# Continuous-Time Sequential Recommendation with Temporal Graph Collaborative Transformer

Ziwei Fan*, Zhiwei Liu*
Department of Computer Science,
University of Illinois at Chicago
USA
{zfan20,zliu213}@uic.edu

Jiawei Zhang
IFM Lab, Department of Computer Science, University of California, Davis
USA
jiawei@ifmlab.org

Yun Xiong
Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University
China
yunx@fudan.edu.cn

Lei Zheng
Pinterest Inc.
USA
lzheng@pinterest.com

Philip S. Yu
Department of Computer Science,
University of Illinois at Chicago
USA
psyu@uic.edu

# Problem Formulation

- **Continuous-Time Sequential Recommendation:** Given user set $\mathcal{U}$, item set $I$, and a set of feature timestamps $\mathcal{T}_u > T$, For a specific user $u$, the continuous-time sequential recommendation is to generate a ranking list of items from $I \backslash I_u(t)$ for every timestamp $t \in \mathcal{T}_u$, where the items that $u$ is interested will be ranked top in the list.

# Motivation

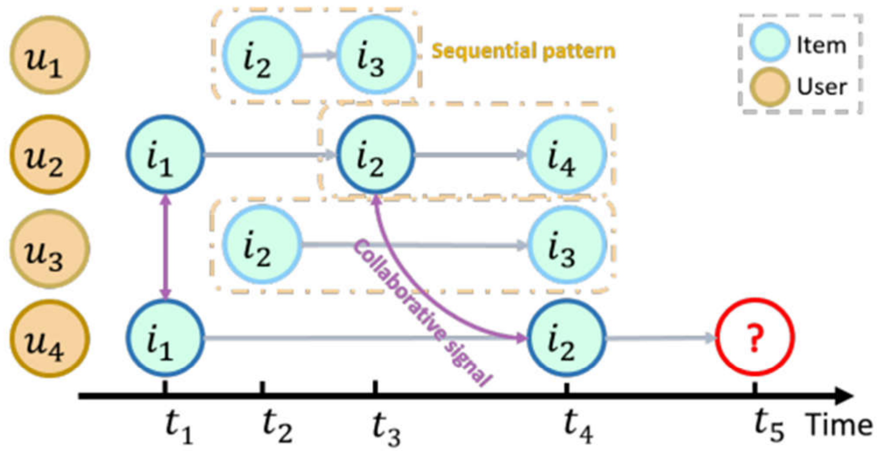
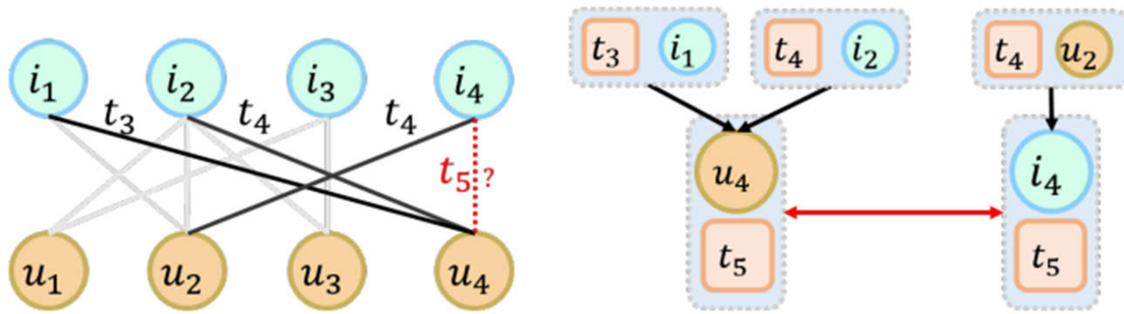



**Figure 1: A toy example of temporal collaborative signals. Given the items that users $u_1, u_2, u_3$ and $u_4$ like in the past timestamps $t_1, t_2, t_3$ and $t_4$, the target is to recommend an item to $u_4$ at $t_5$ as the next item after $i_2$.**

- How to encode collaborative signals and sequential patterns simultaneously?
- How to express the temporal effects of collaborative signals effectively?

