Recommender Transformers with Behavior Pathways

Zhiyu Yao[†], Xinyang Chen[†], Sinan Wang[‡], Qinyan Dai[*],
Yumeng Li[‡], Tanchao Zhu[‡], Mingsheng Long[†](✉)
School of Software, BNRist, Tsinghua University, China[†]
Alibaba Group, China[‡]
Industrial Engineering, Tsinghua University[*]
{yaozy19,chenxiny17,dai-qy18}@mails.tsinghua.edu.cn
mingsheng@mails.tsinghua.edu.cn

Abstract

Sequential recommendation requires the recommender to capture the evolving behavior characteristics from logged user behavior data for accurate recommendations. However, user behavior sequences are viewed as a script with multiple ongoing threads intertwined. We find that only a small set of pivotal behaviors can be evolved into the user's future action. As a result, the future behavior of the user is hard to predict. We conclude this characteristic for sequential behaviors of each user as the *Behavior Pathway*. Different users have their unique behavior pathways. Among existing sequential models, transformers have shown great capacity in capturing global-dependent characteristics. However, these models mainly provide a dense distribution over all previous behaviors using the self-attention mechanism, making the final predictions overwhelmed by the trivial behaviors not adjusted to each user. In this paper, we build the *Recommender Transformer* (RETR) with a novel *Pathway Attention* mechanism. RETR can dynamically plan the behavior pathway specified for each user, and sparingly activate the network through this behavior pathway to effectively capture evolving patterns useful for recommendation. The key design is a learned binary route to prevent the behavior pathway from being overwhelmed by trivial behaviors. We empirically verify the effectiveness of RETR on seven real-world datasets and RETR yields state-of-the-art performance.

1 Introduction

Recommender systems [14, 21, 39] have been widely adopted in real-world industrial applications such as E-commerce and social media. Benefiting from the increase in computing power and model capacity, some recent efforts formulate recommendation as a time-series forecasting problem, known as *sequential recommendation* [16, 28, 6]. The core idea of this field is to infer upcoming actions based on user's historical behaviors, which are reorganized as time-ordered sequences. This intuitive modeling of recommendation is proved time-sensitive and context-aware to make precise predictions.

Recent advanced sequential recommendation models, such as SASRec [16], Bert4Rec [28] and SMRec [6], have achieved significant improvements. Transformers enable these models to recognize global-range sequential patterns, and to model how future behaviors are anchored in historical ones. The self-attention mechanism does make it possible to explore all previous behaviors of each user, with the whole neural network activated. However, misuse of user information, regardless of whether they are informative or not, floods models with trivial ones, makes models dense and inefficient, and results in key behaviors losing voice. And this clearly contradicts with the way our brain works.

The human being has many different parts of the brain specialized for various tasks, yet the brain only calls upon the relevant pieces for a given situation [37]. To some extent, user behavior sequences can

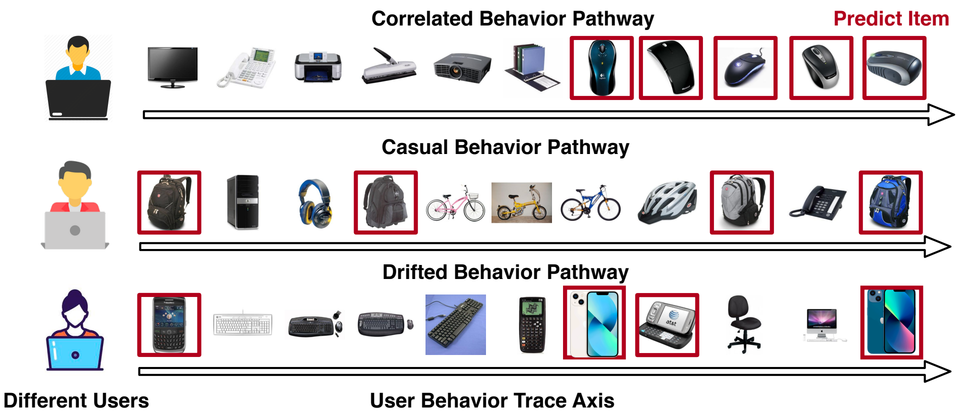

Figure 1: Three main characteristics of the behavior pathway for different users, making sequential recommendation extremely hard. The behavior pathway is outlined by the red boxes.

be viewed as a script with multiple ongoing threads intertwined. And only key clues suggest what will happen next. In sequential recommendation, we find that only a small part of pivotal behaviors can be evolved into the use's future action. And we conclude this characteristics of sequential behaviors as the *Behavior Pathway*.

Different users have their unique behavior pathways and they can be grouped into three categories:

- **Correlated Behavior Pathway**: A user's behavior pathway is closely associated with behaviors at a certain period. As shown in the second line of Figure 1, the mouse is clicked many times recently, leading to the final decision to buy a mouse.
- **Casual Behavior Pathway**: A user's behavior pathway is interested in a specific item at casual times. As shown in the first line of Figure 1, the backpack is randomly clicked sequentially.
- **Drifted Behavior Pathway**: A user's behavior pathway in a particular brand might drift over time. As shown in the third line of Figure 1, the user was initially interested in a keyboard, but suddenly became interested in buying a phone at last.

It's challenging to capture these aspects dynamically for each user to make precise recommendations.

