# MORE THAN A MATCH: STATE-INFORMED ATTENTION FOR UNIFIED RANKING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Unified ranking models use heterogeneous features and behavioral histories as evidence for predicting user feedback. Standard attention selects this evidence through contextualized query–key compatibility, expressing content matching and predictive usefulness through the same score. We introduce RISE (Ranking with Input-dependent State Enhancement), a plug-in attention module that provides an explicit score-level channel for contextual importance. RISE combines Amplitude Modulation (AM), which accumulates input-dependent decay controls to regulate auxiliary interaction strength, and Phase Alignment (PA), which accumulates input-dependent phase controls to adjust the relative orientation of low-dimensional auxiliary projections. The transformed projections produce a signed, query-specific correction to the original attention score, while preserving the backbone's native score-to-weight rule and value aggregation. Experiments across multiple backbones and datasets show that RISE improves average AUC over the corresponding baselines. These results support explicit contextual-importance modeling alongside the representations learned by ranking backbones.

## 1 INTRODUCTION

In recommendation, ranking models predict user feedback on candidate items from user and item features and behavioral histories. Feature-interaction architectures combine wide and deep components, factorized interactions, or explicit cross networks (Cheng et al., 2016; Guo et al., 2017; Wang et al., 2021). Feature-interaction models such as RankMixer (Zhu et al., 2025) encode histories through a separate sequence module and combine the resulting representations with user, candidate, and cross features through token mixing. Their mixing blocks operate on feature groups, leaving event-level temporal modeling to the sequence module. Recent unified models such as HyFormer (Huang et al., 2026) and OneTrans (Zhang et al., 2026b) bring sequence modeling and feature interaction into a common backbone. OneRank (Tang et al., 2026) extends unified modeling to task-specific encoding and context-aware candidate prediction.

Historical interactions provide evidence about a user's interests, and their predictive value depends on both the candidate item and the surrounding behaviors. DIEN (Zhou et al., 2019) models target-conditioned interest evolution, while DSIN (Feng et al., 2019) captures patterns within sessions and dependencies across sessions. For example, the same tent interaction may provide different evidence for a tent candidate when followed by sleeping-bag and camping-stove interactions than when followed by suitcase and shirt interactions(Figure 1). The surrounding behaviors place the shared interaction in different historical contexts. This raises a question about *contextual importance*: can an explicit cumulative context signal complement learned representations in selecting evidence for a candidate item?

**Proposed Method.** We introduce **RISE** (Ranking with Input-dependent State Enhancement), a plug-in method that incorporates cumulative context into ranking scores (Figure 2). RISE comprises two complementary modules: Amplitude Modulation (AM) and Phase Alignment (PA). AM accumulates input-dependent decay controls to regulate auxiliary interaction strength. PA accumulates input-dependent phase controls to rotate auxiliary query and key vectors, adjusting their relative orientation. Both modules follow the backbone's packed-token order and jointly produce a signed correction to the original interaction score. The backbone retains its native weighting rule and value aggregation.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

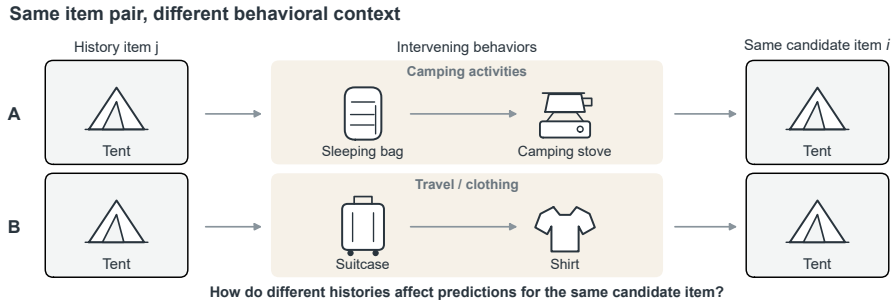



Figure 1: Illustrative behavioral histories for the same candidate item. Both histories contain an earlier tent interaction, while the intervening camping and travel-related behaviors provide different contexts.

We evaluate RISE on HiFormer, HyFormer, RankMixer, Zenith, and OneTrans across four datasets, comparing each backbone with its RISE-enhanced counterpart. RISE improves label-averaged ROC-AUC in all 20 model–dataset settings, with an absolute gain of 0.005881 averaged equally across settings. A component ablation on HyFormer further examines AM and PA: their combination achieves higher macro AUC than either transformation alone or the auxiliary content-only channel.

Our contributions are threefold:

- We develop RISE, a plug-in method that combines AM and PA to model cumulative context through a low-dimensional score correction. We specify its integration with heterogeneous ranking backbones, including token layouts and decay references, while retaining their native weighting rules and value aggregation.
- We characterize the distinct scale and orientation roles of AM and PA, including the composition of cumulative interaction maps. We bound the resulting redistribution of softmax weights and value aggregation, connecting the correction range to changes in the backbone's evidence selection.
- We evaluate RISE across five ranking backbones and four datasets, observing higher label-averaged ROC-AUC in all 20 settings under the reported aggregation protocol. A component study on HyFormer–MerRec further examines AM and PA, showing that their combination outperforms either module alone and the auxiliary content-only channel in macro AUC.

## 2 Related Work

**Unified ranking and contextual behavior modeling.** Unified ranking combines behavioral evidence with heterogeneous feature interactions. Self-attention captures sequential context (Kang & McAuley, 2018), while candidate-conditioned interest and session models capture evolving preferences (Zhou et al., 2018; 2019; Feng et al., 2019). Alternative encoders use selective state spaces for recommendation or retention for general sequence modeling (Liu et al., 2024; Sun et al., 2023). Candidate-aware retrieval extends history modeling to longer sequences (Pi et al., 2020; Chang et al., 2023). OneTrans and HyFormer integrate sequence and feature modeling (Zhang et al., 2026b; Huang et al., 2026), whereas RankMixer (Zhu et al., 2025) scales feature interaction through token mixing. RISE complements these designs through cumulative-context scores at existing attention sites, rather than replacing their contextual encoders.

**Separating selection from representation.** Side-information methods vary which features determine selection and which contribute to representations. FDSA and MSSR learn item- and attribute-specific representations (Zhang et al., 2019; Lin et al., 2024); NOVA and DIF-SR incorporate attributes into attention scores while preserving item-centered values (Liu et al., 2021; Xie et al., 2022). ASIF aligns source representations, while DIFF combines early and intermediate fusion

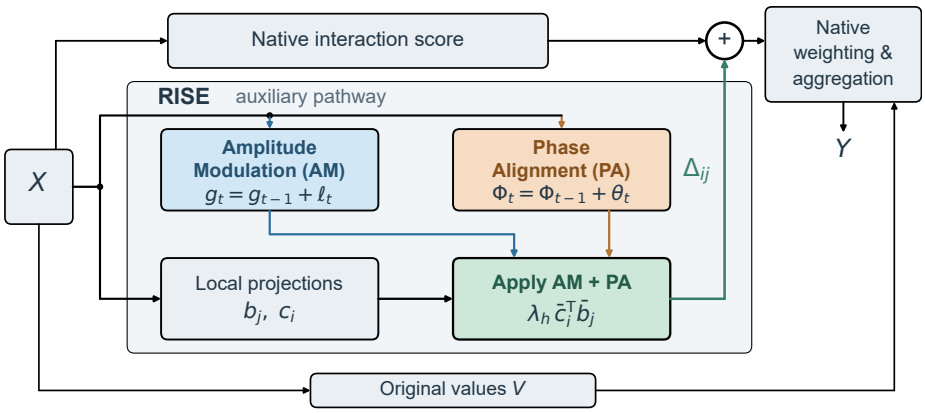



Figure 2: RISE's two-module auxiliary pathway. From token representations $X$, Amplitude Modulation (AM) accumulates decay controls and Phase Alignment (PA) accumulates phase controls along the packed-token order. Their states scale and rotate local auxiliary vectors; the head-scaled inner product gives the signed correction $\Delta_{ij}$. The native score, weighting rule, and value pathway $V$ are retained. AM uses the site-specific decay reference and numerical stabilization in Appendix A.

