Scenario-Wise Rec: A Multi-Scenario Recommendation Benchmark

Xiaopeng Li*
City University of Hong Kong
Hong Kong, China
xiaopli2-c@my.cityu.edu.hk

Jingtong Gao*
City University of Hong Kong
Hong Kong, China
jt.g@my.cityu.edu.hk

Pengyue Jia
City University of Hong Kong
Hong Kong, China
jia.pengyue@my.cityu.edu.hk

Yichao Wang
Huawei Noah's Ark Lab
Shenzhen, China
wangyichao5@huawei.com

Wanyu Wang
City University of Hong Kong
Hong Kong, China
wanyuwang4-c@my.cityu.edu.hk

Yejing Wang
City University of Hong Kong
Hong Kong, China
yejing.wang@my.cityu.edu.hk

Yuhao Wang
City University of Hong Kong
Hong Kong, China
yhwang25-c@my.cityu.edu.hk

Huifeng Guo
Huawei Noah's Ark Lab
Shenzhen, China
huifeng.guo@huawei.com

ABSTRACT

Multi Scenario Recommendation (MSR) tasks, referring to building a unified model to enhance performance across all recommendation scenarios, have recently gained much attention. However, current research in MSR faces two significant challenges that hinder the field's development: the absence of uniform procedures for multi-domain dataset processing, thus hindering fair comparisons, and most models being closed-sourced, which complicates comparisons with current SOTA models. Consequently, we introduce our benchmark, Scenario-Wise Rec, which comprises four public datasets and eight benchmark models, along with a training and evaluation pipeline. We have also validated our benchmark using the Huawei industrial advertising dataset, further enhancing its reliability. We aim for this benchmark to provide researchers with valuable insights from prior works, enabling the development of novel models based on our benchmark and thereby fostering a collaborative research ecosystem in MSR. Our source code is available at https://github.com/Xiaopengli1/Scenario-Wise-Rec.

CCS CONCEPTS

• **Information systems** → **Recommender systems**;

KEYWORDS

Multi Scenario Recommendation, Recommendation Systems, CTR Prediction

*Both authors contributed equally to the paper.

KDD '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Xiaopeng Li, Jingtong Gao, Pengyue Jia, Yichao Wang, Wanyu Wang, Yejing Wang, Yuhao Wang, and Huifeng Guo. 2023. Scenario-Wise Rec: A Multi-Scenario Recommendation Benchmark. In *Proceedings of Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*. ACM, New York, NY, USA, 16 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Recommender systems, deeply integrated into the digital world, play a crucial role in mitigating data overload and personalizing user experiences across diverse online platforms [7, 36, 37]. Current recommender systems leverage user profiles, behavior sequences, and contextual features to produce customized recommendations for specific user and item scenarios/domains [39]. In the face of varied real-world applications, there is growing research on the development of models capable of managing multiple recommendation scenarios simultaneously, known as the Multi-Scenario Recommendation (MSR) task. MSR models, tailored to unique user and item domains, dynamically learn to transfer knowledge across scenarios. This strategy not only addresses data scarcity in less populated domains but enhances overall recommendation performance [8, 32].

More specifically, multi-scenario recommendations involve designing a unified model capable of generating recommendations across multiple scenarios [24, 29, 33]. These scenarios often represent distinct predefined domains, such as various advertising areas, product pages, or manually defined business units shown in Figure 1. The model's primary objective is to harness knowledge transfer across domains to improve domain-specific performance. Central to these models is the ability to balance shared information and specific information across different scenarios, thereby enhancing the model's overall predictive accuracy. This capability is especially crucial for real-life deployments, where enterprises frequently face the challenge of executing recommendation tasks across multiple scenarios [35].

With the development of deep recommender systems [1, 36] and cross-domain studies [10, 42], we have witnessed the rapid

(a) App Icon Slot (b) Stream Video Slot (c) Open Screen Slot

Figure 1: An MSR example in business application: multi-scenario advertising recommendations from Huawei. Each slot is treated as a specific scenario in modeling.

growth of multi-scenario recommendation methods. Many models, such as STAR [24], Adasparse [33], and Causalint [29], among others, have been proposed and effectively implemented. However, there is still a lack of a widely universally recognized benchmark in this area, which poses significant challenges: Firstly, there is a lack of a standardized pipeline for domain data processing, model training, and model performance evaluation to make fair comparisons between models. Secondly, many current MSR models are closed-sourced due to corporate privacy protection policies, which complicates reproducibility for researchers, thereby impeding the field's progression in multi-scenario recommendations.

Given these challenges, the demand for a well-defined benchmark, specifically tailored for multi-scenario recommendations, grows increasingly urgent. This benchmark should provide standardized procedures for data processing, evaluation, and model interfaces, thereby establishing uniform research norms. In this paper, we propose **Scenario-Wise Rec**, the first benchmark dedicated to MSR. Our benchmark incorporates data preprocessing and evaluation protocols for four public domain datasets, providing a structured framework for model comparison and ensuring equitable evaluation conditions. We have developed a uniform model interface and reproduced eight well-recognized MSR models, including three multi-task-related models and five multi-scenario models. Furthermore, to validate our benchmark's applicability and robustness, we have also applied it to an industrial dataset from Huawei's advertising platform, demonstrating its real-world performance. Our comprehensive approach not only enables researchers to derive valuable insights from existing works but also aims to nurture a collaborative research environment within the MSR field. The main challenges could be listed as follows:

- To the best of our knowledge, this is the first benchmark for cutting-edge multi-scenario recommendation studies;
- We provide a thorough code implementation process for the MSR task, encompassing data processing, model training, and metric evaluations. Our framework includes interfaces for four public datasets and eight well-recognized MSR models, facilitating equitable comparisons and enhancing reproducibility for researchers in the field. Also, we conducted tests on an industrial advertising dataset from Huawei to bolster the benchmark's credibility.
- We have made our benchmark publicly available, thereby facilitating researchers' ability to gain valuable insights and conduct MSR experiments more easily. This initiative aims to promote the development and prosperity of the MSR community.

2 RELATED WORK

In this section, we provide a work overview of topics related to our topic. Firstly, an introduction to recent advancements in Multi-Scenario Recommendation tasks (MSR) is presented. This is succeeded by a review of literature reviews to Click-Through Rate (CTR) prediction tasks.

2.1 Multi-Scenario Recommendation

In recent periods we have seen a surge in interest in multi-scenario recommendation tasks, driven by the rapid growth in both user numbers and web content. Platform providers are keen to segment user groups and content themes into distinct scenarios based on varied attributes. This segmentation strategy bears a resemblance to multi-task learning. Researchers, recognizing this similarity, have been investigating the application of domain-transfer technologies for these challenges. Notable among these efforts are works like [17, 26, 40], which employ a Mixture-of-Expert (MoE) structure to manage diversity across domains. Mario [26] is proposed to capture domain information through feature scaling modules and dynamically capture domain signals using a MoE structure. HiNet is designed with hierarchical structures that construct multiple layers of information extraction to effectively capture domain information while simultaneously preserving domain-specific features. In addition, alternative methodologies have been proposed. For instance, PEPnet [3] processes bottom-level inputs using gating units and introduced EPNet for domain feature selection, alongside PPnet for integrating multi-task information.

Other researchers have sought to address these problems from various perspectives. In the realm of domain modeling, STAR [24] introduces a unified model featuring both a domain-specific and a domain-shared tower. This dual structure effectively captures both unique and shared information, thereby enhancing performance. Similarly, SAR-Net [23] and SAML [4] leverage attention mechanisms to model domain features. They utilize attention layers to facilitate knowledge transfer across domains, ultimately improving overall performance. Furthermore, HAMUR [18] employs domain adapters to improve distribution adaptation in different domains, enhancing prediction outcomes. PLATE [30] adopts prompt technology, widely used in NLP, to boost domain adaptation. From an embedding perspective, AdaptDHM[16] aims to distinguish domain communities and variances through an adaptation module, allowing different clusters to implement unique domain strategies. Additionally, research such as [21] explores domain knowledge transfer through embedding alignment. CausalInt[29] addresses multi-domain recommendations via causal inference, while Adasparse[33] applies distinct pruning strategies across domains to further domain adaptation. In recent advancements, D3 [13] shifts the paradigm by evaluating the impact of the autonomous scenario-splitting method, challenging the conventional approach of the manual scenario-splitting strategy. The MDRAU [14] introduces a task that leverages "seen" domains to aid "unseen" domains. This task employs a novel

encoder-decoder model named DRIP, trained with "masked domain modeling" strategies to capture preferences at both the domain and item levels. M-scan [43] incorporates two principal components to address MSR challenges: a Scenario-Aware Co-Attention mechanism and a Scenario Bias Eliminator. The Co-Attention mechanism extracts user interests from scenarios similar to the current one, while the Scenario Bias Eliminator applies causal counterfactual inference to mitigate biases from varied scenario data. And Uni-CTR [9] is proposed for solving MSR via LLMs by extracting layer-wise semantic representations across different scenarios. For future research, several potential topics are noteworthy. Firstly, refining the application of LLMs for fine-grained scenario alignment is crucial, as Uni-CTR [9] offers a foundational approach, yet it does not explicitly extract scenario commonalities, thereby constraining scenario expansion. Secondly, while current MSR research predominantly focuses on CTR tasks, other areas, such as sequential recommendations for diverse scenarios and trustworthy recommendations within MSR, remain underexplored. Finally, developing a joint model that simultaneously considers multiple tasks, scenarios, behaviors, and interests could pave the way for a more generalized recommendation system. Our paper proposes to establish a foundational benchmark for contemporary models, thereby enabling scholars to make significant contributions in this field.