Motivated by the Pathways [8], a new way of thinking about AI, which builds a single model that is sparsely activated for all tasks with small pathways through the network called into action as needed, we propose a novel *Recommender Transformer* (RETR) with a *Pathway Attention* mechanism. RETR dynamically explores behavior pathways for different users and then captures evolving patterns through these pathways effectively. To be specific, the user-dependent pathway attention, which incorporates a pathway router, determines whether or not a behavior token will be maintained in the behavior pathway. Technically, the pathway router generates a customized binary route for each token based on their information redundancy. Recommender Transformers are stacked, and successive pathway routers constitute a hierarchical evolution pathway of user behaviors. To enable the pathway router modules to be end-to-end optimized, we adopt the Gumbel-Softmax [15] sampling strategy to overcome the non-differentiable problem of sampling from a Bernoulli distribution.

To effectively capture the evolving patterns via the behavior pathway, our pathway attention mechanism makes our RETR mainly attend to the obtained pathway. We force the model to focus on the most informative behaviors by using the query routed through the behavior pathway. We cut off the interaction from the off-pathway behaviors of the query. Compared with using all previous behaviors, our pathway attention mechanism is obviously more effective and can avoid the most informative tokens being overwhelmed by trivial behaviors. To validate the effectiveness of our approach, we conduct experiments on seven real-world competitive datasets for sequential recommendations and RETR achieves state-of-the-art performance. Our main contributions can be summarized as follows:

- Our work is the first to propose the concept of behavior pathway for sequential recommendation. We find the key to the recommender is to dynamically capture the behavior pathway for each user.

- We propose the novel recommender transformer (RETR) with a novel pathway attention mechanism, which can generate the behavior pathway hierarchically and capture the evolving patterns dynamically through the pathway.
- We validate the effectiveness of RETR on seven real-world datasets of different scales across different scenarios for sequential recommendations and achieve state-of-the-art performance.

2 Related Work

Traditional recommendation approaches. Capturing evolving behavior characteristics is crucial for many online applications, such as advertising, social media and E-commerce, and it is the key challenge for sequential recommendation [1, 16, 7, 36, 9, 24, 5, 41, 20]. Traditional recommendation approach, such as the collaborative filtering (CF) [13] based on matrix approximation [18, 19], always assumes that the user's behavior is static. However, in practice, user behaviors often change over time due to various reasons, making the CF deteriorate in a real-world application.

Sequential recommendation approaches. To overcome this challenge, some methods, such as FPMC [12] and HRM [30], use Markov chains to capture sequential patterns by learning user-specific transition matrices. Higher-order Markov Chains assume the next action is related to several previous actions. Benefit from this strong inductive bias, MC-based methods [12, 11] show superior performance in capturing short-term patterns. At the same time, there is a potential state space explosion problem when these approaches are faced with different possible sequences [31]. In recent years, many works have been using the deep neural network for sequential recommendation. For example, the GRURec [14] and the RepeatNet [25] adopt the recurrent network to capture dynamic patterns from the user behaviors dependent on sequence positions. The RNN-based models achieve competitive performance in capturing short-term behavior patterns but cannot capture long-term behavior patterns effectively. The CNN-based model, such as Caser [29], applies convolutional operations to extract transitions while tending to overlook the intrinsic relationship across user behaviors. The GNN-based methods, such as SRGNN [33] and GCSAN [34], model behavior sequences as graph-structured data and incorporate an attention mechanism for a session-based recommendation. In addition, DIN [40] uses the gate mechanism to weight different user behaviors. However, concatenating all behaviors makes these models overlook the sequential characteristics.

Transformer-based models for Sequential Recommendation. The transformer-based models have the strong capacity to capture behavior patterns via the attention mechanism, achieving state-of-the-art performance while involving many parameters. SASRec [16], BertRec [28] and SSE-PT [32] introduce the transformer architecture into sequential recommendation, which might lead to the over-parameterized architecture of Transformer-based methods. These models capture the evolving patterns by the self-attention mechanism, interacting with all previous behaviors. However, dense interactions will make the model not adapt to different users and overwhelm behavior pathways. To tackle this challenge, our paper builds the Recommender Transformer (RETR) with a new Pathway Attention mechanism that is dynamically activated for the behavior pathway of all users. Distinct from the previous routing architecture like Switch Transformer [10] using the MoE [27] structure for natural language tasks, our RETR is designed explicitly for sequential recommendation. Our RETR uses the pathway router to adaptively route the sequential behavior of each user rather than routing the experts of feed-forward networks in switch transformer.

3 Method

Suppose that we have a set of users and items, denoted by $\mathcal{U}$ and $\mathcal{I}$ respectively. In the task of sequential recommendation, chronologically-ordered behaviors of a user $u \in \mathcal{U}$ could be represented by a user-interacted item sequence: $\{i_1, \cdots, i_n\}$. Formally, given a user u with her or his behavior sequence $\{i_1, \cdots, i_n\}$, the goal of sequential recommendation is to predict the next item the user u would interact with at the $(n+1)$-th step, denoted as $p\left(i_{n+1} \mid i_{1:n}\right)$.

As aforementioned, we highlight the key to sequential recommendation as the exploration of user-tailored behavior pathways, through which evolving characteristics could be learned. Motivated by this, we propose a novel *Recommender Transformer* (RETR) with a new *Pathway Attention*, the core subassembly of which is a pathway router. Besides the modification of architecture, we additionally introduce a hierarchical update strategy for the behavior pathway in the feed-forward procedure.

