One for Dozens: Adaptive REcommendation for All Domains with Counterfactual Augmentation

Huishi Luo[1], Yiwen Chen[1], Yiqing Wu[2], Fuzhen Zhuang[1,3*], Deqing Wang[3]

[1] Institute of Artificial Intelligence, Beihang University
[2] Institute of Computing Technology, Chinese Academy of Sciences
[3] SKLSDE, School of Computer Science, Beihang University
{hsluo2000, yiwenchen}@buaa.edu.cn, wuyiqing20s@ict.ac.cn, {zhuangfuzhen, dqwang}@buaa.edu.cn

Abstract

Multi-domain recommendation (MDR) aims to enhance recommendation performance across various domains. However, real-world recommender systems in online platforms often need to handle dozens or even hundreds of domains, far exceeding the capabilities of traditional MDR algorithms, which typically focus on fewer than five domains. Key challenges include a substantial increase in parameter count, high maintenance costs, and intricate knowledge transfer patterns across domains. Furthermore, minor domains often suffer from data sparsity, leading to inadequate training in classical methods. To address these issues, we propose Adaptive REcommendation for All Domains with counterfactual augmentation (AREAD). AREAD employs a hierarchical structure with a limited number of expert networks at several layers, to effectively capture domain knowledge at different granularities. To adaptively capture the knowledge transfer pattern across domains, we generate and iteratively prune a hierarchical expert network selection mask for each domain during training. Additionally, counterfactual assumptions are used to augment data in minor domains, supporting their iterative mask pruning. Our experiments on two public datasets, each encompassing over twenty domains, demonstrate AREAD's effectiveness, especially in data-sparse domains.

Code — https://github.com/Chrissie-Law/AREAD-Multi-Domain-Recommendation

1 Introduction

Recommender systems (RSs) have become essential in many web applications to offer personalized recommendations and combat information overload. With an increasing variety of items, platforms like Amazon and Alibaba segment their offerings into different channels or content pages, necessitating that RSs efficiently manage recommendations across these diverse domains. Multi-domain recommendation (MDR) systems mitigate the high costs of maintaining separate models for each domain by capturing cross-domain user interests more effectively in a unified model, thus enhancing recommendation performance across various domains (Chang et al. 2023; Li et al. 2023; Zhao et al. 2023; Zhang et al. 2023; Gan et al. 2024).

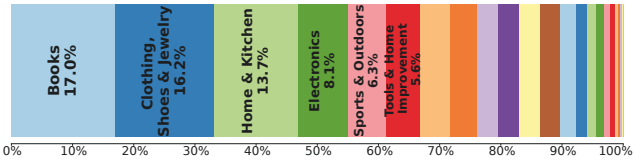

Figure 1: Sample size across 25 Amazon dataset domains, with 12 minor comprising less than 2% of total samples.

However, most MDR algorithms are limited to a handful of domains, typically fewer than five (Wang et al. 2022; Zhang et al. 2022b; Chang et al. 2023). In contrast, real-world large-scale recommendation scenarios involve dozens or even hundreds of domains. For example, on Amazon[1], there are as many as 25 domains just identified using the coarsest item categorization (refer to Figure 1). Furthermore, these domains exhibit a significant long-tail effect, with major domains having substantially more data samples than minor ones. Therefore, training a multi-domain recommendation algorithm capable of handling such a diverse and extensive range of domain data faces the following challenges:

Challenge 1: Scalability in Systems with Numerous Domains. Traditional RS methods (Guo et al. 2017; Song et al. 2019; Wang et al. 2021) process multi-domain data in a single-domain manner, optimizing processing efficiency but sacrificing important domain specificity. To better capture the characteristics of each domain, most MDR algorithms refer to multi-task learning solutions, employing a separate "tower" network structure for each domain before output (Ma et al. 2018a; Sheng et al. 2021; Zou et al. 2022; Zhou et al. 2023). Nevertheless, this strategy significantly increases the model's parameter count, and leads to inadequate convergence of these domain-specific networks as the number of domains grows. Additionally, pre-training-based MDR algorithms (Gu et al. 2021; Zhang et al. 2022b) incur unacceptably high maintenance costs for fine-tuned models of each domain, and their direct application is often ineffective in domains with low cross-domain relevance.

Challenge 2: Complex Cross-domain Knowledge Transfer Pattern. Traditional MDR typically categorizes knowledge as either domain-shared or domain-specific.

*Corresponding author.

[1] https://nijianmo.github.io/amazon/index.html

However, researchers (Standley et al. 2020) highlight the directional nature of knowledge transfer, where beneficial knowledge in one domain may not be applicable in another, thus challenging this binary categorization framework. As domain numbers increase, the shortcomings of this framework become more evident, compounding the complexities of cross-domain knowledge transfer. Consequently, a crucial challenge in MDR with numerous domains is determining *which domains should and should not be learned together*, a problem beyond the scope of traditional binary knowledge classification. While clustering domains and treating each cluster as a single domain offers a solution, this approach tends to overlook intra-cluster domain variations.

Challenge 3: Large Variance in Sample Size across Domains. As shown in Figure 1, nearly half of the domains account for less than 2% of the total samples. This disparity leads to the dominance of data-abundant domains in the optimization process of existing MDR models (Zhang et al. 2023), impacting models regardless of their framework, whether those with explicit domain-specific parameters (Sheng et al. 2021; Zhou et al. 2023) or those inspired by meta-learning (Zhu et al. 2021; Guan et al. 2022) and hypernetworks (Liu et al. 2023; Chang et al. 2023). Consequently, these models often suffer from insufficient optimization in sparse domains, which are particularly prevalent in datasets spanning a large number of domains.

To address the above challenges, we propose Adaptive REcommendation for All Domains with counterfactual augmentation (AREAD), a unified framework that adaptively models relationships across a vast array of domains at different granularities. AREAD is inspired by the hierarchical clustering technique and employs a hierarchical structure. Specifically, the structure utilizes a few expert networks at lower levels to capture coarse-grained domain knowledge, and a greater, yet still limited, number of expert networks at higher levels for finer-grained knowledge, thereby alleviating the parameter overhead in numerous domains (Challenge 1). AREAD leverages the Lottery Ticket Hypothesis (Frankle and Carbin 2019), generating and iteratively pruning a hierarchical expert selection mask for each domain during training, effectively capturing the complex knowledge transfer patterns across domains (Challenge 2). Furthermore, AREAD performs data augmentation for minor domains according to popularity-based counterfactual assumptions (Challenge 3). Finally, each domain utilizes its specific mask to determine the appropriate experts for the inference stage. We highlight our contributions as follows:

• We propose the AREAD, addressing multi-domain recommendation involving dozens of domains.

