be retrained on new items, which could not satisfy the quick adaption to new items at the inference stage. Second, *real-world items exhibit varying degrees of modality missingness*. For example, in short video-sharing platforms, authors may omit introductions for freshly uploaded videos, leading to a dearth of textual modality. Some new videos may intentionally lack audio features due to stylistic choices [7]. Nearly all previous works relied on modality completeness for recommendation or simply tackled this issue with preprocessing techniques to impute missing modalities. Therefore, the recommendation performance is hindered by the inferior preprocessing quality. In summary, how to tackle the new item and the modality missing issues in multimedia recommendation is important and has not been well studied before.

In this paper, we study the problem of multimedia-based new item recommendation. We provide an intuitive experiment to demonstrate the challenges of this problem. As shown in Figure 1, we assess the performance of a classical multimedia-based recommendation model DUIF [10] under different settings in Amazon Baby dataset [12, 20]. All samples in the original dataset have two modalities. The experimental settings are as follows: (1) Original dataset with no modality missing. (2) For the training and testing sets, samples randomly lose one modality. (3) Only samples in the testing set randomly lose one modality. Features for missing modalities are filled with the mean value. We can clearly observe a decline in model performance in *Setting 2* and *3*, indicating that modality missingness has a substantial detrimental effect on the new item recommendation. It should be noticed that the performance in *Setting 3* is much worse than *Setting 2*, showing that when there is a discrepancy in modality missingness between the training and testing sets, the model's performance experiences a more significant decline. It also demonstrates that simple data preprocessing and removing training samples with modality missing are ineffective, and may even exacerbate the situation.

Essentially, the difficulty is that there are multiple combinations with modality missing. When faced with new items of incomplete

[1]https://www.statista.com/

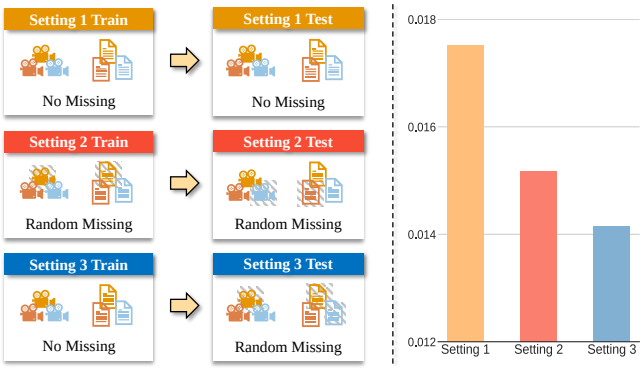


Figure 1: New Item Recommendation with Missing Modalities

patterns, the inference stage distribution changes compared to the training stage. In fact, users' inherent content preference is stable and better kept invariant with arbitraty modality missing. Therefore, an idea recommendation model is encouraged to predict each user's preference as invariant as possible.

To achieve this goal, we draw inspiration from invariant learning [1, 19, 39], which can achieve guaranteed performance under distribution shifts and received great attention in recent years. Invariant learning models correlations invariant across different training environments, where environments are variables that should not affect the prediction. Analogously, we want to learn users' inherent content preference that is invariant to any modality missing. However, implementing this analogy is challenging. As users' interaction records are limited, how to construct environments to generalize any modality missing is a challenge.

In this work, we propose a novel **M**ultimodality **I**nvariant **L**earning re**C**ommendation (a.k.a. *MILK*) framework for multimedia-based new item recommendation. The main idea of *MILK* is to encourage the users' inherent content preference stable and kept invariant to arbitrary modality missing scenarios for any new item recommendation. The *MILK* consists of two modules, the cross-modality alignment module for better item representation learning and the cross-environment invariance module for invariant preference prediction. Specifically, in the cross-modality alignment module, we devise alignment functions that allow one modality to capture content signals from other modalities. This module ensures that the absence of a specific modality does not hinder the extraction of features from other modalities. In the cross-environment invariance module, we design cyclic mixup to create multi-modal heterogeneous environments and employ invariant learning to enhance the model's generalization capability. In cyclic mixup, we use the Dirichlet distribution to create diverse environments, allowing for imbalanced modality proportions and comprehensive consideration of each modality.

Our contributions are summarized as:

- We emphasize the significance of addressing the modality missing issue in multimedia-based new item recommendations. We argue that users' underlying preference is invariant to arbitrary modality missing environments, and tackle this problem from an invariant learning perspective.

- We propose *MILK* for the challenging problem. We design a novel cyclic mixup method, which constructs heterogeneous environments with different modal information proportions, thereby adapting invariant learning with continuous values and augmenting limited training data for invariant user preference modeling.
- Extensive experiments on three real-world datasets demonstrate the superiority and effectiveness of *MILK* in multimedia-based new item recommendation.

2 RELATED WORK

2.1 Multimedia-Based Recommendation

The recommendation task aims to provide personalized recommendations to users[35–37]. Multimedia-based recommendations utilize multimodal contents (e.g., text, image, audio) of items to assist with the recommendation task. The information-rich multimodal content greatly improves item characteristics and user preference modeling. Early works incorporate the item visual features as side information into the models [4, 11, 13]. For example, VBPR [11] adds a visual representation of items based on matrix factorization. Subsequently, some researchers utilize graph neural networks to perform embedding propagation on interaction graphs with different modality data, thereby capturing user preferences on different modalities [16, 18, 33, 34, 42]. For instance, MMGCN [34] conducts graph convolutional operations on a modal-specific graph and captures the modal-specific user preference. Although these works play important roles in exploring the use of multimodal information, most of them are transductive and rely heavily on CF information. Thus, they cannot flexibly deal with the constantly emerging new items.

To address this challenge, some works manage to enable models to make new item recommendations. These methods do not rely on CF information and have the ability to make recommendations for new items directly using multimedia features. Hybrid recommendation methods combine CF signals and multimedia information in the training stage to obtain hybrid preference representations [2, 3, 29, 43, 44]. For example, GoRec [2] directly models the distribution of pre-trained preference representations to generate representations for new items guided by multimedia features. Content-based recommendation methods focus solely on item modeling based on multimedia features [10, 17, 24, 40, 41]. For example, DUIF [10] generates item representations using multimedia features and directly models user preferences for these features, enabling recommendations for new items based solely on their multimedia information. MICRO [41] constructs and fuses multiple item-item relation graphs to explicitly mine the semantic information between items. However, these methods often assume that the data is consistently complete and of high quality, a condition rarely met in the real world. In this paper, we address new item recommendations in scenarios where modalities are missing and introduce a more practical model to tackle this challenge.