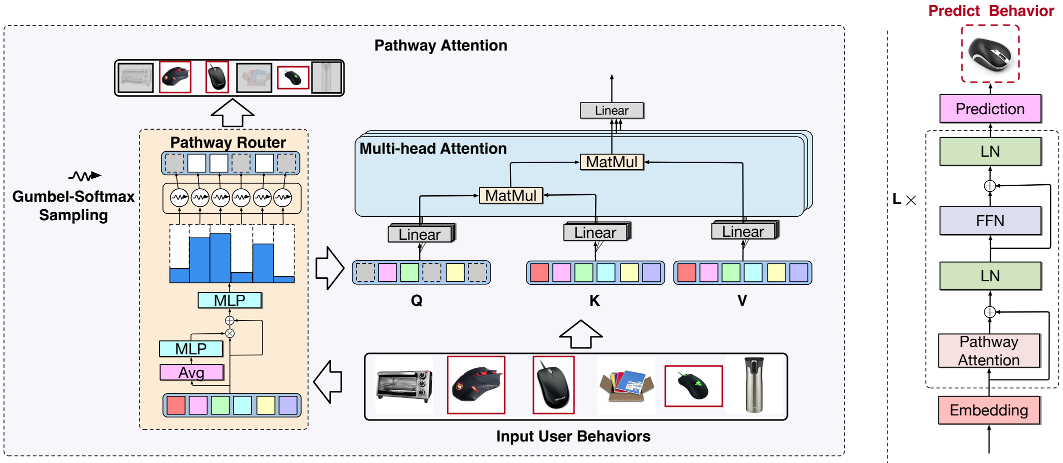

Figure 2: Recommender Transformer architecture (right). Pathway Attention (left) explores the behavior pathway by the pathway router (orange module) and captures the evolving sequential characteristics by the multi-head attention.

3.1 Recommender Transformer

Considering the limitation of Transformers [4] for sequential recommendation, we renovate the vanilla architecture to the Recommender Transformer (Figure 2) with a Pathway Attention mechanism.

Model inputs. To obtain the model inputs, we follow the sliding window practice and transform the user's behavior sequence into a fixed-length-N sequence $s = (s_1, s_2, \ldots, s_N)$. Then we produce an item embedding matrix $\mathcal{E}_{\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}|\times d}$, where d is the embedding dimensionality. We perform a look-up operation from $\mathcal{E}_{\mathcal{I}}$ to retrieve the input embedding matrix $\mathcal{E}_s \in \mathbb{R}^{N\times d}$ for sequence s. Besides, we also add a learnable position embedding $\mathcal{P}_s \in \mathbb{R}^{N\times d}$ for sequence s. Finally, we can generate the input embedding of each behavior sequence s as $\mathcal{X}_s = \mathcal{E}_s + \mathcal{P}_s \in \mathbb{R}^{N\times d}$.

Overall architecture. Recommender Transformer is characterized by stacking the Pathway Attention blocks and feed-forward layers alternately, containing L blocks. This stacking structure is conducive to learning behavior representations hierarchically. The overall equations of block l are formalized as:

$$\begin{aligned}\widehat{\mathcal{Z}}^l, \mathcal{R}^l &= \text{Path-MSA}\left(\mathcal{Z}^{l-1}, \mathcal{R}^{l-1}\right)\\ \widehat{\mathcal{Z}}^l &= \text{LN}\left(\widehat{\mathcal{Z}}^l + \mathcal{Z}^{l-1}\right)\\ \mathcal{Z}^l &= \text{LN}\left(\text{FFN}\left(\widehat{\mathcal{Z}}^l\right) + \widehat{\mathcal{Z}}^l\right),\end{aligned} \tag{1}$$

where $\mathcal{Z}^l \in \mathbb{R}^{N\times d}, l \in \{1, \cdots, L\}$ denotes the output of the l-th block. The initial input $\mathcal{Z}^0 = \mathcal{X}_s \in \mathbb{R}^{N\times d}$ represents the raw behavior embedding. $\mathcal{R}^{l-1}$ is the previous route from the block $l-1$ and we initialize all elements in the route $\mathcal{R}^0$ to 1. Path-MSA$(\cdot)$ is to conduct the pathway attention. LN$(\cdot)$ is to conduct layer normalization [3] and FFN represents the point-wise feed-forward network [4].

3.1.1 Pathway Attention

Note that the single-branch self-attention mechanism [4] in vanilla transformer cannot model the behavior pathway dynamically, resulting in key behaviors being overwhelmed by these non-pivotal ones. To solve this problem, we propose the Pathway Attention mechanism, as shown in Figure 2, which can dynamically attend to the behavior pathway of pivotal behavior tokens.

Pathway router. The pathway attention employs a sequence-adaptive pathway router to custom-tailor behavior pathway routes for users. The router generates a binary route $\mathcal{R}^l \in \{0, 1\}^N$ to determine whether a behavior token would be part of the behavior pathway or not. Each router takes the pre-order route $\mathcal{R}^{l-1}$ and user behavior tokens $\mathcal{Z}^{l-1} \in \mathbb{R}^{N\times d}$ of the block $l-1$ as its inputs. All elements in the route are initialized by 1 and are updated progressively in training.

Foremost, to suppress the potential disturbance to the model caused by the local drifted interest (Figure 1), it is crucial to incorporate the *global* information in the route generation. We apply the average pooling to all the preserved behavior tokens routed by $\mathcal{R}^{l-1}$, and produce the global sequential representation via a multilayer perceptrons (MLP) module. Then, we combine this global representation with the inputs and employ a residual connection to maintain the original input information. Finally, we feed them to another MLP layer to predict the probabilities of keeping or dropping the behavior tokens. The above procedure can be formulated as follows:

$$\begin{aligned} \mathcal{Z}^{l}_{\text{emb}} &= \mathcal{Z}^{l-1} + \mathcal{Z}^{l-1} \odot \text{MLP}\left(\frac{\sum_{i=1}^{N} \mathcal{R}^{l-1}_{i} \mathcal{Z}^{l-1}_{i}}{\sum_{i=1}^{N} \mathcal{R}^{l-1}_{i}}\right) \\ \boldsymbol{\pi} &= \text{Softmax}\left(\text{MLP}(\mathcal{Z}^{l}_{\text{emb}})\right) \in \mathbb{R}^{N\times 2}, \end{aligned} \tag{2}$$

where $\odot$ is the Hadamard product. For $t \in \{1, 2, \cdots, N\}$, we let $\boldsymbol{\pi}_t = [1-\alpha_t, \alpha_t]$, where the logit α_t denotes the probability that the t-th behavior token is kept for the behavior pathway.

