WaySeq: Production Lessons from User Sequence Modeling for Sponsored Product Click and Conversion Prediction

Dongyue Xie, Sergey Kolbin, Manavender Malgireddy, Iurii Shcherbak, Weixing Tang, Arushi Jain, Wanchen Shao, Kurt Zimmer, Raja Hafiz Affandi, Matt Lindsay
{dxie,skolbin,mmalgireddy,ishcherbak,wtang,ajain31,wshao,kzimmer,raffandi,mlindsay}@wayfair.com
Wayfair LLC
USA

Abstract

User sequence modeling is widely used in recommendation and advertising. Sequence modules improve click and conversion prediction while satisfying auction-time serving constraints in large-scale sponsored ranking systems. We present WaySeq, Wayfair's first production framework for sequence-aware sponsored-products ranking, and use it to study which sequence modeling choices are most effective under these constraints. Starting from a lightweight transformer backbone, we evaluate sequence coverage, embedding representation, sequence summarization, tokenization, and attention structure. Our results show that the largest gains come from broader behavioral coverage, stronger product embeddings, and candidate-aware sequence summarization, while heavier behavior-aware attention variants add serving cost without consistent ranking-quality gains. The final production model delivers consistent online gains across engagement and conversion-oriented metrics. These findings suggest that, in industrial sponsored ranking, the interface between user history and the ad candidate, together with representation quality, is more important than simply increasing architectural complexity.

CCS Concepts

• **Information systems** → **Sponsored search advertising**.

Keywords

Sponsored Products, Advertising, Ranking Optimization, Transformer, User Sequence Modeling, Sequence-aware Ranking

ACM Reference Format:
Dongyue Xie, Sergey Kolbin, Manavender Malgireddy, Iurii Shcherbak, Weixing Tang, Arushi Jain, Wanchen Shao, Kurt Zimmer, Raja Hafiz Affandi, Matt Lindsay. 2026. WaySeq: Production Lessons from User Sequence Modeling for Sponsored Product Click and Conversion Prediction. In *Proceedings of KDD '26*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
KDD '26, Jeju, Korea
© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Ranking quality is a key driver of performance in sponsored products advertising and general recommendation tasks, where the model typically predicts both immediate engagement (CTR) and downstream outcomes such as conversion (CVR). Modern production rankers integrate high-capacity feature interaction modules, dense item representations, and behavior-aware components, and recent work has continued to push this trend toward larger and more unified ranking backbones [5, 9, 13, 14, 16, 17, 21].

Within this landscape, user sequence modeling is an increasingly standard ingredient of industrial recommendation and advertising systems. Transformer-based sequence encoders have shown strong performance in large-scale production settings, including e-commerce recommendation and real-time feed ranking [2, 6, 12, 16, 17]. One key question is which sequence-modeling components materially improve CTR and CVR prediction, since prior work often bundles architectural changes [5, 13, 16, 17, 21]. For sponsored products, the module must also respect auction-time probability calibration and per-request latency while scoring many candidates; therefore, our emphasis is on production design choices rather than a new transformer block.

In this work, we present WaySeq, Wayfair's first production framework for sequence-aware sponsored products ranking. Starting from a lightweight TransAct-style [16] backbone, rather than proposing a larger transformer architecture, we use WaySeq to identify which practical sequence-modeling choices matter most in this setting. Our key contributions are

(1) WaySeq, the first production sequence modeling framework at Wayfair, incorporating multi-action histories into a multi-task ranker.
(2) a production ablation study over sequence context, embeddings, summaries, tokenization, and attention variants, including quality-cost trade-offs;
(3) candidate-aware pooling as a soft-retrieval mechanism conditioned on the ad candidate being scored;
(4) an online A/B test on live sponsored-products browse traffic, showing gains across ad revenue and conversion-oriented metrics.

2 Related Work

Modern industrial rankers now combine large-scale sparse and dense features, deep interaction modules, pretrained item representations, and transformer-based backbones [5, 9, 13, 14, 21]. User behavior modeling is a core ingredient for CTR and CVR prediction [3, 7, 22, 23], and sequential recommendation has largely shifted to

transformer-based encoders. SASRec and BERT4Rec established the effectiveness of self-attention for sequence modeling, and recent surveys reflect its central role in recommender systems [6, 10, 12].