• AREAD incorporates a novel hierarchical expert sharing structure, with an iterative mask pruning approach to learn domain-specific sub-networks of experts for each domain. Additionally, we employ popularity-based counterfactual assumptions to augment data for minor domains.

• Extensive experiments on two widely recognized public datasets, each with over 20 domains, demonstrate AREAD's significant improvements across diverse domains. AREAD excels in enhancing attention to minor domains while simultaneously improving overall performance.

2 Methodology

2.1 Preliminaries and Background

Consider a set of recommendation domains $\mathcal{D} = \{1, 2, \cdots, D\}$, with input general features $\boldsymbol{x}$ and a domain indicator $d \in \mathcal{D}$. The corresponding label $y(\boldsymbol{x}, d) \in \{0, 1\}$ indicates a user's positive interaction in domain d.

In real-world recommendation platforms, the number of domains D is often very large, and the sample size varies greatly among domains. We focus on cases where $D > 20$. There exist domains $d_1, d_2 \in \mathcal{D}$ with a substantial disparity in their sample sizes, as expressed by the equation:

$$|\{(\boldsymbol{x}, d, y)|d = d_1\}| \gg |\{(\boldsymbol{x}, d, y)|d = d_2\}|. \tag{1}$$

The optimization objective is formulated as follows:

$$\Theta = \arg\min_{\Theta} \sum_{d \in \mathcal{D}} \mathcal{L}(f(\boldsymbol{x}, d), y(\boldsymbol{x}, d)), \tag{2}$$

where f is the MDR model and $\mathcal{L}$ denotes the loss function,

2.2 Overall Framework

The AREAD framework consists of three main components:

• Hierarchical Expert Integration (HEI) is built upon the bottom Base Recommender, comprising multi-layer expert networks and gating units. It extracts and integrates domain knowledge of varying granularities to generate predictions, effectively reducing the parameter count by avoiding the need for separate output networks for each domain.

• Hierarchical Expert Mask Pruning (HEMP) iteratively prunes the masks for selecting domain-specific experts, thereby identifying the most effective knowledge transfer patterns (Algorithm 1). This process significantly reduces the time complexity of searching for the optimal transferable knowledge for the current domain from HEI.

• Popularity-based Counterfactual Augmenter operates under the counterfactual assumption, suggesting that interactions with unpopular items in major domains are likely to occur similarly in minor domains. This is based on the rationale that if a user engages with a less popular item, the driving force is likely rooted in genuine preference rather than conformity, and such genuine preference remains unchanged across different domains.

2.3 Hierarchical Expert Integration

Why HEI In MDR systems encompassing D domains, existing models typically adopt a multi-task learning framework, constructing D separate tower networks to compute domain-specific outputs. This architecture leads to a catastrophic linear increase in the size of model parameters as D grows. In the context of a large number of domains, we believe that clustering domains is an effective strategy to reduce maintenance costs and ensure recommendation performance. However, *how to measure the similarity between domains* and *how to conduct clustering* are key problems to consider. For the former, the common practice is calculating similarity based on the output loss or gradient information of each domain (Bai and Zhao 2022; Wang et al. 2023). For the latter, using clustering algorithms to treat each cluster as a single domain for multi-domain learning is a feasible

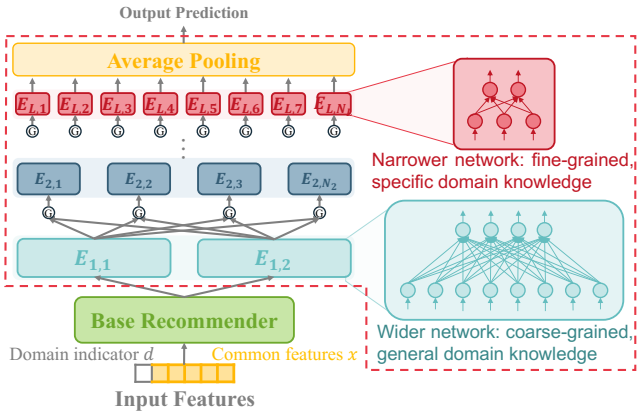

Figure 2: Hierarchical Expert Integration (HEI) uses multi-layer expert networks atop a base recommender to extract and integrate domain knowledge of varying granularities.

option. Nevertheless, these clustering methods exhibit three main drawbacks: 1) they are two-stage rather than end-to-end; 2) the calculation of similarity is inaccurate for minor domains due to insufficient optimization; and 3) intra-cluster domain individual characteristics are neglected.

Main Description To tackle these clustering challenges, we propose a novel Hierarchical Expert Integration (HEI) module (Figure 2), inspired by hierarchical clustering. HEI is comprised of L layers of expert networks and corresponding gating networks, in which each expert is implemented with a two-layer Multi-Layer Perceptron (MLP) and ReLU as the activation function. At the first level, HEI employs a small number of expert networks for coarse-grained clustering, where each expert captures knowledge of similar domains. To address more subtle differences within intra-cluster domains, we introduce higher layers of experts. As the number of layers increases, so does the number of experts per layer, yet still remaining significantly lower than the total number of domains D. These finer experts capture less knowledge, thus each network is structured to be narrower, requiring a reduced parameter count. Even with its multi-layered structure, HEI maintains a lower parameter count than assigning individual towers for each domain.

The hierarchical structure of HEI ensures the extraction of knowledge across dozens of domains at varying granularities. The narrower experts in the higher layer support the adaptive, domain-specific learning of minor domains while minimizing interference from major domains. Moreover, the fully connected arrangement of expert networks between layers, combined with the subsequent Hierarchical Expert Mask Pruning, provides a more flexible exploration for learning each domain's transfer pattern.