(Wang et al., 2024; Kim et al., 2025). CARCA follows profile encoding with target-level cross-attention (Rashed et al., 2022), and TIN distinguishes target-aware attention from target-aware behavior representations (Zhou et al., 2024). DARE explicitly separates correlation and representation embedding tables to address competing optimization demands (Feng et al., 2025). These approaches reorganize information sources and computational roles; RISE retains native scoring and value aggregation while adding cumulative context to the selection score.

**Context- and state-dependent attention scores.** Positional and temporal attention mechanisms encode relative displacement or interaction times (Shaw et al., 2018; Li et al., 2020; Zhai et al., 2024), including distance biases and query/key rotations (Press et al., 2021; Su et al., 2024). CoPE makes positional addressing context dependent through accumulated query-conditioned gates (Golovneva et al., 2024). Related controls update content-bearing recurrent memories through gates in Mamba and GLA (Gu & Dao, 2023; Yang et al., 2023), or rotations in Mamba-3 (Lahoti et al., 2026). The Forgetting Transformer adds cumulative input-dependent log-forget biases to softmax logits (Lin et al., 2025). RISE uses cumulative AM and PA controls to transform auxiliary scoring projections for heterogeneous ranking backbones, while retaining native scoring, weighting, and value aggregation.

## 3 State-Informed Attention for Unified Ranking

At a ranking attention site, let $X = (x_1, \ldots, x_L)^\top \in \mathbb{R}^{L \times D}$ contain the backbone's current token representations. RISE adds a signed, query-specific correction to the native compatibility score:

$$S_{ij} = \underbrace{S_{ij}^{\text{base}}}_{\text{base attention}} + \underbrace{\Delta_{ij}}_{\text{contextual importance}} . \tag{1}$$

The native score retains its original scaling, typically $q_i^\top k_j / \sqrt{d_h}$ for query $q_i$, key $k_j$, and head dimension $d_h$ (Vaswani et al., 2017). RISE preserves the backbone's weighting rule and value aggregation (Figure 2). Both score pathways may encode contextual relevance.

### 3.1 Auxiliary Projections and Control Signals

RISE maintains the native query/key projections and uses a separate low-dimensional correction channel. The per-head auxiliary width $r$ is an even-valued hyperparameter chosen independently of

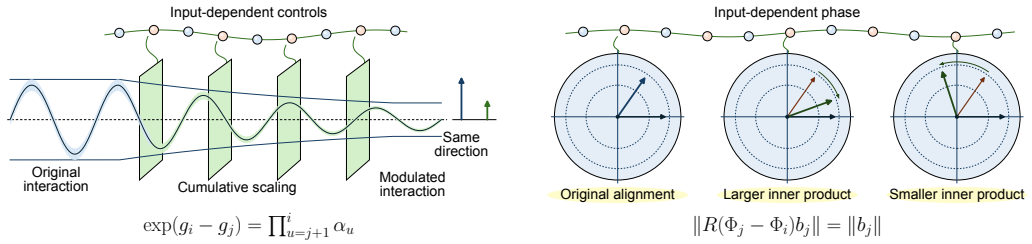



Figure 3: Amplitude and alignment in RISE. Left: AM scales auxiliary interaction strength, illustrated for an earlier key and later query with query reference and inactive clipping. Right: PA preserves vector norms while changing relative orientation; rotation can increase or decrease the inner product with a fixed query. Blue and orange vectors denote original keys, and green denotes rotated keys. Curves and angles are schematic.

the native head dimension $d_h$. For each valid token,

$$b_t = W_b x_t, \qquad c_t = W_c x_t, \tag{2}$$

$$\ell_t = -\operatorname{softplus}(w_\alpha^\top x_t + \beta_\alpha), \qquad \theta_t = W_\theta x_t. \tag{3}$$

Here $W_b, W_c \in \mathbb{R}^{r\times D}$ and $W_\theta \in \mathbb{R}^{(r/2)\times D}$. Vectors $c_i$ and $b_j$ provide the query and key sides of the correction. The scalar $\ell_t$ controls amplitude uniformly across auxiliary coordinates, whereas $\theta_t$ provides one phase increment per coordinate pair. For valid tokens, $\alpha_t = \exp(\ell_t) \in (0, 1)$; at padded positions, auxiliary vectors and control increments are zero, giving $\alpha_t = 1$.

### 3.2 Amplitude Modulation and Phase Alignment

AM and PA form reusable prefix states from the token-local controls:

$$g_t = \sum_{u=1}^{t} \ell_u, \qquad \Phi_t = \sum_{u=1}^{t} \theta_u. \tag{4}$$

Only controls accumulate: $b_t$, $c_t$, and values remain local. The scan follows packed-token order without resetting between feature groups; this order need not follow behavioral time. Prefix differences condition endpoint interactions on intervening controls without pooling token vectors or values.

**Amplitude Modulation (AM).** With query reference $g_i$ and inactive clipping, AM scales auxiliary interactions by $\exp(g_i - g_j)$ (Figure 3, left); the backbone's native mask determines which pairs contribute.

**Phase Alignment (PA).** PA rotates consecutive coordinate pairs by cumulative phase. The block-diagonal rotation is

$$R(\Phi_t) = \operatorname{blkdiag}\big(R_2(\Phi_{t,1}), \ldots, R_2(\Phi_{t,r/2})\big), \qquad R_2(\phi) = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}.$$

Since $R(\Phi_i)^\top R(\Phi_j) = R(\Phi_j - \Phi_i)$, rotating both endpoints yields their relative rotation (Figure 3, right). Rotation preserves norms while changing orientation and hence the signed dot product.

Together, AM and PA give the unclipped query-reference correction

$$\Delta_{ij} = \lambda_h \exp(g_i - g_j) c_i^\top R(\Phi_j - \Phi_i) b_j, \tag{5}$$

where $\lambda_h > 0$ is a learned head scale. For $i \geq j$, the amplitude is $\prod_{u=j+1}^{i} \alpha_u$ and the relative phase is $-\sum_{u=j+1}^{i} \theta_u$; the shared prefix cancels, leaving interval controls. For $i = j$, the amplitude is one and the relative phase is zero. Because the correction is signed, reducing its magnitude need not reduce the final attention weight.

Some sites use a sequence-end decay reference. Appendix A gives this variant and the exact score with endpoint clipping; Appendix B.2 gives the softmax selection identities.

### 3.3 THEORETICAL ANALYSIS

We analyze the auxiliary operator and its effect on selection with local projections, native scores, and the positive head scale held fixed.

**Proposition 1 (Amplitude and orientation of the pair map).** *For query-reference scoring with inactive clipping, define*

$$\eta_{ij} = \exp(g_i - g_j), \qquad T_{ij} = \eta_{ij} R(\Phi_j - \Phi_i). \tag{6}$$

*Then* $T_{ij}^\top T_{ij} = \eta_{ij}^2 I$ *and* $\|T_{ij}\|_2 = \eta_{ij}$. *For* $j \le i$, $0 < \eta_{ij} \le 1$, *and*

$$|\Delta_{ij}| \le \lambda_h \eta_{ij} \|c_i\|_2 \|b_j\|_2, \qquad T_{ik} = T_{ij} T_{jk} \quad (k \le j \le i). \tag{7}$$

*Moreover, equality of two positive-scale rotation maps* $\eta R$ *and* $\eta' R'$ *implies* $\eta = \eta'$ *and* $R = R'$. *Thus, amplitude and rotation have distinct roles at the induced-map level.*

Proposition 1 separates AM's scaling role from PA's orientation. At a softmax site, the correction range also controls how much the attention distribution and its aggregation of fixed values can change.