The most relevant prior work is industrial sequence modeling. BST applied transformer-based behavior modeling to e-commerce recommendation, while TransAct and TransAct V2 demonstrated real-time and longer-horizon user action modeling in large-scale ranking systems. More recent work such as OneTrans, TokenMixer-Large, and SORT points toward unified feature-interaction and sequence backbones at larger scale [2, 5, 13, 16, 17, 21]. Our work is closest to this industrial line, but focuses on the production design space of sequence modeling for sponsored products ranking, where CTR/CVR objectives, candidate conditioning, calibration, and real-time serving constraints must be considered jointly. The contribution is therefore a production evidence study: which known sequence components survive these constraints, which do not, and why.

3 Ranking Model at Wayfair

In this section, we describe the production ranking model used at Wayfair and the sequence-augmented architecture studied in this work. Table 1 summarizes notations.

3.1 Production Ranking

Wayfair's sponsored products ranking model is a multi-task deep neural network that estimates the likelihood that a customer will click and purchase a candidate product. Before auction integration, the final CTR and CVR predictions are calibrated using the post-hoc isotonic regression pipeline. These predictions are then consumed by the downstream ranking and auction stack to compute the final ranking score for sponsored products [8].

The baseline ranker consumes a heterogeneous feature set including customer, product, and request context. These features capture static and semi-static signals, and their interactions are modeled by Wukong [20] crossing layers. Like many conventional ranking models, it represents user behavior primarily through aggregate statistics and static features. To capture short-term user intent and unlock deeper personalization, we augment the baseline ranker with a sequence encoder over recent user interactions. The resulting architecture, which we call WaySeq, encodes a user's recent multi-action behavior sequence together with the candidate SKU being scored, and produces a sequence representation that is injected into the main ranking model. Figure 1 shows where WaySeq is integrated; it complements, rather than replaces, existing ranking features.

3.2 Baseline Sequence Module

We begin from a TransAct-style [16] sequence encoder, which serves as the reference formulation for this work. For each impression, we form a user behavior sequence that includes six action types: click, quick view, hot-deal page visit, favorite, add-to-cart, and purchase. The sequence is ordered by recency and drawn from a one-year lookback window. Each token is constructed using an early-fusion formulation that combines action type, historical SKU identity, candidate SKU identity, and position:

$$\mathbf{x}_i = \text{concat}\big[E_{\text{action}}(a_i) \parallel E_{\text{sku}}(s_i) \parallel E_{\text{sku}}(c)\big] + E_{\text{pos}}(i) \in \mathbb{R}^D,$$

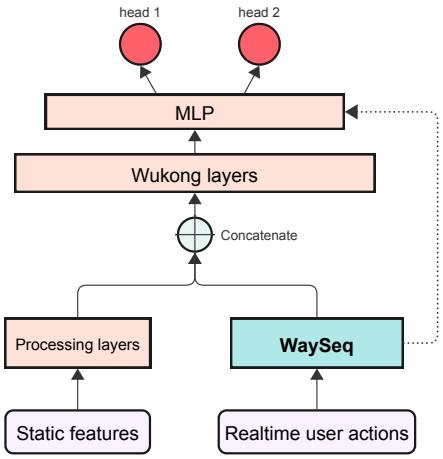

Figure 1: Overview of the Wayfair ranking model.

Table 1: Notation used in the WaySeq sequence module.

Symbol	Description
L	Maximum user sequence length
$l_{\min}$	Minimum user sequence length
K	Number of most recent sequence outputs
D	Transformer hidden dimension
$d_{\text{sku}}, d_{\text{action}}$	Dimensions of the SKU and action embedding
$E_{\text{action}}, E_{\text{sku}}, E_{\text{pos}}$	Action, SKU, and position embeddings
$\mathbf{x}_i$	Input token at sequence position i
A, M, K_{proj}	Action types, MI queries, projection rank
$\mathbf{H} \in \mathbb{R}^{L\times D}$	Transformer output; $\mathbf{h}_i$ is row i
$\mathbf{m} \in \{0,1\}^L$	Padding mask (m_i=1 for valid positions)
a_i, c	Action id and candidate SKU index

where $D = d_{\text{action}} + 2d_{\text{sku}}$. The token sequence is passed to a lightweight 2-layer transformer encoder with padding mask $\mathbf{m}$:

$$\mathbf{H} = \text{Transformer}(\mathbf{x}_{1:L}; \mathbf{m}) \in \mathbb{R}^{L\times D},$$

The baseline summary concatenates the K most recent output rows $\mathbf{R} = [\mathbf{h}_1; \ldots; \mathbf{h}_K]$ with elementwise max-pooling over valid positions. The baseline configuration uses $L = 100$, $l_{\min} = 10$, $K = 10$, and frozen 32-dimensional SKU embeddings from a pretrained semantic search model compressed using PCA.

