entropy

MDPI

Article

Invariant Representation Learning in Multimedia Recommendation with Modality Alignment and Model Fusion

Xinghang Hu [1] and Haiteng Zhang [2,*]

[1] School of Materials Science and Engineering, Sichuan University, Chengdu 610065, China; 2017141061004@stu.scu.edu.cn
[2] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
* Correspondence: zhanghaiteng22@mails.ucas.ac.cn

Abstract: Multimedia recommendation systems aim to accurately predict user preferences from multimodal data. However, existing methods may learn a recommendation model from spurious features, i.e., appearing to be related to an outcome but actually having no causal relationship with the outcome, leading to poor generalization ability. While previous approaches have adopted invariant learning to address this issue, they simply concatenate multimodal data without proper alignment, resulting in information loss or redundancy. To overcome these challenges, we propose a framework called M^3-InvRL, designed to enhance recommendation system performance through common and modality-specific representation learning, invariant learning, and model merging. Specifically, our approach begins by learning modality-specific representations along with a common representation for each modality. To achieve this, we introduce a novel contrastive loss that aligns representations and imposes mutual information constraints to extract modality-specific features, thereby preventing generalization issues within the same representation space. Next, we generate invariant masks based on the identification of heterogeneous environments to learn invariant representations. Finally, we integrate both invariant-specific and shared invariant representations for each modality to train models and fuse them in the output space, reducing uncertainty and enhancing generalization performance. Experiments on real-world datasets demonstrate the effectiveness of our approach.

Keywords: multimedia recommendation; model fusion; multimodal representation

Academic Editor: Deniz Gençağa

Received: 17 November 2024
Revised: 20 December 2024
Accepted: 4 January 2025
Published: 10 January 2025

Citation: Hu, X.; Zhang, H. Invariant Representation Learning in Multimedia Recommendation with Modality Alignment and Model Fusion. *Entropy* **2025**, *27*, 56. https://doi.org/10.3390/e27010056

1. Introduction

Recommendation systems is a useful tool to address information overload [1–3]. Multimedia recommendation systems (MRS) utilize user-item interactions and multimodal features such as text, images, audio, and videos to provide content recommendations based on user preferences [4–7]. They play a crucial role in platforms like e-commerce [8], social media [9], and video sharing [10], enhancing recommendation accuracy by capturing user preferences at the fine-grained level [4,11,12]. Early methods, such as VBPR [13] and DeepStyle [14], integrated multimodal information into traditional collaborative filtering paradigms but overlooked high-order user-item interaction connectivity [15]. Recent approaches, including MMGCN [16], GRCN [17], LATTICE [5], and DualGNN [18], employ graph convolution network (GCN) to better represent user-item interactions and improve recommendation performance [19,20].

Despite progress, many multimedia recommendation methods face out-of-distribution (OOD) generalization issues, where models trained on one data distribution perform poorly when applied to data from a different distribution [21–25]. For instance, as shown

in Figure 1, the user likes dinosaur movies, especially Jurassic Park, but if the movie is recommended based on the director, Spielberg, the user's true preference for dinosaur themes is ignored. In this case, the association of Spielberg's label with the user's preferences is misleading, resulting in inaccurate recommendations. In other words, the Spielberg is the spurious texture feature, and the dinosaur is the causal texture feature.

Figure 1. Schematic diagram of spurious correlation in MRS.

To address these issues, invariant representation learning (IRL) has been proposed, aiming to learn features consistent across different environments [26], such as invariant risk minimization (IRM) [27–29]. However, in multimedia recommendation systems, methods like InvRL [30] and PaInvRL [31] may fail to fully align and interact between modalities, limiting recommendation performance.

Aligning modality-specific information is crucial for effective recommendations. However, simply aligning all modalities in a shared space is insufficient, as different modalities, such as audio, text, and images, capture various aspects of user preferences [32]. For example, in recommending an action movie, intense sound effects may indicate a preference for a tense atmosphere, text descriptions might reveal interest in the storyline, and posters or images could highlight an affinity for visual elements. Thus, integrating modality-specific information can prevent generalization issues associated with a single shared space. Another challenge associated with the alignment of modality-specific information is determining the contribution of each modality to the final prediction. To address this, a weighted fusion method is proposed, allowing the flexible adjustment of modality weights to ensure effective integration without over-reliance on any single modality.

In this paper, we propose the invariant representation learning in multimedia recommendation with modality alignment and model fusion framework (M^3-InvRL), which integrates multimodal representation, invariant learning, and model merging. Our approach introduces a novel contrastive representation learning method that decomposes each modality into common and specific components, extracting invariant features through environment identification and mask generation. These features are then merged and predicted for each modality, followed by weighted model merging in the output space. The main contributions of this paper are summarized as follows:

- We propose to learn both shared and modality-specific representations to mitigate the generalization issues of relying on a single shared space. By aligning individual modality representations with the complete set of modalities, the framework effectively integrates and complements information across modalities.
- We introduce a new multimedia recommendation framework, M^3-InvRL, which maps modality features into shared and specific spaces to learn invariant representations for

each component. We utilize model merging to fully leverage all available invariant information, adaptively adjusting the weights of different modality predictors to enhance the model's generalization ability.
- We conduct extensive experiments on two real-world datasets to demonstrate the effectiveness of our proposed framework.

2. Related Work

2.1. Collaborative Filtering for Recommendation

Collaborative filtering (CF) is a foundational approach in recommendation systems, modeling the similarity between items and users to recommend similar items to similar users [33]. The core model in CF is matrix factorization (MF) [34], where each user and item is assigned a latent embedding, and similarity is assessed via the inner product of these embeddings. NCF [35] introduces neural networks to model similarities and proves that MF is a special case of NCF. NGCF [36] encodes high-hop neighbor information among users and items into embeddings using graph convolutional network (GCN). LightGCN [37] simplifies NGCF by removing feature transformations and nonlinear activations in the original NGCF architecture that are unsuitable for CF tasks. UltraGCN [38] further enhances efficiency by bypassing infinite layers of message passing in NGCF and LightGCN. We adopt UltraGCN as our backbone due to its simplicity and efficiency.

2.2. Multimedia Recommendation

Multimedia recommendation systems utilize multimodal information, such as visual, acoustic, and textual data, to enhance performance by better capturing user preferences [14,39,40]. Early works like VBPR integrated visual and item ID embeddings into a unified item embedding for further training [13]. DVBPR [41] extends the idea of VBPR by proposing an end-to-end architecture for jointly learning image representations and user-item embeddings. Later approaches introduced attention mechanisms to adaptively select multimodal features [42,43]. For instance, VECF [44] learns attention to sub-areas of images to make better image representations. UVCAN [45] uses attention mechanisms to learn multimodal information from both user and item perspectives. MAML [46] models each user's attention to different aspects of an item by extracting multimodal features using an attention neural network. Recently, graph neural networks have been employed to model higher-order user-item interactions. MMGCN [16] learns modal-specific representations to better capture user preferences via the message-passing idea of GNN. LATTICE [5] constructs item–item graphs to improve item embeddings. However, these methods may fail when facing a distribution shift between training and test data, resulting in sub-optimal test performance.

