

Diagnosing and Mitigating Mid-Sequence Degradation in Recommender Systems

Linjiang Guo
University of Technology Sydney
Sydney, Australia

Nitin Bisht
University of Technology Sydney
Sydney, Australia

Shiqing Wu
City University of Macau
Macau, China

Huan Huo
University of Technology Sydney
Sydney, Australia

Xianzhi Wang
University of Technology Sydney
Sydney, Australia

Guandong Xu*
The Education University of Hong
Kong
Hong Kong, China

Abstract

Although attention-based sequential recommenders have achieved strong accuracy, they can exhibit a systematic failure mode, i.e., lost-in-the-middle, where attention disproportionately concentrates on sequence boundaries to informative middle behaviors. Through controlled perturbations, we find that this U-shaped bias persists even when collaborative signals are moved, indicating an architectural effect rather than a positional-encoding artifact. To mitigate this bias, we propose Disentangling and Calibrating Position Bias framework, which is a lightweight, plug-and-play calibration procedure that combines bias-adaptive masking during training with counterfactual logit calibration at inference to remove residual position-only effects. Experiments on diverse datasets demonstrate consistent improvements in ranking quality. Further ablations verify the contributions of both masking and calibration components.

CCS Concepts

• Information systems → Recommender systems.

Keywords

Attention mechanisms; Positional effects; Lost-in-the-middle; Sequential recommendation

ACM Reference Format:

Linjiang Guo, Nitin Bisht, Shiqing Wu, Huan Huo, Xianzhi Wang, and Guandong Xu. 2026. Diagnosing and Mitigating Mid-Sequence Degradation in Recommender Systems. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26)*, July 20–24, 2026, Melbourne, VIC, Australia. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3805712.3809738>

1 Introduction

Sequential recommendation aims to predict a user’s next interaction by modeling the evolution of user interests from historical behavior sequences [8]. Modern sequential recommenders use attention mechanisms to aggregate historical interactions, where attention

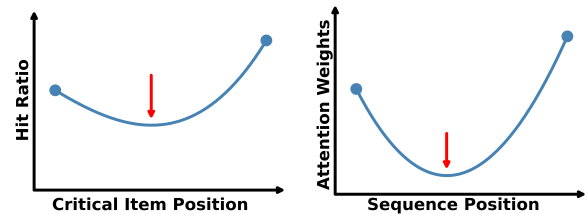


Figure 1: Left: Hit Ratio of prediction grouped by the position of the item with the strongest collaborative signal in a sequence. Right: Attention weights across sequence positions. Both exhibit a U-shaped distribution (on MovieLens).

weights determine how different past items contribute to the resulting user interest embedding and are often interpreted as indicators of user preference strength [2]. Consequently, a wide range of architectures have been adopted, utilizing diverse formulations such as self-attention with uni-directional [18, 26] or bi-directional dependency modeling [6, 38]. Despite these advances, the distribution of attention weights across different positions in a sequence, and whether they accurately reflect collaborative relevance, remains insufficiently understood in sequential recommendation.

A closely related phenomenon, commonly referred to as “lost-in-the-middle,” has been extensively studied in Large Language Models (LLMs) [11, 25]. Recent research reveals that these models exhibit an intrinsic positional bias, heavily prioritizing information at sequence boundaries while underutilizing content in the middle [17]. This observation raises a fundamental question: does this positional bias, where models struggle to retrieve information unless it appears at the start or end of a context window, also plague sequential recommendation systems? As illustrated in Figure 1, sequential models also suffer from a “lost-in-the-middle” structural bias. The prediction accuracy (Hit Ratio (HR)) (Left) drops significantly when the critical item (defined as the historical interaction with the highest similarity to the target) appears mid-sequence. This trend aligns with the scarcity of attention weights (Right) allocated to these middle positions (U-shaped attention). These results reveal a critical failure mode: despite their high relevance, mid-sequence interactions are systematically suppressed by the model’s architectural tendency, thereby undermining the quality of user embeddings.

*Corresponding author.



This work is licensed under a Creative Commons Attribution 4.0 International License. *SIGIR '26, Melbourne, VIC, Australia.*

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2599-9/2026/07

<https://doi.org/10.1145/3805712.3809738>

While LLMs and sequential recommenders share architectural similarities, these insights and solutions, such as repetitive information injection [25], modifying positional encodings [37], or padding with non-informative tokens [10], do not transfer straightforwardly due to fundamental data discrepancies. First, unlike LLMs, which process dense semantic tokens, sequential recommenders operate on sparse item IDs. Second, item sequences exhibit strict chronological dependencies, unlike many LLM long-context tasks, in which the input segments are not strictly chronological. Motivated by this observation, we perform controlled perturbations, including sequence swapping and synthetic signal injection, together with quantitative analysis to verify that the boundary-dominant, U-shaped attention pattern is a robust architectural artifact across datasets and model variants. Having empirically validated that this pattern persists under these perturbations and across datasets, we propose Disentangling and Calibrating Position Bias (DCPB), a model-agnostic framework that disentangles positional bias in attention from collaborative signals. Specifically, DCPB estimates an architecture-induced positional prior via a constrained projection that removes target- and context-aligned components, forming a structurally neutral probe whose attention profile reflects position bias. It then mitigates this bias through two complementary mechanisms. Firstly, we use a bias-adaptive masking during training that preferentially regularizes overreliance on boundaries to improve representation learning. Next, we use counterfactual calibration of attention logits at inference to subtract the estimated positional prior and restore attention to collaboratively relevant mid-sequence interactions, ensuring robust recommendations without disrupting temporal causality. Our contributions are:

- We identify a *lost-in-the-middle* failure mode in attention-based sequential recommendation, where both attention allocation and prediction accuracy exhibit a U-shaped pattern and performance degrades when the critical historical interaction with the strongest collaborative signal appears mid-sequence.
- We propose Disentangling and Calibrating Position Bias, a model-agnostic framework that disentangles positional bias in attention from collaborative signals by estimating an architecture-induced positional prior via constrained projection, and mitigating the bias through bias-adaptive masking during training and counterfactual calibration of attention logits at inference, without disrupting temporal causality.
- We provide a theoretical analysis for DCPB. We formulate a scale-invariant bias metric for comparing U-shaped effects across sequence lengths and a closed-form solver for the constrained projection used to estimate the positional prior.
- Extensive experiments across multiple benchmarks and model variants demonstrate consistent improvements over strong baselines, and ablations confirm the individual contributions of bias-adaptive masking and counterfactual calibration.

2 Related Work

2.1 Positional Influence in Sequential Recommendation

Sequential recommender systems (SRS) aim to capture dynamic user preferences from chronological interactions. Early approaches

based on Markov Chains (MCs) and Recurrent Neural Networks (RNNs) inherently encode order through their recursive structures [16, 34]. The advent of Transformer-based models, such as SASRec [18] and BERT4Rec [38], introduced self-attention mechanisms that significantly outperform their predecessors by modeling global dependencies. Since self-attention is permutation-invariant, these models rely on absolute or relative positional encodings to incorporate sequence order [36, 40]. Research on positional influence in SRS has predominantly focused on the recency effect, where users' most recent interactions are assumed to be most indicative of their future intent [13, 21, 46]. Consequently, various studies have designed decay mechanisms or weighted loss functions to prioritize recent items [5, 29]. However, this focus on sequence tails often overlooks the structural handling of long-term context. In the parallel domain of LLMs, recent findings have identified a "lost-in-the-middle" phenomenon, where models effectively utilize information at the start and end of a context window but neglect the middle [25, 41, 47]. While this U-shaped bias has been analyzed in NLP, its existence and implications in sequential recommendation remain underexplored. Existing SRS studies typically attribute performance drops in long sequences to noise or drift [24, 42], failing to diagnose the architectural tendency of attention mechanisms to systematically suppress mid-sequence signals regardless of their collaborative relevance.

