MAC: A Conversion Rate Prediction Benchmark Featuring Labels Under Multiple Attribution Mechanisms

Jinqi Wu*
jinqiwu001@gmail.com
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
School of Intelligence Science and Technology, Nanjing University
Suzhou, China

Sishuo Chen*
Zhangming Chan
Yong Bai
chensishuo.css@alibaba-inc.com
zhangming.czm@alibaba-inc.com
baiyong.by@alibaba-inc.com
Taobao & Tmall Group of Alibaba
Beijing, China

Lei Zhang
Sheng Chen
Chenghuan Hou
zl165646@alibaba-inc.com
chensheng.cs@alibaba-inc.com
jinyao@alibaba-inc.com
Taobao & Tmall Group of Alibaba
Beijing, China

Xiang-Rong Sheng
Han Zhu
Jian Xu
xiangrong.sxr@alibaba-inc.com
zhuhan.zh@alibaba-inc.com
xiyu.xj@alibaba-inc.com
Taobao & Tmall Group of Alibaba
Beijing, China

Bo Zheng
bozheng@alibaba-inc.com
Taobao & Tmall Group of Alibaba
Beijing, China

Chaoyou Fu†
bradyfu24@gmail.com
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
School of Intelligence Science and Technology, Nanjing University
Suzhou, China

Abstract

Multi-attribution learning (MAL), which enhances model performance by learning from conversion labels yielded by multiple attribution mechanisms, has emerged as a promising learning paradigm for conversion rate (CVR) prediction. However, the conversion labels in public CVR datasets are generated by a single attribution mechanism, hindering the development of MAL approaches.

To address this data gap, we establish the **M**ulti-**A**ttribution Ben**C**hmark (**MAC**), the first public CVR dataset featuring labels from multiple attribution mechanisms. Besides, to promote reproducible research on MAL, we develop **PyMAL**, an open-source library covering a wide array of baseline methods. We conduct comprehensive experimental analyses on **MAC** and reveal three key insights: **(1)** MAL brings consistent performance gains across different attribution settings, especially for users featuring long conversion paths. **(2)** The performance growth scales up with objective complexity in most settings; however, when predicting first-click conversion targets, simply adding auxiliary objectives is counterproductive, underscoring the necessity of careful selection of auxiliary objectives. **(3)** Two architectural design principles are paramount: first, to fully learn the multi-attribution knowledge, and second, to fully leverage this knowledge to serve the main task. Motivated by these findings, we propose **M**ixture **o**f **A**symmetric **E**xperts (**MoAE**), an effective MAL approach incorporating multi-attribution knowledge learning and main task-centric knowledge utilization. Experiments on **MAC** show that **MoAE** substantially surpasses the existing state-of-the-art MAL method. We believe that our benchmark and insights will foster future research in the MAL field. Our **MAC** benchmark and the **PyMAL** algorithm library are publicly available at https://github.com/alimama-tech/PyMAL.

CCS Concepts

• **Applied computing** → **Electronic commerce**; • **Information systems** → **Computational advertising**; • **Computing methodologies** → **Ranking**.

Keywords

Multi-Attribution Learning, Conversion Rate Prediction, Online Advertising, Dataset and Benchmark

ACM Reference Format:
Jinqi Wu, Sishuo Chen, Zhangming Chan, Yong Bai, Lei Zhang, Sheng Chen, Chenghuan Hou, Xiang-Rong Sheng, Han Zhu, Jian Xu, Bo Zheng, and Chaoyou Fu. 2026. MAC: A Conversion Rate Prediction Benchmark Featuring Labels Under Multiple Attribution Mechanisms. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26), August 09–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3770855.3817488

*Equal contribution.
†Corresponding author.

This work is licensed under a Creative Commons Attribution 4.0 International License.
KDD '26, Jeju Island, Republic of Korea
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3817488

1 Introduction

Conversion rate (CVR) prediction is a cornerstone of online advertising systems, directly influencing bidding efficiency, traffic allocation, and ultimately, platform revenue [6, 21]. At its core, the logic of label generation and model optimization is determined by

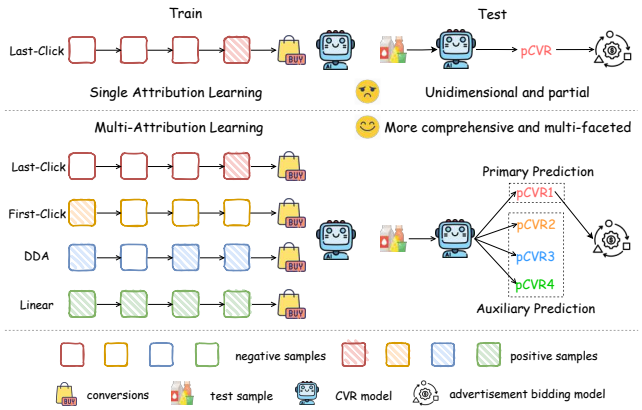


Figure 1: Illustration of single-attribution learning and multi-attribution learning (MAL) for CVR prediction. Our work paves the way for studying MAL by providing the first public benchmark, systematic evaluation, and valuable insights.

attribution mechanisms [28, 40]–the protocols allocating conversion credits across user touchpoints (usually operationalized as ad clicks).

Currently, nearly all existing CVR prediction models, in both industry and academia, are trained and evaluated under a **single-attribution assumption** [7], where conversion labels are derived from a single attribution mechanism. This paradigm constrains models to learning a partial view of touchpoint value, failing to capture the user's complex conversion mindset holistically.

Recently, **multi-attribution learning (MAL)** [7] has emerged as a promising new paradigm that enhances model performance by jointly learning from conversion labels generated by multiple attribution mechanisms, such as first-click, last-click, and linear attribution rules, and data-driven attribution (DDA) models [14, 40]. By integrating multi-faceted auxiliary signals from diverse attribution views, MAL enables the model to develop a more comprehensive and holistic understanding of touchpoint contributions, which leads to significantly higher prediction accuracy under the system-optimized attribution mechanism and ultimately drives substantial growth in user satisfaction and platform revenue. Figure 1 compares the conventional single-attribution learning and the novel multi-attribution learning paradigm.

However, despite its considerable promise, the advancement of research on MAL is constrained by a critical data bottleneck: to the best of our knowledge, all publicly available CVR datasets, such as Criteo [6], Ali-CCP [24], and Taobao [48], only provide conversion labels from a single attribution mechanism.

To address this, we establish the **M**ulti-**A**ttribution Ben**C**hmark (**MAC**), the first CVR prediction benchmark **featuring conversion labels from multiple attribution mechanisms** to our knowledge. For each ad click, **MAC** provides conversion labels yielded by four typical attribution mechanisms: *last-click, first-click, linear, and data-driven attribution (DDA)*, which are illustrated in Figure 2. **MAC** is collected from the advertising system of Taobao, Alibaba, and rigorously anonymized for public release, providing a reliable testbed for assessing MAL approaches for CVR prediction.

Moreover, to facilitate reproducible research on MAL, we release **PyMAL**, an open-source library covering representative baselines, ranging from classic multi-task models [23, 34, 36] to the cutting-edge NATAL model [7][1]. Using **PyMAL**, we perform a systematic evaluation on **MAC** and discover three key insights:

(1) **The Generality of MAL**: MAL consistently improves prediction accuracy across different target attribution mechanisms, with the most notable gains observed in users featuring long conversion journeys.
(2) **The Effect of Auxiliary Objective Choice in MAL**: While stacking more auxiliary targets from different attribution mechanisms typically enhances performance, it is not always the case. Under the first-click setting, indiscriminate addition of auxiliary tasks hurts model generalization, highlighting the importance of strategically selecting complementary attribution views.
(3) **Structure Design Principles for Effective MAL**: We identify two key principles: (i) **fully learning multi-attribution knowledge**, as achieved by the mixture-of-experts (MoE) structures in MTL models like MMoE [23] and PLE [34]; (ii) **leveraging this knowledge in a main-task-prioritized manner**, as realized through asymmetric knowledge transfer in NATAL [7].

