INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN BUSINESS & SOCIAL SCIENCES

Vol 16, Issue 4, (2026) E-ISSN: 2222-6990

SwaDeepFM: Reliable CTR Prediction for Marketing Decision-Making via Stage-wise Attention Selection and Post-fusion Redundancy Removal

Qi Bao, Nadia Farhana

[1]PhD Researcher, Binary University of Management & Entrepreneurship, Malaysia,
[2]Associate Professor, Stamford University Bangladesh

DOI Link: http://dx.doi.org/10.6007/IJARBSS/v16-i4/28054

Published Date: 28 April 2026

Abstract

Click-through rate (CTR) prediction is central to marketing decision-making, where predicted probabilities are directly used for bidding, filtering, and budget allocation. Practical systems therefore require not only strong ranking performance, but also reliable probabilities that remain stable under noisy fields, shifting contexts, and redundant fused representations. However, most existing CTR models primarily optimize ranking accuracy, while paying less attention to probability reliability and stability, which limits their usefulness in real-world marketing decision-making. In this work, we propose SwaDeepFM, a stage-wise attention CTR model based on a DeepFM backbone. SwaDeepFM introduces three lightweight modules at different stages of the prediction pipeline. First, SE performs field-wise reweighting at the embedding stage to suppress noisy fields before interactions are formed. Second, a Context Transformer (CoT) aggregates context-conditioned dependencies among field tokens to stabilize high-order interaction selection. Third, CBAM refines the fused representation before prediction to remove redundancy and improve probabilistic robustness. Experiments on three public CTR benchmarks, Criteo, Avazu, and KDD12, against several representative baselines show consistent improvements in AUC and LogLoss. Fixed ablations further confirm that each module contributes complementary gains. Additional analyses based on field-weight visualization, attention inspection, long-tail bucketing, and calibration and stability protocols explain when the method helps most and why the resulting probabilities are more reliable for marketing decisions.

Keywords: SwaDeepFM, CTR Prediction, Marketing Decision-Making, Multi-Field Sparse Features, Stage-Wise Attention, Feature Interaction Learning, Probability Calibration

Introduction

Click-through rate (CTR) prediction is central to marketing decision-making, where predicted probabilities are directly used for bidding, traffic filtering, and budget allocation (Guan et al., 2025). In this setting, models must deliver not only better ranking performance but also

reliable and stable probabilities, because systematic probability errors can be amplified by downstream policies and lead to suboptimal efficiency and cost control (Dai et al., 2025).

Real-world advertising and recommendation data are typically represented by multi-field sparse categorical features with high-dimensional combinations and long-tail distributions (Qu et al., 2018). Even when interaction modeling becomes more expressive, three sources of instability often remain: noisy fields can enter representations before interactions are formed and then propagate interference; interaction patterns may shift with contexts, making high-order combination learning brittle (Zhang et al., 2023); and fused representations across branches can contain redundant co-occurrences and spurious correlations, which increases probability fluctuations across datasets and long-tail samples (Wu et al., 2023).

Recent CTR research has advanced through explicit interactions, attention mechanisms, and deep architectures, yet the above difficulties are often addressed by localized modifications rather than a systematic control pipeline from input to fusion (Zhang et al., 2023). For example, DeepFM provides a strong and efficient baseline by combining linear effects, second-order interactions, and deep networks, but its default field usage, interaction selection, and post-fusion refinement remain largely implicit, which makes it difficult to ensure reliable probabilities under noise and long-tail conditions (Wu et al., 2023).

To address this gap, we propose SwaDeepFM, a stage-wise attention model built on a DeepFM backbone, which introduces three lightweight modules at key stages of the prediction pipeline to enable controllable selection (Guo et al., 2018). First, SE performs field-wise reweighting at the embedding stage to suppress noisy fields before interactions are formed, satisfying the requirement of field denoising and importance reweighting. Second, CoT aggregates context-conditioned dependencies among field tokens to stabilize high-order interaction selection under context shifts and rare combinations, satisfying the requirement of context-conditioned interaction selection. Third, CBAM refines the fused representation before prediction to reduce redundancy and improve probabilistic robustness, satisfying the requirement of post-fusion redundancy removal and reliable probabilities(Woo et al., 2018). The motivation of this study is that, in real-world marketing systems, predicted CTR values are not only used for ranking but are also directly involved in bidding, filtering, and budget allocation. Therefore, even small probability deviations may accumulate and lead to inefficient traffic delivery or suboptimal spending decisions. This makes probability reliability an important but insufficiently emphasized objective in CTR prediction research.

