Attention Calibration for Transformer-based Sequential Recommendation

Peilin Zhou*
The Hong Kong University of Science and Technology (Guangzhou)
zhoupalin@gmail.com

Qichen Ye*
Peking University
yeeeqichen@pku.edu.cn

Yueqi Xie
The Hong Kong University of Science and Technology
yxieay@connect.ust.hk

Jingqi Gao
Upstage
mrgao.ary@gmail.com

Shoujin Wang
University of Technology Sydney
shoujin.wang@uts.edu.au

Jae Boum Kim
The Hong Kong University of Science and Technology
jbkim@cse.ust.hk

Chenyu You
Yale University
chenyu.you@yale.edu

Sunghun Kim†
The Hong Kong University of Science and Technology (Guangzhou)
hunkim@cse.ust.hk

code: https://github.com/AIM-SE/AC-TSR.

CIKM 2023

Reported by Minqin Li

Introduction

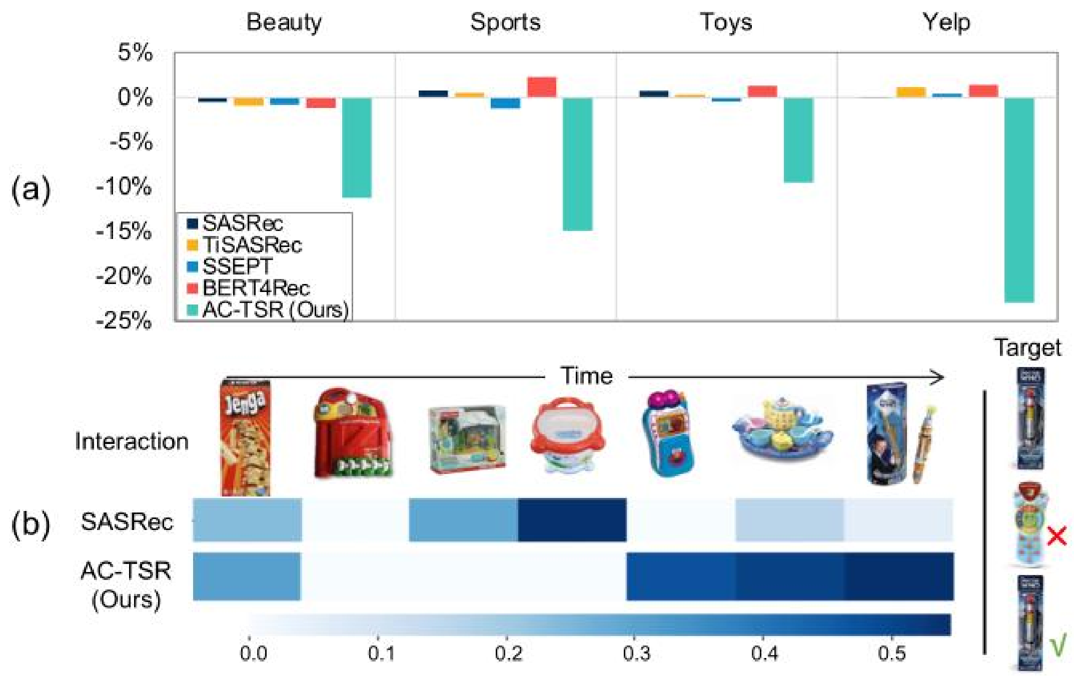


Figure 1: (a) Removing the highest attention weight from transformer-based SRS does not lead to a significant decrease in model performance and even improves performance in some cases; (b) Visualization of the attention weights from SASRec and our proposed AC-TSR.

After careful and in-depth analysis, we found the aforementioned unreliable or inaccurate assignment of attention weights could be mainly attributed to the following two factors:

(1) Sub-optimal position encoding.

(2) Noisy input.