2.2 Calibration and Debiasing Strategies

Recommender systems frequently suffer from biases introduced by data collection and model feedback loops, such as popularity bias [20], exposure bias [4], and selection bias [23]. Mitigation strategies broadly fall into data-centric and model-centric categories. Data-centric approaches, such as Inverse Propensity Weighting (IPW) [35] and data augmentation [26], attempt to reconstruct unbiased training distributions. Model-centric approaches often leverage causal inference [9] or disentangled representation learning [28] to separate user interest from confounding factors. Recently, contrastive learning frameworks like CL4SRec [45] and DuoRec [32] have been employed to enhance representation robustness against noise and sparsity. Calibration techniques, originally designed to align predicted probabilities with true correctness (e.g., Temperature Scaling [12]), have also been adapted for debiasing. For instance, recent works apply calibration to correct over-confidence in popular items [1]. However, existing calibration and debiasing methods primarily target item-level attributes (e.g., popularity) or global prediction confidence. They do not address the positional bias inherent in the attention architecture itself [43]. Unlike prior works that treat position as a static feature or a decay factor, our framework formulates positional bias as a structural confounder.

3 Setup

We follow the standard formulation of sequential recommendation. Given a user u , their historical interaction sequence is denoted as $S_u = (i_1, i_2, \dots, i_t)$, where i_k represents an item interacted with at step k . For attention measurement, we adopt the approach from prior research [25, 44], averaging attention weights across all layers and heads. For the prediction, the sequence representation is summarized into a final user embedding u . Given a candidate item

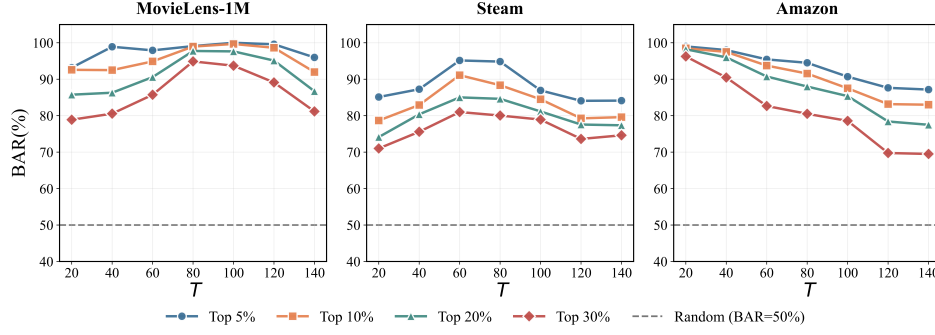


Figure 2: Consistently high BAR values across datasets and sequence lengths indicate robust boundary dominance, and the trend is insensitive to the boundary threshold r .

i with embedding $\mathbf{v}_i$, its interaction probability is computed via a softmax over dot products:

$$P(i | S_u) = \frac{\exp(\mathbf{u}^\top \mathbf{v}_i)}{\sum_{j \in I} \exp(\mathbf{u}^\top \mathbf{v}_j)}. \quad (1)$$

We examine both Transformer attention models (e.g., Bert4Rec), which allow bidirectional interactions among items in the sequence, and causal Transformer models (e.g., SASRec), which apply a causal mask to prevent information leakage from future items.

3.1 Diagnosing Positional Bias

We verify this phenomenon in sequential recommendation using controlled perturbations and quantitative analysis. To verify whether attention aligns with collaborative signals, we identify a critical item i^c in the sequence that is most similar to the target item i^* . Specifically, we define: $i^c = \arg \max_{i_k \in S_u} \cos(\mathbf{v}_k, \mathbf{v}_*)$, where $\mathbf{v}_k$ and $\mathbf{v}_*$ denote the item embeddings of i_k and i^* , respectively. Let $\mathbf{x}_k = \mathbf{v}_k + \mathbf{p}_k$ denote the input embedding of models at step k , where $\mathbf{p}_k$ represents the position embedding.

3.2 U-shape attention bias in recommendation

We characterize the U-shaped attention bias through a progressive analysis. We first provide a qualitative visualization to highlight the phenomenon, and then use controlled sequence perturbations to test whether the pattern is driven by data characteristics or is indeed a model-induced artifact. For causal models such as SASRec, we measure U-shaped attention by the visibility-normalized received attention of each position, i.e., averaging the attention received by a position over all query positions that can causally attend to it, rather than using only the final-query attention row. This avoids the visibility imbalance introduced by causal masking and separates global boundary dominance from the recency preference of the final prediction step. We use ML-1M and fix the sequence length to 35 for controlled comparison across mechanisms. Inspired by previous work [25], we select average attention weights across different segments of the input sequences.

Sequence Swapping. We perform a controlled sequence perturbation to test whether the observed U-shaped attention pattern is driven by data characteristics or induced by the model. Given an original sequence S_u , we generate a permuted sequence S'_u by

swapping its two halves while preserving the internal order within each half. Specifically, let $S_{u,\text{first}} = (i_1, \dots, i_{\lfloor t/2 \rfloor})$ and $S_{u,\text{second}} = (i_{\lfloor t/2 \rfloor + 1}, \dots, i_t)$, and construct

$$S'_u = (S_{u,\text{second}}, S_{u,\text{first}}). \quad (2)$$

If attention primarily follows collaborative signal, relocating items to new positions should shift attention mass accordingly; otherwise, a persistent boundary-heavy pattern indicates an architectural positional bias. Figure 3 shows that even after this swap, the U-shaped distribution of attention weights remains evident, suggesting that the pattern is not explained by boundary items being inherently more informative.

Synthetic Signal Injection. While sequence swapping relocates content, it is still limited by the intrinsic quality of items already present in the sequence. To decouple positional bias from signal strength, we design a Signal Injection experiment. Specifically, we replace the middle item with a synthetic embedding i_{syn} , optimized to maximize similarity to the target while preserving the exact similarity (Sequential Dependency) between the original middle item and its context (see Section 4.3). This construction yields a controlled input that is contextually consistent yet highly predictive. As Figure 3 shows, the model still fails to attend to i_{syn} , indicating that positional bias can suppress collaborative signals even when the middle position contains an intentionally strengthened signal.

3.3 Quantitative Analysis of U-shape Bias

Boundary-based metric (BAR). To quantify the extent to which attention concentrates at sequence extremities, we design a metric *Boundary Attention Ratio* (BAR). Let the model have L layers and H attention heads, and let $\alpha_{l,h,t}$ denote the attention weight assigned to position t by head h in layer l .¹ We treat an attention weight as *significant* if it exceeds the uniform baseline $1/T$, where T is the sequence length. Let $\mathcal{B}$ denote the boundary set (defined below). We define:

$$\text{BAR} = \frac{1}{LH|\mathcal{B}|} \sum_{l=1}^L \sum_{h=1}^H \sum_{t \in \mathcal{B}} \mathbb{I}\left(\alpha_{l,h,t} > \frac{1}{T}\right), \quad (3)$$

¹We use the same attention aggregation protocol as in Section 3 (averaging across layers and heads).

