

# Correlation-Attention Masked Temporal Transformer for User Identity Linkage Using Heterogeneous Mobility Data

Ziang Yan<sup>1</sup>, Xingyu Zhao<sup>1</sup>, Hanqing Ma<sup>1</sup>,  
Wei Chen<sup>2</sup>, Jianpeng Qi<sup>1</sup>, Yanwei Yu<sup>1\*</sup>, Junyu Dong<sup>1</sup>

<sup>1</sup>Faculty of Information Science and Engineering, Ocean University of China, Qingdao, China

<sup>2</sup>The Hong Kong University of Science and Technology (Guangzhou)

{yza, zhaoxingyu, mahanqing}@stu.ouc.edu.cn, onedeanxxx@gmail.com,

{qjianpeng, yuyanwei, dongjunyu}@ouc.edu.cn

## Abstract

With the rise of social media and Location-Based Social Networks (LBSN), check-in data across platforms has become crucial for User Identity Linkage (UIL). These data not only reveal users' spatio-temporal information but also provide insights into their behavior patterns and interests. However, cross-platform identity linkage faces challenges like poor data quality, high sparsity, and noise interference, which hinder existing methods from extracting cross-platform user information. To address these issues, we propose a Correlation-Attention **M**asked **T**ransformer for User Identity **L**inkage Network (MT-Link), a transformer-based framework to enhance model performance by learning spatio-temporal co-occurrence patterns of cross-platform users. Our model effectively captures spatio-temporal co-occurrence in cross-platform user check-in sequences. It employs a correlation attention mechanism to detect the spatio-temporal co-occurrence between user check-in sequences. Guided by attention weight maps, the model focuses on co-occurrence points while filtering out noise, ultimately improving classification performance. Experimental results show that our model significantly outperforms state-of-the-art baselines by 12.92%~17.76% and 5.80%~8.38% improvements in terms of Macro-F1 and Area Under Curve (AUC).

## Introduction

Location-Based Social Network (LBSN) services like Twitter and Foursquare have made daily life more convenient, generating vast amounts of spatio-temporal human mobility data (Qin et al. 2023), such as point-of-interest (POI) check-in sequences (Zhao et al. 2020). The availability of such spatio-temporal data provides a foundation for exploring User Identity Linkage (UIL) tasks (Riederer et al. 2016), which holds significant potential in areas such as recommendation systems (Chen et al. 2020), user behavior analysis, and privacy protection (Qi et al. 2018). Several studies (Goga et al. 2013; Rossi and Musolesi 2014; Naini et al. 2015) have leveraged spatio-temporal check-in data to match user identities based on mobility trajectories, demonstrating the effectiveness of such data.

Previous methods have had some success (Goga et al. 2013; Rossi and Musolesi 2014; Naini et al. 2015) but are

\*Corresponding author: Yanwei Yu.



Figure 1: Spatio-temporal check-ins of the same user across platforms, along with different behavioral patterns.

limited by the discrete and sparse nature of the data, restricting their ability to capture continuous behavior and model cross-platform interactions. The performance of these methods is often constrained by the dataset distribution and lacks a deep exploration of hidden patterns within user check-in sequences. Despite these limitations, data mining approaches have made significant progress in addressing the spatio-temporal UIL problem (Riederer et al. 2016; Basik et al. 2017; Li, Zhu, and Xie 2019; Basik, Ferhatosmanoğlu, and Gedik 2020). These methods effectively link the same user across different platforms by focusing on spatial and temporal correlations. Recently, clustering-based methods (Ding et al. 2020; Ma et al. 2022; Chen et al. 2017; Xue et al. 2021) and kernel density estimation methods (Chen et al. 2018, 2023) have further attempted to solve the challenges posed by sparse spatio-temporal check-in data, thereby improving identification accuracy. Notably, deep learning approaches (Feng et al. 2019, 2020; Li et al. 2023) have also achieved impressive success in capturing abstract representations of user check-in sequences, demonstrating significant potential in modeling complex spatio-temporal data.

