# Dual-interest Factorization-heads Attention for Sequential Recommendation

Guanyu Lin[1], Chen Gao[1†], Yu Zheng[1], Jianxin Chang[2], Yanan Niu[2], Yang Song[2],
Zhiheng Li[1], Depeng Jin[1], Yong Li[1]
[1]Tsinghua University
[2]Beijing Kuaishou Technology Co., Ltd.

code:[https://github.com/tsinghua-fib-lab/WWW2023-DFAR.](https://github.com/tsinghua-fib-lab/WWW2023-DFAR)

WWW 2023

**Reported by Minqin Li**

# Introduction

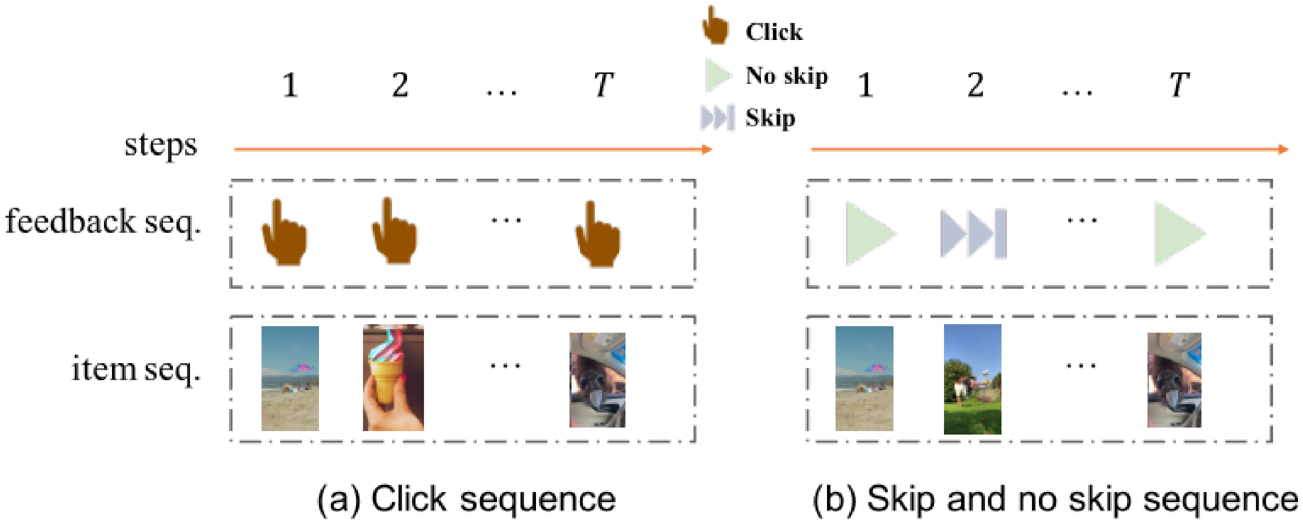
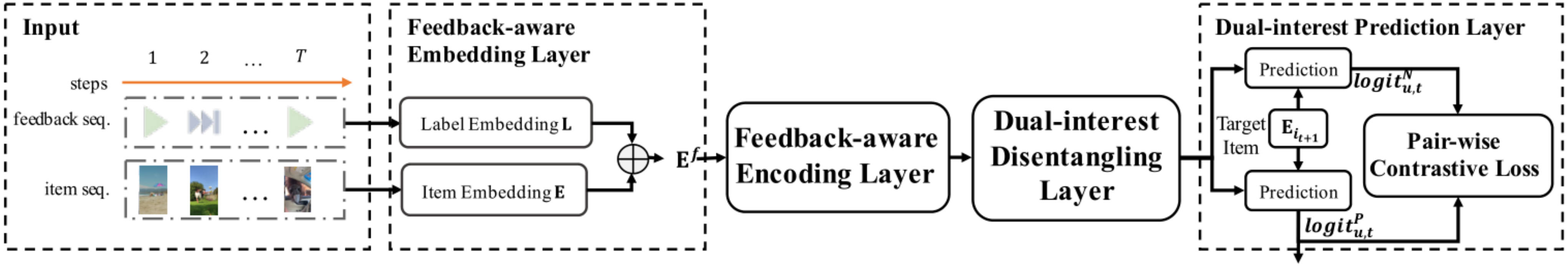
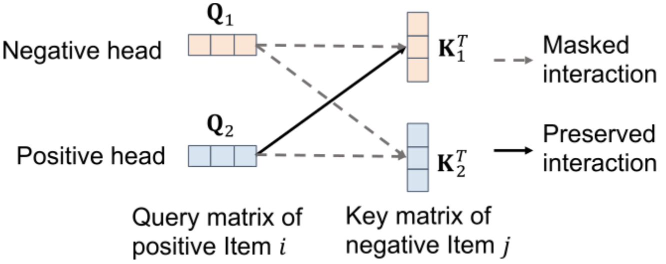
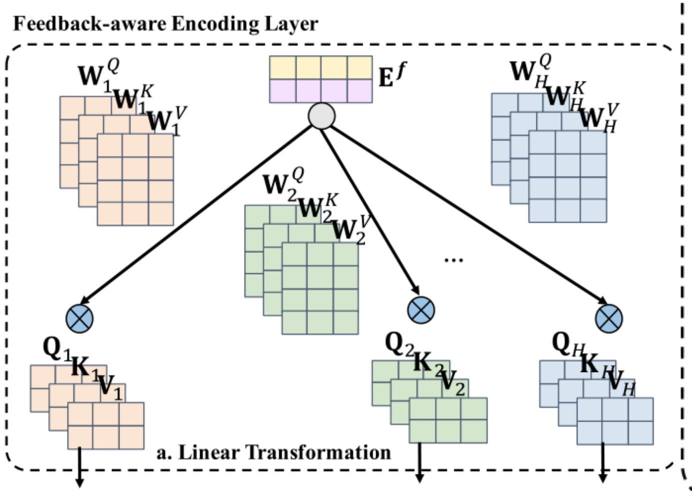


