Stable and Adaptive Fusion for Multi-domain Multi-task Recommendation

Ke Fei, Da Luo, Kangyi Lin, Zibin Zhang, Jingjing Li*

Wechat, Tencent
Guangzhou, China
q191106702@gmail.com, {lodaluo, plancklin, bingoozhang}@tencent.com, lijin117@yeah.net

Abstract

Multi-Domain Multi-Task (MDMT) recommendation aims to provide personalized recommendations by leveraging information across multiple domains and tasks. However, existing methods often suffer from spurious correlations between irrelevant features and the target, leading to negative transfer. To address this, we propose a Stable and Adaptive Fusion (SAF) framework for MDMT recommendation. SAF introduces a weighted Hilbert-Schmidt Independence Criterion (HSIC) loss to decorrelate irrelevant features from the target, learning sample weights that promote stable (i.e., robust to spurious correlations) representations in both bottom and expert layers. We employ Random Fourier Features (RFF) to enable scalable computation of the HSIC loss. We further employ adaptive feature and expert gating to select these stable features, enabling the model to capture intricate cross-domain and cross-task dependencies. The learned sample weights are also used to reweight the MDMT loss during training. Experiments on large-scale datasets show that SAF outperforms state-of-the-art baselines by up to 2% in AUC. To facilitate further research, we release a new industrial dataset with 30 million interactions across 3 domains and 2 tasks, with 300 features.

Extended version —
https://github.com/q1179897215/SAF

1 Introduction

The proliferation of digital platforms has dramatically increased the volume and diversity of information available to users, spanning multiple content types (e.g., images, articles) and user engagement objectives (e.g., clicks, likes). This information deluge poses significant challenges for effective content discovery and personalization. Multi-Domain Multi-Task (MDMT) recommendations have emerged as a promising solution, aiming to deliver tailored suggestions by jointly modeling user preferences across different domains and tasks, thereby alleviating information overload (Alhijawi and Kilani 2020; Aceto, Persico, and Pescapé 2020).

Existing approaches to Multi-Domain (MDR), Multi-Task (MTR), and MDMT recommendation typically rely on parameter sharing to capture commonalities, while employing domain- or task-specific modules to model unique characteristics. Routing mechanisms, such as gating or expert selection, are then used to combine these representations. For instance, STAR (Sheng et al. 2021) utilizes shared-specific Multi-Layer Perceptron (MLP) with privatized normalization, and PLE (Tang et al. 2020) employs gating to route shared and task-specific experts. More recent models like M^3oE (Zhang et al. 2024) and PEPNet (Chang et al. 2023) extend these ideas to the MDMT setting.

However, a critical limitation of these methods is their inability to distinguish causally relevant features from those merely correlated in the training data, making them prone to **spurious correlations** that can cause **negative transfer** and degrade recommendation quality (Marx et al. 2005; Wang et al. 2019; Lake et al. 2017; Marcus 2018; Lopez-Paz et al. 2017; Arjovsky et al. 2019). This issue is especially pronounced in MDMT settings, where overlapping domains and tasks amplify such risks. To illustrate, Figure 1 shows that in our industrial dataset, user histories in domain A (e.g., articles) contain both relevant (item category) and irrelevant (language) features that are correlated. Baseline models tend to overemphasize the irrelevant feature (language) when recommending in domain B (e.g., pictures), leading to misaligned recommendations.

Moreover, in real-world recommender systems, users often have interactions in some domains or tasks but not others. Joint training can therefore cause users to inherit patterns from more active domains, which may not reflect their true preferences elsewhere. Such data sparsity and imbalance further complicate disentangling causal relationships from spurious ones, underscoring the need for methods that can actively disentangle and stabilize feature learning across domains and tasks for robust generalization.

To tackle this challenge, we propose a Stable and Adaptive Fusion (SAF) framework for MDMT recommendation, which explicitly mitigates spurious correlations during training. SAF introduces a two-pronged approach: (1) it learns sample weights by minimizing a weighted Hilbert-Schmidt Independence Criterion (HSIC) loss at both shared and expert layers, thereby decorrelating feature representations and promoting invariance to irrelevant factors; and (2) it employs adaptive gating mechanisms to selectively fuse stable features, enabling nuanced modeling of cross-domain

*Corresponding author

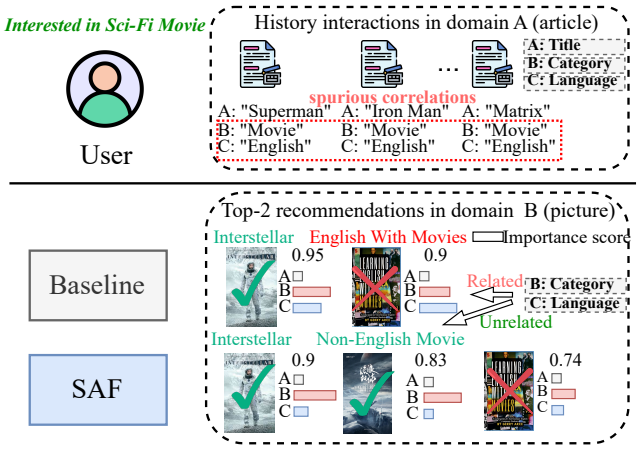

Figure 1: Illustration of spurious correlations in a real industrial dataset. In domain A, relevant features (e.g., item category) and irrelevant features (e.g., language) are statistically dependent. Baseline models trained on this data tend to overemphasize the irrelevant feature (blue bar) when making predictions in domain B, leading to recommendations that do not match user interests. In contrast, our proposed model reduces these spurious dependencies, assigns greater importance to the true causal feature in domain B (red bar), and generates recommendations that better align with user preferences.

and cross-task dependencies. To ensure scalability to large datasets, we approximate kernel computations using Random Fourier Features (RFF). The learned sample weights are further used to reweight the overall MDMT loss, reinforcing the learning of robust and generalizable representations.

