Pareto Invariant Representation Learning for Multimedia Recommendation

Shanshan Huang*
Chongqing University
Chongqing, China
shanshanhuang@cqu.edu.cn

Haoxuan Li*
Peking University
Beijing, China
hxli@stu.pku.edu.cn

Qingsong Li
Chongqing University
Chongqing, China
liqingsong@stu.cqu.edu.cn

Chunyuan Zheng
University of California, San Diego
San Diego, USA
czheng@ucsd.edu

Li Liu†
Chongqing University
Chongqing, China
dcsliuli@cqu.edu.cn

ABSTRACT

Multimedia recommendation involves personalized ranking tasks, where multimedia content is usually represented using a generic encoder. However, these generic representations introduce spurious correlations that fail to reveal users' true preferences. Existing works attempt to alleviate this problem by learning invariant representations, but overlook the balance between independent and identically distributed (IID) and out-of-distribution (OOD) generalization. In this paper, we propose a framework called Pareto Invariant Representation Learning (PaInvRL) to mitigate the impact of spurious correlations from an IID-OOD multi-objective optimization perspective, by learning invariant representations (intrinsic factors that attract user attention) and variant representations (other factors) simultaneously. Specifically, PaInvRL includes three iteratively executed modules: (i) heterogeneous identification module, which identifies the heterogeneous environments to reflect distributional shifts for user-item interactions; (ii) invariant mask generation module, which learns invariant masks based on the Pareto-optimal solutions that minimize the adaptive weighted Invariant Risk Minimization (IRM) and Empirical Risk (ERM) losses; (iii) convert module, which generates both variant representations and item-invariant representations for training a multi-modal recommendation model that mitigates spurious correlations and balances the generalization performance within and cross the environmental distributions. We compare the proposed PaInvRL with state-of-the-art recommendation models on three public multimedia recommendation datasets (Movielens, Tiktok, and Kwai), and the experimental results validate the effectiveness of PaInvRL for both within- and cross-environmental learning.

CCS CONCEPTS

• **Information systems** → **Multimedia information systems**.

KEYWORDS

Multimedia Recommendation, Multimedia Representation Learning, Invariant Learning, Multi-objective Optimization

ACM Reference Format:
Shanshan Huang, Haoxuan Li, Qingsong Li, Chunyuan Zheng, and Li Liu. 2023. Pareto Invariant Representation Learning for Multimedia Recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3612591

1 INTRODUCTION

With the rapid development of the internet, multimedia recommendation systems have become indispensable tools to help users find their interesting items, and have been widely used in many online applications, such as e-commerce platforms, social media, and instant video platforms. For multimedia recommendation, item content includes multiple modalities, including visual, acoustic, and textual representations. These multi-modal data may reflect user preferences at the fine-grained modality level. The core of multimedia recommendation is to use the historical interactions between users and items and the auxiliary multi-modal item representations to improve recommendation performance.

Collaborative filtering (CF) serves as the foundation of personalized recommendation systems, which leverages historical user-item interactions to learn user and item representations and provides recommendations based on these representations [56, 65]. Extending to multimedia tasks, previous studies, e.g., VBPR [15], DeepStyle [42], incorporate multi-modal contents as side information in addition to id embeddings of items to learn the user preference. However, these methods have limited expressiveness as they neglect high-order user-item semantic relations [93]. Inspired by the recent advances in graph neural networks, recent studies [12, 27, 43, 66, 72, 81] take advantage of powerful graph convolution networks (GCNs) to model user-item relationships as bipartite graphs to improve the performance of CF-based recommendation systems. Further, many researchers have also attempted to apply GCNs to incorporate modality information into the message passing for inferring user and item representations, such as MMGCN [79], GRCN [78], LATTICE [87], MICRO [87] and HCGCN [49].

*Both authors contributed equally to this research.
†Corresponding author.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612591

Despite achieving promising performance, previous approaches often use encoder architectures designed for general content understanding tasks [20] (including image classification, object recognition, image colorization, and text classification, etc.), e.g., pretrained VGG19 [63], ResNet50 [14], VILBERT [44], and sentence-transformer [53], to encode multimedia content. The use of these generic encoders may introduce spurious correlations (i.e., some learned representations may affect the recommendation results, but are irrelevant to user's true preferences from a causal perspective), making it difficult for recommendation models to capture user's true preferences and provide accurate recommendations. To alleviate this issue, existing studies mainly rely on preference-aware representations [39, 61, 84], which were extracted with specifically designed multimedia models for specific recommendation tasks. Therefore, the existing methods face the limitation of domain-specific analysis and design, and thus can hardly be generalized.

To address this issue, a recent research work, named InvRL [10], introduced invariant risk minimization (IRM) to multimedia recommendation, by learning invariant item representations to alleviate the impact of spurious correlations. Although experimentally promising, it is widely known that there is a conflict between independent and identical distributed (IID) tasks (where the source and target environments are similar) and out-of-distribution (OOD) tasks (where there is a significant difference between the source and target environments), which may lead to significant degradation of model performance on IID tasks. We verify empirically that the superiority of InvRL is only guaranteed in OOD tasks, whereas empirical risk minimization (ERM) typically outperforms in IID tasks, which motivates us to balance this conflict between IID and OOD. Specifically, in this paper, we formalize the IID-OOD task as a multi-objective optimization problem [89] and adaptively weight the ERM and IRM losses via a gradient approach to obtain the Pareto optimal solution. We theoretically prove that our solution cannot be dominated by other solutions, i.e., there does not exist any solution that performs better compared to our solution on both tasks at the same time. Specifically, we divide the raw multimedia representations into two parts: variant and invariant representations, where the variant representations account for spurious correlations while the invariant representations reflect the user's true preferences.

The main contributions of this paper are summarized as follows.

- We first formalize the IID-OOD task as a multi-objective optimization problem and adaptively weight the ERM and IRM losses using a gradient-based representation learning approach to obtain the Pareto optimal solution, i.e., there does not exist any solution that outperforms compared to our solution on both IID and OOD tasks.
- We propose a new multimedia recommendation framework, called PaInvRL, that aims to obtain a Pareto solution between IID and OOD tasks via a gradient-based updating method, where the gradient is shown to be either 0 when there are no other solution in its neighborhood can have lower values in both ERM and IRM losses, or the gradient givens a descent direction that improves both IID and OOD generalization by reducing ERM and IRM losses simultaneously.
- We instantiate the framework over UltraGCN and conduct extensive experiments over three public datasets, verifying the rationality and effectiveness of PaInvRL.

2 RELATED WORK

2.1 Multimedia Recommendation

The multi-modal recommendation system aims to learn informative representations of users and items by leveraging multi-modal representations. Many efforts [45–47, 77] have been devoted to enhancing recommendation systems by incorporating multimedia content. VBPR [15] is the first model that considers introducing visual representations into the recommendation system by concatenating visual embeddings with id embeddings as the item representations. DVBPR [23] attempts to jointly train the image representations as well as the parameters in a recommendation model. In recent years, graph neural networks have been demonstrated as powerful solutions for multimedia recommendation by capturing high-order dependent structures among users and items. For example, MMGCN [79] constructs a modal-specific graph and conducts graph convolution operations, to capture the modal-specific user preference and distills the item representations simultaneously. MGAT [67] based on the MMGCN framework utilizes the original GCN to do aggregation and the same way to combine the aggregated result. To manage the information transmission for each modality, it added a new gated attention mechanism. DualGNN [70] also introduces a model preference learning module and draws the user's attention to various modalities. InvRL [10] introduces IRM to learn invariant item representations, which reduces the impact of spurious correlations and improves the recommendation performance of multi-modal recommendation models. DRAGON [92] learns dual representations of users and items by constructing homogeneous graphs to enhance the relationship between the two parties, enabling multi-modal recommendations. Different from these works, for robust multi-modal user preference learning, this paper proposes a new framework for invariant representation learning, which first views the IID-OOD task in the multi-modal recommendation as a multi-objective optimization problem, and then adaptively weights the IRM and ERM losses and uses gradient-based methods to seek Pareto optimal solutions for learning invariant representations.

2.2 Invariant Representation Learning

Invariant representation learning aims to learn the essential representations of data, and improve the generalization ability and robustness of models. Recently, some studies have been conducted, among which IRM [2] is an earlier method proposed based on invariant principle [52], which aims to learn representations with invariance in different environments. Several works [1, 25, 36, 94] further develop several variants of IRM by introducing game theory, regret minimization, variance penalization, etc., and [82, 83] try to learn invariant representations by coupled adversarial neural networks. Other approaches [41, 88] attempt to learn invariant representations without providing explicit environment indicators. Liu et al. [40] proposed the HRM to achieve joint learning of latent heterogeneity and invariant relationships in the data, resulting in stable predictions despite distributional shifts. Furthermore, they extended HRM to the representation level using kernel tricks [41].

