# Multi-Behavior Hypergraph-Enhanced Transformer for Sequential Recommendation

Yuhao Yang
University of Hong Kong
Hong Kong, China
yuhao-yang@outlook.com

Chao Huang*
University of Hong Kong
Hong Kong, China
chaohuang75@gmail.com

Lianghao Xia
University of Hong Kong
Hong Kong, China
aka_xia@foxmail.com

Yuxuan Liang
National University of Singapore
Singapore, Singapore
yuxliang@outlook.com

Yanwei Yu
Ocean University of China
Qingdao, China
yuyanwei@ouc.edu.cn

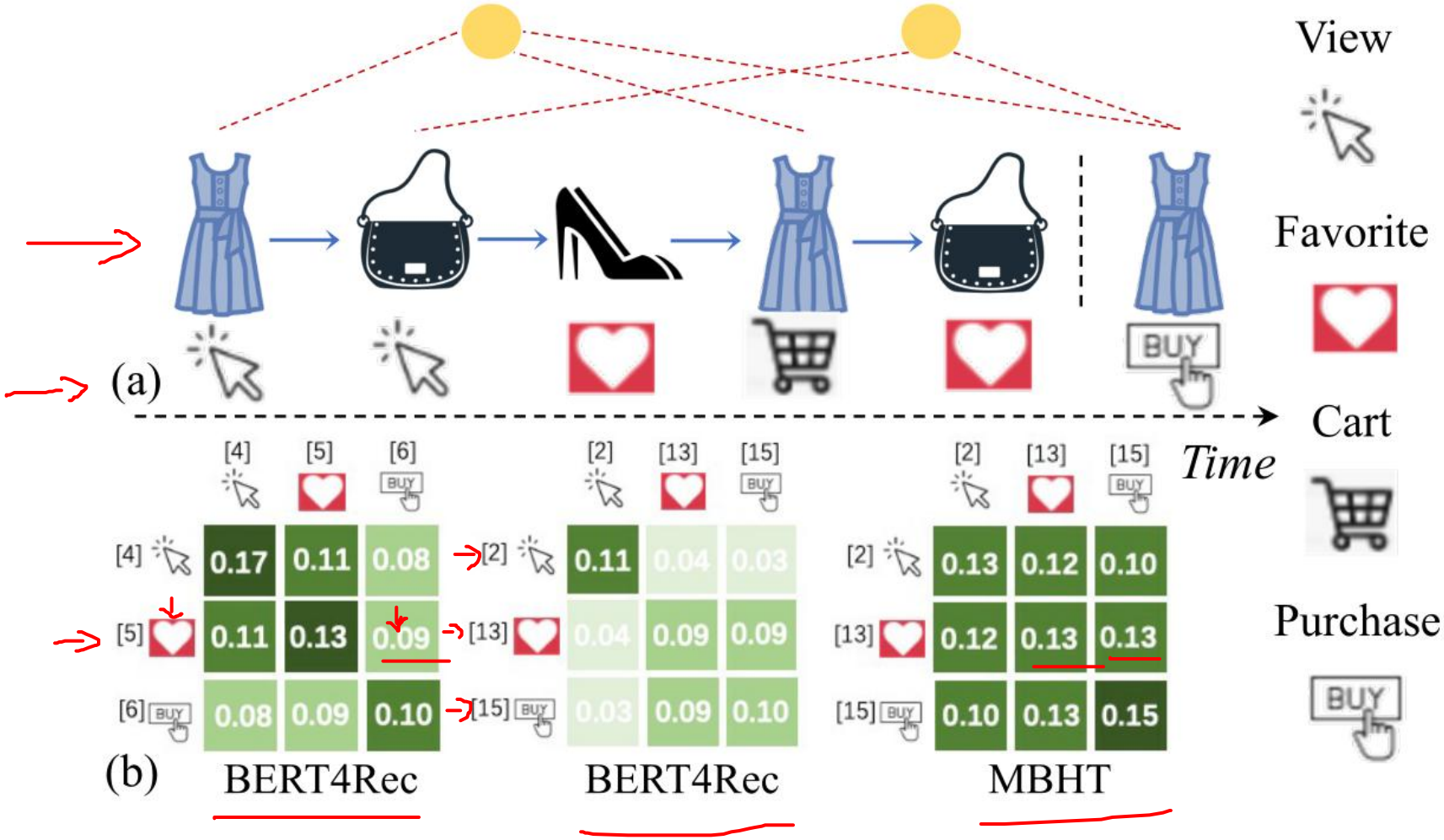
Chenliang Li
Wuhan University
Wuhan, China
cllee@whu.edu.cn