Inspired by the identified principles, we propose **M**ixture **o**f **A**symmetric **E**xperts (**MoAE**), a novel MAL model that simultaneously satisfies both design principles by incorporating an MoE backbone on the bottom for comprehensive multi-attribution knowledge learning, complemented by a main-task-prioritized asymmetric transfer module on the top. Extensive experiments on **MAC** show that **MoAE** significantly outperforms existing MAL models across all settings, raising the GAUC metric by up to **0.39pt**[2]. Besides improving the main task metric, **MoAE** also surpasses the existing state-of-the-art model NATAL [7] in terms of auxiliary task metrics. This confirms that our MoAE method more effectively captures the incremental knowledge from multi-attribution labels.

We summarize our contributions as follows:

- **Data Resource**: We release **the first multi-attribution CVR benchmark** named **MAC**. Covering conversion labels derived from four representative attribution mechanisms, **MAC** fills the key data gap in multi-attribution learning and provides a foundational benchmark for future research.
- **Software Toolkits and Comprehensive Benchmarking**: We release **PyMAL**, an open-source library covering typical MAL approaches, and benchmark them on **MAC**. Our thorough analysis of the benchmarking results confirms the universality of MAL and offers actionable insights for designing critical components in MAL, such as auxiliary objectives and model architecture.
- **Effective Approach**: We propose **MoAE**, a novel MAL model that integrates the MoE architecture for efficiently capturing multi-attribution knowledge and a main-task-oriented knowledge transfer mechanism. Extensive evaluation results show that **MoAE** substantially outperforms existing methods on **MAC**.

[1] NATAL denotes the k**N**owledge **A**ggregation with the car**T**esian **A**uxi**L**iary training model proposed in Chen et al. [7].
[2] "pt" is short for percentage points throughout this paper.

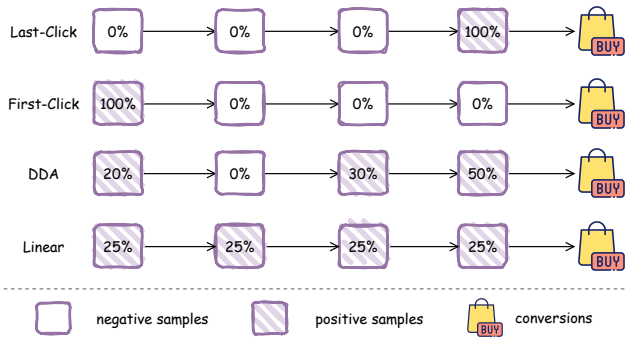


Figure 2: The four conversion attribution mechanisms covered in our MAC benchmark.

2 Background and Related Works

2.1 CVR Prediction and Attribution Mechanism

Conversion rate (CVR) prediction [5, 6, 21, 24] plays a vital role in online advertising. Researchers have pursued numerous avenues to improve CVR prediction models, such as joint learning with other tasks [22, 24, 35, 46], model structure [49, 50], and delayed feedback modeling [8, 11, 19, 39]. Although **attribution mechanisms** [28, 40], namely the rules for allocating conversion credits to ad clicks, directly determine the learning goals of CVR prediction models, they are overlooked in existing studies, which all rely on a single attribution mechanism. Chen et al. [7] proposes multi-attribution learning (MAL) and achieves significant performance gains on in-house data, highlighting the potential of the MAL paradigm. Nevertheless, existing public CVR datasets [6, 24, 48] all use a single attribution mechanism to generate labels, hindering advances in MAL. To the best of our knowledge, we release the first CVR dataset with multi-attribution labels, and further contribute the first algorithmic codebase and in-depth analysis for MAL approaches.

2.2 Multi-Task Learning for Recommendation

Multi-task learning (MTL) [3, 25, 31, 44] seeks to train a single model for predicting multiple targets simultaneously. Researchers have made extensive efforts to apply MTL to recommendation tasks, ranging from general structure design [15, 23, 34, 36, 41] to specific task grouping [4, 7, 17, 24, 35, 46]. Chen et al. [7] regard MAL as an MTL problem and apply classical MTL models to MAL, such as MMoE [23] and PLE [34], where they underperform the NATAL model tailored for MAL. In this study, we will delve deeper into the strengths and weaknesses of the MTL methods within the MAL context, aiming to collect insights for future architectural design.

3 The MAC Benchmark

3.1 The MAL Problem Setup

In CVR prediction, the conventional paradigm assumes a **single attribution mechanism** $\mathcal{A}$ (e.g., last-click or linear) that assigns a binary label $y^{\mathcal{A}} \in \{0, 1\}$ to each ad click, indicating whether the click is credited for a subsequent conversion within a fixed attribution window. However, this view is inherently limited: different attribution mechanisms encode distinct perspectives on user intent and touchpoint contribution, and no single rule fully captures the true conversion causality. To overcome this limitation, we formalize **multi-attribution learning (MAL)** as follows. Let $\mathcal{M} = \{\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_K\}$ be a set of K attribution mechanisms (e.g., last-click, first-click, linear, and data-driven attribution (DDA)). For each click instance $\mathbf{x}$, there is a set of **continuous attribution weights** $\{w^{\mathcal{A}_1}, w^{\mathcal{A}_2}, \ldots, w^{\mathcal{A}_K}\}$, where $w^{\mathcal{A}_i} \in \mathbb{R}_{\geq 0}$ denotes the contribution of the click to related conversions under the mechanism $\mathcal{A}_i$, computed over a fixed attribution window.

In MAL, one attribution mechanism $\mathcal{A}_t \in \mathcal{M}$ is designated as the **target attribution mechanism** according to business requirements. Predicting the conversion labels under $\mathcal{A}_t$ is the **primary task**, while the remaining $K - 1$ mechanisms serve as **auxiliary attribution mechanisms**, and the labels yielded by them are **auxiliary targets**. The goal is to train a prediction model $f_\theta(\cdot)$ such that the estimate $\hat{y}^{\mathcal{A}_t} = f_\theta(\mathbf{x})$ accurately predicts the binary conversion label derived from the attribution weight $w^{\mathcal{A}_t}$ by jointly leveraging supervision signals from all K attribution views. This setup enables the model to acquire a more holistic understanding of post-click conversion behavior.

Table 1: Summary of existing CVR prediction datasets and our MAC benchmark. (Notes: BS = Behavior Sequence, MF = Multimodal Feature, MA = Multi-Attribution Labels.)

Dataset	#Clicks	#Items	#Users	BS	MF	MA
Criteo [6]	3.6M	15.9M	5.0K	✗	✗	✗
Ali-CCP [24]	3.4M	4.3M	0.4M	✗	✗	✗
Taobao [48]	100M	4.0M	1.0M	✗	✗	✗
MAC	79M	15.1M	0.8M	✓	✓	✓

3.2 Benchmark Construction

3.2.1 Overview. Existing public CVR datasets [6, 24, 48] only provide labels under one fixed attribution mechanism, making systematic research on MAL infeasible. To break the data bottleneck, we introduce the Multi-Attribution BenChmark (**MAC**), **the first open-source CVR dataset that provides multi-dimensional attribution weights under different attribution mechanisms**. **MAC** is sampled from real-world industrial logs of the advertising system of Taobao and covers four types of popular attribution mechanisms: last-click, first-click, linear, and data-driven attribution (DDA) [40]. This characteristic enables a systematic investigation of MAL approaches for the research community. We summarize the statistics and key features of **MAC** and existing benchmarks in Table 1 and list the positive sample ratios under different mechanisms in **MAC** in Table 2. The rules of conversion attribution in each attribution mechanisms are given as follows:

- **Last-Click**: Concentrates 100% attribution weight on the final click preceding conversion.
- **First-Click**: Attributes full credit to the initial click.
- **Linear**: Distributes weights uniformly across all touchpoints.
- **Data-driven attribution (DDA)**: Learns attribution weight allocation through causal inference models [1, 13, 16, 40].