An alternative class of methods for learning invariant representations are causality-based approaches with debiased loss [4–6, 30, 32, 34, 37, 38, 71, 74, 75, 80, 90], such as outcome regression methods [13, 21, 48, 64, 86], propensity-based weighting methods [17, 19, 31, 31, 50, 57, 85], doubly robust learning methods [8, 11,

24, 29, 33, 54, 55, 69, 73], multiple robust learning method [28], and representation learning methods [7, 22, 26, 58, 60, 62, 68, 91]. However, these previous approaches failed in obtaining Pareto solutions between IID and OOD tasks [35, 59]. In this paper, we aim to learn representations corresponding to the Pareto solutions between within- and cross-environmental learning to improve the model's generalization performance in multimedia recommendations.

3 METHODOLOGY

3.1 Preliminaries

Considering a multimedia recommendation system, we denote the set of users and items as $\mathcal{U}$ and $\mathcal{I}$, respectively. For each user-item pairs $(u, i) \in \mathcal{U} \times \mathcal{I}$, denote $r_{u,i} = 1$ if user u make a positive feedback on item i, and $r_{u,i} = 0$ otherwise. In addition to user-item interactions, we also have access to multi-modal representations that provide content information about items. We represent the modality representation of item i as $\mathbf{f}_{r,i} \in \mathbb{R}^{d_r}$, where d_r is the dimension of the modality representation, $r \in R = \{V, T, A\}$ denotes the modality, and R is the set of all modalities. In this paper, R includes visual (V), textual (T), and acoustic (A) modalities, let $\mathbf{f}_i = concat(\mathbf{f}_{V,i}, \mathbf{f}_{T,i}, \mathbf{f}_{A,i}) \in \mathbb{R}^d$, where $d = d_V + d_T + d_A$ and $concat(\cdot)$ indicates the concatenation operation. The multi-modal recommendation aims to learn a model $\Gamma(u, i, \mathbf{f}_i|\Theta)$ parameterized by Θ to predict users' true preferences, which can be formalized as

$$\arg\min_{\Theta} \mathcal{L}(\Gamma(u, i, \mathbf{f}_i|\Theta)|\mathcal{R}^{tr}), \tag{1}$$

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 $\mathcal{R}^- = \{(u, i) : r_{u,i} = 0\}$.

3.2 Model Overview

We now present the proposed PaInvRL model, the architecture of which is illustrated in Figure 1. There are four components in the framework: (1) the generic feature extraction network that is used to extract multi-modal representations, including visual, acoustic, and visual representations; (2) the heterogeneous identification module (HIM) that is designed to partition the input historical user-item dataset interaction into multiple heterogeneous environments for invariant representation learning, each reflecting a spurious correlation in user-item interactions; (3) the invariant mask generation module and (4) the convert module work together to select representations that have stable and invariant relationships across environments. Specifically, the generic feature extraction module adopts a pre-trained model and is not the focus of this paper, and we therefore provide only a brief introduction to this module in Section 4. The HIM and the invariant mask generation module promote each other: on one hand, the invariant mask generation module uses the heterogeneous environment identified by HIM to learn the invariant mask $\mathbf{m}$, which leads to the corresponding invariant representations $\mathbf{\Phi}_i$ and variant representations $\mathbf{\Psi}_i$ using the learned invariant mask; on the other hand, the variant representations are utilized to enhance the training of HIM. The convert module divides the raw multimedia representations into invariant representations and variant representations. Finally, we use the learned invariant representations to learn the final multi-modal recommendation model with both promising IID and OOD generalization.

Different from InvRL [10] that utilizes the invariant mask generation module to generate invariant masks used to generate the corresponding invariant representations with superior performance under the OOD task, we propose to generate the invariant mask corresponding to a Pareto solution between IID and OOD tasks via a gradient-based updating method. The proposed invariant mask update gradient is either 0 when no neighboring solution can offer lower values in both ERM and IRM losses, or it provides a descent direction enhancing IID and OOD generalization through simultaneous reduction of ERM and IRM losses.

3.3 Heterogeneous Environment Identification

Heterogeneous identification module (HIM), which takes in the historical user-item interactions and outputs an environment set $\mathcal{E}$ for invariant mask generation [10]. This module comprises two phases: an environment learning phase and a user-item interaction partitioning phase. Specifically, in the environment learning phase, we learn different environments $e \in \mathcal{E}$ by training a recommendation model $\Gamma_{(e)}(u, i, \mathbf{\Psi}_i|\Theta_e)$ for each environment $e \in \mathcal{E}$, where Θ_e denotes the parameters of the recommendation model $\Gamma_{(e)}$, and can be optimized by

$$\arg\min_{\Theta_e} \mathcal{L}(\Gamma_{(e)}(u, i, \mathbf{\Psi}_i|\Theta_e)|\mathcal{R}_e^{tr}), \tag{2}$$

where the variant representations $\mathbf{\Psi}_i$ are obtained by initializing the invariant mask $\mathbf{m}$ by 0.5. We employ UltraGCN [47] as the recommendation model and drive the representations through a graph-based loss function to encode the user-item graph by

$$\mathcal{L} = \mathcal{L}_O + \eta\mathcal{L}_U + \kappa\mathcal{L}_I, \tag{3}$$

where $\mathcal{L}_O$ is used as the main optimization objective of the recommendation model $\Gamma(u, i, \mathbf{\Psi}_i)$, and $\mathcal{L}_U$ and $\mathcal{L}_I$ are used as constraints to learn better user-item graphs, and item-item graphs, respectively. η and κ are used as weights of $\mathcal{L}_U$ and $\mathcal{L}_I$ to adjust the relative importance of user-item and item-item relationships. Following [47], we choose the binary cross entropy loss to calculate $\mathcal{L}_O$ by

$$\mathcal{L}_O = -\sum_{(u,i)\in\mathcal{R}^+} \log(\sigma(\Gamma(u, i, \mathbf{\Psi}_i))) - \sum_{(u,j)\in\mathcal{R}^-} \log(\sigma(-\Gamma(u, j, \mathbf{\Psi}_j))), \tag{4}$$

where σ is the sigmoid function. $\mathcal{L}_U$ is derived by negative log-likelihood, as

$$\begin{aligned}\mathcal{L}_U = &-\sum_{(u,i)\in\mathcal{R}^+} v_{u,i} \log(\sigma(\Gamma(u, i, \mathbf{\Psi}_i))) \\ &-\sum_{(u,j)\in\mathcal{R}^-} v_{u,j} \log(\sigma(-\Gamma(u, j, \mathbf{\Psi}_j))),\end{aligned} \tag{5}$$

where $v_{u,i}$ and $v_{u,j}$ can be derived from the user-item graph by

$$v_{u,i} = \frac{1}{d_u}\sqrt{\frac{d_u + 1}{d_i + 1}}, \tag{6}$$

where d_u and d_i denote the degrees for the corresponding nodes. The term $\mathcal{L}_I$ induced from item-item graph can be calculated by

$$\mathcal{L}_I = -\sum_{(u,i)\in\mathcal{R}^+} \sum_{j\in S(i)} s_{i,j} \log(\sigma(\Gamma(u, j, \mathbf{\Psi}_j))), \tag{7}$$

where $S(i)$ include K weighted positive sample pairs (u, k) corresponding to each positive sample pair (u, i), which are selected