2.3. Invariant Representation Learning

Invariant representation learning seeks to develop representations vital for downstream tasks, particularly by addressing distribution shifts between training and test data through consistent representations across diverse environments, thus improving generalization [47]. Invariant risk minimization (IRM) [27] is a seminal approach, with extensions in information theory [48,49], regularization [29,50], and sparsity [51]. Methods like EIIL [52] and HRM [53] automatically partition environments when labels are unavailable. Another approach involves constructing unbiased losses and optimizing models accordingly [54–57], including propensity score-based [58], doubly robust [59–63], and data fusion debiasing methods [64,65]. In this work, we capture invariant features using soft masks from heterogeneous environments and different modalities.

3. Preliminaries

In our multimedia recommendation model, the data mainly consists of two parts: users u and items i, which are represented by sets $\mathcal{U}$ and $\mathcal{I}$, respectively. An interaction between a user and an item is represented as $(u, i) \in \mathcal{U} \times \mathcal{I}$, where r_{ui} represents the result of the interaction. If the interaction is positive, $r_{ui} = 1$; otherwise, it is 0. The modal information for item i is represented as $\left\{x^i_{M+1} = \left(x^i_1, \cdots, x^i_M\right)\right\}^N_{i=1}$, where each $x^i_m \in \mathcal{R}^{d_m}$ corresponds to a specific modality m. The parameter d_m denotes the dimension of each modality. The multi-modal recommendation aims to learn a model $\Gamma\left(u, i, x^i_{M+1} \mid \Theta\right)$, where Θ denotes the parameters of the recommendation model Γ, to predict users' true preferences.

$$\arg\min_{\Theta} \mathcal{L}\left(\Gamma\left(u, i, x^i_{M+1} \mid \Theta\right) \mid \mathcal{R}^{tr}\right), \tag{1}$$

where $\mathcal{L}(\cdot)$ denotes the recommendation loss, and $\mathcal{R}^{tr}$ denotes the training set, with both positive samples $\mathcal{R}^+ = \{(u, i) : r_{u,i} = 1\}$ and negative samples $\mathcal{R}^- = \{(u, i) : r_{u,i} = 0\}$. For easy reading, we provide the descriptions of all used variable in Table 1.

Table 1. List of all variables used in this paper and their corresponding descriptions.

Variable	Description
$u, \mathcal{U}$	User u in the recommendation system, and $\mathcal{U}$ is the set of all users.
$i, \mathcal{I}$	Item i in the recommendation system, and $\mathcal{I}$ is the set of all items.
r_{ui}	Binary interaction: $r_{ui} = 1$ if user u positively interacts with item i, 0 otherwise.
$\mathcal{R}$	Set of all user-item interactions, where $\mathcal{R}^+$ denotes positive samples ($r_{ui} = 1$), and $\mathcal{R}^-$ denotes negative samples ($r_{ui} = 0$).
x^i_m	Feature of item i for modality m, $m \in \{1, \ldots, M\}$.
d_m	Dimension of the feature vector for modality m.
$\Gamma(\cdot)$	Recommendation model predicting user preferences.
Θ	Parameters of the recommendation model Γ.
$f_r(\cdot)$	Base encoder for the r-th modality, $r = 1, \ldots, M + 1$.
h_r	Representation generated by $f_r(\cdot)$ for the r-th modality.
$g(\cdot)$	Shared head mapping representations to a common space $\mathcal{Z}$.
z_m, z_{M+1}	Shared representations for modality m and all modalities, respectively.
$k_m(\cdot)$	Specific head generating modality-specific representations.
v_m	Modality-specific representation for modality m.
$s_{m,n}(i, j)$	Similarity score between modality m (sample i) and modality n (sample j).
$\mathcal{L}_{\text{com}}(B)$, $\mathcal{L}_{\text{MI}}$	Common loss across modalities and mutual information loss.
Φ^i_r, Ψ^i_r	Invariant and variant representations for modality r of item i.
$\mathbf{m}$	Invariant mask in $[0, 1]$, used to generate invariant representations.
Q_m	Combined invariant representations for modality m, $Q_m = [\Phi_m; \Phi_{M+1}]$.
$\Gamma^*_m(\cdot)$	Final recommendation model for modality m, trained on Q_m.
λ_m	Entropy-based uncertainty for the m-th modality.
ω_m	Importance weight for the m-th modality.
Y_{avg}	Final prediction by aggregating all predictors.

4. Methods

In this section, we introduce the overall framework of M^3-InvRL, as illustrated in Figure 2, which includes multimedia representation learning, invariant learning, and model merging.

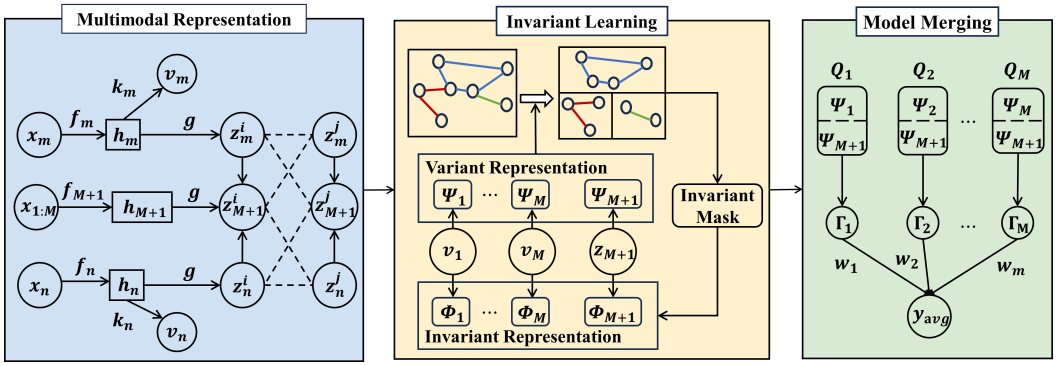

Figure 2. Overall framework of M^3-InvRL includes multimedia representation, invariant representation, and model merging.

4.1. Multimodal Representation for Recommendation

In this section, we first describe the modal-specific representation and one common representation for each modality. We introduce a novel contrastive loss that aligns the representation and imposes mutual information constraints to extract modality-specific features, preventing generalization issues within the same representation space. Next, we discuss the details of our method.

We use base encoders f_r to generate d-dimensional representations $h_r = f_r(x_r; \zeta_r)$ for $r = 1, \ldots, M+1$, where h_{M+1} represents the intermediate representation of $x_{1:M}$. The shared head g maps these to a common space $\mathcal{Z}$, generating shared representations $z_m = g(h_m; \theta)$ for each modality and complete common representation $z_{M+1} = g(h_{M+1}; \theta)$ for all modalities. Specific heads k_m generate modality-specific representations $v_m = k_m(h_m; \eta_m)$ for $m = 1, \ldots, M$.

We define $\mathrm{sim}(u, v)$ as the similarity measure between vectors u and v, such as cosine similarity $\mathrm{sim}(u, v) = \frac{u \cdot v}{\|u\| \|v\|}$. The similarity is scaled by a learnable temperature hyperparameter τ to yield the similarity score, where a larger τ reduces the distinction between similar and dissimilar samples, and a smaller τ enhances this difference. In our paper, τ helps balance the influence of positive and negative sample pairs.

$$s_{m,n}(i, j) = \exp(\mathrm{sim}(z_m^i, z_n^j)/\tau), \tag{2}$$

where z_m^i and z_n^j are the representations of the mth and nth modalities corresponding to the ith and jth samples from a mini-batch B, respectively.