The main contributions of this study are threefold. First, we propose SwaDeepFM, a stage-wise attention CTR framework built on DeepFM, which introduces controllable selection mechanisms from input embedding to final prediction. Second, we design three lightweight modules, namely SE, CoT, and CBAM, to address three practical sources of instability: noisy fields, context-sensitive interaction shifts, and redundant fused representations. Third, extensive experiments on three public CTR benchmarks demonstrate that SwaDeepFM consistently improves not only ranking performance but also calibration and stability, making it more suitable for marketing decision-making scenarios.

Method

Problem Formulation and Notation

In practical CTR prediction, each impression is represented by multiple feature fields, such as user attributes, ad identifiers, and delivery-context signals, where each field usually activates only one or a very small number of values for a given sample.Therefore, the input is high-dimensional at the global feature space level but highly sparse at the individual impression level.

Given multi-field sparse features $x = \{x^i\}_{i=1}^{F}$ for an impression, the goal is to predict the click probability $\hat{y} = p(y = 1|x)$, where $y \in \{0,1\}$. Beyond predictive accuracy, the design target of SwaDeepFM is to produce reliable and stable CTR estimates under noisy input fields, context-sensitive interaction shifts, and redundant fused representations. To this end, the model introduces stage-wise control over field denoising, interaction modeling, and post-fusion refinement before prediction.

Each field value x^i is mapped to an embedding:

$$e_i = Emb_i(x_i) \in R^d$$

Stacking embeddings yields:

$$E = [e_1; e_2; \ldots; e_F] \in R^{F \times D}$$

Flattening and concatenation gives:

$$v = concat(e_1, \ldots, e_F) \in R^{Fd}.$$

The model output is:

$$\hat{y} = \sigma(\text{logit}(x))$$

F, d, E, and v are reused in all subsequent derivations and match the data flow in Fig1.

SwaDeepFM: Stage-wise Attention DeepFM Framework

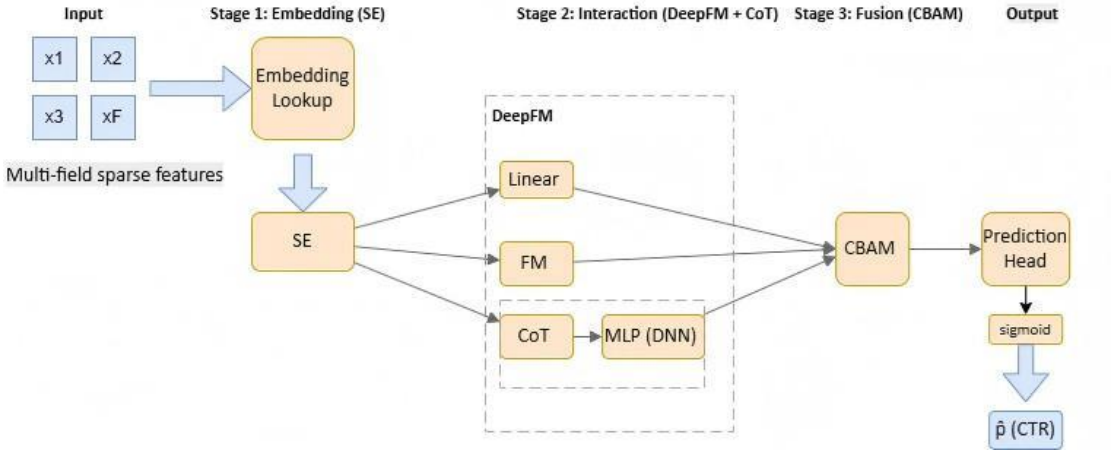

Figure 1 SwaDeepFM

Figure 1 illustrates the overall architecture of SwaDeepFM. We adopt DeepFM as the backbone, which models CTR signals through: (i) a linear term for main effects, (ii) an explicit second-order interaction (FM) branch, and (iii) an implicit high-order interaction (DNN) branch, and fuses them to form the final logit.