KDD 2022

2022.7.24 • ChongQing

Reported by Gu Tang

# Introduction



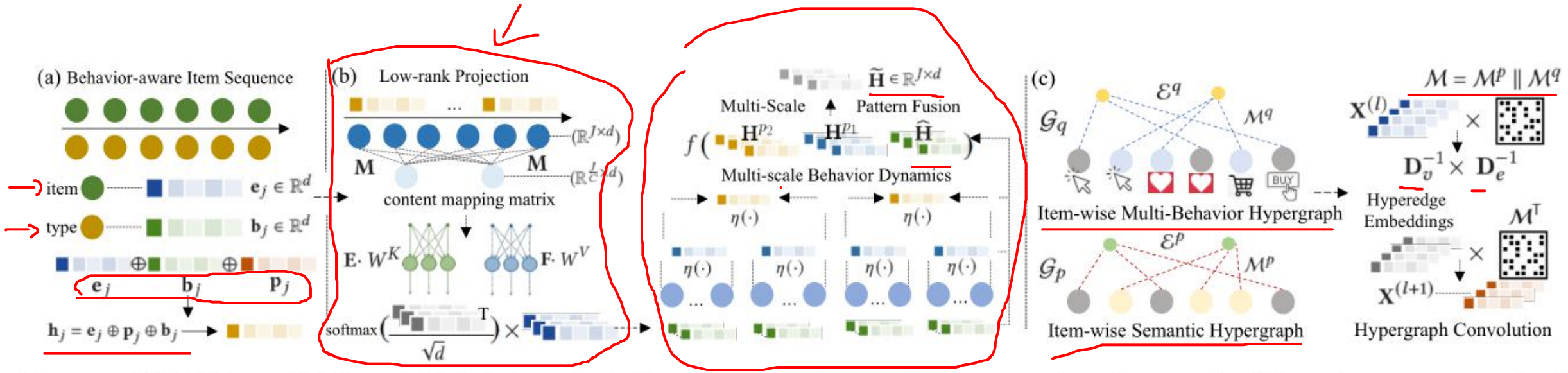
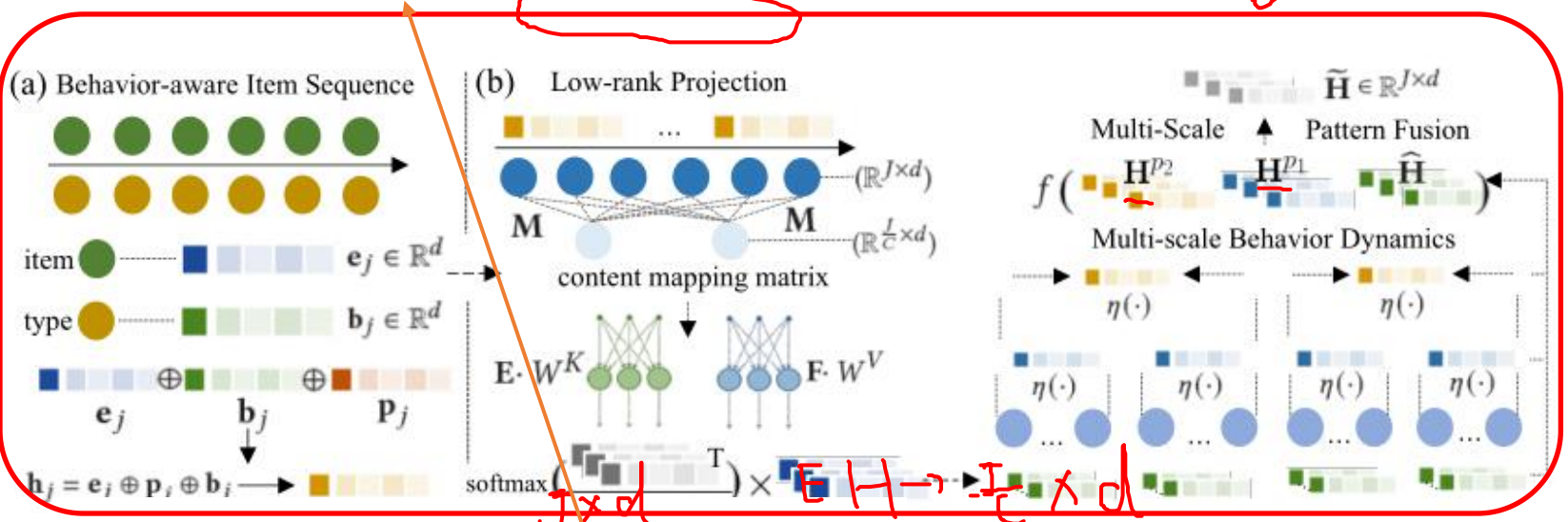
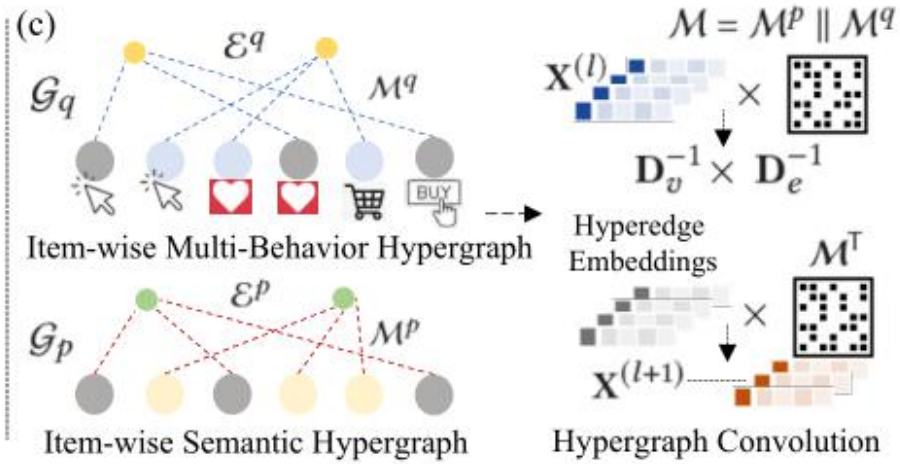