Gumbel-Softmax sampling from $\boldsymbol{\pi}$ for router. Our goal is to generate the binary route from $\boldsymbol{\pi}$. However, sampling from $\boldsymbol{\pi}$ directly is non-differentiable, and it will impede the gradient-based training. Thus, we apply the Gumbel-Softmax [15] technique to such sample. Gumbel-Softmax is an effective way to approximate the original non-differentiable sample from a discrete distribution with a differentiable sample from a Gumbel-Softmax distribution. Instead of directly sampling a keep-or-drop decision $\widehat{\mathcal{R}}^{l}_{t}$ for the t-th behavior token from the distribution $\boldsymbol{\pi}_t$, we generate it as:

$$\widehat{\mathcal{R}}^{l}_{t} = \underset{j\in\{0,1\}}{\arg\max} \left(\log \pi_t(j) + G_t(j)\right), \tag{3}$$

where $G_t = -\log(-\log U_t)$ is a standard Gumbel distribution, and U_t is sampled i.i.d. from a uniform distribution Uniform$(0, 1)$. To remove the non-differentiable argmax operation in Eq 3, the Gumbel-Softmax uses the reparameterization trick [15] as a differentiable approximation to relax one-hot $\widehat{\mathcal{R}}^{l}_{t} \in \{0, 1\}$ to $v_t \in \mathbb{R}^2$:

$$v_t(j) = \frac{\exp((\log \pi_t(j) + G_t(j))/\tau)}{\sum_{i\in\{0,1\}} \exp((\log \pi_t(i) + G_t(i))/\tau)}, j \in \{0, 1\}, \tag{4}$$

where τ is the temperature parameter of the Softmax, which is commonly set to 1 in deep networks.

Hierarchical update strategy for router. The preliminary route $\widehat{\mathcal{R}}^{l}$, sampled from $\boldsymbol{\pi}$, is not a final decision. In our design, once a token fails to be routed in a certain block, it would permanently lose the privilege to be part of the behavior pathway in the following feed-forward procedure, constituting a hierarchical pathway router strategy. Thus finally we formulate the route $\mathcal{R}^{l}$ as the Hadamard product of $\widehat{\mathcal{R}}^{l}$ and the pre-order route $\mathcal{R}^{l-1}$ in the block $l-1$:

$$\mathcal{R}^{l} = \widehat{\mathcal{R}}^{l} \odot \mathcal{R}^{l-1}. \tag{5}$$

Multi-head pathway attention. The standard self-attention mechanism retrieves sequential characteristics exploiting all behavior tokens, making the behavior pathway overwhelmed by the trivial behaviors. In the new pathway attention, the pathway router would be firstly applied to the input behavior tokens to route information. The pathway router would not pare down the number of tokens, but only the interactions between the off-pathway and on-pathway tokens, as these off-pathway tokens may also convey contextual information.

Specifically, for the query, key, and value in the pathway attention: the query is routed by the pathway router, to prevent the pathway from being overwhelmed and to force the pathway attention to attend to the behavior pathway; the key and value are the original input behavior tokens, to ensure that the contextual information from off-pathway behavior tokens can be captured as well:

$$\begin{aligned} \mathcal{Q}_m, \mathcal{K}_m, \mathcal{V}_m &= (\mathcal{Z}^{l-1} \odot \mathcal{R}^{l}) W^{l}_{\mathcal{Q}_m}, \mathcal{Z}^{l-1} W^{l}_{\mathcal{K}_m}, \mathcal{Z}^{l-1} W^{l}_{\mathcal{V}_m} \\ \widehat{\mathcal{Z}}^{l}_{m} &= \text{Softmax}\left(\frac{\mathcal{Q}_m \mathcal{K}^{\mathrm{T}}_{m}}{\sqrt{d/h}}\right) \mathcal{V}^{l}_{m}, \end{aligned} \tag{6}$$

where $m \in \{1, 2, \cdots, h\}$ is the index of head in the multi-head self-attention; $W^{l}_{\mathcal{Q}_m}, W^{l}_{\mathcal{K}_m}, W^{l}_{\mathcal{V}_m} \in \mathbb{R}^{d\times d/h}$ are transformation matrices learned from data. Finally, the outputs $\left\{\widehat{\mathcal{Z}}^{l}_{m} \in \mathbb{R}^{N\times d/h}\right\}_{1\leq m\leq h}$

of multiple heads are concatenated into $\widehat{\mathcal{Z}}^l \in \mathbb{R}^{N\times d}$. We use $\widehat{\mathcal{Z}}^l, \mathcal{R}^l = \text{Path-MSA}\ (\mathcal{Z}^{l-1}, \mathcal{R}^{l-1})$ to summarize the above pathway attention. Its output is further transformed by Eq. (1) to form the final output of the l-th block $\mathcal{Z}^l \in \mathbb{R}^{N\times d}$.

Causality. In the prediction of the $(t+1)$-th behavior, only the first t observable behaviors should be taken into account. To avoid a future information leak and ensure causality, a look-ahead mask is employed and all links between $\mathcal{Q}_j$ and $\mathcal{K}_i$ $(j > i)$ are removed.