2.2 CTR Prediction

Click-Through Rate (CTR) serves as a primary metric for gauging user click behavior. This indicator is instrumental for online platforms in discerning user preferences regarding content. Furthermore, it plays a crucial role in assessing the efficacy of content dissemination strategies, thereby enhancing both the quality of platform services and revenue profit.

Traditional CTR prediction models can be broadly categorized into four groups. The first group encompasses statistical models, with Logistic Regression (LR) [15] as a prime example. These models are valued for their interpretability and ease of deployment and are extensively utilized by numerous online service providers. The second group, Factorization Machines (FMs), are machine learning algorithms based on matrix factorization, first introduced in [22]. FMs, notable for their effectiveness in sparse data scenarios and linear complexity, have become a staple in advertising prediction models. The third category includes tree-based models such as GBDT and XGBoost, renowned for their adeptness at complex feature combinations and interaction exploration. The final category comprises deep learning models, which demonstrate remarkable capability for feature interactions. For instance, DeepFM[11] integrates DNN with FM, while Wide&Deep[5] combines LR with DNN. Further developments include models like DCN [27], which substitute DNN in Wide&Deep with a Deep-Cross Net to enhance performance, and xDeepFM [19] and DCNv2 [28], which incorporate DIN with DNN and a mixture of low-rank DCN, achieving significant improvements. Our benchmark concentrates on the task of CTR prediction, incorporating several MSR models designed to enhance click rates. A detailed introduction to these models is provided in Section 4.3. Additionally, the metrics employed for evaluation are also thoroughly described in Section 5.1.2.

3 PIPELINE

In this section, we give a detailed introduction to the components of the proposed benchmark as shown in Figure 3.

3.1 Task: Click-Through Rate Prediction

The overall recommendation task of the proposed benchmark is Click-Through Rate (CTR) prediction, which is widely used in many MSR works and serves as a primary task in many online platforms, such as search engines, social networks, and online recommendations. The goal of this task is to forecast whether a user would click on a provided recommendation or not, thereby assisting in the personalization of content delivered to each user.

Formally, given a user u and an item i, the click-through rate $\hat{y}$ is estimated as a probability in the range $[0, 1]$, representing the likelihood that user u would click on item i.

Mathematically, this can be expressed as:

$$\hat{y}_{ui} = f_\theta(x_{ui}), \tag{1}$$

where $f_\theta(.)$ is the recommendation model parameterized by θ, x_{ui} refers to the input features that represent user u and item i, and $\hat{y}_{ui}$ is the predicted click-through rate.

The output range of $\hat{y}_{ui}$ is set to $[0, 1]$, making $\hat{y}_{ui}$ interpretable as a probability. Therefore, the task of predicting CTR essentially boils down to learning the parameters of the function $f_\theta(.)$ based on historical interactions, such that the predicted click-through rate $\hat{y}_{ui}$ is as close as possible to the observed outcome y_{ui}, which can be 0 or 1 indicating whether user u clicked on item i or not.

The loss function to learn the parameters θ is often the Binary Cross Entropy loss (BCE loss), which could be formulated as follows:

$$L(\theta) = -\frac{1}{N}\sum_{u,i} y_{ui}\log(\hat{y}_{ui}) + (1 - y_{ui})\log(1 - \hat{y}_{ui}), \tag{2}$$

where N is the total number of user-item pairs. In the context of CTR prediction, minimizing this loss will guide the recommendation model to generate higher likelihoods for cases where the user clicked on the item, and lower likelihoods otherwise.

3.2 Open Datasets

Open datasets are valuable in facilitating research in the area of recommendations. Currently, there are numerous open datasets available for recommendations. However, different studies may utilize different parts of the dataset in different forms, which hinders the fair comparison of these studies. Therefore, the proposed benchmark aims to offer a unified data loading interface, enabling a standardized method for accessing open datasets to ensure fair comparison between different studies. Specifically, our proposed benchmark provides several open datasets for comparison. Four commonly used public datasets, MoveiLens, KuaiRand, Ali-CCP, and Tenrec have been tested and evaluated under the proposed benchmark. The proposed unified data loading interface also offers convenient extensibility, and using additional publicly available datasets for experimentation and evaluation under the proposed benchmark is strongly encouraged, more information could refer 4.1.

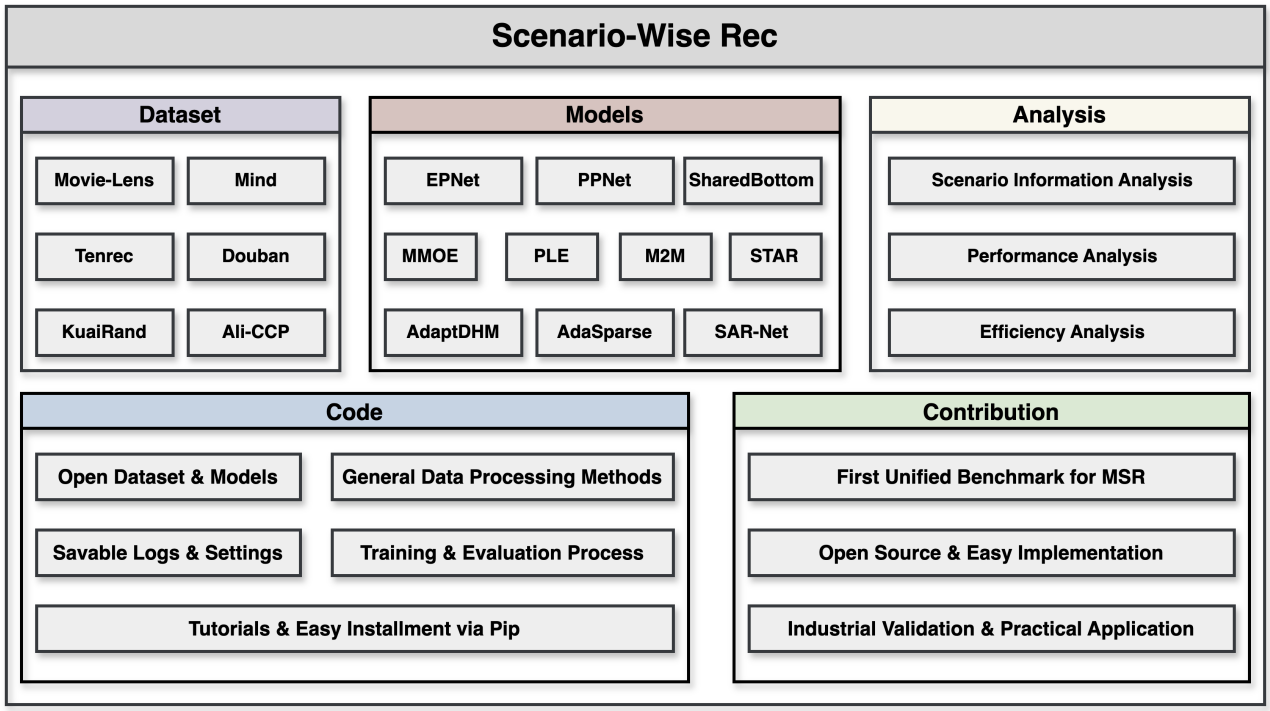

Figure 2: The structure of benchmark Scenario-Wise Rec.

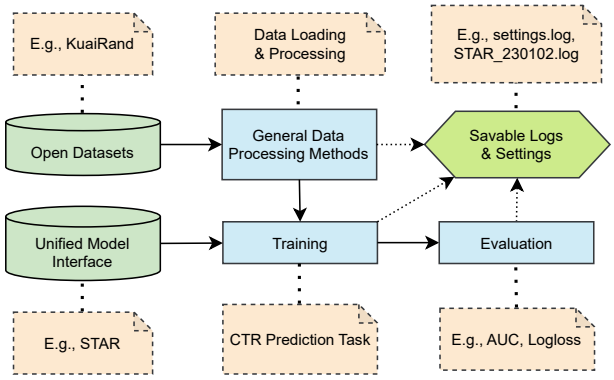

Figure 3: Overall pipeline of Scenario-Wise Rec.

3.3 General Data Processing Methods

Although many open datasets have been applied to performance validation in multiple studies, there is a large degree of variation in the way they are processed in different literature, which leads to inconsistencies in the results of existing papers. This is because most studies use their own data processing methods and fail to provide processed data or a detailed overview of data processing methods, making it difficult for others to directly reuse the same data. In some existing recommendation systems benchmarks, such as RecBole [38], data under the same data set are usually processed by a common method in a unified manner, and corresponding scripts are provided, which greatly facilitates repeatability research. So, our work aims to establish such the same reproducible data processing paradigm for multiple scenarios to ensure a fair comparison of models and repeatable experiments. To achieve this goal, a unified data processing method for each data set has been applied, such as scenario feature declaration, common feature filtering, and processing. These general data processing methods allow the community to conduct diverse research on data that can be directly loaded and processed, while also ensuring fair evaluation results.

3.4 Unified Model Interface

Currently, there are two primary ways to obtain open-source models. The first is through the publication of open-source code by the model's author papers. The second is by reproducing open-source code by others for baseline comparisons or other research purposes. Some commonly used multi-scenario models, such as MMOE [20], already have multiple available open-source model codes. However, with the increasing research in this area, there are still many multi-scenario models that do not have open-source code. Additionally, many open-source codes may use different packages (such as PyTorch) and are written in different manners. These differences can lead to variations in output for similar modules, like Multi-Layer Perceptrons (MLPs), due to complex and varied initializations. The above challenges make reproducible research and fair comparisons difficult. Therefore, our benchmark aims to implement the underlying modules in a standardized way with a consistent model setup and call interface to ensure reproducible model implementations and fair performance comparisons through simple hyper-parameter

settings. So far, we have implemented eight cutting-edge and commonly used models in the field of multi-scenario recommendation through this interface and conducted fair evaluation, which proves the effectiveness of the unified model interface.