Our main contributions are as follows:

- We introduce SAF, a novel framework for MDMT recommendation that explicitly addresses negative transfer by decorrelating spurious features and adaptively fusing stable representations.
- We demonstrate the effectiveness of SAF on both public and large-scale industrial datasets, achieving absolute AUC improvements of 1.5% and 2.5% over state-of-the-art baselines, respectively.
- We provide theoretical analysis and empirical evidence that SAF enhances feature stability and mitigates errors arising from spurious correlations.
- We release a new, large-scale industrial dataset with 30 million interactions and 300 features, establishing a challenging benchmark for future MDMT research.

2 Related Works

2.1 Multi-Domain and Multi-Task Recommendation

Multi-domain recommendation (MDR) and multi-task recommendation (MTR) aim to leverage both shared and domain/task-specific information to improve recommendation performance. The central challenge is to enable effective knowledge transfer while avoiding negative transfer caused by domain/task heterogeneity.

Early MDR approaches primarily rely on parameter sharing to transfer knowledge across domains (Dredze, Kulesza, and Crammer 2010; Joshi et al. 2012; Sheng et al. 2021; Li et al. 2023b), with domain-specific parameters capturing intra-domain nuances. These methods are typically categorized into hard sharing and soft sharing paradigms.

Hard sharing methods use fixed shared and domain-specific parameters, often within a multi-task learning framework (Zhang and Yang 2021). For instance, STAR (Sheng et al. 2021) constructs shared and specific MLPs with privatized normalization, and SARNet (Shen et al. 2021) employs shared and specific experts, domain-specific transformation layers, and gating mechanisms. However, hard sharing can lead to negative transfer when domains are not sufficiently related, as irrelevant information may be forced into shared representations.

Soft sharing methods, on the other hand, share generating modules such as cells or units and inject domain information to enhance flexibility. Hamur (Li et al. 2023b) utilizes a domain-shared hyper-network and domain-specific adapters, while ADIN (Jiang et al. 2022) introduces a domain interest adaptation layer for feature-level adaptation and sharing. Despite their flexibility, soft sharing methods can increase model complexity and still struggle to disentangle domain-specific from shared features, particularly in highly heterogeneous domains.

In the context of MTR, the focus is on improving task performance by mitigating task competition and domination. SharedBottom (Caruana 1997) employs a shared bottom layer and multi-tower structure, but fixed parameter sharing can cause negative transfer when tasks are dissimilar. More recent solutions are either architecture-based (Ma et al. 2018; Tang et al. 2020; Li et al. 2023a) or optimization-based (Yang et al. 2023; Yu et al. 2020). MMOE (Ma et al. 2018) uses shared experts and task-specific gates to reduce competition, while PLE (Tang et al. 2020) introduces both task-specific and shared experts. Optimization-based methods such as AdaTask (Yang et al. 2023) and PCGrad (Yu et al. 2020) adapt learning rates or gradients to avoid task domination and negative transfer.

Recently, unified models for multi-domain multi-task recommendation have been proposed to jointly model inter-domain and inter-task relations. M2M (Zhang et al. 2022) employs meta units to generate attention weights for different domains and tasks, while PEPNET (Chang et al. 2023) leverages both domain and personalized information using gates to dynamically scale features. M^{3}OE (Zhang et al. 2024) adopts a two-level structure, stacking domain-shared and domain-specific experts to capture complex relationships. However, most existing methods do not systematically address the challenge of spurious correlations and stable generalization across both domains and tasks, especially under data sparsity and imbalance.

2.2 Stable Learning

The correlation between relevant and irrelevant features can introduce spurious associations, leading to negative transfer and poor generalization. Stable learning aims to develop models that generalize well across domains or tasks by reducing the influence of such correlations. Typically, stable learning methods involve two main steps: independence-driven weight estimation and weighted model training (Xu et al. 2022).

Shen et al. (Shen et al. 2020) propose a sample reweighting approach to reduce collinearity among input variables, while StableNet (Zhang et al. 2021) extends stable learning to deep neural networks for out-of-distribution image classification. Zhang et al. (Zhang et al. 2023) apply stable learning to multi-modal recommendation, enabling the model to capture stable user preferences from heterogeneous information sources. Inspired by these works, we incorporate feature decorrelation insights from stable learning to mitigate negative transfer in multi-domain and multi-task recommendation settings.

3 Methodology

3.1 Problem Definition

Modern recommender systems must account for heterogeneous user behaviors—such as clicks, purchases, and likes—across multiple platforms or content categories, referred to as *domains*. In this work, we focus on the joint prediction of click-through rate (CTR) and conversion rate (CVR) across D domains, formulating the problem as a multi-domain, multi-task learning task. Here, the CVR task refers to user actions following a click, such as making a purchase, liking content, or subscribing.

Let $\mathcal{U} = \{u_1, \ldots, u_n\}$ be the set of users and $\mathcal{I} = \{i_1, \ldots, i_m\}$ the set of items. Suppose there are T user behavior prediction tasks (e.g., CTR, CVR) and D domains (e.g., different platforms or content types). For each domain $d \in \{1, \ldots, D\}$ and task $t \in \{1, \ldots, T\}$, we observe user-item interactions $(u, i) \in \mathcal{D}_d \subseteq \mathcal{U} \times \mathcal{I}$, each with a ground-truth label $y_{u,i}^{(t,d)} \in \{0, 1\}$ indicating whether the behavior of interest occurred. Our goal is to develop a unified model that, for any user-item pair (u, i), domain d, and task t, predicts the probability $\hat{y}_{u,i}^{(t,d)}$ of the corresponding user behavior. The model is trained jointly across all domains and tasks to leverage shared patterns while capturing domain- and task-specific variations.