3.2 Prediction Layer and Training Objective

Prediction layer. In the final layer of our RETR, we calculate the user's preference score for the item k in the step $(t+1)$ in the context of user behavior history as $p\,(i_{t+1} = k \mid i_{1:t}) = e_k \cdot \mathcal{Z}_t^L$, where e_k is the representation of item k from item embedding matrix $\mathcal{E}_{\mathcal{I}}$, and $\mathcal{Z}_t^L$ is the output of the L-th RETR blocks at step t (L is the number of RETR blocks).

Training objective. We adopt the pairwise ranking loss to optimize the model parameters as:

$$\mathcal{L} = -\sum_{u\in\mathcal{U}}\sum_{t=1}^{n} \log\sigma(p(i_{t+1} \mid i_{1:t}) - p(i_{t+1}^{-} \mid i_{1:t})), \tag{7}$$

where we pair each ground-truth item i_{t+1} with a randomly sampled negative item i_{t+1}^{-}. In each epoch, we randomly generate one negative item for each time step in each sequence. This pairwise ranking loss is widely adopted in previous literature of sequential recommendation [16, 41].

4 Experiments

We extensively evaluate the proposed Recommender Transformer on seven real-world benchmarks.

Table 1: Statistics of the datasets.

Dataset	Users	Items	Actions
Beauty	22,363	12,101	19,8502
Sports	25,598	18,357	29,6337
Toys	19,412	11,924	16,7597
Yelp	30,431	20,033	31,6354
MovieLens-1M	6,040	3,416	1,000,000
Tmall	66,909	37,367	42,7797
Steam	334,730	13,047	3,700,000

Datasets. Here are descriptions of the seven datasets: (1) **Beauty, Sports, and Toys**: these three datasets are three subcategories obtained from Amazon review [22] datasets. (2) **Yelp** [2]: Yelp is a dataset for business recommendation. We only use the transaction records after January 1st, 2019. (3) **Tmall**: Tmall contains users' shopping logs on Tmall online shopping platform, which is from the IJCAI-15 competition. (4) **Steam** [16]: Steam dataset is collected from a large online video game distribution platform. This dataset includes 2,567,538 users, 15,474 games and 7,793,069 English reviews from October 2010 to January 2018. (5) **MovieLens**: this is a widely used benchmark dataset for evaluating collaborative filtering algorithms. The version we use is MovieLens-1M, which includes 1 million user ratings.

We group the interaction records by users or sessions for all datasets and sort them by the timestamps in ascending order. We follow the operation in SASRec [16] and split the historical sequence for each user into three parts: (1) the most recent behavior for testing, (2) the second most recent behavior for validation, and (3) all remaining behaviors for training. During testing, the input sequences contain training behaviors and validation behaviors. We filter less popular items and inactive users with fewer than five interaction records. The statistics of the above seven datasets are summarized in Table 1. Beauty, Sports, Toys and Yelp have fewer average actions per user and item. MovieLens, Tmall and Steam have more actions per item.

Evaluation metrics. Following the previous literature [38, 41, 16], we apply top-k Hit Ratio (HR@k), top-k Normalized Discounted Cumulative Gain (NDCG@k) and Mean Reciprocal Rank (MRR) for evaluation. We report HR@10, NDCG@10 and MRR of the results. Besides, following the standard strategy in SASRec [16], we pair the ground-truth item with 100 randomly sampled negative items that the user has not interacted with. All metrics are calculated according to the ranking of the items and we report the average score.

Baseline methods. We compare our approach with seven baseline methods. The PopRec is a simple baseline that ranks items according to their popularity. Besides, we select four latest state of-the-art transformer-based models: SASRec [16], BertRec [28], SASRec+ [35] and SMRec [6]. We notice that SASRec [16] applies the transformer [4] into the sequential recommendation task; BertRec [28] introduces a Cloze objective loss for sequential recommendation; SASRec+ [35] improves SASRec with time embedding, which is inspired by the positional embedding in transformer; SMRec [6] adopts SASRec as the backbone, and it is associated with a novel self-modulating attention. These recent transformer-based methods have achieved state-of-the-art performance in many benchmarks. Besides transformer-based methods, we also adopt an RNN-based model GRURec [14] and a CNN-based model Caser [29] as baseline models. The GRURec [14] applies GRU to model item sequences and the Caser [29] applies convolution operations for sequential recommendation. All baseline methods are configured using default parameters of the original paper or optimal parameters which can produce the best results in a grid search.

Implementation details. Our model is supervised by the pairwise rank loss in Equ 7, using the ADAM [17] optimizer with an initial learning rate of 0.001. Batch size is set to 512. The maximum number of training epochs for all methods is set to 200. All hyperparameters are tuned on the validation set. The training process is early stopped within 10 epochs. Our RETR has $L = 2$ layers, and each layer has $h = 4$ heads and d is set to be 256. The maximum sequence length N is set to 200 for MovieLens-1m and 100 for the other six datasets. All experiments are repeated three times, implemented in PyTorch [23], and conducted on a single NVIDIA 3090 GPU.

Table 2: Performance comparison of the baselines (PopRec, Caser [29], GRURec [14], BERTRec [28], SASRec [16], SASRec+ [35] and SMRec [6]) and our method on the Beauty, Sports, Toys, Yelp, MovieLens, Tmall and Steam datasets. We use HR@10, NDCG@10 and MRR as our metrics. For these three metrics, a higher value indicates a better performance.