3.3 Design Dimensions

We study sequence context and scale $\{L, l_{\min}, d_{\text{sku}}, K\}$, summary design, attention structure, SKU representation, and tokenization/position encoding. Figure 2 shows the overall architecture.

Sequence Summary. For richer summaries that are injected into the ranking backbone: one option is *sequence projection* (SP), which learns a linear projection over the full set of sequence outputs: $[L, D] \to [K_{\text{proj}}, D]$; a second option is *candidate-aware* (CA) pooling, which summarizes the sequence as:

$$\mathbf{p}^{\text{CA}} = \sum_{i=1}^{L} \alpha_i \mathbf{h}_i \in \mathbb{R}^D,$$

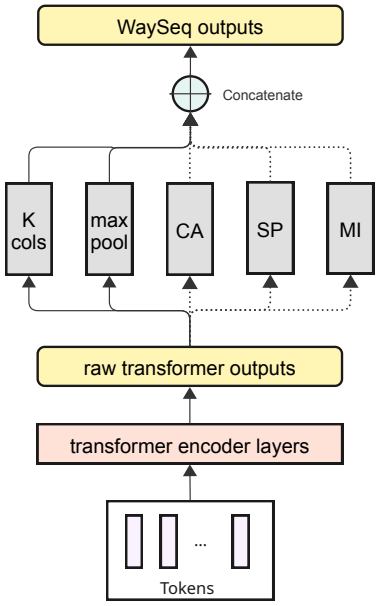

Figure 2: WaySeq architecture.

where

$$\alpha_i = \frac{\mathbf{1}[m_i = 1]\exp(s_i^{\mathrm{CA}})}{\sum_{j=1}^{L}\mathbf{1}[m_j = 1]\exp(s_j^{\mathrm{CA}})}, \quad s_i^{\mathrm{CA}} = \mathbf{h}_i^{\top}\tilde{\mathbf{c}}, \quad \tilde{\mathbf{c}} = \mathbf{W}_c E_{\mathrm{sku}}(c).$$

We also consider *multi-interest* (MI) pooling, which uses M learned queries to extract intent vectors: $\mathbf{p}_m^{\mathrm{MI}} = \sum_{i=1}^{L}\beta_{i,m}\mathbf{h}_i$ where $\beta_{i,m}$ is the softmax score similar to α_i above, while $s_{i,m}^{\mathrm{MI}} = \mathbf{h}_i^{\top}\mathbf{q}_m/\sqrt{D}$.

These summary mechanisms are not arbitrary variants: SP tests whether a learned low-rank compression can recover signals lost by fixed top-K and max pooling; CA pooling tests a candidate-conditioned retrieval view: the useful part of a user's history is the subset that is semantically and behaviorally aligned. MI pooling tests a broader shopping-mission view.

We also study *bias-aware* summary variants that inject action-type or positional priors into the summary scores. For example, action-type bias modifies the CA scores as $s_i^{\mathrm{CA}} \leftarrow \mathbf{h}_i^{\top}\tilde{\mathbf{c}} + b_{a_i}^{\mathrm{a,CA}}$, where $b^{\mathrm{a,CA}} \in \mathbb{R}^A$ is a learned scalar bias table.

Attention structure and task-specific adaptation. At the encoder level, we study behavior-aware attention variants inspired by MB-STR [18], including action-specific query/key/value projections, attention bias, value gating, and action-specific feed-forward networks.

At the prediction level, we study two task-specific adaptations, in which each task $t \in \{\text{click}, \text{conversion}\}$ uses its own sequence output $\mathbf{h}_t$ in addition to the shared ranking representation $\mathbf{z}$ from the Wukong layers. Then the final prediction is given by $\hat{y}_t = \mathrm{MLP}_t([\mathbf{z}\|\mathbf{h}_t])$, as shown in the dotted line in Figure 1. The first is a *head-specific summary*, where each task t has its own sequence summary $\mathbf{h}_t$ that is chosen from the sequence summary design. The second is *per-task gating*, which reweights a shared sequence output $\mathbf{O}^{\mathrm{shared}} = [\mathbf{o}_1, \ldots, \mathbf{o}_N]$ using task-dependent gates: $\mathbf{g}_t = \sigma(\mathbf{W}_g^{(t)}\mathbf{z}), \mathbf{h}_t = \sum_{n=1}^{N} g_{t,n}\mathbf{o}_n$.