Method

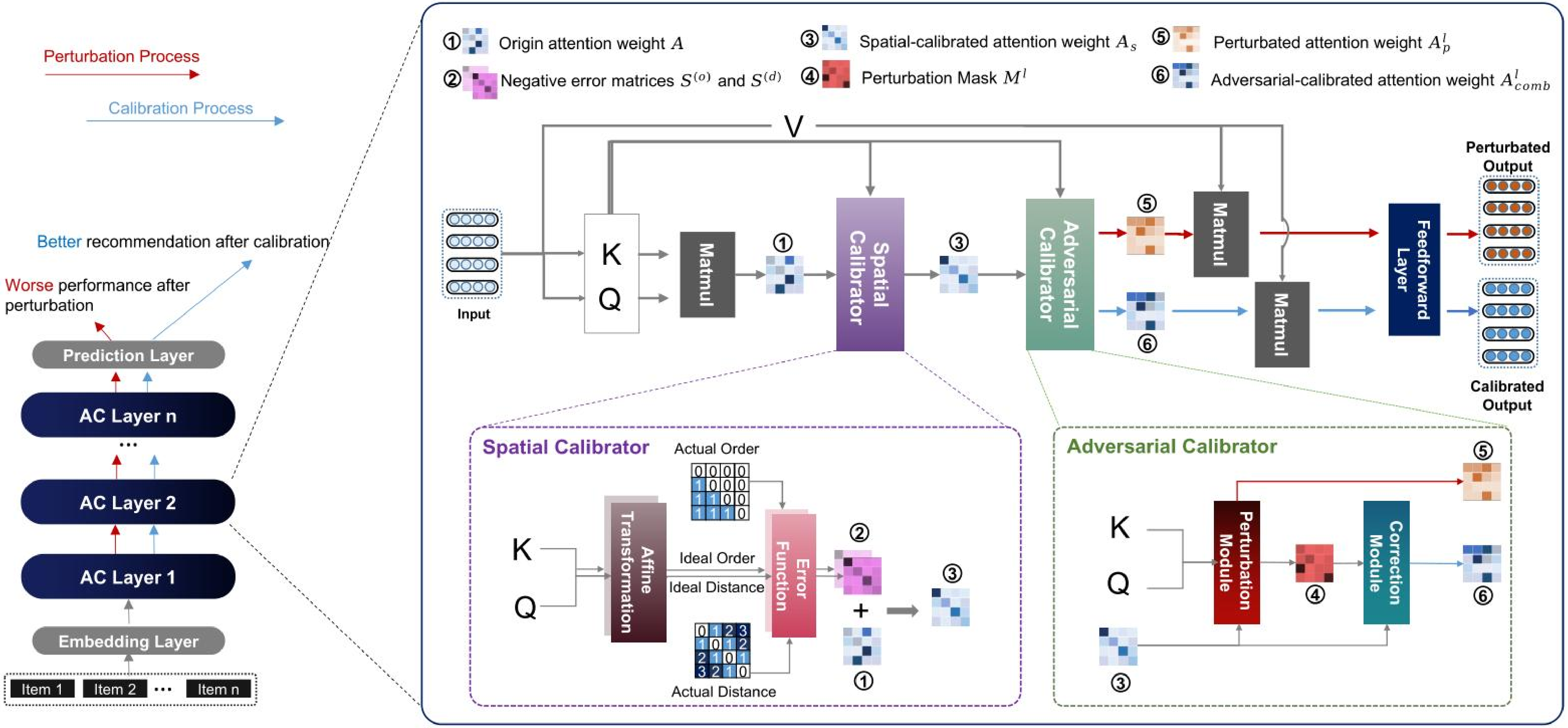
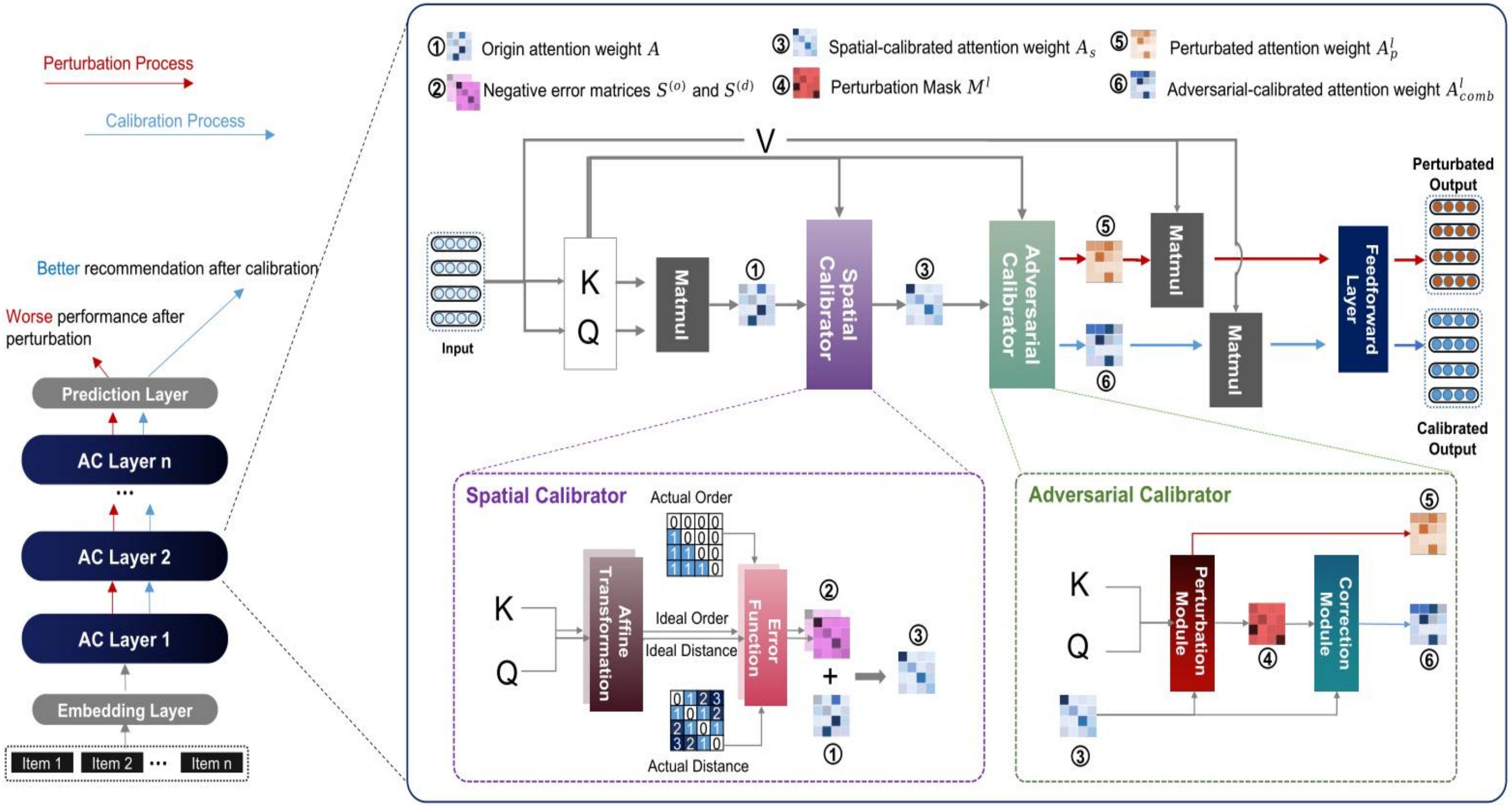


Figure 2: Overview of the proposed AC-TSR framework.