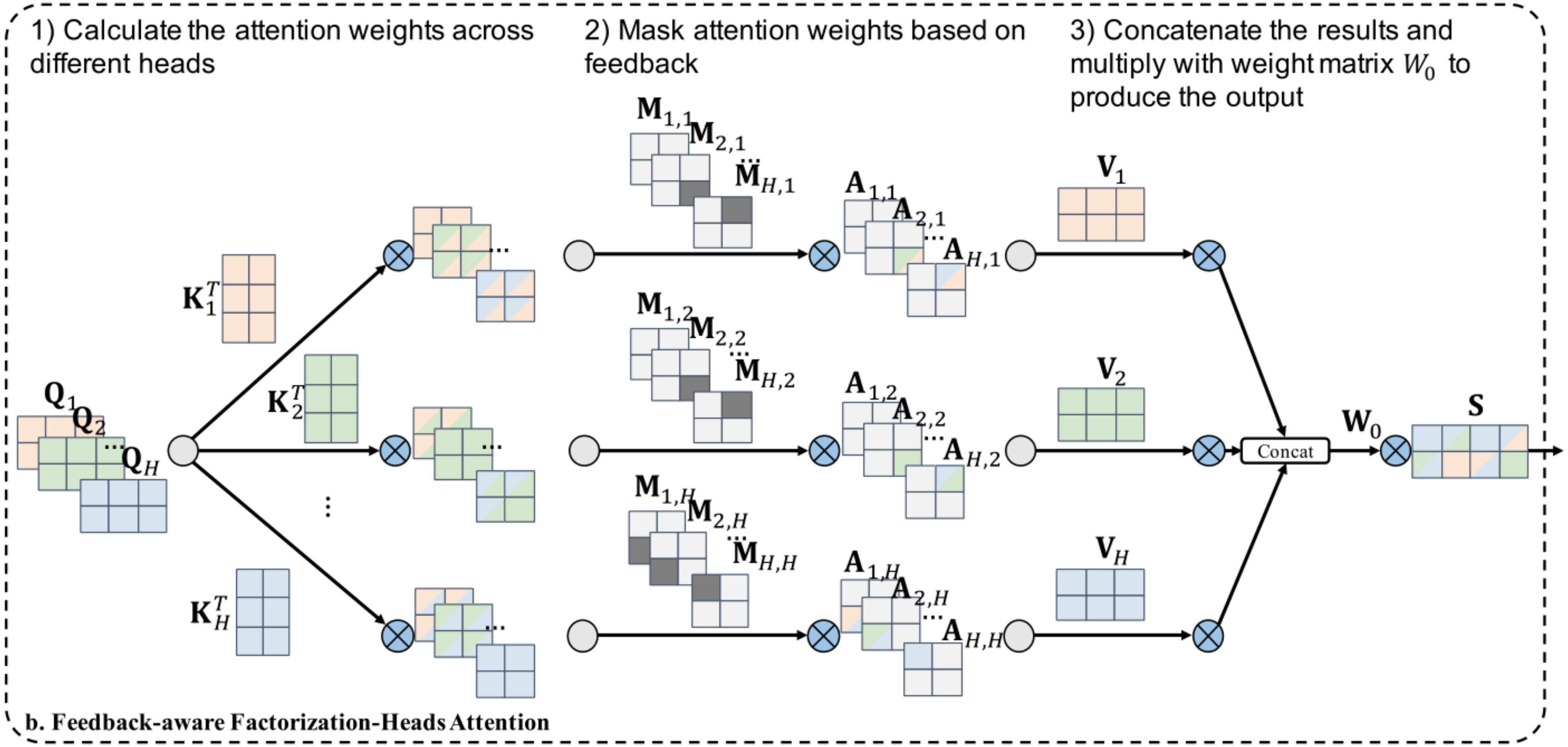

**Figure 1: Illustration of click-based sequential recommendation and our dual-interest sequential recommendation which is hybrid with positive and negative feedback.**