3.2 Framework Overview

To address negative transfer from spurious correlations in MDMT recommendation, we propose the SAF framework, which combines stable feature learning with adaptive feature fusion. As shown in Figure 2, SAF integrates three components: (1) a joint MDMT loss for coordinated parameter optimization, (2) a feature decorrelation loss based on the Hilbert-Schmidt Independence Criterion (HSIC) with Random Fourier Features (RFF), and (3) a dynamic gating mechanism for expert routing and feature fusion.

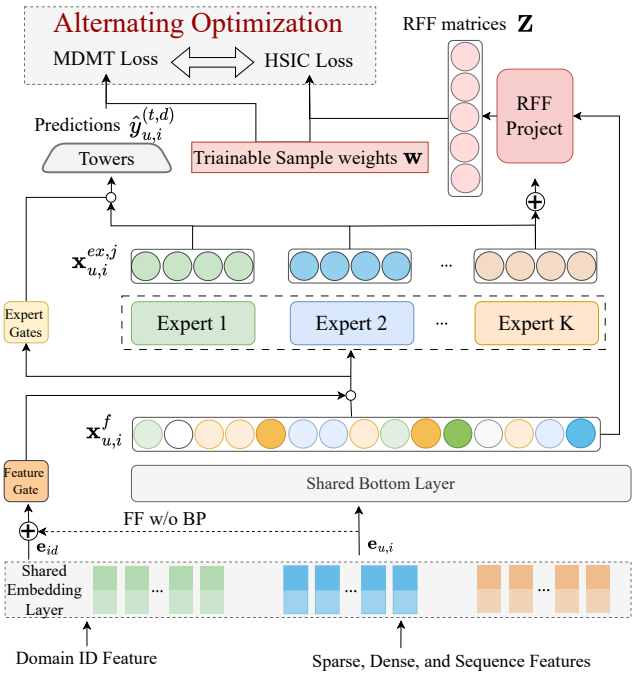

Figure 2: Illustration of the SAF framework for multi-domain multi-task learning. SAF integrates three components: (1) a multi-domain multi-task loss for joint parameter optimization, (2) a feature decorrelation loss based on HSIC with RFF, and (3) a dynamic gating mechanism for stable feature and expert routing. During training, SAF alternates between optimizing the MDMT and HSIC losses, while the gating mechanism dynamically routes features and experts for each domain and task, effectively mitigating negative transfer. $\oplus$ denotes vector concatenation and $\circ$ denotes element-wise multiplication.

The joint MDMT loss enables knowledge sharing across domains and tasks, while the RFF-based HSIC loss adaptively reweights samples and penalizes statistical dependencies between features, suppressing spurious correlations. The dynamic gating mechanism further routes decorrelated features and expert modules to each domain and task, capturing complex inter-domain and inter-task relationships.

A key innovation of SAF is its alternating optimization: the model iteratively minimizes the HSIC loss to refine sample weights and decorrelate features, then optimizes the MDMT loss to update model parameters. This ensures feature representations remain stable and robust, while knowledge transfer is both effective and selective. By explicitly combining stable learning with adaptive fusion, SAF overcomes the limitations of prior methods that rely only on implicit disentanglement of shared and task-specific parameters.

3.3 Stable Feature Learning

Spurious correlations between features can lead to negative transfer and poor generalization in MDMT settings. To mitigate this, we aim to reduce statistical dependencies among

Dataset	AUC for Each Domain and Task											Overall Performance			
	Ali-CCP					Industrial						Ali-CCP		Industrial	
	d1,ctr	d1,cvr	d2,ctr	d2,cvr	d3,ctr	d1,ctr	d1,cvr	d2,ctr	d2,cvr	d3,ctr	d3,cvr	AUC	Logloss	AUC	Logloss
MLP	0.6210	0.6153	0.6213	0.6251	0.5566	0.7120	0.6678	0.7604	0.6701	0.7470	0.6509	0.6078	0.1666	0.7013	0.2763
Shbot	0.6246	0.6508	0.6209	0.6535	0.5625	0.7121	0.6796	0.7643	0.6713	0.7453	0.7024	0.6225	0.1634	0.7125	0.2694
MMOE	0.6240	0.6464	0.6181	0.6422	0.5839	0.7062	0.6730	0.7524	0.6546	0.7475	0.6725	0.6229	0.1633	0.7010	0.2764
PLE	0.6262	0.6482	0.6209	0.6565	0.5702	0.7123	0.7097	0.7653	0.7020	0.7473	0.7110	0.6244	0.1633	0.7124	0.2700
STAR	0.6170	0.4404	0.6138	0.4363	0.5813	0.6993	0.5671	0.7612	0.5438	0.7412	0.5911	0.5378	0.1675	0.6506	0.2757
Adasparse	0.6037	0.4299	0.6005	0.4247	0.5830	0.7072	0.5344	0.7638	0.5310	0.7373	0.6028	0.5284	0.1681	0.6461	0.2850
AdaTask	0.6212	0.6671	0.6141	0.6600	0.5702	0.7061	0.6822	0.7501	0.7142	0.7402	0.6932	0.6265	0.1628	0.7143	0.2692
Hamur	0.6235	0.6313	0.6176	0.6288	0.5767	**0.7123**	0.6588	0.7618	0.6161	0.7485	0.6209	0.6156	0.1655	0.6864	0.2734
PEPNet	0.6263	0.6442	0.6207	0.6601	0.5869	0.7074	0.6905	0.7651	0.6762	0.7476	0.7013	0.6276	0.1629	0.7146	0.2690
M2M	0.6232	0.6423	0.6214	0.6457	0.5823	0.7118	0.6700	0.7613	0.6917	0.7476	0.6909	0.6229	0.1632	0.7122	0.2696
M^3OE	0.6264	0.6477	0.6204	0.6543	0.5901	0.7121	0.6933	0.7664	0.7158	0.7496	0.6521	0.6282	0.1625	0.7148	0.2689
SAF	**0.6275**	**0.6691***	**0.6274***	**0.6658***	**0.5995***	0.7106	**0.7237***	**0.7669**	**0.7212***	**0.7515**	**0.7232***	**0.6378***	**0.1620***	**0.7328***	**0.2673***