Although significant progress has been made in spatio-temporal user identity linkage, we identify three key limitations in most existing methods, which hinder the accuracy and robustness of the linkage process. *First, current research mainly focuses on solving the problem of data sparsity, but introduces inherent noise in the check-in sequence and ignores the information of co-occurrence points.* For example, Figure 1 clearly shows that the check-in frequency of the same user varies across different platforms. We define check-ins that occur at the same time and place as spatio-temporal co-occurrence points. If we consider all check-in points to determine whether these check-in points are generated by the same user, random check-ins (*i.e.*, the red points)

may interfere with the judgment, increasing the likelihood of errors. However, by focusing on co-occurrence points, we can clearly see that platform A and platform B share the same check-in patterns at specific times and locations (*i.e.*, the blue points), and cross-platform user co-occurrence information is easier to mine. *Second, current deep learning approaches typically model each user’s spatio-temporal sequence independently, without simultaneously considering the spatio-temporal co-occurrence of key points across different user trajectories.* For example, in the different trajectories on A and B shown in Figure 1, if we independently learn the specific patterns of these trajectories, noise interference and the lack of co-occurrence capture at key points could negatively impact the model’s classification performance, leading to incorrect associations between the current user and unrelated users. *Third, the process of manually constructing features and frequent pattern mining of existing data mining methods is not only very time-consuming, but also fails to effectively utilize known real labels in the data.* Lack of true value labels leads to the inability to learn the relationship between input sample pairs and users, which affects the performance of the model.

To address these challenges, we propose a transformer-based symmetric deep learning model called Correlation Attention **M**asked **T**ransformer for User Identity **L**inkage Network (MT-Link). Specifically, we design a temporal transformer encoder and a masked transformer encoder. The dense representations from the spatio-temporal embedding layer are fed into the temporal transformer encoder to model the check-in sequences. Then, we use a correlation attention block to collaboratively capture the spatio-temporal co-occurrences between the check-in sequence representations of users on different platforms. Based on the captured spatio-temporal co-occurrence, we guide the masking of the tokens with the lowest attention in the dense representations, helping us to filter out noise points and retain co-occurrence points. Finally, the user identity linkage layer links the outputs of the user check-in sequences from both platforms and produces the prediction probability. The results on four real-world cross-platform datasets demonstrate that our model outperforms the latest deep learning methods and data mining approaches in the spatio-temporal UIL task. In summary, our contributions are as follows:

- We propose a novel MT-Link method that combines a masking mechanism with a masked transformer encoder to retain key information and ignore noise, enhancing the co-occurrence between information, thereby improving user identity linkage performance.
- We propose the correlation attention mechanism to capture spatio-temporal co-occurrences between cross-platform user check-in sequence representations, better guiding our masking process.
- We conduct extensive experimental evaluations on four cross-platform datasets. Experimental results show that our model significantly outperforms state-of-the-art baselines by 12.92%~17.76% and 5.80%~8.38% improvements in terms of Macro-F1 and Area Under Curve (AUC).

## Related Work

The existing research on UIL can be roughly divided into two categories: UIL based on traditional data mining methods and UIL based on deep learning methods.

**Data Mining-Based Methods.** In research based on traditional data mining methods, (Naini et al. 2015) focuses on calculating the frequency of a user’s visits to each location, then uses Kullback-Leibler divergence to define the similarity score between two histograms. (Riederer et al. 2016) is an alignment algorithm that calculates an affinity score based on timestamped location data, and then uses a maximum weighted matching scheme to identify the most likely matching user identities. STUL (Chen et al. 2017) extracts spatio-temporal features using density clustering and Gaussian mixture models to measure user similarity. Subsequently, (Chen et al. 2018) employs kernel density estimation to improve accuracy by mitigating sparsity issues. (Wang et al. 2018) uses a set matching algorithm to identify candidate sets of the same user, employing Bayesian inference for ranking confidence scores to determine cross-platform data linkage. CP-Link (Ding et al. 2020) and its extensions (Ma et al. 2022) utilize behavior patterns to mine frequent locations and apply an improved dynamic time-warping method for similarity calculation.