# Method

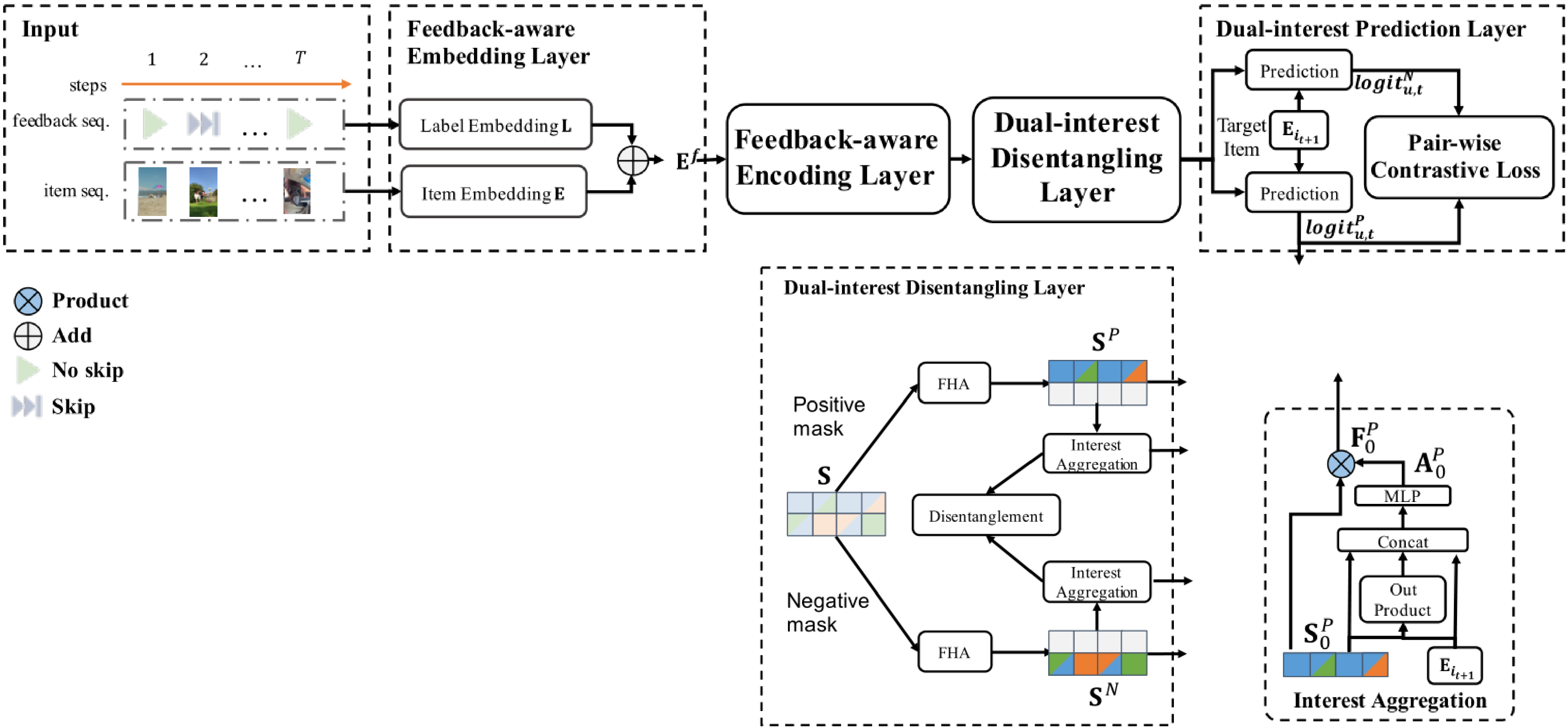

Input
steps
1
2
...
T
feedback seq.
item seq.
Feedback-aware Embedding Layer
Label Embedding L
Item Embedding E
$E^f$
Feedback-aware Encoding Layer
Dual-interest Disentangling Layer
Dual-interest Prediction Layer
Prediction
$logit^N_{u,t}$
Target Item
$E_{i_{t+1}}$
Prediction
$logit^P_{u,t}$
Pair-wise Contrastive Loss


Negative head
Positive head
$Q_1$
$Q_2$
$K_1^T$
$K_2^T$
Masked interaction
Preserved interaction
Query matrix of positive Item $i$
Key matrix of negative Item $j$

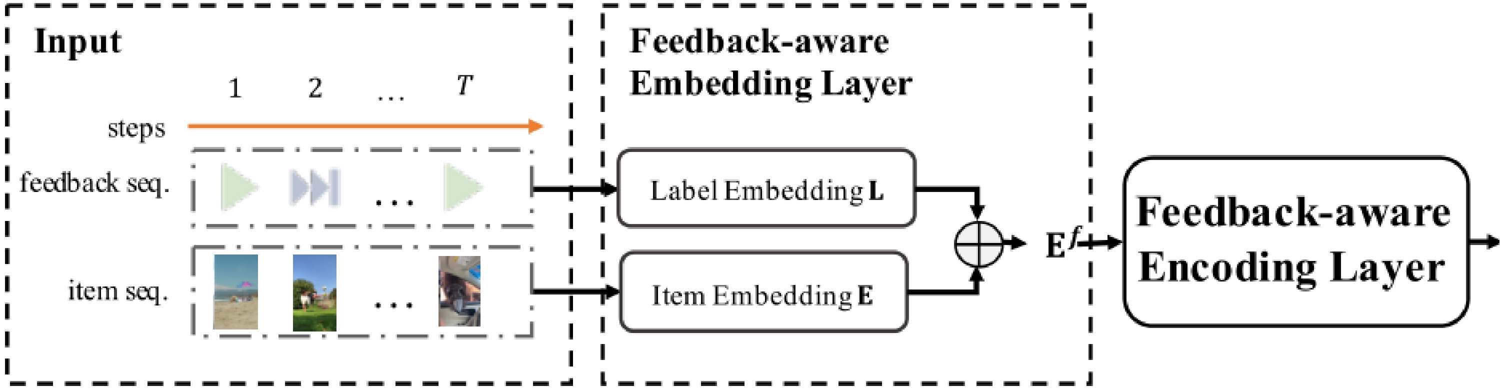

Feedback-aware Encoding Layer
$E^f$
$W_1^Q$ $W_1^K$ $W_1^V$
$W_2^Q$ $W_2^K$ $W_2^V$
$W_H^Q$ $W_H^K$ $W_H^V$
...
$Q_1$ $K_1$ $V_1$
$Q_2$ $K_2$ $V_2$
$Q_H$ $K_H$ $V_H$
a. Linear Transformation