# Continuous Time Bipartite Graph (CTBG)




(a) Example of CTBG (b) Temporal Inference

**Figure 2: The associated CTBG of Figure 1 and the inference of temporal embeddings of $u_4$ and $i_4$ at $t_5$.**

- Based on timestamps and neighbor items of user preserve sequential patterns, the CTBG is constructed for sequential patterns and collaborative signals unification.
- Devising Temporal Collaborative Transformer (TCT) layer, which adopts collaborative attention among user-item interactions to capture temporal collaborative signals.

# Temporal Graph Sequential Recommender (TGSRec)

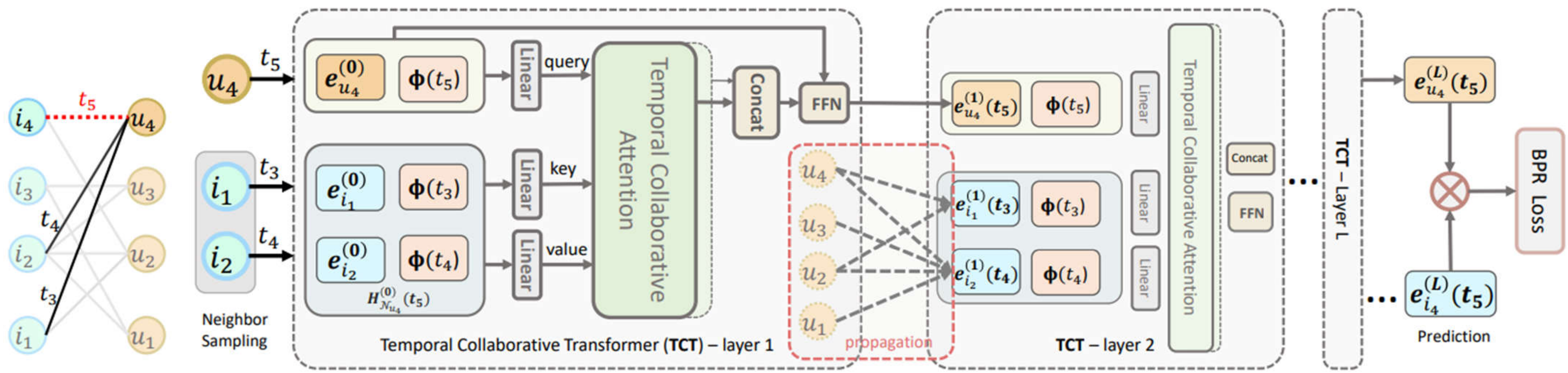



**Figure 3: The framework of TGSRec. The query node is $u_4$, whose final temporal embedding at time $t_5$ is $h_{u_4}^{(2)}(t_5)$. The TCT layer samples its neighbor nodes and edges. Timestamps on edges are encoded as vectors by using mapping function $\Phi$. Node embeddings for the first TCT layer are long-term embeddings. Node embeddings for other TCT layers (*e.g.* layer 2) are propagated from the previous TCT layer, thus being temporal node embeddings.**

# Embedding Layer

- **User/Item Embeddings**

The user(item) node is parameterized by a vector $\boldsymbol{e_u}(\boldsymbol{e_i}) \in \mathbb{R}^d$, which can be generated by indexing an embedding table $\boldsymbol{E} = [\boldsymbol{E_{\mathcal{U}}}; \boldsymbol{E_I}] \in \mathbb{R}^{d\times|\mathcal{V}|}$, where $\mathcal{V} = \mathcal{U} \cup I$. Note that $\boldsymbol{E}$ can be optimized in CTBG.

- **Continuous-Time Embedding**

To explicitly representing the temporal feature, the temporal embedding can be represented by:

$$\Phi(t) \mapsto \sqrt{\frac{1}{d_T}}[cos(\omega_1 t), \sin(\omega_1 t), \ldots, \cos(\omega_{d_T} t), sin(\omega_{d_T} t)]^T$$

where $\omega = [\omega_1, \ldots, \omega_{d_T}]^T$ are learnable and $d_T$ is the dimension.

# Temporal Collaborative Transformer

- **Information Construction**

➢ For user $u$, the information representation can be formulated by:

$$\boldsymbol{h}_u^{(l-1)}(t) = \boldsymbol{e}_u^{(l-1)}(t)||\Phi(t)$$

where $l = 1,2,..,L$. $\boldsymbol{h}_u^{(l-1)}(t) \in \mathbb{R}^{d+d_T}$ is the information for user $u$ at $t$, $\boldsymbol{e}_u^{(l-1)}(t) \in \mathbb{R}^d$ is the temporal embedding of $u$. When $l = 1$, the temporal embedding $\boldsymbol{e}_u^{(0)}(t) = \boldsymbol{E}_u$.