Table 2: Reference model for each comparison block.

Block	Reference for deltas
Scaling	Initial WaySeq: L=100, $l_{\min}$=10, K=10; PCA-32 SKU embedding.
Embedding	Same setup, replacing PCA-32 with AE-compressed semantic embedding, sem32.
Summary/token	sem32 WaySeq; default settings except the tested axis.
Attention/heads	sem32 + candidate-aware summary + action bias.
Online/final	Production ranker without WaySeq.

Embedding representation. This design concerns the SKU representation used throughout the sequence module. The baseline model uses a compressed semantic embedding. We additionally evaluate alternative pretrained internal product representations, including ProductDNA and Resonate embeddings, as well as concatenated representations such as $E_{\mathrm{sku}} = \mathrm{concat}[E_{\mathrm{res32}}\|E_{\mathrm{sem32}}]$.

Tokenization and positional encoding. Here we study how behavioral, semantic, and positional information are combined at the token level. We compare the baseline formulation with two alternatives. The first is a TransAct-v2-style [17] separated formulation:

$$\mathbf{x}_i^{\mathrm{sep}} = \mathrm{concat}[E_{\mathrm{sku}}(s_i)\|E_{\mathrm{sku}}(c)] + E_{\mathrm{action}}(a_i) + E_{\mathrm{pos}}(i) \in \mathbb{R}^D,$$

where $D = 2d_{\mathrm{sku}}$. The second removes absolute positional embeddings and instead introduces a relative position bias directly in attention:

$$e_{ij} = \frac{(\mathbf{W}_Q\mathbf{x}_i)^{\top}(\mathbf{W}_K\mathbf{x}_j)}{\sqrt{D}} + b(i-j).$$

4 Model Design Study

We perform a systematic ablation study on a Wayfair internal dataset, and report relative changes in percentage in overall AUC and grouped AUC (gAUC) for both click and conversion prediction, with module size and inference latency when relevant. Because the calibration layer is shared across variants and independent of the sequence module, the ablation study focuses on discrimination metrics. Due to business confidentiality constraints, all results are presented as relative deltas rather than absolute values.

Table 2 gives the reference model in each study. The staged design was intentional: early studies search broad capacity and context, later studies use the stronger embedding because that representation became the new internal baseline, and the attention study tests whether expensive action-aware mechanisms still help after the best lightweight summary. Therefore, numbers are comparable within a table but should not be summed across tables.

Scaling Context and Encoder Capacity. As shown in Table 3 , the strongest gains come from increasing the effective sequence context. Increasing L exposes more historical evidence, while lowering $l_{\min}$ broadens the share of sparse-history users eligible for sequence modeling. This distinction matters operationally: in the current candidate-conditioned design, increasing L raises per-candidate

Table 3: Relative change of AUC in the scaling study.

Metric	L=300	$l_{\min}$=1	d_{sku}=128	Large attn	K=20
Click AUC	**+0.441**	+0.384	+0.158	+0.173	+0.058
Order AUC	**+1.204**	+1.203	-0.151	+0.313	+0.012
Click gAUC	**+0.555**	+0.432	+0.266	+0.234	+0.109
Order gAUC	+0.226	**+0.381**	-1.091	+0.127	-0.652
Module size	+0	+0	+639	+161	+0
Latency	+62.6	+0.0	+33.3	+25.0	+5.0

Table 4: Relative change of AUC in the embedding study.

Metric	pDNA32	pDNA256	pDNA‖ sem	Res32	Res‖sem
Click AUC	-1.957	-2.326	+0.199	-0.681	**+0.681**
Order AUC	-0.767	-1.396	+0.011	-0.263	**+0.286**
Click gAUC	-2.304	-2.826	-0.353	-0.783	**+0.891**
Order gAUC	-1.939	-2.951	+0.225	-0.717	**+0.913**
Model size	+0	+2405	+148	+0	+148
Latency	+0	+100	+13	+0	+13

transformer work, whereas reducing $l_{\min}$ increases coverage without adding latency. By contrast, increasing SKU embedding dimension and attention capacity provide smaller or less consistent gains.

Embedding Representation. Prior to this study, we replaced the PCA-compressed embedding with an autoencoder-compressed version of a newer semantic embedding model. This single change improved click gAUC by 1.3% and conversion by 0.4%. Unless otherwise noted, later ablations use this updated semantic embedding, denoted as `sem32`.