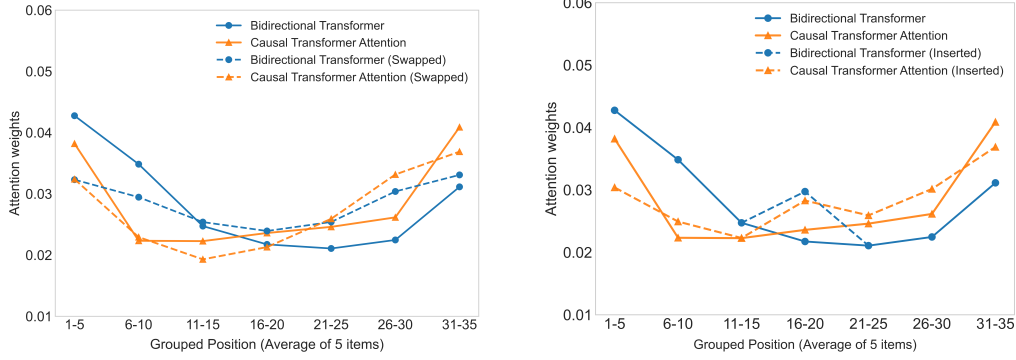


Figure 3: Attention weight distributions under controlled perturbations. Left: sequence swapping. Right: synthetic signal insertion.

where $\mathbb{I}(\cdot)$ is the indicator function and $|\mathcal{B}|$ is the size of the boundary set. A higher BAR indicates that the model consistently allocates above-average attention to boundary positions.

Boundary choice. Based on the scale-invariance and variance-stabilization properties, we define the boundary set $\mathcal{B}$ using a boundary ratio r , taking the first and last $\lceil rT \rceil$ positions of a sequence (i.e., $\mathcal{B} = \{1, \dots, \lceil rT \rceil\} \cup \{T - \lceil rT \rceil + 1, \dots, T\}$). Since the choice of r is heuristic, we evaluate BAR across $r \in \{5\%, 10\%, 20\%, 30\%\}$ in Figure 2. Boundary dominance is consistent across all settings.

Quantitative verification across datasets and perturbations. To generalize beyond the controlled visualization, we compute BAR (with $r = 5\%$) on three benchmarks under the *Original*, *Swapped*, and *Signal Injection* protocols. Table 1 summarizes the results. Across all datasets, BAR is consistently well above the random baseline of 0.5, indicating systematic boundary dominance. Crucially, this effect remains statistically significant ($p < 0.001$) even under *Swapped* and *Signal Injection*, where collaborative signals are displaced or intentionally strengthened. These results provide quantitative evidence that the U-shaped attention pattern is a robust architectural artifact, rather than a consequence of item content or ordering. This motivates the next section, where we explicitly estimate a position-only bias component and calibrate attention to recover collaborative signals in the middle of the sequence.

Table 1: Quantitative verification of U-shape attention weights.

Setting	ML-1M		Steam		Amazon	
	BAR(%)	<i>t</i> -val	BAR(%)	<i>t</i> -val	BAR(%)	<i>t</i> -val
Original	93.52	158.36*	73.39	612.85*	81.81	163.66*
Swapped	74.71	118.39*	62.90	490.17*	67.25	78.92*
Injection	76.37	119.22*	68.51	569.04*	75.46	132.75*

Note. * indicates statistical significance at $p < 0.001$ (one-sample *t*-test against BAR = 0.5).

4 Disentangling and Calibrating Position Bias

We established in Section 3.1 that boundary-dominant attention persists under both sequence swapping and synthetic signal injection,

indicating a systematic positional bias that suppresses collaborative signals in the middle of the sequence. We mitigate this bias with a two-stage causal intervention framework (Figure 4). The remainder of this section is organized as follows: We first formalize attention as a confounded outcome of collaborative, sequential, and positional factors in Section 4.1. Then we empirically verify the separability of these factors in Section 4.2, providing the necessary premise for our approach. In Section 4.3, we construct a neutral probe for isolating pure positional effects. Finally, Section 4.4 details our mitigation strategy, combining bias-adaptive masking during training with counterfactual calibration at inference.

4.1 Factorized Attention Modeling

We encode sequential order using Multi-scale Rotary Position Embeddings (M-RoPE), which rotate query and key representations with multiple frequency scales across attention heads. This provides a relative-position-aware backbone for capturing both short-term transitions and longer-range dependencies without introducing additional item-specific signals.

We view attention generation through a causal lens and treat the raw attention logit $Att(\mathbf{x}_k, k)$ as a confounded outcome of three latent factors: the **Collaborative Signal** $\mathcal{C}(\mathbf{x}_k, \mathbf{v}_*)$ (item–target relevance), the **Sequential Dependency** $\mathcal{D}(S_{\text{ctx}})$ ((consistency with the surrounding context $S_{\text{ctx}} = S_u \setminus \{i_k\}$)), and the **Positional Bias** $\mathcal{P}(k)$ (structural prior). Concretely,

$$Att(\mathbf{x}_k, k) = \mathcal{F}(\mathcal{C}(\mathbf{x}_k, \mathbf{v}_*), \mathcal{D}(S_{\text{ctx}}), \mathcal{P}(k)). \quad (4)$$

Standard attention mechanisms entangle these factors in the observed logits. Our goal is to recover a position-only component associated with $\mathcal{P}(k)$ and use it to calibrate attention for ranking.

4.2 Empirical Justification: Bias Separability

A projection-based estimator is meaningful if positional effects are approximately separable from collaborative signal and sequential dependency. If $\mathcal{P}$ is strongly entangled with either factor, then the projected “bias” term may inadvertently absorb content- or context-driven information.

We therefore test two independence hypotheses about $\mathcal{P}$ using invariance-style conditions:

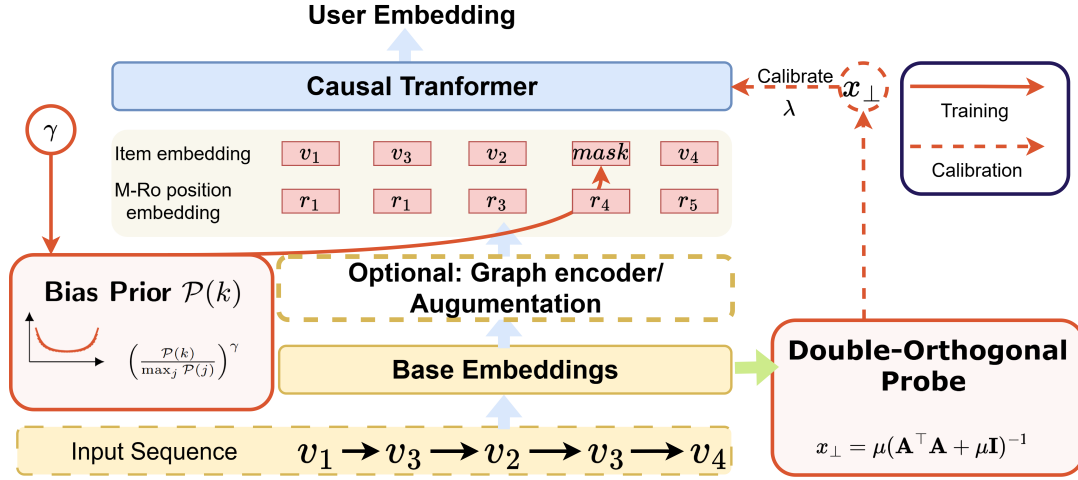


Figure 4: Overview of the proposed framework. We (i) construct controlled embeddings via constrained projection to either inject signal or isolate position-only bias, (ii) apply bias-adaptive masking during training to reduce boundary over-reliance, and (iii) perform counterfactual calibration at inference by subtracting an estimated positional bias term.