**Proposition 2 (Bounded redistribution at softmax sites).** *For a fixed query* $i$, *let* $p_j = \mathrm{softmax}(s)_j$ *and* $\tilde{p}_j = \mathrm{softmax}(s + \Delta)_j$ *on the same finite, nonempty allowed key set, with finite logits. Define the correction range*

$$\Omega_i = \max_{j \in \mathcal{A}_i} \Delta_{ij} - \min_{j \in \mathcal{A}_i} \Delta_{ij}. \tag{8}$$

*Before dropout, the total variation distance obeys the sharp bound*

$$\mathrm{TV}(p, \tilde{p}) = \frac{1}{2} \sum_j |\tilde{p}_j - p_j| \le \tanh(\Omega_i / 4). \tag{9}$$

*For fixed values satisfying* $\|v_j\|_2 \le B_v$, *the aggregation change satisfies*

$$\left\| \sum_j (\tilde{p}_j - p_j) v_j \right\|_2 \le 2 B_v \tanh(\Omega_i / 4). \tag{10}$$

*These bounds apply to any reference mode and clipping choice producing finite corrections.*

A native logit margin exceeding $\Omega_i$ preserves the corresponding key ordering within the attention row. Appendix B provides proofs, control sensitivity, and the treatment of clipping and reference modes.

### 3.4 BACKBONE INTEGRATION AND COMPUTATIONAL COST

RISE augments existing attention sites while retaining their native masks, dropout, and value pathway. For target attention, the candidate follows the history in the packed input. Its $c$ vector interacts with the histories' $b$ vectors, and its own control increments participate in the scan. Appendix A.2 documents site-specific layouts and references.

In HiFormer, RISE augments target-attention pooling over history; attention in the interaction layers retains its native scores. In Zenith, RISE augments target-attention pooling and native attention in every interaction layer. In HyFormer, it augments both history self-attention and global-query-to-history attention in every layer. In RankMixer, only target-attention pooling is augmented; the token-mixing blocks retain their original computation. In OneTrans, RISE augments the mixed attention over sequential and non-sequential tokens in every backbone layer, preserving the native pyramid reduction. Each insertion site has independent trainable RISE parameters; within a site, each head's auxiliary projections are shared across token positions. Appendix A.2 specifies layer coverage and the OneTrans interface.

Auxiliary pairwise scoring costs $O(HN_qN_kr)$ for $H$ heads, $N_q$ queries, and $N_k$ keys. Compatible dot-product sites can incorporate it through augmented Q/K channels. Prefix reuse avoids repeated interval summation; dense self-attention retains its quadratic pairwise term. Full construction costs appear in Appendix A.3; matched latency and memory measurements remain outside the present experiments.

Table 1: Test macro ROC-AUC, averaged over action labels within each backbone–dataset cell. Blue shading identifies RISE, bold marks the higher AUC, and Δ denotes RISE minus baseline. The "Mean" row averages the five backbones equally, whereas RISE entries are reported as means over three runs with distinct random seeds.

| | | ROC-AUC ↑ | | | |
|---|---|---|---|---|---|
| **Backbone** | **Variant** | **QK-Video** | **KuaiRand** | **TAAC-25** | **MerRec** |
| **HiFormer** (Gui et al., 2023) | Baseline | 0.933163 | 0.858482 | 0.855235 | 0.801097 |
| | + RISE | **0.934270** | **0.864227** | **0.858847** | **0.806951** |
| | Δ AUC | +0.001106 | +0.005746 | +0.003612 | +0.005854 |
| **HyFormer** (Huang et al., 2026) | Baseline | 0.930243 | 0.857190 | 0.851237 | 0.793737 |
| | + RISE | **0.931238** | **0.862748** | **0.868615** | **0.804529** |
| | Δ AUC | +0.000995 | +0.005557 | +0.017378 | +0.010791 |
| **RankMixer** (Zhu et al., 2025) | Baseline | 0.928907 | 0.852376 | 0.844863 | 0.788629 |
| | + RISE | **0.929470** | **0.862167** | **0.851183** | **0.798453** |
| | Δ AUC | +0.000562 | +0.009791 | +0.006320 | +0.009824 |
| **Zenith** (Zhang et al., 2026a) | Baseline | 0.934919 | 0.859663 | 0.839489 | 0.806044 |
| | + RISE | **0.935301** | **0.863164** | **0.856290** | **0.808353** |
| | Δ AUC | +0.000383 | +0.003502 | +0.016802 | +0.002310 |
| **OneTrans** (Zhang et al., 2026b) | Baseline | 0.933618 | 0.857342 | 0.855624 | 0.808730 |
| | + RISE | **0.934123** | **0.863997** | **0.858627** | **0.815654** |
| | Δ AUC | +0.000505 | +0.006655 | +0.003003 | +0.006924 |
| **Mean** | Baseline | 0.932170 | 0.857010 | 0.849290 | 0.799647 |
| | + RISE | **0.932880** | **0.863261** | **0.858712** | **0.806788** |
| | Δ AUC | +0.000710 | +0.006250 | +0.009423 | +0.007141 |

# 4 EXPERIMENTS

We evaluate whether augmenting compatibility scores with the cumulative importance channel improves feedback prediction. The primary comparison evaluates RISE at the specified attention sites against each backbone's local baseline.

## 4.1 EXPERIMENTAL SETUP

**Datasets and tasks.** We evaluate on four multi-feedback datasets. QK-Video, the QK-video subset of Tenrec (Yuan et al., 2022), contains click, follow, like, and share feedback on videos. KuaiRand (Gao et al., 2022) provides multiple video-feedback signals, with randomized video exposures included in the source dataset. TAAC-25 uses the processed TencentGR-10M dataset from the Tencent Advertising Algorithm Challenge 2025 (Pan et al., 2026), with click and conversion targets. MerRec (Li et al., 2025) records consumer-to-consumer shopping feedback, ranging from likes to purchases. Each dataset defines a multi-task binary feedback-prediction problem. Each backbone uses the same dataset-specific label set, with histories limited to 100 records excluding the target item. Dataset statistics, label definitions, preprocessing, and split protocols are detailed in Appendix C.1.

**Backbones and comparison.** We evaluate five backbones, each in its native form and with RISE, giving 20 model–dataset cells. HiFormer (Gui et al., 2023) models heterogeneous feature interactions using Transformers. HyFormer (Huang et al., 2026) combines sequence modeling and feature interaction for click-through rate prediction. RankMixer (Zhu et al., 2025) uses token-mixing blocks to model feature interactions for ranking. Zenith (Zhang et al., 2026a) is a scalable ranking architecture developed for livestreaming recommendation. OneTrans (Zhang et al., 2026b) unifies sequential and non-sequential feature modeling in one Transformer. Section 3.4 summarizes how RISE is ap-

plied to each backbone; Appendix A.2 specifies the token interfaces, layer coverage, and parameter sharing. The default auxiliary dimension is 16 per head, and the initial head weight is 0.1.

**Training.** The five backbones use feature embeddings of dimension 16, three backbone layers, a token dimension of 256, and task towers with hidden widths [128,64] and ReLU activations. Dense parameters use AdamW (Loshchilov & Hutter, 2019); sparse parameters use Adagrad (Duchi et al., 2011). The loss is the equally weighted sum of the task-specific binary cross-entropies, each first averaged over the examples in a batch.

**Evaluation and aggregation.** We evaluate **ROC-AUC** for each action label, measuring how well positive examples rank above negative examples without fixing a classification threshold. For backbone–dataset cell $c$, its label set $\mathcal{L}_c$, and variant $v$ (base or RISE), we define

$$\bar{A}_c^{(v)} = \frac{1}{|\mathcal{L}_c|} \sum_{\ell \in \mathcal{L}_c} A_{c\ell}^{(v)}, \qquad \Delta_{\text{cell}} = \frac{1}{20} \sum_{c=1}^{20} \left( \bar{A}_c^{(\text{RISE})} - \bar{A}_c^{(\text{base})} \right). \tag{11}$$

Here, $A_{c\ell}^{(v)}$ is the test ROC-AUC for one label. Each label receives equal weight within its cell, and each cell receives equal weight in the overall comparison. This two-level averaging prevents datasets with more feedback labels from receiving greater overall weight. Comparing each RISE-augmented model with its corresponding baseline measures the gain from adding RISE within that architecture. All AUC gains are reported as absolute differences.