Method



$$\mathbf{H} = \text{Self-Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \operatorname{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V} \quad (1)$$

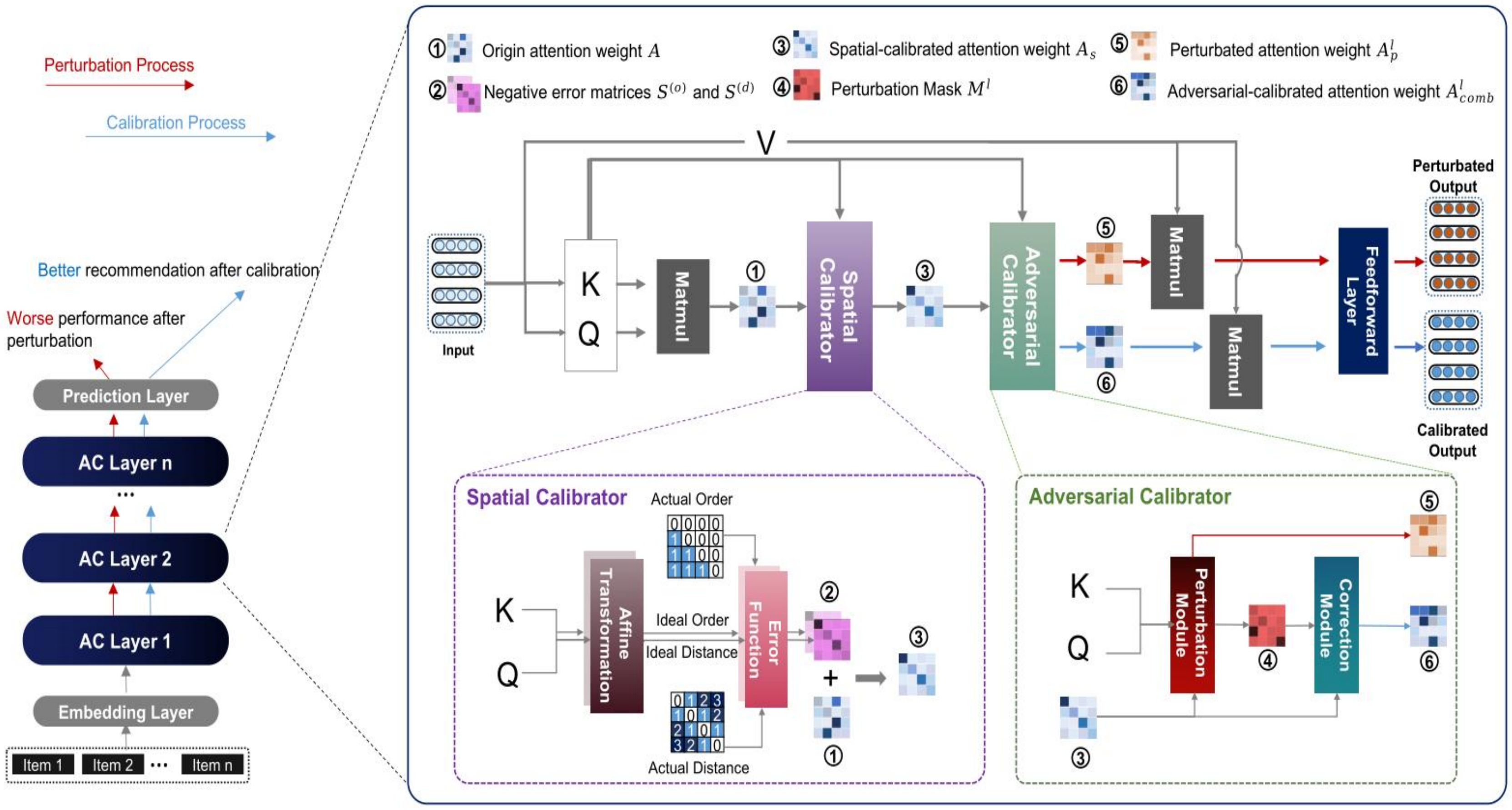
$$\mathbf{F}_i = \mathrm{FFN}(\mathbf{H}_i) = \mathrm{ReLU}(\mathbf{H}_i \mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2 \quad (2)$$

$$\mathbf{F}_i^L = \mathrm{FFN}\left(\mathbf{H}_i^L\right) \quad (3)$$

$$\hat{\mathbf{y}} = \operatorname{softmax}\left(\mathbf{T}{\mathbf{F}_n^L}^T\right) \quad (4)$$

$$\mathcal{L} = -\sum_{i=1}^{|\mathcal{I}|} y_i \log(\hat{y}_i) \quad (5)$$

Method



$$o_{ij} = \mathbb{I}(i < j) = \begin{cases} 1, & i < j \\ 0, & otherwise \end{cases} \tag{6}$$

