

Uncovering Selective State Space Model's Capabilities in Lifelong Sequential Recommendation

Jiyuan Yang
 jiyuan.yang@mail.sdu.edu.cn
 Shandong University
 Qingdao, China

Hanbing Wang
 hanbingwang01@gmail.com
 Michigan State University
 East Lansing, USA

Zhaochun Ren
 z.ren@liacs.leidenuniv.nl
 Leiden University
 Leiden, The Netherlands

Zhumin Chen
 chenzhumin@sdu.edu.cn
 Shandong University
 Qingdao, China

Yuanzi Li
 liyuanzi@mail.sdu.edu.cn
 Shandong University
 Qingdao, China

Muyang Ma
 muyang0331@gmail.com
 Shandong University
 Qingdao, China

Mengqi Zhang
 mengqi.zhang@sdu.edu.cn
 Shandong University
 Qingdao, China

Jingyu Zhao
 jingyu.zhao@mail.sdu.edu.cn
 Shandong University
 Qingdao, China

Jun Ma
 majun@mail.sdu.edu.cn
 Shandong University
 Qingdao, China

Xin Xin
 xinxin@sdu.edu.cn
 Shandong University
 Qingdao, China

Pengjie Ren*
 renpengjie@sdu.edu.cn
 Shandong University
 Qingdao, China

Codes and data are available at <https://github.com/nancheng58/RecMamba>.

KEYWORDS

Sequential Recommendation, Long-term Recommendation, State Space Models

ACM Reference Format:

Jiyuan Yang, Yuanzi Li, Jingyu Zhao, Hanbing Wang, Muyang Ma, Jun Ma, Zhaochun Ren, Mengqi Zhang, Xin Xin, Zhumin Chen, and Pengjie Ren. 2018. Uncovering Selective State Space Model's Capabilities in Lifelong Sequential Recommendation. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 5 pages. <https://doi.org/XXXXXX.XXXXXXX>

1 INTRODUCTION

Over the past decade, recommender systems have been widely applied in various online services, e.g., online shopping [10], video or music platforms [29], news recommendation [20], etc., with the primary aim of offering users the most compelling items through the modeling of user interests. Sequential recommendation, a burgeoning field within recommendation systems, focuses on extracting dynamic user interests from their sequential interactions. The key to sequential recommendation is to capture user interests by modeling user interaction sequences. Early studies in sequential recommendations utilize Markov Chains [6, 21] for modeling user interaction sequences. Over the past few years, plenty of deep learning-based sequential recommendation models have been proposed, including recurrent neural networks (RNN) [8, 9], convolutional neural networks (CNN) [25, 29], graph neural networks (GNN) [28, 30], attention-based methods [12, 13, 24] and Transformer-based methods [2].

ABSTRACT

Sequential Recommenders have been widely applied in various online services, aiming to model users' dynamic interests from their sequential interactions. With users increasingly engaging with online platforms, vast amounts of lifelong user behavioral sequences have been generated. However, existing sequential recommender models often struggle to handle such lifelong sequences. The primary challenges stem from computational complexity and the ability to capture long-range dependencies within the sequence.

Recently, a state space model featuring a selective mechanism (i.e., Mamba) has emerged. In this work, we investigate the performance of Mamba for lifelong sequential recommendation (i.e., $\text{length} \geq 2k$). More specifically, we leverage the Mamba block to model lifelong user sequences selectively. We conduct extensive experiments to evaluate the performance of representative sequential recommendation models in the setting of lifelong sequences. Experiments on two real-world datasets demonstrate the superiority of Mamba. We found that RecMamba achieves performance comparable to the representative model while significantly reducing training duration by approximately 70% and memory costs by 80%.

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 978-1-4503-XXXX-X/18/06
<https://doi.org/XXXXXX.XXXXXXX>

However, the aforementioned methods are constrained in their ability to model long user interaction sequences for several reasons. For instance, Markov Chains-based methods [6, 21] assume that the next interaction is dependent only on the previous interaction (or a few preceding ones), making it challenging to characterize long-range item transitions. GRU4Rec [7] may not be optimal for lifelong sequence recommendation due to information forgetting and inherent vanishing gradients. SASRec [12] encounters challenges in modeling long sequences due to the huge memory requirements and the quadratic complexity in the input length of self-attention. Currently, the majority of research only allows their models to accept about 200 user interaction records [14].