**Figure 2: MBHT's model flow. (a) We inject the behavior-aware interaction context into item embeddings $\mathbf{h}_j = \mathbf{e}_j \oplus \mathbf{p}_j \oplus \mathbf{b}_j$. (b) Multi-scale transformer architecture to capture behavior-aware transitional patterns via low-rank self-attention and multi-scale sequence aggregation. Scale-specific behavior patterns are fused through the fusion function $\widetilde{\mathbf{H}} = f(\widehat{\mathbf{H}} \parallel \mathbf{H}^{p_1} \parallel \mathbf{H}^{p_2})$. (c) We capture the global and personalized multi-behavior dependency learning with our hypergraph neural architecture over $\mathcal{G}$.**

可以通过调整不同大小的C，可以得到Multi-scale Multi-head Attention.




$$\mathbf{h}_j = \mathbf{e}_j \oplus \mathbf{p}_j \oplus \mathbf{b}_j \quad (1)$$

$$\widehat{\mathbf{H}} = \text{softmax}\left(\frac{\mathbf{H}\cdot W^Q (\mathbf{E}\cdot\mathbf{H}\cdot W^K)^\top}{\sqrt{d}}\right)\cdot \mathbf{F}\cdot\mathbf{H}\cdot W^V \quad (2)$$

$\boldsymbol{E} \in \mathbb{R}^{\frac{J}{C}\times J}$ and $\boldsymbol{F} \in \mathbb{R}^{\frac{J}{C}\times J}$ to

$$\Gamma^p = \{\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_{\frac{J}{p}}\} = [\eta(\mathbf{h}_1, ..., \mathbf{h}_p); ...; \eta(\mathbf{h}_{J-p+1}, ..., \mathbf{h}_J)] \quad (3)$$

$$\mathbf{H}^p = \text{softmax}\left(\frac{\Gamma^p \cdot W_p^Q (\Gamma^p \cdot W_p^K)^\top}{\sqrt{d}}\right)\cdot \Gamma^p \cdot W_p^V \quad (4)$$

$$\widetilde{\mathbf{H}} = f(\widehat{\mathbf{H}} \,\|\, \mathbf{H}^{p_1} \,\|\, \mathbf{H}^{p_2}) \quad (5)$$

$\widehat{\mathbf{H}} \in \mathbb{R}^{J\times d}$, $\mathbf{H}^{p_1} \in \mathbb{R}^{\frac{J}{p_1}\times d}$, $\mathbf{H}^{p_2} \in \mathbb{R}^{\frac{J}{p_2}\times d}$.