Table 1: Experimental results on public and industrial datasets. All baselines are MDMT versions except MLP (single-domain, single-task). "d1,ctr" denotes domain 1, CTR task. Bold indicates best, underlined indicates second-best. Results are averaged over 5 runs with different random seeds. "*" denotes statistically significant improvement over the second-best ($p < 0.05$, two-sided t-test).

features during training, thereby encouraging the model to focus on stable, causally relevant information.

Inspired by stable learning (Kuang et al. 2020; Shen et al. 2020; Zhang et al. 2021), we introduce trainable sample weights that directly minimize dependencies between all feature pairs. Specifically, we employ the HSIC (Greenfeld and Shalit 2020) to quantify and penalize statistical dependence between features in Reproducing Kernel Hilbert Spaces (RKHS). By optimizing these sample weights, the model suppresses spurious associations and promotes the learning of stable parameters.

Formally, for random variables A and B (features), HSIC is defined as the squared Hilbert-Schmidt norm of the cross-covariance operator:

$$\mathrm{HSIC}(P_{AB}, \mathcal{F}, \mathcal{G}) = \|C_{AB}\|^2_{\mathrm{HS}}, \tag{1}$$

where $C_{AB} = \mathbb{E}_{AB}[(\phi(A) - \mu_A) \otimes (\psi(B) - \mu_B)]$, with ϕ and ψ denoting feature maps into RKHSs $\mathcal{F}$ and $\mathcal{G}$, and μ_A, μ_B their respective mean embeddings. HSIC captures both linear and nonlinear dependencies, and is zero if and only if A and B are independent.

To make HSIC computation scalable, we approximate kernel mappings using Random Fourier Features (RFF) (Rahimi and Recht 2007). For a feature x, its RFF mapping is:

$$z_x = \sqrt{\frac{2}{n_x}} \left[\cos(\omega_1^\top x + b_1), \ldots, \cos(\omega_{n_x}^\top x + b_{n_x})\right]^\top, \tag{2}$$

where $\omega_i \sim \mathcal{N}(0, 1)$ and $b_i \sim \mathrm{Uniform}[0, 2\pi]$. This enables efficient kernel approximation via inner products.

Let $\mathbf{Z}_A, \mathbf{Z}_B \in \mathbb{R}^{n \times n_x}$ be the RFF matrices for features A and B over n samples. We introduce a trainable sample weight vector $\mathbf{w} \in \mathbb{R}^n$, and define the RFF-based HSIC as:

$$\mathrm{HSIC}_{\mathrm{RFF}} = \frac{1}{n^2} \left\| (\mathbf{H}\,\mathrm{diag}(\mathbf{w})\mathbf{Z}_A)^\top (\mathbf{H}\mathbf{Z}_B) \right\|_F^2, \tag{3}$$

where $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ is the centering matrix. By minimizing $\mathrm{HSIC}_{\mathrm{RFF}}$ for each feature pair, we decorrelate features and suppress spurious dependencies.

The sample weights $\mathbf{w}$ are optimized in an alternating fashion with the model parameters: after each MDMT loss update, we update $\mathbf{w}$ to further reduce feature dependencies, ensuring that the learned representations remain stable throughout training.

3.4 Adaptive Feature Fusion

While feature decorrelation suppresses spurious associations, effective knowledge transfer in MDMT settings also requires adaptive parameter sharing. Indiscriminate sharing can entangle domain- or task-specific features, leading to negative transfer. To address this, we propose an adaptive parameter fusion strategy that selectively leverages stable, decorrelated features for each domain and task.

Unlike prior mixture-of-experts (MoE) or progressive layered extraction (PLE) approaches, our method explicitly enforces expert diversity via RFF-based HSIC regularization and introduces domain- and task-specific gating mechanisms for adaptive feature fusion. This enables more effective disentanglement of shared and domain/task-specific knowledge, which is not achieved by existing architectures.

Concretely, we first embed user-item features into $\mathbf{e}_{u,i}$ and the domain identifier into $\mathbf{e}_{id}$ using dedicated embedding layers. The user-item embedding $\mathbf{e}_{u,i}$ is projected into a shared feature representation $\mathbf{x}^f_{u,i}$ via a shared linear trans-

formation:

$$\mathbf{x}_{u,i}^{f} = \mathrm{ReLU}(\mathrm{BatchNorm}(\mathbf{e}_{u,i}\mathbf{W}_{sh} + \mathbf{b}_{sh})). \quad (4)$$

For a batch of n samples, the matrix of shared features is $\mathbf{X}^{f} \in \mathbb{R}^{n\times d}$. These features are used to compute the RFF-based HSIC, with sample weighting to further decorrelate representations.