### 4.2 Main Results

**Overall performance.** Table 1 shows higher label-averaged AUC in **all 20 model–dataset cells**. Weighting the 20 cells equally raises macro AUC from $0.859529$ to $0.865410$, an absolute gain of **0.005881** (**0.5881 percentage points**). The median cell gain is $0.005652$. The reported RISE values are means over three runs with distinct random seeds.

**Consistency across backbones.** RISE improves label-averaged AUC on all four datasets for each of the five backbones. The gains cover token-mixing architectures such as RankMixer and models that jointly encode sequential and non-sequential features, such as OneTrans. RISE acts through their existing attention scores, allowing cumulative controls to complement the representations supplied by each backbone while retaining its native aggregation pathway. This pattern supports the applicability of the additional scoring channel across the evaluated architectures. Section 5 further examines the role of this channel through dedicated mechanism analyses.

**Variation across datasets and backbones.** **TAAC-25 has the largest average gain** across backbones ($+0.009423$), followed by MerRec ($+0.007141$), KuaiRand ($+0.006250$), and QK-Video ($+0.000710$). QK-Video has the smallest gain for every backbone. Averaging over the four datasets, HyFormer has the largest gain relative to its own baseline ($+0.008681$). RISE generates corrections from current representations at each insertion site, so the resulting reweighting depends on the backbone and prediction task. This dependence offers a plausible explanation for the unequal gains across model–dataset combinations.

### 4.3 Component Ablation on HyFormer–MerRec

**Design and comparison.** Motivated by the scale and orientation roles in Section 3.3, we evaluate four configurations in a $2 \times 2$ component design: content only (neither AM nor PA), AM only (decay), PA only (rotation), and full RISE (both AM and PA). All four retain the auxiliary content projections, per-head scaling, and the same insertion sites. Thus, the content-only configuration retains the auxiliary content-score channel and is distinct from the backbone baseline.

**Contribution of AM and PA.** Both the AM-only and PA-only variants improve label-macro AUC over the content-only channel (Figure 4). Full RISE achieves the highest macro AUC, **0.8045287**. Removing PA from the full configuration lowers macro AUC by $0.0086487$; removing AM lowers it by $0.0092985$. Because the content-only configuration retains the auxiliary content-score channel, this comparison assesses the additional contribution of AM and PA within the same integration

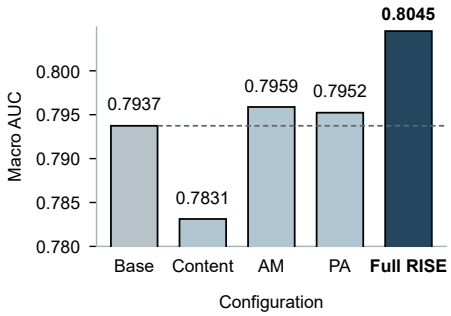



Figure 4: HyFormer–MerRec component ablation. Bars show absolute macro AUC; the dashed line marks the original HyFormer baseline. The axis starts at 0.780.

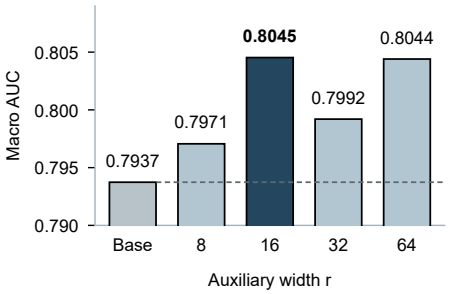



Figure 5: HyFormer–MerRec auxiliary width. Bars show absolute macro AUC; the dashed line marks the original HyFormer baseline. The axis starts at 0.790.

design. This result supports using AM and PA together in the evaluated HyFormer–MerRec configuration.

### 4.4 Sensitivity to Auxiliary Dimension

We vary the per-head auxiliary width $r \in \{8, 16, 32, 64\}$ on HyFormer–MerRec, retaining both AM and PA at all six insertion sites. Figure 5 shows absolute test AUC averaged equally over the five labels, including the original HyFormer baseline from Table 1 as the first bar.

**Observed performance.** All four auxiliary-width AUCs are above the original HyFormer baseline reference of $0.7937372$. Among the evaluated widths, $r = 16$ yields the highest observed macro AUC (0.8045), closely followed by $r = 64$ (0.8044).; the $r = 8, 32$ settings have lower AUCs. Appendix D describes the comparison protocol.

## 5 Discussion

We examine how RISE changes history pooling and which aspects of its input conditioning support prediction. The analyses use fixed RankMixer–KuaiRand and full-site HiFormer–TAAC-25 checkpoints. The HiFormer checkpoint differs from the history-aggregation variant used in the main benchmark. For each model, we change only the correction at the inspected target-pooling site. All model parameters and the inspected site's native Q/K/V remain fixed. We evaluate click binary cross-entropy (BCE) on the same retained interactions.

### 5.1 History Order as a Prediction Cue

At the two pooling sites examined here, keys and values come directly from per-event embeddings, without upstream sequence or positional encoding. Native content matching scores each history record against the candidate. Jointly permuting records, attributes and masks moves their native scores and values together, leaving the weighted sum unchanged in exact arithmetic. A check on 64 real inputs per model agrees with this property: the maximum FP32 pooling difference is below $4 \times 10^{-8}$. RISE's cumulative controls make each correction depend on the arrangement of the records. Moving a record changes which preceding records contribute to its control state. Its correction can therefore change even when its local embedding remains unchanged. RISE thus adds an order-sensitive scoring path over the existing history inputs.

The on/off comparison in Table 2 tests whether this additional scoring path supports prediction. Activating the original correction lowers click BCE from $0.517533$ to $0.510379$ for RankMixer. For HiFormer, the corresponding change is from $0.047422$ to $0.045461$. With the remaining parameters fixed, both trained predictors benefit from the correction at the inspected site.

Table 2: Click BCE with the inspected RISE correction disabled or unchanged (lower is better). RankMixer and HiFormer are evaluated on KuaiRand and TAAC-25, respectively. All model parameters and input records are fixed. HiFormer uses the full-site checkpoint.

| Model | Correction off | Original RISE |
|---|---|---|
| RankMixer | 0.517533 | **0.510379** |
| HiFormer | 0.047422 | **0.045461** |

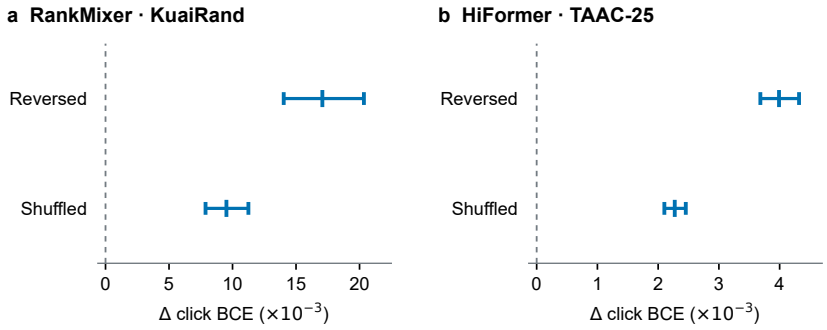



Figure 6: History-order effects with native Q/K/V fixed. Ticks show paired mean click-BCE changes (reordered − original); whiskers are 95% user-cluster intervals conditional on each fixed checkpoint and sample. HiFormer uses the full-site checkpoint. Axes have different ranges; sample details are in Appendix E.

To test whether the correction uses history order, we recompute it after reversing or shuffling valid history records. We map each recomputed score back to its associated record's original location. Replaying these scores with the original keys and values preserves the native aggregation inputs. The intervention changes the order used to generate the correction while holding the native content path fixed. Both perturbations increase click BCE in both models, with larger increases under reversal (Figure 6). All four conditional user-cluster intervals lie above zero. The data-provided order therefore yields lower click loss than either tested rearrangement. Exact differences and interval endpoints are reported in Appendix E.2, Table 6.