Datasets	Meric	PopRec	Caser	GRU4Rec	BERT4Rec	SASRec	SASRec+	SMRec	RETR
Beauty	HR@10	0.3386	0.3942	0.4106	0.4739	0.4696	0.4798	0.4826	**0.5034**
	NDCG@10	0.1803	0.2512	0.2584	0.2975	0.3156	0.3261	0.3238	**0.3425**
	MRR	0.1558	0.2263	0.2308	0.2614	0.2852	0.2901	0.2918	**0.3067**
Sports	HR@10	0.3423	0.4014	0.4299	0.4722	0.4622	0.4776	0.4853	**0.5083**
	NDCG@10	0.1902	0.2390	0.2527	0.2775	0.2869	0.2987	0.3061	**0.3175**
	MRR	0.1660	0.2100	0.2191	0.2378	0.2520	0.2635	0.2665	**0.2768**
Toys	HR@10	0.3008	0.3540	0.3896	0.4493	0.4663	0.4729	0.4754	**0.5104**
	NDCG@10	0.1618	0.2183	0.2274	0.2698	0.3136	0.3183	0.3198	**0.3395**
	MRR	0.1430	0.1967	0.1973	0.2338	0.2842	0.2912	0.2910	**0.3048**
Yelp	HR@10	0.3609	0.6661	0.7265	0.7597	0.7373	0.7481	0.7548	**0.7730**
	NDCG@10	0.2007	0.4198	0.4375	0.4778	0.4642	0.4757	0.4789	**0.5136**
	MRR	0.1740	0.3595	0.3630	0.4026	0.3927	0.4011	0.4023	**0.4354**
MovieLen	HR@10	0.4329	0.7886	0.5581	0.8269	0.8233	0.8291	0.8302	**0.8467**
	NDCG@10	0.2377	0.5538	0.3381	0.5965	0.5936	0.6057	0.6079	**0.6351**
	MRR	0.1891	0.5178	0.3002	0.5614	0.5573	0.5649	0.5703	**0.5921**
Tmall	HR@10	0.2967	0.5943	0.6432	0.6196	0.6275	0.6392	0.6476	**0.7138**
	NDCG@10	0.1874	0.4513	0.5169	0.5025	0.5049	0.5169	0.5192	**0.6103**
	MRR	0.1723	0.4209	0.4975	0.4026	0.4804	0.4912	0.4934	**0.5822**
Steam	HR@10	0.7172	0.7874	0.4190	0.8656	0.8729	0.8773	0.8792	**0.9001**
	NDCG@10	0.4535	0.5381	0.2691	0.6283	0.6306	0.6397	0.6408	**0.6795**
	MRR	0.4102	0.5091	0.2402	0.5883	0.5925	0.6005	0.6011	**0.6326**

4.1 Main Results

The results of different methods on seven datasets are shown in Table 2. First, the non-sequential recommendation method PopRec performs worse than sequential recommendation methods, indicating that capturing the sequential pattern is essential for sequential recommendation. Next, we can easily find that transformer-based models, SASRec [16], BertRec [28], SASRec+ [35], and SMRec [6], achieve better performance than RNN-based model GRURec [14] and CNN-based model Caser [29] on most datasets. It indicates that the transformer has a better capacity to capture sequential characteristics. The transformer-based models use the attention mechanism to capture the interaction information between all previous user behaviors, indicating that the attention mechanism is crucial for sequential recommendation.

Our RETR can achieve state-of-the-art performance by a large margin on most datasets compared with all baseline models. Specifically, our RETR achieves competitive performance on the Beauty, Sports, Toys, and Yelp. These datasets are sparse, containing less action information. Thus they have lots of noisy logged information. By effectively capturing the behavior pathway, our RETR is not affected by this trivial behavior information and captures the most informative behavior representation to achieve better performance. Besides, our RETR can also achieve better performance on the Tmall, Steam, and MovieLens, which contain a large number of actions. Note that under the Tmall benchmark, RETR gains **7%** HR@10, **12%** NDCG@10 and **14%** MRR against the strongest baseline SMRec [6]. Under the Steam benchmark, our RETR also yields state-of-the-art performance, achieving **6%** MRR gains compared with the SASRec. For the MoveLens benchmarks, our RETR also achieves the best performance among all competing baselines. These results strongly indicate that our RETR is consistently effective on both sparse and dense datasets.

Unlike these transformer-based baselines, our RETR adopts a pathway router to detect the most informative behavior tokens dynamically. These informative behavior tokens are denoted as the behavior pathway. Our RETR uses the behavior pathway and discards the less informative behavior tokens from the query. The substantial performance gains of our RETR indicate that focusing more on the behavior pathway enables RETR to capture sequential characteristics more efficiently and effectively than the vanilla self-attention mechanism, which considers all previous user behaviors.

4.2 Ablation Study

Effectiveness of each model component. In the left column of Table 3, we analyze the efficacy of each component in RETR on the Beauty dataset and have the following observations. First, we remove the pathway router module and randomly choose whether it can be maintained or dropped for each input behavior token. Removing the pathway router decreases the prediction performance a lot (MRR: $0.3067 \rightarrow 0.2703$), showing the necessity of learning behavior pathway effectively based on a data-dependent module. Second, discarding the hierarchical update strategy for the behavior pathway also decreases the prediction performance, suggesting that this strategy is crucial for RETR to get a more accurate behavior pathway. Notably, SASRec is the particular case of removing the pathway router, making routing decision 1 for all behaviors. Our RETR significantly outperforms the SASRec, indicating that capturing the behavior pathway effectively interacts with all previous behaviors.

Number of blocks. In the right column of Table 3, we adjust the number of blocks for RETR on Beauty. We find that the performance first increases rapidly with the growth of the block number and achieves the best performance at $L = 2$. We perform a similar grid search on other datasets.

Table 3: Ablation study of (**Left**) the effectiveness of each model component and (**Right**) the number of blocks for each RETR block. Experiments are conducted on the Beauty Dataset.