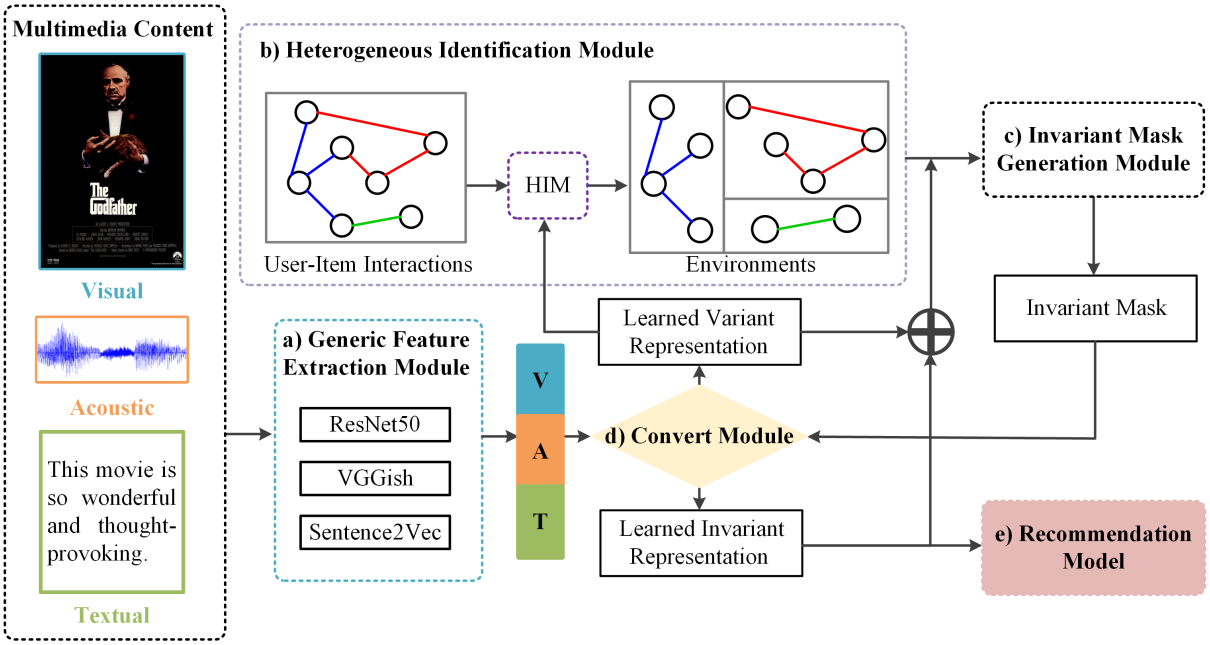


Figure 1: The framework of PaInvRL, where V, A, and T denote the extracted visual representations, acoustic representations, and textual representations, respectively. The symbol ⊕ represents the operation of weighted summation.

from the weighted adjacency matrix of the item-item co-occurrence graph G according to the similarity score $s_{i,j}$. We calculate $s_{i,j}$ by

$$s_{i,j} = \frac{G_{i,j}}{g_i - G_{i,i}} \sqrt{\frac{g_i}{g_j}}, \quad g_i = \sum_{k=1}^{K} G_{i,k}, \tag{8}$$

where $G_{i,j}$ represent the number of co-occurrences of item i and item j, and g_i and g_j denote the degrees of item i and item j in G.

In the user-item interaction partitioning phase, we use the trained recommendation model to partition the user-item interaction into the corresponding environments by

$$e\,(u,i) = \arg\max_{e \in \mathcal{E}} \Gamma_{(e)}(u, i, \Psi_i | \Theta_e). \tag{9}$$

The obtained results $\{\mathcal{R}_{(e)} | e \in \mathcal{E}\}$ are used in the training of the following invariant mask generation module.

3.4 Invariant Mask Generation

Here, we introduce our invariant mask generation module, which takes multiple environments training data $\{\mathcal{R}_{(e)} | e \in \mathcal{E}\}$ as input, and outputs the corresponding invariant mask $\mathbf{m}$. As mentioned above, we learn the invariant mask generation module together with the convert module to generate invariant and variant representations across environments. Following InvRL [10], we approximate $\mathbf{m} = (m_1, m_2, m_3, ..., m_d)^T$ using clipped Gaussian random variable parameterized by $\mu = (\mu_1, \mu_2, \mu_3, ..., \mu_d)^T$ as

$$\mu_i = \max\{0, \min\{1, m_i + \epsilon\}\}, \tag{10}$$

where ϵ is sampled from $\mathcal{N}(0, \sigma^2)$. With this approximation, the objective function of the invariant mask generation module can be written as

$$\begin{aligned} \mathcal{L}_{\text{mask}} &= w_{ERM} \mathbb{E}_{e \in \mathcal{E}} \mathcal{L}^e + w_{IRM} \left\| \text{Var}_{e \in \mathcal{E}} \left(\nabla_{\Theta^{\text{mask}}} \mathcal{L}^e \right) \odot \mu \right\|^2 + \frac{\lambda}{2} \|\mathbf{m}\|^2 \\ &= w_{ERM} \mathcal{L}_{ERM} + w_{IRM} \mathcal{L}_{IRM} + \frac{\lambda}{2} \|\mathbf{m}\|^2, \end{aligned} \tag{11}$$

where λ represents the weight of the regularization term, w_{ERM} and w_{IRM} represent the weights of $\mathcal{L}_{ERM}$ and $\mathcal{L}_{IRM}$, respectively. The first term is the ordinary recommended loss, which is the average loss within environment $\mathcal{E}$ and can be viewed as the ERM loss, i.e.,

$$\mathcal{L}_{ERM} = \mathcal{L}(\Gamma^{mask}(u, i, \mu \odot \mathbf{h}_i | \Theta^{mask}) | \mathcal{R}_e^{tr}), \tag{12}$$

where $\mathbf{h}_i$ symbolizes the weighted representations, Θ^{mask} denotes the parameters of Γ^{mask}, $\odot$ means dot product operation. The second term is the cross-environment constraint, which is the IRM loss. The last term is the regularization term.

To learns invariant masks based on the Pareto-optimal solution, architecturally, instead of just using invariant representations [10], we incorporate an attention mechanism, which empowers us to dynamically assign weights to both the invariant representations $\mathbf{\Phi}_i$ and variant representations $\mathbf{\Psi}_i$. This attention mechanism allows our model to focus on the most relevant representations from both invariant and variant representations. Formally, the weighted representations $\mathbf{h}_i$ can be expressed as

$$\mathbf{h}_i = \alpha_i^{\Phi} \cdot \mathbf{\Phi}_i + \alpha_i^{\Psi} \cdot \mathbf{\Psi}_i, \tag{13}$$

where α_i^{Φ} and α_i^{Ψ} are implemented using multi-layer perceptron (MLP). Specifically, we first concatenate the collaborative embedding and content representations of users and items, and then use two MLPs, respectively, to obtain the weights of the variant and

invariant representations, which can be formalized as

$$\begin{aligned}\alpha_i^{\Phi} &= \mathrm{MLP}_1([\mathbf{p}_u^{(t)}, \mathbf{p}_u^{(f)}, \mathbf{t}_i, \mathbf{f}_i]),\\ \alpha_i^{\Psi} &= \mathrm{MLP}_2([\mathbf{p}_u^{(t)}, \mathbf{p}_u^{(f)}, \mathbf{t}_i, \mathbf{f}_i]),\end{aligned} \tag{14}$$

where $\mathbf{t}_i$ and $\mathbf{f}_i$ denote the collaborative and raw multimedia representations of item i, and $\mathbf{p}_u^{(t)}$ and $\mathbf{p}_u^{(f)}$ denote the corresponding user representations. In such case, the recommendation model $\Gamma(u, i, \mathbf{h}_i)$ can be formalized as

$$\Gamma(u, i, \mathbf{h}_i) = \Gamma(\mathbf{p}_u^{(t)}, \mathbf{p}_u^{(f)}, \mathbf{t}_i, \mathbf{h}_i) = \langle \mathbf{p}_u^{(t)}, \mathbf{t}_i \rangle + \langle \mathbf{p}_u^{(f)}, \mathbf{W} \cdot \mathbf{h}_i \rangle, \tag{15}$$

where $\mathbf{W}$ refers to a projection matrix that is used to compress the dimension of the raw multimedia representations. To obtain the Pareto optimal invariant mask, we require to solve the minimization problem of the loss function $\mathcal{L}_{mask}$ via an adaptive manner, where

$$\begin{aligned}\min_{w_{ERM}, w_{IRM}} \ & \|w_{ERM}\nabla_{\mathbf{m}}\mathcal{L}_{ERM} + w_{IRM}\nabla_{\mathbf{m}}\mathcal{L}_{IRM}\|_2^2,\\ \text{s.t.} \quad & w_{ERM} + w_{IRM} = 1, w_{ERM} \geq 0, w_{IRM} \geq 0,\end{aligned} \tag{16}$$

with an analytical solution

$$w_{ERM}^* = \frac{(\nabla_{\mathbf{m}}\mathcal{L}_{IRM}(\mathbf{m}) - \nabla_{\mathbf{m}}\mathcal{L}_{ERM}(\mathbf{m}))^\top \nabla_{\mathbf{m}}\mathcal{L}_{IRM}(\mathbf{m})}{\|\nabla_{\mathbf{m}}\mathcal{L}_{ERM}(\mathbf{m}) - \nabla_{\mathbf{m}}\mathcal{L}_{IRM}(\mathbf{m})\|_2^2}, \tag{17}$$

and we clip w_{ERM}^* to ensure $0 \leq w_{ERM}^* \leq 1$ after each iteration

$$w_{ERM}^* \leftarrow \max\{0, \min\{1, w_{ERM}^*\}\}, \tag{18}$$

and let $w_{IRM} = 1 - w_{ERM}^*$. Finally, we update $\mathbf{m}$ by

$$\mathbf{m} \leftarrow \mathbf{m} - s(w_{ERM}^*\nabla_{\mathbf{m}}\mathcal{L}_{ERM} + w_{IRM}^*\nabla_{\mathbf{m}}\mathcal{L}_{IRM} + \lambda\mathbf{m}), \tag{19}$$

where s is the step-size for invariant mask update.