➢ Randomly sampling $S$ different interactions of $u$ before time $t$ as $\mathcal{N}_u(t) = \{(i, t_s)|(u, i, t_s) \in \mathcal{E}_t \; and \; t_s < t\}$. For pair $(i, t_s)$, the information representation can be formulated by:

$$\boldsymbol{h}_i^{(l-1)}(t_s) = \boldsymbol{e}_i^{(l-1)}(t_s)||\Phi(t_s)$$

where $\boldsymbol{h}_i^{(l-1)}(t_s)$ is the information for item i at $t_s$. $\boldsymbol{e_i}(t_s)$ denotes the temporal embedding of i at $t_s$. Again, when $l = 1$, $\boldsymbol{e}_i^{(0)}(t_s) = \boldsymbol{E}_i$.

# Temporal Collaborative Transformer

- **Information Propagation**

The temporal embedding $\boldsymbol{e}_{\mathcal{N}_u}^{(l)}(t)$ can be inferred via propagating the information of sampled neighbors $\mathcal{N}_u(t)$,

$$\boldsymbol{e}_{\mathcal{N}_u}^{(l)}(t) = \sum_{(i,t_s)\in\mathcal{N}_u(t)} \pi_t^u(i, t_s)\, \boldsymbol{W}_v^{(l)} \boldsymbol{h}_i^{(l-1)}(t_s)$$

where $\pi_t^u(i, t_s)$ represents the impact of a historical interaction $(u, i, t_s)$ to the temporal inference of user $u$ at time $t$, which is calculated by the temporal collaborative attention. $\boldsymbol{W}_v \in \mathbb{R}^{d\times(d+d_T)}$ is the linear transformation matrix.

# Temporal Collaborative Transformer

- **Temporal Collaborative Attention**

The attention weight $\pi_t^u(i, t_s)$ is formulated as follows:

$$\pi_t^u(i, t_s) = \frac{1}{\sqrt{d + d_T}} (\boldsymbol{W}_k^{(l)} \boldsymbol{h}_i^{(l-1)}(t_s)) \, \boldsymbol{W}_q^{(l)} \boldsymbol{h}_u^{(l-1)}(t)$$

where $\boldsymbol{W}_k^{(l)}$ and $\boldsymbol{W}_q^{(l)}$ are both linear transformation matrices. If we ignore the transformation matrices and the scalar factor, the right-hand side of above equation can be rewritten as:

$$\boldsymbol{e}_u^{(l-1)}(t) \cdot \boldsymbol{e}_i^{(l-1)}(t_s) + \Phi(t) \cdot \Phi(t_s)$$

where the first term denotes the user-item collaborative signal, and the second term models the temporal effect according to the temporal kernel trick $\psi(t_1 - t_2) = \mathcal{K}(t_1, t_2) = \Phi(t_1) \cdot \Phi(t_2)$.

# Temporal Collaborative Transformer

- **Temporal Collaborative Attention**

The attention weights across all sampled interactions can be normalized by employing a softmax function:

$$\pi_t^u(i, t_s) = \frac{\exp(\pi_t^u(i, t_s))}{\sum_{(i', t_s') \in N_u(t)} \exp(\pi_t^u(i', t_s'))}$$

To be more specific, denoting $\boldsymbol{q}_u^{(l-1)}(t) = \boldsymbol{W}_q^{(l)} \boldsymbol{h}_u^{(l-1)}(t)$, $\boldsymbol{K}_u^{(l-1)}(t) = \boldsymbol{W}_k^{(l)} \boldsymbol{H}_{N_u}^{(l-1)}(t)$, and $\boldsymbol{V}_u^{(l-1)}(t) = \boldsymbol{W}_v^{(l)} \boldsymbol{H}_{N_u}^{(l-1)}(t)$ are respectively the key, value, and query input for the temporal collaborative attention module. $\boldsymbol{H}_{N_u}^{(l-1)}(t) \in \mathbb{R}^{(d+d_T) \times S}$ is the stacked information of all sampled interaction $\boldsymbol{h}_i^{(l-1)}(t_s)$. Then, $\boldsymbol{e}_{N_u}$ can be rewritten as:

$$\boldsymbol{e}_{N_u} = \boldsymbol{V}_u \cdot Softmax\left(\boldsymbol{K}_u^T \boldsymbol{q}_u / \sqrt{d + d_T}\right)$$