Native scores measure candidate–record content matching, while the cumulative correction also reflects how those records are arranged. This use of order connects RISE to CoPE and Forgetting Transformer, which incorporate ordering into attention scores (Golovneva et al., 2024; Lin et al., 2025). Our fixed-checkpoint results show how score-level order dependence can improve prediction at these otherwise permutation-invariant sites, using the same history inputs.

## 6 Conclusion

We introduced RISE (Ranking with Input-dependent State Enhancement), a plug-in score-level channel that makes contextual importance explicit in unified ranking. RISE combines Amplitude Modulation (AM), which uses cumulative decay controls to regulate auxiliary interaction strength, with Phase Alignment (PA), which uses cumulative phase controls to adjust the relative orientation of low-dimensional query and key projections. Their signed correction is added to the native attention score, while each backbone retains its original weighting rule and value aggregation. Across five backbones and four datasets, RISE improves label-averaged ROC-AUC in all 20 model–dataset settings, raising the equally weighted macro AUC from 0.859529 to 0.865410, an absolute gain of 0.005881. On HyFormer–MerRec, the full AM+PA configuration outperforms the auxiliary content-only channel and either single-component variant. Together, these results suggest that contextual importance can be modeled as an explicit, compact score pathway alongside learned representations, allowing the evaluated ranking backbones to adapt which historical evidence is emphasized while retaining their native value aggregation.

### AI USE STATEMENT

We used generative AI tools to assist with manuscript drafting and language editing. The authors developed the research methods and implemented and conducted the experiments. The authors take responsibility for the final manuscript, including its scientific claims, experimental results, and AI-assisted text.

### IMPACT STATEMENT

This paper advances machine learning research in recommendation by improving how ranking models use behavioral context to predict user feedback. By explicitly modeling the contextual importance of historical interactions, RISE aims to support more relevant recommendations in applications such as online shopping and content discovery, potentially helping users find items that better match their interests. Its plug-in design allows the approach to be incorporated into existing ranking architectures and supports further research on context-aware recommendation. As with other recommendation methods, its deployment may affect content exposure and reinforce biases present in historical interaction data. The practical impact of these advances depends on the application setting and the objectives of the deployed system. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

## REFERENCES

Jianxin Chang, Chenbin Zhang, Zhiyi Fu, Xiaoxue Zang, Lin Guan, Jing Lu, Yiqun Hui, Dewei Leng, Yanan Niu, Yang Song, and Kun Gai. TWIN: TWo-stage interest network for lifelong user behavior modeling in CTR prediction at kuaishou. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3785–3794. Association for Computing Machinery, 2023. doi: 10.1145/3580305.3599922. URL https://doi.org/10.1145/3580305.3599922.

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

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Ningya Feng, Junwei Pan, Jialong Wu, Baixu Chen, Ximei Wang, QianLi, Xian Hu, Jie Jiang, and Mingsheng Long. Long-sequence recommendation models need decoupled embeddings. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 20123–20144, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/332bcd9f74dfdbb83352efb0df6f8164-Paper-Conference.pdf.

Yufei Feng, Fuyu Lv, Weichen Shen, Menghan Wang, Fei Sun, Yu Zhu, and Keping Yang. Deep session interest network for click-through rate prediction. *arXiv preprint arXiv:1905.06482*, 2019.

Chongming Gao, Shijun Li, Yuan Zhang, Jiawei Chen, Biao Li, Wenqiang Lei, Peng Jiang, and Xiangnan He. Kuairand: An unbiased sequential recommendation dataset with randomly exposed videos. In *Proceedings of the 31st ACM international conference on information & knowledge management*, pp. 3953–3957, 2022.

Olga Golovneva, Tianlu Wang, Jason Weston, and Sainbayar Sukhbaatar. Contextual position encoding: Learning to count what's important. *arXiv preprint arXiv:2405.18719*, 2024.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Huan Gui, Ruoxi Wang, Ke Yin, Long Jin, Maciej Kula, Taibai Xu, Lichan Hong, and Ed H Chi. Hiformer: Heterogeneous feature interactions learning with transformers for recommender systems. *arXiv preprint arXiv:2311.05884*, 2023.

Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. Deepfm: a factorization-machine based neural network for ctr prediction. *arXiv preprint arXiv:1703.04247*, 2017.

Yunwen Huang, Shiyong Hong, Xijun Xiao, Jinqiu Jin, Xuanyuan Luo, Zhe Wang, Zheng Chai, Shikang Wu, Yuchao Zheng, and Jingjian Lin. Hyformer: Revisiting the roles of sequence modeling and feature interaction in ctr prediction. *arXiv preprint arXiv:2601.12681*, 2026.

Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*, pp. 197–206. IEEE, 2018.

Hye-young Kim, Minjin Choi, Sunkyung Lee, Ilwoong Baek, and Jongwuk Lee. Diff: Dual side-information filtering and fusion for sequential recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1624–1633. Association for Computing Machinery, 2025. doi: 10.1145/3726302.3729948. URL https://doi.org/10.1145/3726302.3729948.

Aakash Sunil Lahoti, Kevin Li, Berlin Chen, Caitlin Wang, Aviv Bick, Zico Kolter, Tri Dao, and Albert Gu. Mamba-3: Improved sequence modeling using state space principles. In *International Conference on Learning Representations*, volume 2026, pp. 86173–86202, 2026.

Jiacheng Li, Yujie Wang, and Julian McAuley. Time interval aware self-attention for sequential recommendation. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pp. 322–330. Association for Computing Machinery, 2020. doi: 10.1145/3336191.3371786. URL https://doi.org/10.1145/3336191.3371786.

Lichi Li, Zainul Abi Din, Zhen Tan, Sam London, Tianlong Chen, and Ajay Daptardar. Merrec: A large-scale multipurpose mercari dataset for consumer-to-consumer recommendation systems. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 2371–2382, 2025.

Xiaolin Lin, Jinwei Luo, Junwei Pan, Weike Pan, Zhong Ming, Xun Liu, Shudong Huang, and Jie Jiang. Multi-sequence attentive user representation learning for side-information integrated sequential recommendation. In *Proceedings of the 17th ACM international conference on web search and data mining*, pp. 414–423, 2024.

Zhixuan Lin, Evgenii Nikishin, Xu He, and Aaron Courville. Forgetting transformer: Softmax attention with a forget gate. In *International Conference on Learning Representations*, volume 2025, pp. 69704–69738, 2025.

Chang Liu, Xiaoguang Li, Guohao Cai, Zhenhua Dong, Hong Zhu, and Lifeng Shang. Noninvasive self-attention for side information fusion in sequential recommendation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 4249–4256, 2021.

Chengkai Liu, Jianghao Lin, Jianling Wang, Hanzhou Liu, and James Caverlee. Mamba4rec: Towards efficient sequential recommendation with selective state space models. *arXiv preprint arXiv:2403.03900*, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019.

Junwei Pan, Wei Xue, Chao Zhou, Xing Zhou, Lunan Fan, Yanbo Wang, Haoran Xin, Zhiyu Hu, Yaozheng Wang, Fengye Xu, et al. Tencent advertising algorithm challenge 2025: All-modality generative recommendation. *arXiv preprint arXiv:2604.04976*, 2026.

Qi Pi, Guorui Zhou, Yujing Zhang, Zhe Wang, Lejian Ren, Ying Fan, Xiaoqiang Zhu, and Kun Gai. Search-based user interest modeling with lifelong sequential behavior data for click-through rate prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 2685–2692. Association for Computing Machinery, 2020. doi: 10.1145/3340531.3412744. URL https://doi.org/10.1145/3340531.3412744.

Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021.