The primary challenge in modeling lifelong user interaction sequences revolves around two key aspects: i) How to efficiently handle extended user behavior sequences, while addressing concerns like data sparsity and computational complexity? and ii) How to capture long-range item transition dependencies within the sequence? Recently, Mamba [4], a novel state space model featuring a selective mechanism has emerged, with the aim of achieving the modeling power of Transformer [26] while scaling linearly in sequence length. Some works leverage Mamba for various downstream tasks, such as visual tasks [17, 32], medical applications [18], recommendation [15], and so on. In this scenario, we are curious about examining the performance of both early models and Mamba when modeling long (i.e., length $\geq 2k$) sequences.

In this work, we investigate the performance of Mamba in sequential recommendation, especially for lifelong user behavior sequences. Specifically, we introduce RecMamba, a sequential recommendation framework that utilizes the Mamba block to model user preferences over time. By integrating the capabilities of Mamba into our framework, we aim to improve recommendation performance and better accommodate evolving user preferences in lifelong sequential recommendation scenarios. We conduct experiments to assess the effectiveness of prominent recommendation models for long user sequence (i.e., length $\geq 2k$) scenarios on two real-world datasets. We found that (1) the performance of RecMamba improves as the length increases on two real-world datasets; and (2) RecMamba achieves superior performance than representative sequential models. More specifically, RecMamba achieves comparable performance with the representative model SASRec while greatly reducing about 70% training duration and 80% memory costs.

2 LEVERAGING MAMBA FOR SEQUENTIAL RECOMMENDATION

State Space Models (SSMs) [5, 23] are a class of models for sequence modeling. Recently, a novel space state model incorporating the selective mechanism (i.e., Mamba) [4] has become a hot topic. Compared with prior SSMs, Mamba introduces a data-dependence selection mechanism and utilizes a parallel algorithm optimized for hardware in recurrent mode, enabling effective sequence modeling, particularly for long sequences.

Lifelong sequential recommendation is a common task involving long-sequence modeling. Mamba4Rec [15] is the pioneer in utilizing Mamba for efficient sequential recommendation. Compared to SASRec [12], the principal architectural modification in Mamba4Rec

Table 1: Dataset statistics.

Dataset	KuaiRand	LFM-1b
#users	27,285	120,322
#items	32,038,725	31,634,450
#interactions	322,278,385	1,088,161,692
#avg.len	11,811	9,043

involves substituting the self-attention block with the Mamba block. In contrast to Mamba4Rec, our approach, Rec-Mamba, replaces the Transformer layer with the Mamba block, thereby improving the efficiency of processing lifelong sequences.

3 EXPERIMENTS

3.1 Methods for Comparison

To demonstrate the effectiveness of RecMamba on lifelong sequential recommendation, we compare it with a range of representative recommender models based on Transformers, RNNs, and Linear Transformer:

- **SASRec** [12] is a method that employs the traditional self-attention block, specifically multi-head attention, to generate sequence representations.
- **GRU4Rec** [1] is a pioneering method leveraging RNNs to model user action sequences for session-based recommendation, treating each user’s feedback sequence as a session.
- **Linrec** [16] adopts a linear Transformer architecture that utilizes L2 norm to approximate softmax fitting.

3.2 Datasets

In recent years, plenty of long-term or lifelong recommendation works have emerged [11, 19]. However, the sequence lengths discussed in these studies typically range from 10 to 200. We think that such sequences do not aptly represent a ‘lifelong’ user interests. For instance, in the micro-video scenario, users watch dozens or hundreds of videos daily. Hence, conducting experiments assessing model performance on extended sequences is imperative.

To this end, we choose two real-world datasets with longer user interaction sequences KuaiRand¹ [3] and LFM-1b² [22] to conduct the experiments. All items in the sequence are arranged based on chronological order. We filter out users and items with less than 3 recorded interactions.

Table 1 summarizes the statistics of the two datasets.

- **KuaiRand.** KuaiRand is a large-scale dataset derived from the famous micro-video platform Kuaishou. It uniquely features millions of randomly exposed items inserted in the standard feeds, ensuring unbiased sequential recommendations. This dataset contains 27,285 users and 32,038,725 videos and the average interaction length is 11,811.
- **LFM-1b.** This is a large-scale dataset conducted by the music platform Last.FM, which contains more than one billion music-listening interactions covering 31,634,450 music and a total of 120,322 user clicks. The average interaction length is 9,043.