We define (z_m^i, z_{M+1}^i) for $i = 1, \cdots, B$ as positive pairs, the remaining pairs are the negative pairs,

$$\Omega_m(i) = \sum_{i \neq j} (s_{m,M+1}(i, j) + s_{m,m}(i, j) + s_{M+1,M+1}(i, j)) \tag{3}$$

is the sum of similarities among negative pairs that correspond to the positive pair (z_m^i, z_{M+1}^i), and the contrastive loss for the same pair of samples is

$$l_m(i) = -\log \frac{s_{m,M+1}(i, i)}{\Omega_m(i)}. \tag{4}$$

We combine the loss terms for each modality $m = 1, \cdots, M$ and obtain the common loss

$$\mathcal{L}_{\mathrm{com}}(B) = \sum_{m=1}^{M} \sum_{i=1}^{B} l_m(i). \tag{5}$$

Aligning all modalities in a single shared space can lead to generalization issues and loss of unique modality-specific information. To preserve the distinctiveness of modality-specific representation $v_m(x)$ relative to modality-shared features $z_m(x)$, the goal is to minimize the mutual information

$$\mathcal{L}_{\text{MI}} = \sum_{m=1}^{M} \text{CLUB}(v_m(x), z_m(x)), \tag{6}$$

where $\text{CLUB}(V, W)$ is the estimator for the contrastive log-ratio upper bound of mutual information between two random variables V and W [66].

4.2. Invariant Learning for Recommendation

Invariant learning [28,67] encourages models to concentrate on stable representations across different environments. Within our multimodal framework, it is applied to modality-specific representation v_m and complete common representation z_{M+1} to learn invariant representations $\{\Phi_r\}_{r=1}^{M+1}$.

Environment Identification. We take historical user-item interactions as input and partition them into a set of environments $\mathcal{E}$, which supports the generation of invariant masks for the subsequent stages of learning.

During the environment identification stage, we aim to learn environment-specific representations $e_r \in \mathcal{E}$ by training a recommendation model $\Gamma_{(e_r)}(u, i, \Psi_r^i | \Theta_{e_r})$ for each environment e_r. Here, Ψ_r^i denotes variant representations with item i and Θ_{e_r} represents the model's parameters in the environment e_r:

$$\arg\min_{\Theta_{e_r}} \mathcal{L}\left(\Gamma_{(e_r)}\left(u, i, \Psi_r^i \mid \Theta_{e_r}\right) \mid \mathcal{R}_{e_r}^{tr}\right), \tag{7}$$

where the variant representations Ψ_r^i are obtained by initializing the invariant mask. We employ UltraGCN [38] as the recommendation model and drive the representations through a graph-based loss function $\mathcal{L}$ to encode the user-item graph.

Once the environment-specific representations are learned, the user-item interactions are assigned to the corresponding environments by maximizing the recommendation model output for each interaction:

$$\mathcal{R}_{e_r} = \arg\max_{e_r \in \mathcal{E}} \Gamma_{(e_r)}(u, i, \Psi_r^i | \Theta_{e_r}). \tag{8}$$

The environment-specific interaction sets $\{\mathcal{R}_{e_r} | e_r \in \mathcal{E}\}$ are then used to guide the invariant representation learning.

Invariant Representation Learning. We minimize $\mathcal{L}_r^{mask}$ by optimizing the mask $\mathbf{m}$. To constrain that each $\mathbf{m}_i$ in the mask is between $[0, 1]$, we use the softmax function.

Followed with the prediction model Γ_r^{mask} converging, the invariant representations

$$\begin{aligned} \Phi_r^i &= \mathbf{m}_r^i \odot v_r^i, r = 1, \cdots, M, \\ \Phi_r^i &= \mathbf{m}_r^i \odot z_r^i, r = M + 1, \end{aligned} \tag{9}$$

and variant representations

$$\begin{aligned} \Psi_r^i &= (1 - \mathbf{m}_r^i) \odot v_r^i, r = 1, \cdots, M, \\ \Psi_r^i &= (1 - \mathbf{m}_r^i) \odot z_r^i, r = M + 1. \end{aligned} \tag{10}$$

4.3. Model Merging for Recommendation

In multimodal fusion, a key challenge is that the contribution of each modality to the final prediction is uncertain. To address this, we apply a weighted fusion strategy that adjusts the importance of each modality based on its uncertainty.

Our approach concatenates the invariant representation Φ_{M+1} in the common space and the model-specific invariant representation Φ_m to obtain the combined modality feature Q_m, defined as

$$Q_m = [\Phi_m; \Phi_{M+1}]. \tag{11}$$

Thus, we learn the final recommendation model $\Gamma_m^*(u, i, Q_m|\Theta_m^*)$ based on the combined representation Q_m in each modality. The learning objective in Equation (1) can be rewritten as

$$\arg\min_{\Theta_m} \mathcal{L}(\Gamma_m^*(u, i, Q_m|\Theta_m^* \mid \mathcal{R}^{tr}). \tag{12}$$

When one modality exhibits higher uncertainty in its predictions, it becomes more prone to making incorrect predictions. Consequently, we leverage the prediction uncertainty as a proxy to gauge the importance of each modality.

$$\lambda_m = -p_m^T \log p_m, \tag{13}$$

where $p_m = \text{softmax}(\Gamma_m^*(u, i, Q_m|\Theta_m^*))$.

A higher entropy λ_m indicates lower confidence in the prediction, leading to a smaller importance weight during the model merging process. Based on this, we calculate the importance weight for a mth modality predictor as

$$\omega_m = \frac{\exp(\max_{m=1,\cdots,M} \lambda_m - \lambda_m)}{\sum_{i=1}^{M} \exp(\max_{m=1,\ldots,M} \lambda_m - \lambda_i)}. \tag{14}$$

The final prediction is obtained by aggregating the outputs of all predictors. We use a weighted sum to combine the predictions, ensuring that the weights sum to one. Specifically, the final result Y_{avg} is given by

$$Y_{\text{avg}} = \sum_{m=1}^{M} \omega_m \Gamma_m^*(u, i, Q_m|\Theta_m^*). \tag{15}$$

5. Results

5.1. Datasets

Following previous work [17,30,31,68], we conducted experiments using two publicly available multimedia datasets: **Tiktok** (https://github.com/nickwzk/InvRL, accessed on 10 October 2022) and **Movielens** (https://github.com/nickwzk/InvRL, accessed on 10 October 2022). The **Tiktok** dataset contains short micro-videos, while the **Movielens** dataset consists of user movie viewing histories. Both datasets include multimedia representations extracted from visual, acoustic, and textual content. The representations of the **Tiktok** dataset are extracted and provided officially. The visual, acoustic, and textual representations of the **Movielens** dataset were extracted by [16] with pre-trained ResNet50 for visual representations, VGGish [69] for acoustic representations, and [70] for textual representations. Note that there are many widely used datasets such as **Kwai** (https://github.com/nickwzk/InvRL, accessed on 10 October 2022) included in previous work. However, since such datasets only contain one modality, we excluded this dataset

from our experiment. The summary statistics of the **Tiktok** and **Movielens** datasets are shown in Table 2.