To prove that the gradient-based update in Eq. (16) and Eq. (19) lead to Pareto optimality, i.e., there exists no $\mathbf{m}'$ such that $\mathcal{L}_{ERM}(\mathbf{m}') \leq \mathcal{L}_{ERM}(\mathbf{m})$ and $\mathcal{L}_{IRM}(\mathbf{m}') \leq \mathcal{L}_{IRM}(\mathbf{m})$, we follow [9, 35, 59] to consider the following optimization problem

$$\begin{aligned}(\Delta\mathbf{m}, \zeta) = \ & \arg\min_{\Delta\mathbf{m}, \zeta} \zeta + \frac{1}{2}\|\Delta\mathbf{m}\|_2^2,\\ & \text{s.t.} \quad (\nabla_{\mathbf{m}}\mathcal{L}_{ERM})^T\Delta\mathbf{m} \leq \zeta, (\nabla_{\mathbf{m}}\mathcal{L}_{IRM})^T\Delta\mathbf{m} \leq \zeta.\end{aligned} \tag{20}$$

Then we claim the solution to this optimization problem is either $\Delta\mathbf{m} = 0$ and the resulting point satisfies the Karush-Kuhn-Tucker (KKT) conditions (i.e., no other solution in its neighborhood can have lower values in both $\mathcal{L}_{ERM}$ and $\mathcal{L}_{IRM}$, thus if we want to improve the performance for a specific task, the other task's performance will be deteriorated), or the solution gives a descent direction that improves both IID and OOD generalization by reducing $\mathcal{L}_{ERM}$ and $\mathcal{L}_{IRM}$ simultaneously.

In fact, the Lagrange function of Eq. (20) can be written as

$$\begin{aligned}&\mathcal{L}(\Delta\mathbf{m}, \zeta, w_{ERM}, w_{IRM}) = \zeta + \frac{1}{2}\|\Delta\mathbf{m}\|_2^2\\ &+ w_{ERM}((\nabla_{\mathbf{m}}\mathcal{L}_{ERM})^T\Delta\mathbf{m} - \zeta) + w_{IRM}((\nabla_{\mathbf{m}}\mathcal{L}_{IRM})^T\Delta\mathbf{m} - \zeta),\end{aligned} \tag{21}$$

where $w_{ERM} \geq 0$ and $w_{IRM} \geq 0$ are the Lagrange multipliers. Then

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\Delta_{\mathbf{m}}} &= \Delta_{\mathbf{m}} - w_{ERM}\cdot\nabla_{\mathbf{m}}\mathcal{L}_{ERM} - w_{IRM}\cdot\nabla_{\mathbf{m}}\mathcal{L}_{IRM} = 0,\\ &\Rightarrow \Delta_{\mathbf{m}} = -w_{ERM}\cdot\nabla_{\mathbf{m}}\mathcal{L}_{ERM} - w_{IRM}\cdot\nabla_{\mathbf{m}}\mathcal{L}_{IRM},\\ \frac{\partial\mathcal{L}}{\partial\zeta} &= 1 - w_{ERM} - w_{IRM}, \quad \Rightarrow w_{ERM} + w_{IRM} = 1.\end{aligned} \tag{22}$$

Algorithm 1: The overall training process of PaInvRL.

Input: $\mathcal{R}^+, \mathcal{R}^-, \mathcal{R}^{tr}$.
for $i \leftarrow 1$ *to* T **do**
  while *not converge* **do**
    for $e \in \mathcal{E}$ **do**
      Optimize $\Gamma_{(e)}$ via Eq. (2);
    end
    for $e \in \mathcal{E}$ **do**
      Compute $\mathcal{R}_e$ via Eq. (9);
    end
  end
  while *not converge* **do**
    $w_{ERM}^* = \frac{(\nabla_{\mathbf{m}}\mathcal{L}_{IRM}(\mathbf{m}) - \nabla_{\mathbf{m}}\mathcal{L}_{ERM}(\mathbf{m}))^\top \nabla_{\mathbf{m}}\mathcal{L}_{IRM}(\mathbf{m})}{\|\nabla_{\mathbf{m}}\mathcal{L}_{ERM}(\mathbf{m}) - \nabla_{\mathbf{m}}\mathcal{L}_{IRM}(\mathbf{m})\|_2^2}$;
    $w_{ERM}^* \leftarrow \max\{0, \min\{1, w_{ERM}^*\}\}$;
    $w_{IRM}^* = 1 - w_{ERM}^*$;
    $\mathbf{m} \leftarrow \mathbf{m} - s(w_{ERM}^*\nabla_{\mathbf{m}}\mathcal{L}_{ERM} + w_{IRM}^*\nabla_{\mathbf{m}}\mathcal{L}_{IRM} + \lambda\mathbf{m})$;
  end
end
Optimize $\Gamma^*(u, i, \Phi|\Theta^*)$ via Eq. (26);
Output: Final recommendation model $\Gamma^*(u, i, \Phi|\Theta^*)$.

Notably, the dual problem of Eq. (20) is Eq. (16), and according to KKT condition, we have

$$\begin{aligned}w_{ERM}^*((\nabla_{\mathbf{m}}\mathcal{L}_{ERM})^T\Delta\mathbf{m}^* - \zeta^*) = 0,\\ w_{IRM}^*((\nabla_{\mathbf{m}}\mathcal{L}_{IRM})^T\Delta\mathbf{m}^* - \zeta^*) = 0.\end{aligned} \tag{23}$$

Thus, if $\Delta\mathbf{m}^* = 0$, then $(\nabla_{\mathbf{m}}\mathcal{L}_{ERM})^T\Delta\mathbf{m}^* = (\nabla_{\mathbf{m}}\mathcal{L}_{IRM})^T\Delta\mathbf{m}^* = 0$. If $\Delta\mathbf{m}^* \neq 0$, then we have $-\|\Delta\mathbf{m}^*\|_2^2 - \zeta^* = 0$, which implies that $(\nabla_{\mathbf{m}}\mathcal{L}_{ERM})^T\Delta\mathbf{m}^* \leq \zeta^* \leq -\|\Delta\mathbf{m}^*\|_2^2$ and $(\nabla_{\mathbf{m}}\mathcal{L}_{IRM})^T\Delta\mathbf{m}^* \leq \zeta^* \leq -\|\Delta\mathbf{m}^*\|_2^2$, and reduces $\mathcal{L}_{ERM}$ and $\mathcal{L}_{IRM}$ simultaneously.

3.5 Representation Convertion

Based on the invariant mask obtained by the invariant mask generation module, we use the convert module to divide the raw multimedia representations into variant representations and invariant representations. Specifically, the invariant representations are

$$\mathbf{\Phi}_i = \mathbf{m} \odot \mathbf{f}_i. \tag{24}$$

Correspondingly, the variant representations can be expressed as

$$\mathbf{\Psi}_i = (1 - \mathbf{m}) \odot \mathbf{f}_i, \tag{25}$$

where $\mathbf{m} \in [0, 1]^d$ is the float invariant mask.

3.6 Final Recommendation Model

By repeating T times the workflow shown in Figure 1 until convergence, stable invariant masks are generated. Thus, we learn the final recommendation model $\Gamma^*(u, i, \mathbf{\Phi}_i|\Theta^*)$ parameterized by Θ^* based on the invariant representations generated by the convert module. The learning objective shown in Eq. (1) can be rewritten as

$$\arg\min_{\Theta^*} \mathcal{L}(\Gamma^*(u, i, \mathbf{\Phi}_i|\Theta^*)|\mathcal{R}^{tr}). \tag{26}$$

The whole training process of PaInvRL is described in Algorithm 1.