Hypothesis 1: Independence of Collaborative Signal ($C \perp \mathcal{P}$). We verify that user preference is invariant to item position:

- **Condition 1 (Preference Invariance):** The relative preference induced by target similarity is position-independent. Given positions n, m , if $\text{Att}(x_i, n) > \text{Att}(x_g, n)$, then $\text{Att}(x_i, m) > \text{Att}(x_g, m)$.
- **Condition 2 (Bias Consistency):** Positional attenuation holds regardless of item relevance. Specifically, if $\text{Att}(x_i, n) > \text{Att}(x_i, m)$, then $\text{Att}(x_g, n) > \text{Att}(x_g, m)$.

Hypothesis 2: Independence of Sequential Dependency ($\mathcal{D} \perp \mathcal{P}$). Similarly, we test if positional bias is orthogonal to context:

- **Condition 3 (Bias Stability):** Positional bias remains stable independent of contextual dependency: if $\text{Att}(x_i, n) > \text{Att}(x_i, m)$, then $\text{Att}(x_r, n) > \text{Att}(x_r, m)$.
- **Condition 4 (Ranking Invariance):** The ranking driven by context is invariant to position: if $\text{Att}(x_i, n) > \text{Att}(x_r, n)$, then $\text{Att}(x_i, m) > \text{Att}(x_r, m)$.

Empirical results confirm all four conditions hold ($p < 0.01$), supporting the use of a projection-based estimator for positional bias.

4.3 Generalized Constraint-Based Projection

We next introduce a constrained projection framework that constructs *controlled* embeddings. This tool serves two purposes: it enables synthetic embeddings used in our diagnostic analysis (Section 3.2), and it produces a neutral probe embedding that isolates position-only effects for bias estimation.

4.3.1 Formulation. Let $M = |S_{\text{ctx}}|$ and define a context matrix $\mathbf{A} \in \mathbb{R}^{(M+1) \times d}$ by stacking the target embedding $\mathbf{v}_*$ and the M context embeddings in S_{ctx} . We synthesize a control vector $\mathbf{x}^*$ that matches a desired signal configuration $\mathbf{b} \in \mathbb{R}^{M+1}$ while remaining close to an anchor $\mathbf{v}_{\text{anc}}$:

$$\mathbf{x}^* = \arg \min_{\mathbf{x}} \left(\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \mu \|\mathbf{x} - \mathbf{v}_{\text{anc}}\|_2^2 \right), \quad (5)$$

where $\mathbf{v}_{\text{anc}}$ is an anchor vector (e.g., a random initialization) and $\mu > 0$ controls the projection strength (trading off alignment against deviation from the anchor).

DEFINITION 1 (CONFIGURABLE SIGNAL CONTROL). The configuration vector $\mathbf{b} = [\alpha, \beta, \dots, \beta]^\top$ specifies the desired alignment, where $\alpha, \beta \in \{0, 1\}$ govern whether to preserve **Collaborative Signal** (C) and **Sequential Dependency** ($\mathcal{D}$), respectively.

LEMMA 1 (CLOSED-FORM EQUIVALENT REPRESENTATION). Let $\mathcal{L}(\mathbf{x})$ denote the objective above. The unique minimizer $\mathbf{x}^*$ admits the equivalent dual form:

$$\mathbf{x}^* = \mathbf{v}_{\text{anc}} - \mathbf{A}^\top (\mathbf{A}\mathbf{A}^\top + \mu\mathbf{I})^{-1} (\mathbf{A}\mathbf{v}_{\text{anc}} - \mathbf{b}), \quad (6)$$

which is algebraically equivalent to the primal ridge solution $(\mathbf{A}^\top \mathbf{A} + \mu\mathbf{I})^{-1} (\mathbf{A}^\top \mathbf{b} + \mu\mathbf{v}_{\text{anc}})$.

REMARK 1. Equation (6) replaces the inversion of the $d \times d$ matrix $(\mathbf{A}^\top \mathbf{A} + \mu\mathbf{I})$ with the inversion of the $(M+1) \times (M+1)$ matrix $(\mathbf{A}\mathbf{A}^\top + \mu\mathbf{I})$. Hence, when $M \leq d$, evaluating the dual form is typically cheaper. Using the primal form for $M > d$ and the dual form for $M \leq d$ ensures the per-step computation is dominated by an inversion of size $\min(d, M+1)$, leading to time complexity $O(\min(d, M+1)^3)$.

Dimension-adaptive computation. Using Lemma 1, we compute $\mathbf{x}^*$ by solving the smaller of the primal and dual linear systems:

$$\mathbf{x}^* = \begin{cases} (\mathbf{A}^\top \mathbf{A} + \mu\mathbf{I})^{-1} (\mathbf{A}^\top \mathbf{b} + \mu\mathbf{v}_{\text{anc}}), & \text{if } M > d \text{ (primal),} \\ \mathbf{v}_{\text{anc}} - \mathbf{A}^\top (\mathbf{A}\mathbf{A}^\top + \mu\mathbf{I})^{-1} (\mathbf{A}\mathbf{v}_{\text{anc}} - \mathbf{b}), & \text{if } M \leq d \text{ (dual).} \end{cases} \quad (7)$$

Although S_{ctx} scales with sequence length, optimization remains efficient. By switching between the primal and dual forms in Equation (7), complexity is strictly bounded by $O(\min(d, M+1)^3)$. It is more efficient than the quadratic complexity $O(M^2 d)$ of standard self-attention layers, particularly for long sequences where $M \gg d$. Since the probe operates solely during inference and bypasses gradient backpropagation, the associated overhead is negligible.

4.3.2 Instantiation: Double-Orthogonal Probe. For bias estimation, we design a probe that removes both target- and context-aligned components. We instantiate the framework with $\alpha = 0$ and $\beta = 0$ (i.e., $\mathbf{b} = \mathbf{0}$) and draw $v_{anc} \sim \mathcal{N}(0, \mathbf{I})$:

$$x_{\perp} = \mu(\mathbf{A}^T \mathbf{A} + \mu \mathbf{I})^{-1} v_{anc}. \quad (8)$$

This suppresses the principal components associated with the span of $\mathbf{A}$ while preserving near-null-space information, so that the resulting logit captures positional effects associated with $\mathcal{P}(k)$.

4.4 Mitigation: Bias-Adaptive Augmentation and Counterfactual Calibration

With a position-only probe available, we apply complementary interventions at training and inference. Training reduces boundary over-reliance in representation learning, while inference removes residual positional effects in logits.