3.5 Training

In order to ensure fair comparisons and easy scalability, we have implemented a unified model training procedure. This procedure eliminates most unnecessary disparate treatments of different models during the training process and allows for straightforward extension with various models and datasets. Moreover, to facilitate the reproduction of the study, we provide functions for saving logs, which allow for clear record-keeping of training specifics for each experiment. This facilitates the reproducibility of experiments under the same settings.

3.6 Evaluation

Evaluation metrics are the key to evaluating model performance. Due to the large number of evaluation metrics, it is common for different studies to use different evaluation metrics. This, however, makes fair comparisons between models difficult. Therefore, in the proposed benchmark for the CTR task, we recommend using the two most commonly used evaluation metrics, AUC and Logloss, to evaluate the performance of the model in different scenarios, as well as the overall performance. Moreover, our benchmark provides a consistent evaluation interface for all models to ensure a fairer comparison between different models.

3.7 Savable Logs & Settings

In order to standardize the evaluation process and ensure fair comparison and reproducibility, Scenario-Wise Recprovides a unified interface for hyper-parameter setting. At the same time, hyper-parameter settings and training logs are written in savable files. With these files, users can both clearly understand the performance changes of the model during training and easily reproduce the current results based on saved hyper-parameter settings.

4 BENCHMARKING FOR MULTI-SCENARIO RECOMMENDATION

In this section, a comprehensive overview of the datasets employed in our benchmark is provided, along with an in-depth analysis of scenario-specific information and a description of the multi-scenario baseline model which we implemented in this benchmark.

4.1 Dataset

Adhering to the principles of fair comparison and ease of use, our benchmark selects widely-used multi-scenario open datasets varying in feature numbers and data volumes. Furthermore, the benchmark model is deployed on a real-world dataset from Huawei's advertising platform to augment the reliability and applicability of experimental comparisons. Specifically, for public datasets, we choose MovieLens-1M, KuaiRand, Tenrec, and Ali-CCP, and the industrial advertising dataset from Huawei is derived from daily logs. A detailed introduction of these datasets is elaborated as follows. The statistics of the fourdatasets are listed in Table 1.

Table 1: Dataset Overal Statistics.

Dataset	# Scenario	# User	# Item	# Interaction
Movie-Lens	3	6k	4k	1M
KuaiRand	5	1k	4M	11M
Ali-CCP	3	238k	467k	85M
Tenrec	3	1M	2M	120M
Douban	3	2k	210k	1.7M
Mind	4	748k	20k	56M
Huawei	10	-	-	3M

- **MovieLens**[1]: The MovieLens dataset is a comprehensive collection of movie ratings and information that is widely used for various research and recommender systems. It contains user ratings, demographic information, movie metadata, and user preferences. It consists of 1 million anonymous ratings of approximately 4 thousand movies made by 6 thousand MovieLens users. With the development of recommender systems, it has become an invaluable resource that enables insights into movie preferences and aids in the development of innovative recommendation systems for the benefit of movie enthusiasts worldwide. In the proposed Scenario-Wise Rec, to realize multi-scenario evaluation, interaction samples are divided into three scenarios based on the "age" feature, i.e., "1-24", "25-34", and "35+".
- **KuaiRand**[2]: The KuaiRand dataset is an unbiased recommendation dataset with randomly exposed videos gathered from the Kuaishou App. In Scenario-Wise Rec, KuaiRand has been processed and used for model evaluation. It contains 11 million interactions with 1 thousand users and 4 million videos. In this dataset, different scenarios represent different advertising positions of the Kuaishou App. The scenario identification "tab" has already been given as a feature in the range of [0,14] to indicate the scenario of different interactions. To facilitate the evaluation, we extracted data from the top five scenarios with the most data for training and testing.
- **Ali-CCP**[3]: Ali-CCP is a large-scale CTR recommendation dataset gathered from the real-world traffic logs of the recommender system in Taobao, which is one of the largest online retail platforms in the world. In this dataset, context feature "301" is regarded as a different scenarios indicator, representing an expression of the position the interaction sample is from.
- **Tenrec**[4]: The Tenrec dataset is designed for multiple recommendation tasks, collected from Tencent's two distinct feed recommendation platforms, encompassing video and graphic recommendations. Given that our study focuses on CTR tasks, we adopt the preprocessing setting in paper [34], which specifically processes the dataset from the QQ-KAN platform, focusing on video-watching data for CTR analysis. This extensive dataset includes over 1 million users, nearly 2 million videos, and over 120 million interaction samples. Comprising 20 features, the dataset encompasses user characteristics, video attributes, and other

[1] https://grouplens.org/datasets/movielens/
[2] https://kuairand.com/
[3] https://tianchi.aliyun.com/dataset/408
[4] https://static.qblv.qq.com/qblv/h5/algo-frontend/tenrec_dataset.html

relevant factors. We have chosen the "video_category" as the domain indicator for our analysis.

- **Douban** [41]: The Douban dataset, a real-world collection derived from the Douban platform, is divided into three subsets: Douban-book, Douban-music, and Douban-movie. All subsets share the same users, and we treat each platform as a distinct scenario. In terms of user features, attributes like "living place" and "user ID" are retained. For items, we systematically renumber all items across the three domains and assign new ids. Following the previous work [41], ratings above 3 are considered positive, while those 3 or below are deemed negative.
- **Mind** [31]: The MIcrosoft News Dataset (MIND) is specifically designed for news recommendation by Microsoft. It is a real-world dataset gathered from users of the Microsoft News platform. For our benchmark, we collect the metadata from both training and validation datasets of MIND to create a comprehensive dataset. Regarding item features, we maintain "category" and "subcategory" attributes, labeling "clicks" as positive and "not click" as negative. In terms of domain division, we categorize different genres as separate scenarios. Specifically, we retain the four largest genres, "news", "lifestyle", "sports", and "finance" as distinct domains. This configuration encompasses a total of 748 million users, more than 20k items, and over 56 million interactions.
- **Huawei Industrial Dataset**[5]: The industrial dataset utilized in our paper is a subset, uniformly sampled from the click logs across ten scenarios on the Huawei advertising platform, spanning a nine-day period. We set the initial seven days' data for training, and the data from the eighth and ninth serve as validation and test datasets, respectively. This dataset comprises 108 features, encompassing user features, item features, contextual features, and scenario-specific features. While different scenarios exhibit a common user and item space, they also maintain their unique domain-specific users and items.

4.2 Scenario Information Analysis

As mentioned in the previous section, our study employs five datasets. However, unlike conventional recommendation benchmarks, our research primarily targets multi-scenario recommendation tasks. Accordingly, this section provides a detailed analysis of domain-specific information and statistical data for each dataset.

4.2.1 Scenario Splitting Strategy. Unlike traditional CTR prediction tasks, MSR models emphasize domain-unified prediction, requiring a domain indicator within the dataset features to facilitate dataset splitting. Traditionally, scholars utilize features such as the advertising area, product page number, or other manually defined context features as domain indicators. Specifically, for datasets focusing on multi-scenario recommendations (E.g., Ali-CCP, KuaiRand), the domain indicator is often a predefined feature field provided by the dataset itself, representing different sources of different samples (E.g., different advertising slots). For general datasets (E.g., ML-1M), when applied to multi-scenario recommendations, existing studies often use a feature that can clearly distinguish samples as a domain indicator (E.g., item category). Notably, recent studies, like [12], have begun exploring other domain-splitting features to enhance overall performance. In our benchmark, to advance domain analysis, we implement various splitting strategies, encompassing traditional context feature division, user feature separation, and item feature segmentation across five datasets. Specifically, for the Ali-CCP dataset, we follow the approach of previous studies such as [18, 29], employing the "301" feature, which denotes the display position of items on the screen. In the KuaiRand dataset, segmentation is based on the "tab" feature, indicating whether the recommendation appears on the app's main page or a specific recommendation page. For the MovieLens dataset, user segmentation is achieved using the "age" feature, categorizing users into different groups to investigate the impact of user division on results. Regarding Tenrec, we utilize "video category" as the splitting criterion to examine the influence of item feature division in multi-scenario recommendations. Lastly, for the industrial dataset, we adhere to the guidelines in the manual and segment the dataset using the "ListID", which serves as an indicator of different platforms within Huawei.

[5]We will publicize this dataset upon acceptance to foster the research on this important topic.

Table 2: Dataset Statistics for Multi-Domains.

Dataset	Domain number	Interaction	User	Item
MovieLens	Domain 0	210,747	1,325	3,429
	Domain 1	395,556	2,096	3,508
	Domain 2	393,906	2,619	3,595
KuaiRand	Domain 0	2,407,352	961	1,596,491
	Domain 1	7,760,237	991	2,741,383
	Domain 2	895,385	171	332,210
	Domain 3	402,366	832	547,908
	Domain 4	183,403	832	43,106
Ali-CCP	Domain 0	32,236,951	89,283	465,870
	Domain 1	639,897	2,561	188,610
	Domain 2	52,439,671	150,471	467,122
Tenrec	Domain 0	64,475,979	997,263	1,365,660
	Domain 1	54,277,815	989,911	791,826
	Domain 2	1,588,512	455,636	152,601
Huawei	Domain 0	301,654	-	-
	Domain 1	91,468	-	-
	Domain 2	22,986	-	-
	Domain 3	10,928	-	-

4.2.2 Scenario Analysis. The results of the dataset splitting are detailed in Table 2. Due to space limits, we display only the top four domains out of ten for the Huawei datasets. Complete information could be referred to Table 6. Considering the variability in splitting outcomes across different datasets, we utilize the Coefficient of Variation (COV) [6] to evaluate the uniformity of domain distribution within each dataset. A higher COV value signifies a higher degree of uneven distribution among domains, as depicted in Table 3. Our analysis indicates that KuaiRand exhibits the most uneven domain distribution, and MovieLens displays the most uniform distribution. This observation aligns with our splitting strategy. MovieLens is segmented into relatively evenly distributed age groups. In contrast,

KuaiRand users tend to mainly stay on the homepage, leading to an uneven distribution across different pages. The COV values for the datasets Ali-CCP, Tenrec, and Huawei are approximately 0.9, suggesting a nearly uniform domain distribution across all domains.