Table 1: The statistics of datasets. d_V, d_A, and d_T denote the dimensions of visual, acoustic, and textual modalities.

Dataset	#Interactions	#Items	#Users	Sparsity	d_V	d_A	d_T
Movielens	1,239,508	5,986	55,485	99.63%	2,048	128	100
Tiktok	726,065	76,085	36,656	99.99%	128	128	128
Kwai	298,492	86,483	7,010	99.98%	2,048	-	-

4 EXPERIMENTS

In this section, we conduct experiments on three widely used real-world datasets to answer the following research questions:

- **RQ1**: Can PaInvRL outperform other recommendation methods in both IID and OOD tasks?
- **RQ2**: How masks incorporating attention mechanisms affect learned representations?
- **RQ3**: How does each component in $\mathcal{L}_{\text{mask}}$ affect the performance of PaInvRL in both IID and OOD tasks?
- **RQ4**: How does the number of environments affect the performance of PaInvRL?

4.1 Datasets

We conduct experiments on three publicly available real-world datasets: Movielens, Tiktok, and Kwai. The summary statistics of these datasets are shown in Table 1.

Movielens. This dataset is widely used in personalized recommendation tasks. The dataset is constructed by collecting movie titles and descriptions from the Movielens dataset[1] and retrieving the corresponding trailers. The visual, acoustic, and textual representations were extracted from the pre-trained ResNet50 [14], VGGish [18], and Sentence2Vec [3], respectively.

Tiktok. It is collected from the micro-video sharing platform TikTok[2]. It includes micro-videos with a duration of 3-15 seconds, along with video captions, user information, and user-item interactions. The multi-modal representations include visual, acoustic, and textual representations of micro-videos. All of the multi-modal representations are provided by the official.

Kwai. It is a large-scale micro-video dataset collected from the Kwai platform[3]. Similar to the TikTok dataset, it includes user information, micro-video content representations, and interaction data. We follow the previous work [10] to obtain the raw multimedia representations. It should be noticed that this dataset only includes visual representations.

4.2 Experiment Setup

4.2.1 Baselines. To verify the effectiveness of PaInvRL, we compare it with the following baseline methods:

NGCF [72]. It is based on graph neural networks that explicitly encode collaborative signals as higher-order connections by performing embedding propagation.

UltraGCN [47]. It is an ultra-simplified GCN model that does not perform explicit message passing, but directly approximates the limit of infinite layer graph convolutions by constraining losses.

LightGCN [16]. It is a graph-based model designed to improve the performance and efficiency of recommendations by simplifying the graph convolution networks.

VBPR [15]. It is the first model that considers introducing visual representation into the recommendation system by concatenating visual embeddings with id embeddings as the item representations.

MMGCN [79]. It is a model that builds on the message-passing idea of graph neural networks to generate user and micro-video-specific pattern representations to capture user preferences better.

InvRL [10]. This model introduces IRM to multi-modal recommendations for the first time, which mitigates the effects of spurious correlations by learning invariant item representations.

MMSSL [76]. This method solves the problem of label sparsity in multimedia recommendations by two-stage self-supervised learning to achieve modality-aware data scaling.

4.2.2 Experiment Protocol and Details. Following the previous work [10], three widely-used metrics are adopted to evaluate the ranking performance: Recall@K (R@K), NDCG@K (N@K), and Precision@K (P@K). We set $K = 10$ in our experiments. All the experiments are implemented with PyTorch [51] and Adam is implemented as the optimizer. The embedding size is fixed to 64 for all models. For Movielens, we set d_V = 2,048, d_A = 128, and d_T = 10. For Tiktok, we set d_V = 128, d_A = 128, and d_T = 128. For Kwai, we only use visual representations and set d_V = 4,096. The batch size is set to 512 and the number of environments is set to 10. We also set the parameters λ in Eq. (11) to 1, and the hyper-parameters η and κ in Eq. (3) to 0.0001 and 0.01, respectively. The heterogeneous identification module, invariant mask generation module, and the final recommendation model are trained for 20, 40, and 500 epochs, respectively. To evaluate the performance of PaInvRL in both IID and OOD tasks, we first use UltraGCN to identify two environments using the heterogeneous identification module. We train the model in the environment that contains more data, and test the model in the environment that contains less data for the OOD task. We split the training set for the OOD task into two parts with 9:1 ratio to obtain the training set and test set for the IID task.

4.3 Performance Comparison (RQ1)

We report the performance of various methods on all three datasets in Table 2, where the best-performing baselines are bolded. We have the following observations.

First, multi-modality-based methods outperform single-modality-based methods in both IID and OOD tasks, and MMSSL achieves the most competitive performance among all the baseline methods.

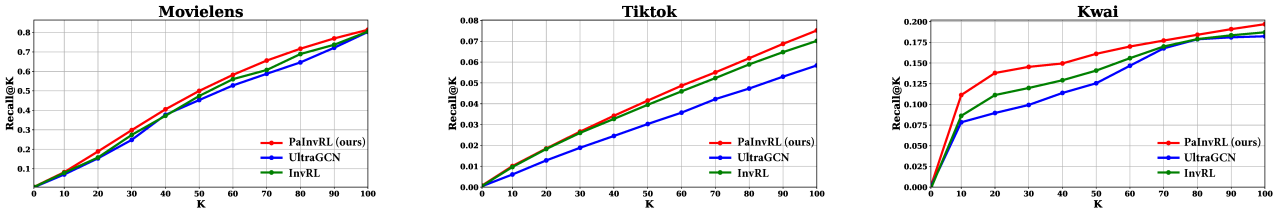
Second, in the OOD recommendation task, it shows that PaInvRL significantly outperforms other methods, due to PaInvRL learning invariant representations and identifying spuriously correlation. In addition, it should be noticed that PaInvRL outperforms InvRL, which is attributed to PaInvRL learning a better mask by considering Pareto optimization and weighting both invariant and variant representation using the attention mechanism.

Third, in the IID task, although InvRL achieved better performance than some single-modality-based methods like NGCF and LightGCN, its performance is not as good as that of other multi-modality-based methods like MMGCN. This is because InvRL only focuses on learning invariant representations, which leads to performance degradation in the IID task. However, the proposed method

[1] https://movielens.org/.
[2] https://www.tiktok.com/.
[3] https://www.kwai.com/.

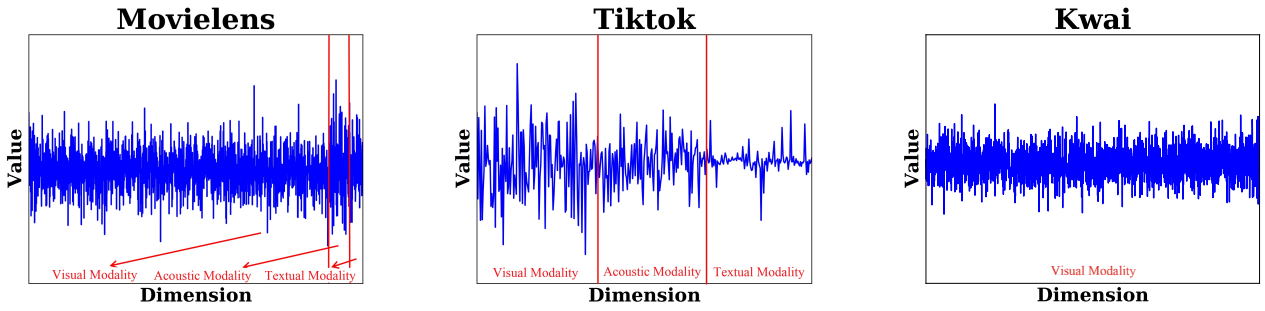
Table 2: Performance comparison on different datasets in terms of Recall@10, Precision@10, and NDCG@10.