¹<https://kuairand.com/>

²<http://www.cp.jku.at/datasets/LFM-1b/>

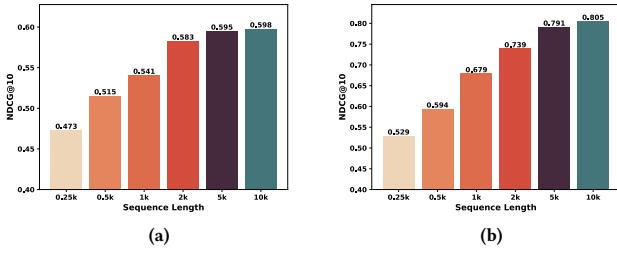


Figure 1: Performance of the user interests modeling about sequence length on KuaiRand (a) and Tracks (b) datasets.

3.3 Implementation Details

We adopted the hyperparameter settings used for sequence recommendation due to the absence of prior work on sequence recommendation with lengths exceeding 2k. The trainable parameters were initialized using Xavier normal distribution. We employed the Adam optimizer for each model. For all models, we used the batch size of 256 for sequence length of 2k, while for the length of 5k, we used a batch size of 256 for RecMamba, Linrec, GRU4Rec, and a batch size of 32 for SASRec to avoid Out-of-Memory. The setting of the learning rate is 0.0004 and 0.0002, respectively. The hidden layer size is set to 50. We pad the sequence with an additional padding item if the sequence length is less than 2k. All models were trained 500 epochs on NVIDIA A800 80G.

3.4 Evaluation Protocols

Following existing sequential recommendation studies [12, 24, 27, 31], we split data by the leave-one-out strategy. For an user u , the interaction sequence $S_u = [v_1^u, v_2^u, \dots, v_{n_u}^u]$, we use $([v_1^u, \dots, v_{n_u-3}^u], v_{n_u-2}^u)$ for training, $([v_1^u, \dots, v_{n_u-2}^u], v_{n_u-1}^u)$ for validation, and $([v_1^u, \dots, v_{n_u-1}^u], v_{n_u}^u)$ for testing, where n is the length of S_u and v_t^u is the item user interact with at the t -th timestamp. We adopt Recall and NDCG metrics to evaluate the ranking performance. We report the average results of three experiments.

4 EXPERIMENTS RESULTS

In this section, we aim to answer the following research questions in terms of lifelong sequential recommendation:

- **RQ1:** How does longer sequence length contribute to the effectiveness of sequence recommendation?
- **RQ2:** How do different recommenders perform in modeling lifelong sequences?
- **RQ3:** How is the efficiency of different recommenders?

4.1 Lengths Comparison (RQ1)

To demonstrate the performance of modeling different sequence lengths, we conduct experiments to evaluate the performance of Rec-Mamaba on two real-world datasets KuaiRand and LFM-1b. We report the main experimental results in Figure 1.

Overall, with RecMamba, we observe that longer sequences consistently outperform shorter sequences on both the KuaiRand and

LFM-1b datasets, indicating that introducing longer sequences benefits modeling user interest. The performance improvement with longer sequences can be attributed to their ability to capture more comprehensive and meaningful patterns, dependencies, and user preferences, enabling the recommender systems to make more accurate predictions and generate more relevant recommendations. In summary, the experimental results provide strong evidence for the effectiveness of lifelong sequences in improving recommendation performance. The findings emphasize the importance of considering longer sequences in sequence recommendation models, as they can capture richer interaction information and lead to more accurate recommendations.

4.2 Overall Performance (RQ2)

In this subsection, we conduct a comprehensive evaluation to verify the sequence modeling performance of RecMamba. We compare it with different representative recommendation models including the RNN-based model GRU4Rec, attention-based model SASRec, and linear attention-based model LinRec. The results are reported in Table 4.2.

We can see that RecMamba and SASRec significantly outperform both LinRec and GRU4Rec with max sequence lengths of 2k and 5k on two datasets. The reason may be that GRU4Rec has not fully addressed the limitations of the RNN-based architecture, such as information forgetting and intrinsic vanishing gradients issues. Besides, the LinRec modeling sequence using approximate softmax could not fit the long-range dependency of item transition. This highlights the superior ability of RecMamba and SASRec in the lifelong sequential recommendation.