To gain a deeper understanding of domain splitting in public datasets, we illustrate the intersection of different domains in each dataset in Table 3. However, for the industrial dataset, owing to data protection and privacy policies, obtaining specific user and item information is not feasible. Our findings indicate that user and item interaction attributes vary significantly across different datasets. In the MovieLens dataset, segmented by users' age groups, we observe that each age group shares a majority of movies while maintaining a distinct preference for a small number of films. For KuaiRand, as depicted in Table 2, we notice a bimodal distribution in domain users and a long-tail distribution in items. This pattern is also reflected in interaction distribution. For example, domains 3 and 4 share 704 users out of a total of 832, suggesting similar user behavior patterns in these domains, yet the interactions with items are notably distinct. In the Ali-CCP dataset, Domain 1 is a quite small domain, accounting for nearly 1% of total interactions, resulting in a skewed domain distribution. Intersection analysis reveals that these three domains maintain distinct attributes, sharing only a small portion of users and items across each pair. For Tenrec, as previously mentioned, domains are split based on item features, resulting in no item intersections, but users are shared across domains. In the case of Domain 2, approximately 30% of the users overlap with Domains 0 and 1, offering limited insights for exploring reach within this dataset.

Table 3: Dataset Statistics for Domain Intersection.

Dataset	COV	Domain	User Intersection	Item Intersection
MovieLens	0.3186	D0 ∩ D1	-	3,320
		D1 ∩ D2	-	3,448
		D0 ∩ D2	-	3,354
KuaiRand	1.3552	D0 ∩ D1	961	380,375
		D0 ∩ D2	160	64,292
		D1 ∩ D2	162	213,106
		D1 ∩ D3	832	264,931
		D2 ∩ D3	141	66,063
		D3 ∩ D4	704	2,721
Ali-CCP	0.9180	D0 ∩ D1	814	188,510
		D1 ∩ D2	515	188,590
		D0 ∩ D2	2,385	465,694
Tenrec	0.8413	D0 ∩ D1	987,743	-
		D1 ∩ D2	454,158	-
		D0 ∩ D2	455,221	-

4.3 Multi-Scenario Recommendation Model

With the rapid development of multi-scenario recommendations, more and more research has arisen. However, due to the different data, parameters, and model implementation methods used in different studies, it is difficult to directly summarize the current frontier research and make a fair comparison. In order to track the most cutting-edge research in the field of multi-scenario recommendation and facilitate fair comparison, in the proposed Scenario-Wise Rec, we reproduce eight cutting-edge models that are commonly used or mentioned in the related studies and evaluate them on the four public datasets. We reproduce these models under the uniform model interface described in Section 3.4, and reproduction details are depicted in Appendix A.2.2. An introduction about these models is described as follows.

- **Shared Bottom [2]:** The Shared Bottom model is an approach for multi-task recommendation tasks. It learns a shared representation from different tasks with a shared network base to capture the patterns and shared information. Afterward, different network towers are applied to different tasks for task-specific modeling. Recently, it has also been applied to multi-scenario recommendations as a commonly used baseline by treating different scenarios as different recommendation tasks [24, 29].
- **MMOE [20]:** Multi-gate Mixture-of-Experts (MMOE) model is a commonly used model for multi-task learning. Different from the Shared Bottom, MMOE applies multiple expert networks named MOE (i.e., Mixture-of-Experts structure) as the bottom structure and uses multiple gating networks to control the connections between different experts and the following task-specific networks. Through a detailed modeling of task relations, MMOE achieves better performance in multi-task recommendations. Similar to other multi-task models. MMOE can also be easily applied to multi-scenario recommendations by treating different scenarios as different recommendation tasks.
- **PLE [25]:** The Progressive Layered Extraction (PLE) model is a solution to the challenges faced by multi-task learning (MTL) models in recommender systems. PLE addresses the issues of negative transfer and complex task correlations by separating shared components and task-specific components explicitly and adopting a progressive routing mechanism to gradually extract deeper semantic knowledge. Through extensive experiments, PLE has outperformed state-of-the-art MTL models significantly in various task correlation scenarios. Similarly, PLE could also be applied as an effective multi-scenario recommendation model by treating different scenarios as different recommendation tasks.
- **SAR-Net [23]:** The Scenario-Aware Ranking Network (SAR-Net) is proposed by Alibaba and designed for the travel marketing platform for multi-scenario recommendation tasks. It tackles the challenge of training a unified model by leveraging specific attention modules that incorporate scenario, item features, and user behavior features. Moreover, SAR-Net handles biased logs resulting from manual intervention during promotion periods through scenario-specific expert networks, scenario-shared expert networks, and a multi-scenario gating module. Experiments and online A/B testing demonstrate the effectiveness of SAR-Net, which has been successfully deployed and serves hundreds of travel scenarios on Alibaba's online travel marketing platform.
- **STAR [24]:** The Star Topology Adaptive Recommender (STAR) model addresses the challenge of making click-through rate (CTR) predictions for multiple scenarios within a large-scale commercial platform. It achieves multi-scenario learning by combining a shared network that captures commonalities between

scenarios (referred to as domains) with domain-specific networks tailored to each domain. The weights of the shared network and the domain-specific network are multiplied to generate a unified network during the inference stage for each domain. STAR effectively learns the shared network from all data and adapts domain-specific parameters to each domain's characteristics. Production data has validated the effectiveness of STAR, with significant improvements in CTR and Revenue Per Mille (RPM) observed since its deployment in Alibaba's display advertising system in late 2020.

- **M2M [35]:** The Multi-Scenario Multi-Task Meta-Learning (M2M) model is a novel approach designed to address the challenges of multi-task and multi-scenario advertiser modeling in e-commerce platforms like Taobao and Amazon. M2M utilizes a meta unit to capture inter-scenario correlations, a meta attention module to capture diverse inter-scenario correlations for different tasks, and a meta tower module to enhance scenario-specific feature representation for different recommendation tasks. In Scenario-Wise Rec, the number of the meta-towers is set to 1 to correspond to the single CTR prediction task.
- **AdaSparse [33]:** AdaSparse is designed for multi-domain CTR prediction and aims to adaptively learn the sparse structures of domain models. Specifically, AdaSparse introduces a lightweight network functioning as a pruner, which operates a domain-pruning process for each layer within individual domain towers. During this pruning process, a novel fusion strategy is employed, combining binary and scale approaches to enhance pruning performance, effectively eliminating as much redundant information as possible. The results demonstrate significant improvements not only in public datasets but also in online A/B tests within Alibaba's advertising system's CTR platform.
- **AdaptDHM [16]:** The Adaptive Distribution Hierarchical Model (AdaptDHM), a novel multi-distribution method, concentrates on multi-domain CTR prediction. It features an end-to-end, hierarchical structure that includes a clustering process and a classification process. The core component, the distribution adaptation module, employs a routing mechanism, adaptively determining the distribution cluster for each sample. This model effectively captures the commonalities and distinctions among various distributions, thereby enhancing the model's representation capability without relying on prior knowledge for predefined data allocation. Extensive experiments are conducted on public datasets, and an industrial dataset from Alibaba's online system consisting of 10 distinct domains. The results demonstrate its effectiveness and efficiency compared to other models.

- **PPNet & EPNet [3]:** PPNet and EPNet are two submodels in the Parameter and Embedding Personalized Network (PEPNet). EPNet performs personalized selection on embedding to fuse features with different importance for different users in multiple domains. PPNet executes personalized modification on DNN parameters to balance targets with different sparsity for different users in multiple tasks. By applying PPNet and EPNet, PEPNet is able to handle multi-task recommendations under multi-scenario settings. In Scenario-Wise Rec, We designed these two models to explore the impact of each on multi-scenario recommendations. Meanwhile, the number of the meta-towers in PPNet is set to the same as the domain number to correspond to the CTR prediction task on each domain.

5 EXPERIMENT

In this section, a comprehensive overview of our benchmarking experiment is provided, including aspects like dataset processing, evaluation metrics, and parameter settings. Subsequent to this, we report testing results and present an analysis of our experiments.

5.1 Benchmarking Settings

In our benchmark, we evaluated eight benchmark models across five datasets and open-sourced our benchmark package for scholars to use. For clear reference, we provide detailed descriptions of how the dataset is processed, the evaluation metrics we use, the parameters employed, and our detailed experimental setup.

5.1.1 Data Processing. For each dataset, we independently apply feature processing. Diverse pre-processing strategies are employed for distinct features, including discretization and bucketing methods. Subsequently, the features are categorized into three groups: sparse features, representing discretized attributes; dense features, corresponding to continuous attributes; and domain features, which specify the domain number for domain-specific operations. We split the dataset into three parts: training, evaluation, and testing with a ratio 8:1:1 for MovieLens, KuaiRand, and Tenrec. For Ali-CCP, it has already been divided into three-fold, and for industrial dataset, we detail the splitting strategy in Section 4.1.

5.1.2 Evaluation Metrics. Our benchmark evaluation follows methodologies from prior work, such as the conventional CTR prediction task in [11]. We employ two metrics: Area Under the ROC Curve (AUC) and Logloss. AUC is the cross-entropy loss to evaluate the classification performance for different models and AUC is used to measure the probability that a random positive sample ranked higher than a negative sample. Generally, a higher AUC value or a lower Log Loss value denotes greater model performance.