Task	Method	Modality	Movielens			Tiktok			Kwai		
			P@10	R@10	N@10	P@10	R@10	N@10	P@10	R@10	N@10
IID	NGCF [72]	Single	0.0180	0.1355	0.0383	0.0138	0.0409	0.0513	0.0425	0.0487	0.0697
	UltraGCN [47]	Single	0.0126	0.1060	0.0418	0.0163	0.0437	0.0543	0.0459	0.0509	0.0729
	LightGCN [16]	Single	0.0215	0.1643	0.0554	0.0164	0.0444	**0.0584**	0.0496	0.0435	0.0688
	VBPR [15]	Multi	0.0176	0.1290	0.0400	0.0142	0.0409	0.0469	0.0409	0.0476	0.0682
	MMGCN [79]	Multi	0.0207	0.1613	0.0641	0.0154	0.0444	**0.0584**	0.0496	0.0535	0.0738
	InvRL [10]	Multi	0.0218	**0.1681**	0.0617	0.0213	0.0440	0.0576	0.0528	0.0549	0.0729
	MMSSL [76]	Multi	0.0237	0.1587	0.0572	0.0202	0.0443	0.0555	0.0523	0.0518	0.0748
	PaInvRL (ours)	Multi	**0.0240**	0.1660	**0.0650**	**0.0229**	**0.0463**	0.0578	**0.0536**	**0.0595**	**0.0815**
OOD	NGCF [72]	Single	0.0191	0.0733	0.0474	0.0048	0.0060	0.0153	0.0411	0.0828	0.1466
	UltraGCN [47]	Single	0.0212	0.0708	0.0508	0.0043	0.0061	0.0174	0.0321	0.0784	0.1464
	LightGCN [16]	Single	0.0159	0.0638	0.0412	0.0053	0.0082	0.0169	0.0420	0.0883	0.1331
	VBPR [15]	Multi	0.0165	0.0649	0.0396	0.0036	0.0057	0.0153	0.0324	0.0763	0.1504
	MMGCN [79]	Multi	0.0188	0.0732	0.0463	0.0044	0.0058	0.0161	0.0402	0.0831	0.1503
	InvRL [10]	Multi	0.0253	0.0791	0.0543	0.0059	0.0097	0.0226	0.0407	0.0862	0.1894
	MMSSL [76]	Multi	0.0225	0.0813	0.0537	0.0055	0.0098	0.0234	0.0462	0.1036	0.1682
	PaInvRL (ours)	Multi	**0.0291**	**0.0825**	**0.0584**	**0.0067**	**0.0107**	**0.0252**	**0.0524**	**0.1113**	**0.2061**



(a) Movielens (b) Tiktok (c) Kwai

Figure 2: The performance comparison between UltraGCN, InvRL and PaInvRL on all three datasets using Recall@K as evaluation metric, where K is varying from 0 to 100 with step-size 10.



(a) Movielens (b) Tiktok (c) Kwai

Figure 3: Visualization of the masks on different modalities and corresponding patterns on all three datasets.

PaInvRL weights the ERM loss $\mathcal{L}_{ERM}$ and IRM loss $\mathcal{L}_{IRM}$ to ensure the learned representations are able to perform well in both IID and OOD tasks. Therefore, PaInvRL also achieves the best performance compared to other methods in the IID task.

Overall speaking, PaInvRL not only outperforms other baseline methods in the OOD task, but also has the best performance in the IID task. In addition, we conduct a more detailed experiment to compare PaInvRL, InvRL, and UltraGCN using Recall@K as the evaluation metric in the OOD task on all three datasets. The results are presented in Figure 2, which indicates that PaInvRL stably outperforms UltraGCN and InvRL across different K values, which further verifies the effectiveness of the proposed method.

Table 3: Performance comparison with different loss components in the IID and OOD tasks.

Task	Loss	Movielens			Tiktok			Kwai		
		P@10	R@10	N@10	P@10	R@10	N@10	P@10	R@10	N@10
IID	$\mathcal{L}_{ERM}$	0.0253	0.1068	0.0410	0.0263	0.0537	0.0643	0.0574	0.0619	0.0837
	$\mathcal{L}_{IRM}$	0.0141	0.1027	0.0311	0.0217	0.0385	0.0572	0.0562	0.0612	0.0831
	$\mathcal{L}_{ERM} + \mathcal{L}_{IRM}$	0.0240	0.1660	0.0650	0.0229	0.0463	0.0578	0.0536	0.0595	0.0815
OOD	$\mathcal{L}_{ERM}$	0.0212	0.0583	0.0380	0.0058	0.0097	0.0234	0.0508	0.1026	0.1983
	$\mathcal{L}_{IRM}$	0.0353	0.1009	0.0693	0.0068	0.0103	0.0235	0.0575	0.1504	0.2522
	$\mathcal{L}_{ERM} + \mathcal{L}_{IRM}$	0.0291	0.0825	0.0584	0.0067	0.0107	0.0252	0.0524	0.1113	0.2061

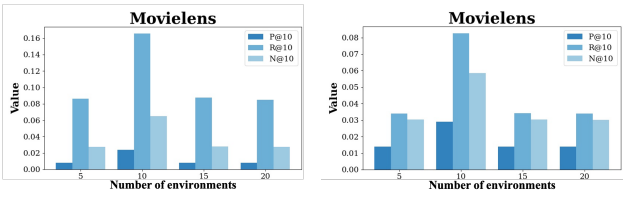
4.4 Ablation Study (RQ2)

In ablation studies, we first investigate the effect of IRM loss $\mathcal{L}_{IRM}$ and ERM loss $\mathcal{L}_{ERM}$, which is used for training the invariant mask generation module. Then we discuss how the generated mask m works. Additionally, we conducted experiments to study the impact of environmental quantity on experimental performance.

We consider three cases in the ablation study: only with the ERM loss $\mathcal{L}_{ERM}$, only with the IRM loss $\mathcal{L}_{IRM}$, and with adaptively weighted ERM loss and IRM loss $\mathcal{L}_{ERM} + \mathcal{L}_{IRM}$ across all three datasets. The experiment results are shown in Table 3. From this table, we can observe that when PaInvRL only with the IRM loss, it is able to achieve the best performance in the OOD task but perform the worst in the IID task. Meanwhile, when only using the ERM loss, it will perform best in the IID task but perform worst in the OOD task. It shows that if only focusing on a single task, though we will obtain a competitive result, the model performance is harmed in another task, which shows the necessity of considering two tasks simultaneously. When using weighted IRM loss and ERM loss together, it has competitive performance in both IID tasks and OOD tasks. Overall, using only one type of loss or simply using a hyper-parameter to weight them directly (i.e., InvRL) cannot achieve good recommendation performance. When we adaptively weight ERM loss and IRM loss together and obtain the weights from a Pareto optimal solution, it obtains competing recommendation performance. This can be attributed to the fact that our solution cannot be dominated by other solutions. In other words, there does not exist any solution that performs better than our solution on both IID and OOD tasks at the same time.

4.5 In Depth Analysis (RQ3, RQ4)

Study on the Generated Mask (RQ3). To study the effect of the generated mask m in the invariant mask generation module, we visualize the invariant mask generated on three datasets, Movielens, Tiktok, and Kwai, as shown in Figure 3. According to the results in Figure 3, the generated masks show different distributions in different modalities, especially, for the Movielens and Tiktok datasets, which contain three different modal representations of visual, acoustic, and textual. Since the Kwai dataset has only one modal representation, the distribution of the masks varies subtly. Additionally, our method demonstrates a more uniform distribution across different modalities compared to InvRL [10]. It can be attributed to the fact that PaInvRL learns a better mask by considering Pareto optimization during mask generation, while InvRL only considers a simple hyper-parameter to weight two losses together.



(a) IID (b) OOD

Figure 4: Experimental comparison of different environment numbers on IID and OOD recommendation tasks.

Study on Number of Environments (RQ4). To investigate the capacity of PaInvRL under different numbers of environments, we conduct several experiments on the Movielens dataset with different numbers of environments. The experimental results are shown in Figure 4. First, PaInvRL performs better under a moderate number of experiments. When the number of environments is small, we cannot effectively separate the variant and invariant information. When the number of environments is large, only a few samples are in each environment. Therefore, either too small or too large number will harm the performance of the proposed method.

5 CONCLUSIONS

In this paper, we provide a fresh perspective on the optimization dilemma in the IID-OOD generalization task of multimedia recommendation from a multi-objective optimization viewpoint. We propose a new Pareto-optimality-based invariant representation learning method, PaInvRL, which adaptively assigns the weights of ERM loss and IRM loss to obtain Pareto-optimal solutions. In contrast to previous approaches like InvRL, our gradient-based invariant mask generation method is shown to provide a descent direction that improves both IID and OOD generalization by reducing ERM and IRM losses simultaneously. This allows the final recommendation model trained on the learned invariant representations to achieve Pareto optimality in both IID and OOD recommendation tasks. Extensive experimental results show that our method achieves significant performance improvements compared to various baselines on three public datasets. In our future work, it is interesting to enhance the explainability of the learned invariant representations by developing a GNN-based explainer to learn causal effects on modality-aware user-item interaction graphs. This will help provide insights into how the invariant representations contribute to the recommendation performance and enable us to make more informed decisions in the recommendation process.