2.4 Hierarchical Expert Mask Pruning

Why HEMP For architecture like HEI with multiple experts, traditional multi-domain recommendation algorithms often explicitly dictate how parameters are shared, designating certain experts as domain-shared and others as domain-specific (Sheng et al. 2021; Zhou et al. 2023). However, for

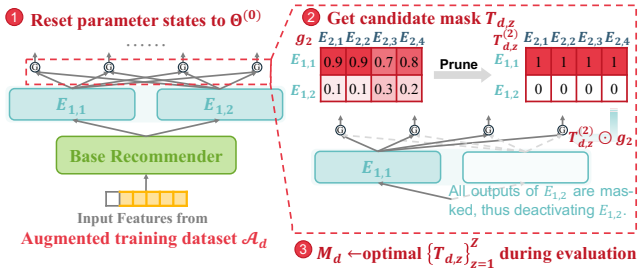

Figure 3: Hierarchical Expert Mask Pruning (HEMP) generates and iteratively prunes to select domain-specific experts.

a large number of domains, this binary division of domain knowledge into shared or specific falls short of capturing the complexities of knowledge transfer. Research by Standley et al. (Standley et al. 2020) supports this, exploring the relationship between jointly trained tasks and searching for the best way to split five example tasks into groups for multi-task learning in experiments. This study reveals that the knowledge gained from jointly training certain task combinations can be more beneficial than completely shared knowledge across all tasks. Such insight is crucial for multi-domain learning involving dozens of domains, highlighting the importance of determining *which domains should be jointly learned together and which should not*.

Directly searching for the optimal domains combination for joint training is practically infeasible due to high complexity. Existing methods rely on gate networks, attention mechanisms, or hyper-networks to adaptively learn the importance of experts, an assumption that may be overly idealistic, especially with imbalanced data volumes across domains. AESM^2 (Zou et al. 2022) provides an automatic expert selection framework. However, the greedy strategy in AESM^2, focusing on immediate layer-specific optimizations, may not always lead to globally optimal results because of the hierarchical knowledge transfer between layers.

Main Description Inspired by the Lottery Ticket Hypothesis (Frankle and Carbin 2019), we propose the Hierarchical Expert Mask Pruning (HEMP), a flexible architecture for selecting domain-specific experts (Figure 3). The lottery hypothesis suggests that within a randomly initialized, dense neural network, there exists a subnetwork (referred to as "winning tickets") that, if trained in isolation, can match the test accuracy of the original network in no more training iterations. Specifically, the winning tickets are identified via iterative pruning. Following this hypothesis, we posit that optimal networks of domain-specific experts exist within the HEI framework, and we utilize HEMP to generate hierarchical selection masks for experts specific to each domain. For domain d, its mask $M_d = [M_d^{(l)} \in \{0,1\}^{N_{l-1} \times N_l}$ for $l = 2, \cdots, L]$ controls the use of the gating weight across hierarchical experts layers, with N_l denoting the expert number in the l-th layer of HEI. Before the application of domain masks, the output of the n-th expert in the l-th layer, denoted $e_{l,n}$, is defined as:

$$e_{l,n} = \text{MLP}_{\Theta_{l,n}}(\sum_{i=1}^{N_{l-1}} g_{l,n}[i] e_{l-1,i}), \tag{3}$$

Algorithm 1 Hierarchical Expert Mask Pruning

Input: AREAD framework with parameter Θ; pruning rate α; initial sparsity threshold S_0; minimal sparsity S; augmented training datasets for D domains $\{\mathcal{A}_1, \cdots, \mathcal{A}_D\}$; training datasets $\mathcal{T}$
Output: Optimized AREAD model; updated masks for each domain $\{M_d\}_{d \in \mathcal{D}}$

Warm up AREAD with training data $\mathcal{T}$.
while AREAD has not converged **do**
  if condition to update domain masks is met **then**
    $\Theta^{(0)} \leftarrow$ current parameter state of AREAD.
    for $d \in \mathcal{D}$ **do**
      for $z \leftarrow 1$ to Z **do**
        Reset Θ to $\Theta^{(0)}$
        Randomly initialize candidate mask $T_{d,z}$ based on gate values and initial sparsity threshold S_0.
        for k steps **do**
          Train AREAD with data sample from augmented training dataset. $\mathcal{A}_d$ under the updating learning rate lr_u.
          Prune $\alpha\%$ of gates with the lowest magnitudes from $[T_{d,z}^{(l)} \odot g_l, \text{ for } l = 2, \ldots, L]$.
          if $\frac{\|T_{d,z}\|_0}{\sum_{l=2}^{L} N_{l-1} N_l} \leq S$ **then**
            break.
        Update M_d from $\{T_{d,z}\}_{z=1}^{Z}$
    Reset Θ to $\Theta^{(0)}$
  else
    Train AREAD with $\{M_d \text{ for } d \in \mathcal{D}\}$ and data sampled from $\mathcal{T}$ under the learning rate lr.

where $g_{l,n} \in [0,1]^{N_{l-1}}$ is the corresponding gating network output, and $g_{l,n}[i]$ is its i-th element, indicating the importance score for the i-th expert in the previous layer. With domain masks applied, the output is reformulated as:

$$e_{l,n} = \mathrm{MLP}_{\Theta_{l,n}} \left(\frac{\sum_{i=1}^{N_{l-1}} M_d^{(l)}[i,n]\, g_{l,n}[i]\, e_{l-1,i}}{\sum_{i=1}^{N_{l-1}} M_d^{(l)}[i,n] g_{l,n}} \right). \quad (4)$$