# Temporal Collaborative Transformer

- **Information Aggregation**

A feed-forward neural network (FFN) is applied for information aggregation:

$$\boldsymbol{e}_u^{(l)}(t) = FFN\left(\boldsymbol{e}_{\mathcal{N}_u}^{(l)}(t) || \boldsymbol{h}_u^{(l-1)}(t)\right)$$

where $\boldsymbol{e}_u^{(l)}(t)$ is the temporal embedding of $u$ at $t$ on $l$-th layer, and FFN consists two linear transformation layers with a ReLU activation function.

# Model Prediction and Optimization

- **Model Prediction**

For each test triplet $(u, i, t)$, it yields temporal embeddings for both $u$ and $i$ at $t$ on the last TCT layer, denoting as $\boldsymbol{e}_u^{(L)}(t)$ and $\boldsymbol{e}_i^{(L)}(t)$, respectively. Then, the prediction score $r(u, i, t)$ is:

$$r(u, i, t) = \boldsymbol{e}_u^{(L)}(t) \cdot \boldsymbol{e}_i^{(L)}(t)$$

- **Model Optimization**

The BRP loss is used for model optimization,

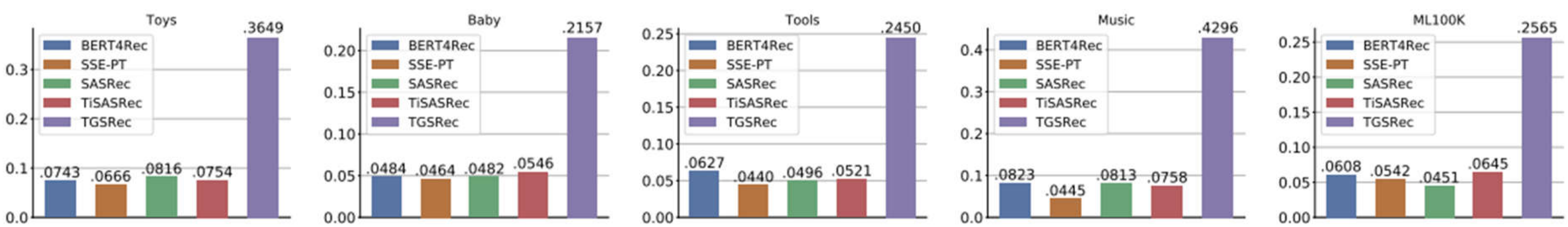
$$\mathcal{L}_{bpr} = \sum_{(u,i,j,t)\in\mathcal{O}_T} -log\sigma\big(r(u, i, t) - r(u, j, t)\big) + \lambda||\Theta||_2^2$$

where $\mathcal{O}_T = \{(u, i, j, t)|(u, i, t) \in \mathcal{E}_T, j \in I\backslash I_u(t)\}$. The positive interaction $(u, i, t)$ comes from the edge set $\mathcal{E}_T$ of CTBG, the negative item $j$ is sampled from unobserved items $I\backslash I_u(t)$ of user $u$ at timestamp $t$. $\Theta$ denotes the training samples.

# Experiments

Table 2: Overall Performance w.r.t. Recall@{10,20} and MRR.