Table 2: The positive sample ratio under different attribution mechanisms in our MAC benchmark.

	Last-Click	First-Click	DDA	Linear
Positive Ratio	1.6%	1.8%	4.3%	5.3%

3.2.2 Data Sampling Criteria. **MAC** covers 21 consecutive days of traffic on Taobao, one of the world's largest e-commerce platforms. To ensure that the dataset size is suitable for academic research, we performed stratified sampling, where we assigned higher sampling rates to highly active users and lower rates to less active ones. The resulting **MAC** dataset includes 0.8 million users, 79 million clicks, and 15.1 million distinct items. As shown in Table 1, **MAC** is comparable to or larger than existing datasets [6, 24, 48].

3.2.3 Feature and Label Schema. Each ad click sample in **MAC** contains three groups of signals:

- *Categorical features*: 7 user features including user ID and profile features, 10 item features such as item, shop, category, and adgroup IDs, and 3 context features such as the ad scenario ID.
- *Behavior sequence features*: the user purchase item sequence with a maximum truncated length of 20. For each item, we provide the item, shop, and category IDs along with its visual similarity score to the target advertised item, using the features from an image encoder pre-trained through contrastive learning [30].
- *Multi-attribution conversion labels*: **MAC** provides continuous attribution weights for each ad click under all four attribution mechanisms as stated in § 3.2.1, derived from the conversion paths within a fixed, confidential attribution window. We use the CausalMTA model [40] as the DDA model and use direct conversion attribution, which means attributing conversions only to clicks on ads within the same item.

3.2.4 Data Compliance and Availability. Our **MAC** benchmark has undergone rigorous anonymization to remove personally sensitive information. All personally identifiable information and business details, such as user, shop, and adgroup identifiers, are irreversibly hashed. The dataset is constructed from historical traffic logs and does not contain real-time operational metrics or business conditions due to the user sampling procedure. We release **MAC** as a public benchmark to support future research on CVR prediction on https://huggingface.co/datasets/alimamaTech/MAC.

3.3 Evaluation Metrics

Following industrial practice [7], we regard CVR prediction as a binary classification problem derived from attribution weights:

- **Positive Samples.** Samples with positive attribution weights are treated as positive samples.
- **Negative Samples.** Samples with zero attribution weights are treated as negative samples.

For evaluating model performance, we report the widely-adopted **AUC** and Group AUC (**GAUC**) metrics [2, 27, 29, 47] of the primary task. AUC is a widely used ranking metric that evaluates a model's ability to distinguish positive and negative samples, making it well-suited for CVR prediction. The definition of AUC is given in Eq. (1)

$$\text{AUC} = \frac{1}{n_+ n_-} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} \left[\mathbb{I}(y_i^+ > y_j^-) + \frac{1}{2} \mathbb{I}(y_i^+ = y_j^-) \right] \tag{1}$$

Here n_+ and n_- denote the numbers of positive and negative samples, respectively, y_i^+ and y_j^- stand for the predicted conversion probabilities for the i-th positive and j-th negative sample, respectively, and $\mathbb{I}(\cdot)$ is the indicator function.

Moreover, we report GAUC, which captures the model's ability to rank ad clicks within individual user groups and serves as the primary evaluation criterion in our production system. It has been empirically shown to align more closely with online performance than AUC. The definition of GAUC is given in Eq. (2):

$$\text{GAUC} = \frac{\sum_{u=1}^{U} \#\text{click}(u) \times \text{AUC}_u}{\sum_{u=1}^{U} \#\text{click}(u)} \tag{2}$$

Here U represents the number of users, $\#\text{click}(u)$ denotes the number of clicks for the u-th user, and AUC_u is the AUC computed on the samples from the u-th user.

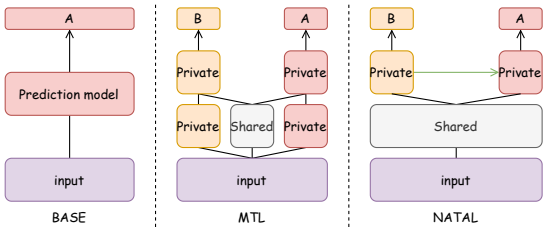


Figure 3: Illustration of three families of baselines. A is the primary target under the target attribution mechanism, and B is an auxiliary target from other attribution mechanisms.

4 The PyMAL Library

To foster reproducible and comparable research on multi-attribution learning (MAL) for CVR prediction, we establish **PyMAL**, an open-source Python library built on top of PyTorch [26] available at GitHub. It provides a unified, modular, and extensible framework that implements a comprehensive suite of MAL baselines, and it can be easily extended to incorporate ongoing efforts in the MAL field. The library supports rapid prototyping and fair comparison of MAL approaches, which paves the way for studies on MAL in the community.

We categorize the implemented models in **PyMAL** into three representative families as follows, reflecting the evolution of modeling paradigms in MAL. Figure 3 illustrates their model structures.

First, the single-attribution learning baseline. The **BASE** model is trained solely on the labels generated by the target attribution mechanism (e.g., last-click), ignoring the auxiliary signals from other attribution mechanisms. Regarding model architecture, it comprises target attention [47] for sequence modeling, SimTier for multimodal modeling [30], and an MLP predictor for outputting the predicted CVR value, following prior industrial practice [7].

Second, the multi-task learning (MTL) models adapted for MAL. These models jointly learn the primary task together with

one or more auxiliary attribution tasks, using shared or structured representations to capture cross-task knowledge:

- **Shared-Bottom** [3]: A classic architecture featuring a shared bottom embedding layer followed by task-specific output towers.
- **MMoE** [23]: A mixture-of-experts model that uses task-specific gating networks to dynamically combine shared expert subnetworks, enabling flexible knowledge sharing.
- **PLE** [34]: A progressive layered extraction architecture that explicitly separates shared and task-specific experts to reduce negative transfer and improve task cooperation.
- **HoME** [36]: An enhanced MoE variant that introduces Self-Gate and Feature-Gate mechanisms to adaptively modulate expert selection based on input feature granularity, coupled with Swish activation for better nonlinearity modeling.

Last, the state-of-the-art NATAL model tailored for MAL. NATAL [7], short for k**N**owledge **A**ggregation with the car**T**esian **A**uxi**L**iary training, goes beyond conventional MTL by explicitly modeling attribution interactions and prioritizing the primary objective. Concretely, it incorporates three key components: (1) *Attribution Knowledge Aggregation* (AKA) for fusing supervision from multiple attribution mechanisms via intermediate feature transfer; (2) *Primary Target Predictor* (PTP) for generating the prediction under the target attribution mechanism; and (3) *Cartesian-based Auxiliary Training* (CAT), a Cartesian product-based auxiliary task that integrates higher-order attribution interactions, whose details are given in Appendix E.

5 Method: Mixture of Asymmetric Experts

5.1 Motivation: Structure Design Principles

To understand the pros and cons of prior methods, we record their performance under both the target and auxiliary attribution mechanisms. As shown in Table 3, the advanced MTL model equipped with the MoE structure, e.g., PLE and HoME, not only raises the primary task performance, but also improves the auxiliary task metrics compared with Shared-Bottom. This is reasonable, as it is known that MoE can capture task commonalities and reduce gradient conflicts, thereby improving the performance of all subtasks in MTL [23, 34, 36]. In contrast, NATAL discards the MoE structure and introduces asymmetric knowledge transfer from auxiliary towers to the main task tower for enhancing main task-prioritized knowledge utilization [7]. As a result, although NATAL gives the best main task performance among baselines, it underperforms PLE and HoME in terms of auxiliary task performance. Therefore, we contend that a successful MAL model should meet two principles:

Structure Design Principles

1. Fully learn multi-attribution knowledge underlying multiple attribution mechanisms, as done in PLE and HoME;
2. Fully leverage this knowledge in a main-task-prioritized manner, as done in NATAL.