Bias-Adaptive Masking for Model Training. Standard masking assumes uniform information density. In contrast, boundary dominance implies that boundary positions can dominate learning signals. We therefore upweight masking at high-bias positions using a stationary bias profile grounded in *Primacy* and *Recency* effects:

$$\mathcal{P}(k) = \frac{1}{k} + \frac{1}{T - k + 1}, \quad (9)$$

where T is the sequence length. The position-specific masking probability is defined as

$$p_{mask}(k) \propto \left(\frac{\mathcal{P}(k)}{\max_j \mathcal{P}(j)} \right)^{\gamma}, \quad (10)$$

where γ controls how sharply masking concentrates on boundary positions. Larger γ increases masking pressure near the boundaries, encouraging the encoder to leverage mid-sequence signals. During training, we minimize the standard Cross-Entropy Loss for fair comparison [19, 30].

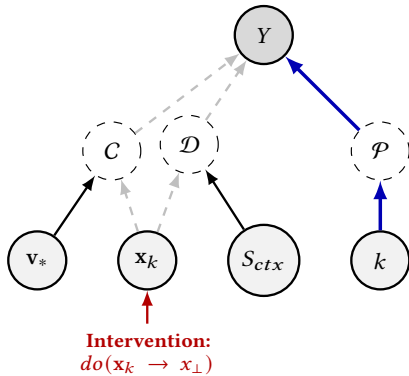


Figure 5: Causal Graph of Attention Generation. We intervene on x_k by replacing it with an orthogonal probe $x_{\perp}$. This suppresses content pathways via C and D , isolating the structural path $k \rightarrow P \rightarrow Y$ for bias estimation.

Counterfactual Attention Calibration for Inference. We calibrate attention at logit level using a probabilistic factorization of positional bias. Under separability assumptions validated in Section 4.2, the attention distribution over position k can be written as:

$$P(k | \mathbf{x}_k, S_{ctx}) \propto \exp(\text{Att}_{cd}(\mathbf{x}_k, S_{ctx})) \cdot \mathcal{P}(k)^{\lambda}, \quad (11)$$

where Att_{cd} captures collaborative and sequential effects, and $\mathcal{P}(k)$ denotes the architecture-induced positional prior.

Positional Prior Estimation. We estimate $\mathcal{P}(k)$ using the neutral probe $x_{\perp}$ (Section 4.3), which removes both target- and context-aligned components. The resulting attention pattern reflects position-only structural bias:

$$\mathcal{P}(k) = \text{Softmax}(\text{Att}(x_{\perp}, k)). \quad (12)$$

Logit-Level Calibration. Taking logs yields an additive decomposition in the logit space. We compute calibrated attention logits:

$$\text{Att}_{calib}(\mathbf{x}_k, k) = \text{Att}(\mathbf{x}_k, k) - \lambda \cdot \log(\mathcal{P}(k) + \epsilon), \quad (13)$$

where λ controls calibration strength and ϵ ensures numerical stability. Final attention weights are obtained via $\text{Softmax}(\text{Att}_{calib})$, which suppresses boundary-dominant structural bias while preserving content-driven ranking.

Prediction. Based on calibrated attention weights $\text{Softmax}(\text{Att}_{calib})$, we obtain the final user embedding $\mathbf{u}$. This representation aggregates the sequence context with mitigated positional bias.

Prediction and Optimization. After stacking L calibrated Transformer layers, we obtain the final hidden states $\mathbf{H}^{(L)} = [\mathbf{h}_1, \dots, \mathbf{h}_T]$. We take the representation at the last position T as the user embedding $\mathbf{u} = \mathbf{h}_T$, which aggregates sequence information with corrected attention distributions. To predict the probability of candidate item i being the next interaction, we compute the dot product between the user embedding $\mathbf{u}$ and item embedding $\mathbf{v}_i$ in Equation (1).

Algorithm 1 DCPB: Disentangling and Calibrating Position Bias Framework

- 1: **Input:** Context sequence S_{ctx} , Candidate set $\mathcal{I}$, Anchor $v_{anc} \sim \mathcal{N}(0, \mathbf{I})$, Hyperparameters μ, γ, λ .
 - 2: **Output:** Recommendation probabilities $P(i | S_u)$ for $i \in \mathcal{I}$.
 - 3: // Phase 1: Bias-Adaptive Training
 - 4: **for** each training step **do**
 - 5: Compute bias profile $\leftarrow$ Equation (9).
 - 6: Set U-shape masking probability $\leftarrow$ Equation (10).
 - 7: Minimizing Cross-Entropy loss.
 - 8: **end for**
 - 9: // Phase 2: Calibrated Inference
 - 10: **Step 1: Constrained Probe Construction**
 - 11: Construct context matrix $\mathbf{A}$ from S_{ctx} .
 - 12: Compute probe $x_{\perp} \leftarrow$ Equation (7).
 - 13: **Step 2: Attention Calibration & Prediction**
 - 14: Estimate position bias $\mathcal{P}(k) \leftarrow$ Equation (12)
 - 15: Compute raw attention logits.
 - 16: Compute final weights by removing $\mathcal{P}(k) \leftarrow$ Equation (13)
 - 17: Return ranked items $\leftarrow$ Equation (1)
-

5 Experiments

We evaluate our framework on three benchmark datasets to answer three research questions:

- **RQ1 (Overall Effectiveness):** Does the proposed full-cycle debiasing framework outperform state-of-the-art sequential recommenders?
- **RQ2 (Ablation Analysis):** How do the bias-adaptive augmentation (training) and counterfactual calibration (inference) individually contribute to the model’s performance?
- **RQ3 (Mechanism Verification):** Does the method effectively mitigate the U-shaped attention bias and recover the representation of middle items?
- **RQ4 (Robustness & Efficiency):** How does the framework perform under varying hyperparameter settings, and is the computational overhead acceptable for real-world deployment?

5.1 Experimental Setup

We conduct experiments on three widely used public datasets: ML-1M [14], Amazon [15], and Steam introduced in [18]. These datasets vary in domain and sparsity, providing a robust evaluation environment. For the Amazon dataset, we use the last 2 million interactions across all categories. Following standard protocols, we adopt the leave-one-out strategy for data splitting [18, 22]. The dataset statistics are summarized in Table 2. Furthermore, we compare our proposed method against a comprehensive set of baselines drawn from four families of sequential recommenders. For RNN/CNN-based methods, we include GRU4Rec [16], DSSRec [27], and Caser [39], and we report BPR [33] as a static reference. For transformer-based models, we consider SASRec [18], Bert4Rec [38], and Mi-aSRec [6]. For graph-based approaches, we include SURGE [3], which constructs an item–item interest graph to capture dynamic user preferences. Finally, we compare against contrastive learning and augmentation-based methods, including CL4SRec [45], DuoRec [32], CoSeRec [26], ICSRec [31], and BASRec [7]. The implementation details and metrics are further elaborated in Appendix.

Table 2: Statistics of the datasets.

Dataset	# Users	# Items	# Interacts	Avg. Len	Density
Amazon	65,262	106,427	1,540,690	23.61	0.02%
ML-1M	6,040	3,416	999,611	165.50	4.84%
Steam	39,795	10,587	1,790,393	44.99	0.42%

5.2 Overall Performance Comparison (RQ1)

We answer RQ1 by examining the performance from two complementary perspectives: (1) **Comparative Superiority**, where we benchmark against state-of-the-art methods; and (2) **Backbone Compatibility**, where we validate the framework’s efficacy as a model-agnostic enhancement.