In particular, with HEMP, the training of hierarchical experts encompasses two distinct phases: *Warm-Up* and *Training with Mask*. During the Warm-Up phase, the model undergoes preliminary training on the initialized hierarchical experts without the use of domain masks, to allow initial adaptation to data characteristics. In the subsequent Training with Mask phase, HEMP is applied independently to each domain to derive specific masks. This involves generating a set of Z candidate temporary masks $T_{d,z}$ with initial sparsity S_0 for each domain and iteratively pruning these masks to achieve a certain sparsity threshold S or a certain number of iterations. For candidate mask initialization, the vast space of potential random masks necessitates a balance between randomness and training efficiency. To achieve this, the initialization space is narrowed down based on previous gate values. Specifically, we calculate the average value of gate weights obtained during previous training steps on domain data, retain the top S_0 percent of largest gates, and further introduce randomness by inverting the masking state of some gates to generate an initial candidate mask for each domain. After these candidate masks are initialized, pruning occurs iteratively during training. α percent of the remaining gates with the lowest values are pruned from $[T_{d,z}^{(l)} \odot g_l \text{ for } l = 2, \cdots, L]$, where $g_l = [g_{l,1}, g_{l,2}, \cdots, g_{l,N_{l-1}}]$ and $\odot$ denotes element-wise multiplication. Upon acquiring Z candidate masks $\{T_{d,z}\}_{z=1}^{Z}$, to select a single mask for subsequent training from these multiple candidates, we adhere to a straightforward principle: choosing the mask that demonstrates the best performance on domain data sampled from the training dataset $\mathcal{T}$. Note that all candidate masks for all domains start from a uniform parameter state for training and pruning to adhere to the lottery hypothesis. The process of searching for hierarchical expert selection masks is detailed in Algorithm 1.

2.5 Popularity-based Counterfactual Augmentation

During the training of candidate masks in HEMP, minor domains struggle to support training for k steps due to limited data, and repetitive sampling has little benefit in uncovering the characteristics of these domains. Inspired by Zheng et al.(Zheng et al. 2021), we analyze the causal relationship between users' real interests and popularity, and on this basis, we perform counterfactual data augmentation(Ying et al. 2023; Chen et al. 2024a,b). For a popular item, a user might click on it simply because it has been clicked by many others, as seen on e-commerce platforms where items are often displayed with their sales figures. These interactions are primarily driven by user conformity rather than genuine interest. As a crucial factor for decision making, conformity describes how users tend to follow other people (Zheng et al. 2021). Hence, we can break down the observed interactions into two user-side factors: *interest* and *conformity*. Their relationship can be depicted using a collider structure in a causal graph, expressed as $interest \rightarrow interact \leftarrow conformity$, meaning a positive interaction may result from either or both causes of interest and conformity. Based on this causal relationship, we propose the following assumption:

Assumption 1: If a user has a positive interaction with a unpopular item, it is highly likely due to the genuine interest rather than the conformity.

This assumption, based on collider bias or the "explain-away" effect (Pearl and Mackenzie 2018), posits that a user does not need both conformity and interest to make a positive interaction; one is sufficient. Thus, with a positive outcome $y = 1$, conformity and interest are spuriously negatively related. Consequently, when interacting with a unpopular item, the level of conformity is low, making it more likely that the interaction is driven by genuine preference, as illustrated in Figure 4. Additionally, we assume:

Assumption 2: A user's genuine interests are consistent across domains.

Based on these, we derive a corresponding corollary to generate the augmented dataset $\{\mathcal{A}_1, \cdots, \mathcal{A}_D\}$:

Corollary 1: If a user u has a positive interaction $y = 1$ with a unpopular item i in a major domain, then it is also likely to occur in a minor domain:

$$y(\boldsymbol{x}, d \in \mathcal{D}_a | p(i) < \rho) = 1 \implies y(\boldsymbol{x}, d' \in \mathcal{D}_b | p(i) < \rho) = 1, \quad (5)$$

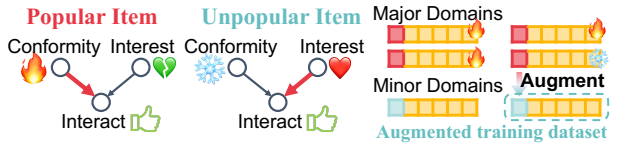

Figure 4: Popularity-based Counterfactual Augmenter utilizes counterfactual reasoning to infer genuine user interest across domains, augmenting interactions with unpopular items in major domains to minor domains.

where $\mathcal{D}_a$ represents the set of major domains and $\mathcal{D}_b$ represents the set of minor domains. $p(i)$ means the popularity of i correspond to $\boldsymbol{x}$ and ρ is a certain popularity threshold.

For instance, a user purchasing niche philosophical books is likely to enjoy similarly niche philosophical art films. Although transferring items across domains can introduce noise, the augmentation's benefits significantly outweigh this noise at controlled augmentation ratios, as demonstrated in our hyper-parameter study. The strength of our counterfactual augmenter lies in its reliance on straightforward yet highly reasonable assumptions, enabling rapid augmentation implementation even in large-scale datasets.

2.6 Model Optimization

MMoE (Ma et al. 2018a) serves as the base recommender system, facilitating shared feature interactions through embeddings corresponding to the input features. During the Warm-Up phase, data samples drawn from the training dataset are processed through both the foundational MMoE layer and the subsequent HEI layer, ultimately producing N_L outputs. The average of these N_L outputs is taken as the predicted value for the data sample:

$$\mathcal{L}(f(\boldsymbol{x},d),y(\boldsymbol{x},d)) = \mathcal{L}(\text{Avg}\{\sigma(e_{L,i})\}_{i=1}^{N_L}, y(\boldsymbol{x},d)), \quad (6)$$

where σ denotes the Sigmoid function. In the Training with Mask phase, for data (x,d), the number of outputs calculated under the hierarchical expert selection masks M_d is denoted by $|\mathcal{K}(M_d)|$, where:

$$\mathcal{K}(M_d) = \{j \in \{1,\cdots,N_L\} \mid \text{any}(M_d^{(L)}[:,j] > 0)\} \quad (7)$$

To ensure that experts with smaller outputs receive adequate optimization, a strategy akin to Bagging (Kuncheva 2014) in ensemble learning is adopted. During training, the loss between each output's prediction and the actual value is individually computed for gradient back-propagation:

$$\mathcal{L}(f(\boldsymbol{x},d),y(\boldsymbol{x},d)) = \sum_{j\in\mathcal{K}(M_d)} \mathcal{L}(\sigma(e_{L,j}), y(\boldsymbol{x},d)). \quad (8)$$

At inference, the outputs are averaged:

$$f(\boldsymbol{x},d) = \text{Avg}\{\sigma(e_{L,j})\}_{j\in\mathcal{K}(M_d)}. \quad (9)$$

The approach for utilizing candidate masks follows similarly, by substituting M_d with $T_{d,z}$. In our experiments, the loss function $\mathcal{L}(\cdot)$ is set to binary cross-entropy loss.