2.2 Invariant Learning for Recommendation

Invariant learning (IL) [1, 5, 21] improves the robustness of models to distribution shifts. IL is based on the assumption that the causal mechanism keeps invariant across various environments. By penalizing the variance of model prediction across environments, models are then encouraged to capture the causal mechanism instead of spurious correlations. The common process of IL is to first split the training data into groups (i.e., environments), here the groups need to reflect spurious correlations. Then, by ensuring consistent performances across different environments, the purpose of learning invariant representations is achieved. The environment assignments play an important role in IL. Early works [1] assume the environment labels are given in the dataset. Recently, IL has been introduced to the scenario where environment labels are unknown [6, 25]. These methods utilize prior knowledge of spurious correlations to split the training data. For example, Teney *et al.* [25] cluster training samples with their predefined spurious features. EIIL [6] splits training data into the majority/minority sets on which the spurious feature conditioned label distribution varies maximally. IL has been introduced into recommender systems for building more trustworthy models nowadays. InvPref [30] estimates heterogeneous environments corresponding to different types of latent bias and uses IL to handle different unknown data biases in a unified framework. InvRL [9] utilizes environments to reflect the spurious correlations and then learns invariant representations to make a consistent prediction of user-item interaction across various environments. In our work, we are inspired by IL, making the users' inherent content preference stable and kept invariant to arbitrary modality missing. We present a novel cyclic mixup heterogeneous environments construction method. Cyclic mixup mimics infinite real-world scenarios with finite training samples.

3 PROBLEM FORMULATION

3.1 New Item Recommendation

Let $\mathcal{U}$ and $\mathcal{V}$ denote the sets of users and items. Since implicit feedback is very common, we use $\mathbf{R} \in \mathbb{R}^{|\mathcal{U}|\times|\mathcal{V}|}$ to denote the user-item interaction matrix, $r_{ij} = 1$ if user i has interacted with item j, otherwise $r_{ij} = 0$. Beyond the interaction signal, the multimodal features of items are extracted from their content, such as the visual, textual, acoustic modalities, and so on. We can generate the multimodality features into representations via generic feature extractors. For an item $j \in \mathcal{V}$, we denote its feature vector as $\mathbf{x}_j \in \mathbb{R}^{M\times d_x}$, where M is the number of modalities and d_x is the dimension of the vector. Taking the user ID i and multimodality representation $\mathbf{x}_j$ of item j as input, the recommendation model $\mathcal{F}_\Phi$ aims to infer the probability of user i will interact with item j. Optimizing $\mathcal{F}_\Phi$ is to minimize the loss function $\mathcal{L}$ (e.g., the BPR loss [23]) given the observed interactions:

$$\Phi^* = \arg\min_{\Phi} \mathbb{E}_{(i,\mathbf{x}_j)\sim \mathbf{P}_{train}} \mathcal{L}(\mathcal{F}_\Phi(i, \mathbf{x}_j); \mathbf{R}). \tag{1}$$

The expectation is calculated in the training data distribution $\mathbf{P}_{train}$. In the inference stage, we aim for the new item recommendation model $\mathcal{F}_{\Phi^*}$ can perform well on the new items set $\mathcal{V}_{\text{new}}$ without any historical interactions.

3.2 Modality Missing Issue

The new item recommendation models effectively infer the probability $\hat{y}_{ij}$ of user i will interact with new item j_{new} under the

modality completeness assumption. Real-world scenarios often deviate from this idealized setting. Many samples exhibit varying degrees of modality missingness. We expect the recommendation model to remain effective in this scenario.

We denote the modality representation of items' as $\mathbf{X} \in \mathbb{R}^{|V| \times M \times d_x}$. We use $\mathbf{A} \in \mathbb{R}^{|V| \times M}$ to denote the item modality indication matrix, where $\mathbf{a}_{jm} = 1$ if the m^{th} modalities of item j is available, and otherwise $\mathbf{a}_{jm} = 0$. During the training stage, input data $(i, \mathbf{x}_j)$ can be considered to be drawn from a joint distribution :

$$(i, \mathbf{x}_j) \sim \mathbf{P}_{train}(i, \mathbb{I}(\mathbf{X}_j, \mathbf{A}_j)). \tag{2}$$

$\mathbb{I}(\mathbf{X}_j, \mathbf{A}_j)$ represents some modality in $\mathbf{X}_j$ is missing indicated by $\mathbf{A}_j$. In the inference stage, the unavailability of certain modalities in a new item leads to data being drawn from another joint distribution:

$$(i, \mathbf{x}_{j_{\text{new}}}) \sim \mathbf{P}_{test}(i, \mathbb{I}(\mathbf{X}_{\text{new}}, \mathbf{A}_{j_{\text{new}}})). \tag{3}$$

The multimodal data in the training set is typically complete, while the missingness of multimodal data in the testing phase is unknown, namely $\mathbf{A} \neq \mathbf{A}_{\text{new}}$. The distribution shift between $\mathbf{P}_{test}$ and $\mathbf{P}_{train}$ challenges the performance of new item recommendation models built on the empirical risk minimization (ERM) in the training set.

Our goal is to develop an optimal new item recommendation model capable of generalizing well to multimedia features drawn from the test distribution $\mathbf{P}_{test}$, where $\mathbf{P}_{test} \neq \mathbf{P}_{train}$. The optimization objective is to minimize the loss $\mathcal{L}(\mathcal{F}(i, \mathbf{x}_j); \mathbf{R})$ with respect to the model's parameters Φ, where the expectation is taken over the test data distribution $\mathbf{P}_{test}$. Note that $\mathbf{P}_{test}$ is unknown during the training stage.

$$\arg\min_{\Phi} \mathbb{E}_{(i,\mathbf{x}_j) \sim \mathbf{P}_{test}} \mathcal{L}(\mathcal{F}_{\Phi}(i, \mathbf{x}_j); \mathbf{R}). \tag{4}$$

4 THE PROPOSED *MILK* FRAMEWORK

4.1 Overview of *MILK*

As illustrated in Figure 2, we present the overall framework of our proposed *MILK* . Essentially, *MILK* aims to encourage stable user preference prediction for new items, which is guaranteed by invariant preference learning under missing modality scenarios. To achieve this goal, *MILK* consists of two elaborated modules: Cross-Modality Alignment Module (CMAM) and Cross-Environment Invariance Module (CEIM).

Specifically, CMAM aims to learn an informative multimodal representation under potential modality missing scenarios. We implement CMAM by narrowing each modality representation, thus each modality representation can supplemented from other modality features, to tackle the insufficient multimodal representation issue under missing modality. After multimodal representation alignment, we execute CEIM as follows: heterogeneous environment construction and invariant user preference learning in various environments. Particularly, we devise a flexible and unique environment construction based on a cyclic modality mixup strategy, enabling adaptation to varying proportions of missing modalities in test data. Given the constructed environments, we conduct stable user preference learning based invariant risk minimization (IRM) principle. Next, we introduce each module in detail.

4.2 Cross-Modality Alignment Module

Modal-Specific Extractors. Most existing works [32, 44] concatenate multi-modal features and then convert them into item representation through an only extractor. This means that the absence of any modality has an impact on the whole extractor, which in turn adversely affects the extraction of other modal information.