5.2.1 Comparison with State-of-the-Arts. Table 3 summarizes the performance of our framework compared to the baselines. The results reveal three key observations of the proposed method:

- **Superiority over State-of-the-Art Baselines:** As reported in Table 3, DCPB consistently surpasses all baselines across

datasets. Notably, on ML-1M, our approach achieves a substantial **34.38%** improvement in MRR@5 compared to the most competitive baseline (BASRec). While existing methods effectively model sequential dependencies, they remain constrained by inherent positional biases. These results validate that explicitly calibrating positional logits mitigates the “lost-in-the-middle” phenomenon, effectively recovering informative interactions in the middle that are suppressed by structural bias.

- **Advantage over Contrastive Baselines:** Compared to contrastive learning methods (e.g., CL4SRec, DuoRec) that rely on stochastic data augmentation (e.g., cropping or masking) to enhance representations, DCPB offers a more deterministic and stable solution. On the Amazon dataset, we surpass the BASRec by **23.37%** in MRR@10. This indicates that our counterfactual calibration provides a more targeted supervisory signal than generic augmentation, effectively guiding the decoder to rectify attention patterns rather than merely expanding the view through constructing negative samples.
- **Robustness in Sparse Scenarios:** The performance gap is most pronounced on the sparse Amazon dataset, where we observe a **30.25%** leap in MRR@5. In sparse settings, standard models tend to overfit to sequence boundaries due to limited collaborative signals. Our counterfactual calibration effectively removes this position-only bias during inference, demonstrating superior robustness when interaction data is scarce.

5.2.2 Model Agnostic Analysis. Beyond direct competition, a core advantage of our framework is its flexibility. We treat the proposed training-inference as a model-agnostic plugin that can be integrated into existing architectures. Specifically, prior graph-based or augmentation-based methods could integrate into optional module (in Figure 4). Synergy results in Table 4 verify our method as a versatile, plug-and-play solution,

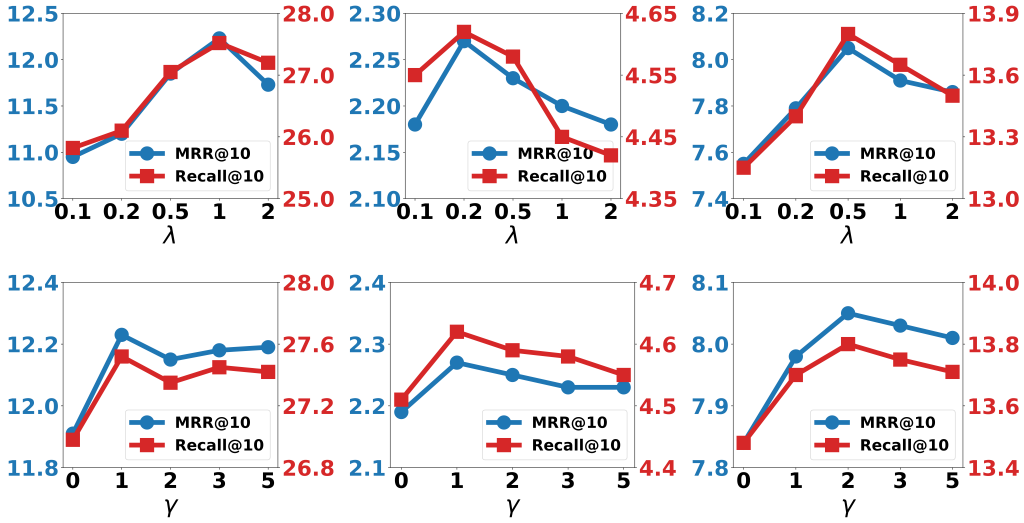
- **Enhancing Augmentation Frameworks:** We integrate our method with **CoSeRec**, a state-of-the-art model relying on data augmentation and contrastive learning. On ML-1M dataset, this combination yields a boost. It indicates that while CoSeRec’s augmentation addresses data sparsity and noise, it does not correct structural positional biases. Our method shows the structural debiasing is complementary to data augmentation.
- **Enhancing Graph-based Models:** We further validate compatibility with **SURGE**, which models user interests via dynamic graph neural networks. The integration results in an improvement in Recall@10 on ML-1M. It could attributed to the complementary nature of the two approaches: while SURGE excels at capturing item transitions via graph topology, its node aggregation mechanism inherently dilutes strict sequential signals. Our framework compensates for this by explicitly calibrating positional importance, ensuring that the temporal evolution of interests is preserved alongside structural relationships.

5.3 Ablation Study (RQ2)

To disentangle the contributions of our approach, we conduct an ablation study by removing or replacing components in the proposed framework. The results, summarized in Table 5, yield the following observations:

Table 3: Performance comparison on three datasets. We report the mean results with standard deviation. Best results are bolded, and second-best are underlined. Improvement is calculated relative to the best baseline.

Dataset	Metric	GRU4Rec	Caser	DSSRec	Bert4Rec	SASRec	MiaSRec	CL4SRec	DuoRec	ICSRec	BASRec	DCPB	Improv.
ML-1M	Recall@5	2.44	6.89	7.21	9.96	10.19	11.80	9.95	11.58	10.70	15.75±0.14	18.89±0.15	+19.94%
	Recall@10	3.29	11.27	14.35	18.55	16.63	19.75	16.30	17.41	17.25	<u>24.75±0.19</u>	27.52±0.21	+11.19%
	MRR@5	1.44	3.59	3.11	4.18	4.92	5.50	5.08	5.71	5.33	<u>8.26±0.08</u>	11.10±0.11	+34.38%
	MRR@10	1.56	4.55	3.97	5.33	5.75	6.50	5.92	5.94	6.19	<u>9.44±0.10</u>	12.23±0.13	+29.56%
Amazon	Recall@5	1.95	2.29	2.23	1.20	2.76	2.20	<u>2.89±0.11</u>	1.51	2.80	2.75	3.37±0.05	+16.61%
	Recall@10	2.85	3.17	3.68	2.27	4.23	3.41	4.15	2.52	<u>4.32±0.06</u>	4.27	4.62±0.07	+6.94%
	MRR@5	1.15	1.48	0.62	0.51	1.41	1.15	<u>1.62±0.05</u>	0.77	1.47	1.44	2.11±0.03	+30.25%
	MRR@10	1.35	1.59	0.72	0.64	1.61	1.31	<u>1.84±0.09</u>	0.91	1.67	1.64	2.27±0.05	+23.37%
Steam	Recall@5	7.48	8.89	8.20	9.58	9.82	9.71	9.37	7.31	9.70	<u>9.93±0.11</u>	10.35±0.10	+4.23%
	Recall@10	10.23	11.74	12.50	12.74	13.05	12.81	12.71	10.66	13.28	<u>13.41±0.13</u>	13.80±0.15	+2.91%
	MRR@5	5.19	6.64	6.88	7.08	7.13	7.16	6.74	4.55	7.12	<u>7.21±0.09</u>	7.60±0.06	+5.41%
	MRR@10	5.50	7.01	7.39	7.49	7.59	7.57	7.17	5.00	7.60	<u>7.64±0.07</u>	8.05±0.08	+5.37%

**Figure 6: Hyper-parameter sensitivity analysis of λ and γ on ML, Amazon, and Steam datasets.****Table 4: Plug-and-Play comparison. We observe significant improvements over both backbones.**

Method	ML-1M		Steam	
	Recall@10	MRR@10	Recall@10	MRR@10
CoSeRec	20.99	9.66	12.88	7.65
+ Ours	26.75	11.65	13.73	7.96
Improv.	+27.44%	+20.60%	+6.60%	+4.05%
SURGE	23.69	10.40	12.92	7.19
+ Ours	27.11	12.10	13.51	7.75
Improv.	+14.44%	+16.35%	+4.57%	+7.79%

- **Impact of Training Structure (w/o U-Mask):** Removing U-shaped masking strategy results in a consistent performance drop

(e.g., Recall@10 declines from 27.52 to 26.98 on ML-1M). This indicates standard causal masking is insufficient for alleviating position bias. Our U-shaped design effectively augments the training signal, forcing the model to attend to informative historical interactions rather than over-relying on recent positions.