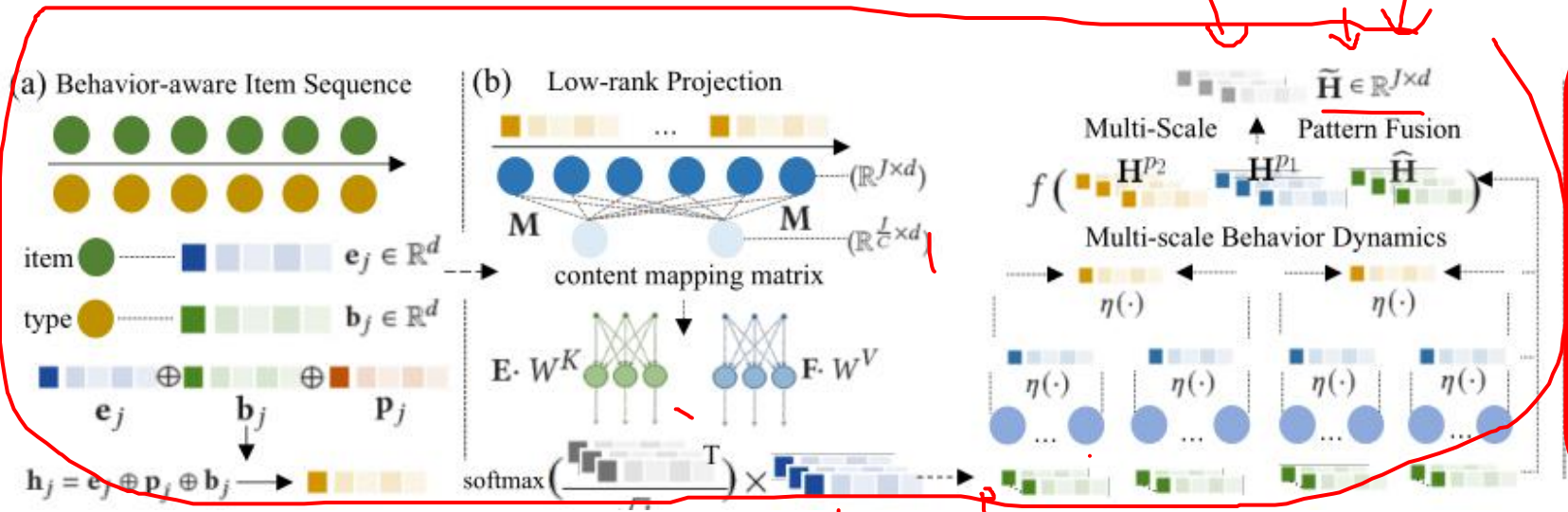
$$\widetilde{\mathbf{H}} = (\text{head}_1 \,\|\, \text{head}_2 \,\|\, \cdots \,\|\, \text{head}_N)\boldsymbol{W}^D \quad (6)$$

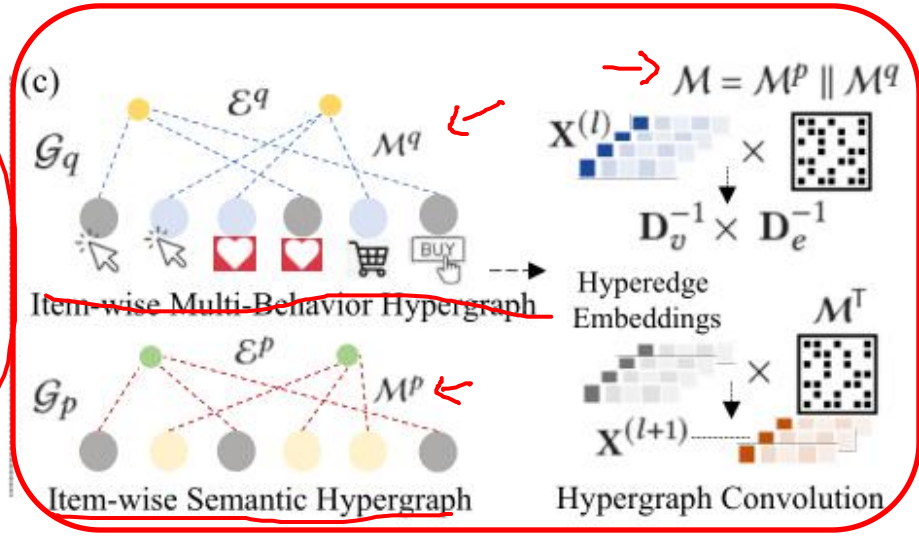
$$\text{head}_n = f(\widehat{\mathbf{H}}_n \,\|\, \mathbf{H}_n^{p_1} \,\|\, \mathbf{H}_n^{p_2})$$

$$\text{PFFN}(\widetilde{\mathbf{H}}^{(l)}) = [\text{FFN}(\tilde{\mathbf{h}}_1^{(l)})^\top, \cdots, \text{FFN}(\tilde{\mathbf{h}}_t^{(l)})^\top] \quad (7)$$

$$\text{FFN}(\mathbf{x}) = \text{GELU}(\mathbf{x}\mathbf{W}_1^{(l)} + \mathbf{b}_1^{(l)})\mathbf{W}_2^{(l)} + \mathbf{b}_2^{(l)},$$

Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-Attention with Linear Complexity. arXiv:2006.04768 [cs.LG]




$$\beta_{j,j'} = \frac{1}{N}\sum_{n=1}^{N}\hat{\beta}^n_{j,j'}; \mathbf{v}_j = \mathbf{e}_j \oplus \mathbf{b}_j \quad (8)$$

$$\hat{\beta}^n_{j,j'} = \tau(\mathbf{w}^n \odot \mathbf{v}_j, \mathbf{w}^n \odot \mathbf{v}_{j'}) \quad (9)$$

$$m^p(v_j, \epsilon_{j'}) = \begin{cases} \beta_{j,j'} & v_{j'} \in A_j; \\ 0 & otherwise: \end{cases} \quad (10)$$

$$m^q(v^b_j, \epsilon^{b'}_j) = \begin{cases} 1 & v^b_j \in \mathcal{E}^q_j; \\ 0 & otherwise; \end{cases} \quad (11)$$

$A_j$ represents the set of top-$k$ semantic correlated items of a specific item $v_j$. We define the connection matrix between items and hyperedges as $\mathcal{M}^p \in \mathbb{R}^{J\times|\mathcal{E}^p|}$ in which each entry $m^p(v_j, \epsilon_{j'})$ is:

i.e., $\mathcal{M} = \mathcal{M}^p \,\|\, \mathcal{M}^q$. $\mathcal{M} \in \mathbb{R}^{J\times(|\mathcal{E}^p|+|\mathcal{E}^q|)}, \mathcal{M}^q \in \mathbb{R}^{J\times|\mathcal{E}^q|}$ $\mathcal{M}^p \in \mathbb{R}^{J\times|\mathcal{E}^p|}$

$$\mathbf{X}^{(l+1)} = \mathbf{D}_v^{-1} \cdot \mathcal{M} \cdot \mathbf{D}_e^{-1} \cdot \mathcal{M}^\mathsf{T} \cdot \mathbf{X}^{(l)} \quad (12)$$

$$\mathbf{x}^{(0)} = (\mathbf{v}_j \oplus \mathbf{b}_j) \odot \text{sigmoid}((\mathbf{v}_j \oplus \mathbf{b}_j) \cdot \mathbf{w} + \mathbf{r}).$$

$$\alpha_i = \text{Attn}(\boldsymbol{e}_i) = \frac{\exp(\boldsymbol{a}^\mathsf{T} \cdot \boldsymbol{W}_a \boldsymbol{e}_i)}{\sum_i \exp(\boldsymbol{a}^\mathsf{T} \cdot \boldsymbol{W}_a \boldsymbol{e}_i)} \quad (13)$$

$$\boldsymbol{e}_i \in \{\tilde{\mathbf{h}}_i, \tilde{\mathbf{x}}_i\}; \ \mathbf{g}_i = \alpha_1 \cdot \tilde{\mathbf{h}}_i \oplus \alpha_2 \cdot \tilde{\mathbf{x}}_i$$

$\tilde{\mathbf{x}}_i$ is the average of $\mathbf{x}^{(l)}$ across all convolutional layers.

$$\mathcal{L} = \frac{1}{|T|}\sum_{t\in T, m\in M} -\log\left(\frac{\exp \hat{y}_{m,t}}{\sum_{j\in V}\exp \hat{y}_{m,j}}\right) \quad (14)$$

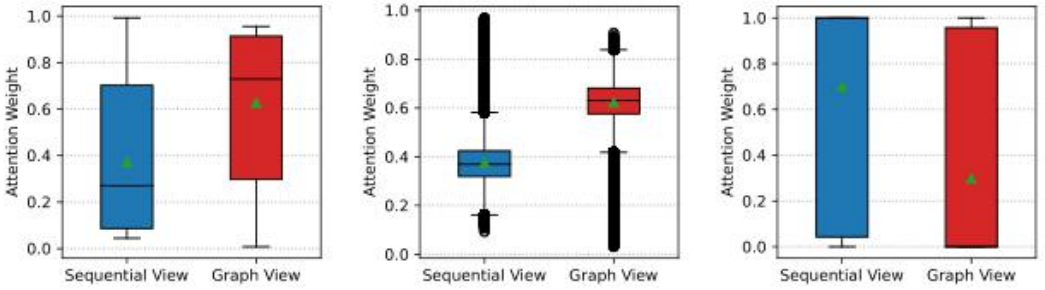
**Table 1: Statistical information of experimented datasets.**

| Stats. | Taobao | Retailrocket | IJCAI |
|---|---|---|---|
| # Users | 147, 892 | 11, 649 | 200, 000 |
| # Items | 99, 038 | 36, 223 | 808, 354 |
| # Interactions | 7, 092, 362 | 87, 822 | 13, 072, 940 |
| # Average Length | 48.23 | 14.55 | 78.58 |
| # Density | $5 \times 10^{-6}$ | $1 \times 10^{-6}$ | $7 \times 10^{-7}$ |
| # Behavior Types | [buy, cart, fav, pv] | [buy, cart, pv] | [buy, cart, fav, pv] |