5.2 Our MoAE Model

To satisfy the two aforementioned principles, we devise **M**ixture **o**f **A**symmetric **E**xperts (**MoAE**), a new multi-attribution learning model integrating the strengths of the baselines analyzed in § 5.1. Concretely, as illustrated in Figure 4 [3], **MoAE** consists of three components: **(1)** An MoE backbone that learns rich attribution-specific representations. Inspired by PLE [34], we introduce a shared expert for learning public conversion patterns and attribution-specific experts for learning unique patterns underlying different attribution mechanisms. **(2)** A main–task–centric feature transfer module that aggregates auxiliary knowledge into the main task predictor in an asymmetric way. **(3)** MLP predictors that produce predictions for each attribution mechanism.

As shown in Table 3, our MoAE model not only reaches the best primary task performance, but also reaches superior performance in terms of the GAUC metrics under the auxiliary attribution mechanisms, which suggests that **MoAE** more fully fits the objectives under each attribution mechanism, aligning with its design motivation. We further show the generality of **MoAE** in the following experiments.

Table 3: The model performance in terms of GAUC when last-click serves as the target attribution mechanism. MoE denotes the mixture-of-experts structure, and KT stands for main-task-prioritized knowledge transfer. The best results are indicated in bold and the second best results are underlined.

Model	MoE	KT	Primary	Auxiliary		
			Last	First	DDA	Linear
Base	✗	✗	0.7424	-	-	-
Shared-Bottom	✗	✗	0.7571	0.6844	0.7579	0.7619
MMoE	✓	✗	0.7571	0.6843	0.7586	0.7626
PLE	✓	✗	0.7583	**0.6900**	0.7620	0.7660
HoME	✓	✗	0.7594	0.6871	<u>0.7624</u>	<u>0.7666</u>
NATAL	✗	✓	<u>0.7613</u>	0.6842	0.7620	0.7650
MoAE (ours)	✓	✓	**0.7636**	<u>0.6877</u>	**0.7637**	**0.7668**

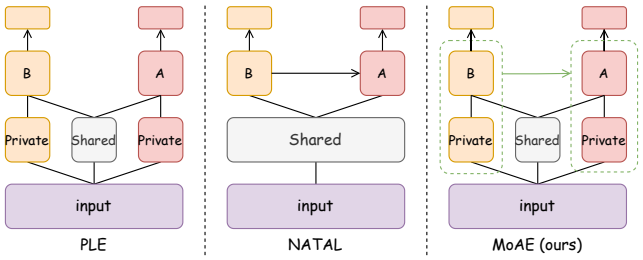


Figure 4: Information flow comparison between cutting-edge baselines (PLE and NATAL) and our MoAE model. A denotes the main target under the target attribution mechanism, and B is an auxiliary target from other attribution mechanisms.

Table 4: The CVR ranking performance under different target attribution mechanisms.

Primary Target	Last-Click		First-Click		DDA		Linear	
Method/Metrics	**AUC ↑**	**GAUC ↑**	**AUC ↑**	**GAUC ↑**	**AUC ↑**	**GAUC ↑**	**AUC ↑**	**GAUC ↑**
BASE	0.8494	0.7424	0.8311	0.6969	0.8435	0.7470	0.8503	0.7602
Shared-Bottom [3]	0.8628	0.7571	0.8303	0.6979	0.8514	0.7575	0.8515	0.7631
MMoE [23]	0.8642	0.7571	0.8314	0.6977	0.8507	0.7555	0.8512	0.7634
PLE [34]	0.8645	0.7583	0.8316	0.6986	0.8509	0.7574	0.8527	0.7638
HoME [36]	0.8661	0.7594	0.8316	0.6985	0.8524	0.7589	0.8542	0.7643
NATAL [7]	0.8671	0.7613	0.8319	0.6990	0.8546	0.7617	0.8548	0.7643
MoAE (ours)	**0.8694**	**0.7636**	**0.8325**	**0.7003**	**0.8576**	**0.7644**	**0.8568**	**0.7682**
Δ (vs BASE)	**+2.00pt**	**+2.12pt**	**+0.14pt**	**+0.34pt**	**+1.41pt**	**+1.74pt**	**+0.65pt**	**+0.80pt**
Δ (vs NATAL)	**+0.23pt**	**+0.23pt**	**+0.06pt**	**+0.13pt**	**+0.30pt**	**+0.27pt**	**+0.20pt**	**+0.39pt**

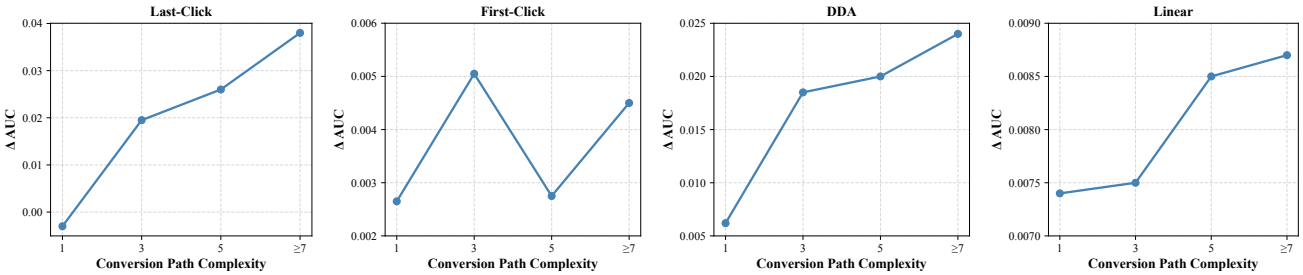


Figure 5: The AUC growth of users grouped by conversion path complexity.

6 Experimental Results and Analysis

6.1 Experimental Setup

6.1.1 Training Hyperparameters. We train all models using the Adam optimizer [12] for one epoch [45] on the **MAC**, where the last-day data serve as the test set and the preceding data are used for training. The batch size is 4096 for all models, as exploratory experiments showed that this value works well across different settings. Regarding the learning rate, we search for the best value over the set $\text{lr} \in \{0.1x,\ x,\ 10x \mid x \in \{0.0030, 0.0035, 0.0040\}\}$ based on the validation performance[4].

6.1.2 Objective Formulation. The CVR prediction tasks under all attribution mechanisms are formulated as binary classification, where the loss is defined as:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\Big[y_i \log p_i + (1-y_i)\log(1-p_i)\Big], \tag{3}$$

where N is the batch size, $y_i \in \{0,1\}$ is the ground-truth label of the i-th instance, and $p_i \in (0,1)$ is the predicted conversion probability. The overall training objective combines the primary task loss under the target attribution mechanism with auxiliary losses from other attribution mechanisms:

$$\mathcal{L} = \mathcal{L}_{\text{primary}} + \lambda \sum_{k} \mathcal{L}_{\text{aux}}^{(k)}, \tag{4}$$

where $\mathcal{L}_{\text{primary}}$ denotes the primary task loss, and $\mathcal{L}_{\text{aux}}^{(k)}$ denotes the loss for the k-th auxiliary task. Each loss term is instantiated as the binary cross-entropy loss in Eq. (3). To find the best λ value for each model, we search over the set $\{0.1, 0.2, 0.3, 0.4\}$ based on validation performance. We also test an automatic method GCS [9] dynamically setting different weights for each task, but we find it does not make significant contributions, which will be shown in § 7.2.

6.1.3 The Choice of Auxiliary Tasks. For each target attribution setting, we further search over the combination of auxiliary attribution objectives and report the results under the best setting. We give the details of our greedy forward-selection search in Appendix A.