Table 4 evaluates alternative SKU embeddings relative to this baseline. The most consistent gains come from the concatenated `Res32+sem32` representation, which improves all four AUC metrics with only modest increases in size and latency. In contrast, ProductDNA-only variants perform substantially worse, even at larger embedding sizes. We also tested using the same frozen SKU embedding as a static feature in the main ranking model. This variant improved click grouped AUC by 0.474% and order grouped AUC by 0.168%. Overall, these results show that embedding quality is a first-order factor for sequence modeling, and combining complementary pretrained representations is effective.

Sequence Summary Design. Table 5 compares summary mechanisms. The three individual summary mechanisms improve most metrics over the baseline. Candidate-aware summarization provides the strongest and most balanced gains among the individual variants, especially on conversion gAUC. Adding action bias further improves robustness and yields the best overall click-side improvements. All summary mechanisms add little model size and negligible latency. The combined summary improves click metrics but hurts conversion quality, which is consistent with label density rather than necessarily contradictory. Clicks are dense and immediate, so redundant SP/CA/MI features can help capture visual and semantic affinity. Orders are sparser and more sensitive to high-intent actions and candidate compatibility; adding several correlated summary paths can overfit frequent browsing behavior that explains

Table 5: Relative change of AUC in the sequence summary.

Metric	SP	CA	MI	combo	CA+act bias
Click AUC	+0.227	+0.270	+0.199	+0.298	**+0.340**
Order AUC	+0.092	+0.092	+0.011	-0.252	+0.092
Click gAUC	+0.338	+0.276	+0.184	**+0.445**	**+0.445**
Order gAUC	+0.000	**+0.422**	+0.155	-0.014	+0.225
Model size	+1.00	+3.05	+3.05	+4.44	+3.06

Table 6: Relative change of AUC in the tokenization and positional encoding study.

Metric	TransAct-v2	Rel pos bias	Pos bias
Click AUC	-0.128	-0.227	-0.141
Order AUC	-0.195	-0.126	+0.023
Click gAUC	-0.061	-0.122	-0.138
Order gAUC	-0.742	-0.084	+0.056
Model size	-43.0	+0.2	+0.1

clicks but not purchases. CA with action bias is therefore a better bias-variance trade-off for the final model.

Tokenization and Positional Encoding. Table 6 shows that none of the alternative tokenization or positional variants consistently improves over the baseline. Both the TransAct-v2-style token and relative position bias degrade most metrics, while positional bias in sequence summary (recall that the formulation is similar to the action bias) yields only marginal gains on conversion. This negative result should not be read as evidence that time is unimportant. It shows only that ordinal position variants did not help once recency-ordered actions and candidate-conditioned tokens were already present. Time-gap features, session boundaries, and event freshness buckets are a more plausible next step because they encode purchase-cycle dynamics that absolute or relative event index cannot represent.

Attention Structure. Table 7 evaluates richer attention mechanisms and task-specific sequence heads. The reference is the baseline model with the sem32 embedding and candidate-aware summary with action bias. Here, MBSTR-full enables all components, while MBSTR-qkv and MBSTR-ffn enable only the corresponding action-aware sub-module. The Task summary variant uses a task-specific summary based on candidate-aware pooling, multi-interest pooling, and action bias.

The MBSTR-style variants substantially increase model size and latency while degrading ranking quality. We interpret this as a mismatch between action-specific parameterization and the action distribution. Multi-action histories contain dense low-intent behaviors and much sparser high-intent behaviors; fully action-specific attention branches can therefore be estimated reliably for frequent actions but poorly for rare conversion-adjacent actions. If query/key/value or feed-forward parameters are split by action type, the rare high intent actions receive fewer training examples and the frequent low-intent actions dominate both optimization and attention mass. This explanation is qualitative, so we state it as an interpretation rather than as a diagnosed causal claim.

Table 7: Relative change of AUC in the attention structure.

Metric	MBSTR full	MBSTR qkv	MBSTR ffn	Task summary	Task gating
Click AUC	-2.106	-0.170	-0.396	+0.014	-0.057
Order AUC	-1.966	-0.103	-0.274	-0.103	-0.069
Click gAUC	-2.324	-0.229	-0.520	+0.061	-0.214
Order gAUC	-3.323	-0.210	-0.561	+0.014	+0.126
Model size	+463	+378	+85	+36	+0
Latency	+54.0	+43.0	+28.0	+6.0	+0.6

Table 8: Relative lift of WaySeq over the production baseline in online A/B testing on browse traffic.