ACKNOWLEDGMENTS

This work was supported by grants from the National Major Science and Technology Projects of China (grant no: 2022YFB3303302), the National Natural Science Foundation of China (grant nos: 61977012, 62207007) and the Central Universities Project in China for the Digital Cinema Art Theory and Technology Lab at Chongqing University (grant nos: 2021CDJYGRH011, 2020CDJSK06PT14).

REFERENCES

[1] Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. 2020. Invariant Risk Minimization Games. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 145–155.

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2020. Invariant Risk Minimization. arXiv:1907.02893 [stat.ML]

[3] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *International conference on learning representations*.

[4] Stephen Bonner and Flavian Vasile. 2018. Causal embeddings for recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 104–112.

[5] Jiawei Chen, Hande Dong, Yang Qiu, Xiangnan He, Xin Xin, Liang Chen, Guli Lin, and Keping Yang. 2021. AutoDebias: Learning to Debias for Recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[6] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems* 41, 3 (2023), 1–39.

[7] Hugh A Chipman, Edward I George, and Robert E McCulloch. 2010. BART: Bayesian additive regression trees. *The Annals of Applied Statistics* 4, 1 (2010), 266–298.

[8] Quanyu Dai, Haoxuan Li, Peng Wu, Zhenhua Dong, Xiao-Hua Zhou, Rui Zhang, Xiuqiang He, Rui Zhang, and Jie Sun. 2022. A Generalized Doubly Robust Learning Framework for Debiasing Post-Click Conversion Rate Prediction. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

[9] Jean-Antoine Désidéri. 2012. Multiple-gradient descent algorithm (MGDA) for multiobjective optimization. *Comptes Rendus Mathematique* 350, 5-6 (2012), 313–318.

[10] Xiaoyu Du, Zike Wu, Fuli Feng, Xiangnan He, and Jinhui Tang. 2022. Invariant Representation Learning for Multimedia Recommendation. In *Proceedings of the 30th ACM International Conference on Multimedia*. 619–628.

[11] Siyuan Guo, Lixin Zou, Yiding Liu, Wenwen Ye, Suqi Cheng, Shuaiqiang Wang, Hechang Chen, Dawei Yin, and Yi Chang. 2021. Enhanced Doubly Robust Learning for Debiasing Post-Click Conversion Rate Estimation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[12] Wei Guo, Yang Yang, Yaochen Hu, Chuyuan Wang, Huifeng Guo, Yingxue Zhang, Ruiming Tang, Weinan Zhang, and Xiuqiang He. 2021. Deep graph convolutional networks with hybrid normalization for accurate and diverse recommendation. In *Proceedings of 3rd Workshop on Deep Learning Practice for High-Dimensional Sparse Data with KDD*.

[13] Behram Hansotia and Brad Rukstales. 2002. Incremental value modeling. *Journal of Interactive Marketing* 16 (2002), 35–46.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[15] Ruining He and Julian McAuley. 2016. VBPR: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 30.

[16] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 639–648.

[17] Miguel Hernán and Jamie Robins. 2020. *Causal Inference: What If*. Boca Raton: Chapman and Hall/CRC.

[18] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 131–135.

[19] Daniel G Horvitz and Donovan J Thompson. 1952. A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association* 47 (1952), 663–685.

[20] Shanshan Huang, Xin Jin, Qian Jiang, and Li Liu. 2022. Deep learning for image colorization: Current and future prospects. *Engineering Applications of Artificial Intelligence* 114 (2022), 105006.

[21] Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

[22] Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *International conference on machine learning*. PMLR, 3020–3029.

[23] Wang-Cheng Kang, Chen Fang, Zhaowen Wang, and Julian McAuley. 2017. Visually-aware fashion recommendation and design with generative image models. In *2017 IEEE international conference on data mining*. IEEE, 207–216.

[24] Edward H. Kennedy. 2020. Optimal doubly robust estimation of heterogeneous causal effects. *https://arxiv.org/abs/2004.14497v1* (2020).

[25] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. 2021. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*. PMLR, 5815–5826.

[26] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences* 116, 10 (2019), 4156–4165.

[27] Fei Lei, Zhongqi Cao, Yuning Yang, Yibo Ding, and Cong Zhang. 2023. Learning the User's Deeper Preferences for Multi-modal Recommendation Systems. *ACM Transactions on Multimedia Computing, Communications and Applications* 19, 3s (2023), 1–18.

[28] Haoxuan Li, Quanyu Dai, Yuru Li, Yan Lyu, Zhenhua Dong, Xiao-Hua Zhou, and Peng Wu. 2023. Multiple Robust Learning for Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[29] Haoxuan Li, Yan Lyu, Chunyuan Zheng, and Peng Wu. 2023. TDR-CL: Targeted Doubly Robust Collaborative Learning for Debiased Recommendations. In *International Conference on Learning Representations*.

[30] Haoxuan Li, Yanghao Xiao, Chunyuan Zheng, and Peng Wu. 2023. Balancing unobserved confounding with a few unbiased ratings in debiased recommendations. In *Proceedings of the ACM Web Conference 2023*. 1305–1313.

[31] Haoxuan Li, Yanghao Xiao, Chunyuan Zheng, Peng Wu, and Peng Cui. 2023. Propensity Matters: Measuring and Enhancing Balancing for Recommendation. In *International Conference on Machine Learning*.

[32] Haoxuan Li, Chunyuan Zheng, Yixiao Cao, Zhi Geng, Yue Liu, and Peng Wu. 2023. Trustworthy Policy Learning under the Counterfactual No-Harm Criterion. In *International Conference on Machine Learning*.

[33] Haoxuan Li, Chunyuan Zheng, and Peng Wu. 2023. StableDR: Stabilized Doubly Robust Learning for Recommendation on Data Missing Not at Random. In *International Conference on Learning Representations*.

[34] Haoxuan Li, Chunyuan Zheng, Peng Wu, Kun Kuang, Yue Liu, and Peng Cui. 2023. Who should be Given Incentives? Counterfactual Optimal Treatment Regimes Learning for Recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

[35] Xi Lin, Hui-Ling Zhen, Zhenhua Li, Qing-Fu Zhang, and Sam Kwong. 2019. Pareto multi-task learning. *Advances in neural information processing systems* 32 (2019).

[36] Yong Lin, Shengyu Zhu, Lu Tan, and Peng Cui. 2022. ZIN: When and How to Learn Invariance Without Environment Partition? *Advances in Neural Information Processing Systems* 35 (2022), 24529–24542.

[37] Dugang Liu, Pengxiang Cheng, Zhenhua Dong, Xiuqiang He, Weike Pan, and Zhong Ming. 2020. A general knowledge distillation framework for counterfactual recommendation via uniform data. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 831–840.

[38] Dugang Liu, Pengxiang Cheng, Hong Zhu, Zhenhua Dong, Xiuqiang He, Weike Pan, and Zhong Ming. 2021. Mitigating Confounding Bias in Recommendation via Information Bottleneck. In *Proceedings of the 15th ACM conference on Recommender systems*.

[39] Fan Liu, Zhiyong Cheng, Changchang Sun, Yinglong Wang, Liqiang Nie, and Mohan Kankanhalli. 2019. User diverse preference modeling by multimodal attentive metric learning. In *Proceedings of the 27th ACM international conference on multimedia*. 1526–1534.

[40] Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. 2021. Heterogeneous risk minimization. In *International Conference on Machine Learning*. PMLR, 6804–6814.

[41] Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. 2021. Kernelized heterogeneous risk minimization. (2021).

[42] Qiang Liu, Shu Wu, and Liang Wang. 2017. Deepstyle: Learning user preferences for visual recommendation. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*. 841–844.

[43] Xiaohao Liu, Zhulin Tao, Jiahong Shao, Lifang Yang, and Xianglin Huang. 2022. EliMRec: Eliminating Single-modal Bias in Multimedia Recommendation. In *Proceedings of the 30th ACM International Conference on Multimedia*. 687–695.

[44] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems* 32 (2019).

[45] Jianxin Ma, Peng Cui, Kun Kuang, Xin Wang, and Wenwu Zhu. 2019. Disentangled graph convolutional networks. In *International conference on machine learning*. PMLR, 4212–4221.

[46] Jianxin Ma, Chang Zhou, Peng Cui, Hongxia Yang, and Wenwu Zhu. 2019. Learning disentangled representations for recommendation. *Advances in neural information processing systems* 32 (2019).