1) Calculate the attention weights across different heads
2) Mask attention weights based on feedback
3) Concatenate the results and multiply with weight matrix $W_0$ to produce the output
$Q_1$ $Q_2$ ... $Q_H$
$K_1^T$
$K_2^T$
$K_H^T$
$M_{1,1}$ $M_{2,1}$ ... $M_{H,1}$
$M_{1,2}$ $M_{2,2}$ ... $M_{H,2}$
$M_{1,H}$ $M_{2,H}$ ... $M_{H,H}$
$A_{1,1}$ $A_{2,1}$ ... $A_{H,1}$
$A_{1,2}$ $A_{2,2}$ ... $A_{H,2}$
$A_{1,H}$ $A_{2,H}$ ... $A_{H,H}$
$V_1$
$V_2$
$V_H$
Concat
$W_0$
S
b. Feedback-aware Factorization-Heads Attention

# Method


Input
1
2
…
T
steps
feedback seq.
item seq.
Feedback-aware Embedding Layer
Label Embedding L
Item Embedding E
E^f
Feedback-aware Encoding Layer
Dual-interest Disentangling Layer
Dual-interest Prediction Layer
Prediction
logit^N_{u,t}
Target Item
E_{i_{t+1}}
Prediction
logit^P_{u,t}
Pair-wise Contrastive Loss
Product
Add
No skip
Skip
Dual-interest Disentangling Layer
Positive mask
S
Negative mask
FHA
S^P
Interest Aggregation
Disentanglement
Interest Aggregation
FHA
S^N
F^P_0
A^P_0
MLP
Concat
Out Product
S^P_0
E_{i_{t+1}}
Interest Aggregation

# Method




$$\mathbf{E}^f = [\mathbf{E}_{i_1}, \mathbf{E}_{i_2}, \ldots, \mathbf{E}_{i_t}] + [\mathbf{L}_{y_{u,i_1}}, \mathbf{L}_{y_{u,i_2}}, \ldots, \mathbf{L}_{y_{u,i_t}}] \quad (1)$$

$$\mathbf{S} = \mathrm{MHA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \left[\mathbf{A}_1^{MHA}\mathbf{V}_1, \ldots, \mathbf{A}_H^{MHA}\mathbf{V}_H\right] \mathbf{W}_0 \quad (2)$$

$$\mathbf{A}_h^{MHA} = \mathrm{softmax}\left(\frac{\mathbf{Q}_h {\mathbf{K}_h}^T}{\sqrt{d}}\right) \quad (3)$$

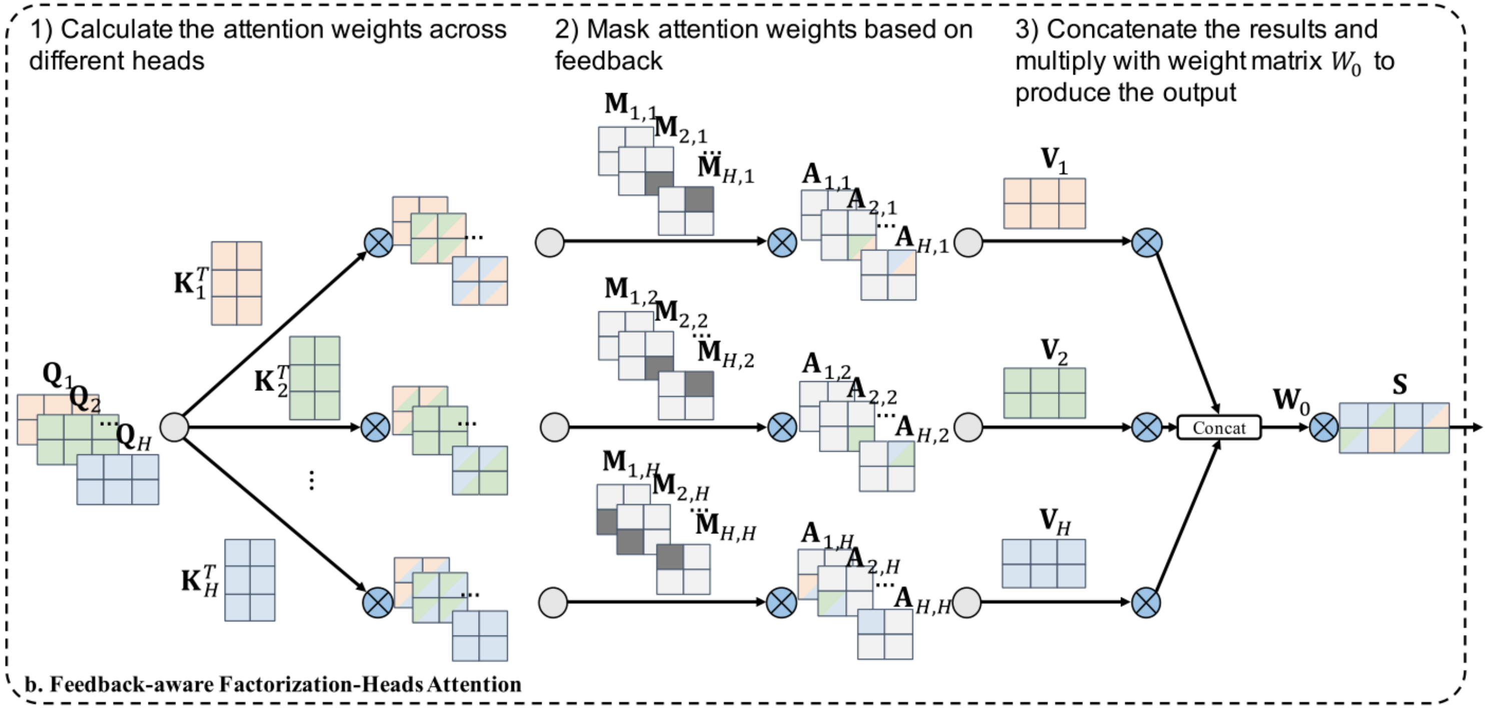
$$\mathbf{Q}_h = \mathbf{Q}\mathbf{W}_h^Q, \mathbf{K}_h = \mathbf{K}\mathbf{W}_h^K, \mathbf{V}_h = \mathbf{V}\mathbf{W}_h^V \quad (4)$$