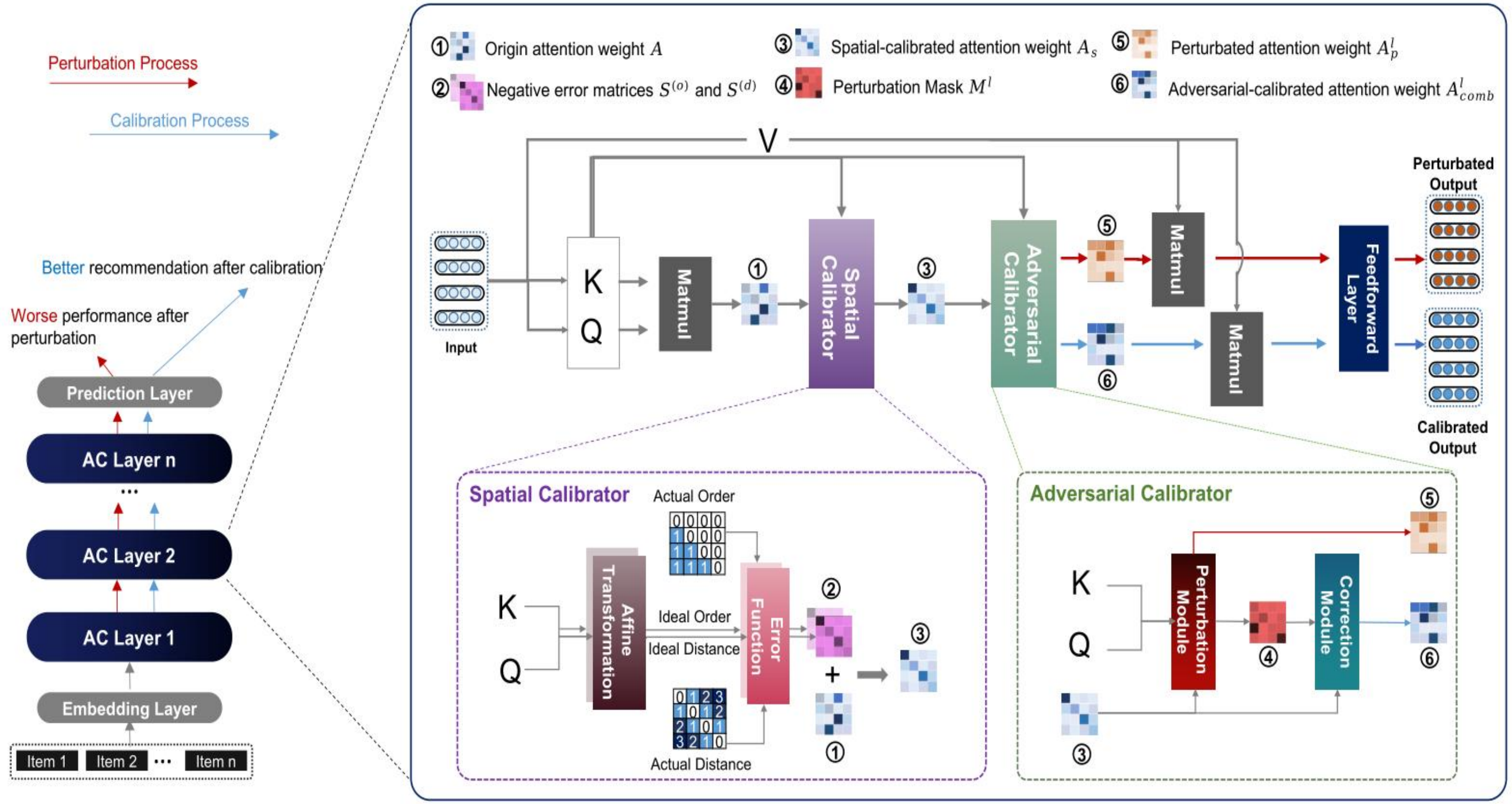
$$d_{ij} = \ln\left(1 + |i - j|\right) \tag{7}$$

$$\hat{o}_{ij} = \text{sigmoid}\left(\text{affine}^{(o)}\left(\left[\mathbf{q}_i^l; \mathbf{k}_j^l\right]\right)\right) \tag{8}$$

$$\hat{d}_{ij} = \text{affine}^{(d)}\left(\left[\mathbf{q}_i^l; \mathbf{k}_j^l\right]\right) \tag{9}$$

$$s_{ij}^{(o)} = o_{ij}\ln\left(\hat{o}_{ij}\right) + (1 - o_{ij})(1 - \ln\left(\hat{o}_{ij}\right)) \tag{10}$$

Method



$$s_{ij}^{(d)} = -\frac{\theta^2\left(d_{ij}-\hat{d}_{ij}\right)^2}{2} \tag{11}$$