(i) Linear term (main effects)

$s_{lin} = w_0 + \sum_{i=1}^{F} w_i\,(x_i).$ (1)

(ii) Explicit second-order interactions (FM branch)

$s_{fm} = \sum_{1 \le i < j \le F} e_i^T e_j.$ (2)

(iii) Implicit high-order interactions (DNN branch)
Let $h_0 = v$, For $\ell = 1, \ldots, \mathrm{L}$,
$h_l = \varphi(W_l h_{l-1} + b_l)$ (3)
$s_{dnn} = w^T h_L + b$ (4)
(iv) Fusion and prediction
$logit(x) = s_{lin} + s_{fm} + s_{dnn}$ (5)
$\hat{y} = \sigma(\mathrm{logit}(x))$ (6)

SE: Embedding-stage Field Denoising
The SE module recalibrates field tokens at the embedding stage to suppress noisy fields and emphasize informative ones, ensuring that subsequent interaction learning is built on denoised inputs. Noisy fields can enter representations before interactions are formed and then propagate interference. The backbone tends to treat fields nearly equally, allowing weak fields to participate in interactions. SE learns an explicit field gate a to control field visibility. Input $E \in R^{(F\times d)}$, output $E^{SE} \in \mathrm{R}^{F\times \mathrm{d}}$. SE extracts a scalar descriptor per field, generates a field gate, and rescales embeddings.
(i) Squeeze
$z_i = \frac{1}{d}\sum_{k=1}^{d} e_i\,[k],\ z \in R^F$ (7)
(ii) Excitation
$a = \sigma(W_2\delta(W_1 z)),\quad a \in (0,1)^F$ (8)
(iii) Reweight
$e_i^{SE} = a_i e_i$ (9)
$E^{SE} = a \odot E$ (10)

CoT: Context-conditioned Dependency Aggregation for Stable High-order Interactions
Based on the SE-enhanced tokens, the backbone computes low-order signals through the linear and second-order interaction branches. For high-order modeling, we insert the CoT module before the multilayer perceptron to aggregate context-conditioned dependencies among field tokens, improving the stability of interaction selection under context shifts and rare combinations. Interaction patterns may shift with contexts, and an implicit MLP may not stably select key dependencies for each sample. The backbone encodes interaction selection into parameters, which becomes brittle for rare combinations. CoT uses self-attention to explicitly model context-conditioned dependencies among field tokens. We update token representations using an attention matrix A before feeding them into the DNN branch. CoT takes $E^{SE} \in \mathrm{R}^{F\times \mathrm{d}}$ and outputs $E^{CoT} \in R^{F\times d}$. We flatten E^{CoT} into $v^{CoT} \in \mathrm{R}^{Fd}$ as the DNN input.
(i) Q, K, V projections

$$Q = E^{SE}W_Q\ ,\quad K = E^{SE}W_K,\quad V = E^{SE}W_V$$

(ii) Attention and aggregation
$A = \mathrm{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)$ (11)
$\tilde{E} = AV$ (12)

(iii) Output with stabilization
$E^{CoT} = \ln(E^{SE} + \tilde{E}W_O)$ (13)

CBAM: Post-fusion Redundancy Removal before Prediction

We fuse low-order and high-order information and apply CBAM before prediction to refine the fused representation, reducing redundant co-occurrences and spurious correlations, which improves probabilistic robustness and usability for marketing decisions[7]. Fused representations may contain redundant co-occurrences and spurious correlations, which harms LogLoss and calibration stability. The backbone fusion mainly sums branch outputs and lacks explicit post-processing selection. CBAM applies channel attention and spatial attention to refine the fused representation. We adjust the emphasis of fused features using gates M_c and M_s. We construct a fused vector u from high-order features and low-order signals, reshape it into a one-dimensional feature map $U \in R^{C\times S}$, apply CBAM to obtain U^{CBAM}, and then flatten it for the prediction head.

(i) Channel attention

$M_c = \sigma(\mathrm{MLP}(\mathrm{AvgPool}_S(U)) + \mathrm{MLP}(\mathrm{MaxPool}_S(U)).$ (14)

$U_c = M_c \odot U.$ (15)

(ii) Spatial attention

$M_s = \sigma(Conv1D([AvgPool_C(U_c); MaxPool_C(U_c)]))$ (16)

$U^{\mathrm{CBAM}} = M_s \odot U_c.$ (17)

Training Objective and Complexity

We optimize binary cross-entropy with L2 regularization.

$\mathcal{L}_{bce} = -\frac{1}{N}\sum_{n=1}^{N}\left(y^{(n)} \log \hat{y}^{(n)} + (1-y^{(n)}) \log(1-\hat{y}^{(n)})\right)$ (18)

$\mathcal{L} = \mathcal{L}_{bce} + \lambda\|\theta\|_2^2$ (19)

The FM term can be computed in $\mathrm{O(Fd)}$. SE adds lightweight pooling and a small gating network, typically cheaper than the DNN branch. CoT adds an attention cost dominated by $O(F^2 d_k)$, which is manageable for typical CTR field counts. CBAM consists of lightweight pooling and gating before prediction.