3 Experiments

3.1 Experimental Setup

Datasets We select two widely recognized datasets for our experiments: the **Amazon dataset** (Ni, Li, and McAuley 2019) and the **AliCCP dataset** (Ma et al. 2018b), organized by item category to define domains, as shown in Table 1.

Datasets	Amazon	AliCCP
#Samples	17,664,862	4,454,814
#Positive samples	8,790,575	329,416
#Domains	25	30
Majority ratio	17%	60.51%
#Minor domains	12	23
Augmentation ratio	10%	10%

Table 1: Statistics of the datasets. The "Majority ratio" refers to the sample ratio in the largest domain, and "Minor domains" represent domains with less than 2% of the samples.

Baselines We compare AREAD against state-of-the-art single-domain and multi-domain models. **(1) General Recommendation Models**: **DeepFM** (Guo et al. 2017), **DCN** (Wang et al. 2017), **AutoInt** (Song et al. 2019), and **DCNv2** (Wang et al. 2021). **EPNet-S** uses EPNet (Chang et al. 2023) for personalized input embedding selection and shares one top DNN tower across domains. The **Isolated** model, independently training and testing within individual domains, optimizes each domain's model and hyperparameters through grid search. **(2) Multi-Domain Recommendation Models**: **MMoE** (Ma et al. 2018a), **PLE** (Tang et al. 2020), **STAR** (Sheng et al. 2021), **AdaSparse** (Yang et al. 2022), **HiNet** (Zhou et al. 2023), **EPNet**, **PEPNet** (Chang et al. 2023), **ADL** (Li et al. 2023), and **MAMDR** (Luo et al. 2023). Due to the high cost of training separate-tower models across many domains, we pre-cluster domains and conduct multi-domain learning within each cluster as a single domain. ADL and MAMDR, which has validated scalability, eliminate the need for pre-clustering.

Pre-Clustering Domains We first train a single-domain DCN model using a combined dataset from all domains, chosen for its superior performance. Next, we record the distribution of losses for each domain on the test dataset and calculate the Kullback-Leibler (KL) divergence of these loss distributions to determine the distance between domains. Finally, we perform K-Means clustering based on the distances computed among domains. For both datasets, the number of clusters is set to 3. AREAD does not require pre-clustering.

Metrics We utilize **AUC** for evaluation, a standard metric for binary classification tasks. In MDR, where item interactions within each domain are independently assessed, aggregate AUC provides limited insights. Thus, we use **Domain-AUC**, which measures the AUC for each domain separately, averaged and weighted by their sample sizes. Additionally, we assess performance with **Major5AUC**, **Minor10AUC**, and **Minor5AUC** to evaluate effectiveness in data-rich and data-sparse environments; these represent the weighted average AUCs of the largest and smallest domains, respectively. Notably, prior research (Song et al. 2019; Jia et al. 2024) indicates that even a minor improvement of **0.001** in AUC can produce significant positive benefits online.

Implementation Details We use Adam optimizer with learning rates [$5e^{-4}$, $1e^{-3}$, $1.5e^{-3}$, $2e^{-3}$, $3e^{-3}$] and batch sizes [1024, 2048, 4096] optimized through grid search. Regularization coefficients are set at $1e^{-5}$. In AREAD, the

Method	Amazon					AliCCP				
	AUC	DomainAUC	Major5AUC	Minor10AUC	Minor5AUC	AUC	DomainAUC	Major5AUC	Minor10AUC	Minor5AUC
DeepFM	0.7036	0.7111	0.7130	0.7251	0.7026	0.6003	0.5853	0.5869	0.5516	0.5669
DCN	0.7040	0.7114	0.7135	0.7256	0.6962	0.6084	0.5939	0.5952	0.5556	0.5709
AutoInt	0.7039	0.7112	0.7132	0.7256	0.6954	0.6080	0.5930	0.5945	0.5520	0.5701
DCNv2	0.7040	0.7114	0.7134	0.7262	0.7009	0.6082	0.5939	0.5954	0.5540	0.5730
EPNet-S	0.7033	0.7110	0.7131	0.7267	0.7026	0.6083	0.5946	0.5961	0.5531	0.5705
Isolated	-	**0.7137**	**0.7150**	0.7229	0.7199	-	0.5851	0.5907	0.5536	0.5628
MMoE	0.7038	0.7113	0.7136	0.7255	0.7001	0.6091	0.5947	0.5963	0.5552	0.5733
PLE	0.7025	0.7099	0.7119	0.7248	0.6995	0.6090	0.5946	0.5959	0.5563	0.5765
STAR	0.7032	0.7032	0.7029	0.7042	0.6949	0.6075	0.6075	0.6078	0.6071	0.6036
AdaSparse	0.7053	0.7102	0.7136	0.7258	0.6934	0.6074	0.5912	0.5927	0.5519	0.5695
HiNet	0.7049	0.7109	0.7121	0.7247	0.6998	0.6075	0.5915	0.5928	0.5545	0.5738
EPNet	0.7030	0.7106	0.7128	0.7251	0.6976	0.6084	0.5942	0.5957	0.5559	0.5769
PEPNet	0.6995	0.7062	0.7073	0.7178	0.6891	0.6083	0.5941	0.5954	0.5584	0.5806
ADL	0.7033	0.7109	0.7130	0.7240	0.6991	0.6086	0.6086	0.6087	0.6151	0.6131
MAMDR	0.6898	0.7079	0.7066	0.7194	0.6982	0.5957	0.5869	0.5869	0.5572	0.5691
AREAD	**0.7120***	0.7131	0.7147	**0.7298***	**0.7218***	**0.6122***	**0.6170***	**0.6165***	**0.6200***	**0.6264***

Table 2: Performance comparison of different methods using five multi-domain metrics on Amazon and AliCCP datasets. Best and second-best results are highlighted in bold and underlined, respectively. $*$ indicates statistically significant differences (p-value < 0.01) from the second-best result. Results are averaged over five runs.