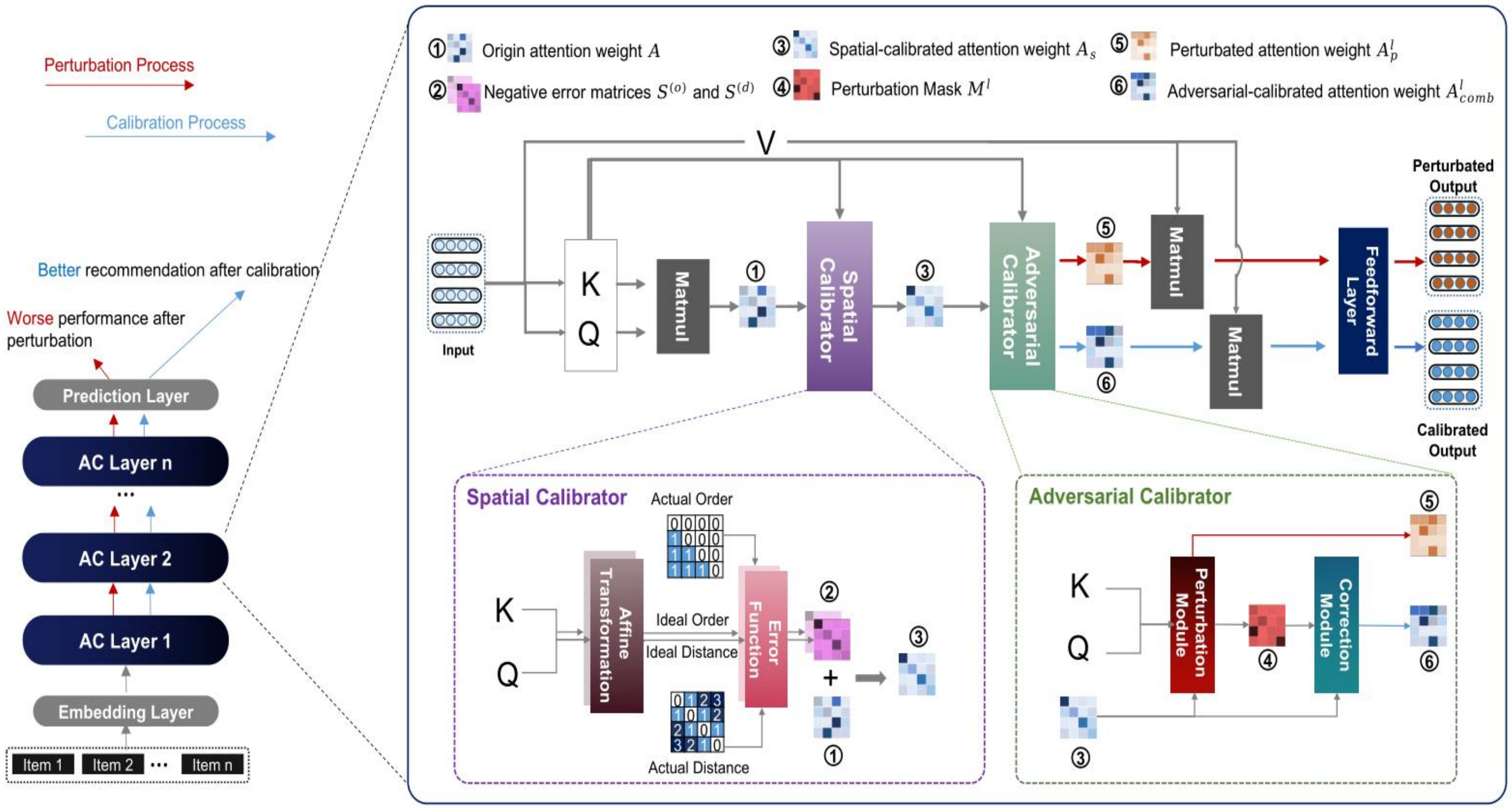
$$\mathbf{A}_s = \mathrm{softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{d}} + \mathbf{s}^{(o)} + \mathbf{s}^{(d)}\right) \tag{12}$$

$$\mathbf{A}_p^l = \mathbf{M}^l \odot \mathbf{A}_s^l + (1-\mathbf{M}^l) \odot \mu \tag{13}$$

$$\mathbf{M}^l = \mathrm{sigmoid}\left(\frac{\mathbf{Q}^l\mathbf{W}_{Q_p}^l\left(\mathbf{K}^l\mathbf{W}_{K_p}^l\right)^T}{\sqrt{d}}\right) \tag{14}$$

$$\mathbf{A}_c^l = \mathbf{A}_s^l \odot e^{1-\mathbf{M}^l} \tag{15}$$

Method



$$\mathbf{A}^l_{comb} = \mathbf{g} * \mathbf{A}^l_s + (1-\mathbf{g}) * \mathbf{A}^l_c \tag{16}$$

$$\mathbf{g} = \sigma\left(\mathbf{Q}^l \mathbf{W}^l_g + \mathbf{b}^l_g\right) \tag{17}$$

$$\mathcal{L}_P = -\sum_{i=1}^{|\mathcal{I}|} y_i \log\left(\hat{y}_i^P\right) \tag{18}$$

$$\mathcal{L}_C = -\sum_{i=1}^{|\mathcal{I}|} y_i \log\left(\hat{y}_i^C\right) \tag{19}$$

$$\mathcal{L}_{P_{final}}\left(\theta^P\right) = -\mathcal{L}_P(\theta) + \alpha \mathcal{L}_{norm}\left(\theta^P\right) \tag{20}$$

$$\mathcal{L}_{norm}\left(\theta^P\right) = \sum_{l=0}^{L} ||1-\mathbf{m}^l||_2 \tag{21}$$

$$\mathcal{L}_{final} = \mathcal{L}_{P_{final}} + \mathcal{L}_C \tag{22}$$

Experiments

Table 1: Overall performance. The highest results are denoted in bold, while the runner-up results are underscored. "*" denotes the statistical siginificance for $p < 0.01$ compared to the best baseline methods with paired t-test.