Experiments

Datasets and Preprocessing

We evaluate SwaDeepFM on three public CTR benchmarks: Criteo Display Advertising Challenge (Tien et al., 2014), Avazu CTR Prediction (Wang & Cukierski, 2014), and KDD Cup 2012 Track 2 (Aden & Wang, 2012). Together, these datasets cover display advertising, mobile advertising, and sponsored search, enabling us to assess whether the model generalizes across traffic sources under multi-field sparsity and long-tail feature combinations.

To ensure fair and reproducible comparisons, we apply a standard multi-field preprocessing pipeline across all benchmarks. For missing values, we assign a default value or an explicit missing indicator for numerical features, while mapping missing categorical entries to a dedicated unknown token. For categorical encoding, we either construct field-wise vocabularies or use feature hashing, and we optionally map rare values to the unknown token to control vocabulary size and reduce noise. When numerical fields are present, we apply a $log\ (1+x)$ transformation or discretization to mitigate heavy-tailed distributions and align feature representations with embedding-based modeling. Finally, we use official dataset splits whenever available; otherwise, we create a validation set from the training data for early stopping and hyperparameter tuning, and reserve the test set strictly for final reporting.

Implementation Details

For fair comparisons, we use the same preprocessing, splits, and training protocol for all baselines, the DeepFM backbone, and SwaDeepFM. Unless otherwise stated, we set the embedding dimension to 16, and additionally test 8 and 32 in sensitivity analyses. The DeepFM backbone uses a three-layer MLP with hidden size 256, ReLU activations, and dropout 0.2. For SE, we use a reduction ratio of 4 with a sigmoid gate. For CoT, we use one self-attention block with four heads, residual connections, and layer normalization. For CBAM, we apply channel attention and spatial attention on a reshaped one-dimensional feature map with reduction ratio 4 and kernel size 7, and we use zero padding when reshaping does not align exactly.

We optimize with Adam using an initial learning rate of 1×10^{-3} and select the L2 regularization coefficient from $\{10^{-6}, 10^{-5}\}$. Batch size ranges from 2048 to 4096 depending on dataset scale. We train for up to 10 epochs and apply early stopping based on validation AUC, stopping when there is no improvement for two consecutive epochs and restoring the best checkpoint. Each setting is run with three random seeds, and we report mean performance, with standard deviations provided when stability is discussed.

We report AUC and LogLoss as the primary metrics. AUC measures discrimination under class imbalance and is widely used in CTR prediction. LogLoss evaluates probabilistic accuracy and aligns with the training objective.

Main Results

Table 1 reports the main results on Criteo, Avazu, and KDD12. Under the same evaluation protocol, SwaDeepFM consistently outperforms all representative baselines across the three benchmarks in terms of AUC and LogLoss, demonstrating strong effectiveness and transferability across display advertising, mobile advertising, and sponsored-search scenarios. The only exception is that on KDD12, SwaDeepFM achieves a LogLoss that is on par with DDT, while still attaining the best AUC. This indicates that SwaDeepFM improves discrimination without sacrificing probabilistic accuracy in sponsored-search settings.

These results support the benefit of a stage-wise control pipeline that integrates field-wise denoising, context-conditioned dependency aggregation, and post-fusion refinement. By explicitly controlling feature visibility, stabilizing high-order interaction selection under shifting contexts and long-tail combinations, and removing redundancy before prediction, SwaDeepFM yields more robust and reliable CTR estimates for downstream marketing decision-making.

Table 1
Main results on CTR benchmarks

Model	Criteo AUC	Criteo LogLoss	Avazu AUC	Avazu LogLoss	KDD12 AUC	KDD12 LogLoss
xDeepFM(Li et al., 2023)	0.807	0.4447	0.777	0.3823	0.782	0.156
DSAN(Wang et al., 2024)	0.8083	0.4434	0.7774	0.3811	0.7898	0.1543
DDT(Zhang et al., 2025)	0.8103	0.4423	0.7832	0.3786	0.792	0.152
DCN^2(Škrlj et al., 2025)	0.8085	0.4451	0.7894	0.3759	0.8012	0.1531
SwaDeepFM(Ours)	0.812	0.441	0.791	0.3745	0.803	0.152

Ablation Study

Table 2 presents the ablation results of SwaDeepFM on Criteo, Avazu, and KDD12. The full model consistently achieves the best overall performance across the three benchmarks. Removing any single module leads to clear degradation in AUC and/or LogLoss, confirming that the three components contribute complementary benefits rather than redundant effects. Among the modules, SE is responsible for embedding-stage field denoising and importance reweighting, which improves robustness to noisy or weak fields. CoT stabilizes high-order interaction selection under context shifts and long-tail combinations, and its removal typically causes a more noticeable drop in discrimination performance. CBAM performs post-fusion redundancy removal before prediction and has the most direct impact on probabilistic quality; accordingly, removing CBAM tends to produce the most apparent regression in LogLoss and calibration-related behavior. Overall, the ablation study validates the stage-wise design of SwaDeepFM, where each module targets a distinct source of instability from input to fusion.