Ahmed Rashed, Shereen Elsayed, and Lars Schmidt-Thieme. Context and attribute-aware sequential recommendation via cross-attention. In *Proceedings of the 16th ACM Conference on Recommender Systems*, pp. 71–80. Association for Computing Machinery, 2022. doi: 10.1145/3523227.3546777. URL https://doi.org/10.1145/3523227.3546777.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 464–468, 2018.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023.

Jiakai Tang, Sunhao Dai, Kun Wang, Zhiluohan Guo, Yu Zhao, Cong Fu, Kangle Wu, Yabo Ni, Anxiang Zeng, Xu Chen, et al. Onerank: Unified transformer-native ranking architecture for multi-task recommendation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 8053–8064, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Ruoxi Wang, Rakesh Shivanna, Derek Cheng, Sagar Jain, Dong Lin, Lichan Hong, and Ed Chi. Dcn v2: Improved deep & cross network and practical lessons for web-scale learning to rank systems. In *Proceedings of the web conference 2021*, pp. 1785–1797, 2021.

Shuhan Wang, Bin Shen, Xu Min, Yong He, Xiaolu Zhang, Liang Zhang, Jun Zhou, and Linjian Mo. Aligned side information fusion method for sequential recommendation. In *Companion Proceedings of the ACM Web Conference 2024*, pp. 112–120. Association for Computing Machinery, 2024. doi: 10.1145/3589335.3648308. URL https://doi.org/10.1145/3589335.3648308.

Yueqi Xie, Peilin Zhou, and Sunghun Kim. Decoupled side information fusion for sequential recommendation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pp. 1611–1621, 2022.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. *arXiv preprint arXiv:2312.06635*, 2023.

Guanghu Yuan, Fajie Yuan, Yudong Li, Beibei Kong, Shujie Li, Lei Chen, Min Yang, Chenyun Yu, Bo Hu, Zang Li, et al. Tenrec: A large-scale multipurpose benchmark dataset for recommender systems. *Advances in Neural Information Processing Systems*, 35:11480–11493, 2022.

Jiaqi Zhai, Lucy Liao, Xing Liu, Yueming Wang, Rui Li, Xuan Cao, Leon Gao, Zhaojie Gong, Fangda Gu, Michael He, et al. Actions speak louder than words: Trillion-parameter sequential transducers for generative recommendations. *arXiv preprint arXiv:2402.17152*, 2024.

Ruifeng Zhang, Zexi Huang, Zikai Wang, Ke Sun, Bohang Zheng, Yuchen Jiang, Zhe Chen, Zhen Ouyang, Huimin Xie, Phil Shen, et al. Zenith: Scaling up ranking models for billion-scale livestreaming recommendation. *arXiv preprint arXiv:2601.21285*, 2026a.

Tingting Zhang, Pengpeng Zhao, Yanchi Liu, Victor S. Sheng, Jiajie Xu, Deqing Wang, Guanfeng Liu, and Xiaofang Zhou. Feature-level deeper self-attention network for sequential recommendation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 4320–4326. International Joint Conferences on Artificial Intelligence Organization, 2019. doi: 10.24963/ijcai.2019/600. URL https://www.ijcai.org/proceedings/2019/600.

Zhaoqi Zhang, Haolei Pei, Jun Guo, Tianyu Wang, Yufei Feng, Hui Sun, Shaowei Liu, and Aixin Sun. Onetrans: Unified feature interaction and sequence modeling with one transformer in industrial recommender. In *Proceedings of the ACM Web Conference 2026*, pp. 8162–8170, 2026b.

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

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

Haolin Zhou, Junwei Pan, Xinyi Zhou, Xihua Chen, Jie Jiang, Xiaofeng Gao, and Guihai Chen. Temporal interest network for user response prediction. In *Companion Proceedings of the ACM Web Conference 2024*, pp. 413–422. Association for Computing Machinery, 2024. doi: 10.1145/3589335.3648340. URL https://doi.org/10.1145/3589335.3648340.

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

# A METHOD IMPLEMENTATION AND BACKBONE INTEGRATION

## A.1 NUMERICAL COMPUTATION AND STABILIZATION

The main-text projections are written for valid tokens. At padded positions, auxiliary vectors and both control increments are zeroed before the prefix scan. The scalar gate at a valid position satisfies $\alpha_t = \exp(\ell_t) = \mathrm{sigmoid}(-(w_\alpha^\top x_t + \beta_\alpha))$. The cumulative states parameterize the amplitude and phase transformations of auxiliary scoring vectors.

In PA, the matrix $R(\Phi_t) \in \mathbb{R}^{r\times r}$ is block diagonal, with each consecutive coordinate pair rotated by its corresponding phase:

$$R(\phi) = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}. \tag{12}$$

This gives $R(\Phi_i)^\top R(\Phi_j) = R(\Phi_j - \Phi_i)$, with blockwise phase differences. For $i \geq j$, the inclusive prefix definitions imply

$$g_i - g_j = \sum_{u=j+1}^{i} \ell_u, \qquad \Phi_j - \Phi_i = -\sum_{u=j+1}^{i} \theta_u. \tag{13}$$

For $i > j$, the remaining increments include the query position; for $i = j$, both differences are zero. For allowed pairs with $i < j$, the query-reference amplitude can exceed one. PA preserves vector norms, while AM controls the magnitude of the auxiliary score and does not impose a universal rule for suppressing past behavior.

AM supports query reference $\gamma_i = g_i$ and sequence-end reference $\gamma_i = g_{\mathrm{end}}$. Let $\mathcal{V}$ be the nonempty set of valid positions. Because log-decay increments are non-positive, $g_{\mathrm{end}} = \min_{t\in\mathcal{V}} g_t$ is the state at the last valid position. A sequence-end reference replaces $g_i - g_j$ by $g_{\mathrm{end}} - g_j$ in the unclipped amplitude; the phase difference remains $\Phi_j - \Phi_i$.

For numerical stability, we center valid cumulative values and clip each endpoint separately:

$$\mu = \frac{\min_{t\in\mathcal{V}} g_t + \max_{t\in\mathcal{V}} g_t}{2}, \qquad f(z) = \mathrm{clip}(z - \mu, -\kappa, \kappa), \tag{14}$$

where $\kappa > 0$. The transformed auxiliary vectors are

$$\bar{c}_i = \exp(f(\gamma_i))R(\Phi_i)c_i, \qquad \bar{b}_j = \exp(-f(g_j))R(\Phi_j)b_j. \tag{15}$$

The learned positive head scale is $\lambda_h = \mathrm{softplus}(\omega_h) + \varepsilon$, with $\varepsilon = 10^{-8}$. The exact correction under either reference is

$$\Delta_{ij} = \lambda_h \bar{c}_i^\top \bar{b}_j = \lambda_h \exp(f(\gamma_i) - f(g_j))\, c_i^\top R(\Phi_j - \Phi_i) b_j. \tag{16}$$

Inactive clipping and query reference recover (5), because the common centering term cancels. With clipping, the two endpoint terms must remain separate; their difference is not a clipped pairwise difference.

## A.2 BACKBONE INTEGRATION AND PARAMETER SHARING

**Attention-site interfaces.** Table 3 summarizes the input layouts, query/key roles, decay references, and weighting rules at the RISE insertion sites. The evaluated implementations use these interfaces, with HiFormer restricted to target-attention pooling.

**Layer coverage and parameter sharing.** RISE is enabled at every instance of the attention sites listed in Table 3. HiFormer has one RISE module at target-attention pooling over history, with no RISE correction in its interaction layers. With the three-layer configuration, Zenith has four RISE modules: one at target-attention pooling before the interaction stack and one in each interaction layer. HyFormer has six modules, covering history self-attention and query-to-history decoding in all three layers. RankMixer has one module at target-attention pooling and none in its token-mixing blocks. OneTrans has one module in each mixed-attention layer. Each module owns its auxiliary content projections, decay and phase projections, and head scales; no trainable RISE parameters are tied across layers or attention sites. Within a module, parameters are head-specific and shared across token positions. Cumulative controls are recomputed from that site's current input, rather than passed between layers.