Table 2. The statistics of datasets. d_V , d_A, and d_T denote the dimensions of visual, acoustic, and textual modalities. # means "the numbers of".

Dataset	#Interactions	#Items	#Users	Sparsity	d_V	d_A	d_T
Movielens	1,239,508	5986	55,485	99.63%	2048	128	100
Tiktok	726,065	76,085	36,656	99.99%	128	128	128

5.2. Experiment Details

We adopted Adam [71] as the optimizer and implemented our models using PyTorch 1.11.0, running on an NVIDIA V100 GPU. The batch size was set to 512, and the number of environments was selected from {5, 10, 15, 20, 25}. The learning rate was tuned within the set {0.01, 0.001, 0.0001}. For the regularization parameters, λ_{com} was chosen from {0.1, 1, 2, 5, 10} and λ_{MI} from {0.01, 0.1, 1, 10}. Additionally, γ and ρ were selected from {0.01, 0.1, 0.5, 1, 5}, while κ and ν were chosen from {0.1, 1, 5}. The temperature hyperparameter τ was tuned within the range {0.1, 0.5, 1, 5, 10}. The iteration parameter T was initially set to 5, and training was conducted for 200 epochs.

5.3. Baselines

We evaluated our model against several state-of-the-art multimedia recommendation methods. The M-CF models, including VBPR [13], CB2CF [72], and DUIF [73], integrate multimedia content into traditional collaborative filtering approaches. G-NCF models, such as DisenGCN [74], MacridVAE [75], and NGCF [36], employ neural networks to capture complex user-item interactions. M-NCF models, including HUIGN [68], GRCN [17], and MMGCN [16], specialize in neural CF for multimedia content. InvRL models, such as InvRL [30], introduce invariant learning. UltraGCN [38] served as the backbone, simplifying graph CF through regularization and improving efficiency.

5.4. Evaluation Metrics

We used three widely-used evaluation metrics: Precision@K (P@K), Recall@K (R@K), and NDCG@K (N@K), to measure the ranking performance of our proposed method. Precision@K calculates the average of the proportion of the corrected recommended items among the top K predicted items for each user. Recall@K calculates the average of the proportion of the corrected recommended items among the sum of the corrected recommended items and the wrongly missed items in the top K predicted items for each user. NDCG@K, short for normalized discounted cumulative gain at K, measures the order of the corrected recommended items in the top K predicted items. Higher values of the three metrics indicate better ranking performance of our proposed method. In our experiments, K was set to 10.

5.5. Overall Performance

We report the performance of various methods on both **Tiktok** and **Movielens** datasets in Table 3, where the best-performing method is bolded for each metric. We have the following observations.

Firstly, multi-modality-based methods outperform single-modality-based methods, emphasizing the critical role of integrating multi-modality information to enhance recommendation performance. M^3-InvRL achieves the most competitive performance among all the methods.

Secondly, compared to the Naive-UltraGCN, the incorporation of InvRL on UltraGCN (InvRL) enhances the recommendation performance through the introduction of invariant representation learning. On the other hand, our proposed M^3-InvRL further enhances the recommendation performance on InvRL. On the **Movielens** dataset, M^3-InvRL outperforms InvRL by 4.65% in Precision@10, 6.11% in Recall@10, and 0.89% in NDCG@10. On the **Tiktok** dataset, M^3-InvRL surpasses InvRL with a 3.13% increase in Precision@10, 3.49% increase in Recall@10, and 3.79% increase in NDCG@10. We can conclude that unlike InvRL's direct concatenation of representations, M^3-InvRL achieves higher performance by aligning modalities through multimodal contrastive representation learning and applying model merging in each modality prediction model.

Table 3. Performance comparison across datasets using Precision@10, Recall@10, and NDCG@10. The best result is bold. The second best result is underlined.

Category	Methods	Movielens			Tiktok		
		P@10	**R@10**	**N@10**	**P@10**	**R@10**	**N@10**
M-CF	VBPR	0.0512	0.1990	0.2261	0.0118	0.0628	0.0574
	DUIF	0.0538	0.2167	0.2341	0.0087	0.0483	0.0434
	CB2CF	0.0548	0.2265	0.2505	0.0109	0.0642	0.0613
G-NCF	NGCF	0.0547	0.2196	0.2342	0.0135	0.0780	0.0661
	DisenGCN	0.0555	0.2222	0.2401	0.0145	0.0760	0.0639
	MacridVAE	0.0576	0.2286	0.2437	0.0152	0.0813	0.0686
M-NCF	MMGCN	0.0581	0.2345	0.2517	0.0144	0.0808	0.0674
	HUIGN	0.0619	0.2522	0.2677	0.0164	0.0884	0.0769
	GRCN	0.0639	0.2569	0.2754	0.0195	0.1048	0.0938
UltraGCN	Naive-UltraGCN	0.0624	0.2547	0.2691	0.0183	0.0981	0.0878
	UltraGCN + InvRL	0.0645	0.2615	0.2815	0.0192	0.1062	0.0922
	M^3-InvRL(Ours)	**0.0675**	**0.2775**	**0.2840**	**0.0198**	**0.1099**	**0.0957**
%Improvement over Naive-UltraGCN		8.17%	8.95%	5.55%	8.20%	12.03%	9.00%
%Improvement over UltraGCN + InvRL		4.65%	6.11%	0.89%	3.13%	3.49%	3.79%

5.6. Performance Comparison with Different Values of K

To highlight the improvements of M^3-InvRL, we conducted a comparative analysis between M^3-InvRL and its backbone model, UltraGCN, by evaluating their top-*K* scores. Figure 3 illustrates the curves for NDCG, prediction, and recall scores on the **TikTok** dataset.

M^3-InvRL consistently outperforms Naive-UltraGCN and UltraGCN + InvRL across all three metrics. Specifically, in Precision@K, M^3-InvRL demonstrates higher accuracy and maintains superior prediction scores across various K values, indicating its effectiveness in identifying the most relevant items at the top of the recommendation list. In Recall@K, M^3-InvRL achieves higher recall, particularly as K increases, showcasing its ability to retrieve more relevant items in scenarios where maximizing relevant item retrieval is essential. Finally, in NDCG@K, which considers both the relevance and ranking of recommended items, M^3-InvRL not only identifies relevant items but also ranks them more effectively, leading to significant performance improvements over other approaches. These consistent enhancements across different evaluation metrics underscore the robustness and effectiveness of M^3-InvRL in delivering high-quality recommendations.