Compared with SASRec, RecMamba achieves suboptimal performance on sequences of length 2k, whereas it outperforms SASRec on sequences of length 5k in most cases. This indicates that RecMamba is capable of effectively modeling longer user interest sequences, potentially due to its ability to select relevant items for modeling user interests.

4.3 Efficiency Comparison (RQ3)

In this subsection, we conduct experiments to compare the efficiency between RecMamba and other representative sequential recommenders on the lifelong sequential recommendation. More specifically, we delve into crucial efficiency metrics, encompassing GPU memory consumption, training duration, and inference time. We used the batch size of 256 for sequence length of 2k for all models. To investigate the model efficiency, we evaluate the computational cost, including GPU memory consumption, training duration, and inference time as shown in Table 3.

We can observe that RecMamba notably reduces GPU memory footprint and significantly slashes both inference and training times. Compared with SASRec on the LFM-1b(2k) dataset, RecMamba reduces about 73% training duration, 61% inference time, and 80% memory costs. We can observe similar results on the KuaiRand(2k) dataset. For the two datasets of the 5k version, SASRec has encountered an out-of-memory (OOM) issue and RecMamba achieves better efficiency compared with LinRec. These results demonstrate RecMamba's prowess in handling lifelong sequence recommendation tasks with remarkable efficiency.

Table 2: Overall performance comparison of different methods on KuaiRand and LFM-1b datasets. Boldface indicates the best result. Underscore indicates the suboptimal results. 2k and 5k denote the max sequence length.

Datasets	Method	Recall@5	NDCG@5	Recall@10	NDCG@10	Recall@20	NDCG@20
KuaiRand(2k)	GRU4Rec	0.4062	0.3080	0.4439	0.3184	0.4617	0.3230
	LinRec	0.4984	0.3774	0.6167	0.4152	0.7242	0.4422
	SASRec	0.7143	0.5875	0.7857	0.6126	0.8361	0.6254
	RecMamba	<u>0.6907</u>	<u>0.5584</u>	<u>0.7652</u>	<u>0.5828</u>	<u>0.8172</u>	<u>0.5960</u>
LFM-1b(2k)	GRU4Rec	0.3119	0.2300	0.3597	0.2401	0.3864	0.2469
	LinRec	0.6163	0.5052	0.6955	0.5327	0.7668	0.5501
	SASRec	0.8112	0.7448	0.8388	0.7537	0.8598	0.7591
	RecMamba	<u>0.8019</u>	<u>0.7487</u>	<u>0.8319</u>	<u>0.7393</u>	<u>0.8570</u>	<u>0.7456</u>
KuaiRand(5k)	GRU4Rec	0.2593	0.2200	0.2957	0.3080	0.2318	0.2350
	LinRec	0.5163	0.3965	0.6278	0.4327	0.7287	0.4582
	SASRec	<u>0.6422</u>	<u>0.5224</u>	<u>0.7370</u>	<u>0.5468</u>	<u>0.8042</u>	<u>0.5639</u>
	RecMamba	0.7004	0.5713	0.7733	0.5951	0.8253	0.6082
LFM-1b(5k)	GRU4Rec	0.3009	0.2113	0.3800	0.2366	0.4055	0.2432
	LinRec	0.6362	0.5296	0.7135	0.5543	0.7776	0.5717
	SASRec	0.8542	0.7820	0.8794	0.7905	0.8969	0.7950
	RecMamba	<u>0.8485</u>	0.7823	<u>0.8727</u>	0.7908	<u>0.8904</u>	0.7953

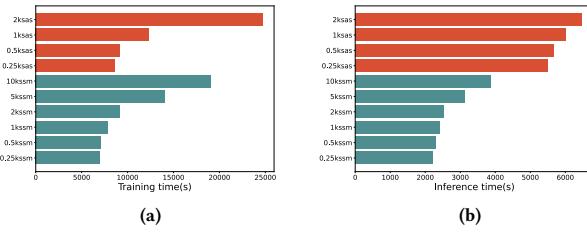


Figure 2: Training time (a) and Inference time (b) comparison on LFM-1b dataset.