SR Model	Beauty				Sports				Toys				Yelp			
	Recall		NDCG		Recall		NDCG		Recall		NDCG		Recall		NDCG	
	@10	@20	@10	@20	@10	@20	@10	@20	@10	@20	@10	@20	@10	@20	@10	@20
PopRec	0.0157	0.0242	0.0076	0.0097	0.0146	0.0244	0.0078	0.0103	0.0105	0.0172	0.0060	0.0077	0.0099	0.0161	0.0051	0.0067
BPR	0.0375	0.0590	0.0168	0.0222	0.0302	0.0480	0.0144	0.0188	0.0344	0.0560	0.0151	0.0205	0.0589	0.0830	0.0324	0.0384
GRU4Rec	0.0654	0.1002	0.0322	0.0410	0.0386	0.0609	0.0195	0.0251	0.0449	0.0708	0.0221	0.0287	0.0418	0.0679	0.0206	0.0271
Caser	0.0474	0.0731	0.0239	0.0304	0.0227	0.0364	0.0118	0.0153	0.0361	0.0566	0.0186	0.0238	0.0380	0.0608	0.0197	0.0255
LightSANS	0.0770	0.1177	0.0358	0.0461	0.0509	0.0781	0.0226	0.0294	0.0768	0.1116	0.0354	0.0442	0.0630	0.0904	0.0385	0.0453
Locker	0.0802	0.1197	0.0365	0.0464	0.0508	0.0753	0.0225	0.0286	0.0755	0.1094	0.0345	0.0430	0.0603	0.0869	0.0380	0.0446
SASRec	0.0779	0.1152	0.0353	0.0447	0.0504	0.0760	0.0224	0.0289	0.0776	0.1100	0.0352	0.0434	0.0618	0.0879	0.0387	0.0453
w/ AC	0.0817*	0.1218*	**0.0375***	0.0454*	0.0532*	0.0817*	0.0235*	0.0307*	0.0825*	0.1166*	0.0371*	0.0456*	**0.0664***	**0.0955***	**0.0407***	**0.0480***
Improve.	4.88%	5.73%	6.23%	1.57%	5.56%	7.50%	4.91%	6.23%	6.31%	6.00%	5.40%	5.07%	7.44%	8.65%	5.17%	5.96%
BERT4Rec	0.0557	0.0868	0.0279	0.0358	0.0313	0.0502	0.0155	0.0202	0.0489	0.0769	0.0253	0.0324	0.0467	0.0710	0.0264	0.0325
w/ AC	0.0628*	0.0929*	0.0318*	0.0394*	0.0381*	0.0607*	0.0196*	0.0253*	0.0643*	0.0924*	0.0339*	0.0410*	0.0481*	0.0769*	0.0265*	0.0337*
Improve.	12.73%	7.03%	13.98%	10.06%	21.73%	20.92%	26.45%	25.25%	31.49%	20.16%	33.99%	26.54%	3.00%	8.31%	0.38%	3.69%
SSE-PT	0.0587	0.0936	0.0278	0.0366	0.0363	0.0580	0.0184	0.0239	0.0560	0.0837	0.0255	0.0325	0.0556	0.0779	0.0323	0.0379
w/ AC	0.0629*	0.1001*	0.0293*	0.0387*	0.0379*	0.0589*	0.0191*	0.0244*	0.0614*	0.0896*	0.0282*	0.0353*	0.0565*	0.0821*	0.0330*	0.0394*
Improve.	7.16%	6.94%	5.40%	5.74%	4.41%	1.55%	3.80%	2.09%	9.64%	7.05%	10.59%	8.62%	1.62%	5.39%	2.17%	3.96%
TiSASRec	0.0794	0.1208	0.0356	0.0461	0.0523	0.0799	0.0230	0.0300	0.0819	0.1171	0.0367	0.0456	0.0618	0.0909	0.0387	0.0460
w/ AC	**0.0823***	**0.1227***	0.0373*	**0.0474***	**0.0548***	**0.0837***	**0.0241***	**0.0313***	**0.0831***	**0.1208***	**0.0375***	**0.0470***	0.0654*	0.0939*	0.0401*	0.0473*
Improve.	3.65%	1.57%	4.78%	2.82%	4.78%	4.76%	4.78%	4.33%	1.47%	3.16%	2.18%	3.07%	5.83%	3.30%	3.62%	2.83%

Experiments

Table 2: Model Complexity.

Model	# Parameters	Inference speed	Recall@20			
			Beauty	Sports	Toys	Yelp
SASRec	0.87M	2482.33/s	0.1152	0.0760	0.1100	0.0879
AC-SASRec	0.90M	917.54/s	0.1218	0.0817	0.1166	0.0955
AC-SASRec-lite	0.87M	2482.33/s	0.1164	0.0768	0.1150	0.0913

Experiments

Table 3: Ablation study of AC-TSR on Beauty dataset.

Settings	Spatial		Adv.	Recall		NDCG	
	order	distance		@10	@20	@10	@20
(A)	✔	✔	✔	0.0817	0.1218	0.0375	0.0476
(B)	✔	✗	✔	0.0791	0.1210	0.0364	0.0470
(C)	✗	✔	✔	0.0792	0.1201	0.0372	0.0475
(D)	✗	✗	✔	0.0800	0.1202	0.0367	0.0469
(E)	✔	✔	✗	0.0802	0.1197	0.0365	0.0464
(F)	✔	✗	✗	0.0776	0.1168	0.0353	0.0452
(G)	✗	✔	✗	0.0806	0.1188	0.0367	0.0463
(H)	✗	✗	✗	0.0779	0.1152	0.0353	0.0447

Experiments

Table 4: Impact of different positional encoding strategies. The SASRec is chosen as the backbone.

Position Encoding Strategy	Sports		Toys	
	Recall@20	NDCG@20	Recall@20	NDCG@20
Remove Position	0.0775	0.0294	0.1170	0.0456
Absolute Position	0.0760	0.0289	0.1100	0.0434
Relative Position	0.0753	0.0285	0.1172	0.0461
Decoupled Position	0.0769	0.0295	0.1153	0.0449
Spatial Calibrator (Ours)	**0.0785**	**0.0298**	**0.1193**	**0.0462**