Metric	Relative change (%)
CTR	+4.7
CPC	-2.8
Ad spend	+1.7
Orders / impression	+5.9
ROAS	+2.3
Customer-level CVR	+0.8

4.1 Practical Takeaways and Unified Model

The staged ablations suggest several practical lessons for deploying user sequence modeling in sponsored-products ranking. First, SKU representation quality is a first-order factor. Second, candidate-aware summarization provides a low-latency interface between user history and the sponsored-product candidate being scored. Finally, richer action-specific attention mechanisms can fail under severe action imbalance.

Relative to the baseline introduced in Section 3.1, the final configuration was selected based on joint validation of the full model rather than by greedily combining the best single-axis ablation result. It uses a default sequence context ($L = 100, l_{\min} = 1$), larger attention capacity, candidate-aware summarization with action bias, concat`[sem32‖res32]` as SKU embeddings, the same frozen SKU embedding as a static feature in the main ranker, and a task-specific summary based on candidate-aware and multi-interest pooling.

The final model improves click overall AUC by 2.79%, click gAUC by 3.85%, order overall AUC by 2.22%, and order gAUC by 2.20%, at the cost of a 622.98% increase in model size and a 44.89% increase in inference latency. Owing to the efficient serving architecture described in Section 5, this additional inference cost translates to only a small single-digit increase in end-to-end p99 latency.

5 Model Serving

WaySeq is served in the real-time sponsored-products ranking path, where each request requires scoring a large candidate set under tight latency constraints. The serving architecture separates feature orchestration from model execution. An online scoring service handles request routing and feature collection, while neural inference is performed by a dedicated GPU-backed inference service. This separation allows the front-end service to scale with CPU and I/O demand, while the inference back end scales independently with GPU utilization.

The main addition required by WaySeq is online construction of user sequence features. For each request, the scoring service retrieves recent customer actions. These events are then merged into a reverse-chronological sequence. We highlight that the user-action service is refreshed in real time, allowing the sequence tensors to reflect the customer's most recent behavior at request time. To keep latency within budget, the serving graph executes independent feature fetches in parallel, uses batch retrieval for product features, and batches candidate scoring in the inference back end. For users with empty histories, the sequence output is zeroed and the ranker falls back to its non-sequence behavior; aggregate online lift should therefore be interpreted as the combined effect across users with and without usable histories.

The current implementation embeds the candidate in each sequence token, so transformer work scales with the candidate set. This design is simple and effective for first deployment, but it is not the only serving design. Longer-history systems can decouple history retrieval from candidate scoring using approximate nearest-neighbor or hierarchical compression before sequence encoding. We view this as the main serving limitation and the most important path for extending beyond the production $L = 100$ setting.

6 Online Experiments

We evaluate the final WaySeq model through online A/B testing. The experiment was conducted on live sponsored-products browse traffic for five weeks using standard user-level randomization, with traffic split evenly between the production baseline and the WaySeq variant. The experiment was evaluated using Wayfair's online experimentation framework, with the final launch decision based on a standard review of online ranking metrics.

As shown in Table 8, WaySeq delivers consistent gains over the production baseline across the primary business metrics, with all reported improvements statistically significant at the 5% level. In particular, we observe higher click-through rate, higher customer-level CVR and orders per impression. These results show that the offline gains from sequence modeling transfer to the live ranking environment and generate measurable business value. CPC decreases, which is favorable to advertisers but could be negative for platform revenue if volume did not compensate. In this experiment, ad spend, orders per impression, and ROAS increase together, indicating that improved matching increased total value despite the lower average unit click price.

An alternative design is to inject a standalone MB-STR-style personalization score directly into the ranking model as a feature. We do not report comparable deltas for that system in this short paper. For deployment, we favored integrated sequence modeling because the sequence representation is learned jointly with the CTR/CVR ranker and can interact with candidate, request, and product-cross features inside the same model.

At the time of writing, we are exploring how the sequence-aware ranking framework may generalize to other recommendation and discovery contexts beyond the initial deployment setting.

7 Discussion and Future Work

The main lesson from WaySeq is that effective sequence modeling for sponsored-products ranking depends less on architectural complexity and more on designing the right interface between user history, candidate products, and production constraints. In our setting, broader behavior coverage, stronger SKU representations, and candidate-aware summarization provide the most reliable gains, while additional architectural complexity provides smaller returns. In deployment, calibrated probabilities are handled by the shared isotonic pipeline, and users without histories fall back through zeroed sequence outputs.