CMAM uses independent extractors to guarantee stable modality feature extraction, which means each extractor is exclusively related to a specific modality. Such an approach ensures the accuracy of the underlying understanding of existing modal features. The process is as follows:

$$\mathbf{c}_j^m = \mathcal{G}^m(\mathbf{x}_j^m) = \mathbf{W}^m \mathbf{x}_j^m + \mathbf{b}^m, \tag{5}$$

where $\mathbf{x}_j^m$ denote the m^{th} modalitys' original feature of item j, $\mathcal{G}^m$ denote the representation generator of m^{th} modality and $\mathbf{W}^m$ and $\mathbf{b}^m$ are used to parameterize this process.

Cross-Modality Alignment. While the modal-specific extractors ensure the stable extraction of modal information, they also increase the risk that different modal information is mapped into different representation spaces. We use alignment across modalities to guarantee the semantic consistency between representations of different modalities. At the same time, the alignment across modalities makes the information between the modalities be transferred to each other. When a certain modality is unavailable, other modalities can provide a supplement. In *MILK*, we achieve this goal by optimizing the following alignment objective:

$$\mathcal{L}_{align} = \sum_{j=1}^{|\mathcal{V}|} \sum_{m=1}^{M-1} \sum_{m'=m+1}^{M} \mathbf{a}_{jm} \mathbf{a}_{jm'} \|\mathbf{c}_j^m - \mathbf{c}_j^{m'}\|^2. \tag{6}$$

4.3 Cross-Environment Invariant Module

Embedding and Fusion. CEIM obtains user representation and item representation by embedding layer $\mathcal{P}$ and fusion function $\mathcal{Q}$. For users, CEIM directly uses an embedding layer to convert the ID of user i into a user representation $\mathbf{u}_i$. Such an embedding layer can directly capture the user's interest from their interaction records. For items, CMAM already generates M representations $\mathbf{c}_j^m$ for each item. Then CEIM fuse these representations by the function $\mathcal{Q}$ to obtain the representation of item j:

$$\mathbf{z}_j = \mathcal{Q}(\mathbf{c}_j^1, \ldots, \mathbf{c}_j^M), \tag{7}$$

In order to display and control the proportion of modal information more intuitively, we use weighted summation as the fusion function. We first determine the weights $\Theta = \{\theta^1, \ldots, \theta^M\}$ and fuse the multimedia representations as follows:

$$\mathbf{z}_j = \theta^1 \times \mathbf{c}_j^1 + \ldots + \theta^M \times \mathbf{c}_j^M. \tag{8}$$

Heterogeneous Environment Construction. To enable the model the ability to consistently perform well in complex modal missing scenarios, we assume the existence of multiple heterogeneous environments, each with multimedia features drawn from a different distribution. In our context, the environments should simulate the different modality missingness situations, i.e., let $(i, \mathbf{x}_j)_e \sim \mathbf{P}_{train_e}(i, \mathbb{I}(\mathbf{X}_j, \mathbf{A}_{j_e}))$ indicates the training data belongs to environment e, $\forall e \neq e', \mathbf{P}_{train_e} \neq \mathbf{P}_{train_{e'}}$. We encourage the model to maintain good performance and eventually learn users' inherent

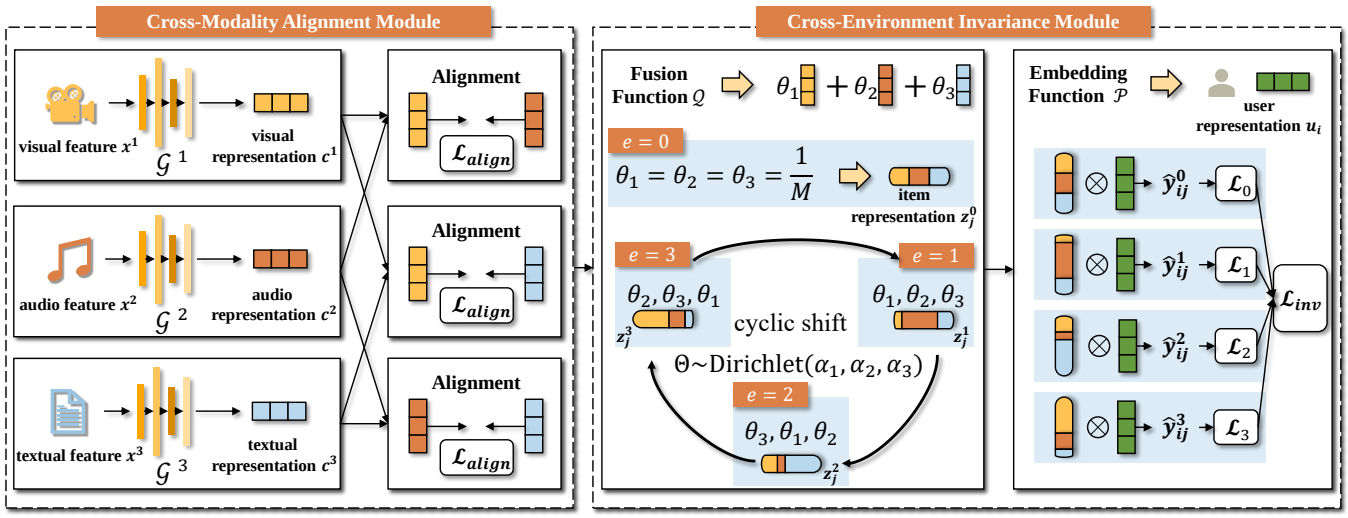


Figure 2: Model overview. *MILK* is consisted of Cross-Modality Alignment Module (CMAM) and Cross-Environment Invariant Module (CEIM). CMAM obtains the modality representations c^m through independent feature extractors $\mathcal{G}^m$ and then imposes alignment between any two modalities. CEIM converts user ID into user representation by embedding function $\mathcal{P}$ and generates item representations through fusion functions $\mathcal{Q}$. CEIM generates multiple sets of weights as heterogeneous environments through cyclic mixup and aggregates multi-modal representations into item representations z_j^e in each environment e. Finally, CEIM optimizes the model under the invariant learning paradigm.

content preferences across heterogeneous environments. We naturally simulate and control the heterogeneity of the environment by adjusting the weights Θ.

Fusing different modality representations using equal weights is the basic strategy. In the testing phase, this fusion strategy will be used since no clear judgment can be made about the quality of information of a certain modality. We take equal weights $\Theta_0 = \{\theta_0^1, \ldots, \theta_0^M\} = \{\frac{1}{M}\}^M$ into account as an environment during the training phase to ensure that this strategy is always exposed to the model, which ensures the basic performance of the model [22].