Experiments

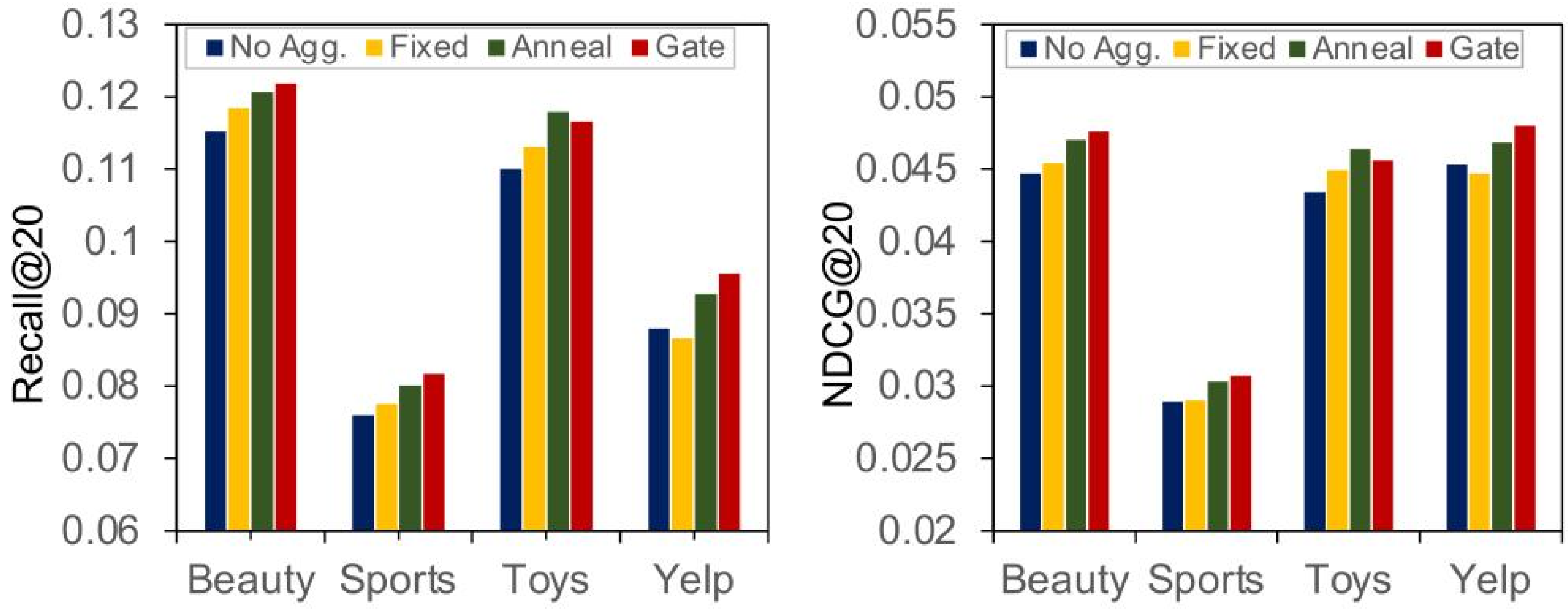


Figure 3: Impact of different aggregation strategies in Correction Module.

Experiments

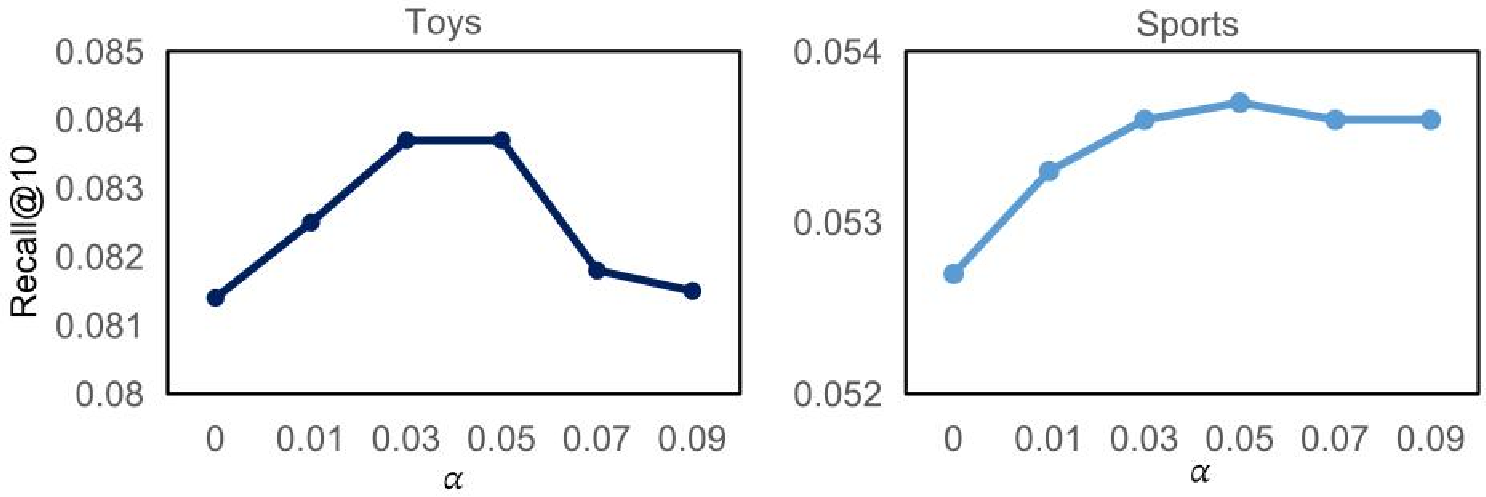


Figure 4: Effect of balance parameter α.