Three directions appear especially promising. First, extending the model to much longer user histories may capture richer long- and short-term preferences, as recent industrial systems have begun to support ultra-long behavior sequences through hierarchical compression and more efficient serving designs [1, 11]. Second, the sequence can be enriched with stronger multimodal and contextual signals, including visual content, richer text representations, and fine-grained temporal context, following recent progress in multimodal sequential recommendation and vision-language recommendation systems [4, 15]. Third, generative and pretraining-based approaches to user behavior modeling are becoming increasingly attractive, suggesting a path toward sequence models trained with next-action or sequence-transduction objectives rather than only discriminative ranking losses [19]. Taken together, these directions suggest that the next generation of industrial sponsored ranking systems will likely combine longer-context sequence modeling, richer product understanding, and more general sequence pretraining objectives.

References

[1] Zheng Chai, Qin Ren, Xijun Xiao, Huizhi Yang, Bo Han, Sijun Zhang, Di Chen, Hui Lu, Wenlin Zhao, Lele Yu, Xionghang Xie, Shiru Ren, Xiang Sun, Yaocheng Tan, Peng Xu, Yuchao Zheng, and Di Wu. 2025. LONGER: Scaling Up Long Sequence Modeling in Industrial Recommenders. *CoRR* abs/2505.04421 (2025). doi:10.48550/arXiv.2505.04421

[2] Qiwei Chen, Huan Zhao, Wei Li, Pipei Huang, and Wenwu Ou. 2019. Behavior Sequence Transformer for E-commerce Recommendation in Alibaba. *CoRR* abs/1905.06874 (2019). doi:10.48550/arXiv.1905.06874

[3] Yufei Feng, Fuyu Lv, Weichen Shen, Menghan Wang, Fei Sun, Yu Zhu, and Keping Yang. 2019. Deep Session Interest Network for Click-Through Rate Prediction. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. 2301–2307. doi:10.24963/ijcai.2019/319

[4] Ramin Giahi, Kehui Yao, Sriram Kollipara, Kai Zhao, Vahid Mirjalili, Jianpeng Xu, Topojoy Biswas, Evren Korpeoglu, and Kannan Achan. 2025. VL-CLIP: Enhancing Multimodal Recommendations via Visual Grounding and LLM-Augmented CLIP Embeddings. In *Proceedings of the 19th ACM Conference on Recommender Systems (RecSys '25)*. doi:10.1145/3705328.3748064

[5] Yuchen Jiang, Jie Zhu, Xintian Han, Hui Lu, Kunmin Bai, Mingyu Yang, Shikang Wu, Ruihao Zhang, Wenlin Zhao, Shipeng Bai, Sijin Zhou, Huizhi Yang, Tianyi Liu, Wenda Liu, Ziyan Gong, Haoran Ding, Zheng Chai, Deping Xie, Zhe Chen, Yuchao Zheng, and Peng Xu. 2026. TokenMixer-Large: Scaling Up Large Ranking Models in Industrial Recommenders. *CoRR* abs/2602.06563 (2026). doi:10.48550/arXiv.2602.06563

[6] Wang-Cheng Kang and Julian McAuley. 2018. Self-Attentive Sequential Recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*. 197–206.

[7] 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. *CoRR* abs/1804.07931 (2018). doi:10.48550/arXiv.1804.07931

[8] Manavender Malgireddy, Dongyue Xie, Sergey Kolbin, Kurt Zimmer, Masoum Mosmer, and Patrick Phelps. 2025. Profit Aware Ad Ranking with Relevance. In *Computational Advertising: 19th International Workshop, AdKDD 2025, Toronto, ON, Canada, August 4, 2025, Proceedings*. Springer Nature, 71.

[9] Maxim Naumov, Dheevatsa Mudigere, Hao-Jui Michael Shi, Jianyu Huang, Narayanan Sundaram, Jongsoo Park, Xiaodong Wang, Udit Gupta, Carole-Jean Wu, Alisson G. Azzolini, Dmytro Dzhulgakov, Andrey Mallevich, Ilia Cherniavskii, Yinghai Lu, Raghuraman Krishnamoorthi, Ansha Yu, Volodymyr Kondratenko, Stephanie Pereira, Xianjie Chen, Wenlin Chen, Vijay Rao, Bill Jia, Liang Xiong, and Misha Smelyanskiy. 2019. Deep Learning Recommendation Model for Personalization and Recommendation Systems. *CoRR* abs/1906.00091 (2019). doi:10.48550/arXiv.1906.00091