6.2 Main Results and Key Findings

We present the main results in Table 4 and provide an in-depth analysis by answering the following research questions (RQs):

- **RQ1** (§6.2.1): Does MAL bring substantial performance gains across different target attribution mechanisms?
- **RQ2** (§6.2.2): Which user group gains the most from MAL?
- **RQ3** (§6.2.3): Does adding more auxiliary targets from different attribution mechanisms always yield performance growth?
- **RQ4** (§6.2.4): Does our MoAE model, which adheres to the key design principles, achieve superior performance?

[3]We only show one auxiliary attribution mechanism here for brevity. When there are multiple auxiliary attribution mechanisms, **MoAE** allocates one private expert and an MLP predictor for each auxiliary attribution mechanism.

[4]Supposing there are T days of training data, we test the day T-1 model on the data of day T to get validation performance.

6.2.1 The Generality of MAL. Considering that Chen et al. [7] only demonstrate the contribution of MAL under last-click and DDA attribution mechanisms, it remains unknown whether MAL brings stable performance gains across different target attribution mechanisms. It is an important question as other attribution mechanisms, e.g., linear attribution, are also widely adopted in recommendation systems. As shown in Table 4, the cutting-edge MAL models, e.g., NATAL [7] and our **MoAE**, achieve substantial improvements under all four target attribution settings: last-click, first-click, DDA, and linear. The performance lifts across diverse attribution mechanisms demonstrate that **MAL is a general, robust, and highly promising learning paradigm for conversion modeling**.

Furthermore, regarding the improvements under different target attribution mechanisms, we observe that **last-click and DDA settings see the most significant GAUC** increases at 2.12pt and 1.74pt, respectively. In contrast, in the **linear attribution** setting where MAL does not increase the number of positive samples, we observe **a smaller GAUC uplift** at 0.80pt, which suggests that the causal relationships in DDA labels and the positional information underlying last-click and first-click labels also make contributions.

Notably, in the **first-click** setting, **MAL yields the smallest GAUC improvement** at 0.34pt. We attribute this more modest improvement to the inherently noisy nature of first-click labels. Specifically, the extended time gap between the initial click and the final conversion introduces many confounding factors (e.g., subsequent promotions) unknown to the model. This weakens the label's correlation with the user's original intent, thereby constraining the potential improvement from MAL under the first-click setting. The fact that the first-click setting yields the lowest absolute AUC and GAUC scores, as shown in Table 4, underscores the noisy property of the first-click labels, a claim we will further substantiate with empirical evidence in Appendix B.

6.2.2 Impact Analysis on User Groups. To answer **RQ2**, we present the AUC lift brought by our **MoAE** over the single-attribution baseline across **user groups categorized by conversion path complexity** in Figure 5. Specifically, the conversion path complexity is proxied by the ratio of the number of positive samples under the linear-click setting to that under the last-click setting. A higher value of this ratio indicates that users click more ads on average before conversion, signifying a longer conversion path.

We observe that **multi-attribution learning delivers greater AUC improvements for users with more complex conversion paths**, a trend that holds across most of the evaluated target attribution mechanisms: last-click, linear, and DDA. This correlation suggests that **MAL is particularly beneficial for users with complex multi-touch conversion journeys**, where richer information gains underlie multi-attribution labels. In contrast, when first-click is the target attribution mechanism, the AUC improvements brought by MAL do not exhibit a clear monotonic trend with respect to conversion path complexity, which is likely caused by the inherently noisy nature of first-click labels.

6.2.3 The Effect of Auxiliary Objective Choice. We progressively add different auxiliary targets to **MoAE** under each target attribution mechanism and report the resulting model performance in Table 5, where we make two key observations as follows.

Table 5: The CVR ranking performance when labels from different attribution methods are leveraged as auxiliary targets. "None" denotes the MoAE variant trained without any auxiliary target

Auxiliary Targets	AUC	GAUC
Last-Click as Target		
None	0.8521	0.7429
First-Click	0.8621	0.7572
DDA	0.8543	0.7506
Linear	0.8642	0.7583
First-Click & Linear & DDA	0.8666	0.7612
First-Click & Linear & DDA & CAT	**0.8694**	**0.7636**
First-Click as Target		
None	0.8314	0.6971
Last-Click	**0.8325**	**0.7003**
DDA	0.8248	0.6932
Linear	0.8182	0.6886
Last-Click & DDA & Linear	0.8171	0.6898
Last-Click & DDA & Linear & CAT	0.8182	0.6878
DDA as Target		
None	0.8447	0.7481
Last-Click	0.8467	0.7494
First-Click	0.8502	0.7530
Linear	0.8520	0.7570
Last-Click & First-Click & Linear	0.8521	0.7592
Last-Click & First-Click & Linear & CAT	**0.8576**	**0.7644**
Linear as Target		
None	0.8524	0.7611
Last-Click	0.8545	0.7649
First-Click	0.8534	0.7618
DDA	0.8535	0.7632
Last-Click & First-Click & DDA	0.8552	0.7661
Last-Click & First-Click & DDA & CAT	**0.8568**	**0.7682**

First, the model's performance lift scales with the number of auxiliary targets under most of target attribution mechanisms, and the Cartesian product auxiliary target (CAT) [7] yields a further significant improvement, demonstrating *the potential of incorporating more sophisticated auxiliary targets.*

Second, in contrast, the first-click CVR prediction model is sensitive to the choice of auxiliary objectives. Concretely, only the last-click auxiliary target yields significant performance gains, while other choices of auxiliary targets undermine model performance. We speculate that this phenomenon is due to the unique position of the last-click target. Because it is temporally closest to the conversion, it offers the greatest discrepancy from the first-click signal, which in turn helps the model learn the user's intrinsic intent more effectively.

*6.2.4 The Superiority of Our **MoAE** Model.* As shown in Table 4, **MoAE consistently outperforms all baselines under all four target attribution mechanisms.** Compared with the existing

Table 6: The CVR ranking performance with/without auxiliary labels from multiple attribution mechanisms (mal). Last-click is the target attribution mechanism.

Method/Metrics	AUC ↑	GAUC ↑
Base	0.8494	0.7424
Shared-Bottom	0.8628 (+1.34pt)	0.7571 (+1.47pt)
- w/o mal	0.8497 (+0.03pt)	0.7420 (-0.04pt)
MMoE	0.8642 (+1.48pt)	0.7571 (+1.47pt)
- w/o mal	0.8501 (+0.07pt)	0.7428 (+0.04pt)
PLE	0.8645 (+1.51pt)	0.7583 (+1.59pt)
- w/o mal	0.8494 (0.00pt)	0.7424 (0.00pt)
HoME	0.8661 (+1.67pt)	0.7594 (+1.70pt)
- w/o mal	0.8523 (+0.29pt)	0.7421 (-0.03pt)
NATAL	0.8671 (+1.77pt)	0.7613 (+1.89pt)
- w/o mal	0.8527 (+0.33pt)	0.7437 (+0.13pt)
MoAE (Ours)	0.8694 (+2.00pt)	0.7636 (+2.12pt)
- w/o mal	0.8521 (+0.27pt)	0.7429 (+0.05pt)

state-of-the-art model NATAL [7], **MoAE** achieves GAUC gains of +0.23pt, +0.13pt, +0.27pt, and +0.39pt under last-click, first-click, DDA, and linear settings, respectively. We further validate the GAUC improvement with five random seeds under the last-click setting, as reported in Appendix C. The superiority of **MoAE** substantiates the power of combining the two architecture design principles in § 6.2.4, namely **(1) fully learning multi-attribution knowledge** underlying multi-attribution targets and **(2) fully leveraging multi-attribution knowledge** to maximize the main task performance. We are confident that these findings will serve as valuable guidelines for future MAL architectural designs.

7 Further Discussion

In this section, we discuss two research questions that are not covered by the major experiments:

- **RQ5**: Does the performance growth brought by MAL come from the increase of model parameters or the information gain underlying multi-attribution labels?
- **RQ6**: How do auxiliary-task learning (ATL) techniques affect the performance of MAL models?