**Table 2: The performance of our method and the best performed baseline are presented with bold and underlined, respectively. Superscript $*$ indicates the significant improvement between our MBHT and the best performed baseline with $p$ value $< 0.01$.**

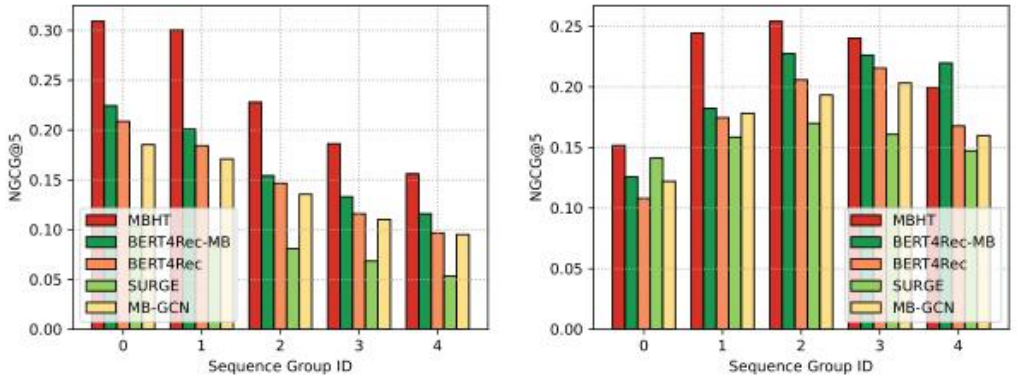
| Model | Taobao | | | | | Retailrocket | | | | | IJCAI | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | HR@5 | NDCG@5 | HR@10 | NDCG@10 | MRR | HR@5 | NDCG@5 | HR@10 | NDCG@10 | MRR | HR@5 | NDCG@5 | HR@10 | NDCG@10 | MRR |
| *General Sequential Recommendation Methods* | | | | | | | | | | | | | | | |
| Caser | 0.082 | 0.058 | 0.123 | 0.071 | 0.070 | 0.632 | 0.539 | 0.754 | 0.578 | 0.535 | 0.134 | 0.092 | 0.167 | 0.104 | 0.109 |
| HPMN | 0.162 | 0.130 | 0.219 | 0.141 | 0.139 | 0.664 | 0.633 | 0.711 | 0.587 | 0.602 | 0.144 | 0.085 | 0.197 | 0.124 | 0.123 |
| GRU4Rec | 0.147 | 0.105 | 0.209 | 0.125 | 0.118 | 0.640 | 0.575 | 0.708 | 0.597 | 0.572 | 0.141 | 0.100 | 0.200 | 0.119 | 0.113 |
| SASRec | 0.150 | 0.110 | 0.206 | 0.128 | 0.123 | 0.669 | 0.644 | 0.689 | 0.650 | 0.645 | 0.146 | 0.110 | 0.191 | 0.124 | 0.122 |
| BERT4Rec | 0.198 | 0.153 | 0.254 | 0.171 | 0.163 | 0.808 | 0.670 | 0.881 | 0.694 | 0.639 | 0.297 | 0.220 | 0.402 | 0.253 | 0.227 |
| *Graph-based Sequential Recommender Systems* | | | | | | | | | | | | | | | |
| SR-GNN | 0.102 | 0.071 | 0.153 | 0.087 | 0.086 | 0.848 | 0.780 | 0.891 | 0.793 | 0.767 | 0.072 | 0.048 | 0.118 | 0.062 | 0.064 |
| GCSAN | 0.217 | 0.160 | 0.305 | 0.188 | 0.173 | 0.872 | 0.846 | 0.890 | 0.851 | 0.842 | 0.119 | 0.086 | 0.175 | 0.104 | 0.101 |
| HyperRec | 0.145 | 0.130 | 0.224 | 0.133 | 0.129 | 0.860 | 0.705 | 0.833 | 0.820 | 0.816 | 0.140 | 0.109 | 0.236 | 0.144 | 0.132 |
| SURGE | 0.122 | 0.078 | 0.193 | 0.100 | 0.093 | 0.906 | 0.879 | 0.878 | 0.887 | 0.870 | 0.226 | 0.159 | 0.322 | 0.190 | 0.171 |
| *Multi-Behavior Recommendation Models* | | | | | | | | | | | | | | | |
| BERT4Rec-MB | 0.211 | 0.169 | 0.263 | 0.186 | 0.178 | 0.875 | 0.858 | 0.889 | 0.863 | 0.857 | 0.257 | 0.189 | 0.342 | 0.216 | 0.197 |
| MB-GCN | 0.185 | 0.103 | 0.309 | 0.143 | 0.149 | 0.878 | 0.735 | 0.844 | 0.752 | 0.739 | 0.218 | 0.145 | 0.335 | 0.182 | 0.177 |
| NMTR | 0.125 | 0.082 | 0.174 | 0.097 | 0.103 | 0.858 | 0.697 | 0.827 | 0.724 | 0.741 | 0.109 | 0.076 | 0.184 | 0.099 | 0.106 |
| MB-GMN | 0.196 | 0.115 | 0.319 | 0.154 | 0.151 | 0.901 | 0.762 | 0.853 | 0.830 | 0.822 | 0.235 | 0.161 | 0.337 | 0.193 | 0.176 |
| **MBHT** | **0.323**$^*$ | **0.257**$^*$ | **0.405**$^*$ | **0.283**$^*$ | **0.262**$^*$ | **0.956**$^*$ | **0.950**$^*$ | **0.931**$^*$ | **0.933**$^*$ | **0.929**$^*$ | **0.346**$^*$ | **0.268**$^*$ | **0.437**$^*$ | **0.297**$^*$ | **0.272**$^*$ |
| # Improve | 48.84% | 52.07% | 26.95% | 50.53% | 47.19% | 5.17% | 8.07% | 6.36% | 5.19% | 6.78% | 16.50% | 21.82% | 8.71% | 17.39% | 19.82% |