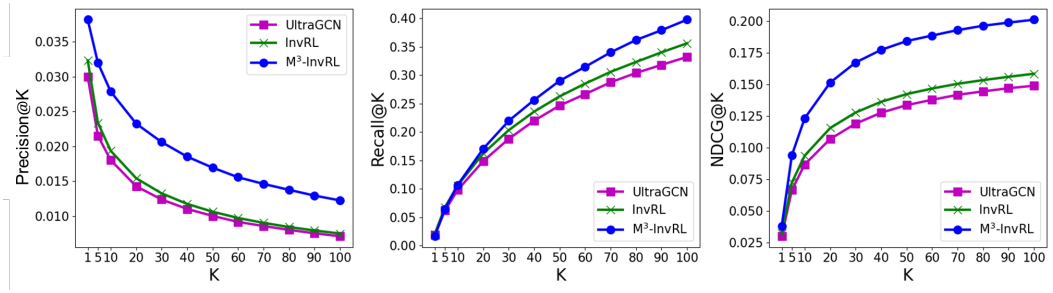

Figure 3. The comparison among Naive-UltraGCN (UltraGCN), UltraGCN + InvRL (InvRL) and M^3-InvRL on **Tiktok** datasets with respect to Precision@K, Recall@K, NCDG@K.

5.7. Effect of $\mathcal{L}_{\text{com}}$ and $\mathcal{L}_{\text{MI}}$

In this section, we examine the impact of the common loss $\mathcal{L}_{\text{com}}$ and the mutual information loss $\mathcal{L}_{\text{MI}}$ on the model's performance. We do this by removing each loss during the training process of M^3-InvRL. For comparative purposes, we evaluate the following three models: M^3-InvRL without the common loss $\mathcal{L}_{\text{com}}$ (denoted as M^3-InvRL w/o $\mathcal{L}_{\text{com}}$), M^3-InvRL without the mutual information loss $\mathcal{L}_{\text{MI}}$ (denoted as M^3-InvRL w/o $\mathcal{L}_{\text{MI}}$), and the original M^3-InvRL model. The experimental results are presented in Table 4.

Our observations indicate that removing the common loss $\mathcal{L}_{\text{com}}$ leads to a performance decline across both datasets. This highlights the crucial role of aligning common representations in multimodal representation learning. Similarly, the removal of the mutual information loss $\mathcal{L}_{\text{MI}}$ negatively affects the model's performance. This suggests that relying on a single shared representation space may restrict the model's generalization capabilities, underscoring the importance of $\mathcal{L}_{\text{MI}}$ in effectively capturing modality-specific features.

Furthermore, we note that the performance drop in M^3-InvRL w/o $\mathcal{L}_{\text{com}}$ is more pronounced than in M^3-InvRL w/o $\mathcal{L}_{\text{MI}}$. This demonstrates that common representations are pivotal in determining user preferences, while modality-specific representations play a significant supplementary role.

Table 4. Performance comparison with different loss components. The best result is bold. The second best result is underlined.

	Movielens			**Tiktok**		
	P@10	**R@10**	**N@10**	**P@10**	**R@10**	**N@10**
M^3-InvRL w/o $\mathcal{L}_{\text{com}}$	0.0642	0.2648	0.2792	0.0190	0.1030	0.0925
M^3-InvRL w/o $\mathcal{L}_{\text{MI}}$	<u>0.0667</u>	<u>0.2753</u>	<u>0.2836</u>	<u>0.0194</u>	<u>0.1093</u>	<u>0.0931</u>
M^3-InvRL	**0.0675**	**0.2775**	**0.2840**	**0.0198**	**0.1099**	**0.0957**

5.8. Different Model Merging Strategy

To validate the effectiveness of our proposed model merging strategy, we conducted experiments using three additional weighting methods: equal weighting (E-weight), loss-based weighting (L-weight), and attention mechanism-based weighting (A-weight). In the equal weighting strategy, each modality model is assigned an equal weight of 1/3. The loss-based weighting strategy builds upon this by assigning weights based on the ratio of each modality's loss to the total loss across all modalities, thereby giving more importance to modalities that contribute less error. The attention mechanism-based weighting further enhances the approach by dynamically adjusting weights according to the relevance of each modality's information.

As shown in Table 5, the loss-based strategy performs almost identically to equal weighting, indicating that merely acknowledging the differences between modalities does not improve overall performance. However, the attention mechanism strategy significantly

enhances model performance compared to both the loss-based and equal weighting strategies. This suggests that by dynamically adjusting weights based on the importance of each modality in varying contexts, the merging mechanism can improve model performance. Furthermore, the M^3-InvRL model achieves the best performance across all metrics. By employing entropy-based weights as a proxy for model uncertainty, M^3-InvRL dynamically and accurately allocates weights, effectively leveraging the strengths of each modality and reducing uncertainty. This leads to superior overall performance, as demonstrated by the experimental results.

Table 5. Performance comparison on different weight strategies. The best result is bold. The second best result is underlined.

	Movielens			**Tiktok**		
	P@10	**R@10**	**N@10**	**P@10**	**R@10**	**N@10**
E-weight	0.0652	0.2731	0.2829	0.0192	0.1080	0.0911
L-weight	0.0648	0.2719	0.2817	0.0193	0.1073	0.0937
A-weight	<u>0.0670</u>	<u>0.2761</u>	<u>0.2834</u>	<u>0.0195</u>	**0.1105**	<u>0.0955</u>
M^3-InvRL	**0.0675**	**0.2775**	**0.2840**	**0.0198**	<u>0.1099</u>	**0.0957**

5.9. Study on the Number of Environments

To assess how the number of environments impacts the performance of M^3-InvRL compared to InvRL, we conducted experiments on the **Tiktok** and **Movielens** datasets with varying numbers of environments. As illustrated in Figure 4, M^3-InvRL consistently surpasses InvRL in NDCG@10 across different environment counts. A key advantage of M^3-InvRL is its use of weighted averaging after adapting to each modality's environment, which reduces the model uncertainty and enhances flexibility. In contrast, InvRL simply concatenates modes as a representation to learn invariant representations. By learning invariant representations separately from specific and common complete representations, M^3-InvRL facilitates easier differentiation between environments. In the **Tiktok** dataset, using approximately 10 environments yields the best performance, as this number allows M^3-InvRL to effectively distinguish between variant and invariant information, thereby enhancing recommendation quality. For the **Movielens** dataset, the performance improves with an increasing number of environments, suggesting that a larger number of environments is more suitable for this dataset.

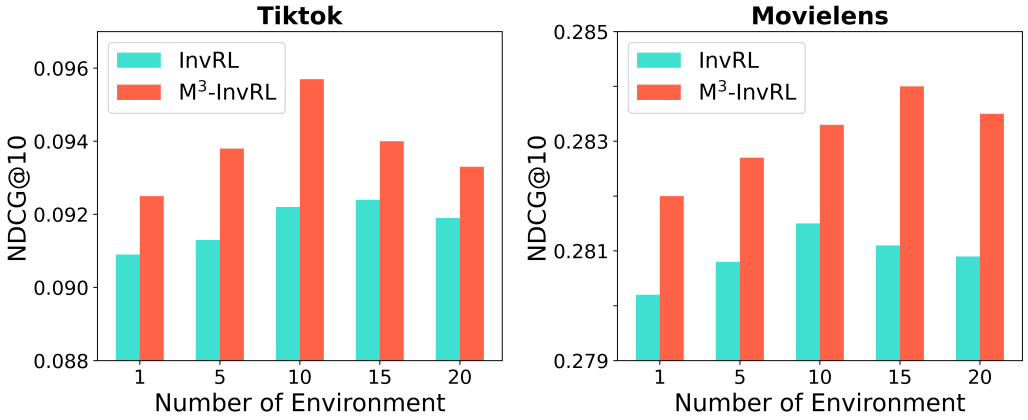

Figure 4. Experimental comparison of different environment numbers $|\mathcal{E}|$.