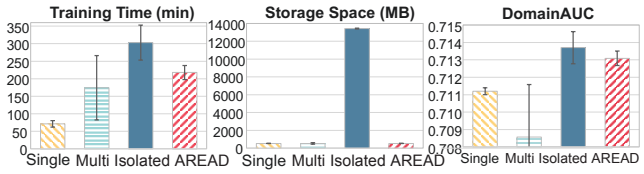

Figure 5: Comparison of performance, storage space, and training time across Single-domain, Multi-domain models, Isolated method, and AREAD on the Amazon dataset.

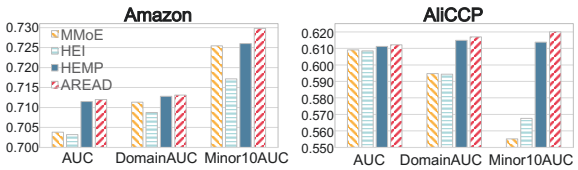

Figure 6: Effectiveness of the sub-modules HEI, HEMP, and Counterfactual Augmentater in the AREAD framework.

Warm-Up phase trains 100 batches. During the Train with Mask phase, masks update every 2000 batches, exploring $Z = 10$ candidate masks. Each mask $T_{d,z}$ is trained for $k = 5$ batches on augmented data $\mathcal{A}_d$ of domain d. Initial mask sparsity is $S_0 = 0.7$, targeting $S = 0.4$ with a pruning ratio of $\alpha = 0.05$ per iteration.

3.2 Overall Performance

Table 2 shows the experimental results across dozens of domains in two datasets. Our observations include:

(1) **The proposed method consistently achieves the best performance across general metrics.** AREAD achieves AUC improvements of 6.7‰ on Amazon and 3‰ on AliCCP, both with p-values < 0.01. For the crucial DomainAUC metric, AREAD shows improvements of 2‰ on Amazon and 8.4‰ on AliCCP, compared to the best baseline trained on mixed-domain data. This underscores its ability to effectively leverage domain characteristics in large-scale recommendation scenarios, leading to marked performance gains across all domains.

(2) **Our model delivers outstanding results at acceptable maintenance costs.** On the Amazon dataset, AREAD outperforms all mixed-domain methods, though it doesn't exceed the "Isolated" method. However, the Isolated method incurs impractically high training and maintenance costs (see Figure 5), and performs less effectively as domain similarities increase (see AliCCP results in Table 2). AREAD strikes an optimal balance between performance and cost, leading among existing models and matching the Isolated method's results, even under significant domain variability.

(3) **Our model demonstrates exceptional performance on minor domains.** In the Amazon dataset, AREAD significantly enhances Minor10AUC by 3‰ and Minor5AUC by 2‰. In the AliCCP dataset, improvements are 5‰ and 13‰, respectively. These gains stem from the HEI and HEMP strategies in AREAD, which reduce interference from major domains and boost knowledge transfer. Counterfactual augmentation also mitigates data scarcity in minor domains, further enhancing performance.

3.3 Ablation Study

To further validate the effectiveness of the sub-modules in the AREAD framework, we evaluated the performance of the base recommender MMoE, the models with the HEI module, with the HEMP, and the complete model on both datasets, as shown in Figure 6. Results indicate a slight performance decline when integrating the HEI module compared to the base recommender alone, confirming our previous hypothesis that without the use of expert selection masks, even a well-designed hierarchical expert architecture cannot ideally adapt and learn knowledge transfer patterns

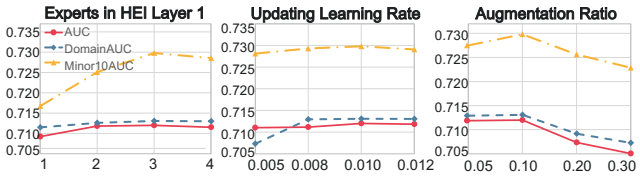

Figure 7: Hyperparameter study on Amazon dataset.

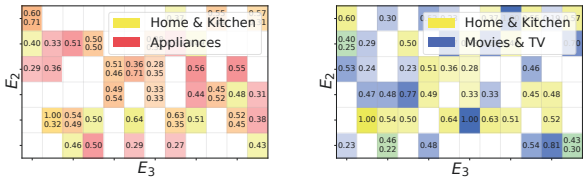

Figure 8: Expert utilization in HEI of AREAD, with the left OR at 0.405 ("Home & Kitchen" vs. "Appliances"), and the right OR at 0.065 ("Home & Kitchen" vs. "Movies & TV").

across dozens of domains. Integrating HEMP significantly enhances both AUC and DomainAUC, validating that domain masks substantially facilitate beneficial domain knowledge transfer. Further incorporation of the Popularity-based Counterfactual Augmentation results in notable improvements in the Minor10AUC metric, demonstrating that counterfactual data augmentation is particularly advantageous for the optimization of minor domains.

3.4 Hyper-parameter Study

To investigate the impact of various configurations within AREAD, we explored different hyperparameters: the number of experts (N_1) in the 1st layer of HEI, the updating learning rate (lr_u) for model updates under candidate masks as specified in Algorithm 1, and the percentage of augmented data (r_{aug}). Results on the Amazon dataset are shown in Figure 7. We maintain three layers in HEI, with each layer doubling the experts of the previous one. With fewer experts, AUC and DomainAUC performed well as the major domains dominated the optimization of HEI; however, performance in minor domains declined due to a lack of sufficient specific experts to integrate useful knowledge. The optimal lr_u is 0.01, with the metrics remaining relatively stable around this optimum. Lowering lr_u diminishes the model's capacity to quickly discern the effectiveness of candidate masks, adversely affecting DomainAUC. The augmentation ratio r_{aug} directly influences the extent to which augmented data can affect model performance. Increasing r_{aug} from 0.05 to 0.1 improves Minor10AUC without affecting AUC and DomainAUC. However, further increasing r_{aug} introduces a substantial amount of noisy augmented samples, significantly impairing model performance.