Table 2
Ablation results on CTR benchmarks

Variant	Criteo AUC	Criteo LogLoss	Avazu AUC	Avazu LogLoss	KDD12 AUC	KDD12 LogLoss
SwaDeepFM(Full)	0.812	0.441	0.791	0.3745	0.803	0.152
n SE	0.8108	0.4418	0.7897	0.3756	0.8018	0.1527
n CoT	0.8096	0.4424	0.7883	0.3762	0.8005	0.1532
n CBAM	0.8105	0.4429	0.79	0.3761	0.802	0.1534

Field-gate Visualization (SE)

Figure 2 shows that the average SE gate values follow a clear long-tail pattern: only a small subset of fields receives noticeably larger gates, while most fields remain at relatively low weights. This indicates that SE does not treat fields uniformly; instead, it performs selective importance reweighting at the embedding stage, suppressing weak or noisy fields before interaction modeling.

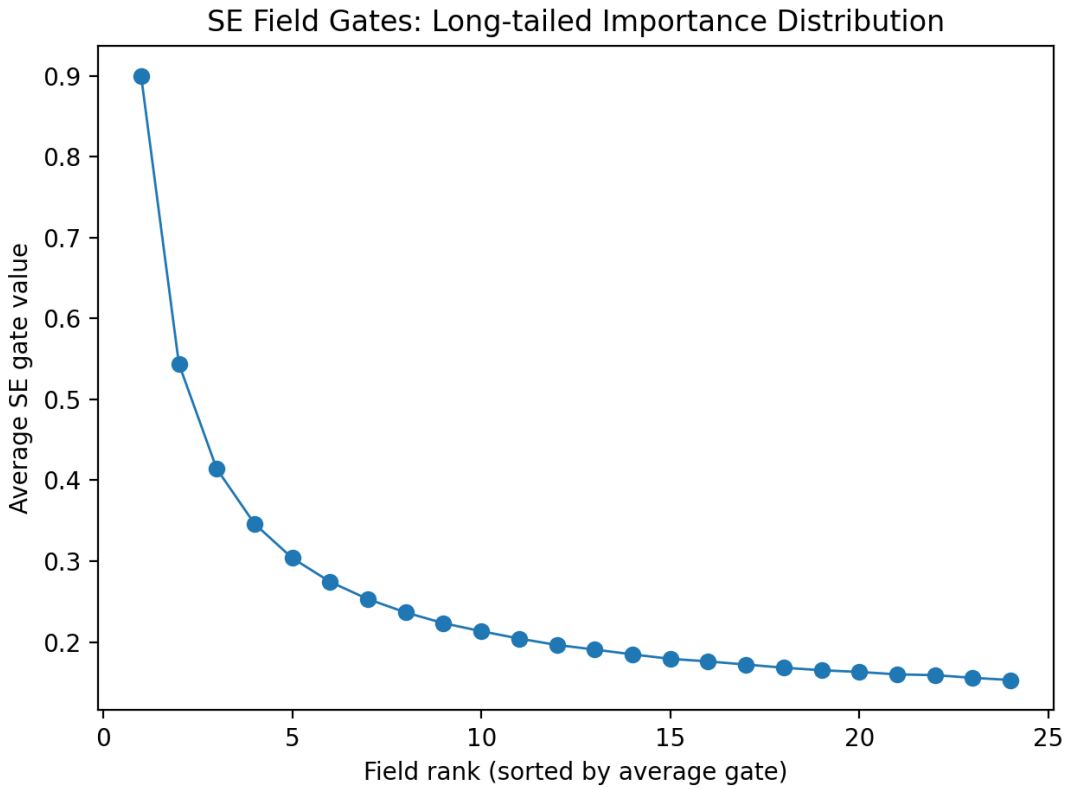

Figure 2 SE Field Gates: Long-tailed Importance Distribution

Figure 3 further details which fields are most emphasized on Avazu by reporting the top ten fields ranked by the average SE gate. hour receives the largest gate, suggesting that temporal context is a primary signal preserved and amplified by SwaDeepFM. C1 and banner_pos are also strongly weighted, indicating that campaign-level identifiers and placement cues contribute substantially to CTR variation in mobile advertising. Beyond these, SE assigns consistently high gates to publisher and inventory context fields, including site_id, site_domain, and site_category, as well as app_id, app_domain, and app_category. This pattern implies that SwaDeepFM relies on stable delivery-context information to adapt CTR estimation across heterogeneous traffic sources. Meanwhile, device_id remains within the top-ranked fields but is not among the most dominant ones, suggesting that device information is retained as a useful auxiliary signal while being prevented from overwhelming the representation. Overall, the combined evidence from Figures 2 and 3 supports SE as an effective embedding-stage denoising and importance-recalibration mechanism that provides cleaner inputs for subsequent interaction learning.