Then, we construct heterogeneous environments by adjusting the weight. We expect these environments to have some important properties: (1) *Unbalance Proportion*: On the one hand, all modalities should be included in each environment, otherwise, the model will be encouraged to ignore modality, resulting in an overall decrease in performance. On the other hand, the proportion of modalities should be unbalanced to simulate the complex real situation. (2) *Full Modality Consideration*: Different environments should be dominated by different modalities, and each modality should play a major role in some environments. This guarantees that the model does not establish too strong associations with specific modalities and that the information of each modality is fully learned. (3) *Diversity for Generalization*: Enough environments should be simulated to improve the model's ability. The model should be exposed to a large number of heterogeneous environments. If the number of environments is small, it may lead to spurious associations between the model and the limited pattern.

Inspired by mixup [27, 38], a data augmentation method that mixes original samples to generate new samples, we propose a novel cyclic mixup method to construct environments satisfying the above properties. Specifically, we first determine a Dirichlet distribution and then sample a set of weights $\Theta_1 = \{\theta_1^1, \ldots, \theta_1^M\}$ from this distribution:

$$\Theta_1 \sim Dirichlet(\alpha_1, \ldots, \alpha_M), \tag{9}$$

where $\alpha_1, \ldots, \alpha_M$ is used to adjust the Dirichlet distribution. Since the Dirichlet distribution represents a probabilistic simplex in M-dimensional space, there is no need for additional normalization of the weights and we can extend this method to more scenarios with any number of modalities.

We then generate the other weights by cyclic shift:

$$\Theta_m = circle\_shift(\Theta_{m-1}), \tag{10}$$

where Θ_m denotes the resulting weight after $m-1$ cyclic shift. Each time through the shift, we move all the weights in Θ_{m-1} back one bit and place the last weight at the front. The weight changes for performing one cyclic shift are as follows:

$$\{\theta^1, \theta^2, \ldots, \theta^m, \ldots, \theta^M\} \longrightarrow \{\theta^M, \theta^1, \ldots, \theta^{m-1}, \ldots, \theta^{M-1}\}. \tag{11}$$

We can construct M environments based on Θ_1 by M-1 cyclic shift. The weights sampled from the Dirichlet distribution guarantee the *Unbalance Proportion* that the weights are nonzero and unequal for different modes. Cyclic shift guarantees the property of *Full Modality Consideration*, where the dominant mode is different in each environment, and each mode will dominate an environment. We perform this process at each iteration, which ensures the *Diversity for Generalization* of the environment through randomness.

We denote environments by e, where $e = 0$ is the environment with equal weights, and $e = 1, \ldots, M$ denotes the M environments obtained by cyclic mixup. This fusion function uses each environment e's weight to generate the item representation $\mathbf{z}_j^e$:

$$\mathbf{z}_j^e = \mathcal{Q}(\mathbf{c}_j^1, \ldots, \mathbf{c}_j^M) = \theta_e^1 \times \mathbf{c}_j^1 + \ldots + \theta_e^M \times \mathbf{c}_j^M. \tag{12}$$

Invariant optimization. With the synergy of multiple modules, the model can predict the user's liking for the item based on the

user ID and item multimedia features in each environment:

$$\hat{y}_{ij}^{e} = \mathcal{F}_{\Phi}(\mathbf{u}_i, \mathbf{z}_j^{e}), \tag{13}$$

herein, $\mathcal{F}_{\Phi}$ estimates the probability that a user likes an item by the inner product operation.

We argue that users' inherent content preference is stable and better kept invariant to arbitrary modality information distribution. Hence, we encourage the model to maintain performance in heterogeneous environments. To do this, we train the model using the invariant learning paradigm to learn an invariant preference prediction mechanism. We use one of the common optimization objectives under the invariant learning paradigm as follows [15]:

$$\mathcal{L}_{inv} = \mathbb{E}_{e\in\mathcal{E}}\mathcal{L}_e + \beta \mathbf{Var}_{e\in\mathcal{E}}(\mathcal{L}_e), \tag{14}$$

where the first term is the task-dependent loss used to guarantee the performance of the model on the target task. The second term is the constraint over the loss variance across environments, which encourages the model to be stable across different environments. In the context of our work, the environment e is represented by different weights Θ_m, and $\mathcal{L}_e$ is the average recommendation loss value inside the environment e, and the widely used BPR loss is adopted in this paper:

$$\mathcal{L}_e = \sum_{(i,j,j')\in\mathcal{U}\cup\mathcal{V}} -\ln\sigma\left(\hat{y}_{ij}^{e} - \hat{y}_{ij'}^{e}\right), \tag{15}$$

where j' is the negative sample sampled item of user i.

4.4 Model Optimization and Inference

In summary, our final optimization objective is:

$$\mathcal{L} = \mathbb{E}_{e\in\mathcal{E}}\mathcal{L}_e + \beta \mathbf{Var}_{e\in\mathcal{E}}(\mathcal{L}_e) + \lambda\mathcal{L}_{align} + \|\Phi\|^2, \tag{16}$$

where λ is a hyperparameter that controls the weight of the alignment loss, and Φ includes all model parameters. We optimize the model parameters in the training stage by :

$$\Phi^{*} = \arg\min_{\Phi} \mathbb{E}_{(i,\mathbf{x}_j)\sim\mathbf{P}_{train}} \mathcal{L}(\mathcal{F}_{\Phi}(i, \mathbf{x}_j); \mathbf{R}). \tag{17}$$

In the inference stage, *MILK* can directly apply to new items with varying degrees of modality missingness. When a new item j_{new} appears, the multimedia feature $\mathbf{X}_{j_{\text{new}}}$ is first processed using mean imputation. Then we can predict the preference score of user i to j_{new} as:

$$\hat{y}_{ij_{\text{new}}} = \mathcal{F}_{\Phi^*}(\mathbf{u}_i, \frac{1}{M}\sum_{m=1}^{M}\mathcal{G}^m(\mathbf{x}_{j_{\text{new}}}^m)). \tag{18}$$

5 EXPERIMENTS

In this section, we conduct extensive experiments on three real-world datasets, which aim to answer the following questions:

- **Q1**: How does our model perform compared with state-of-the-art new item recommendation methods?
- **Q2**: How does our model improve new item recommendation performance with missing modalities?
- **Q3**: How do all modules in our model make positive effects on the performance?

Table 1: The statistics of datasets.