6. Discussion

In this work, we propose an invariant representation learning framework (M^3-InvRL) to enhance the generalization ability of multimedia recommendation systems, particularly in the presence of distribution shifts between training and testing data. Our method achieves up to an 8.95% improvement in ranking performance on the **Movielens** dataset and a 12.03% improvement on the **Tiktok** dataset over the Naive-UltraGCN model. Compared to the UltraGCN + InvRL method, our approach yields up to a 6.11% improvement on the **Movielens** dataset and 3.79% on the **Tiktok** dataset. These improvements stem from three key components of our framework.

The first component involves the separation of common and modality-specific representations. For each modality, we use different heads to transform the original representation into common and modality-specific parts. A common loss and a mutual information loss are then combined to enhance the representation capabilities of the common representation and the distinctiveness of the modality-specific representation relative to the shared representation. This separation guides the model to learn more representative features for downstream tasks. Experiments demonstrate that both types of representations contribute to the model's performance. The second component is invariant representation learning applied to both common and modality-specific representations. This approach endows our model with the ability to maintain robustness when faced with distribution shifts between training and testing data. The third component involves model merging through an ensemble of modality-level predictions. Unlike existing works [30] that train a single model on concatenated features from multiple modalities, we train a distinct model for each modality to capture modality-specific information and merge the results based on their importance to overall performance. This enables our model to learn and adjust its focus on relevant information.

Despite the advantages of our proposed method, there are still improvements that can be made in the future. The first one is the determination of the number of environments. In this work, we predefined the number of environments, but optimal numbers vary across datasets. Developing an adaptive method to automatically determine the number of environments would be beneficial. Second, we may enhance the efficiency of our method, as dividing it into three consecutive parts may increase the training costs. An end-to-end approach that integrates these components could significantly improve the efficiency in the future.

7. Conclusions

Our M^3-InvRL framework enhances the generalization ability of multimedia recommendation systems in the presence of distribution shifts between training and testing data. Specifically, our approach learns both shared and modality-specific invariant representations. By utilizing modal-specific and common representations, invariant learning, and adaptive model merging techniques, our method effectively addresses issues related to spurious feature learning and misalignment.

Author Contributions: Conceptualization, X.H. and H.Z.; methodology, H.Z.; software, X.H.; validation, X.H.; formal analysis, H.Z.; writing—original draft preparation, X.H. and H.Z.; writing—review and editing, X.H. and H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Restrictions apply to the availability of these data. Data were obtained from GitHub and are available https://github.com/nickwzk/InvRL (accessed on 10 October 2022) with the permission of GitHub.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lv, Z.; Zhang, W.; Chen, Z.; Zhang, S.; Kuang, K. Intelligent model update strategy for sequential recommendation. In Proceedings of the WWW, Singapore, Singapore, 13–17 May 2024.
2. Lv, Z.; He, S.; Zhan, T.; Zhang, S.; Zhang, W.; Chen, J.; Zhao, Z.; Wu, F. Semantic Codebook Learning for Dynamic Recommendation Models. In Proceedings of the ACM MM, Melbourne, Australia, 28 October–1 November 2024.
3. Wang, F.; Zhu, H.; Srivastava, G.; Li, S.; Khosravi, M.R.; Qi, L. Robust collaborative filtering recommendation with user-item-trust records. *IEEE Trans. Comput. Soc. Syst.* **2021**, *9*, 986–996. [CrossRef]
4. Mu, Z.; Zhuang, Y.; Tan, J.; Xiao, J.; Tang, S. Learning hybrid behavior patterns for multimedia recommendation. In Proceedings of the ACM MM, Lisboa, Portugal, 10–14 October 2022.
5. Zhang, J.; Zhu, Y.; Liu, Q.; Wu, S.; Wang, S.; Wang, L. Mining latent structures for multimedia recommendation. In Proceedings of the ACM MM, Virtual, 20–24 October 2021.
6. Lin, Z.; Tan, Y.; Zhan, Y.; Liu, W.; Wang, F.; Chen, C.; Wang, S.; Yang, C. Contrastive intra-and inter-modality generation for enhancing incomplete multimedia recommendation. In Proceedings of the ACM MM, Ottawa, ON, Canada, 29 October–3 November 2023.
7. Zhang, X.; Weng, H.; Wei, Y.; Wang, D.; Chen, J.; Liang, T.; Yin, Y. Multivariate Hawkes Spatio-Temporal Point Process with attention for point of interest recommendation. *Neurocomputing* **2024**, *619* , 129161. [CrossRef]
8. Chen, X.; Lei, C.; Liu, D.; Wang, G.; Tang, H.; Zha, Z.J.; Li, H. E-commerce storytelling recommendation using attentional domain-transfer network and adversarial pre-training. *IEEE Trans. Multimed.* **2021**, *24*, 506–518. [CrossRef]
9. Baek, J.W.; Chung, K.Y. Multimedia recommendation using Word2Vec-based social relationship mining. *Multimed. Tools Appl.* **2021**, *80*, 34499–34515. [CrossRef]
10. Deldjoo, Y. Enhancing video recommendation using multimedia content. In *Special Topics in Information Technology*; Springer: Cham, Switzerland, 2020; pp. 77–89.
11. Xia, F.; Asabere, N.Y.; Ahmed, A.M.; Li, J.; Kong, X. Mobile multimedia recommendation in smart communities: A survey. *IEEE Access* **2013**, *1*, 606–624.
12. Lin, J.; Li, Q.; Xie, G.; Guan, Z.; Jiang, Y.; Xu, T.; Zhang, Z.; Zhao, P. Mitigating Sample Selection Bias with Robust Domain Adaption in Multimedia Recommendation. In Proceedings of the ACM MM, Melbourne, VIC, Australia, 28 October–1 November 2024.
13. He, R.; McAuley, J. VBPR: Visual bayesian personalized ranking from implicit feedback. In Proceedings of the AAAI, Phoenix, AZ, USA, 12–17 February 2016.
14. Liu, Q.; Wu, S.; Wang, L. Deepstyle: Learning user preferences for visual recommendation. In Proceedings of the SIGIR, Tokyo, Japan, 7–11 August 2017.
15. Wei, W.; Huang, C.; Xia, L.; Zhang, C. Multi-modal self-supervised learning for recommendation. In Proceedings of the WWW, Austin, TX, USA, 30 April–4 May 2023.
16. Wei, Y.; Wang, X.; Nie, L.; He, X.; Hong, R.; Chua, T.S. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In Proceedings of the ACM MM, Nice, France, 21–25 October 2019.
17. Wei, Y.; Wang, X.; Nie, L.; He, X.; Chua, T.S. Graph-refined convolutional network for multimedia recommendation with implicit feedback. In Proceedings of the ACM MM, Seattle, WA, USA, 12–16 October 2020.
18. Wang, Q.; Wei, Y.; Yin, J.; Wu, J.; Song, X.; Nie, L. Dualgnn: Dual graph neural network for multimedia recommendation. *IEEE Trans. Multimed.* **2021**, *25*, 1074–1084. [CrossRef]
19. Li, S.; Guo, D.; Liu, K.; Hong, R.; Xue, F. Multimodal Counterfactual Learning Network for Multimedia-based Recommendation. In Proceedings of the SIGIR, Taipei, Taiwan, 23–27 July 2023.
20. Yang, Y.; Zhou, S.; Weng, H.; Wang, D.; Zhang, X.; Yu, D.; Deng, S. Siamese learning based on graph differential equation for Next-POI recommendation. *Appl. Soft Comput.* **2024**, *150*, 111086. [CrossRef]
21. Volpi, R.; Namkoong, H.; Sener, O.; Duchi, J.C.; Murino, V.; Savarese, S. Generalizing to unseen domains via adversarial data augmentation. In Proceedings of the NeurIPS, Montreal, QC, Canada, 3–8 December 2018.
22. Li, H.; Zheng, C.; Wu, P. StableDR: Stabilized Doubly Robust Learning for Recommendation on Data Missing Not at Random. In Proceedings of the ICLR, Kigali, Rwanda, 1–5 May 2023.
23. Li, H.; Zheng, C.; Xiao, Y.; Wu, P.; Geng, Z.; Chen, X.; Cui, P. Debiased collaborative filtering with kernel-based causal balancing. In Proceedings of the ICLR, Vienna, Austria, 7–11 May 2024.
24. Wang, J.; Li, H.; Zhang, C.; Liang, D.; Yu, E.; Ou, W.; Wang, W. Counterclr: Counterfactual contrastive learning with non-random missing data in recommendation. In Proceedings of the ICDM, Shanghai, China, 1–4 December 2023.
25. Li, H.; Zheng, C.; Ding, S.; Feng, F.; He, X.; Geng, Z.; Wu, P. Be Aware of the Neighborhood Effect: Modeling Selection Bias under Interference for Recommendation. In Proceedings of the ICLR, Vienna, Austria, 7–11 May 2024.