To enhance domain adaptability, we introduce a domain scaling gate $g(\cdot, \theta_{fg})$, implemented as a MLP. This gate generates domain-specific scaling factors $\boldsymbol{\pi}$ for each sample:

$$\boldsymbol{\pi} = g(\mathrm{detach}(\mathbf{e}_{u,i}) \oplus \mathbf{e}_{id}; \theta_{fg}), \quad (5)$$

where $\oplus$ denotes vector concatenation and detach prevents gradients from flowing into the user-item embedding during gate training. This design encourages the model to learn generalizable scaling patterns across domains.

We employ K MLP experts f_j, parameterized by θ_{ex}^{j}, to project the domain-scaled features $\mathbf{x}_{u,i}^{f} \circ \boldsymbol{\pi}$ into different subspaces:

$$\mathbf{x}_{u,i}^{ex,j} = f_j(\mathbf{x}_{u,i}^{f} \circ \boldsymbol{\pi}; \theta_{ex}^{j}), \quad j = 1, \ldots, K. \quad (6)$$

To encourage diversity among expert features, we optimize the RFF-based HSIC over the concatenated expert features $[\mathbf{x}_{u,i}^{ex,1}; \ldots; \mathbf{x}_{u,i}^{ex,K}]$.

The resulting stable expert features are then adaptively fused for each task and domain using domain/task-specific gates, implemented as MLPs $g_j^{t,d}$. This enables the model to dynamically select and combine expert knowledge tailored to each prediction scenario:

$$\mathbf{x}_{u,i}^{t,d} = \sum_{j=1}^{K} g_j^{t,d}(\mathbf{x}_{u,i}^{f} \circ \boldsymbol{\pi}; \theta_{eg}^{t,d,j}) \cdot \mathbf{x}_{u,i}^{ex,j}. \quad (7)$$

The fused expert features $\mathbf{x}_{u,i}^{t,d}$ are subsequently fed into task- and domain-specific prediction towers to generate outputs for each task in each domain. For each task t and domain d, we employ a tower network $h_{t,d}(\cdot; \theta_{t,d})$ to produce the final prediction:

$$\hat{y}_{u,i}^{(t,d)} = h_{t,d}(\mathbf{x}_{u,i}^{t,d}; \theta_{t,d}), \quad (8)$$

where $\hat{y}_{u,i}^{(t,d)}$ denotes the predicted output for user-item pair (u, i) on task t in domain d.

3.5 Theoretical Analysis

We provide a theoretical foundation for the SAF approach, which aims to mitigate negative transfer in multi-domain, multi-task learning by suppressing the influence of spurious features. Our analysis builds on the result from (Xu et al. 2022), which shows that sample reweighting can recover the minimal stable (causal) feature set for robust prediction.

Theorem 1 (Minimal Stable Feature Recovery, (Xu et al. 2022))**.** *Let $w \in \mathcal{W}$ be a sample weighting function such that the reweighted distribution $P_w(\boldsymbol{X})$ renders the features $X_i \in \boldsymbol{X}$ mutually independent. Suppose a model is perfectly trained on $P_w(\boldsymbol{X})$ to predict Y. Let $\boldsymbol{X}_s \subseteq \boldsymbol{X}$ denote the minimal stable (causal) variable set such that Y is conditionally independent of $\boldsymbol{X} \setminus \boldsymbol{X}_s$ given $\boldsymbol{X}_s$. Then, the model assigns nonzero importance only to features in $\boldsymbol{X}_s$, and zero importance to all other features.*

Motivated by this result, we propose to learn sample weights w by minimizing the empirical RFF-based HSIC between the spurious features $\boldsymbol{X}_{ir}$ and the stable features $\boldsymbol{X}_s$. Formally, we solve:

$$\min_{w\in\mathcal{W}} \mathrm{HSIC}_{\mathrm{RFF}}^{w}(\boldsymbol{X}_{ir}, \boldsymbol{X}_s) \quad (9)$$

where $\mathrm{HSIC}_{\mathrm{RFF}}^{w}$ denotes the empirical HSIC computed under the reweighted sample distribution.

Proposition 1 (HSIC Minimization Induces Independence)**.** *If the sample weights w^* achieve $\mathrm{HSIC}_{\mathrm{RFF}}^{w^*}(\boldsymbol{X}_{ir}, \boldsymbol{X}_s) = 0$, then under the reweighted distribution P_{w^*}, the spurious features $\boldsymbol{X}_{ir}$ are independent of the stable features $\boldsymbol{X}_s$.*

Proof. HSIC is zero if and only if the two variables are independent under the considered distribution. Thus, minimizing HSIC to zero via sample weights ensures independence between $\boldsymbol{X}_{ir}$ and $\boldsymbol{X}_s$ under P_{w^*}. □

By enforcing independence between spurious and stable features, the reweighted distribution P_{w^*} removes spurious correlations that could otherwise be exploited by the model. According to Theorem 1, if the features are mutually independent under P_{w^*}, a model trained on P_{w^*} will assign nonzero importance only to the minimal stable feature set $\boldsymbol{X}_s$, and zero importance to all other features. Thus, our HSIC-based sample reweighting serves as a practical surrogate for the ideal sample weighting function in (Xu et al. 2022).

Let $h_{t,d}$ denote the model for task t and domain d, and let $h_{s,t,d}^{*}$ be the Bayes optimal classifier using only the stable features $\boldsymbol{X}_s^{(t,d)}$. The excess risk from including irrelevant features is

$$\begin{aligned} &R_{t,d}(h_{t,d}) - R_{t,d}(h_{s,t,d}^{*}) \\ &= \mathbb{E}_{\boldsymbol{X}_s, \boldsymbol{X}_{ir}\sim P_{t,d}} \left[D_{\mathrm{KL}}(P_{t,d}(Y|\boldsymbol{X}_s) \| h_{t,d}(\boldsymbol{X}_s, \boldsymbol{X}_{ir})) \right]. \end{aligned} \quad (10)$$

(See Appendix B.1 for derivation.) If HSIC-based sample reweighting enforces independence between stable and spurious features, $h_{t,d}$ approaches $h_{s,t,d}^{*}$, eliminating excess risk and negative transfer.