Dataset		Baby	Clothing	TikTok
Train	# Users	19442	39384	9319
	# Items	5640	18427	5368
	# Interactions	128963	222759	55126
	Density	0.118%	0.031%	0.110%
Val	# Users	10342	19801	2380
	# Items	705	2303	671
Test	# Users	10474	19858	2960
	# Items	705	2303	671

5.1 Experimental Settings

5.1.1 Datasets. **Datasets Description.** We conduct experiments on three widely used real-world datasets, including (a) Amazon Baby [12, 20], (b) Amazon Clothing, Shoes, and Jewelry [12, 20], and (c) TikTok[2]. To simplify reading, they are called Baby, Clothing, and TikTok. Baby and Clothing include both visual and textual modalities. TikTok is collected from the TikTok platform to log the viewed short videos of users. The multi-modal features are visual, acoustic, and title textual features of videos. For a fair comparison, all models use the pretrained multi-modal features as input [31, 40]. The statistics of the pre-processed datasets are listed in Table 1.
New Item Setting. We randomly select 20% items and delete their historical interactions in the training process to simulate new items. Among them, we further divide half as validation and the remaining as test items. These items are entirely unseen in the training set.
Modality Missing Setting. On three datasets, we validate the effect of our model under two settings. In the first Full Training Missing Test (FTMT) setting, we assume that the quality of the training data can be guaranteed. We use the complete multimedia features for training in the training phase, and randomly select 50% of the items in the testing phase, assuming that they randomly miss one modality feature. The missing modality is randomly selected and filled in using the mean imputation to pre-process (We use a simple padding way to keep the input data in a consistent format, and we provide detailed experiments on the imputation method in section 5.3.2). In the second more realistic setting Missing Training Missing Test (MTMT), we assume that there are also 30% items in the training set that randomly miss one modality feature.

5.1.2 Evaluation Metrics. We select two metrics that are widely used in personalized recommender systems: Recall (Recall@K) and Normalized Discounted Cumulative Gain (NDCG@K). The higher these two metrics are, the better the model is performing.

5.1.3 Baselines. To verify the effectiveness of *MILK*, we select multiple SOTA models suitable for task scenarios for comparison:

- **DUIF** [10] is a content-based method. It transforms heterogeneous user-content networks into homogeneous low-dimensional space for unified representation learning.
- **MICRO** [41] is a hybrid method. It learns item-item relationships for each modality, and it utilizes contrastive learning for better item-level multimodal fusion.

[2]http://ai-lab-challenge.bytedance.com/tce/vc/

- **DropoutNet** [28] is designed for new item recommendation. It randomly drops the partial preference representation in the training stage to simulate the new item scenario.
- **MTPR** [8] address the new item multimedia recommendation issue by strategically replacing the preference representation with the all-zero vector in the training phase.
- **Heater** [44] is a new item recommendation method that uses the sum squared error loss to align CF signal and content representation.
- **CLCRec** [32] uses contrastive learning to constrain CF signal and content representation to maximize the mutual information between them.
- **CCFCRec** [43] capitalizes on the co-occurrence collaborative signals in warm training data to alleviate the issue of blurry collaborative embeddings.
- **GAR** [3] trains a generator and a recommender adversarially, generating the new item's representation which is similar to the old item's representation.
- **GoRec** [2] directly models the distribution of pre-trained preference representations to generate representations for new items guided by multimedia features.

We comprehensively selected multiple content-based recommendation methods and multimedia new-item recommendation methods as baselines. Note that some graph-based multimedia recommendation methods (e.g., MMGCN) are not compared because they require graph construction based on user-item interaction records, which makes it difficult be apply in our challenging scenario.

5.1.4 Hyper-Parameter Settings. We implement our *MILK* and all baselines with Pytorch framework[3]. The dimension of preference representation is fixed as 64. The batch size is set to 2048. During training, we employ Adam [14] as the optimizer and set the learning rate at 0.001, the early stop strategy is employed to avoid overfitting. We carefully search the best parameter of β and λ and find *MILK* achieves the best performance when $\beta = 1000$ and $\lambda = 0.05$ on Baby, $\beta = 50$ and $\lambda = 0.05$ on Clothing, and $\beta = 50$ and $\lambda = 0.5$ on TikTok dataset. For the choice of the Dirichlet distribution when constructing the environment, we set $\alpha_1 = \ldots = \alpha_M$ and search α_m from $[0.01, 0.1, 1, 10, 100]$. For all baselines, we search the parameters carefully for fair comparisons. We repeat all experiments 5 times and report the average results.

5.2 Overall Comparisons (Q1)

As shown in Table 2, we compare our model with other baselines on three datasets. We have the following observations:

(1) On all three datasets, *MILK* shows a significant improvement over all baselines in the FTMT setting. Specifically, *MILK* improves the strongest baseline *w.r.t* NDCG@20 by 11.2%, 14.0%, and 2.8% on Baby, Clothing, and TikTok dataset, respectively. Extensive empirical studies verify the effectiveness of the proposed *MILK* . We attribute this improvement to the two modules we designed, which greatly enhanced the generalization ability of the model.

(2) In the MTMT setting, we consider scenarios that are more realistic. The lack of partial item modality in the training phase puts forward higher requirements for the fitting and generalization ability of the model. *MILK* still performs best on all three datasets. *MILK* improves the strongest baseline *w.r.t* NDCG@20 by 13.5%, 9.6%, and 4.3% on Baby, Clothing, and TikTok dataset, respectively. The experimental results further validate the practicability of *MILK*.

(3) DUIF, MICRO, and *MILK* only model the characteristics of new items through multi-modal features, without introducing additional CF representation. Our model performs well on all settings and datasets, which fully demonstrates that our design is working. For instance, on Baby dataset *MILK* improves DUIF *w.r.t* NDCG@20 by 69.9%, 87.2%, in FTMT setting and MTMT setting.

5.3 Improvements on Modality Missingness (Q2)

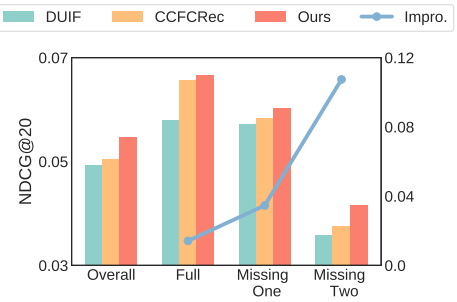


Figure 3: Performance on different missing scenarios.

5.3.1 Performance on Different Missing Scenarios. In this section, we further explore how *MILK* improves performance when the modalities are missing. The Tiktok dataset has richer modalities, based on which we construct different modality missing scenarios. We use the full training and validation sets. In the test set, 10% items are missing no modality, 45% are randomly missing one modality and 45% are randomly missing two modalities. We report the performance of the model on all samples and on a specific group of items separately. We select the top 2 baselines on the Tiktok dataset in Section 5.2 for comparison. The bar charts in Figure 3 show the results of the three models under different groups. *Overall* represents the results on all test sets, *full* represents the results on the item group with complete modality, *missing one* and *missing two* represent the item group with one and to modality missing, respectively. For ease of presentation, we show the numbers after two decimal places of NDCG@20. The line chart demonstrates the percentage of our method improving the strongest baseline. We can observe that the models outperform baselines on the whole test set. At the same time, we find that the improvement of the *MILK* is more obvious in the item group with missing modalities, and the improvement of the group with missing two modalities is far more than the rest of the groups. This shows that our model really improves the recommendation performance of the model on modal missing items, and has stronger generalization ability.