| Datasets | Metric | BPR | LightGCN | SR-GNN | GRU4Rec | Caser | SSE-PT | BERT4Rec | SASRec | TiSASRec | CDTNE | TGSRec | Improv. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Toys | Recall@10 | 0.0021 | 0.0016 | 0.0020 | 0.0274 | 0.0138 | 0.1213 | 0.1273 | 0.1452 | 0.1361 | 0.0016 | **0.3650** | 0.2198 |
| | Recall@20 | 0.0036 | 0.0026 | 0.0033 | 0.0288 | 0.0238 | 0.1719 | 0.1865 | 0.2044 | 0.1931 | 0.0045 | **0.3714** | 0.1670 |
| | MRR | 0.0024 | 0.0018 | 0.0018 | 0.0277 | 0.0082 | 0.0595 | 0.0643 | 0.0732 | 0.0671 | 0.0025 | **0.3661** | 0.2929 |
| Baby | Recall@10 | 0.0028 | 0.0036 | 0.0030 | 0.0036 | 0.0077 | 0.0911 | 0.0884 | 0.0975 | 0.1040 | 0.0218 | **0.2235** | 0.1195 |
| | Recall@20 | 0.0039 | 0.0045 | 0.0062 | 0.0048 | 0.0193 | 0.1418 | 0.1634 | 0.1610 | 0.1662 | 0.0292 | **0.2295** | 0.0663 |
| | MRR | 0.0019 | 0.0024 | 0.0024 | 0.0028 | 0.0071 | 0.0434 | 0.0511 | 0.0455 | 0.0521 | 0.0157 | **0.2147** | 0.1626 |
| Tools | Recall@10 | 0.0023 | 0.0021 | 0.0051 | 0.0048 | 0.0077 | 0.0775 | 0.1296 | 0.0913 | 0.0946 | 0.0186 | **0.2457** | 0.1161 |
| | Recall@20 | 0.0036 | 0.0035 | 0.0092 | 0.0059 | 0.0161 | 0.1155 | 0.1784 | 0.1337 | 0.1356 | 0.0271 | **0.2559** | 0.0775 |
| | MRR | 0.0026 | 0.0023 | 0.0028 | 0.0051 | 0.0068 | 0.0419 | 0.0628 | 0.0460 | 0.0480 | 0.0203 | **0.2468** | 0.1840 |
| Music | Recall@10 | 0.0122 | 0.0142 | 0.0051 | 0.0549 | 0.0183 | 0.0915 | 0.1352 | 0.1372 | 0.1372 | 0.0071 | **0.5935** | 0.4563 |
| | Recall@20 | 0.0152 | 0.0183 | 0.0092 | 0.0589 | 0.0346 | 0.1494 | 0.2093 | 0.2094 | 0.1951 | 0.0163 | **0.5986** | 0.3892 |
| | MRR | 0.0057 | 0.0064 | 0.0028 | 0.0540 | 0.0106 | 0.0423 | 0.0824 | 0.0768 | 0.0681 | 0.0037 | **0.3820** | 0.2996 |
| ML100k | Recall@10 | 0.0461 | 0.0565 | 0.0045 | 0.0996 | 0.0246 | 0.1079 | 0.1116 | 0.09450 | 0.1332 | 0.0350 | **0.3118** | 0.1786 |
| | Recall@20 | 0.0766 | 0.0960 | 0.0060 | 0.1168 | 0.0417 | 0.1801 | 0.1786 | 0.1808 | 0.2232 | 0.0536 | **0.3252** | 0.1020 |
| | MRR | 0.0213 | 0.0252 | 0.0012 | 0.0938 | 0.0147 | 0.0519 | 0.0600 | 0.0448 | 0.0605 | 0.0162 | **0.2416** | 0.1478 |




(a) NDCG@10 in Toys (b) NDCG@10 in Baby (c) NDCG@10 in Tools (d) NDCG@10 in Music (e) NDCG@10 in ML100K

Figure 4: NDCG@10 Performance in all Datasets. We ignore other methods because of their low values.