Besides, we investigate the efficiency comparison for modeling different lengths. Figure 2 shows the training time and inference time comparison between RecMamba and SASRec on the LFM-1b dataset. Compared with SASRec, RecMamba demonstrates a significant efficiency advantage in modeling sequences of any length reported, including both training and inference time. This advantage becomes more pronounced as the sequence length increases.

To conclude, the experimental findings provide compelling evidence of RecMamba’s superior efficiency as a sequence recommendation framework. Its ability to efficiently reduce GPU memory consumption and optimize both training and inference times positions RecMamba as an exceedingly promising solution for building efficient and scalable recommendation systems. This superiority is evident not only in terms of performance but also in terms of efficiency, including GPU memory consumption and inference time. This implies that RecMamba achieves better results on lifelong sequences with fewer computational resources, making it more efficient and cost-effective for deployment in recommendation systems.

Table 3: Efficiency comparison. Boldface indicates the best result. 2k and 5k denote the max sequence length. \times denotes the OOM issue.

Datasets	Model	GPU memory(GB)	Time cost(s/epoch)	
			Training	Evaluation
KuaiRand(2k)	LinRec	11.85G	22.49s	54.72s
	SASRec	37.86G	49.42s	64.81s
	RecMamba	8.36G	18.21s	25.39s
KuaiRand(5K)	LinRec	20.02G	39.54s	61.4s
	SASRec	\times	\times	\times
	RecMamba	14.68G	28.09s	31.35s
LFM-1b(2k)	LinRec	10.69G	18.19s	52.95s
	SASRec	39.85G	41.24s	59.95s
	RecMamba	7.60G	11.75s	23.12s
LFM-1b(5k)	LinRec	19.83G	32.76s	56.65s
	SASRec	\times	\times	\times
	RecMamba	14.46G	17.62s	26.54s

5 CONCLUSION AND FUTURE WORK

In this paper, we have investigated how the selective state space model (i.e., Mamba) performs lifelong sequential recommendation. More specifically, we have leveraged Mamba to adopt the sequential recommendation task. We have conducted experiments to verify the performance and efficiency of representative recommendation models for longer user sequences (i.e., $\text{length} \geq 2\text{k}$) scenarios. Extensive experiments and analysis on two real-world datasets have demonstrated that Mamba achieves superior performance than representative sequential models.

In future work, we plan to refine the Mamba to better perform the subfield of recommendation. For instance, leveraging Mamba for multi-behavior recommendations would be a potential research direction. Additionally, how to effectively deal with longer side information (e.g., title and description of items) would also be an

interesting research direction. We hope the findings in this work could facilitate future research on using Mamba for sequential recommendations.