$$\mathbf{S} = \mathrm{THA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \left[\mathbf{A}_1^{THA}\mathbf{V}_1, \ldots, \mathbf{A}_H^{THA}\mathbf{V}_H\right] \mathbf{W}_0 \quad (5)$$

$$\begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \\ \vdots \\ \mathbf{A}_{H'} \end{bmatrix} = \mathbf{W}_{THA} \begin{bmatrix} \frac{\mathbf{Q}_1 {\mathbf{K}_1}^T}{\sqrt{d}} \\ \frac{\mathbf{Q}_2 {\mathbf{K}_2}^T}{\sqrt{d}} \\ \vdots \\ \frac{\mathbf{Q}_H {\mathbf{K}_H}^T}{\sqrt{d}} \end{bmatrix} \quad (6)$$

$$\begin{bmatrix} \mathbf{A}_1^{THA} \\ \mathbf{A}_2^{THA} \\ \vdots \\ \mathbf{A}_H^{THA} \end{bmatrix} = \mathbf{W}_{THA}^S \begin{bmatrix} \mathrm{softmax}(\mathbf{A}_1) \\ \mathrm{softmax}(\mathbf{A}_2) \\ \vdots \\ \mathrm{softmax}(\mathbf{A}_{H'}) \end{bmatrix} \quad (7)$$

# Method



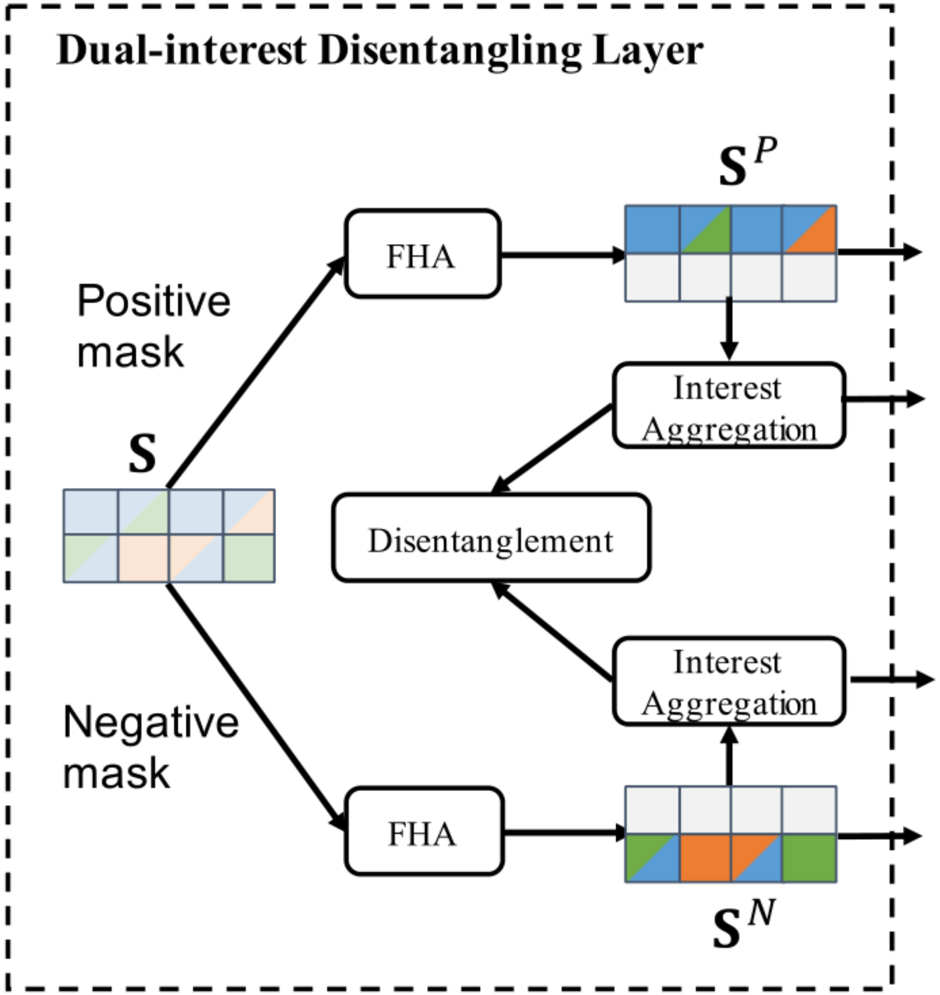

$$\mathbf{S} = \mathrm{FHA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \left[\mathbf{A}_{1,1}^{FHA}\mathbf{V}_1, \ldots, \mathbf{A}_{H,H}^{FHA}\mathbf{V}_H\right]\mathbf{W}_0 \tag{8}$$

$$\mathbf{A}_{h_1,h_2}^{FHA} = \mathrm{softmax}\left(\frac{\mathbf{Q}_{h_1}\mathbf{K}_{h_2}{}^T}{\sqrt{d}}\right) \tag{9}$$

$$\mathbf{S} = \mathrm{FFHA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \left[\mathbf{A}_{1,1}^{FFHA}\mathbf{V}_1, \ldots, \mathbf{A}_{H,H}^{FFHA}\mathbf{V}_H\right]\mathbf{W}_0 \tag{10}$$

$$\mathbf{A}_{h_1,h_2}^{FFHA} = \mathrm{softmax}\left(\mathbf{M}_{h_1,h_2}\frac{\mathbf{Q}_{h_1}\mathbf{K}_{h_2}{}^T}{\sqrt{d}}\right) \tag{11}$$