[47] Kelong Mao, Jieming Zhu, Xi Xiao, Biao Lu, Zhaowei Wang, and Xiuqiang He. 2021. UltraGCN: ultra simplification of graph convolutional networks for recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 1253–1262.

[48] Benjamin Marlin, Richard S Zemel, Sam Roweis, and Malcolm Slaney. 2007. Collaborative filtering and the missing at random assumption. *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence* (2007).

[49] Zongshen Mu, Yueting Zhuang, Jie Tan, Jun Xiao, and Siliang Tang. 2022. Learning Hybrid Behavior Patterns for Multimedia Recommendation. In *Proceedings of the 30th ACM International Conference on Multimedia*. 376–384.

[50] Xinkun Nie and Stefan Wager. 2021. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* 108, 2 (2021), 299–319.

[51] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[52] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. 2016. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* (2016), 947–1012.

[53] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[54] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. 1994. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association* 89, 427 (1994), 846–866.

[55] Yuta Saito. 2020. Doubly robust estimator for ranking metrics with post-click conversions. In *Fourteenth ACM Conference on Recommender Systems*. 92–100.

[56] J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. 2007. Collaborative filtering recommender systems. In *The adaptive web: methods and strategies of web personalization*. Springer, 291–324.

[57] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as Treatments: Debiasing Learning and Evaluation. In *International Conference on Machine Learning*.

[58] Patrick Schwab, Lorenz Linhardt, and Walter Karlen. 2018. Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv preprint arXiv:1810.00656* (2018).

[59] Ozan Sener and Vladlen Koltun. 2018. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems* 31 (2018).

[60] Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*. PMLR, 3076–3085.

[61] Tiancheng Shen, Jia Jia, Yan Li, Hanjie Wang, and Bo Chen. 2020. Enhancing music recommendation with social media content: an attentive multimodal autoencoder approach. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[62] Claudia Shi, David Blei, and Victor Veitch. 2019. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems* 32 (2019).

[63] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. (2015), 1–14.

[64] Harald Steck. 2010. Training and testing of recommender systems on data missing not at random. In *Proceedings of the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

[65] Xiaoyuan Su and Taghi M Khoshgoftaar. 2009. A survey of collaborative filtering techniques. *Advances in artificial intelligence* 2009 (2009).

[66] Rui Sun, Xuezhi Cao, Yan Zhao, Junchen Wan, Kun Zhou, Fuzheng Zhang, Zhongyuan Wang, and Kai Zheng. 2020. Multi-modal knowledge graphs for recommender systems. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 1405–1414.

[67] Zhulin Tao, Yinwei Wei, Xiang Wang, Xiangnan He, Xianglin Huang, and Tat-Seng Chua. 2020. Mgat: Multimodal graph attention network for recommendation. *Information Processing & Management* 57, 5 (2020), 102277.

[68] Stefan Wager and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* 113, 523 (2018), 1228–1242.

[69] Hao Wang, Tai-Wei Chang, Tianqiao Liu, Jianmin Huang, Zhichao Chen, Chao Yu, Ruopeng Li, and Wei Chu. 2022. Escm2: Entire space counterfactual multi-task model for post-click conversion rate estimation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 363–372.

[70] Qifan Wang, Yinwei Wei, Jianhua Yin, Jianlong Wu, Xuemeng Song, and Liqiang Nie. 2021. Dualgnn: Dual graph neural network for multimedia recommendation. *IEEE Transactions on Multimedia* (2021).

[71] Wenjie Wang, Yang Zhang, Haoxuan Li, Peng Wu, Fuli Feng, and Xiangnan He. 2023. Causal recommendation: Progresses and future directions. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3432–3435.

[72] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural graph collaborative filtering. In *Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval*. 165–174.

[73] Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. 2019. Doubly Robust Joint Learning for Recommendation on Data Missing Not at Random. In *International Conference on Machine Learning*.

[74] Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. 2021. Combating Selection Biases in Recommender Systems with A Few Unbiased Ratings. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*.

[75] Zifeng Wang, Xi Chen, Rui Wen, Shao-Lun Huang, Ercan Kuruoglu, and Yefeng Zheng. 2020. Information theoretic counterfactual learning from missing-not-at-random feedback. *Advances in Neural Information Processing Systems* (2020), 1854–1864.

[76] Wei Wei, Chao Huang, Lianghao Xia, and Chuxu Zhang. 2023. Multi-Modal Self-Supervised Learning for Recommendation. In *Proceedings of the ACM Web Conference 2023* (Austin, TX, USA). Association for Computing Machinery, New York, NY, USA, 790–800.

[77] Yinwei Wei, Xiang Wang, Xiangnan He, Liqiang Nie, Yong Rui, and Tat-Seng Chua. 2021. Hierarchical user intent graph network for multimedia recommendation. *IEEE Transactions on Multimedia* 24 (2021), 2701–2712.

[78] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2020. Graph-refined convolutional network for multimedia recommendation with implicit feedback. In *Proceedings of the 28th ACM international conference on multimedia*. 3541–3549.

[79] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM international conference on multimedia*. 1437–1445.

[80] Peng Wu, Haoxuan Li, Yuhao Deng, Wenjie Hu, Quanyu Dai, Zhenhua Dong, Jie Sun, Rui Zhang, and Xiao-Hua Zhou. 2022. On the Opportunity of Causal Learning in Recommendation Systems: Foundation, Estimation, Prediction and Challenges. In *Proceedings of the 31st International Conference on International Joint Conferences on Artificial Intelligence*.

[81] Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. 2019. Session-based recommendation with graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 346–353.

[82] Renzhe Xu, Xingxuan Zhang, Peng Cui, Bo Li, Zheyan Shen, and Jiazheng Xu. 2022. Regulatory instruments for fair personalized pricing. In *Proceedings of the ACM Web Conference 2022*. 4–15.

[83] Yilun Xu and Tommi Jaakkola. 2021. Learning representations that support robust transfer of predictors. *arXiv preprint arXiv:2110.09940* (2021).

[84] Wenhui Yu, Huidi Zhang, Xiangnan He, Xu Chen, Li Xiong, and Zheng Qin. 2018. Aesthetic-based clothing recommendation. In *Proceedings of the 2018 world wide web conference*. 649–658.

[85] Bowen Yuan, Jui-Yang Hsia, Meng-Yuan Yang, Hong Zhu, Chih-Yao Chang, Zhenhua Dong, and Chih-Jen Lin. 2019. Improving ad click prediction by considering non-displayed events. In *Proceedings of the 28th ACM International Conference on Information & Knowledge Management*. 329–338.

[86] Bowen Yuan, Yaxu Liu, Jui-Yang Hsia, Zhenhua Dong, and Chih-Jen Lin. 2020. Unbiased Ad Click Prediction for Position-aware Advertising Systems. In *Fourteenth ACM Conference on Recommender Systems*. 368–377.

[87] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. 2021. Mining latent structures for multimedia recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3872–3880.

[88] Xingxuan Zhang, Zekai Xu, Renzhe Xu, Jiashuo Liu, Peng Cui, Weitao Wan, Chong Sun, and Chen Li. 2022. Towards domain generalization in object detection. *arXiv preprint arXiv:2203.14387* (2022).

[89] Yu Zhang and Qiang Yang. 2021. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering* 34, 12 (2021), 5586–5609.

[90] Zeyu Zhang, Heyang Gao, Hao Yang, and Xu Chen. 2023. Hierarchical Invariant Learning for Domain Generalization Recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3470–3479.

[91] Kailiang Zhong, Fengtong Xiao, Yan Ren, Yaorong Liang, Wenqing Yao, Xiaofeng Yang, and Ling Cen. 2022. DESCN: Deep Entire Space Cross Networks for Individual Treatment Effect Estimation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4612–4620.

[92] Hongyu Zhou, Xin Zhou, and Zhiqi Shen. 2023. Enhancing Dyadic Relations with Homogeneous Graphs for Multimodal Recommendation. *arXiv preprint arXiv:2301.12097* (2023).

[93] Hongyu Zhou, Xin Zhou, Zhiwei Zeng, Lingzi Zhang, and Zhiqi Shen. 2023. A Comprehensive Survey on Multimodal Recommender Systems: Taxonomy, Evaluation, and Future Directions. *arXiv preprint arXiv:2302.04473* (2023).

[94] Xiao Zhou, Yong Lin, Weizhong Zhang, and Tong Zhang. 2022. Sparse invariant risk minimization. In *International Conference on Machine Learning*. PMLR, 27222–27244.