7.1 Effect of Parameter Growth

Given that MAL models such as NATAL [7] and our **MoAE** introduce more parameters for fitting multiple attribution labels compared to the single-attribution baseline, the performance growth of MAL models may be attributed to **parameter scaling** rather than **multi-attribution knowledge utilization**.

To verify the source of performance gains brought by MAL, we disentangle the effect of multi-attribution labels (**mal** for short) from parameter scaling by setting the auxiliary tasks weights to zeros, and list the resulting model metrics in Table 6.

We observe that **vanilla parameter scaling yields no or marginal GAUC improvements**, and **the performance growth brought by MAL mainly results from multi-attribution supervision signals.** Specifically, when the auxiliary task weights are set to zero, namely the "w/o mal" variants in Table 6, although the parameter scales of resulting models are larger than the single-attribution Base and equal to corresponding MAL models, they significantly underperform the MAL counterparts learning from auxiliary labels, and perform almost on par with the Base model. The trend is consistent across the multi-task family (Shared-Bottom, MMoE, PLE, and HoME), NATAL, and our **MoAE** model, suggesting that harnessing multi-attribution knowledge underlying diverse attribution mechanisms contributes to the performance advantage of MAL models over the single-attribution baseline.

7.2 Effect of Auxiliary-Task Learning Methods

Auxiliary-task learning (ATL) [9, 20, 42] aims to maximize the target task performance by leveraging the knowledge underlying auxiliary tasks, which aligns with our MAL problem setup. ATL is a developed research area with various approaches, such as gradient manipulation [37, 42], auxiliary loss weight adjustment [9, 18, 32], and auxiliary task grouping [10, 33], and we believe that incorporating ATL techniques is promising for future work on multi-attribution learning. In this study, we take the first step to investigate the effect of ATL techniques on multi-attribution learning models. Concretely, we implement the classic GCS [9] and PCGrad [42] approaches as **optional training features in our PyMAL** library, and examine their effects on MAL models. GCS adjusts auxiliary loss weights using gradient similarities between target and auxiliary targets, and PCGrad eliminates gradient conflicts through gradient projection.

We present the results in Table 7 and make two key observations. **First**, GCS and PCGrad yield small performance gains when attached to weak baselines such as Shared-Bottom and MMoE, showing their effectiveness in mitigating negative transfer, namely the negative effects of auxiliary tasks on the target task. **Second**, when applied to strong MAL models such as NATAL and our **MoAE** tailored for maximizing the target task performance, GCS and PCGrad show marginal negative effects. Overall, our **MoAE** without GCS or PCGrad achieves the best performance among all model variants. These results indicate that **well-designed architectures are more effective than ATL techniques based on gradient projection and loss weight adjustment in multi-attribution learning for CVR prediction**. We believe that it is easy to incorporate more advanced ATL approaches to our **PyMAL** library, which are promising for future study.

8 Conclusion

In this work, we introduce **MAC**, the first public CVR prediction benchmark providing **conversion labels derived from multiple attribution mechanisms**. This dataset enables systematic investigation into **multi-attribution learning (MAL)**, a promising paradigm previously hindered by the unavailability of multi-attribution conversion labels. Through comprehensive experiments on **MAC** using our open-source **PyMAL** library, we gather three key insights. **First**, MAL consistently enhances model performance across diverse target attribution settings, with the most significant gains observed for users exhibiting long and complex conversion

Table 7: The effect of ATL techniques GCS and PCGrad.

Method/Metrics	AUC ↑	GAUC ↑
Shared-Bottom	0.8628	0.7571
+ GCS	0.8632 (+0.04pt)	0.7589 (+0.18pt)
+ PCGrad	0.8636 (+0.08pt)	0.7591 (+0.20pt)
MMoE	0.8642	0.7571
+ GCS	0.8636 (-0.06pt)	0.7590 (+0.19pt)
+ PCGrad	0.8640 (-0.02pt)	0.7573 (+0.02pt)
PLE	0.8645	0.7583
+ GCS	0.8638 (-0.07pt)	0.7572 (-0.11pt)
+ PCGrad	0.8636 (-0.09pt)	0.7568 (-0.15pt)
HoME	0.8661	0.7594
+ GCS	0.8626 (-0.35pt)	0.7569 (-0.25pt)
+ PCGrad	0.8653 (-0.08pt)	0.7609 (+0.15pt)
NATAL	0.8671	0.7613
+ GCS	0.8628 (-0.43pt)	0.7571 (-0.42pt)
+ PCGrad	0.8663 (-0.08pt)	0.7618 (+0.05pt)
MoAE (Ours)	0.8694	0.7636
+ GCS	0.8673 (-0.21pt)	0.7604 (-0.32pt)
+ PCGrad	0.8680 (-0.14pt)	0.7625 (-0.11pt)

paths. **Second**, the choice of auxiliary objectives in MAL is crucial and task-dependent. **Third**, effective MAL architectures must satisfy two core design principles: (i) fully learn the knowledge embedded in multi-attribution labels, and (ii) effectively leverage this knowledge in a main-task-prioritized manner. Motivated by these insights, we propose **MoAE**, an effective MAL model that integrates an MoE backbone for knowledge acquisition with an asymmetric transfer module for main-task-centric utilization, whose superiority is demonstrated by extensive empirical evaluation.

We believe the **MAC** benchmark, the **PyMAL** toolkit, and the gathered actionable insights will serve as valuable resources for the research community, accelerating future innovation in multi-attribution learning and more holistic conversion modeling. For future work, we believe that incorporating stronger foundation model capabilities, such as chain-of-thought reasoning [38], and more comprehensive conversion objectives [43] holds great promise.

Acknowledgments

This work is funded by Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM902), National Natural Science Foundation of China (Grant No. 62506158 and No. 62441234), Basic Research Program of Jiangsu (BK20251183), and CCF-ALIMAMA TECH Kangaroo Fund (No. CCF-ALIMAMA of 2025004).

References

[1] John Bencina, Erkut Aykutlug, Yue Chen, Zerui Zhang, Stephanie Sorenson, Shao Tang, and Changshuai Wei. 2025. LiDDA: Data Driven Attribution at LinkedIn. arXiv:2505.09861 [cs.LG] https://arxiv.org/abs/2505.09861

[2] Weijie Bian, Kailun Wu, Lejian Ren, Qi Pi, Yujing Zhang, Can Xiao, Xiang-Rong Sheng, Yong-Nan Zhu, Zhangming Chan, Na Mou, Xinchen Luo, Shiming Xiang, Guorui Zhou, Xiaoqiang Zhu, and Hongbo Deng. 2022. CAN: Feature Co-Action Network for Click-Through Rate Prediction. In *Proceedings of the 15th ACM International Conference on Web Search and Data Mining*. 57–65.

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

[4] Zhangming Chan, Yuchi Zhang, Xiuying Chen, Shen Gao, Zhiqiang Zhang, Dongyan Zhao, and Rui Yan. 2020. Selection and generation: Learning towards multi-product advertisement post generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 3818–3829.

[5] Zhangming Chan, Yu Zhang, Shuguang Han, Yong Bai, Xiang-Rong Sheng, Siyuan Lou, Jiacen Hu, Baolin Liu, Yuning Jiang, Jian Xu, et al. 2023. Capturing conversion rate fluctuation during sales promotions: A novel historical data reuse approach. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3774–3784.

[6] Olivier Chapelle. 2014. Modeling delayed feedback in display advertising. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1097–1105.

[7] Sishuo Chen, Zhangming Chan, Xiang-Rong Sheng, Lei Zhang, Sheng Chen, Chenghuan Hou, Han Zhu, Jian Xu, and Bo Zheng. 2025. See Beyond a Single View: Multi-Attribution Learning Leads to Better Conversion Rate Prediction. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management* (Seoul, Republic of Korea) *(CIKM '25)*. Association for Computing Machinery, New York, NY, USA, 5600–5608.

[8] Yu Chen, Jiaqi Jin, Hui Zhao, Pengjie Wang, Guojun Liu, Jian Xu, and Bo Zheng. 2022. Asymptotically unbiased estimation for delayed feedback modeling via label correction. In *Proceedings of the ACM Web Conference 2022*. 369–379.

[9] Yunshu Du, Wojciech M. Czarnecki, Siddhant M. Jayakumar, Mehrdad Farajtabar, Razvan Pascanu, and Balaji Lakshminarayanan. 2020. Adapting Auxiliary Losses Using Gradient Similarity. arXiv:1812.02224 [stat.ML] https://arxiv.org/abs/1812.02224

[10] Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. 2021. Efficiently identifying task groupings for multi-task learning. *Advances in Neural Information Processing Systems* 34 (2021), 27503–27516.

[11] Siyu Gu, Xiang-Rong Sheng, Ying Fan, Guorui Zhou, and Xiaoqiang Zhu. 2021. Real negatives matter: Continuous training with real negatives for delayed feedback modeling. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2890–2898.

[12] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG] https://arxiv.org/abs/1412.6980

[13] Sachin Kumar, Garima Gupta, Ranjitha Prasad, Arnab Chatterjee, Lovekesh Vig, and Gautam Shroff. 2020. Camta: Causal attention model for multi-touch attribution. In *ICDMW 2020*. IEEE, 79–86.

[14] Randall Lewis, Florian Zettelmeyer, Brett R. Gordon, Cristobal Garib, Johannes Hermle, Mike Perry, Henrique Romero, and German Schnaidt. 2025. Amazon Ads Multi-Touch Attribution. arXiv:2508.08209 [econ.EM] https://arxiv.org/abs/2508.08209

[15] Danwei Li, Zhengyu Zhang, Siyang Yuan, Mingze Gao, Weilin Zhang, Chaofei Yang, Xi Liu, and Jiyan Yang. 2023. 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.

[16] Ning Li, Sai Kumar Arava, Chen Dong, Zhenyu Yan, and Abhishek Pani. 2018. Deep Neural Net with Attention for Multi-Channel Multi-Touch Attribution. arXiv:1809.02230 [cs.LG] https://arxiv.org/abs/1809.02230

[17] Xinyu Li, Sishuo Chen, Guipeng Xv, Li Zhang, Mingxuan Luo, Zhangming Chan, Xiang-Rong Sheng, Han Zhu, Jian Xu, and Chen Lin. 2026. Delayed Feedback Modeling for Post-Click Gross Merchandise Volume Prediction: Benchmark, Insights and Approaches. In *Proceedings of the ACM Web Conference 2026* (United Arab Emirates) *(WWW '26)*. Association for Computing Machinery, New York, NY, USA, 6299–6308.

[18] Xingyu Lin, Harjatin Singh Baweja, George Kantor, and David Held. 2019. *Adaptive auxiliary task weighting for reinforcement learning*. Curran Associates Inc., Red Hook, NY, USA.

[19] Qiming Liu, Xiang Ao, Yuyao Guo, and Qing He. 2024. Online conversion rate prediction via multi-interval screening and synthesizing under delayed feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 8796–8804.

[20] Shikun Liu, Stephen James, Andrew Davison, and Edward Johns. 2022. Auto-Lambda: Disentangling Dynamic Task Relationships. *Transactions on Machine Learning Research* (2022).

[21] Quan Lu, Shengjun Pan, Liang Wang, Junwei Pan, Fengdan Wan, and Hongxia Yang. 2017. A Practical Framework of Conversion Rate Prediction for Online Display Advertising. In *Proceedings of the ADKDD'17* (Halifax, NS, Canada) *(ADKDD'17)*. Association for Computing Machinery, New York, NY, USA, Article 9, 9 pages.

[22] Mingxuan Luo, Guipeng Xv, Sishuo Chen, Xinyu Li, Li Zhang, Zhangming Chan, Xiang-Rong Sheng, Han Zhu, Jian Xu, Bo Zheng, and Chen Lin. 2026. Modeling Cascaded Delay Feedback for Online Net Conversion Rate Prediction: Benchmark, Insights and Solutions. In *Proceedings of the ACM Web Conference 2026* (United Arab Emirates) *(WWW '26)*. Association for Computing Machinery, New York, NY, USA, 6252–6263.

[23] 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.

[24] Xiao Ma, Liqin Zhao, Guan Huang, Zhi Wang, Zelin Hu, Xiaoqiang Zhu, and Kun Gai. 2018. Entire space multi-task model: An effective approach for estimating post-click conversion rate. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1137–1140.

[25] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. 2016. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3994–4003.

[26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Naresh Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, Vol. 32. 8024–8035.

[27] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*. Vol. 39. Cambridge University Press Cambridge.

[28] Xuhui Shao and Lexin Li. 2011. Data-driven multi-touch attribution models. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 258–264.

[29] Xiang-Rong Sheng, Jingyue Gao, Yueyao Cheng, Siran Yang, Shuguang Han, Hongbo Deng, Yuning Jiang, Jian Xu, and Bo Zheng. 2023. Joint Optimization of Ranking and Calibration with Contextualized Hybrid Model. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4813–4822.

[30] Xiang-Rong Sheng, Feifan Yang, Litong Gong, Biao Wang, Zhangming Chan, Yujing Zhang, Yueyao Cheng, Yong-Nan Zhu, Tiezheng Ge, Han Zhu, et al. 2024. Enhancing Taobao Display Advertising with Multimodal Representations: Challenges, Approaches and Insights. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 4858–4865.

[31] 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.

[32] Baifeng Shi, Judy Hoffman, Kate Saenko, Trevor Darrell, and Huijuan Xu. 2020. Auxiliary task reweighting for minimum-data learning. *Advances in Neural Information Processing Systems* 33 (2020), 7148–7160.

[33] Xiaozhuang Song, Shun Zheng, Wei Cao, James Yu, and Jiang Bian. 2022. Efficient and effective multi-task grouping via meta learning on task combinations. *Advances in Neural Information Processing Systems* 35 (2022), 37647–37659.

[34] 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.

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

[36] Xu Wang, Jiangxia Cao, Zhiyi Fu, Kun Gai, and Guorui Zhou. 2025. HoME: Hierarchy of Multi-Gate Experts for Multi-Task Learning at Kuaishou. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1* (Toronto ON, Canada) *(KDD '25)*. Association for Computing Machinery, New York, NY, USA, 2638–2647.

[37] Zirui Wang, Yulia Tsvetkov, Orhan Firat, and Yuan Cao. 2021. Gradient Vaccine: Investigating and Improving Multi-task Optimization in Massively Multilingual Models. In *International Conference on Learning Representations*.

[38] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[39] Jia-Qi Yang, Xiang Li, Shuguang Han, Tao Zhuang, De-Chuan Zhan, Xiaoyi Zeng, and Bin Tong. 2021. Capturing delayed feedback in conversion rate prediction via elapsed-time sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 4582–4589.

[40] Di Yao, Chang Gong, Lei Zhang, Sheng Chen, and Jingping Bi. 2022. CausalMTA: Eliminating the user confounding bias for causal multi-touch attribution. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4342–4352.

[41] Meng-Hsuan Yu, Juntao Li, Zhangming Chan, Rui Yan, and Dongyan Zhao. 2021. Content learning with structure-aware writing: A graph-infused dual conditional variational autoencoder for automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 6021–6029.

[42] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient Surgery for Multi-Task Learning. arXiv:2001.06782 [cs.LG] https://arxiv.org/abs/2001.06782

[43] Xiangyu Zeng, Amit Jaspal, Bin Liu, Goutham Panneeru, Kevin Huang, Nicolas Bievre, Mohit Jaggi, Prathap Maniraju, and Ankur Jain. 2025. Click A, Buy B: Rethinking Conversion Attribution in E- Commerce Recommendations. arXiv:2507.15113 [cs.IR] https://arxiv.org/abs/2507.15113

[44] Yu Zhang and Qiang Yang. 2021. A survey on multi-task learning. *IEEE transactions on knowledge and data engineering* 34, 12 (2021), 5586–5609.