26. Muandet, K.; Balduzzi, D.; Schölkopf, B. Domain generalization via invariant feature representation. In Proceedings of the ICML, Atlanta, GA, USA, 17–19 June 2013.
27. Arjovsky, M.; Bottou, L.; Gulrajani, I.; Lopez-Paz, D. Invariant risk minimization. *arXiv* **2019**, arXiv:1907.02893.
28. Ahuja, K.; Shanmugam, K.; Varshney, K.; Dhurandhar, A. Invariant risk minimization games. In Proceedings of the ICML, Virtual, 13–18 July 2020.
29. Krueger, D.; Caballero, E.; Jacobsen, J.H.; Zhang, A.; Binas, J.; Zhang, D.; Le Priol, R.; Courville, A. Out-of-distribution generalization via risk extrapolation (rex). In Proceedings of the ICML, Virtual, 18–24 July 2021.
30. Du, X.; Wu, Z.; Feng, F.; He, X.; Tang, J. Invariant representation learning for multimedia recommendation. In Proceedings of the ACM MM, Lisboa, Portugal, 10–14 October 2022.
31. Huang, S.; Li, H.; Li, Q.; Zheng, C.; Liu, L. Pareto invariant representation learning for multimedia recommendation. In Proceedings of the ACM MM, Ottawa, ON, Canada, 29 October–3 November 2023.
32. Bai, H.; Wu, L.; Hou, M.; Cai, M.; He, Z.; Zhou, Y.; Hong, R.; Wang, M. Multimodality Invariant Learning for Multimedia-Based New Item Recommendation. In Proceedings of the SIGIR, Washington DC, USA, 14–18 July 2024.
33. Lv, Z.; Zhang, W.; Zhang, S.; Kuang, K.; Wang, F.; Wang, Y.; Chen, Z.; Shen, T.; Yang, H.; Ooi, B.C.; et al. DUET: A Tuning-Free Device-Cloud Collaborative Parameters Generation Framework for Efficient Device Model Generalization. In Proceedings of the WWW, Austin, USA, 30 April – 4 May 2023.
34. He, X.; Zhang, H.; Kan, M.Y.; Chua, T.S. Fast matrix factorization for online recommendation with implicit feedback. In Proceedings of the SIGIR, Pisa, Italy, 17–21 July 2016.
35. He, X.; Liao, L.; Zhang, H.; Nie, L.; Hu, X.; Chua, T.S. Neural collaborative filtering. In Proceedings of the WWW, Perth, Australia, 3–7 April 2017.
36. Wang, X.; He, X.; Wang, M.; Feng, F.; Chua, T.S. Neural graph collaborative filtering. In Proceedings of the SIGIR, Paris, France, 21–25 July 2019.
37. He, X.; Deng, K.; Wang, X.; Li, Y.; Zhang, Y.; Wang, M. Lightgcn: Simplifying and powering graph convolution network for recommendation. In Proceedings of the SIGIR, Virtual, 25 July 2020.
38. Mao, K.; Zhu, J.; Xiao, X.; Lu, B.; Wang, Z.; He, X. UltraGCN: Ultra simplification of graph convolutional networks for recommendation. In Proceedings of the CIKM, Virtual, 1–5 November 2021.
39. Zhang, F.; Yuan, N.J.; Lian, D.; Xie, X.; Ma, W.Y. Collaborative knowledge base embedding for recommender systems. In Proceedings of the KDD, San Francisco, CA, USA, 13–17 August 2016.
40. Zhang, J.; Liu, G.; Liu, Q.; Wu, S.; Wang, L. Modality-Balanced Learning for Multimedia Recommendation. In Proceedings of the ACM MM, Melbourne, VIC, Australia, 28 October–1 November 2024.
41. Kang, W.C.; Fang, C.; Wang, Z.; McAuley, J. Visually-aware fashion recommendation and design with generative image models. In Proceedings of the ICDM, New Orleans, LA, USA, 18–21 November 2017.
42. Wang, D.; Du, R.; Yang, Q.; Yu, D.; Wan, F.; Gong, X.; Xu, G.; Deng, S. Category-aware self-supervised graph neural network for session-based recommendation. *World Wide Web* **2024**, *27*, 61. [CrossRef]
43. Li, S.; Xue, F.; Liu, K.; Guo, D.; Hong, R. Multimodal graph causal embedding for multimedia-based recommendation. *IEEE Trans. Knowl. Data Eng.* **2024**, *36*, 8842–8858. [CrossRef]
44. Chen, X.; Chen, H.; Xu, H.; Zhang, Y.; Cao, Y.; Qin, Z.; Zha, H. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In Proceedings of the SIGIR, Paris, France, 21–25 July 2019.
45. Liu, S.; Chen, Z.; Liu, H.; Hu, X. User-video co-attention network for personalized micro-video recommendation. In Proceedings of the WWW, Palma, Spain, 1–3 July 2019.
46. Liu, F.; Cheng, Z.; Sun, C.; Wang, Y.; Nie, L.; Kankanhalli, M. User diverse preference modeling by multimodal attentive metric learning. In Proceedings of the ACM MM, Nice, France, 21 October 2019.
47. Lu, C.; Wu, Y.; Hernández-Lobato, J.M.; Schölkopf, B. Invariant causal representation learning for out-of-distribution generalization. In Proceedings of the ICLR, Vienna, Austria, 4 May 2021.
48. Ahuja, K.; Caballero, E.; Zhang, D.; Gagnon-Audet, J.C.; Bengio, Y.; Mitliagkas, I.; Rish, I. Invariance principle meets information bottleneck for out-of-distribution generalization. In Proceedings of the NeurIPS, Online, 6–14 December 2021.
49. Li, B.; Shen, Y.; Wang, Y.; Zhu, W.; Li, D.; Keutzer, K.; Zhao, H. Invariant information bottleneck for domain generalization. In Proceedings of the AAAI, Online, 22 February–1 March 2022.
50. Yu, R.; Zhu, H.; Li, K.; Hong, L.; Zhang, R.; Ye, N.; Huang, S.L.; He, X. Regularization Penalty Optimization for Addressing Data Quality Variance in OoD Algorithms. In Proceedings of the AAAI, Online, 22 February–1 March 2022.
51. Zhou, X.; Lin, Y.; Zhang, W.; Zhang, T. Sparse invariant risk minimization. In Proceedings of the ICML, Baltimore, MD, USA, 17–23 July 2022.
52. Creager, E.; Jacobsen, J.H.; Zemel, R. Environment inference for invariant learning. In Proceedings of the ICML, Virtual, 18–24 July 2021.

