Table 12. Performance of models on Recall@10 with different batch size. "-" indicates OOM.

	Models	256	512	1024	2048	4096	8192
Sampling	a	0.1934	0.1962	0.1989	0.1955	0.1991	**0.2002**
	b	0.1627	0.162	0.1615	**0.166**	0.1619	0.159
	c	0.1892	0.1923	**0.1943**	0.1904	0.1891	0.1926
	d	**0.2041**	0.2004	0.2029	0.2028	0.2038	0.1978
	e	0.2002	**0.2035**	0.2024	0.2013	0.2	-
	f	0.1947	0.1954	0.1956	0.1962	**0.1971**	-
	g	0.1785	0.181	0.1797	0.1809	**0.1815**	-
Non-sampling	h	0.2016	0.2025	0.2043	**0.2057**	-	-
	i	0.2005	0.2022	**0.2022**	0.1956	-	-
	j	0.2013	**0.2027**	0.2022	0.1992	-	-

Table 13. The correlation of RO_RS and TO_RS methods with and without cross validation on ML-1M.

Setting	**SRC**	**OR@5**
w/ CV	0.7±0.06	0.8 ±0.05
w/o CV	0.73	0.83

A APPENDIX

We conduct several additional experiments on some other factors that not been fully discussed in the paper including batch size, cross-validation, cutoff of recommendation list and performance of non-sampling models on AMZ_Elec dataset. We show all the experimental results in this supplementary material.

A.1 Impact of *batch size*

To test the performance of algorithm with different *batch size*, we conduct experiments on 7 sampling based models and 3 non-sampling models and list the results in Table 12.

As shown in the table, though the performance of a model would slightly change with different *batch size*, the relative ranking of the model could be maintained for most cases. Nearly all the model's performance would firstly increase then decrease when we enlarge the *batch size* while the performance change are mostly less than 5%. Besides, in most of the surveyed papers, *batch size* is fixed to 2048 or 4096 in the experiments, so we think it is not a crucial factor in the evaluation and follow the setting in previous papers is suggested.

A.2 Splitting dataset for 5-fold cross-validation

In some of the surveyed paper, the researchers use 5-fold cross-validation to evaluate the algorithm. To study the correlation of different splitting method, we conduct 5-fold CV with RO_RS and TO_RS splitting and report the SRC and OR@5 score between this two splitting method in Table 13.

As shown in the table above, the SRC and OR@5 between two configurations(i.e., RO_RS and TO_RS) of ml-1m are 0.7 and 0.8 respectively. It is very close to results reported in the paper(i.e. 0.73 and 0.83).

Table 14. Algorithm performance with different types of loss functions on AMZ_Elec dataset.

		AMZ_Elec		
	algorithm	Recall@10	NDCG@10	AUC
Non-Sampling	a	0.0248	0.0134	0.7173
	b	**0.0264**	0.0143	0.7169
	c	0.0225	0.0121	0.6925
BCE-based	d	0.0218	0.0119	0.7397
	e	0.0248	**0.0147**	0.636
	f	0.0205	0.011	0.7668
BPR-based	g	0.025	0.0135	0.7745
	h	0.0234	0.0125	**0.8013**
	i	0.026	0.0142	0.7745

A.3 Performance comparison of loss function on long-tailed dataset

We conduct additional experiments of Section 6.1.2 in our paper on AMZ_Elec dataset and list the results in Table 14.

The experimental results are quite counterintuitive as we do not see the superiority of non-sampling models and some of them even perform worse than sampling models. A possible reason is that items in AMZ_Elec dataset are large and diverse while those non-sampling models treat all the non-interacted items as training instance. More noise is introduced and it is hard to capture those few positive items.

A.4 Cutoff metric

To further study how the cutoff length affect the correlation, we conduct additional experiment on NDCG (written as N in the Figure 7) and AUC where we vary the cutoff length from 10 to 10k to calculate NDCG. For datasets with less than 10k items (ML-1M, LastFM, AMZ_video and AMZ_toys), we compute the metric on the full item ranking list and the results are written as N@full in the figure. For other large-scale dataset, we do not compute N@full because of the limited GPU RAM. The SRC score of two metric is averaged across all datasets and shown in Figure 7.

From the heatmap above, we can observe that NDCG@full has very high correlation with other metrics on those small dataset while NDCG@10k has relatively low correlation with NDCG@10 and NDCG@100 which indicates that the performance ranking may be disturbed by the cutoff length. As for AUC, we can see that it has low correlation with 10 and 100 cutoff and the correlation becomes higher when the cutoff length increase. However the SRC between NDCG@full and AUC is still less than 0.75. These results show that both the cutoff on other metrics and the inherent difference of AUC result in the low correlation between AUC and other metrics.

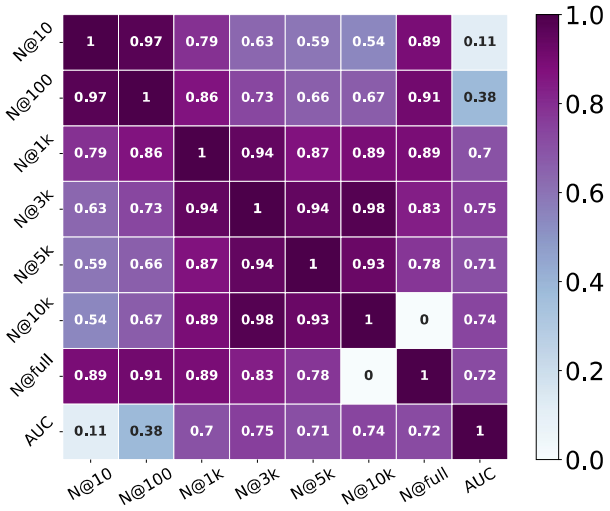

Fig. 7. Visualization of metric correlations under different cutoff. Each cell indicates the computed SRC score between two metrics (a darker color indicates a larger value).