# Experiment




(a) Taobao (b) IJCAI (c) Retailrocket

**Figure 3: Distributions of the learned attentive view-specific contributions. Green triangles and black line in the showed boxes denote the mean and median values, respectively.**




(a) Taobao (b) IJCAI

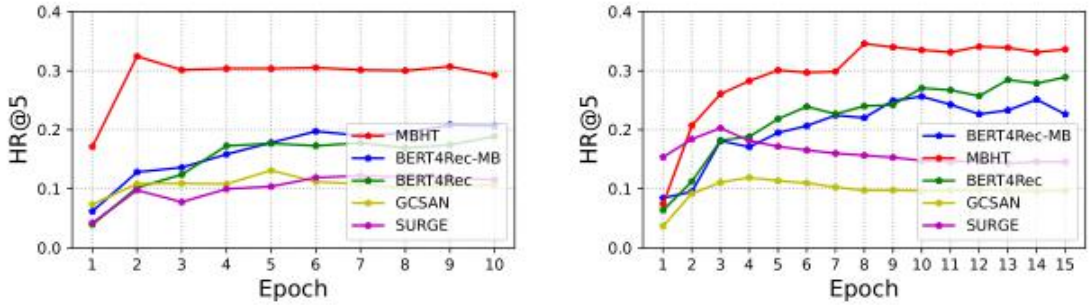
**Figure 4: Performance *w.r.t* different sequence lengths.**

**Table 3: Ablation study with key modules.**

| Model Variants | Taobao | | Retailrocket | | IJCAI | |
|---|---|---|---|---|---|---|
| | HR@5 | NDCG@5 | HR@5 | NDCG@5 | HR@5 | NDCG@5 |
| MBHT | **0.323** | **0.257** | **0.956** | **0.950** | **0.346** | **0.268** |
| *(-) MB-Hyper* | 0.261 | 0.206 | 0.883 | 0.861 | 0.320 | 0.249 |
| *(-) ML-Hyper* | 0.271 | 0.212 | 0.898 | 0.874 | 0.328 | 0.256 |
| *(-) Hypergraph* | 0.246 | 0.194 | 0.813 | 0.839 | 0.301 | 0.234 |
| *(-) MS-Attention* | 0.253 | 0.200 | 0.816 | 0.832 | 0.329 | 0.256 |