[10] Liwei Pan, Weike Pan, Meiyan Wei, Hongzhi Yin, and Zhong Ming. 2024. A Survey on Sequential Recommendation. *CoRR* abs/2412.12770 (2024). doi:10.48550/arXiv.2412.12770

[11] Zihua Si, Lin Guan, ZhongXiang Sun, Xiaoxue Zang, Jing Lu, Yiqun Hui, Xingchao Cao, Zeyu Yang, Yichen Zheng, Dewei Leng, Kai Zheng, Chenbin Zhang, Yanan Niu, Yang Song, and Kun Gai. 2024. TWIN V2: Scaling Ultra-Long User Behavior Sequence Modeling for Enhanced CTR Prediction at Kuaishou. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24)*. doi:10.1145/3627673.3680030

[12] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 1441–1450. doi:10.1145/3357384.3357895

[13] Chunqi Wang, Bingchao Wu, Taotian Pang, Jiahao Wang, Jie Yang, Jia Liu, Hao Zhang, Hai Zhu, Lei Shen, Shizhun Wang, Bing Wang, and Xiaoyi Zeng. 2026. SORT: A Systematically Optimized Ranking Transformer for Industrial-scale Recommenders. *CoRR* abs/2603.03988 (2026). doi:10.48550/arXiv.2603.03988

[14] Ruoxi Wang, Rakesh Shivanna, Derek Zhiyuan Cheng, Sagar Jain, Dong Lin, Lichan Hong, and Ed H. Chi. 2021. DCN V2: Improved Deep & Cross Network and Practical Lessons for Web-scale Learning to Rank Systems. In *Proceedings of the Web Conference 2021*. 1785–1797. doi:10.1145/3442381.3450078

[15] Yuxiang Wang, Xin Shi, and Xueqing Zhao. 2025. MLLM4Rec: multimodal information enhancing LLM for sequential recommendation. *Journal of Intelligent Information Systems* 63 (2025), 745–761. doi:10.1007/s10844-024-00915-3

[16] Xue Xia, Pong Eksombatchai, Nikil Pancha, Dhruvil Deven Badani, Po-Wei Wang, Neng Gu, Saurabh Vishwas Joshi, Nazanin Farahpour, Zhiyuan Zhang, and Andrew Zhai. 2023. TransAct: Transformer-based Realtime User Action Model for Recommendation at Pinterest. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5249–5259.

[17] Xue Xia, Saurabh Vishwas Joshi, Kousik Rajesh, Kangnan Li, Yangyi Lu, Nikil Pancha, Dhruvil Deven Badani, Jiajing Xu, and Pong Eksombatchai. 2025. TransAct V2: Lifelong User Action Sequence Modeling on Pinterest Recommendation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*. doi:10.48550/arXiv.2506.02267

[18] Enming Yuan, Wei Guo, Zhicheng He, Huifeng Guo, Chengkai Liu, and Ruiming Tang. 2022. Multi-behavior sequential transformer recommender. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*. 1642–1652.

[19] Jiaqi Zhai, Lucy Liao, Xing Liu, Yueming Wang, Rui Li, Xuan Cao, Leon Gao, Zhaojie Gong, Fangda Gu, Jiayuan He, Yinghai Lu, and Yu Shi. 2024. Actions Speak Louder than Words: Trillion-Parameter Sequential Transducers for Generative Recommendations. In *Proceedings of the 41st International Conference on Machine Learning (PMLR, Vol. 235)*. 58484–58509. https://proceedings.mlr.press/v235/zhai24a.html

[20] Buyun Zhang, Liang Luo, Yuxin Chen, Jade Nie, Xi Liu, Daifeng Guo, Yanli Zhao, Shen Li, Yuchen Hao, Yantao Yao, et al. 2024. Wukong: Towards a scaling law for large-scale recommendation. *arXiv preprint arXiv:2403.02545* (2024).

[21] Zhaoqi Zhang, Haolei Pei, Jun Guo, Tianyu Wang, Yufei Feng, Hui Sun, Shaowei Liu, and Aixin Sun. 2026. OneTrans: Unified Feature Interaction and Sequence Modeling with One Transformer in Industrial Recommender. In *Proceedings of The Web Conference 2026*. doi:10.48550/arXiv.2510.26104 Accepted at WWW 2026; arXiv:2510.26104.

[22] Guorui Zhou, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai. 2019. Deep Interest Evolution Network for Click-Through Rate Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 5941–5948. doi:10.1609/aaai.v33i01.33015941

[23] Guorui Zhou, Chengru Song, Xiaoqiang 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. doi:10.1145/3219819.3219823