5.3.2 Comparison with common data imputation methods. We further designed experiments to verify the effects of different data imputation operations in the inference stage. We conduct experiments on Baby and Clothing datasets in the FTMT setting and report the result in Table 3. *Zero* means that we delete the alignment part,

[3]https://pytorch.org/

Table 2: Performance comparisons with different Top-K values different settings.

Settings		Full Training Missing Test (FTMT)						Missing Training Missing Test (MTMT)					
Datasets		Baby		Clothing		Tiktok		Baby		Clothing		Tiktok	
Metric	Models	K=10	K=20	K=10	K=20	K=10	K=20	K=10	K=20	K=10	K=20	K=10	K=20
Recall@K	DUIF	0.0217	0.0381	0.0189	0.0318	0.1814	0.1921	0.0172	0.0308	0.0188	0.0323	0.1706	0.1804
	DropoutNet	0.0178	0.0208	0.0155	0.0269	0.1221	0.1276	0.0130	0.0227	0.0150	0.0254	0.1011	0.1169
	MTPR	0.0261	0.0328	0.0224	0.0371	0.1526	0.1615	0.0187	0.0325	0.0190	0.0317	0.1444	0.1661
	Heater	0.0303	0.0492	0.0338	0.0522	0.1260	0.1327	0.0259	0.0438	0.0308	0.0515	0.1148	0.1349
	CLCRec	0.0247	0.0418	0.0281	0.0453	0.1529	0.1584	0.0190	0.0328	0.0265	0.0403	0.1228	0.1459
	CCFCRec	0.0295	0.0464	0.0355	0.0545	0.1802	0.1920	0.0210	0.0355	0.0291	0.0453	0.1715	0.1809
	GAR	0.0307	0.0485	0.0352	0.0555	0.1486	0.1691	0.0271	0.0449	0.0319	0.0523	0.1292	0.1512
	GoRec	0.0335	0.0544	0.0429	0.0639	0.1465	0.1605	0.0282	0.0473	0.0430	0.0634	0.1416	0.1582
	Ours	**0.0381**	**0.0628**	**0.0484**	**0.0716**	**0.1886**	**0.2002**	**0.0328**	**0.0517**	**0.0470**	**0.0684**	**0.1793**	**0.1884**
NDCG@K	DUIF	0.0117	0.0163	0.0100	0.0137	0.1792	0.1803	0.0088	0.0125	0.0104	0.0140	0.1682	0.1706
	DropoutNet	0.0057	0.0071	0.0093	0.0123	0.0979	0.0956	0.0071	0.0098	0.0077	0.0100	0.0824	0.0860
	MTPR	0.0156	0.0208	0.0125	0.0165	0.1271	0.1211	0.0100	0.0139	0.0106	0.0140	0.1177	0.1229
	Heater	0.0178	0.0222	0.0186	0.0237	0.1082	0.1016	0.0136	0.0185	0.0166	0.0223	0.0937	0.0975
	CLCRec	0.0133	0.0181	0.0163	0.0210	0.1321	0.1249	0.0105	0.0143	0.0148	0.0185	0.1036	0.1080
	CCFCRec	0.0159	0.0209	0.0207	0.0259	0.1799	0.1802	0.0114	0.0154	0.0163	0.0207	0.1567	0.1596
	GAR	0.0163	0.0210	0.0200	0.0255	0.1133	0.1173	0.0147	0.0196	0.0176	0.0232	0.0956	0.1004
	GoRec	0.0179	0.0249	0.0234	0.0292	0.1096	0.1124	0.0153	0.0206	0.0234	0.0300	0.0952	0.0988
	Ours	**0.0215**	**0.0277**	**0.0279**	**0.0343**	**0.1830**	**0.1852**	**0.0182**	**0.0234**	**0.0270**	**0.0329**	**0.1774**	**0.1779**

Table 3: Comparison with common data imputation methods.

Datasets	Baby		Clothing	
Metrics	Recall@20	NDCG@20	Recall@20	NDCG@20
Zero	0.0501	0.0215	0.0670	0.0322
Mean	0.0525	0.0228	0.0665	0.0320
Map	0.0507	0.0220	0.0619	0.0294
CMAM	0.0585	0.0255	0.0682	0.0330
CEIM	0.0609	0.0264	0.0686	0.0335
Ours	0.0628	0.0277	0.0716	0.0343

heterogeneous environment construction and invariant learning part in *MILK*, and fill the missing modalities in the test set with zeros. *Mean* indicates padding with the mean value. *Map* indicates we pre-train the mapping function between the two modalities. In the testing phase, for the missing image modality, we use the text_to_image mapping function to calculate image representation from the text representation. The missing text representation is computed from the image representation using the image_to_text mapping function. *CMAM* and *CEIM* indicate that we used only one module from *MILK*. The experimental results show that the preprocessing method does not improve the model performance in our problem scenario. Sometimes a well-designed imputation method can further degrade the performance of the model, such as the *Mean* and *Map* imputation on the Clothing dataset. In contrast, each component of *MILK* can effectively improve the performance of the model in the absence of modalities.

5.3.3 Impact of Constructed Environment. *MILK* constructs differentiated environments by adjusting modal weights and learns preference prediction mechanisms that are invariant across environments. We further explore the impact of different environmenta construction strategies. *w/o e=0* does not additionally consider the

Table 4: Different strategies for constructing environments.

Dataset	Baby		Clothing	
Metric	Recall@20	NDCG@20	Recall@20	NDCG@20
ours	0.0628	0.0277	0.0716	0.0343
w/o e=0	0.0433	0.0179	0.0685	0.0333
w/o cs	0.0487	0.0205	0.0656	0.0319
w/o random	0.0555	0.0237	0.0643	0.0314

case with the same weights. *w/o cs* indicates that no cyclic shift is used to generate the weights, and weights are sampled from the Dirichlet distribution for each environment while keeping the number of environments. *w/o random* means that weights are sampled and cyclically shifted to construct the environment only at the beginning of model training, and weights are fixed afterward. Table 4 shows that our method outperforms the three variants of the constructive environment method. Our strategy for constructing the environment is reasonable, necessary and effective.