- **Necessity of Inference Correction (w/o Calib.):** Eliminating inference-time calibration leads to a substantial degradation (Recall@10 drops to 24.81 on ML-1M). This validates our core hypothesis: structural bias inherent in Transformers heavily distorts attention scores. The counterfactual calibration helps in rectifying these shifted logits to reflect genuine user intent.
- **Effectiveness of Position Encoding (w/o M-Ro):** Replacing our Multi-scale RoPE with a standard solution ('w/o M-Ro') causes the most severe drop among structural variants. This underscores that fine-grained, multi-frequency position modeling is essential for capturing long-term dependencies in user behavior sequences, which fixed-scale encodings fail to preserve.

Table 5: We evaluate the contribution of each component (top), the impact of probe strategies (middle), and calibration operations (bottom).

Variant	ML-1M		Steam	
	Recall@10	MRR@10	Recall@10	MRR@10
DCPB (Full)	27.52	12.23	13.80	8.05
(1) w/o U-Mask	26.98	11.91	13.48	7.84
(2) w/o Calib.	24.81	10.14	12.74	7.15
(3) w/o M-Ro	23.81	9.98	10.25	6.66
(4) w/ Zero Probe	25.15	10.28	12.92	7.28
(5) w/ Avg. Probe	25.76	10.27	12.94	7.41
(6) w/ Rand. Probe	24.02	9.20	12.22	6.59
(7) w/ Div. Calib. (/)	3.75	1.29	3.48	1.19
(8) w/ Mul. Calib. (×)	4.42	1.46	4.71	1.70

Table 6: Efficiency comparison across three datasets. We report the total wall-clock time (in seconds) for the training phase and the closed-form calibration phase.

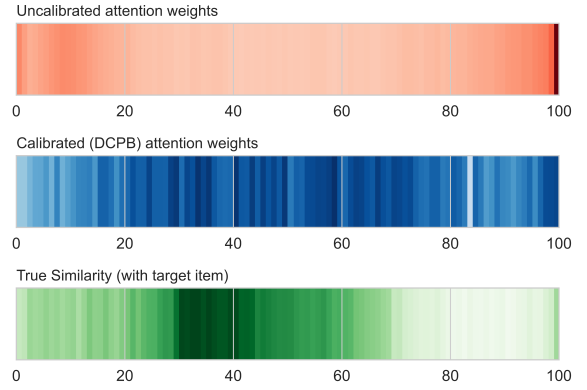
Dataset	Model	Training (s)	Calibration (s)
ML-1M	SASRec ⁺	64	15
	BERT ⁺	241	17
Amazon	SASRec ⁺	2587	221
	BERT ⁺	16135	289
Steam	SASRec ⁺	332	74
	BERT ⁺	1944	83

- **Probe Strategy Analysis ((4)-(6)):** We evaluate various probes for bias estimation. The random probe (w/ Rand. Probe) introduces noise, degrading performance below the uncalibrated baseline. In contrast, static heuristics, specifically the Zero Probe and history Average Probe, yield moderate improvements by providing stable reference points. However, both lag significantly behind our learnable anchor strategy (DCPB Full).
- **Validity of Logit Operations ((7) and (8)):** We further verify the mathematical basis of our calibration. Applying multiplicative (‘×’) or divisive (‘/’) operations leads to catastrophic failure (single-digit Recall). Since bias acts multiplicatively in the probability space, it corresponds to an additive term in the logit space.

5.4 Analysis of De-biasing Efficacy (RQ3)

Visualizing Attention Distribution. To verify the effectiveness of our method, we visualize the attention distribution for sequences where the critical item (i.e., the one with the strongest collaborative signal) is located in the middle 60% of the sequence. As shown in Figure 7, the uncalibrated model exhibits a distinct U-shaped bias: even though the middle items are highly informative, the attention mechanism fails to focus on them, instead disproportionately allocating weights to the sequence boundaries.

In contrast, our calibrated model successfully rectifies this distortion. The attention peak shifts significantly towards the true signal

**Figure 7: Attention weights for sequences where the item with the strongest collaborative signal appears in the middle 60% of the relative position on ML-1M. The x-axis denotes the relative item position in the sequence (percentage).**

in the middle, demonstrating the model’s recovered ability to capture key information. While the bias towards the initial positions is suppressed, the model retains a moderate attention weight on the most recent items (approaching the 100% relative position). This residual weight is acceptable and even desirable, as it reflects the short-term interest of users in recommendation systems.

5.5 Robustness and Efficiency Analysis (RQ4)

Hyperparameter Sensitivity. We analyze the sensitivity of two key hyperparameters: calibration strength λ (Equation (13)) and augmentation focus γ (Equation (10)). As illustrated in Figure 6:

- **Calibration Strength (λ):** We observe a positive correlation between the optimal λ and the average sequence length of the dataset. Longer sequences tend to suffer from more severe structural U-shaped bias, necessitating a stronger calibration ($\lambda \rightarrow 1$) to suppress these artifacts.
- **Augmentation Intensity (γ):** Increasing γ encourages model to rely on mid-sequence patterns. However, excessively high γ yields a steeper masking curve, causing the masking probability to converge almost exclusively on the head and tail items. This over-suppression leads to a slight degradation in performance.

Computational Efficiency. We report the computational efficiency in Table 6. The calibration step is lightweight and incurs significantly lower time cost than backbone model training.

6 Conclusion

This work identifies a *lost-in-the-middle* failure mode in attention-based sequential recommendation and shows that it is robust to boundary-dominant positional bias via controlled perturbations and quantitative analysis. We propose DCPB, which estimates architecture-induced positional prior via constrained projection, mitigates bias via bias-adaptive masking and counterfactual logit calibration, and preserves temporal causality. Moreover, our approach is plug-and-play, and can be integrated into existing attention-based recommendation models without significant overhead.