Model	MRR	Model (# number of blocks)	MRR
RETR	**0.3067**	RETR ($L = 1$)	0.2966
RETR w/o Pathway Router	0.2793	RETR ($L = 2$)	**0.3067**
RETR w/o hierarchical update	0.2957	RETR ($L = 3$)	0.3058
SASRec	0.2852	**RETR** ($L = 4$)	0.3052

Table 4: Ablation study of (**Left**) the effectiveness of different temperatures; Comparison Parameters and GFLOPs (**Right**). All ablation study experiments are conducted on the Yelp Dataset.

Model (temperature)	MRR	Model	Parameters (M)	GFLOPs	MRR
RETR ($\tau = 0.4$)	0.4312	RETR	5.021	9.558	**0.4354**
RETR ($\tau = 0.8$)	**0.4354**	SASRec [16]	4.916	9.552	0.3927
RETR ($\tau = 1$)	0.4292	SASRec+ [35]	5.133	9.942	0.4011
RETR ($\tau = 2$)	0.4183	SMRec [6]	5.173	9.864	0.4023

Effectiveness of temperature. In the left column of Table 4, we analyze the efficacy of different temperatures for Gumbel-Softmax sampling in RETR on the Yelp dataset. We observe that the performance first increases rapidly with the growth of the temperature and achieves the best performance when $\tau = 0.8$, while the performance degenerates a lot when $\tau > 1$. The temperature τ softens the softmax with $\tau > 1$. However, when $\tau \to \infty$, the Gumbel-Softmax distribution $p_\tau(y_t) \to 0.5$ becomes more smooth, leading to the maximum uncertainty. To make the sampling results more convincing, we apply the temperature calibration $\tau < 1$ during training to avoid overconfident predictions. These results show that Gumble-Softmax sampling with lower temperature ($\tau < 1$) avoids overconfident predictions, leading to better performance.

Evaluation on efficiency. The efficiency is compared between SASRec, SASRec+ and SMRec on the Yelp dataset. The computation cost is measured with gigabit floating-point operations (GFLOPs) on the self-attention module with position encoding. Meanwhile, the model scale measured with parameters is also presented. As shown in Table 4, our RETR has almost the same number of parameters or GFLOPs, compared with SASRec, indicating that our pathway router is a light-weight module. Our pathway attention does not bring more costs. It's worth noticing that the parameter scales and GLOPs of other competing transformers (apart from SASRec) are larger than RETR, but our RETR achieves higher performance. This result shows that our RETR is more efficient and effective than other competing transformer-based models.

4.3 Visualization

Setups. We also provide qualitative visualizations for our RETR, SASRec [16] and SMRec [6]. Technically, we use the GradCAM [26] to generate behavior heat maps of the output of the last layer in each model. A random user's historical behaviors in the Steam dataset are given in sequential order (row-wise). The last item is an RPG game *The Dwarves* in the test set. We provide attention heatmaps on the last ten positions at the last ten time steps in Figure 3. As shown in Figure 3, two main behavior pathway characteristics exist in this sequence: (1) **Casual Behavior Pathway:** RPG games are randomly clicked by the user, while the user has a continuing interest in RPGs. (2) **Drifted Behavior Pathway:** The user has recently been interested in adventure games but chooses the RPG at last. We hope that a sequential recommendation model can figure out the exact behavior pathway.

Visualzation results. As shown in Figure 3, we can find that the baseline method can not mitigate the drifted behavior pathway and capture the casual pathway adaptively. It seems that these models can be misled by the recent *Correlated Behavior Pathway*. In the third row of Figure 3, the SASRec focuses more on the recent adventure games and overlooks the user's casual interest in RPGs, which indicates the SASRec is easy to be overwhelmed by the drifted behavior pathway. The SASRec cannot effectively mitigate the drifted behavior pathway and capture the casual behavior pathway. For SMRec, even though it adopts a time weight mechanism. However, only using the time weight mechanism can not help it capture the practical behavior pathway. On the contrary, as shown in the second row of Figure 3, our RETR can capture the casual behavior pathway precisely with higher scores, and avoid being misled by the drifted behavior pathway. This visualization results strongly show that our RETR can capture the behavior pathway dynamically for each user.

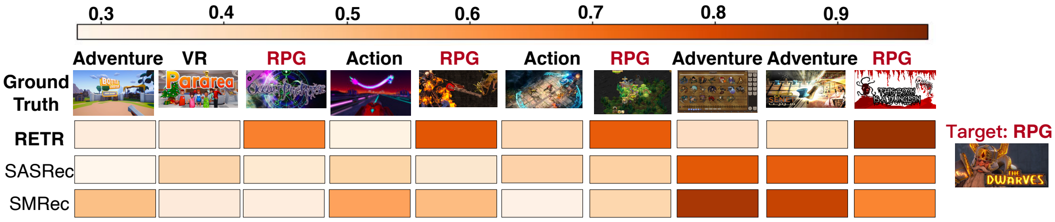

Figure 3: Illustration of how RETR, SASRec and SMRec differs on utilizing the historical behaviors of a random User in Steam Dataset. We provide the visualization of behavior heat maps for RETR, SASRec and SMRec of a random user in Steam dataset.

5 Conclusion

A sequential recommender is designed to make accurate recommendations based on users' historical behaviors. The sequential recommendation system has benefited many practical applications such as online advertising and social media. However, the users' behaviors are dynamic and come in a continually evolving manner due to various reasons. We conclude these sequential characteristics as the behavior pathway. Previous models cannot capture the behavior pathway dynamically. We propose the Recommender Transformer (RETR) with a novel pathway attention mechanism to tackle these challenges. The pathway attention develops a pathway router to dynamically get the behavior pathway for each user and capture the evolving patterns. Our RETR achieves state-of-the-art performance on seven real-world datasets for sequential recommendation.