In practice, perfect independence is limited by data, model capacity, and the expressiveness of the weighting function class $\mathcal{W}$. However, as empirical HSIC decreases, the influence of spurious features is reduced, and the model's reliance on stable features increases. Our Adaptive Feature Fusion (AFF) module further enhances robustness by adaptively selecting and combining stable features across tasks and domains, which complements the theoretical guarantee by promoting specialization and generalization in the learned representations.

The overall effectiveness of this approach depends on three factors: (1) the expressiveness of the weighting function class $\mathcal{W}$, (2) the accuracy of RFF-based HSIC estimation, and (3) the AFF module's ability to identify and utilize stable features. These factors jointly determine the degree to which negative transfer is mitigated in practice.

Figure 3: Saliency maps of PEPNet (P), M^{3}OE (M), and SAF (S) on the bottom-layer features across industrial datasets. Brighter pixels indicate greater contribution. 'D' and 'T' denote domain and task.

3.6 Overall Optimization

We employ an alternating optimization strategy to jointly decorrelate features and optimize predictive performance. Specifically, we first update the sample weights by minimizing the RFF-based HSIC for m iterations, thereby decorrelating features and suppressing spurious associations. Next, we update the model parameters by minimizing the MDMT loss, weighted by the learned sample weights, for one iteration. This process is repeated throughout training.

To ensure that feature decorrelation is effective at both the batch and dataset level, we maintain exponential moving averages of both features and weights. For each mini-batch, we concatenate the global features $\mathbf{X}^{\text{global}}$ and current batch features $\mathbf{X}$, as well as the corresponding weights $\mathbf{w}^{\text{global}}$ and $\mathbf{w}$. The RFF-based HSIC loss is computed as

$$\mathcal{L}_{\text{HSIC}} = \text{HSIC}_{\text{RFF}}(\mathbf{X}^{\text{global}} \oplus \mathbf{X},\ \mathbf{w}^{\text{global}} \oplus \mathbf{w}), \tag{11}$$

where $\oplus$ denotes concatenation. We minimize $\mathcal{L}_{\text{HSIC}}$ to update only the current sample weights $\mathbf{w}$ (including $\mathbf{w}^{ex}$ and $\mathbf{w}^{f}$), while maintaining an exponential moving average of both features and weights within each mini-batch:

$$\mathbf{X}^{\text{global}} = \alpha \mathbf{X}^{\text{global}} + (1-\alpha)\mathbf{X}, \tag{12}$$

$$\mathbf{w}^{\text{global}} = \alpha \mathbf{w}^{\text{global}} + (1-\alpha)\mathbf{w}, \tag{13}$$

where α is the momentum coefficient.

The MDMT loss is defined as the mean weighted loss over all user-item pairs, tasks, and domains:

$$\mathcal{L}_{\text{MDMT}} = \frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} \sum_{t=1}^{T} \sum_{d=1}^{D} w_{u,i}^{f} \cdot w_{u,i}^{ex} \cdot \ell\left(\hat{y}_{u,i}^{(t,d)},\ y_{u,i}^{(t,d)}\right), \tag{14}$$

where $\mathcal{D}$ denotes the set of user-item pairs and $|\mathcal{D}|$ its cardinality. $w_{u,i}^{f}$ and $w_{u,i}^{ex}$ are elements of the weight vectors $\mathbf{w}^{f}$ and $\mathbf{w}^{ex}$, respectively, corresponding to the user-item pair (u,i).

During inference, only the trained backbone model is used for recommendations, without the need for sample reweighting. Following Adatask (Yang et al. 2023), we apply adaptive learning rates to the bottom and embedding layers to further mitigate gradient domination and ensure balanced optimization across all components.

4 Experiment

We evaluate the effectiveness of the proposedSAF framework on both public and industrial real-world datasets. SAF is compared with 11 baseline methods, encompassing MTR, MDR and MDMT methods. See Appendix A.1 for dataset information, Appendix A.2 for baseline details, and Appendix A.3 for implementation details.

Model	Saliency Entropy	Gini Coefficient	Correlation
PEPNet	5.21 ± 0.12	0.44 ± 0.03	0.58 ± 0.03
M3OE	5.32 ± 0.10	0.41 ± 0.04	0.52 ± 0.02
SAF	2.83 ± 0.09	0.62 ± 0.02	0.21 ± 0.02

Table 2: Saliency map statistics (mean $\pm$ std) across 1,000,000 test samples.

4.1 Overall Performance Comparison

We compare the overall performance of our proposed SAF method with 11 representative baselines, as summarized in Table 1. The key findings are as follows:

- SAF consistently outperforms all baseline models in both AUC and Logloss across all domains and tasks on the AliCCP and Industrial datasets. Specifically, on the AliCCP dataset, SAF achieves AUC scores of 0.6275 (d1, ctr), 0.6691 (d1, cvr), 0.6274 (d2, ctr), 0.6658 (d2, cvr), and 0.5995 (d3, ctr). On the Industrial dataset, SAF attains AUC scores of 0.7106 (d1, ctr), 0.7237 (d1, cvr), 0.7669 (d2, ctr), 0.7213 (d2, cvr), 0.7515 (d3, ctr), and 0.7232 (d3, cvr). Averaged over all tasks, SAF achieves AUCs of 0.6378 and 0.7328 on the AliCCP and Industrial datasets, respectively, surpassing the second-best methods, which obtain 0.6282 and 0.7148. This corresponds to relative improvements of 1.5% and 2.5%. In terms of Logloss, SAF also achieves the lowest values, with 0.1620 and 0.2673 on the AliCCP and Industrial datasets, compared to 0.1625 and 0.2689 for the best baselines.
- SAF demonstrates substantial improvements on tasks and domains with limited samples, such as domain 3 and CVR tasks (each comprising less than 7% of the data), while maintaining strong performance on the major domains and tasks (domain 1, domain 2, and CTR task). This indicates that SAF effectively mitigates common challenges in multi-domain and multi-task learning, such as negative transfer and the dominance of large domains. In contrast, methods employing gating mechanisms, such as PEPnet (0.6672 vs. 0.6934) and M3OE (0.6718 vs. 0.6934), perform worse on smaller domains and tasks. Similarly, Adatask (0.6728 vs. 0.6831), which utilizes adaptive learning rates, exhibits reduced accuracy on tasks with more abundant data. These results demonstrate that SAF is more robust than existing baselines.

4.2 In-depth Analysis

We provide an in-depth analysis to assess both the effectiveness and scalability of SAF in multi-domain, multi-task learning scenarios.

First, we examine feature attributions in the shared bottom layer using smoothed saliency maps (Adebayo et al. 2018)

Dataset	AliCCP	Industrial
SAF w/o all	0.6217	0.7176
SAF w/o SL	0.6309	0.7233
SAF w/o FG	0.6323	0.7302
SAF w/o EG	0.6330	0.7299
SAF w/o SL FG	0.6321	0.7243
SAF	**0.6378***	**0.7328***

Table 3: Ablation study results of SAF. Averaged AUC of all tasks and domains over 5 runs with different random seeds are reported. "*" indicates statistically significant improvements over the ablation versions, as determined by a two-sided t-test with $p < 0.05$.

for PEPNet (Chang et al. 2023), M3oE (Zhang et al. 2024), and SAF. As shown in Figure 3, the saliency maps for M3oE and PEPNet are diffuse, indicating that these models rely on a broad and correlated set of features. In contrast, SAF produces sharply focused saliency maps, consistently highlighting a compact subset of stable, decorrelated features across domains and tasks. This qualitative difference suggests that SAF is more effective at isolating the most relevant features and suppressing spurious correlations.

To quantify these patterns, Table 2 reports three metrics: saliency entropy (Shannon 1948), Gini coefficient (Ceriani and Verme 2024), and feature correlation. Lower entropy indicates more decisive feature selection, higher Gini reflects greater sparsity, and lower correlation signifies more independent attributions. SAF achieves the lowest entropy (2.83), highest Gini (0.62), and lowest correlation (0.21), significantly outperforming the baselines ($p < 0.05$). These results demonstrate that SAF assigns sparse, decorrelated, and decisive attributions for each domain and task.

The observed attribution patterns directly support the theoretical claims in Section 3.5: by decorrelating features and suppressing spurious associations, SAF encourages the model to focus on a minimal, stable set of features, thereby reducing negative transfer. This is further corroborated by the lowest logloss error achieved by SAF in Table 1, indicating improved generalization across domains and tasks.

In terms of scalability, SAF adds only 4% to training time in our online system, with no increase in inference time. This efficiency is achieved because feature decorrelation training is applied only during training, leaving the inference pipeline unchanged. Furthermore, the use of random Fourier features makes the HSIC loss computation highly efficient, particularly since industrial batch sizes are typically under 10,000. Thus, SAF is practical for large-scale, real-time applications.

4.3 Ablation Study

We conduct ablation experiments to assess the contributions of three key SAF components: stable learning (SL), feature adaptive gating (FG), and expert gating (EG). As shown in Table 3, removing all components reduces the average AUC from 0.6378 to 0.6217 on AliCCP and from 0.7328 to 0.7176 on the Industrial dataset, corresponding to improvements of 2.58% and 2.11%, respectively. Excluding SL, FG, EG, or both gating mechanisms results in AUC drops of 1.08%, 0.86%, 0.75%, and 0.89% on AliCCP, and 1.29%, 0.35%, 0.39%, and 1.15% on the Industrial dataset. We also compare adaptive learning rate and task loss tuning (0.6378 vs. 0.6356 on AliCCP, 0.7328 vs. 0.7300 on the Industrial dataset), which yields minor improvements for our method, but no significant gains for PEPNET, M3OE, or other baselines. These results underscore the importance of both stable learning and adaptive fusion: SL mitigates spurious correlations, while adaptive fusion enables robust feature selection across domains and tasks.

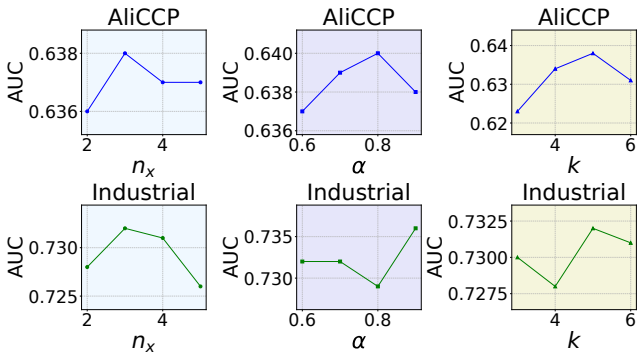

Figure 4: Effects of varying RFF mapping dimension n_x, moving average parameter α, number of experts K on public and industrial datasets. We report the average scores of AUC for each dataset.

4.4 Hyper-Parameter Sensitivity Analysis

We analyze three key hyperparameters: RFF mapping dimension n_x, moving average parameter α, and number of experts K. Performance peaks at $n_x = 3$ for both datasets. The model is relatively insensitive to α, with optimal values at 0.8 (AliCCP) and 0.9 (Industrial). Increasing K to 5 improves AUC, with no further gains beyond that. We use $n_x = 3$, $\alpha = 0.8/0.9$, and $K = 5$ in all experiments (see Figure 4).

5 Conclusion

In this work, we tackle negative transfer caused by spurious correlations in multi-domain multi-task recommendation. We propose SAF, a stable and adaptive fusion framework that first minimizes a weighted RFF-based HSIC loss to decorrelate features under the weighted data distribution. The resulting sample weights are then used to reweight the multi-domain multi-task loss, promoting stable feature learning. SAF also incorporates feature and expert gating mechanisms to adaptively select stable features and model complex cross-domain and cross-task relationships. Extensive experiments on large-scale datasets show that SAF consistently outperforms existing methods.

Acknowledgements

This work was supported by Tencent Rhino-Bird Focused Research Program.

References

Aceto, G.; Persico, V.; and Pescapé, A. 2020. Industry 4.0 and health: Internet of things, big data, and cloud computing for healthcare 4.0. *Journal of Industrial Information Integration*, 18: 100129.

Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; and Kim, B. 2018. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31.

Alhijawi, B.; and Kilani, Y. 2020. The recommender system: a survey. *International Journal of Advanced Intelligence Paradigms*, 15: 229–251.

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Caruana, R. 1997. Multitask learning. *Machine learning*, 28: 41–75.

Ceriani, L.; and Verme, P. 2024. Gini on mutability. *METRON*, 82: 269–292.

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.

Dredze, M.; Kulesza, A.; and Crammer, K. 2010. Multi-domain learning by confidence-weighted parameter combination. *Machine Learning*, 79: 123–149.

Greenfeld, D.; and Shalit, U. 2020. Robust learning with the hilbert-schmidt independence criterion. In *International Conference on Machine Learning*, 3759–3768. PMLR.

Jiang, Y.; Li, Q.; Zhu, H.; Yu, J.; Li, J.; Xu, Z.; Dong, H.; and Zheng, B. 2022. Adaptive domain interest network for multi-domain recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 3212–3221.

Joshi, M.; Dredze, M.; Cohen, W.; and Rose, C. 2012. Multi-domain learning: when do domains matter? In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 1302–1312.

Kuang, K.; Xiong, R.; Cui, P.; Athey, S.; and Li, B. 2020. Stable prediction with model misspecification and agnostic distribution shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 4485–4492.

Lake, B. M.; Ullman, T. D.; Tenenbaum, J. B.; and Gershman, S. J. 2017. Building machines that learn and think like people. *Behavioral and brain sciences*, 40: e253.

Li, D.; Zhang, Z.; Yuan, S.; Gao, M.; Zhang, W.; Yang, C.; Liu, X.; and Yang, J. 2023a. AdaTT: Adaptive Task-to-Task Fusion Network for Multitask Learning in Recommendations. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4370–4379.

Li, X.; Yan, F.; Zhao, X.; Wang, Y.; Chen, B.; Guo, H.; and Tang, R. 2023b. Hamur: Hyper adapter for multi-domain recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 1268–1277.

Lopez-Paz, D.; Nishihara, R.; Chintala, S.; Scholkopf, B.; and Bottou, L. 2017. Discovering causal signals in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6979–6987.

Ma, J.; Zhao, Z.; Yi, X.; Chen, J.; Hong, L.; and Chi, E. H. 2018. 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.

Marcus, G. 2018. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*.

Marx, Z.; Rosenstein, M. T.; Kaelbling, L. P.; and Dietterich, T. G. 2005. Transfer learning with an ensemble of background tasks. *Inductive Transfer*, 10.

Rahimi, A.; and Recht, B. 2007. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20.

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27: 379–423.

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.

Shen, Z.; Cui, P.; Zhang, T.; and Kunag, K. 2020. Stable learning via sample reweighting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 5692–5699.

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.

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, Z.; Dai, Z.; Póczos, B.; and Carbonell, J. 2019. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11293–11302.

Xu, R.; Zhang, X.; Shen, Z.; Zhang, T.; and Cui, P. 2022. A theoretical analysis on independence-driven importance weighting for covariate-shift generalization. In *International Conference on Machine Learning*, 24803–24829. PMLR.

Yang, E.; Pan, J.; Wang, X.; Yu, H.; Shen, L.; Chen, X.; Xiao, L.; Jiang, J.; and Guo, G. 2023. Adatask: A task-aware adaptive learning rate approach to multi-task learning.

In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 10745–10753.

Yu, T.; Kumar, S.; Gupta, A.; Levine, S.; Hausman, K.; and Finn, C. 2020. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836.

Zhang, J.; Liu, Q.; Wu, S.; and Wang, L. 2023. Mining stable preferences: Adaptive modality decorrelation for multimedia recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 443–452.

Zhang, Q.; Liao, X.; Liu, Q.; Xu, J.; and Zheng, B. 2022. 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*, 1368–1376.

Zhang, X.; Cui, P.; Xu, R.; Zhou, L.; He, Y.; and Shen, Z. 2021. Deep stable learning for out-of-distribution generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5372–5382.

Zhang, Y.; and Yang, Q. 2021. A survey on multi-task learning. *IEEE transactions on knowledge and data engineering*, 34: 5586–5609.

Zhang, Z.; Liu, S.; Yu, J.; Cai, Q.; Zhao, X.; Zhang, C.; Liu, Z.; Liu, Q.; Zhao, H.; Hu, L.; et al. 2024. M3oe: Multi-domain multi-task mixture-of experts recommendation framework. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 893–902.