3.5 Mask Analysis

To explore AREAD's expert integration across domains, we analyze expert utilization on the Amazon dataset. AREAD uses three layers with $[3, 6, 12]$ experts each in HEI, with a target sparsity of $S = 0.4$ in HEMP. After training on 25 domains, we focus on "Home & Kitchen" (HK), "Appliances" (AP), and "Movies & TV" (MT), which constitute 0.137, 0.018, and 0.004 of the dataset's total volume, respectively.

Figure 8 shows the average gate weights between the second and third layers for the domain pairs "HK vs. AP" and "HK vs. MT". We calculate the Mask Overlap Ratio (OR) to analyze domain relatedness (Sun et al. 2020): $\mathrm{OR}\left(M_{d_1}^{(l)}, M_{d_2}^{(l)}\right) = \left\|M_{d_1}^{(l)} \cap M_{d_2}^{(l)}\right\|_0 / \left\|M_{d_1}^{(l)} \cup M_{d_2}^{(l)}\right\|_0$. The OR for "HK vs. AP" is 0.405, notably higher than the random overlap of 0.25, highlighting similar functional features. In contrast, the OR for "HK vs. MT" is only 0.065, indicating their relative non-relevance. Interestingly, despite the high similarity, HK and AP are often assigned to different groups in the pre-clustering stage. This underscores AREAD's ability to capture complex knowledge transfer patterns tailored to specific domain characteristics, which traditional methods might overlook. Furthermore, despite data volume disparities in the two smaller domains, the interpretability of expert utilization confirms both the reliability of Popularity-based Counterfactual Augmentation and AREAD's robust learning in minor domains.

4 Related Work

Multi-domain recommendation systems aim to utilize diverse domain data to boost performance across all included domains. Inspired by multi-task learning (Ma et al. 2018a; Tang et al. 2020), models often feature a domain-general base and domain-specific tower networks (Ma et al. 2018a; Tang et al. 2020; Sheng et al. 2021; Shen et al. 2021; Zou et al. 2022; Zhou et al. 2023; Chang et al. 2023). This architecture is supplemented by gating (Ma et al. 2018a; Zou et al. 2022; Chang et al. 2023), attention mechanisms (Shen et al. 2021), hypernetworks (Chang et al. 2023; Liu et al. 2023), and dynamic weight parameters (Zhang et al. 2022a; Yan et al. 2022). Another approach, "Pre-Training + Fine-Tuning" (Gu et al. 2021; Zhang et al. 2022b; Luo et al. 2023), enables cross-domain knowledge transfer through temporal parameter inheritance. Nonetheless, these models typically handle few domains and struggle to scale with the increasing number of real-world domains. A recent study (Li et al. 2023) has attempted to cluster scenes into groups for multi-domain adaptation, but this often results in suboptimal performance due to overlooked intra-cluster variations.

5 Conclusion

In this paper, we explore the issue of multi-domain recommendation across dozens of domains, and propose the Adaptive REcommendation for All Domains with counterfactual augmentation (AREAD) framework. AREAD employs Hierarchical Expert Integration to capture domain transfer knowledge at varying granularities within hierarchical expert networks. Building on this, the framework utilizes the Hierarchical Expert Mask Pruning algorithm to learn knowledge transfer patterns across numerous domains. Moreover, Popularity-based Counterfactual Augmentation is adopted to augment data for minor domains. Finally, experiments conducted on two public datasets, each encompassing over twenty domains, confirm the effectiveness of our approach.

Acknowledgements

The research work is supported by the National Key Research and Development Program of China under Grant No. 2021ZD0113602, the National Natural Science Foundation of China under Grant Nos. 62176014 and 62276015, and the Fundamental Research Funds for the Central Universities.

References

Bai, G.; and Zhao, L. 2022. Saliency-regularized deep multi-task learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 15–25.

Chang, J.; Zhang, C.; Hui, Y.; Leng, D.; Niu, Y.; Song, Y.; and Gai, K. 2023. Pepnet: Parameter and embedding personalized network for infusing with personalized prior information. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3795–3804.

Chen, W.; Wu, Y.; Zhang, Z.; Zhuang, F.; He, Z.; Xie, R.; and Xia, F. 2024a. FairGap: Fairness-aware Recommendation via Generating Counterfactual Graph. *ACM Transactions on Information Systems*, 42(4): 1–25.

Chen, W.; Yuan, M.; Zhang, Z.; Xie, R.; Zhuang, F.; Wang, D.; and Liu, R. 2024b. FairDgcl: Fairness-aware Recommendation with Dynamic Graph Contrastive Learning. *arXiv preprint arXiv:2410.17555*.

Frankle, J.; and Carbin, M. 2019. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. In *International Conference on Learning Representations*.

Gan, C.; Huang, B.; Hu, B.; Ma, J.; Zhang, Z.; Zhou, J.; Zhang, G.; and Zhong, W. 2024. PEACE: Prototype lEarning Augmented transferable framework for Cross-domain rEcommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 228–237.

Gu, Y.; Bao, W.; Ou, D.; Li, X.; Cui, B.; Ma, B.; Huang, H.; Liu, Q.; and Zeng, X. 2021. Self-Supervised Learning on Users' Spontaneous Behaviors for Multi-Scenario Ranking in E-commerce. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 3828–3837.

Guan, R.; Pang, H.; Giunchiglia, F.; Liang, Y.; and Feng, X. 2022. Cross-Domain Meta-Learner for Cold-Start Recommendation. *IEEE Transactions on Knowledge and Data Engineering*.

Guo, H.; Tang, R.; Ye, Y.; Li, Z.; and He, X. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 1725–1731.

Jia, P.; Wang, Y.; Lin, S.; Li, X.; Zhao, X.; Guo, H.; and Tang, R. 2024. D3: A Methodological Exploration of Domain Division, Modeling, and Balance in Multi-Domain Recommendations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 8553–8561.

Kuncheva, L. I. 2014. *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons.

Li, J.; Zheng, H.; Liu, Y.; Lu, M.; Wu, L.; and Hu, H. 2023. ADL: Adaptive Distribution Learning Framework for Multi-Scenario CTR Prediction. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, 1786–1790. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394086.

Liu, Q.; Zhou, Z.; Jiang, G.; Ge, T.; and Lian, D. 2023. Deep Task-specific Bottom Representation Network for Multi-Task Recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 1637–1646.