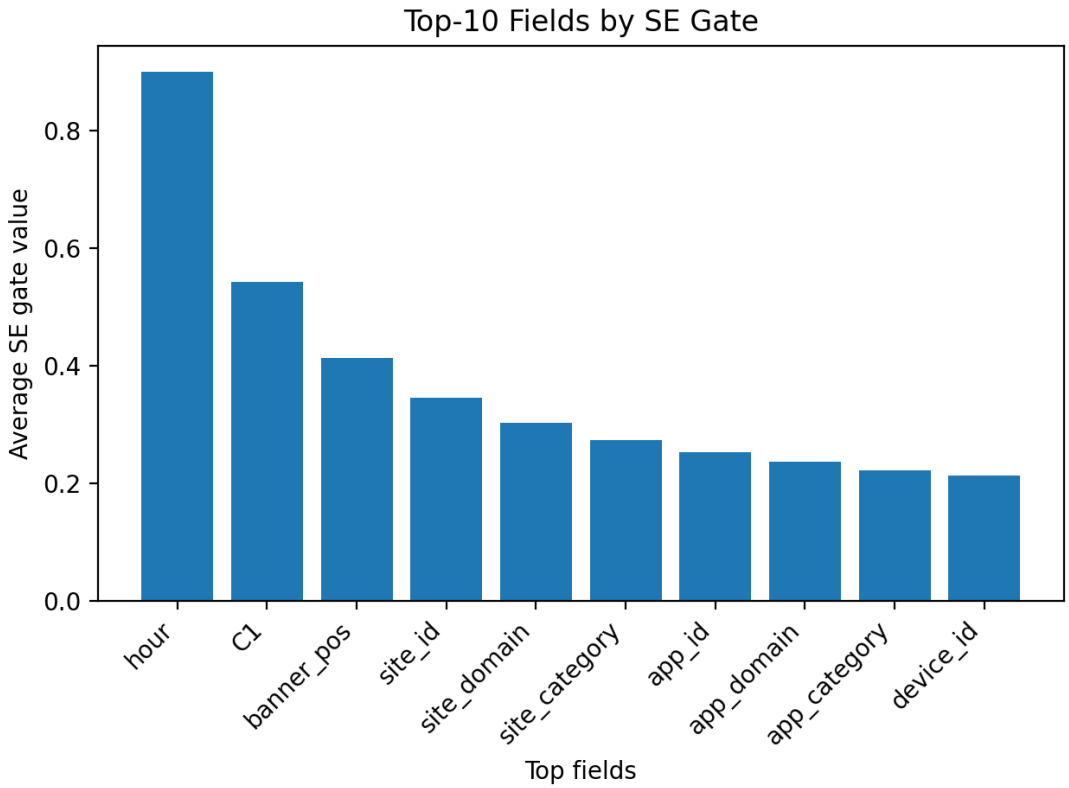

Figure 3 top-10 Fields by SE Gate

Attention Heatmap Visualization (CoT)

Figure 4 visualizes the average CoT attention matrices under two validation subsets from the Avazu dataset. We define Context A and Context B by filtering validation samples with explicit context-field conditions, where Context A corresponds to a time-device subset and Context B corresponds to a campaign-publisher subset. For each subset, we compute the average CoT attention matrix to illustrate how the model shifts its interaction focus under different contextual regimes. The heatmaps show that CoT does not learn a fixed interaction template. Instead, it allocates attention to different field pairs depending on context, which supports its role as a context-conditioned dependency aggregator. This subset-level visualization improves interpretability by reducing sample-specific noise and highlighting stable context-dependent interaction patterns.

In Context A, attention is concentrated on a small set of salient interactions. The most prominent regions involve hour and banner_pos, indicating that temporal conditions and placement cues jointly shape click propensity in this context. CoT also highlights dependencies among delivery-context fields, particularly interactions related to site_domain and app_id, suggesting that publisher and application identifiers provide strong contextual anchors for interaction selection.