Experiments

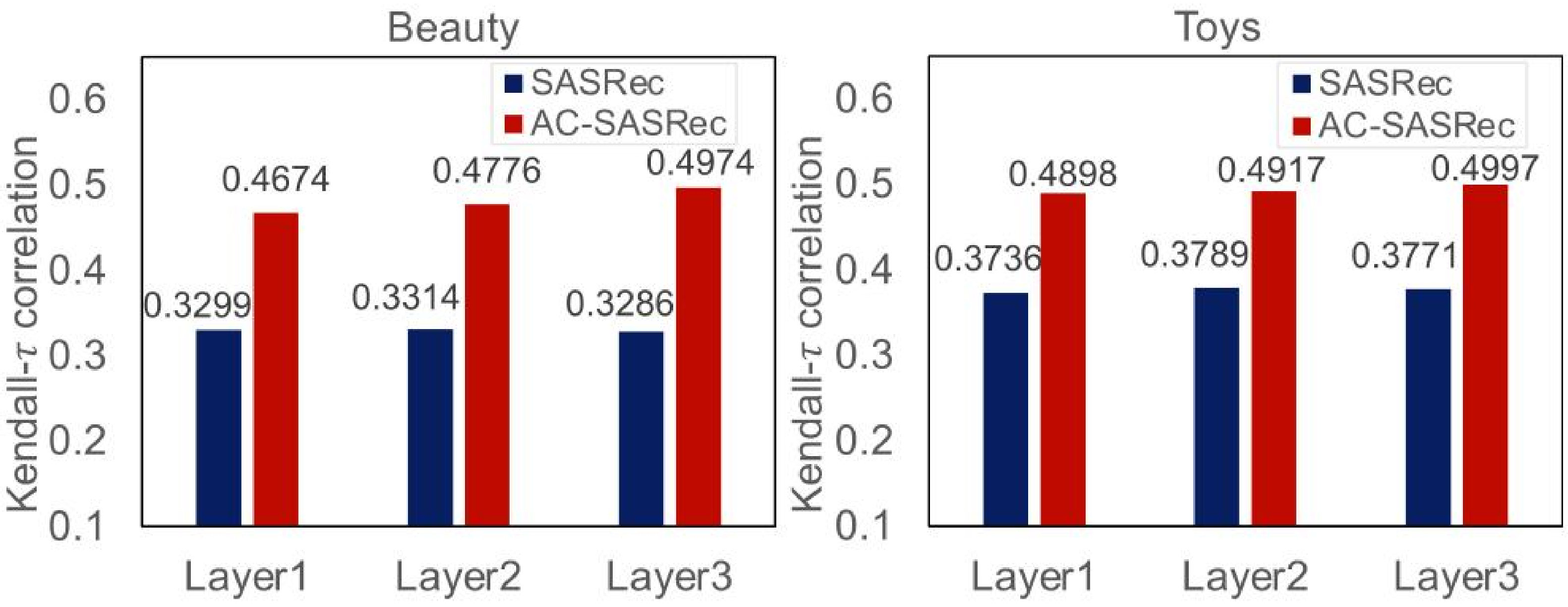


Figure 5: Comparison of the mean Kendall-τ correlation between attention weights and gradient importance measures. The results verify that our AC method can improve Kendall-τ correlation by a large margin.

Experiments

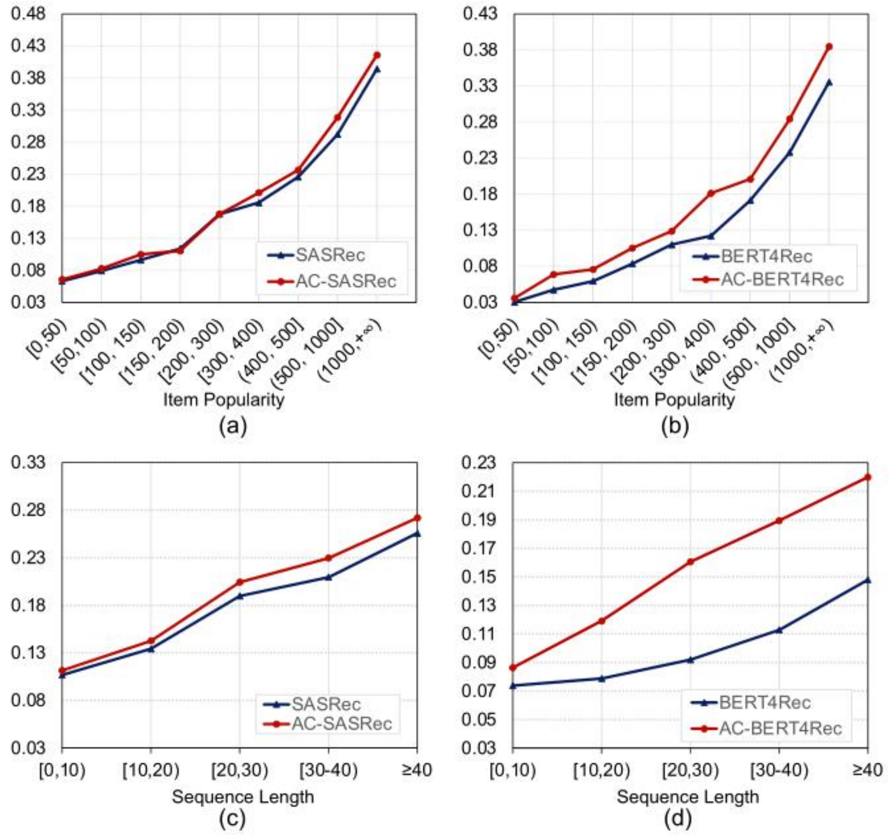


Figure 6: Performance comparison (Recall@20) between AC-TSR and TSR under different sequence lengths (*i.e.*, number of training interactions of users) and item popularity (*i.e.*, number of training interactions of items) on Amazon Beauty.

Thanks