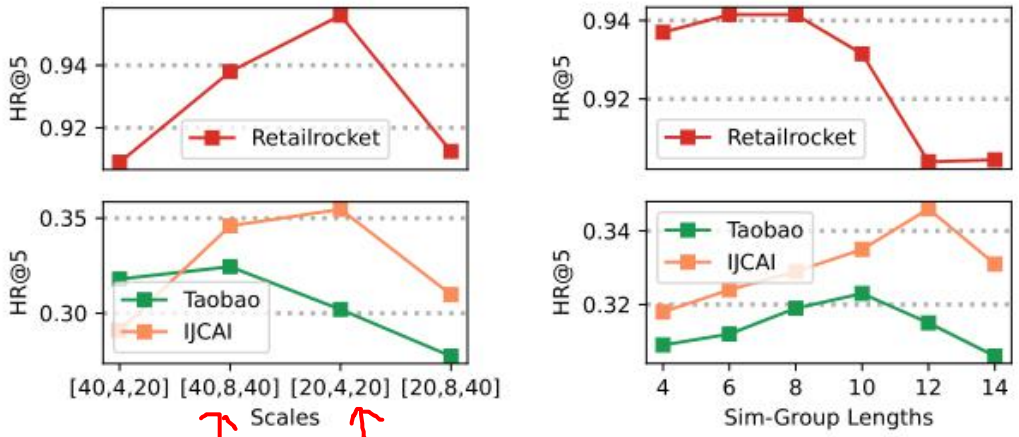
- *(-) MB-Hyper.* This variant does not include the hypergraph of item-wise multi-behavior dependency to capture the long-range cross-type behavior correlations.
- *(-) ML-Hyper.* In this variant, we remove the hypergraph message passing over the hyperedges of item semantic dependence (encoded with the metric learning component).
- *(-) Hypergraph.* This variant disables the entire hypergraph item-wise dependency learning, and only relies on the multi-scale Transformer to model the sequential behavior patterns.
- *(-) MS-Attention.* For this variant, we replace our multi-scale attention layer with the original multi-head attentional operation.

# Experiment




(a) Taobao (b) IJCAI

**Figure 5: Training curves evaluated by testing Hit Rate.**




(a) Sensitivity to attention scales (b) Sensitivity to sim-group length

**Figure 7: Hyperparameter study of MBHT framework.**

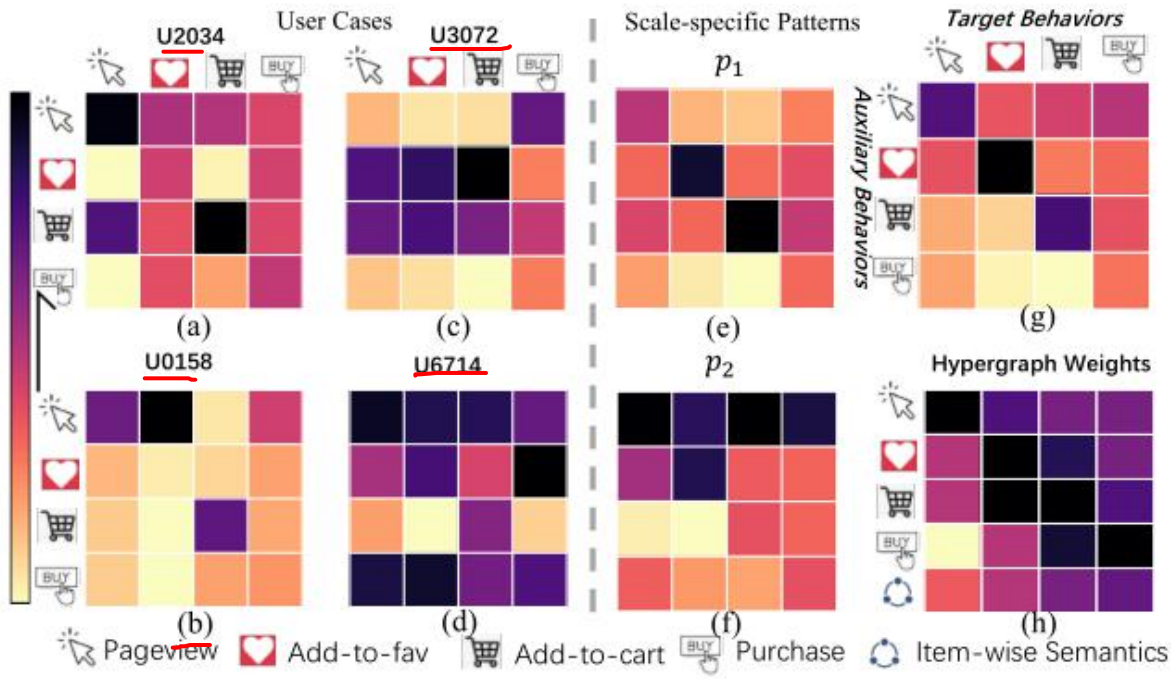



**Figure 6: Case studies with cross-type behavior dependencies.**

# Thanks