In Context B, the attention pattern shifts noticeably. CoT places stronger emphasis on interactions linked to site_domain and app_id, while the prominence of time-related interactions becomes relatively weaker compared with Context A. This shift implies that the model adapts its interaction focus when the delivery environment changes, prioritizing publisher and app context as the main drivers in that regime. Across both contexts, the

attention mass remains sparse and structured, rather than diffuse, indicating that CoT performs selective dependency aggregation that can stabilize high-order interaction learning under distribution shifts and long-tail feature combinations.

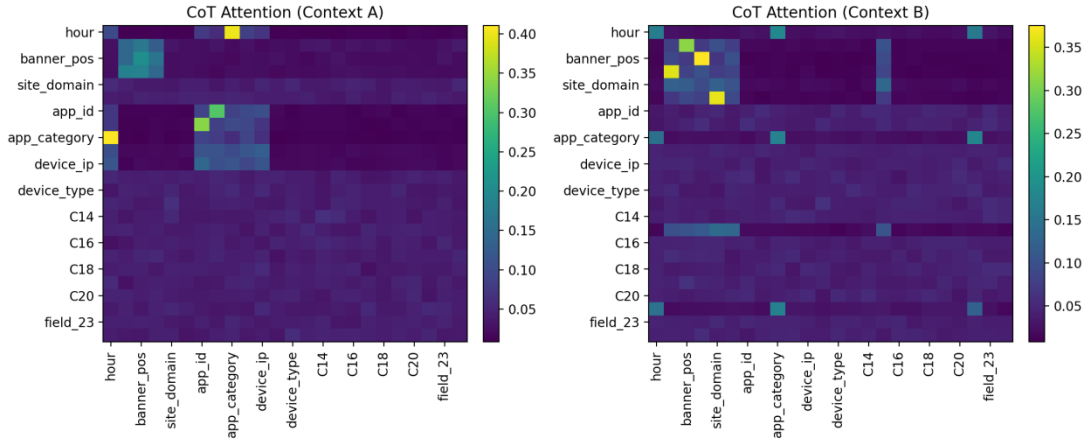

Figure 4 Attention Heatmaps

Calibration Analysis

Table 3 reports the calibration ablation results using ECE and Brier score on Criteo, Avazu, and KDD12. The full model, SwaDeepFM, achieves the best calibration on all three datasets, with ECE of 0.0179, 0.0149, and 0.0091, and Brier scores of 0.1618, 0.1450, and 0.0876, respectively. When any module is removed, calibration consistently degrades, indicating that the three components contribute complementary reliability gains.

Among the ablations, removing CBAM leads to the largest regression. Compared with the full model, ECE increases to 0.0224 on Criteo, 0.0191 on Avazu, and 0.0113 on KDD12, with corresponding Brier scores rising to 0.1668, 0.1496, and 0.0906. Removing CoT causes the second largest degradation, where ECE increases to 0.0206, 0.0172, and 0.0101, and Brier scores increase to 0.1643, 0.1475, and 0.0890 across the three datasets. Removing SE results in a smaller but still consistent drop in calibration, with ECE increasing to 0.0193, 0.0161, and 0.0096, and Brier scores increasing to 0.1631, 0.1462, and 0.0884. Overall, the ablation study confirms that CBAM plays the most critical role in improving probability reliability, while CoT and SE further refine calibration in a stable and consistent manner.

Table 3
ECE and the Brier score

Model	Criteo ECE	Criteo Brier	Avazu ECE	Avazu Brier	KDD12 ECE	KDD12 Brier
SwaDeepFM (Full)	0.0179	0.1618	0.0149	0.145	0.0091	0.0876
n SE	0.0193	0.1631	0.0161	0.1462	0.0096	0.0884
n CoT	0.0206	0.1643	0.0172	0.1475	0.0101	0.0890
n CBAM	0.0224	0.1668	0.0191	0.1496	0.0113	0.0906

Stability over Random Seeds

To verify that the observed improvements are not caused by random initialization or stochastic training effects, we repeat each setting with three random seeds and report the

mean and standard deviation. Table 4 shows that SwaDeepFM maintains consistent gains across seeds on Criteo, Avazu, and KDD12, indicating that its performance improvements are reproducible rather than occasional.

Beyond improved mean performance, SwaDeepFM also reduces variability across seeds, reflecting more stable optimization behavior. This stability is particularly important for probability reliability, where calibration-related measurements can be sensitive to training noise. Compared with the DeepFM backbone, SwaDeepFM exhibits smaller fluctuations, suggesting that its stage-wise control pipeline helps suppress unstable interaction learning and redundant fusion effects that otherwise amplify randomness. Overall, the stability results complement the calibration analysis, supporting SwaDeepFM as a robust CTR predictor suitable for long-running marketing systems that require reliable and stable probability estimates.