53. Liu, J.; Hu, Z.; Cui, P.; Li, B.; Shen, Z. Heterogeneous risk minimization. In Proceedings of the ICML, Virtual, 18–24 July 2021.
54. Chen, J.; Dong, H.; Wang, X.; Feng, F.; Wang, M.; He, X. Bias and Debias in Recommender System: A Survey and Future Directions. *ACM Trans. Inf. Syst.* **2023**, *41*, 1–39. [CrossRef]
55. Wang, F.; Chen, C.; Liu, W.; Fan, T.; Liao, X.; Tan, Y.; Qi, L.; Zheng, X. CE-RCFR: Robust Counterfactual Regression for Consensus-Enabled Treatment Effect Estimation. In Proceedings of the KDD, Barcelona, Spain, 28 August 25–29 2024.
56. Wu, P.; Li, H.; Deng, Y.; Hu, W.; Dai, Q.; Dong, Z.; Sun, J.; Zhang, R.; Zhou, X.H. On the Opportunity of Causal Learning in Recommendation Systems: Foundation, Estimation, Prediction and Challenges. In Proceedings of the IJCAI, Vienna, Austria, 23–29 July 2022.
57. Wang, W.; Zhang, Y.; Li, H.; Wu, P.; Feng, F.; He, X. Causal Recommendation: Progresses and Future Directions. In Proceedings of the SIGIR, Taipei, Taiwan, 23–27 July 2023.
58. Schnabel, T.; Swaminathan, A.; Singh, A.; Chandak, N.; Joachims, T. Recommendations as treatments: Debiasing learning and evaluation. In Proceedings of the ICML, New York, NY, USA, 19–24 June 2016.
59. Wang, X.; Zhang, R.; Sun, Y.; Qi, J. Doubly robust joint learning for recommendation on data missing not at random. In Proceedings of the ICML, Long Beach, CA, USA, 9–15 June 2019.
60. Li, H.; Lyu, Y.; Zheng, C.; Wu, P. TDR-CL: Targeted Doubly Robust Collaborative Learning for Debiased Recommendations. In Proceedings of the ICLR, Kigali, Rwanda, 1–5 May 2023.
61. Wang, H.; Chang, T.W.; Liu, T.; Huang, J.; Chen, Z.; Yu, C.; Li, R.; Chu, W. Escm2: Entire space counterfactual multi-task model for post-click conversion rate estimation. In Proceedings of the SIGIR, Madrid, Spain, 11–15 July 2022.
62. Li, H.; Zheng, C.; Wang, S.; Wu, K.; Wang, E.; Wu, P.; Geng, Z.; Chen, X.; Zhou, X.H. Relaxing the Accurate Imputation Assumption in Doubly Robust Learning for Debiased Collaborative Filtering. In Proceedings of the ICML, Vienna, Austria, 21–27 July 2024.
63. Li, H.; Xiao, Y.; Zheng, C.; Wu, P.; Cui, P. Propensity Matters: Measuring and Enhancing Balancing for Recommendation. In Proceedings of the ICML, Honolulu, HI, USA, 23–29 July 2023.
64. Li, H.; Xiao, Y.; Zheng, C.; Wu, P. Balancing unobserved confounding with a few unbiased ratings in debiased recommendations. In Proceedings of the WWW, Austin, TX, USA, 30 April–4 May 2023.
65. Li, H.; Wu, K.; Zheng, C.; Xiao, Y.; Wang, H.; Geng, Z.; Feng, F.; He, X.; Wu, P. Removing hidden confounding in recommendation: A unified multi-task learning approach. In Proceedings of the NeurIPS, New Orleans, LA, USA, 10–16 December 2023.
66. Cheng, P.; Hao, W.; Dai, S.; Liu, J.; Gan, Z.; Carin, L. Club: A contrastive log-ratio upper bound of mutual information. In Proceedings of the ICML, Virtual, 13–18 July 2020.
67. Liu, J.; Hu, Z.; Cui, P.; Li, B.; Shen, Z. Kernelized heterogeneous risk minimization. In Proceedings of the NeurIPS, Online, 6–14 December 2021.
68. Wei, Y.; Wang, X.; He, X.; Nie, L.; Rui, Y.; Chua, T.S. Hierarchical user intent graph network for multimedia recommendation. *IEEE Trans. Multimed.* **2021**, *24*, 2701–2712. [CrossRef]
69. Hershey, S.; Chaudhuri, S.; Ellis, D.P.; Gemmeke, J.F.; Jansen, A.; Moore, R.C.; Plakal, M.; Platt, D.; Saurous, R.A.; Seybold, B.; et al. CNN architectures for large-scale audio classification. In Proceedings of the ICASSP, New Orleans, LA, USA, 5–9 March 2017.
70. Arora, S.; Liang, Y.; Ma, T. A simple but tough-to-beat baseline for sentence embeddings. In Proceedings of the ICLR, Toulon, France, 24–26 April 2017.
71. Kingma, D.P. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
72. Barkan, O.; Koenigstein, N.; Yogev, E.; Katz, O. CB2CF: A neural multiview content-to-collaborative filtering model for completely cold item recommendations. In Proceedings of the RecSys, Copenhagen, Denmark, 16–20 September 2019.
73. Geng, X.; Zhang, H.; Bian, J.; Chua, T.S. Learning image and user features for recommendation in social networks. In Proceedings of the ICCV, Santiago, Chile, 7–13 December 2015.
74. Ma, J.; Cui, P.; Kuang, K.; Wang, X.; Zhu, W. Disentangled graph convolutional networks. In Proceedings of the ICML, Long Beach, CA, USA, 9–15 June 2019.
75. Ma, J.; Zhou, C.; Cui, P.; Yang, H.; Zhu, W. Learning disentangled representations for recommendation. In Proceedings of the NeurIPS, Vancouver, BC, Canada, 8–14 December 2019.