Table 3: RISE attention-site interfaces for all five backbones. "All" refers to positions in the current packed input, subject to native pairwise masks and before OneTrans output selection. Token representations include any normalization already applied at the native site.

| Attention site | Packed input order | Queries / keys | Decay reference | Weight rule |
|---|---|---|---|---|
| HiFormer, RankMixer, Zenith: target pooling | History, then candidate | Candidate / history | Query | Configured; default softmax |
| HyFormer: history self-attention | Normalized history tokens | History / history | Sequence end | Softmax |
| HyFormer: query decoding | Normalized history, then normalized global queries | Global queries / history | Query | Softmax |
| Zenith: token attention | User ID, item ID, user-context chunks, item-attribute chunks | All / all | Sequence end | Softmax |
| OneTrans: mixed attention | Current sequential tokens, then non-sequential tokens | All / all | Query | Configured; default softmax |

**OneTrans interface.** At each layer, OneTrans packs the current normalized sequential tokens (S), followed by normalized non-sequential tokens (NS) encoding user/context and candidate attributes. RISE scans this combined sequence with query-referenced decay and adds its correction to the native mixed-attention scores. The native causal pairwise mask, S-shared and NS-specific Q/K/V projections, and value aggregation remain in place. Attention is computed over the full current packed input; the native pyramid rule then retains the selected suffix of S outputs and all NS outputs. The original feed-forward blocks and prediction from the final NS tokens are preserved.

Padding is excluded from the auxiliary vectors and cumulative controls. At HyFormer and OneTrans sites without a native pairwise padding mask, this exclusion does not remove padded keys from softmax normalization. OneTrans's causal pairwise mask also does not, by itself, establish prefix-causal auxiliary scoring; Appendix A.3 discusses the role of centering and clipping.

HyFormer query decoding concatenates all global queries after the history. Consequently, a later query's cumulative state includes control increments from earlier packed queries. This layout defines an ordering for score modulation and does not assign behavioral timestamps to the global queries.

### A.3 Dot-product integration, cost, and input availability

At compatible sites using native scale $1/\sqrt{d_h}$, concatenating scaled auxiliary vectors to the original queries and keys incorporates the correction into the same scoring operation. Appending $d_h^{1/4}\sqrt{\lambda_h}\bar{c}_i$ and $d_h^{1/4}\sqrt{\lambda_h}\bar{b}_j$ reproduces (1) when the original scale is retained. This widens the dot-product channels to $d_h + r$ and preserves the original value pathway without a separate value-aggregation branch.

For $H$ heads, local projections cost $O(LDHr)$. The parallel prefix computation uses $O(HLr \log L)$ arithmetic work, followed by $O(HLr)$ rotations and scaling. Auxiliary pairwise scores cost $O(HN_qN_kr)$, whether formed separately or through augmented Q/K. Actual latency and memory depend on the attention implementation and token layout.

Native masks determine allowed pairwise value connections, whereas the supplied tokens determine the controls' information dependencies. Sequence-end decay reads the supplied endpoint, and centering uses all valid tokens. Centering cancels from the unclipped product, but may affect scores when clipping is active. A causal pairwise mask alone therefore does not establish prefix-causal scoring; token availability, reference mode, and clipping must be considered together.

# B Theoretical Properties and Proofs

Propositions 1 and 2 are stated in Section 3.3. Here we give their proofs and analyze control sensitivity, clipping, and softmax selection with local projections, native scores, and the positive head scale held fixed. This separates the algebraic role of each control from changes introduced by jointly training the backbone.

## B.1 Properties of amplitude and phase transformations

*Proof of Proposition 1.* Each two-dimensional rotation block is orthogonal, so $R^\top R = I$. Multiplication gives $T_{ij}^\top T_{ij} = \eta_{ij}^2 I$, establishing the norm identity. Since $g_i - g_j = \sum_{u=j+1}^{i} \ell_u \le 0$ for $j \le i$, the amplitude lies in $(0, 1]$. Cauchy–Schwarz bounds $\lambda_h c_i^\top T_{ij} b_j$. For composition, scalar exponents telescope and block rotations add their angles:

$$T_{ij}T_{jk} = e^{g_i - g_k} R\big((\Phi_j - \Phi_i) + (\Phi_k - \Phi_j)\big) = T_{ik}. \tag{17}$$

Finally, $\eta R = \eta' R'$ implies $\eta^2 I = (\eta')^2 I$. Positivity gives $\eta = \eta'$, and hence $R = R'$. □

AM therefore controls the singular values of the auxiliary map, whereas PA controls its orientation. Equality of the associated bilinear scores for every pair of local vectors implies equality of the maps. This statement concerns the operator: a single scalar score need not identify either control, angles are defined modulo $2\pi$, and trainable projections permit additional parameterization freedom. The operator-norm bound permits a particular scalar interaction to increase as alignment changes.

**Clipping and reference modes.** For the implemented correction, replace $\eta_{ij}$ by $\exp(f(\gamma_i) - f(g_j))$. Because $f$ is monotone, this amplitude is at most one for a past key with query reference, and for every valid key with sequence-end reference. It is bounded by $e^{2\kappa}$ in general. Query-reference maps still compose within the same fixed input, since the endpoint differences of $f$ telescope. Sequence-end-reference maps generally do not satisfy this composition identity.

**Sensitivity to the two controls.** Let $\psi = \Phi_j - \Phi_i$ and consider two positive amplitudes $\eta, \eta'$ and phase vectors $\psi, \psi'$ with the same $c_i, b_j, \lambda_h$. Then

$$\begin{aligned} |\Delta'_{ij} - \Delta_{ij}| &\le \lambda_h \|c_i\|_2 \|b_j\|_2 \big(|\eta' - \eta| + \min(\eta, \eta') d_R\big), \\ d_R &= 2 \max_m \left| \sin \frac{\psi'_m - \psi_m}{2} \right|. \end{aligned} \tag{18}$$

To obtain the bound, decompose $\eta' R(\psi') - \eta R(\psi)$ into a scale difference and a rotation difference, using the smaller amplitude for the latter. For block rotations, $\|R(\psi') - R(\psi)\|_2 = d_R$. The bound applies to fixed-vector control sensitivity; an input perturbation can also change the local projections.

## B.2 Softmax reweighting and redistribution bounds

Let $\mathcal{A}_i$ be the nonempty allowed key set for query $i$. Before attention dropout, the softmax weights and aggregation of the original values satisfy

$$a_{ij} = \frac{\exp(S_{ij}^{\text{base}} + \Delta_{ij})}{\sum_{k \in \mathcal{A}_i} \exp(S_{ik}^{\text{base}} + \Delta_{ik})}, \quad j \in \mathcal{A}_i, \qquad y_i = \sum_{j \in \mathcal{A}_i} a_{ij} v_j, \tag{19}$$

where $v_j$ is the native value vector. Let $a_{ij}^{(0)}$ denote the attention weight obtained by removing $\Delta$ while holding current representations, native scores, and allowed keys fixed. This within-model comparison does not describe a separately trained baseline. Substitution gives

$$a_{ij} = \frac{a_{ij}^{(0)} \exp(\Delta_{ij})}{\sum_{k \in \mathcal{A}_i} a_{ik}^{(0)} \exp(\Delta_{ik})}. \tag{20}$$

For any two allowed keys $j, k$ with finite logits,

$$\log \frac{a_{ij}}{a_{ik}} = \log \frac{a_{ij}^{(0)}}{a_{ik}^{(0)}} + \Delta_{ij} - \Delta_{ik}. \tag{21}$$

Thus, $\Delta_{ij} - \Delta_{ik}$ adjusts the relative preference for key $j$ over key $k$. A row-constant correction leaves the softmax distribution unchanged. These identities do not apply to elementwise score activations, which also occur in recommendation models such as HSTU (Zhai et al., 2024). RISE retains the weighting rule of each insertion site.