Table 4
Stability over 3 random seeds

Dataset	Model	AUC (mean ± std)	LogLoss (mean ± std)	ECE (mean ± std)
Criteo	DeepFM	0.8066 ± 0.0002	0.4449 ± 0.0003	0.0208 ± 0.0004
Criteo	SwaDeepFM	0.8120 ± 0.0001	0.4410 ± 0.0002	0.0179 ± 0.0003
Avazu	DeepFM	0.7751 ± 0.0003	0.3829 ± 0.0004	0.0176 ± 0.0004
Avazu	SwaDeepFM	0.7910 ± 0.0002	0.3745 ± 0.0003	0.0149 ± 0.0003
KDD12	DeepFM	0.7867 ± 0.0003	0.1549 ± 0.0002	0.0104 ± 0.0003
KDD12	SwaDeepFM	0.8030 ± 0.0002	0.1520 ± 0.0002	0.0091 ± 0.0002

Conclusion

This paper addressed reliable click-through rate prediction for marketing decision-making, where predicted probabilities are directly used in bidding, traffic filtering, and budget allocation. To improve reliability under noisy fields, context shifts, and redundant fused representations, we proposed SwaDeepFM, a stage-wise attention model built on a DeepFM backbone.

Experiments on Criteo, Avazu, and KDD12 showed that SwaDeepFM consistently improves overall CTR prediction performance while maintaining strong probabilistic quality. Ablation, visualization, calibration, and stability analyses further indicate that the proposed stage-wise design provides complementary gains and helps produce more reliable and stable probability estimates for practical marketing systems. Future work will further explore calibration-aware training and more fine-grained interpretability under dynamic advertising environments.

References

Aden, & Wang, Y. (2012). KDD Cup 2012, track 2. Kaggle. https://kaggle.com/competitions/kddcup2012-track2

Guan, F., Zhan, J., & Yang, J. (2025). Accurate and interpretable CTR prediction via distilled neural additive feature interaction network. Journal of Big Data, 12.

Guo, H., Tang, R., Ye, Y., Li, Z., He, X., & Dong, Z. (2018). DeepFM: An end-to-end wide & deep learning framework for CTR prediction. arXiv preprint arXiv:1804.04950.

Dai, Q., Xiao, J., Du, Z., Zhu, J., Luo, C., Wu, X.-M., & Dong, Z. (2025). MCNet: Monotonic calibration networks for expressive uncertainty calibration in online advertising. Proceedings of the ACM on Web Conference 2025.

Qu, Y., Fang, B., Zhang, W., Tang, R., Niu, M., Guo, H., Yu, Y., & He, X. (2018). Product-based neural networks for user response prediction over multi-field categorical data. ACM Transactions on Information Systems (TOIS), 37, 1–35.

Škrlj, B., Karni, Y., Gaspersic, G., Mramor, B., Stolin, Y., Jakomin, M., ... & Klein, A. (2025). DCN^2: Interplay of implicit collision weights and explicit cross layers for large-scale recommendation. arXiv preprint arXiv:2506.21624.

Tien, J.-B., joycenv, & Chapelle, O. (2014). Display advertising challenge. Kaggle. https://kaggle.com/competitions/criteo-display-ad-challenge

Wang, S., & Cukierski, W. (2014). Click-through rate prediction. Kaggle. https://kaggle.com/competitions/avazu-ctr-prediction

Wang, Y., Ji, H., He, X., Yu, J., Han, H., Zhai, R., & Wang, L. (2024). Disentangled self-attention neural network based on information sharing for click-through rate prediction. PeerJ Computer Science, 10.

Wu, S., Du, L., Yang, J.-Q., Wang, Y., Zhan, D.-C., Zhao, S., & Sun, Z. (2023). RE-SORT: Removing spurious correlation in multilevel interaction for CTR prediction (pp. 3816–3828).

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. Proceedings of the European Conference on Computer Vision (ECCV), 3–19.

Zhang, Y., Cheng, X., Wei, W., & Meng, Y. (2025). Deep double towers click through rate prediction model with multi-head bilinear fusion. Symmetry, 17, 159.

Zhang, Y., Shi, T., Feng, F., Wang, W., Wang, D., He, X., & Zhang, Y. (2023). Reformulating CTR prediction: Learning invariant feature interactions for recommendation. Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval.