A INFORMATION FOR REPRODUCIBILITY

A.1 The preprocessing steps for datasets.

For the Amazon datasets, we follow the common practice [44, 45] which labels the data samples whose ratings are higher than 3 as positive while others are negative. Since the ratio of these negative items is small, we randomly sample other five items as negative samples for each positive sample as in [44, 45]. Note that HIM focuses on the recommendations for the long-tailed users, therefore, we only filter out the users who have no positive interaction and the items with fewer than 5 interactions. We divide the whole dataset by 70% as training set, 10% as validation and 20% as testing. When it comes to the Advertising dataset, its samples range from 2017-05-06 to 2017-05-13 and each sample contains the user's historical behaviors in four weeks. The clicked and unclicked interactions are treated as positive and negative samples, respectively. We regard the samples ranging from 2017-05-06 to 2017-05-11 as training set. The samples in 2017-05-12 and 2017-05-13 are regarded as validation set and test set, respectively.

A.2 Implementation Details.

As mentioned in Section 3.3, we divide users' behaviors into subsequences by the statistics of each dataset. We choose the same timestamps for all users, which contains recent 10%, 30%, 50% number of total interactions. Therefore, we choose $\{14d, 6m, 12m, all\}$, $\{3m, 9m, 18m, all\}$, and $\{3d, 7d, 14d, 30d\}$, $\{3d, 7d, 14d, 30d\}$ for Instruments, Electronics, Advertising, and Industrial dataset, respectively, where d and m are the abbreviation of day and month. Since we mainly focus on the long-tailed users, we also divide users into head and long-tailed users based on their interaction frequency. We regarded users whose interactions are less than 10 in one month as long-tailed users in all datasets.

We conduct all experiments by TensorFlow on 16 Tesla P100 GPUs with 500 GB memory. And we utilize Adam [19] to optimize the parameters with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and linear decay of the learning rate. For the public datasets, the batch size and the learning rate are set to 256 and $1e-3$, respectively. The item embedding size is test on $[8, 16, 32, 64, 128]$, and the group embedding dimension is $2\times$ of the item vector. We test group number k on $[1, 2, \cdots, 32]$ When it comes to the industrial dataset, the batch size and learning rate are 1024 and $1e-4$, respectively. The item embedding size and the group embedding size are set to 32 and 128, respectively. Other hyper-parameter settings of baselines are the same as the original settings proposed in the papers. All experiments are repeated 10 times. The mean values and standard deviations of the results are finally reported.