*Proof of Proposition 2.* Let $m = \min_j \Delta_{ij}$, $M = \max_j \Delta_{ij}$, $A = e^m$, $B = e^M$, and let $X$ take value $e^{\Delta_{ij}}$ with probability $p_j$. Writing $Z = \mathbb{E}_p X$ gives $\tilde{p}_j = p_j e^{\Delta_{ij}}/Z$. If $A = B$, the correction is constant and $p = \tilde{p}$. Otherwise, convexity bounds $|X - Z|$ by the chord joining its values at $A$ and $B$. Taking expectations yields

$$\mathrm{TV}(p, \tilde{p}) = \frac{\mathbb{E}_p|X - Z|}{2Z} \le \frac{(B - Z)(Z - A)}{Z(B - A)}. \tag{22}$$

The right-hand side is maximized at $Z = \sqrt{AB}$, where it equals

$$\frac{\sqrt{B} - \sqrt{A}}{\sqrt{B} + \sqrt{A}} = \tanh\big((M - m)/4\big). \tag{23}$$

Two keys with $X \in \{A, B\}$ and probability $\sqrt{B}/(\sqrt{A} + \sqrt{B})$ on $A$ attain equality, proving sharpness. The aggregation bound follows from $\| \sum_j (\tilde{p}_j - p_j) v_j \|_2 \le B_v \sum_j |\tilde{p}_j - p_j|$. □

**Native-order preservation and scope.** For two allowed keys, the corrected logit gap is $s_j - s_k + \Delta_{ij} - \Delta_{ik}$. A native gap $s_j - s_k > \Omega_i$ therefore remains positive. Also, $\Omega_i \le 2\lambda_h \|c_i\|_2 \max_j \eta_{ij} \|b_j\|_2$, connecting the correction scale to the redistribution bound. The comparison holds within one model at fixed representations and values. It applies to pre-dropout softmax selection, independently of the recommendation AUC comparison between separately trained models.

# C Experimental Setup

This appendix describes the datasets, feedback labels, and split construction used in the experiments.

## C.1 Datasets, labels, and splits

Table 4 in this appendix reports the processed training, validation, and test split sizes. The label inventory and dataset-specific protocols are provided below.

Table 4: Processed interaction counts for the training, validation, and test splits. These counts describe the experimental datasets, not raw downloads or the number of unique users/items.

| Dataset | Training | Validation | Test | Labels |
|---|---|---|---|---|
| QK-Video | 394,647,318 | 50,079,027 | 48,579,958 | 4 |
| KuaiRand | 256,110,233 | 34,334,869 | 33,019,342 | 6 |
| TAAC-25 | 586,117,494 | 17,846,140 | 153,243,512 | 2 |
| MerRec | 144,108,235 | 16,515,570 | 11,681,154 | 5 |

Table 5: Action labels for the four datasets. Label capitalization is preserved.

| Dataset | Labels |
|---|---|
| QK-Video | click, follow, like, share |
| KuaiRand | is_click, is_follow, is_like, is_comment, is_forward, long_view |
| TAAC-25 | is_click, is_conversion |
| MerRec | Like, Cart, Offer, Checkout, Purchase |

Table 5 lists all 17 dataset–label tasks.

**QK-Video.** The configured subset originates from Tenrec (Yuan et al., 2022). Preprocessing divides each user's existing sequence order into training, validation, and test portions, targeting proportions of 0.8/0.1/0.1. Integer allocation for short sequences means that aggregate proportions need not equal these targets. This is an ordered per-user split; the archived description does not establish a common global calendar boundary.

**KuaiRand.** The processed data are derived from KuaiRand (Gao et al., 2022). Split metadata records April 7–May 2, 2022 for training, May 3–5 for validation, and May 6–8 for testing. The presence of randomized exposures in the source dataset does not by itself make the processed evaluation an unbiased estimate of a deployment policy's performance.

**TAAC-25.** The configuration uses the TencentGR-10M source associated with TAAC 2025 (Pan et al., 2026). Its preprocessing metadata records training boundaries of October 15, 2024–May 28, 2025, validation on May 29, and testing from May 30–August 5, using Asia/Shanghai. These are the boundary values stored by the preprocessing pipeline; their correspondence to the source data's undisclosed real collection dates has not been established. The split construction is constrained by positive-example counts near the terminal portions, retains users with at least 10 interactions, and filters user sequences without clicks or conversions. Consequently, the study concerns this filtered population. Validation and test intervals differ substantially in duration and label distribution.

**MerRec.** The study uses the October partition of MerRec (Li et al., 2025), with metadata boundaries of October 1–25, 2023 for training, October 26–28 for validation, and October 29–31 for testing. The first `item_view` for each user–item pair is an example anchor. Subsequent Like, Cart, Offer, Checkout, and Purchase actions are combined by logical OR into the anchor's binary labels. Non-view actions supply labels and are not retained as separate examples or history tokens.

# D Supplementary Ablation Settings

**AM/PA component ablation.** Figure 4 compares content-only, AM-only, PA-only, and full RISE on HyFormer–MerRec. Content-only retains native attention, auxiliary content projections, and per-head scaling, with AM and PA disabled. Disabling AM sets $\log \alpha = 0$; disabling PA sets the phase to zero. All variants share the three sequence self-attention and three query-to-sequence sites, architecture, objective, optimizer settings, and data protocol (Appendix C.1).

**Auxiliary-dimension sensitivity.** We vary the per-head auxiliary width $r \in \{8, 16, 32, 64\}$ at the same six sites. The baseline in Figure 5 is the original HyFormer model, distinct from content-only.

# E History-Order Prediction Interventions

## E.1 Checkpoints and fixed samples

The probes use fixed RankMixer–KuaiRand and full-site HiFormer–TAAC-25 checkpoints; the latter differs from the history-aggregation variant in the main benchmark. The fixed evaluation samples contain 131,072 eligible interactions from 107 users for RankMixer and 131,070 from 6,563 users for HiFormer. Eligibility requires at least two allowed history records. These samples were retained without selecting outcomes and are not a random sample of the full test population.

## E.2 History permutations and prediction effects

We reverse valid history records or apply one prespecified random permutation per interaction, jointly moving every event attribute and its masks while keeping candidate and padding positions fixed. Original order denotes the data-provided arrangement, without assuming verified chronology. We compute the RISE correction on the permuted records, inverse-align it to their original identities, and replay it on the original input. Native Q/K/V, mask, and scale at the inspected target-pooling site, all parameters, and other augmentation sites remain fixed. No amplitude or mean matching is applied.

A separate FP32 check jointly permutes 64 real inputs per model with the inspected correction disabled. Maximum absolute pooling differences are $3.73 \times 10^{-8}$ for RankMixer and $2.98 \times 10^{-8}$ for HiFormer, consistent with native permutation invariance up to rounding.

The primary endpoints are paired mean click BCE differences for reversed and shuffled corrections relative to the original order. Table 6 reports higher click loss for both perturbations in both models. Percentile 95% intervals use user-cluster bootstrap resampling, retaining each sampled user's interactions. All four intervals lie above zero; equal-user weighting retains the directions. Effects are not uniform across feedback labels. These intervals condition on the checkpoints, fixed windows, and two fixed perturbations. The interventions test the inspected correction's predictive use of order, without decomposing the total training gain.

Table 6: Click BCE changes after reordering the histories seen by RISE, with native Q/K/V fixed. Values are $10^4$ times BCE(route)$-$BCE(original), with conditional 95% user-cluster intervals; positive values mean higher loss. The probes use 131,072 and 131,070 interactions with at least two valid historical events, respectively. HiFormer uses the full-site checkpoint.

| **Model–dataset** | **Reverse − original** | **Random − original** |
|---|---|---|
| RankMixer–KuaiRand | $+170.732$ $[+140.300, +203.445]$ | $+95.194$ $[+78.745, +112.499]$ |
| HiFormer–TAAC-25 | $+39.898$ $[+36.828, +43.196]$ | $+22.733$ $[+21.023, +24.549]$ |