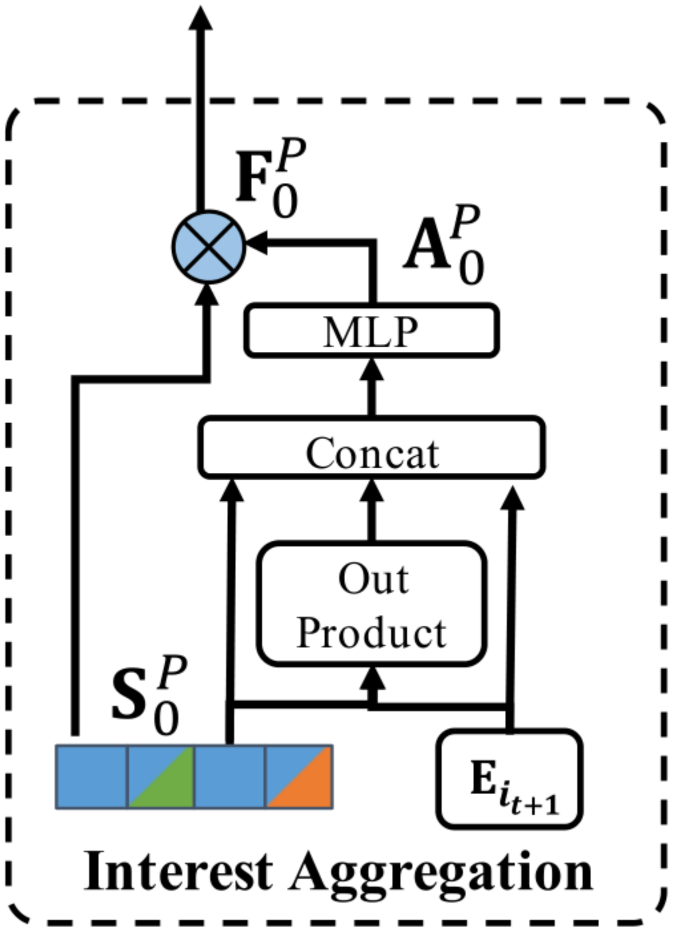
# Method




$$\mathbf{S}^P = [\mathbf{S}_{i_1}, \mathbf{S}_{i_2}, \ldots, \mathbf{S}_{i_t}] * [y_{u,i_1}, y_{u,i_2}, \ldots, y_{u,i_t}],$$
$$\mathbf{S}^N = [\mathbf{S}_{i_1}, \mathbf{S}_{i_2}, \ldots, \mathbf{S}_{i_t}] * (1 - [y_{u,i_1}, y_{u,i_2}, \ldots, y_{u,i_t}]) \quad (12)$$

$$\mathbf{S}^P = \mathrm{FHA}(\mathbf{S}^P, \mathbf{S}^P, \mathbf{S}^P), \mathbf{S}^N = \mathrm{FHA}(\mathbf{S}^N, \mathbf{S}^N, \mathbf{S}^N) \quad (13)$$

# Method



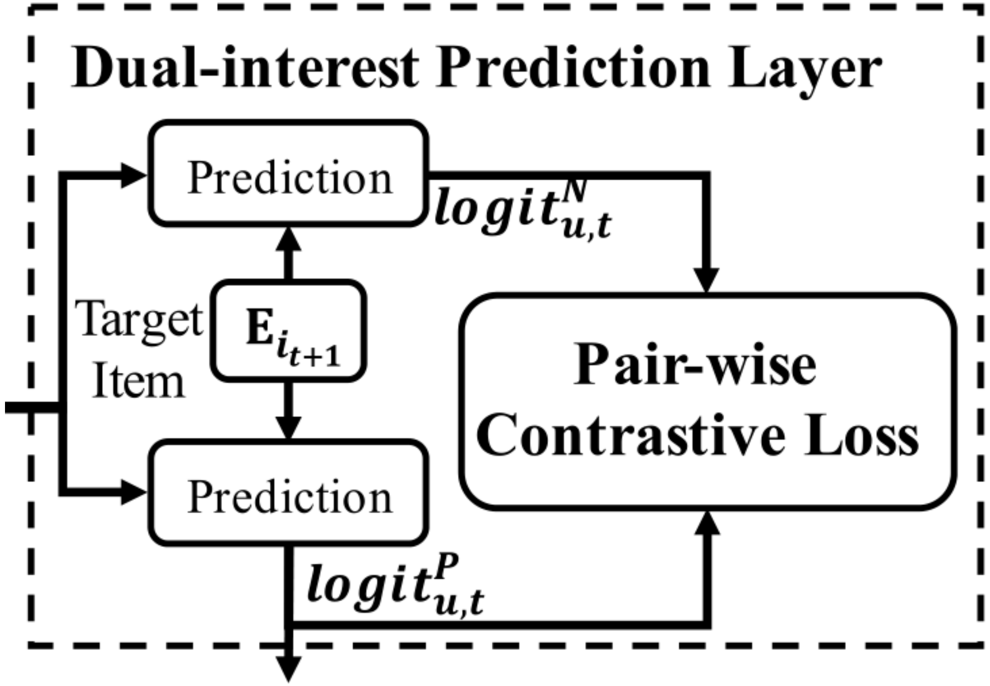

$$\mathbf{A}^P = \mathrm{MLP}\left((\mathbf{E}_{i_{t+1}} + \mathbf{L}_1)\|\mathrm{S}^P\right), \mathbf{A}^N = \mathrm{MLP}\left((\mathbf{E}_{i_{t+1}} + \mathbf{L}_0)\|\mathrm{S}^N\right) \quad (14)$$