References

[1] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering*, 17(6):734–749, 2005.

[2] Nabiha Asghar. Yelp dataset challenge: Review rating prediction. *arXiv preprint arXiv:1605.05362*, 2016.

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[4] Prateep Bhattacharjee and Sukhendu Das. Temporal coherency based criteria for predicting video frames using deep multi-stage generative adversarial networks. In *NeurIPS*, pages 4268–4277, 2017.

[5] Shuqing Bian, Wayne Xin Zhao, Kun Zhou, Jing Cai, Yancheng He, Cunxiang Yin, and Ji-Rong Wen. Contrastive curriculum learning for sequential user behavior modeling via data augmentation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3737–3746, 2021.

[6] Chao Chen, Haoyu Geng, Nianzu Yang, Junchi Yan, Daiyue Xue, Jianping Yu, and Xiaokang Yang. Learning self-modulating attention in continuous time space with applications to sequential recommendation. In *International Conference on Machine Learning*, pages 1606–1616. PMLR, 2021.

[7] Qiang Cui, Shu Wu, Qiang Liu, Wen Zhong, and Liang Wang. Mv-rnn: A multi-view recurrent neural network for sequential recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 32(2):317–331, 2018.

[8] Jeff Dean. Introducing pathways: A nextgeneration ai architecture. *Google Blog*, 2021.

[9] Hui Fang, Danning Zhang, Yiheng Shu, and Guibing Guo. Deep learning for sequential recommendation: Algorithms, influential factors, and evaluations. *ACM Transactions on Information Systems (TOIS)*, 39(1):1–42, 2020.

[10] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv preprint arXiv:2101.03961*, 2021.

[11] Ruining He, Chen Fang, Zhaowen Wang, and Julian McAuley. Vista: A visually, socially, and temporally-aware model for artistic recommendation. In *Proceedings of the 10th acm conference on recommender systems*, pages 309–316, 2016.

[12] Ruining He and Julian McAuley. Fusing similarity models with markov chains for sparse sequential recommendation. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 191–200. IEEE, 2016.

[13] Jonathan L Herlocker, Joseph A Konstan, Al Borchers, and John Riedl. An algorithmic framework for performing collaborative filtering. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 230–237, 1999.

[14] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*, 2015.

[15] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[16] Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 197–206. IEEE, 2018.

[17] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[18] Yehuda Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 426–434, 2008.

[19] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[20] Yuli Liu, Christian Walder, and Lexing Xie. Determinantal point process likelihoods for sequential recommendation. *arXiv preprint arXiv:2204.11562*, 2022.

[21] Jie Lu, Dianshuang Wu, Mingsong Mao, Wei Wang, and Guangquan Zhang. Recommender system application developments: a survey. *Decision Support Systems*, 74:12–32, 2015.

[22] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52, 2015.

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[24] 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*, pages 2685–2692, 2020.

[25] Pengjie Ren, Zhumin Chen, Jing Li, Zhaochun Ren, Jun Ma, and Maarten De Rijke. Repeatnet: A repeat aware neural recommendation machine for session-based recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4806–4813, 2019.

[26] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[27] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[28] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1441–1450, 2019.

[29] Jiaxi Tang and Ke Wang. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 565–573, 2018.

[30] Pengfei Wang, Jiafeng Guo, Yanyan Lan, Jun Xu, Shengxian Wan, and Xueqi Cheng. Learning hierarchical representation model for nextbasket recommendation. In *Proceedings of the 38th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 403–412, 2015.

[31] Chao-Yuan Wu, Amr Ahmed, Alex Beutel, Alexander J Smola, and How Jing. Recurrent recommender networks. In *Proceedings of the tenth ACM international conference on web search and data mining*, pages 495–503, 2017.

[32] Liwei Wu, Shuqing Li, Cho-Jui Hsieh, and James Sharpnack. Sse-pt: Sequential recommendation via personalized transformer. In *Fourteenth ACM Conference on Recommender Systems*, pages 328–337, 2020.

[33] Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. Session-based recommendation with graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 346–353, 2019.

[34] Chengfeng Xu, Pengpeng Zhao, Yanchi Liu, Victor S Sheng, Jiajie Xu, Fuzhen Zhuang, Junhua Fang, and Xiaofang Zhou. Graph contextualized self-attention network for session-based recommendation. In *IJCAI*, pages 3940–3946, 2019.

[35] Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. Self-attention with functional time representation learning. *Advances in neural information processing systems*, 32, 2019.

[36] An Yan, Shuo Cheng, Wang-Cheng Kang, Mengting Wan, and Julian McAuley. Cosrec: 2d convolutional neural networks for sequential recommendation. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2173–2176, 2019.

[37] Zeenat F Zaidi. Gender differences in human brain: a review. *The open anatomy journal*, 2(1), 2010.

[38] Wayne Xin Zhao, Junhua Chen, Pengfei Wang, Qi Gu, and Ji-Rong Wen. Revisiting alternative experimental settings for evaluating top-n item recommendation algorithms. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2329–2332, 2020.

[39] Wayne Xin Zhao, Shanlei Mu, Yupeng Hou, Zihan Lin, Yushuo Chen, Xingyu Pan, Kaiyuan Li, Yujie Lu, Hui Wang, Changxin Tian, et al. Recbole: Towards a unified, comprehensive and efficient framework for recommendation algorithms. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4653–4664, 2021.

[40] 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*, pages 1059–1068, 2018.

[41] Kun Zhou, Hui Yu, Wayne Xin Zhao, and Ji-Rong Wen. Filter-enhanced mlp is all you need for sequential recommendation. In *Proceedings of the ACM Web Conference 2022*, pages 2388–2399, 2022.