[45] Zhaorui Zhang, Xiang-Rong Sheng, Yujing Zhang, Biye Jiang, Shuguang Han, Hongbo Deng, and Bo Zheng. 2022. Towards Understanding the Overfitting Phenomenon of Deep Click-Through Rate Models. *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (2022), 2671–2680.

[46] Yunfeng Zhao, Xu Yan, Xiaoqiang Gui, Shuguang Han, Xiang-Rong Sheng, Guoxian Yu, Jufeng Chen, Zhao Xu, and Bo Zheng. 2023. Entire space cascade delayed feedback modeling for effective conversion rate prediction. In *CIKM*. 4981–4987.

[47] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1059–1068.

[48] Han Zhu, Xiang Li, Pengye Zhang, Guozheng Li, Jie He, Han Li, and Kun Gai. 2018. Learning tree-based deep model for recommender systems. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1079–1088.

[49] Jie Zhu, Zhifang Fan, Xiaoxie Zhu, Yuchen Jiang, Hangyu Wang, Xintian Han, Haoran Ding, Xinmin Wang, Wenlin Zhao, Zhen Gong, et al. 2025. Rankmixer: Scaling up ranking models in industrial recommenders. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*. 6309–6316.

[50] Jinfeng Zhuang, Yinrui Li, Runze Su, Ke Xu, Zhixuan Shao, Kungang Li, Ling Leng, Han Sun, Meng Qi, Yixiong Meng, Yang Tang, Zhifang Liu, Qifei Shen, and Aayush Mudgal. 2025. On the Practice of Deep Hierarchical Ensemble Network for Ad Conversion Rate Prediction. In *Companion Proceedings of the ACM on Web Conference 2025* (Sydney NSW, Australia) *(WWW '25)*. Association for Computing Machinery, New York, NY, USA, 671–680.

A Details of Auxiliary Objective Search

For each model under each attribution setting, we start from the primary objective alone and greedily add auxiliary objectives in a forward-selection manner: (1) at each step, we temporarily add one candidate auxiliary objective, train the model, and measure the GAUC improvement on the validation day; (2) we then select the auxiliary objective that brings the largest positive gain and permanently include it into the auxiliary objective set; (3) this process is repeated on the remaining candidates until either all auxiliary objectives have been used or no further GAUC improvement is observed. This greedy search is conducted independently under each target attribution mechanism.

In our experiments, this greedy search yields consistent patterns across settings. For all models under the last-click, linear, and DDA attribution mechanisms, the optimal configuration includes all candidate auxiliary objectives. Under the first-click attribution mechanism, however, the best configuration for all models uses only the last-click objective as an auxiliary signal.

B Analysis of First-Click Labels

We provide additional evidence of the noisier nature of first-click labels by examining their inconsistency with data-driven attribution (DDA) and linear attribution, summarized in Table 8. The table reports the mean difference between the DDA weight and the linear weight for positive labels (MML). Under the last-click setting, MML is positive (+0.05), indicating that last-click positive samples tend to receive more weight under DDA than under the simpler linear rule—i.e., the DDA view broadly agrees that these touchpoints are important. In contrast, under the first-click setting, MML becomes negative (−0.03), meaning that clicks labeled as positive by first-click are, on average, assigned less weight by DDA than by linear

attribution. Since DDA is learned from causal signals along the entire path, this finding suggests that many first-click "positives" are not strongly supported as influential touchpoints and are even down-weighted relative to a naive linear split. This comparison indicates that first-click labels are noisier and harder to predict.

Table 8: The MML metric comparison between the first-click and last-click positive samples.

	Last-Click	First-Click
MML	0.05	-0.03

C Multi-Seed Validation

To further examine the stability of the performance improvement, we repeat the representative last-click experiment with five random seeds and compare **MoAE** with the strongest existing MAL baseline **NATAL**. As shown in Table 9, **MoAE** consistently outperforms **NATAL** across runs, achieving an average GAUC gain of +0.18pt. A significance test on GAUC gives $p \approx 0.006$, suggesting that the improvement is statistically reliable under this setting.

Table 9: Results on the last-click setting over five random seeds.

Model	AUC	GAUC
NATAL	0.8673 ± 0.0002	0.7612 ± 0.0008
MoAE	0.8691 ± 0.0005	0.7630 ± 0.0002
Δ	+0.18pt	+0.18pt

D Details of DDA Label Generation

The data-driven attribution (DDA) labels in **MAC** are generated by CausalMTA [40], a causal multi-touch attribution model that estimates conversion credits from user conversion paths. Unlike rule-based mechanisms such as last-click, first-click, and linear attribution, CausalMTA estimates the counterfactual contribution of each touchpoint to the final conversion.

CausalMTA first learns a debiased CVR predictor for counterfactual conversion estimation, which is required by Shapley-value-based attribution. To mitigate the confounding effect of user preference on both ad exposure and conversion behavior, CausalMTA combines two debiasing modules: a journey reweighting module based on a variational recurrent autoencoder for static user attributes, and a causal recurrent predictor with a gradient reversal layer for dynamic behavior features. The learned debiased predictor is then used to estimate the Shapley-value attribution credit of each touchpoint, where the DDA weight reflects the marginal contribution of the corresponding click to the conversion probability.

In our benchmark construction, we apply CausalMTA to conversion paths within the fixed attribution window and use the resulting attribution credits as the DDA labels. The same CausalMTA-based attribution system has served large-scale online traffic in Alibaba's advertising platform for several years, supporting the reliability of the generated DDA labels.

E Details of CAT Label Construction

The Cartesian-style Auxiliary Target (CAT) [7] transforms the set of N binary attribution labels into a single multi-class label by enumerating all possible joint outcomes across attribution mechanisms.

Concretely, for each sample we start from the binary labels

$$\mathbf{A} = \big(A_0, A_1, \ldots, A_{N-1}\big), \quad A_i \in \{0, 1\},$$

where A_i is the conversion indicator under the i-th attribution mechanism (e.g., last-click, first-click, DDA, linear). The CAT label is then defined as the integer encoding of $\mathbf{A}$ in little-endian binary form:

$$\mathbf{O} = \sum_{i=0}^{N-1} A_i \cdot 2^i, \quad \mathbf{O} \in \{0, 1, \ldots, 2^N - 1\}. \tag{5}$$

Algorithm 1 gives the corresponding pseudo-code. In our experiments with four attribution mechanisms (last-click, first-click, DDA, linear), CAT becomes a 16-class classification problem. For instance, the combination $(1, 0, 1, 0)$ is mapped to class

$$\mathbf{O} = 1 \cdot 2^0 + 0 \cdot 2^1 + 1 \cdot 2^2 + 0 \cdot 2^3 = 5.$$

Algorithm 1 Computation of the **CAT** Label

Require: Array $\mathbf{A}$ of length N with elements in $\{0, 1\}$, i.e., the conversion indicators under N attribution mechanisms
Ensure: Integer CAT label $\mathbf{O} \in [0, 2^N - 1]$
1: Initialize $\mathbf{O} \leftarrow 0$
2: **for** $i = 0$ to $N - 1$ **do**
3: $\quad \mathbf{O} \leftarrow \mathbf{O} + \mathbf{A}[i] \cdot 2^i$
4: **end for**

During training, we treat $\mathbf{O}$ as the target of a (2^N)-way classification task, optimized with a standard softmax cross-entropy loss. The corresponding conversion knowledge vector $\mathcal{K}_{\text{CAT}}$ is extracted from the CAT prediction tower and fused into the overall conversion representation $\mathcal{K}$ learned by AKA. At inference time, we do not use the CAT logits directly; only the enriched representation $\mathcal{K}$ is consumed by the primary target predictor.