$$\mathrm{F}^P = \mathbf{softmax}\left(\mathrm{A}^P\right)\mathrm{S}^P, \mathrm{F}^N = \mathbf{softmax}\left(\mathrm{A}^N\right)\mathrm{S}^N \quad (15)$$

$$\mathbf{f}^P = \sum_{j=1}^{t} \mathbf{F}_j^P, \mathbf{f}^N = \sum_{j=1}^{t} \mathbf{F}_j^N \quad (16)$$

$$\mathcal{L}^D = \frac{\mathbf{f}^P \cdot \mathbf{f}^N}{\|\mathbf{f}^P\| \times \|\mathbf{f}^N\|} \quad (17)$$

# Method




$$logit^P_{u,t} = \mathbf{MLP}\left(\mathbf{s}\|\mathbf{s}^P\|\mathbf{f}^P\|(\mathbf{E}_{i_{t+1}} + \mathbf{L}_1)\right) \tag{18}$$

$$logit^N_{u,t} = \mathbf{MLP}\left(\mathbf{s}\|\mathbf{s}^N\|\mathbf{f}^N\|(\mathbf{E}_{i_{t+1}} + \mathbf{L}_0)\right) \tag{19}$$

$$\mathcal{L}^{BPR} = \begin{cases} -\log(\sigma(logit^P_{u,t} - logit^N_{u,t})), & y_{u,t} = 1, \\ -\log(\sigma(logit^N_{u,t} - logit^P_{u,t})), & y_{u,t} = 0. \end{cases} \tag{20}$$

$$\mathcal{L} = -\frac{1}{|\mathcal{R}|}\sum_{(u,i_t)\in\mathcal{R}}\left(y_{u,t}\log\hat{y}^P_{u,t} + (1-y_{u,t})\log\left(1-\hat{y}^P_{u,t}\right)\right) \tag{21}$$

$$\mathcal{L}^J = \mathcal{L} + \lambda^{BPR}\mathcal{L}^{BPR} + \lambda^D\mathcal{L}^D + \lambda\|\Theta\| \tag{22}$$

# Experiments

Table 1: Micro-video and Amazon data statistics.

| Dataset | | Micro-video | Amazon |
|---|---|---|---|
| #Users | | 37,497 | 6,919 |
| #Items | | 129,092 | 28,695 |
| #Records | Positive | 6,413,396 | 99,753 |
| | Negative | 5,448,693 | 20,581 |
| Avg. records per user | | 316.35 | 17.39 |

# Experiments

**Table 2: Overall evaluations for DFAR against baselines under Micro-video and Amazon datasets on four metrics. Here Improv. is the improvement. Bold is the highest result and underline is the second highest result.**

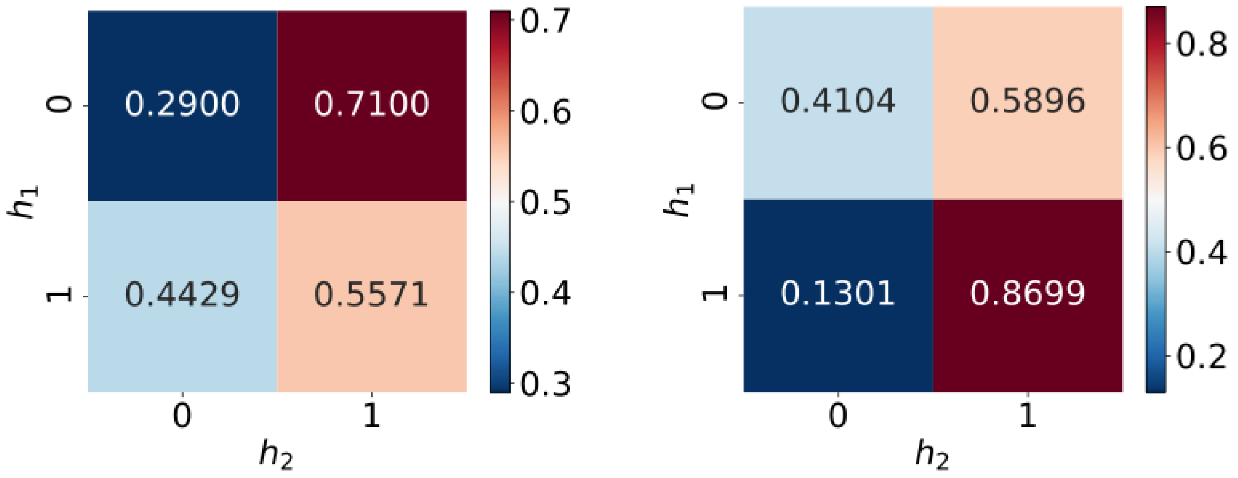
| Models | | DIN | Caser | GRU4REC | DIEN | SASRec | THA4Rec | DFN | FeedRec | Ours | Improv. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Micro-video | AUC | 0.7345 | 0.8113 | 0.7983 | 0.7446 | 0.8053 | 0.8104 | <u>0.8342</u> | 0.8119 | **0.8578** | 2.83% |
| | MRR | 0.5876 | 0.6138 | 0.5927 | 0.5861 | 0.6046 | 0.6080 | <u>0.6321</u> | 0.6095 | **0.6568** | 3.91% |
| | NDCG | 0.6876 | 0.7079 | 0.6916 | 0.6861 | 0.7009 | 0.7035 | <u>0.7222</u> | 0.7047 | **0.7410** | 2.60% |
| | GAUC | 0.7703 | 0.8211 | 0.8041 | 0.7753 | 0.8120 | 0.8138 | <u>0.8362</u> | 0.8180 | **0.8545** | 2.19% |
| Amazon | AUC | 0.6595 | 0.7192 | <u>0.7278</u> | 0.6688 | 0.6903 | 0.7069 | 0.6998 | 0.7037 | **0.7333** | 0.76% |
| | MRR | 0.4344 | 0.4846 | <u>0.4901</u> | 0.4547 | 0.4604 | 0.4599 | 0.4743 | 0.4675 | **0.4980** | 1.61% |
| | NDCG | 0.5669 | 0.6073 | <u>0.6114</u> | 0.5832 | 0.5883 | 0.5879 | 0.5990 | 0.5938 | **0.6175** | 1.00% |
| | GAUC | 0.6618 | 0.7245 | <u>0.7266</u> | 0.6859 | 0.7029 | 0.7021 | 0.7120 | 0.7079 | **0.7305** | 0.54% |