**Deep Learning-Based Methods.** To overcome the feature difficulties caused by the high heterogeneity of mobile data from different sources, the deep learning framework DPLink for UIL is first proposed. DPLink (Feng et al. 2019) and its extensions (Feng et al. 2020) address high data heterogeneity across platforms with a pre-training strategy, pioneering the use of end-to-end deep learning for UIL. Subsequently, (Huang et al. 2024) proposes a graph convolutional network method to learn user representations from location-based social relationships, aggregating information from surrounding nodes and generating embeddings that comprehensively represent users and locations. (Long et al. 2023) proposes using graph convolution aggregation to supplement the missing neighborhoods of nodes in two co-author networks, where the deviation of adjacent nodes is trained from two well-structured head nodes and corrected locally.

*Most existing UIL methods rely on data mining techniques, which have limitations in capturing deep co-occurrence across platforms, particularly with nonlinear, heterogeneous, and complex data. Traditional deep learning approaches also fall short in leveraging advanced models to capture the intrinsic spatio-temporal co-occurrences in cross-platform check-in sequences.*

## Problem Definition

We define the user sets  $\mathcal{U}_A = \{u_1, u_2, \dots, u_m\}$  and  $\mathcal{U}_B = \{u_1, u_2, \dots, u_n\}$  as the collections of registered users on two different social platforms  $A$  and  $B$ , where  $m$  and  $n$  are the number of users in  $\mathcal{U}_A$  and  $\mathcal{U}_B$ , respectively.

**Definition 1** (Check-in Point). *A check-in point is represented by a triplet  $(u, t, p)$ . Here,  $u$  is the user ID on the current social platform,  $t$  is the timestamp of the check-in, and  $p$  is the POI.*

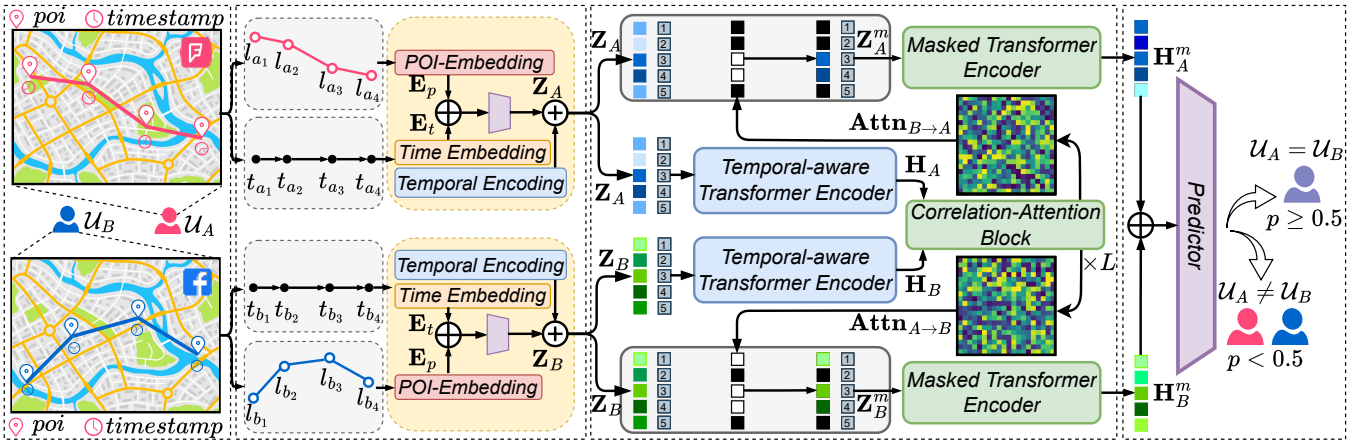


Figure 2: The overview of the proposed framework.