5.4 Detailed Model Analysis (Q3)

5.4.1 Ablation Study. To exploit the effectiveness of each component of the proposed *MILK* , we conduct the ablation study on different datasets. As shown in Figure 4, we compare *MILK* and corresponding variants on Top-20 recommendation performance. MILK-*w/o CMAM* denotes that the Cross-Modality alignment module is removed and the fusion is performed directly after extracting the modal representation. MILK-*w/o CEIM* refers to training under the ERM paradigm. In this setting, we directly use the average strategy to fuse the representations of multiple modalities, and then calculate the propensity score with the user's representation to optimize the objective by BPR. MILK-*w/o BOTH* means that we delete both modules at the same time. Under this setting, we did not perform any design for distribution inconsistency. In Figure 4, we observed that each component of the *MILK* contributed to the final

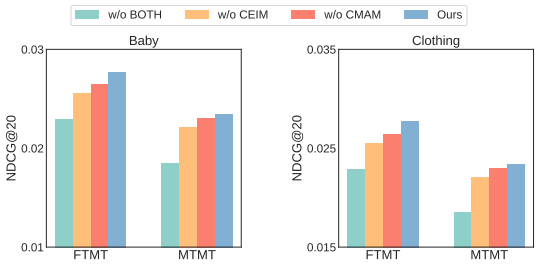
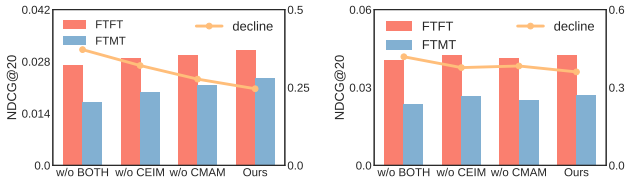


Figure 4: Ablation experiments on Baby and Clothing datasets.



(a) Baby (b) Clothing

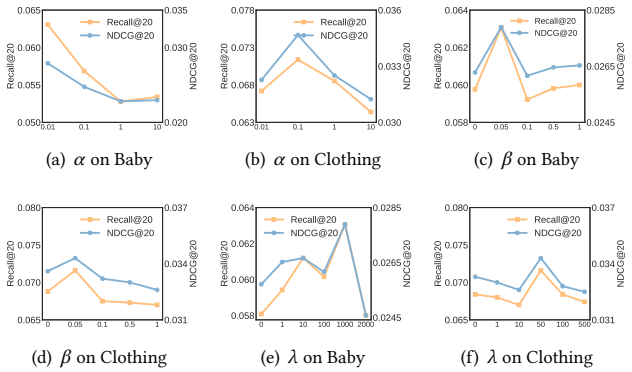
Figure 5: Effect of different modules on the robustness of the model.

superior performance. The boost of either module for MILK-*w/o BOTH* is huge. For example, in the FTMT setting, CMAM and CEIM improve MILK-*w/o BOTH* *w.r.t* NDCG@20 by 19.4%, and 24.27% on the Baby dataset, respectively.

Besides, we focus more on the generalization ability of the model in missing modality scenarios. As shown in Figure 5, we report the impact of different modules on the model in the full scenario (Full modality Training, Full modality Test (FTFT)) and missing modality scenario (FTMT). The bar charts represent the performance of the model variants under different scenarios. We can clearly see that the performance of the model drops significantly in the missing modality scenario. The line chart visually shows the magnitude of performance degradation for different model variants in the two Settings. The addition of the two modules reduces the loss of the model in missing modality scenarios. Specifically, on the Baby data, when no modules are added, the performance loss is 37.15%. After adding CEIM, the performance loss is 27.68%; The performance loss after adding CMAM is 32.06%. When the two modules cooperate, the performance loss is reduced to 24.59%. This shows that our model not only improves the performance of the model but also alleviates the performance degradation and enhances the generalization ability of the model.

5.4.2 Hyper-Parameter Sensitivities.

Effect of Dirichlet Distribution Parameters α. As illustrated in Figure 6(a) and (b), we sample weights from 4 different Dirichlet distributions and observe the performance of the model. Simply put, the smaller α is, the larger the weight gap is likely to be. In the FTMT setting, the model achieves the best results on the Baby and Clothing datasets with $\alpha = 0.01$ and $\alpha = 0.1$, respectively. On both datasets, the optimal value is achieved when α is small, which is due to the fact that one of the weights generated by Dirichlet



(a) α on Baby (b) α on Clothing (c) β on Baby

(d) β on Clothing (e) λ on Baby (f) λ on Clothing

Figure 6: Performance of different hyperparameters.

distribution at this time will be around 1, which guarantees that the heterogeneous environment is dominated by a certain modality. Small α value helps *MILK* to ensure that the modal composition in different environments is sufficiently different.

Effect of Cross-Modality Alignment Loss Weights β. As illustrated in Figure 6(c) and (d), we carefully tune the Cross-Modality alignment loss weights β on the Baby and Clothing datasets. We observe that *MILK* achieves the best performance when $\beta = 0.05$ on both datasets. When the weight is too small, the goal of capturing the information of other modalities cannot be achieved. When the weight is too large, the information of the modality itself may be seriously offset, resulting in performance degradation.

Effect of Cross-Environment Invariant Loss Weights λ. As illustrated in Figure 6(e) and Figure 6(f), we carefully tune the cross-environment consistency loss weights λ on the Baby and Clothing datasets. *MILK* achieves the best performance when $\lambda = 1000$ on both the Baby and $\lambda = 50$ on the Clothing datasets. The hyperparameter λ has a direct impact on the model. Too small λ results in cross-environment invariants where the invariant constraint is not sufficient to achieve our goal. Too large λ will cause the model to focus too much on consistency and not enough on learning the recommendation task itself. In the implementation process, it needs to be carefully tuned to find the optimum.

6 CONCLUSION

In this paper, we focused on the problem of new item recommendations in the missing modality scenario. We pointed out that existing works have too ideal assumptions about the data and are difficult to handle in real-world complex situations. We proposed *MILK* to solve the problem, encouraging the model to guarantee its performance as much as possible when the modality quality changes. Specifically, we designed the cross-modal alignment module so that the single modal representation can capture the signals of other modalities. Then we proposed a novel cyclic mixup method to construct multiple heterogeneous environments. Based on these environments, we optimized our model using invariant learning, encouraging it to learn the users' inherent content preferences that are kept invariant to arbitrary modality missing. Extensive experiments verify the effectiveness of *MILK*.

ACKNOWLEDGMENTS

This work was supported in part by grants from the National Key Research and Development Program of China(Grant No.2021ZD0111802), the National Natural Science Foundation of China(Grant No.U23B2031, 72188101), and the Fundamental Research Funds for the Central Universities (Grant No.JZ2023HGQA0471).