Luo, L.; Li, Y.; Gao, B.; Tang, S.; Wang, S.; Li, J.; Zhu, T.; Liu, J.; Li, Z.; and Pan, S. 2023. MAMDR: A model agnostic learning framework for multi-domain recommendation. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, 3079–3092. IEEE.

Ma, J.; Zhao, Z.; Yi, X.; Chen, J.; Hong, L.; and Chi, E. H. 2018a. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 1930–1939.

Ma, X.; Zhao, L.; Huang, G.; Wang, Z.; Hu, Z.; Zhu, X.; and Gai, K. 2018b. Entire space multi-task model: An effective approach for estimating post-click conversion rate. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 1137–1140.

Ni, J.; Li, J.; and McAuley, J. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, 188–197.

Pearl, J.; and Mackenzie, D. 2018. *The book of why: the new science of cause and effect*. Basic books.

Shen, Q.; Tao, W.; Zhang, J.; Wen, H.; Chen, Z.; and Lu, Q. 2021. Sar-net: a scenario-aware ranking network for personalized fair recommendation in hundreds of travel scenarios. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 4094–4103.

Sheng, X.-R.; Zhao, L.; Zhou, G.; Ding, X.; Dai, B.; Luo, Q.; Yang, S.; Lv, J.; Zhang, C.; Deng, H.; et al. 2021. One model to serve all: Star topology adaptive recommender for multi-domain ctr prediction. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 4104–4113.

Song, W.; Shi, C.; Xiao, Z.; Duan, Z.; Xu, Y.; Zhang, M.; and Tang, J. 2019. Autoint: Automatic feature interaction learning via self-attentive neural networks. In *Proceedings of the 28th ACM international conference on information and knowledge management*, 1161–1170.

Standley, T.; Zamir, A.; Chen, D.; Guibas, L.; Malik, J.; and Savarese, S. 2020. Which tasks should be learned together in multi-task learning? In *International Conference on Machine Learning*, 9120–9132. PMLR.

Sun, T.; Shao, Y.; Li, X.; Liu, P.; Yan, H.; Qiu, X.; and Huang, X. 2020. Learning sparse sharing architectures for

multiple tasks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 8936–8943.

Tang, H.; Liu, J.; Zhao, M.; and Gong, X. 2020. Progressive layered extraction (ple): A novel multi-task learning (mtl) model for personalized recommendations. In *Proceedings of the 14th ACM Conference on Recommender Systems*, 269–278.

Wang, C.; Senjie; Shuli; Wen, S.; Yin, H.; and Xiao, X. 2023. Exploration and Practice of Multi-Scenario Modeling at Meituan. Technical report, Meituan. Accessed: 2024/02/01.

Wang, R.; Fu, B.; Fu, G.; and Wang, M. 2017. Deep & Cross Network for Ad Click Predictions. In *Proceedings of the ADKDD'17*, ADKDD'17. New York, NY, USA: Association for Computing Machinery. ISBN 9781450351942.

Wang, R.; Shivanna, R.; Cheng, D.; Jain, S.; Lin, D.; Hong, L.; and Chi, E. 2021. Dcn v2: Improved deep & cross network and practical lessons for web-scale learning to rank systems. In *Proceedings of the web conference 2021*, 1785–1797.

Wang, Y.; Guo, H.; Chen, B.; Liu, W.; Liu, Z.; Zhang, Q.; He, Z.; Zheng, H.; Yao, W.; Zhang, M.; et al. 2022. Causalint: Causal inspired intervention for multi-scenario recommendation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4090–4099.

Yan, B.; Wang, P.; Zhang, K.; Li, F.; Deng, H.; Xu, J.; and Zheng, B. 2022. Apg: Adaptive parameter generation network for click-through rate prediction. *Advances in Neural Information Processing Systems*, 35: 24740–24752.

Yang, X.; Peng, X.; Wei, P.; Liu, S.; Wang, L.; and Zheng, B. 2022. Adasparse: Learning adaptively sparse structures for multi-domain click-through rate prediction. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 4635–4639.

Ying, Y.; Zhuang, F.; Zhu, Y.; Wang, D.; and Zheng, H. 2023. Camus: attribute-aware counterfactual augmentation for minority users in recommendation. In *Proceedings of the ACM Web Conference 2023*, 1396–1404.

Zhang, Q.; Liao, X.; Liu, Q.; Xu, J.; and Zheng, B. 2022a. Leaving No One Behind: A Multi-Scenario Multi-Task Meta Learning Approach for Advertiser Modeling. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, 1368–1376. New York, NY, USA: Association for Computing Machinery. ISBN 9781450391320.

Zhang, W.; Zhang, P.; Zhang, B.; Wang, X.; and Wang, D. 2023. A Collaborative Transfer Learning Framework for Cross-domain Recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 5576–5585.

Zhang, Y.; Wang, X.; Hu, J.; Gao, K.; Lei, C.; and Fang, F. 2022b. Scenario-Adaptive and Self-Supervised Model for Multi-Scenario Personalized Recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 3674–3683.

Zhao, C.; Zhao, H.; He, M.; Zhang, J.; and Fan, J. 2023. Cross-domain recommendation via user interest alignment. In *Proceedings of the ACM Web Conference 2023*, 887–896.

Zheng, Y.; Gao, C.; Li, X.; He, X.; Li, Y.; and Jin, D. 2021. Disentangling user interest and conformity for recommendation with causal embedding. In *Proceedings of the Web Conference 2021*, 2980–2991.

Zhou, J.; Cao, X.; Li, W.; Bo, L.; Zhang, K.; Luo, C.; and Yu, Q. 2023. Hinet: Novel multi-scenario & multi-task learning with hierarchical information extraction. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, 2969–2975. IEEE.

Zhu, Y.; Ge, K.; Zhuang, F.; Xie, R.; Xi, D.; Zhang, X.; Lin, L.; and He, Q. 2021. Transfer-meta framework for cross-domain recommendation to cold-start users. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1813–1817.

Zou, X.; Hu, Z.; Zhao, Y.; Ding, X.; Liu, Z.; Li, C.; and Sun, A. 2022. Automatic expert selection for multi-scenario and multi-task search. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1535–1544.