# Experiments

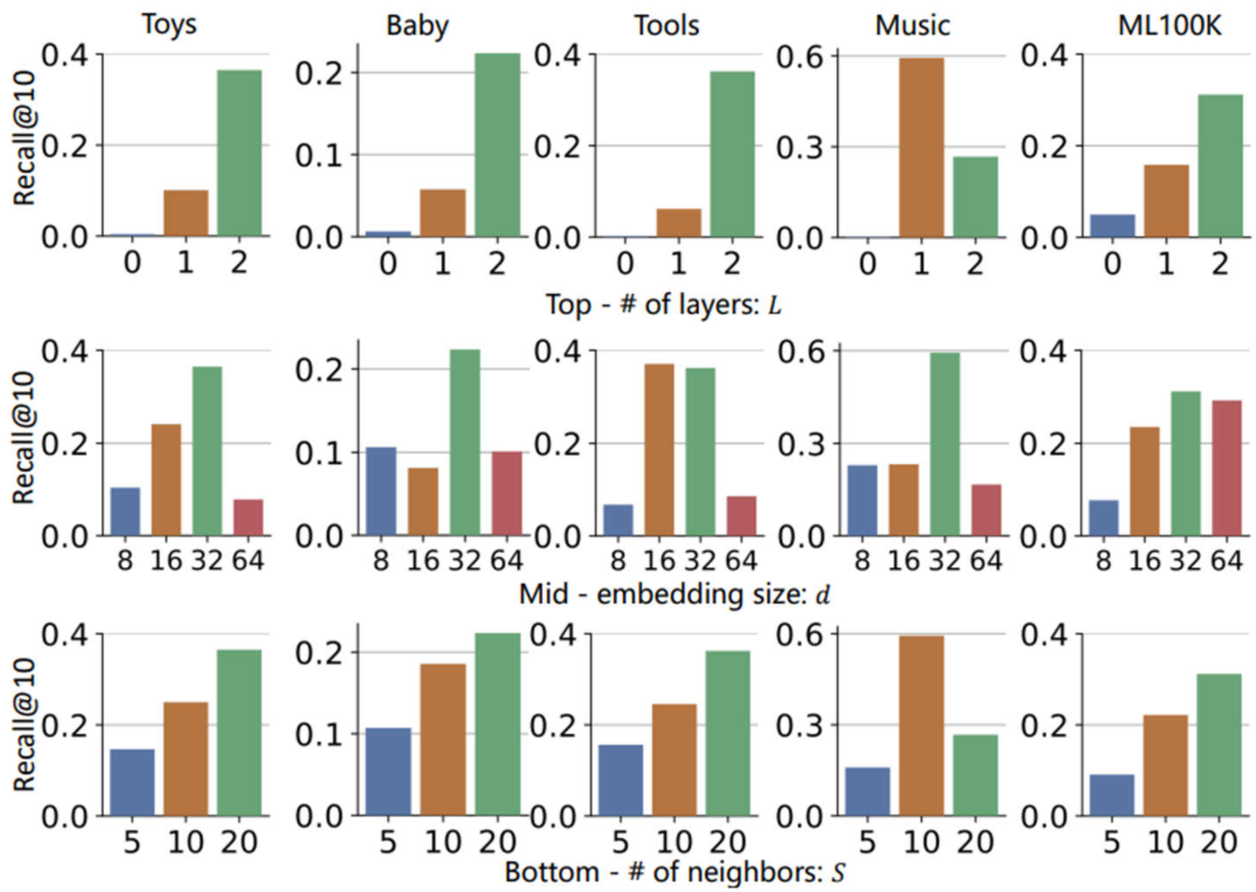



**Figure 5: Recall@10 w.r.t. $L$, $d$ and $S$ on 5 datasets.**

# Experiments

**Table 3: Ablation analysis (Recall@10) on five datasets. Bold score indicates performance better than the default version, while ↓ indicates a performance drop more than 50%.**

| Architecture | Toys | Baby | Tools | Music | ML100K |
|---|---|---|---|---|---|
| (0) Default | **0.3649** | **0.2235** | **0.3623** | **0.5935** | 0.3118 |
| (1) Mean | 0.0027↓ | 0.0210↓ | 0.0055↓ | 0.0051↓ | 0.0647↓ |
| (2) LSTM | 0.0991↓ | 0.1237 | 0.1266↓ | 0.3740 | 0.3088 |
| (3) Fixed $\omega$ | 0.0854↓ | 0.0944↓ | 0.0910↓ | 0.3679 | 0.2789 |
| (4) Position | 0.0380↓ | 0.0243↓ | 0.0209↓ | 0.0742↓ | 0.0878↓ |
| (5) Empty | 0.0139↓ | 0.0240↓ | 0.0018↓ | 0.0346↓ | 0.0603↓ |
| (6) BCELoss | 0.2200 | 0.1916 | 0.1763↓ | 0.4624 | **0.3542** |

**Table 4: Variants of Temporal Information Construction**

| Variant | Toys | Baby | Tools | Music | ML100K |
|---|---|---|---|---|---|
| TGSRec | **0.3649** | **0.2235** | **0.3623** | **0.5935** | **0.3118** |
| $\mathcal{U}$ w/o T | 0.0103 | 0.0138 | 0.0106 | 0.0112 | 0.1555 |
| $\mathcal{I}$ w/o T | <u>0.1013</u> | <u>0.0961</u> | <u>0.0836</u> | <u>0.2785</u> | <u>0.2336</u> |