References

- [1] Himan Abdollahpour, Robin Burke, and Bamshad Mobasher. 2017. Controlling popularity bias in learning-to-rank recommendation. In *Proceedings of the eleventh ACM conference on recommender systems*. 42–46.
- [2] Gianni Brauers and Flavius Frasin. 2021. A general survey on attention mechanisms in deep learning. *IEEE transactions on knowledge and data engineering* 35, 4 (2021), 3279–3298.
- [3] Jianxin Chang, Chen Gao, Yu Zheng, Yiqun Hui, Yanan Niu, Yang Song, Depeng Jin, and Yong Li. 2021. Sequential recommendation with graph neural networks. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 378–387.
- [4] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems* 41, 3 (2023), 1–39.
- [5] Lihua Chen, Ning Yang, and Philip S Yu. 2022. Time lag aware sequential recommendation. In *Proceedings of the 31st ACM international conference on information & knowledge management*. 212–221.
- [6] Minjin Choi, Hye-young Kim, Hyunsouk Cho, and Jongwuk Lee. 2024. Multi-tenant-aware session-based recommendation. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*. 2532–2536.
- [7] Yizhou Dang, Jiahui Zhang, Yuting Liu, Enneng Yang, Yuliang Liang, Guibing Guo, Jianzhe Zhao, and Xingwei Wang. 2025. Augmenting Sequential Recommendation with Balanced Relevance and Diversity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 11563–11571.
- [8] Hui Fang, Danning Zhang, Yiheng Shu, and Guibing Guo. 2020. Deep learning for sequential recommendation: Algorithms, influential factors, and evaluations. *ACM Transactions on Information Systems (TOIS)* 39, 1 (2020), 1–42.
- [9] Chen Gao, Yu Zheng, Wenjie Wang, Fuli Feng, Xiangnan He, and Yong Li. 2024. Causal inference in recommender systems: A survey and future directions. *ACM Transactions on Information Systems* 42, 4 (2024), 1–32.
- [10] Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. [n. d.]. Think before you speak: Training language models with pause tokens. 2024. URL <https://arxiv.org/abs/2310.02226> ([n. d.]).
- [11] Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. 2024. When attention sink emerges in language models: An empirical view. *arXiv preprint arXiv:2410.10781* (2024).
- [12] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*. PMLR, 1321–1330.
- [13] Linjiang Guo, Shiqing Wu, Dan Lu, Longxiang Gao, and Guandong Xu. 2026. Dual-channel time-aware graph attention network for session-based recommendation. *Information Sciences* (2026), 123289.
- [14] F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)* 5, 4 (2015), 1–19.
- [15] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*. 507–517.
- [16] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939* (2015).
- [17] Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, et al. 2024. Found in the middle: Calibrating positional attention bias improves long context utilization. In *Findings of the Association for Computational Linguistics: ACL 2024*. 14982–14995.
- [18] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*. IEEE, 197–206.
- [19] Anton Klenitskiy and Alexey Vasilev. 2023. Turning dross into gold loss: is bert4rec really better than sasrec?. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 1120–1125.
- [20] Anastasiia Klimashevskaya, Dietmar Jannach, Mehdi Elahi, and Christoph Trattner. 2024. A survey on popularity bias in recommender systems. *User Modeling and User-Adapted Interaction* 34, 5 (2024), 1777–1834.
- [21] Jiacheng Li, Yujie Wang, and Julian McAuley. 2020. Time interval aware self-attention for sequential recommendation. In *Proceedings of the 13th international conference on web search and data mining*. 322–330.
- [22] Yicong Li, Hongxu Chen, Xiangguo Sun, Zhenchao Sun, Lin Li, Lizhen Cui, Philip S Yu, and Guandong Xu. 2021. Hyperbolic hypergraphs for sequential recommendation. In *Proceedings of the 30th ACM international conference on information & knowledge management*. 988–997.
- [23] Jiaye Lin, Qing Li, Guorui Xie, Zhongxu Guan, Yong Jiang, Ting Xu, Zhong Zhang, and Peilin Zhao. 2024. Mitigating Sample Selection Bias with Robust Domain Adaption in Multimedia Recommendation. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 7581–7590.
- [24] Xiaolin Lin, Wei Pan, and Zhong Ming. 2025. Towards Interest Drift-driven User Representation Learning in Sequential Recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1541–1551.
- [25] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics* 12 (2024), 157–173.
- [26] Zhiwei Liu, Yongjun Chen, Jia Li, Philip S Yu, Julian McAuley, and Caiming Xiong. 2021. Contrastive self-supervised sequential recommendation with robust augmentation. *arXiv preprint arXiv:2108.06479* (2021).
- [27] Jianxin Ma, Chang Zhou, Hongxia Yang, Peng Cui, Xin Wang, and Wenwu Zhu. 2020. Disentangled self-supervision in sequential recommenders. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 483–491.
- [28] Zhonghong Ou, Xiao Zhang, Yifan Zhu, Shuai Lyu, Jiahao Liu, and Tu Ao. 2025. LS-TGNN: Long and Short-Term Temporal Graph Neural Network for Session-Based Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 12426–12434.
- [29] Aleksandr Petrov and Craig Macdonald. 2024. RSS: effective and efficient training for sequential recommendation using recency sampling. *ACM Transactions on Recommender Systems* 3, 1 (2024), 1–32.
- [30] Aleksandr Vladimirovich Petrov and Craig Macdonald. 2023. gsasrec: Reducing overconfidence in sequential recommendation trained with negative sampling. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 116–128.
- [31] Xiuyuan Qin, Huanhuan Yuan, Pengpeng Zhao, Guanfang Liu, Fuzhen Zhuang, and Victor S Sheng. 2024. Intent contrastive learning with cross subsequences for sequential recommendation. In *Proceedings of the 17th ACM international conference on web search and data mining*. 548–556.
- [32] Ruihong Qiu, Zi Huang, Hongzhi Yin, and Zijian Wang. 2022. Contrastive learning for representation degeneration problem in sequential recommendation. In *Proceedings of the fifteenth ACM international conference on web search and data mining*. 813–823.
- [33] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618* (2012).
- [34] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*. 811–820.
- [35] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as treatments: Debiasing learning and evaluation. In *international conference on machine learning*. PMLR, 1670–1679.
- [36] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155* (2018).
- [37] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* 568 (2024), 127063.
- [38] 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.
- [39] Jiayi Tang and Ke Wang. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 565–573.
- [40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [41] David Wan, Jesse Vig, Mohit Bansal, and Shafiq Joty. 2025. On positional bias of faithfulness for long-form summarization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 8791–8810.
- [42] Bohao Wang, Feng Liu, Changwang Zhang, Jiawei Chen, Yudi Wu, Sheng Zhou, Xingyu Lou, Jun Wang, Yan Feng, Chun Chen, et al. 2025. Llm4dsr: Leveraging large language model for denoising sequential recommendation. *ACM Transactions on Information Systems* 44, 1 (2025), 1–32.
- [43] Xinyi Wu, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. 2025. On the emergence of position bias in transformers. *arXiv preprint arXiv:2502.01951* (2025).
- [44] Guangxuan Xiao, Yundong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453* (2023).
- [45] Xu Xie, Fei Sun, Zhaoyang Liu, Shiwen Wu, Jinyang Gao, Jiandong Zhang, Bolin Ding, and Bin Cui. 2022. Contrastive learning for sequential recommendation. In *2022 IEEE 38th international conference on data engineering (ICDE)*. IEEE, 1259–1273.
- [46] Junjie Zhang, Ruobing Xie, Hongyu Lu, Wenqi Sun, Wayne Xin Zhao, Yu Chen, and Zhanhui Kang. 2025. Frequency-Augmented Mixture-of-Heterogeneous-Experts Framework for Sequential Recommendation. In *Proceedings of the ACM*

- on *Web Conference 2025*. 2596–2605.
- [47] Zhenyu Zhang, Runjin Chen, Shiwei Liu, Zhewei Yao, Olatunji Ruwase, Beidi Chen, Xiaoxia Wu, Zhangyang Wang, et al. 2024. Found in the middle: How

language models use long contexts better via plug-and-play positional encoding. *Advances in Neural Information Processing Systems* 37 (2024), 60755–60775.