5.1.3 Parameter Settings. Considering the unavailability of numerous multi-scenario models, it becomes impractical to ascertain every detail of each model's settings. Consequently, in our benchmark experiment, we maintain a commitment to the principle of fair comparison. For every dataset, we set the parameters for each model within the search space to make sure the number of parameters remains within the same order of magnitude, thereby facilitating a fair comparison. For more details, refer to the Appendix A.2.

We present the results in Table 4, including the performance of multi-scenario models across five datasets, showing both AUC and Logloss. In the following sections, we will provide an analysis of each dataset, comparing the performance of different models.

5.1.4 Analysis for Movie-Lens. As Table 3 demonstrates, the distribution of all domains in the MovieLens dataset is quite balanced. Analyzing the overall performance from Table 4, M2M and AdaSparse emerge as the top performance models. This success is attributed to the design of the meta unit and the sparse pruner, which effectively recognizes domain-specific patterns, allowing the model to adapt across all domains. Table 7 reveals no significant "seesaw

Table 4: Performance Comparison Results.

Model/Datasets	Movie-Lens		KuaiRand		Ali-CCP		Tenrec		Douban		Mind		Industrial Dataset	
	AUC	Logloss	AUC	Logloss	AUC	Logloss	AUC	Logloss	AUC	Logloss	AUC	Logloss	AUC	Logloss
SharedBottom	0.8095	0.5228	0.7793	0.5483	0.6232	0.1628	0.7748	0.4575	0.7993	0.5178	0.7509	0.1600	0.8276	0.1521
MMOE	0.8086	0.5218	0.7794	0.5477	0.6242	0.1621	0.7750	0.4579	0.7978	0.5192	0.7508	0.1600	0.8301	0.1567
PLE	0.8091	0.5257	0.7796	0.5495	0.6250	0.1617	0.7749	0.4575	0.7979	0.5196	0.7503	0.1601	0.8330	0.1496
STAR	0.8096	0.5258	0.7806	0.5404	0.6253	0.1613	0.7737	0.4584	0.7957	0.5218	0.7512	0.1601	0.8310	0.1503
SAR-Net	0.8092	0.5245	0.7816	0.5393	0.6245	0.1616	0.7741	0.4582	0.8033	0.5131	0.7490	0.1604	0.8355	0.1528
M2M	0.8115	0.5213	0.7821	0.5397	0.6257	0.1611	0.7819	0.4527	0.7962	0.5229	0.7508	0.1601	0.8392	0.1494
AdaSparse	0.8108	0.5205	0.7816	0.5399	0.6239	0.1614	0.7752	0.4568	0.7963	0.5216	0.7497	0.1644	0.8224	0.1596
AdaptDHM	0.8083	0.5238	0.7773	0.5436	0.6233	0.1619	0.7748	0.4575	0.8003	0.5187	0.7328	0.1629	0.8358	0.1489
EPNet	0.8097	0.5215	0.7801	0.5411	0.6236	0.1612	0.7817	0.4559	0.7997	0.5182	0.7418	0.1616	-	-
PPNet	0.8063	0.5257	0.7800	0.5408	0.6144	0.1622	0.7749	0.4576	0.7994	0.5175	0.7494	0.1603	-	-

Table 5: Efficiency Performance Results.

	Douban			KuaiRand			Mind		
	Training time (1 epoch)	Inference time	Parameter size	Training time (1 epoch)	Inference time	Parameter size	Training time (1 epoch)	Inference time	Parameter size
SAR-Net	10s	2s	3.44M	5min30s	32s	69.59M	6min50s	33s	12.31M
STAR	11s	2s	3.50M	5min55s	30s	69.90M	7min28s	41s	12.38M
SharedBottom	9s	2s	3.43M	6min12s	31s	69.53M	7min20s	42s	12.35M
MMOE	11s	2s	3.42M	6min38s	45s	69.51M	7min29s	50s	12.31M
PLE	11s	2s	3.43M	6min10s	45s	69.81M	8min57s	53s	12.35M
M2M	18s	2s	3.54M	5min57s	46s	72.87M	9min13s	57s	12.38M
AdaptDHM	10s	2s	3.45M	5min58s	46s	69.56M	7min19s	50s	12.44M
AdaSparse	10s	2s	3.43M	5min33s	41s	69.79M	7min51s	52s	12.34M
ppnet	12s	2s	3.60M	6min20s	50s	70.54M	8min45s	50s	12.52M
epnet	10s	2s	3.43M	6min08s	45s	69.95M	7min30s	50s	12.30M

	MovieLens			Ali-CCP			Tenrec		
	Training time (1 epoch)	Inference time	Parameter size	Training time (1 epoch)	Inference time	Parameter size	Training time (1 epoch)	Inference time	Parameter size
SAR-Net	7s	2s	239.34K	48min00s	26min4s	25.07M	43min34s	5min54s	83.37M
STAR	8s	2s	308.63K	49min52s	27min8s	25.54M	40min27s	5min27s	83.74M
SharedBottom	8s	2s	227.59K	48min38s	25min34s	25.69M	43min14s	5min35s	83.96M
MMOE	9s	2s	217.80K	51min40s	23min12s	25.40M	42min50s	5min12s	83.63M
PLE	8s	2s	224.20K	42min39s	25min43s	25.96M	42min12s	5min39s	84.06M
M2M	11s	2s	372.53K	50min42s	24min36s	26.68M	42min10s	5min10s	84.16M
AdaptDHM	8s	2s	257.49K	53min14s	25min7s	25.52M	43min50s	5min45s	83.75M
AdaSparse	8s	2s	230.32K	48min05s	24min15s	25.33M	41min20s	5min21s	83.58M
ppnet	9s	2s	349.68K	48min30s	23min44s	26.23M	42min43s	5min03s	84.29M
epnet	8s	2s	232.33K	50min14s	25min47s	25.23M	43min13s	5min00s	83.43M

phenomenon", aligning with our dataset splitting strategy. However, structural differences among models result in varied domain emphases. For instance, Shared-Bottom models, which share a bottom tower across all domains, exhibit a more uniform performance than other MSR models.

5.1.5 Analysis for KuaiRand. KuaiRand is a dataset comprising five distinct domains, which, unlike the MovieLens dataset, shows an uneven distribution across domains. Analysis of Table 4 reveals that MSR models such as STAR, SAR-Net, and M2M significantly outperform multi-task models like SharedBottom, MMOE, and PLE.

This underscores the importance of meticulous architecture design for multi-scenario tasks, considering that variations in data distribution across different domains can have a profound impact on overall performance. The "seesaw phenomenon" observed in Table 10 illustrates the disparity in performance across domains, with domains 2# and 4# significantly outperforming the others.

5.1.6 Analysis for Ali-CCP. Ali-CCP is a dataset containing three domains, with a notably uneven distribution due to the small size of domain 1#. Analysis of Table 4 indicates that STAR and M2M lead other models by a narrow margin. This suggests that the design of the star topology and the meta-unit paradigm can effectively address balance across all domains, especially in cases of significant unevenness in domain distribution. Regarding domain-specific results in Table 8, the seesaw effect is evident, particularly since STAR and M2M demonstrate superior performance in the data-sparse domain 1#, outperforming other models significantly.

5.1.7 Analysis for Tenrec. Compared to the previously mentioned datasets, the Tenrec dataset is much larger, containing over 120 million interactions. This scale enables a comprehensive demonstration of the impact of dataset size on MSR tasks. Analysis of each domain's results from Table 9 indicates that performance in the domain # is inferior to the other two domains, attributed to its sparsity. Furthermore, a comparison of model performances reveals that M2M consistently achieves superior results, underscoring the robustness of its structural design across varying dataset sizes.

5.1.8 Analysis for Industrial Dataset. Our industrial dataset, derived from log samples on Huawei's advertising platforms, encompasses ten distinct domains. We present the overall performance results in Table 4 and the domain-specific results in Table 11. In comparison to other datasets, this industrial dataset features a significantly larger number of domains, facilitating our investigation into how domain number influences performance metrics and the observation of the "seesaw phenomenon". It is observed that SAR-Net and M2M exhibit superior performance on this dataset, demonstrating their enhanced ability to capture domain-specific features when faced with a large number of scenarios, attributing to the innovative design of the domain-specific transformer and meta cell.

5.1.9 Comprehensive Analysis. In analyzing Table 4's multi-scenario recommendation models, we underscore the challenge of managing the "seesaw effect" through effective scenario correlation modeling. The key lies in the model's ability to handle varying data distributions across scenarios, preventing overfitting in data-rich environments at the expense of data-sparse ones. This shows the necessity of fine-grained modeling of scenario relations in multi-scenario approaches.

In Table 4, models leveraging an expert structure (E.g., MMoE, PLE, SAR-Net) commonly outperform models that directly model different scenarios (E.g., SharedBottom, AdaptDHM), suggesting the former's superior capability in capturing complex inter-scenario dynamics at deeper network levels. Furthermore, Models that could dynamically adjust major structures or parameters (E.g., M2M, AdaSparse) depending on different scenarios surpass those with static expert structures, indicating a more precise control over hidden structures' influence on scenario performance. This leads to enhanced scenario correlation understanding and overall model performance.

Combined with Tables tables 7 to 11, we could summarize that dataset size does not directly correlate with model performance disparity. Instead, variability in sparse scenario performance significantly impacts model effectiveness, with top models maintaining high performance across scenarios and less effective models showing improvements only in specific sparse scenarios. This highlights the importance of scenario correlation modeling to mitigate scenario-specific data distribution impacts, facilitating stable performance improvements across all scenarios.

5.1.10 Efficiency Analysis. In analyzing the efficiency of different models across various datasets, we conducted extensive and rigorous experiments. These experiments measured the training time for one epoch, the inference time on the test set, and the parameter size for each model, and the results are shown in Table **??**. Adhering to the principles of a fair comparison, we observed that models exhibited a range of parameter sizes, which highlighted the trade-offs between model complexity and efficiency. For relatively small datasets, such as "MovieLens" and "Douban", the training times were notably lower, reflecting the reduced computational load compared to larger datasets like "Aliccp" and "Tenrec". It is evident that model efficiency is influenced not only by algorithmic design but also significantly by the characteristics of the dataset, including the number and intrinsic nature of features. This is a crucial consideration for applications with limited computational resources. Across different models, the model sizes remained within the same order of magnitude, primarily because most parameters in recommendation systems derive from embedding parameters. Our findings underscore the importance of selecting the appropriate model based on both the computational budget and the dataset's specific characteristics. We believe these efficiency results could serve as a reference for scholars to select suitable models or datasets based on their resources in practical machine learning applications.

6 CONCLUSION

In this paper, we introduce Scenario-Wise Rec, a pioneering benchmark designed specifically to tackle the complexities and challenges inherent in multi-scenario recommendation systems. Scenario-Wise Recaims to establish a comprehensive framework for facilitating fair and reproducible comparisons among diverse multi-scenario recommendation models, while also promoting the sharing of insights and advancements within this field. Our contributions are threefold. Firstly, to the best of our knowledge, Scenario-Wise Recis the first benchmark released in the field of multi-scenario recommendation, offering significant benefits for the community by enabling fair comparisons across different models and fostering development. Secondly, we have integrated a pipeline that includes multi-scenario data processing, model training, and evaluation, along with comprehensive logging and open-source practices. Scenario-Wise Recthus sets a new standard for transparency and reproducibility in the field and is friendly for all scholars. Thirdly, we provide the reproduction for eight multi-scenario recommendation models and five distinct multi-scenario datasets, offering scholars diverse angles to test and implement their models in this domain. This facilitates a

deeper understanding of the current landscape and identifies potential avenues for future research. We hope our benchmark will contribute to the field and collectively foster collaboration in the area of Multi-Scenario Recommendation.

REFERENCES

[1] Zeynep Batmaz, Ali Yurekli, Alper Bilge, and Cihan Kaleli. 2019. A review on deep learning for recommender systems: challenges and remedies. *Artificial Intelligence Review* 52 (2019), 1–37.

[2] Rich Caruana. 1997. Multitask learning. *Machine learning* 28 (1997), 41–75.

[3] Jianxin Chang, Chenbin Zhang, Yiqun Hui, Dewei Leng, Yanan Niu, Yang Song, and Kun Gai. 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.

[4] Yuting Chen, Yanshi Wang, Yabo Ni, An-Xiang Zeng, and Lanfen Lin. 2020. Scenario-aware and Mutual-based approach for Multi-scenario Recommendation in E-Commerce. In *2020 International Conference on Data Mining Workshops (ICDMW)*. IEEE, 127–135.

[5] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*. 7–10.

[6] Brian S Everitt and Anders Skrondal. 2010. The Cambridge dictionary of statistics. (2010).

[7] Wenqi Fan, Xiangyu Zhao, Xiao Chen, Jingran Su, Jingtong Gao, Lin Wang, Qidong Liu, Yiqi Wang, Han Xu, Lei Chen, et al. 2022. A comprehensive survey on trustworthy recommender systems. *arXiv preprint arXiv:2209.10117* (2022).

[8] Chenjiao Feng, Jiye Liang, Peng Song, and Zhiqiang Wang. 2020. A fusion collaborative filtering method for sparse data in recommender systems. *Information Sciences* 521 (2020), 365–379.

[9] Zichuan Fu, Xiangyang Li, Chuhan Wu, Yichao Wang, Kuicai Dong, Xiangyu Zhao, Mengchen Zhao, Huifeng Guo, and Ruiming Tang. 2023. A Unified Framework for Multi-Domain CTR Prediction via Large Language Models. *arXiv preprint arXiv:2312.10743* (2023).

[10] Jingtong Gao, Xiangyu Zhao, Bo Chen, Fan Yan, Huifeng Guo, and Ruiming Tang. 2023. AutoTransfer: Instance Transfer for Cross-Domain Recommendations. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1478–1487.

[11] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. *arXiv preprint arXiv:1703.04247* (2017).

[12] Wei Guo, Chenxu Zhu, Fan Yan, Bo Chen, Weiwen Liu, Huifeng Guo, Hongkun Zheng, Yong Liu, and Ruiming Tang. 2023. DFFM: Domain Facilitated Feature Modeling for CTR Prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 4602–4608.

[13] Pengyue Jia, Yichao Wang, Shanru Lin, Xiaopeng Li, Xiangyu Zhao, Huifeng Guo, and Ruiming Tang. 2024. D3: A Methodological Exploration of Domain Division, Modeling, and Balance in Multi-Domain Recommendations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 8553–8561.

[14] Hyunjun Ju, SeongKu Kang, Dongha Lee, Junyoung Hwang, Sanghwan Jang, and Hwanjo Yu. 2024. Multi-Domain Recommendation to Attract Users via Domain Preference Modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 8582–8590.

[15] Rohit Kumar, Sneha Manjunath Naik, Vani D Naik, Smita Shiralli, VG Sunil, and Moula Husain. 2015. Predicting clicks: CTR estimation of advertisements using logistic regression classifier. In *2015 IEEE international advance computing conference (IACC)*. IEEE, 1134–1138.

[16] Jinyun Li, Huiwen Zheng, Yuanlin Liu, Minfang Lu, Lixia Wu, and Haoyuan Hu. 2022. AdaptDHM: Adaptive Distribution Hierarchical Model for Multi-Domain CTR Prediction. *arXiv preprint arXiv:2211.12105* (2022).

[17] Pengcheng Li, Runze Li, Qing Da, An-Xiang Zeng, and Lijun Zhang. 2020. Improving multi-scenario learning to rank in e-commerce by exploiting task relationships in the label space. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2605–2612.

[18] Xiaopeng Li, Fan Yan, Xiangyu Zhao, Yichao Wang, Bo Chen, Huifeng Guo, and Ruiming Tang. 2023. HAMUR: Hyper Adapter for Multi-Domain Recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 1268–1277.

[19] Jianxun Lian, Xiaohuan Zhou, Fuzheng Zhang, Zhongxia Chen, Xing Xie, and Guangzhong Sun. 2018. xdeepfm: Combining explicit and implicit feature interactions for recommender systems. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1754–1763.

[20] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. 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.

[21] Wentao Ning, Xiao Yan, Weiwen Liu, Reynold Cheng, Rui Zhang, and Bo Tang. 2023. Multi-domain Recommendation with Embedding Disentangling and Domain Alignment. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 1917–1927.

[22] Steffen Rendle. 2010. Factorization machines. In *2010 IEEE International conference on data mining*. IEEE, 995–1000.

[23] Qijie Shen, Wanjie Tao, Jing Zhang, Hong Wen, Zulong Chen, and Quan Lu. 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.

[24] Xiang-Rong Sheng, Liqin Zhao, Guorui Zhou, Xinyao Ding, Binding Dai, Qiang Luo, Siran Yang, Jingshan Lv, Chi Zhang, Hongbo Deng, 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.

[25] Hongyan Tang, Junning Liu, Ming Zhao, and Xudong Gong. 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.

[26] Yu Tian, Bofang Li, Si Chen, Xubin Li, Hongbo Deng, Jian Xu, Bo Zheng, Qian Wang, and Chenliang Li. 2023. Multi-Scenario Ranking with Adaptive Feature Learning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 517–526.

[27] Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang. 2017. Deep & cross network for ad click predictions. In *Proceedings of the ADKDD'17*. 1–7.

[28] Ruoxi Wang, Rakesh Shivanna, Derek Cheng, Sagar Jain, Dong Lin, Lichan Hong, and Ed Chi. 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.

[29] Yichao Wang, Huifeng Guo, Bo Chen, Weiwen Liu, Zhirong Liu, Qi Zhang, Zhicheng He, Hongkun Zheng, Weiwei Yao, Muyu Zhang, 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.

[30] Yuhao Wang, Xiangyu Zhao, Bo Chen, Qidong Liu, Huifeng Guo, Huanshuo Liu, Yichao Wang, Rui Zhang, and Ruiming Tang. 2023. PLATE: A Prompt-Enhanced Paradigm for Multi-Scenario Recommendations. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1498–1507.

[31] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, et al. 2020. Mind: A large-scale dataset for news recommendation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. 3597–3606.

[32] Xu Xie, Fei Sun, Zhaoyang Liu, Shiwen Wu, Jinyang Gao, Jiandong Zhang, Bolin Ding, and Bin Cui. 2022. Contrastive learning for sequential recommendation. In *2022 IEEE 38th international conference on data engineering (ICDE)*. IEEE, 1259–1273.

[33] Xuanhua Yang, Xiaoyu Peng, Penghui Wei, Shaoguo Liu, Liang Wang, and Bo Zheng. 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.

[34] Guanghu Yuan, Fajie Yuan, Yudong Li, Beibei Kong, Shujie Li, Lei Chen, Min Yang, Chenyun Yu, Bo Hu, Zang Li, et al. 2022. Tenrec: A Large-scale Multipurpose Benchmark Dataset for Recommender Systems. *Advances in Neural Information Processing Systems* 35 (2022), 11480–11493.

[35] Qianqian Zhang, Xinru Liao, Quan Liu, Jian Xu, and Bo Zheng. 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.

[36] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep learning based recommender system: A survey and new perspectives. *ACM computing surveys (CSUR)* 52, 1 (2019), 1–38.

[37] Weinan Zhang, Jiarui Qin, Wei Guo, Ruiming Tang, and Xiuqiang He. 2021. Deep learning for click-through rate estimation. *arXiv preprint arXiv:2104.10584* (2021).

[38] Wayne Xin Zhao, Shanlei Mu, Yupeng Hou, Zihan Lin, Yushuo Chen, Xingyu Pan, Kaiyuan Li, Yujie Lu, Hui Wang, Changxin Tian, Yingqian Min, Zhichao Feng, Xinyan Fan, Xu Chen, Pengfei Wang, Wendi Ji, Yaliang Li, Xiaoling Wang, and Ji-Rong Wen. 2021. RecBole: Towards a Unified, Comprehensive and Efficient Framework for Recommendation Algorithms. In *CIKM*. ACM, 4653–4664.

[39] Guorui Zhou, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai. 2019. Deep interest evolution network for click-through rate prediction. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 5941–5948.

[40] Jie Zhou, Xianshuai Cao, Wenhao Li, Lin Bo, Kun Zhang, Chuan Luo, and Qian Yu. 2023. Hinet: Novel multi-scenario & multi-task learning with hierarchical information extraction. In *2023 IEEE 39th International Conference on Data*

Engineering (ICDE). IEEE, 2969–2975.

[41] Feng Zhu, Yan Wang, Chaochao Chen, Guanfeng Liu, and Xiaolin Zheng. 2020. A graphical and attentional framework for dual-target cross-domain recommendation.. In *IJCAI*, Vol. 21. 39.

[42] Feng Zhu, Yan Wang, Chaochao Chen, Jun Zhou, Longfei Li, and Guanfeng Liu. 2021. Cross-domain recommendation: challenges, progress, and prospects. In *30th International Joint Conference on Artificial Intelligence, IJCAI 2021*. International Joint Conferences on Artificial Intelligence, 4721–4728.

[43] Jiachen Zhu, Yichao Wang, Jianghao Lin, Jiarui Qin, Ruiming Tang, Weinan Zhang, and Yong Yu. 2024. M-scan: A Multi-Scenario Causal-driven Adaptive Network for Recommendation. *arXiv preprint arXiv:2404.07581* (2024).

A APPENDIX

In the appendix section, we offer supplementary information encompassing performance for each domain within all datasets we discussed before, along with detailed experimental details and settings for model implementation.

A.1 Domain-Specific results

In this part, we present a table of domain interaction distribution for the Huawei Industrial dataset and five tables detailing domain-specific results across all datasets for user reference.

Table 6: Domain Distribution for Huawei Industrial Dataset.

Domain Number	Interaction
Domain 9	655,569
Domain 7	459,370
Domain 6	383,791
Domain 4	316,734
Domain 0	301,654
Domain 1	91,468
Domain 8	87,353
Domain 2	22,986
Domain 5	16,288
Domain 3	10,928

Table 7: The Domain-detailed Results for Movie-Lens.

Models/AUC	Total	0#	1#	2#
Shared Bottom	0.8095	0.8116	0.8128	0.8041
MMOE	0.8086	0.8029	0.8178	0.8016
PLE	0.8091	0.8118	0.8187	0.8002
STAR	0.8096	0.8137	0.8133	0.7979
SAR-Net	0.8092	0.8068	0.8158	0.8026
M2M	0.8115	0.8111	0.8163	0.8057
AdaSparse	0.8108	0.8109	0.8188	0.7947
AdaptDHM	0.8083	0.8074	0.8160	0.7995
EPNet	0.8097	0.8100	0.8148	0.8031
PPNet	0.8063	0.8084	0.8113	0.7994
Models/Logloss	Total	0#	1#	2#
Shared Bottom	0.5228	0.5243	0.5208	0.5239
MMOE	0.5218	0.5239	0.5164	0.5262
PLE	0.5257	0.5335	0.5164	0.5310
STAR	0.5258	0.5239	0.5228	0.5299
SAR-Net	0.5245	0.5337	0.5180	0.5261
M2M	0.5213	0.5321	0.5208	0.5240
AdaSparse	0.5205	0.5248	0.5137	0.5400
AdaptDHM	0.5238	0.5293	0.5162	0.5283
EPNet	0.5215	0.5251	0.5178	0.5234
PPNet	0.5257	0.5266	0.5228	0.5281

Table 8: The Domain-detailed Results for Ali-CCP.

Models/AUC	Total	0#	1#	2#
Shared Bottom	0.6232	0.6279	0.5627	0.6246
MMOE	0.6242	0.6279	0.5744	0.6257
PLE	0.6250	0.6326	0.5841	0.6255
STAR	0.6253	0.6270	0.6041	0.6242
SAR-Net	0.6245	0.6282	0.5900	0.6253
M2M	0.6257	0.6278	0.6018	0.6247
AdaSparse	0.6239	0.6220	0.5926	0.6237
AdaptDHM	0.6233	0.6249	0.5823	0.6222
EPNet	0.6236	0.6257	0.5974	0.6222
PPNet	0.6144	0.6156	0.5591	0.6144
Models/Logloss	Total	0#	1#	2#
Shared Bottom	0.1628	0.1659	0.2001	0.1605
MMOE	0.1621	0.1652	0.1801	0.1600
PLE	0.1617	0.1644	0.1810	0.1597
STAR	0.1613	0.165	0.1786	0.1588
SAR-Net	0.1616	0.1646	0.1797	0.1589
M2M	0.1611	0.1649	0.1788	0.1585
AdaSparse	0.1614	0.1660	0.1793	0.1594
AdaptDHM	0.1619	0.1651	0.1795	0.1587
EPNet	0.1612	0.1648	0.1790	0.1587
PPNet	0.1622	0.1655	0.1881	0.1599

Table 9: The Domain-detailed Results for Tenrec.

Models/AUC	Total	0#	1#	2#
Shared Bottom	0.7748	0.7928	0.7800	0.7413
MMOE	0.7750	0.7929	0.8527	0.7430
PLE	0.7749	0.7928	0.8561	0.7427
STAR	0.7737	0.7918	0.7800	0.7413
SAR-Net	0.7741	0.7919	0.7844	0.7423
M2M	0.7819	0.8003	0.8130	0.7499
AdaSparse	0.7752	0.7936	0.7953	0.7427
AdaptDHM	0.7748	0.7920	0.8428	0.7426
EPNet	0.7817	0.7956	0.7888	0.7447
PPNet	0.7749	0.7930	0.7951	0.7426
Models/Logloss	Total	0#	1#	2#
Shared Bottom	0.4575	0.4446	0.0174	0.4858
MMOE	0.4579	0.4445	0.0167	0.4855
PLE	0.4575	0.4446	0.0169	0.4857
STAR	0.4584	0.4454	0.0186	0.4866
SAR-Net	0.4582	0.4453	0.0204	0.4864
M2M	0.4527	0.4392	0.0173	0.4816
AdaSparse	0.4568	0.5120	0.0220	0.5404
AdaptDHM	0.4575	0.4451	0.0168	0.4855
EPNet	0.4559	0.4416	0.0163	0.4891
PPNet	0.4576	0.4448	0.0167	0.4858

Table 10: The Domain-detailed Results for KuaiRand.

Models/AUC	Total	0#	1#	2#	3#	4#
Shared Bottom	0.7793	0.7117	0.7282	0.7898	0.7293	0.8535
MMOE	0.7794	0.7146	0.7272	0.7773	0.7310	0.8562
PLE	0.7796	0.7104	0.7285	0.7890	0.7298	0.8531
STAR	0.7806	0.7201	0.7305	0.7895	0.7322	0.8055
SAR-Net	0.7816	0.7263	0.7312	0.7921	0.7359	0.8378
M2M	0.7821	0.7248	0.7326	0.7898	0.7339	0.8447
AdaSparse	0.7816	0.7243	0.7314	0.7889	0.7332	0.8227
AdaptDHM	0.7773	0.7258	0.7244	0.7887	0.7349	0.8071
EPNet	0.7801	0.7235	0.7303	0.7883	0.7319	0.7803
PPNet	0.7800	0.7167	0.7285	0.7887	0.7329	0.8642
Models/Logloss	**Total**	**0#**	**1#**	**2#**	**3#**	**4#**
Shared Bottom	0.5483	0.3532	0.6074	0.5357	0.6092	0.3454
MMOE	0.5477	0.3510	0.6069	0.5507	0.6110	0.3344
PLE	0.5495	0.3517	0.6092	0.5479	0.6078	0.3444
STAR	0.5404	0.3335	0.6019	0.5331	0.6003	0.3753
SAR-Net	0.5393	0.3319	0.6014	0.5307	0.6023	0.3467
M2M	0.5397	0.3324	0.6012	0.5340	0.6011	0.3436
AdaSparse	0.5399	0.3333	0.6014	0.5350	0.6015	0.3604
AdaptDHM	0.5436	0.7269	0.6064	0.5330	0.5986	0.3875
EPNet	0.5411	0.3340	0.6022	0.5344	0.6013	0.3942
PPNet	0.5408	0.3353	0.6033	0.5331	0.6006	0.3191

Table 11: The Domain-detailed Results for Industrial Dataset.

Models/AUC	Total	0#	1#	2#	3#	4#	5#	6#	7#	8#	9#
Shared Bottom	0.8276	0.6680	0.7176	0.8194	0.7951	0.8238	0.8740	0.8420	0.6833	0.7653	0.8227
MMOE	0.8301	0.6684	0.7551	0.8808	0.7351	0.8251	0.8501	0.8407	0.7241	0.7752	0.8371
PLE	0.8330	0.6494	0.7240	0.8130	0.7648	0.8195	0.9262	0.8474	0.6999	0.7317	0.8323
STAR	0.8310	0.6649	0.7351	0.8070	0.7179	0.7921	0.8529	0.8191	0.6728	0.7024	0.8109
SAR-Net	0.8355	0.6580	0.7382	0.8903	0.7678	0.8286	0.9598	0.8484	0.7413	0.7581	0.8417
M2M	0.8392	0.6534	0.7114	0.8770	0.7584	0.8257	0.8823	0.8504	0.7256	0.7596	0.8462
AdaSparse	0.8224	0.6482	0.7350	0.8842	0.7489	0.7617	0.9212	0.8387	0.6854	0.7629	0.8230
AdaptDHM	0.8358	0.6592	0.7103	0.8969	0.7705	0.8254	0.9219	0.8534	0.7145	0.7808	0.8460
Models/Logloss	**Total**	**0#**	**1#**	**2#**	**3#**	**4#**	**5#**	**6#**	**7#**	**8#**	**9#**
Shared Bottom	0.1521	0.1405	0.1863	0.0853	0.1406	0.1259	0.0362	0.2062	0.0584	0.1497	0.1959
MMOE	0.1567	0.1408	0.1690	0.0722	0.1705	0.1263	0.0281	0.2038	0.0562	0.1535	0.1948
PLE	0.1496	0.1514	0.1781	0.0713	0.1501	0.1131	0.0311	0.1896	0.0593	0.1521	0.2001
STAR	0.1503	0.1632	0.1793	0.0977	0.2006	0.1198	0.0532	0.2021	0.0719	0.1374	0.2117
SAR-Net	0.1528	0.1509	0.1811	0.0817	0.1941	0.1486	0.0335	0.2336	0.0597	0.1672	0.2108
M2M	0.1494	0.1442	0.1820	0.0840	0.1687	0.1260	0.0314	0.2009	0.0590	1.1488	0.1897
AdaSparse	0.1596	0.1594	0.1867	0.0922	0.1727	0.1508	0.0297	0.2180	0.0792	0.1642	0.1968
AdaptDHM	0.1489	0.1438	0.1843	0.0745	0.1510	0.1253	0.0277	0.1981	0.0551	0.1438	0.1861

A.2 Experiment Settings

A.2.1 Implementation Details. In this part, we present the experiment setting during our experiment. Our framework is implemented using PyTorch. Empirically, we set the feature embedding dimension d to 16. We customized batch sizes for each dataset: 4096 for MovieLens, 9,048 for both Kuairand and the Huawei industrial dataset, and 102,400 for Aliccp and Tenrec. Experiments were conducted on a single GPU of Tesla V100 PCIe 32GB, utilizing the Adam optimizer. The initial learning rate was set to 1e-3. To enhance training performance, we incorporated an early stopping strategy and a learning rate scheduler for optimal adjustment. All

Table 12: The Domain-detailed Results for Douban.

Models/AUC	Total	0#	1#	2#
SharedBottom	0.7993	0.7144	0.7349	0.8119
MMOE	0.7978	0.7098	0.7317	0.8111
PLE	0.7979	0.7142	0.7342	0.8109
STAR	0.7957	0.7080	0.7292	0.8089
SAR-Net	0.8033	0.7220	0.7451	0.8154
M2M	0.7962	0.7004	0.7160	0.8145
AdaSparse	0.7963	0.7073	0.7279	0.8096
AdaptDHM	0.8003	0.7124	0.7287	0.8142
Models/Logloss	Total	0#	1#	2#
SharedBottom	0.5178	0.5531	0.4952	0.5147
MMOE	0.5192	0.5563	0.4981	0.5156
PLE	0.5196	0.5543	0.4955	0.5169
STAR	0.5218	0.5581	0.4998	0.5185
SAR-Net	0.5131	0.5487	0.4895	0.5101
M2M	0.5229	0.5681	0.5147	0.5160
AdaSparse	0.5216	0.5577	0.4997	0.5184
AdaptDHM	0.5187	0.5604	0.5018	0.5137

Table 13: The Domain-detailed Results for Mind.

Models/AUC	Total	0#	1#	2#	3#
SharedBottom	0.7509	0.7675	0.7007	0.7569	0.7356
MMOE	0.7508	0.7670	0.7001	0.7572	0.7358
PLE	0.7503	0.7668	0.6993	0.7565	0.7351
STAR	0.7512	0.7678	0.7007	0.7577	0.7351
SAR-Net	0.7490	0.7653	0.6984	0.7557	0.7338
M2M	0.7508	0.7675	0.7010	0.7566	0.7344
AdaSparse	0.7497	0.7664	0.6999	0.7564	0.7341
AdaptDHM	0.7328	0.7480	0.6737	0.7444	0.7203
Models/Logloss	Total	0#	1#	2#	3#
SharedBottom	0.1600	0.1578	0.1662	0.1815	0.1351
MMOE	0.1600	0.1578	0.1662	0.1814	0.1351
PLE	0.1601	0.1579	0.1662	0.1815	0.1352
STAR	0.1601	0.1578	0.1662	0.1816	0.1352
SAR-Net	0.1604	0.1582	0.1666	0.1818	0.1354
M2M	0.1601	0.1578	0.1661	0.1816	0.1352
AdaSparse	0.1644	0.1622	0.1699	0.1854	0.1407
AdaptDHM	0.1629	0.1611	0.1695	0.1839	0.1368

experiments were conducted three times under different random seeds.

A.2.2 Model Reproduction Details. In this part, we provide the reproduction details for each model, serving as a reference for users.

- **Shared Bottom** Our Shared Bottom code implementation comprises a single-layer MLP at the bottom, followed by domain-specific MLP towers for each domain. Considering the dataset sizes, we configured the MLP towers with three layers for the MovieLens, KuaiRand, and Industrial datasets and six layers for the other two datasets. We search the dimension bottom layer in {128, 256, 512}.
- **MMOE** Our MMOE module is consistent with the original paper [20]. During our experiment, we search the space of expert dimension {128, 256, 512} and for the output tower, we choose three layers of MLP for MovieLens, KuaiRand, and Industrial datasets and six layers for the other two datasets.
- **PLE** In our PLE implementation, unlike the implementation used in multi-task recommendation models, we replaced the task-specific and task-shared experts with domain-specific and domain-shared experts. Our exploration space including CGC layers {1, 2} and expert dimensions {128, 256, 512}. Regarding the output tower design, we adhered to the configurations employed in both MMOE and Shared Bottom models.
- **SAR-Net** In SAR-Net implementation, there are deviations from the method described in the original paper. Specifically, we omitted the cross-scenario behavior extraction layer, a design intended to process user behavior sequences, because our dataset lacks such features. Consequently, this module was excluded from our implementation. Our exploration space for the configuration included domain-shared expert counts within {2, 4, 8} and domain-specific expert counts within {1, 2}.
- **STAR** In reproducing the STAR model, our implementation remains strictly consistent with the specifications outlined in the original paper. We employ a single-layer network for the auxiliary network, and for the domain tower, MLPs are utilized. The configuration of the domain tower is set with three layers for the MovieLens, KuaiRand, and Huawei datasets, while six layers are designated for the Ali-CCP and Tenrec datasets, aligning with previous settings. We explored auxiliary network dimensions within the searching space {8, 16, 32}.
- **M2M** In our reproduction of the M2M model, which originally focus on multi-scenario multi-task problems, our work focuses on a single task—CTR prediction. Thus, accordingly, we adapted it for a single-task tower. Our exploration space comprised the expert output size within {8, 16}, the number of encoding layers within {1, 2}, the number of decoding layers within {2, 3}, and the feedforward dimension within {128, 256, 512}.
- **AdaSparse** In our replication of the AdaSparse model, as detailed in the original paper [33], we initially employ a domain-adaptive pruner module. This module offers three instantiation strategies: "Binarization", "Scaling", and "Fusion". Each represents distinct approaches to computing weighting factors. Subsequently, this adaptive pruning technique is utilized to facilitate a sparse MLP for CTR prediction across varied scenarios, demonstrating its flexibility in handling sparse data environments. We employ the "Fusion" strategy for all datasets, without losing generality. The backbone network is chosen for three and six, respectively, for different datasets like STAR and Shared Bottom. And we set α is 1 and the searching space for β is {2, 3, 4}.
- **AdaptDHM** In the reproduction of the AdaptDHM model, we commence by establishing a shared fully connected network dedicated to modeling correlations across different scenarios. This is complemented by the construction of several scenario-specific fully connected networks, aimed at conducting nuanced, scenario-specific analyses. Furthermore, a Distribution Learning Module (DLM) is developed as illustrated in the original

paper [16], employing a clustering algorithm based on cosine similarities to enable dynamic routing during both training and inference phases, thereby enhancing the model's adaptability to diverse data distributions. For the shared fully-connected-network, we follow the previously mentioned setting: three layers for dataset Movie-Lens, KuaiRand, industrial dataset, and six layers for Ali-CCP and Tenrec. Besides, we search the space of the number of cluster in {3, 4, 5}.

- **EPNet** In constructing the EPNet, we first built the Gate NU module to provide gated scaling signals for the model. Then, we divide the input into domain-side features and domain-agnostic features (i.e., sparse features and dense features), respectively, and embed them into embedding vectors. Afterward, we construct the scaled embedding by input the domain-side embedding and detached domain-agnostic embeddings to the GateNU module and applying the output scaling parameters to the original embedding. To avoid the effects of the PPNet structure, through a simple parameter search, we replace the subsequent network about PPNet in the original paper with a three-layer feedforward structure with different neurons according to different datasets and add an output header to output values between [0, 1].

- **PPNet** In developing the PPNet model, we adhered to the design outlined in paper [3]. Initially, we concatenate ID embeddings and input them into Gate NU modules. The number of Gate NU modules is the same as the number of PPNet layers. Subsequently, we constructed the PPNet tower. Given that PPNet was originally designed for multi-task learning, we adhered to our initial settings, assigning different task-specific architectures within the domain tower. We configured PPNet with multi-layer perceptrons (MLPs) tailored to various dataset distributions. For each instance, the input is directed to an appropriate domain tower based on its "domain indicator."