REFERENCES

- [1] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* (2014).
- [2] Gabriel de Souza Pereira Moreira, Sarai Rabhi, Jeong Min Lee, Ronay Ak, and Even Oldridge. 2021. Transformers4rec: Bridging the gap between nlp and sequential/session-based recommendation. In *Proceedings of the 15th ACM conference on recommender systems*. 143–153.
- [3] Chongming Gao, Shijun Li, Yuan Zhang, Jiawei Chen, Biao Li, Wenqiang Lei, Peng Jiang, and Xiangnan He. 2022. KuaiRand: An Unbiased Sequential Recommendation Dataset with Randomly Exposed Videos. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (Atlanta, GA, USA) (CIKM '22). Association for Computing Machinery, New York, NY, USA, 3953–3957. <https://doi.org/10.1145/3511808.3557624>
- [4] Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752* (2023).
- [5] Albert Gu, Karan Goel, and Christopher Ré. 2021. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396* (2021).
- [6] Ruining He, Chen Fang, Zhaowen Wang, and Julian McAuley. 2016. Vista: A visually, socially, and temporally-aware model for artistic recommendation. In *Proceedings of the 10th ACM conference on recommender systems*. 309–316.
- [7] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939* (2015).
- [8] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2016. Session-based Recommendations with Recurrent Neural Networks. In *ICLR (Poster)*.
- [9] Balázs Hidasi, Massimo Quadrana, Alexandros Karatzoglou, and Domonkos Tikk. 2016. Parallel recurrent neural network architectures for feature-rich session-based recommendations. In *Proceedings of the 10th ACM conference on recommender systems*. 241–248.
- [10] Yujing Hu, Qing Da, Anxiang Zeng, Yang Yu, and Yinghui Xu. 2018. Reinforcement learning to rank in e-commerce search engine: Formalization, analysis, and application. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 368–377.
- [11] Jin Huang, Wayne Xin Zhao, Hongjian Dou, Ji-Rong Wen, and Edward Y Chang. 2018. Improving sequential recommendation with knowledge-enhanced memory networks. In *The 41st international ACM SIGIR conference on research & development in information retrieval*. 505–514.
- [12] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 197–206.
- [13] Jing Li, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tao Lian, and Jun Ma. 2017. Neural attentive session-based recommendation. In *CIKM*. 1419–1428.
- [14] Jianghao Lin, Rong Shan, Chenxu Zhu, Kounianhua Du, Bo Chen, Shigang Quan, Ruiming Tang, Yong Yu, and Weinan Zhang. 2023. Rella: Retrieval-enhanced large language models for lifelong sequential behavior comprehension in recommendation. *arXiv preprint arXiv:2308.11131* (2023).
- [15] Chengkai Liu, Jianghao Lin, Jianling Wang, Hanzhou Liu, and James Caverlee. 2024. Mamba4Rec: Towards Efficient Sequential Recommendation with Selective State Space Models. *arXiv preprint arXiv:2403.03900* (2024).
- [16] Langming Liu, Liu Cai, Chi Zhang, Xiangyu Zhao, Jingtong Gao, Wanyu Wang, Yifu Lv, Wenqi Fan, Yiqi Wang, Ming He, Zitao Liu, and Qing Li. 2023. LinRec: Linear Attention Mechanism for Long-term Sequential Recommender Systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (<conf-loc>, <city>Taipei</city>, <country>Taiwan</country>, </conf-loc>) (SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 289–299. <https://doi.org/10.1145/3539618.3591717>
- [17] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. 2024. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166* (2024).
- [18] Jun Ma, Feifei Li, and Bo Wang. 2024. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722* (2024).
- [19] Qi Pi, Weijie Bian, Guorui Zhou, Xiaoqiang Zhu, and Kun Gai. 2019. Practice on Long Sequential User Behavior Modeling for Click-Through Rate Prediction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) (KDD '19). Association for Computing Machinery, New York, NY, USA, 2671–2679. <https://doi.org/10.1145/3292500.3330666>
- [20] Tao Qi, Fangzhao Wu, Chuhuan Wu, Yongfeng Huang, and Xing Xie. 2020. Privacy-Preserving News Recommendation Model Learning. In *EMNLP (Findings) (Findings of ACL, Vol. EMNLP 2020)*. Association for Computational Linguistics, 1423–1432.
- [21] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized Markov chains for next-basket recommendation. In *WWW*. ACM, 811–820.
- [22] Markus Schedl. 2016. The lfm-1b dataset for music retrieval and recommendation. In *Proceedings of the 2016 ACM on international conference on multimedia retrieval*. 103–110.
- [23] Jimmy TH Smith, Andrew Warrington, and Scott W Linderman. 2022. Simplified state space layers for sequence modeling. *arXiv preprint arXiv:2208.04933* (2022).
- [24] 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 *CIKM*. 1441–1450.
- [25] Jiaxi Tang and Ke Wang. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *WSDM*. ACM, 565–573.
- [26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NIPS*. 5998–6008.
- [27] Hanbing Wang, Xiaorui Liu, Wenqi Fan, Xiangyu Zhao, Venkataramana Kini, Devendra Yadav, Fei Wang, Zhen Wen, Jiliang Tang, and Hui Liu. 2024. Rethinking Large Language Model Architectures for Sequential Recommendations. *arXiv preprint arXiv:2402.09543* (2024).
- [28] Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. 2019. Session-based recommendation with graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 346–353.
- [29] Fajie Yuan, Alexandros Karatzoglou, Ioannis Arapakis, Joemon M Jose, and Xiangnan He. 2019. A Simple Convolutional Generative Network for Next Item Recommendation. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, 582–590.
- [30] Mengqi Zhang, Shu Wu, Xueli Yu, Qiang Liu, and Liang Wang. 2022. Dynamic graph neural networks for sequential recommendation. *IEEE Transactions on Knowledge and Data Engineering* 35, 5 (2022), 4741–4753.
- [31] Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 2020. S3-Rec: Self-Supervised Learning for Sequential Recommendation with Mutual Information Maximization. In *CIKM*. ACM, 1893–1902.
- [32] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. 2024. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417* (2024).