REFERENCES

[1] Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant Risk Minimization. *ArXiv* (2019).
[2] Haoyue Bai, Min Hou, Le Wu, Yonghui Yang, Richang Hong, and Meng Wang. 2023. GoRec: A Generative Cold-Start Recommendation Framework. *MM* (2023).
[3] Hao Chen, Zefan Wang, Feiran Huang, Xiao Huang, Yue Xu, Yishi Lin, Peng He, and Zhoujun Li. 2022. Generative Adversarial Framework for Cold-Start Item Recommendation. *SIGIR* (2022).
[4] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive Collaborative Filtering: Multimedia Recommendation with Item- and Component-Level Attention. *SIGIR* (2017).
[5] Yimeng Chen, Ruibin Xiong, Zhi-Ming Ma, and Yanyan Lan. 2022. When Does Group Invariant Learning Survive Spurious Correlations? *NeurIPS* (2022).
[6] Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. 2021. Environment inference for invariant learning. (2021).
[7] Tiago de Melo and Carlos M.S. Figueiredo. 2020. A first public dataset from Brazilian twitter and news on COVID-19 in Portuguese. *Data in Brief* (2020).
[8] Xiaoyu Du, Xiang Wang, Xiangnan He, Zechao Li, Jinhui Tang, and Tat-Seng Chua. 2020. How to Learn Item Representation for Cold-Start Multimedia Recommendation? *MM* (2020).
[9] Xiaoyu Du, Zike Wu, Fuli Feng, Xiangnan He, and Jinhui Tang. 2022. Invariant Representation Learning for Multimedia Recommendation. *MM* (2022).
[10] Xue Geng, Hanwang Zhang, Jingwen Bian, and Tat-Seng Chua. 2015. Learning image and user features for recommendation in social networks. *ICCV* (2015).
[11] Ruining He and Julian McAuley. 2015. VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback. *AAAI* (2015).
[12] Ruining He and Julian McAuley. 2016. Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. *WWW* (2016).
[13] Wang-Cheng Kang, Chen Fang, Zhaowen Wang, and Julian McAuley. 2017. Visually-Aware Fashion Recommendation and Design with Generative Image Models. *ICDM* (2017).
[14] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *ICLR* (2014).
[15] David Krueger, Ethan Caballero, Jörn-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Rémi Le Priol, and Aaron C. Courville. 2020. Out-of-Distribution Generalization via Risk Extrapolation (REx). *ICML* (2020).
[16] Xingchen Li, Xiang Wang, Xiangnan He, Long Chen, Jun Xiao, and Tat-Seng Chua. 2020. Hierarchical Fashion Graph Network for Personalized Outfit Recommendation. *SIGIR* (2020).
[17] Jianxun Lian, Xiaohuan Zhou, Fuzheng Zhang, Zhongxia Chen, Xing Xie, and Guang zhong Sun. 2018. xDeepFM: Combining Explicit and Implicit Feature Interactions for Recommender Systems. *KDD* (2018).
[18] Xinyu Lin, Wenjie Wang, Jujia Zhao, Yongqi Li, Fuli Feng, and Tat-Seng Chua. 2024. Temporally and Distributionally Robust Optimization for Cold-start Recommendation. *AAAI* (2024).
[19] Jiashuo Liu, Zheyuan Hu, Peng Cui, B. Li, and Zheyan Shen. 2021. Heterogeneous Risk Minimization. *ICML* (2021).
[20] Julian McAuley, Christopher Targett, Javen Qinfeng Shi, and Anton van den Hengel. 2015. Image-Based Recommendations on Styles and Substitutes. *SIGIR* (2015).
[21] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. 2016. Causal inference by using invariant prediction: identification and confidence intervals. *J R Stat Soc* (2016).
[22] Francesco Pinto, Harry Yang, Ser Nam Lim, Philip H. S. Torr, and Puneet Kumar Dokania. 2022. RegMixup: Mixup as a Regularizer Can Surprisingly Improve Accuracy and Out Distribution Robustness. *NeurIPS* (2022).
[23] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. *UAI* (2009).
[24] Suvash Sedhain, Scott Sanner, Darius Braziunas, Lexing Xie, and Jordan Christensen. 2014. Social collaborative filtering for cold-start recommendations. *RecSys* (2014).
[25] Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. 2021. Unshuffling data for improved generalization in visual question answering. (2021).
[26] Aäron van den Oord, Sander Dieleman, and Benjamin Schrauwen. 2013. Deep content-based music recommendation. *NeurIPS* (2013).
[27] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. 2018. Manifold Mixup: Better Representations by Interpolating Hidden States. *ICML* (2018).
[28] Maksims Volkovs, Guangwei Yu, and Tomi Poutanen. 2017. DropoutNet: Addressing Cold Start in Recommender Systems. *NeurIPS* (2017).
[29] Shuai Wang, Kun Zhang, Le Wu, Haiping Ma, Richang Hong, and Meng Wang. 2021. Privileged Graph Distillation for Cold Start Recommendation. *SIGIR* (2021).
[30] Zimu Wang, Yue He, Jiashuo Liu, Wenchao Zou, Philip S. Yu, and Peng Cui. 2022. Invariant Preference Learning for General Debiasing in Recommendation. *KDD* (2022).
[31] Wei Wei, Chao Huang, Lianghao Xia, and Chuxu Zhang. 2023. Multi-Modal Self-Supervised Learning for Recommendation. *WWW* (2023).
[32] Yin wei Wei, Xiang Wang, Qi Li, Liqiang Nie, Yan Li, Xuanping Li, and Tat-Seng Chua. 2021. Contrastive Learning for Cold-Start Recommendation. *MM* (2021).
[33] Yin wei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2020. Graph-Refined Convolutional Network for Multimedia Recommendation with Implicit Feedback. *MM* (2020).
[34] Yin wei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal Graph Convolution Network for Personalized Recommendation of Micro-video. *MM* (2019).
[35] Le Wu, Lei Chen, Pengyang Shao, Richang Hong, Xiting Wang, and Meng Wang. 2021. Learning Fair Representations for Recommendation: A Graph-based Perspective. *WWW* (2021).
[36] Le Wu, Xiangnan He, Xiang Wang, Kun Zhang, and Meng Wang. 2021. A Survey on Accuracy-Oriented Neural Recommendation: From Collaborative Filtering to Information-Rich Recommendation. *TKDE* (2021).
[37] Le Wu, Junwei Li, Peijie Sun, Richang Hong, Yong Ge, and Meng Wang. 2020. DiffNet++: A Neural Influence and Interest Diffusion Network for Social Recommendation. *TKDE* (2020).
[38] Hongyi Zhang, Moustapha Cissé, Yann Dauphin, and David Lopez-Paz. 2018. mixup: Beyond Empirical Risk Minimization. *ICLR* (2018).
[39] Jinghao Zhang, Qiang Liu, Shu Wu, and Liang Wang. 2023. Mining Stable Preferences: Adaptive Modality Decorrelation for Multimedia Recommendation. *SIGIR* (2023).
[40] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. 2021. Mining Latent Structures for Multimedia Recommendation. *MM* (2021).
[41] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Mengqi Zhang, Shu Wu, and Liang Wang. 2023. Latent Structure Mining With Contrastive Modality Fusion for Multimedia Recommendation. *TKDE* (2023).
[42] Xin Zhou and Zhiqi Shen. 2022. A Tale of Two Graphs: Freezing and Denoising Graph Structures for Multimodal Recommendation. *MM* (2022).
[43] Zhihui Zhou, Lili Zhang, and Ning Yang. 2023. Contrastive Collaborative Filtering for Cold-Start Item Recommendation. *WWW* (2023).
[44] Ziwei Zhu, Shahin Sefati, Parsa Saadatpanah, and James Caverlee. 2020. Recommendation for New Users and New Items via Randomized Training and Mixture-of-Experts Transformation. *SIGIR* (2020).