# Experiments

Table 3: Effectiveness study of our proposed components. FHA means factorization-heads attention; MO means label mask operation on heads; IDL means interest disentangling loss on positive and negative representations; IBL means interest BPR loss on positive and negative logits.

| Dataset | Micro-video | | | | |
|---|---|---|---|---|---|
| Methods | w/o FHA | w/o MO | w/o IDL | w/o IBL | Ours |
| AUC | 0.8360 | 0.8473 | 0.8475 | 0.8364 | **0.8578** |
| MRR | 0.6198 | 0.6378 | 0.6377 | 0.6324 | **0.6568** |
| NDCG | 0.7127 | 0.7264 | 0.7264 | 0.7212 | **0.7410** |
| GAUC | 0.8319 | 0.8428 | 0.8436 | 0.8283 | **0.8545** |
| Dataset | Amazon | | | | |
| AUC | 0.7133 | 0.7141 | 0.7284 | 0.7137 | **0.7333** |
| MRR | 0.4782 | 0.4883 | 0.4855 | 0.4839 | **0.4980** |
| NDCG | 0.6016 | 0.6095 | 0.6073 | 0.6057 | **0.6175** |
| GAUC | 0.7054 | 0.7137 | 0.7128 | 0.7047 | **0.7305** |

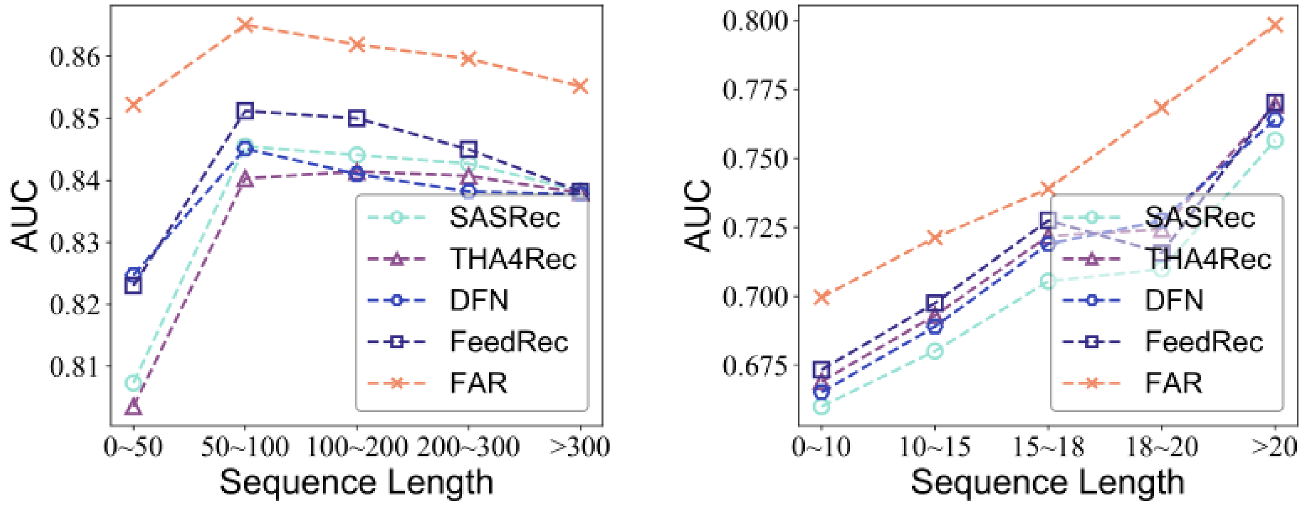
# Experiments




(a) Micro-video (b) Amazon

**Figure 4: Visualization of accumulated attention weights between different heads. Here $h_1$ and $h_2$ represent the heads for the source and target behaviors, respectively (i.e., if the source behavior is negative and target behavior is positive, we have $h_1 = 0$ and $h_2 = 1$). This illustrates our method can factorize and extract the relation between different feedback based on the proposed factorization-heads attention.**

# Experiments




(a) Micro-video (b) Amazon

**Figure 5: AUC performance comparisons under different sequence lengths on the Micro-video and Amazon datasets.**

# Thanks