**Definition 2** (Check-in Sequence). For  $\forall u_i \in \mathcal{U}$  in  $A$  and  $B$ , all check-in points are sorted in chronological order by  $t$ , forming a long spatial temporal sequence  $\mathcal{T}_{u_i} = \{(u_i, t_1, p_1), (u_i, t_2, p_2), \dots, (u_i, t_k, p_k)\}$ , where  $i$  is the  $i$ -th user and  $k$  represents the numbers of triplet in the sequence.

**Problem** (User Identity Linkage).  $\forall u_i^A \in \mathcal{U}_A$  and  $\forall u_j^B \in \mathcal{U}_B$ , they have their respective check-in sequences  $\mathcal{T}_{u_i^A}$  and  $\mathcal{T}_{u_j^B}$ . The goal is to construct a mapping function  $f_\theta(\cdot)$  for these check-in sequences,

$$f(\mathcal{T}_{u_i^A}, \mathcal{T}_{u_j^B}; \theta) = \begin{cases} 1, & u_i^A, u_j^B \text{ is the same user.} \\ 0, & u_i^A, u_j^B \text{ is different user.} \end{cases} \quad (1)$$

where  $\theta$  is the parameter of the function.

## Methodology

The architecture of the MT-Link network is depicted in Figure 2. The primary components of the network include: (1) *Spatial-Temporal Embedding Layer*, (2) *Temporal Transformer Encoder*, (3) *Correlation Attention Block*, (4) *Masked Transformer Encoder*, and (5) *User Identity Linkage Layer*. The following sections will detail each module, explaining its functionality and contribution to the model.

### Spatial-Temporal Embedding Layer

Discrete data are typically high-dimensional and sparse, leading to the curse of dimensionality, which complicates processing and analysis (Berisha et al. 2021). Traditional one-hot encoding is unsuitable for our task due to its inability to capture spatio-temporal semantics and its high computational cost (Rodríguez et al. 2018). To better capture user trajectory semantics across platforms, we model check-in sequences  $\mathcal{T}$  from both spatial and temporal dimensions. Specifically, each POI in the check-in sequence  $\mathcal{T}$ , i.e., POI  $p$ , is represented by  $|\mathcal{P}|$ -dimensional one-hot vector. The time of the check-in sequence, i.e., timestamp  $t$ , is represented by  $|T|$ -dimensional one-hot vector by day of the month. By learning the embeddings of each POI, denoted by  $\mathbf{E}_p \in \mathbb{R}^{|\mathcal{P}| \times d_p}$  and time slot denoted by  $\mathbf{E}_t \in \mathbb{R}^{|T| \times d_t}$ , we can capture the semantic information of POIs within each

time slot and the relationships between different time slots in the check-in sequence  $\mathcal{T}$ . Finally, we obtain the embedded sequence  $\mathbf{X} = f_{st}(\mathcal{T}; \theta_{st}) = (x_1, \dots, x_k) \in \mathbb{R}^{k \times d}$ , where  $f_{st}(\cdot)$  denotes the spatial-temporal embedding layer and  $\theta_{st}$  is the learnable parameters in the embedding.

### Temporal Transformer Encoder

Transformer model (Vaswani et al. 2017), utilizing attention mechanisms, efficiently processes long sequences and overcomes the limitations of traditional RNNs. In our task, it is crucial to model long spatio-temporal sequences while preserving temporal correlations to achieve high-quality intra-sequence and inter-sequence representations. Inspired by (Chen et al. 2022), we replace the original transformer’s positional encoding with a temporal positional encoding, allowing time slot information to reflect the temporal changes in the check-in sequences  $\mathcal{T}$ :

$$[\text{TE}(t_i)]_j = \frac{1}{\sqrt{d}} \cdot \cos(\mathbf{w}_j \cdot t_i + \mathbf{b}_j), \quad (2)$$

where  $w_j$  and  $b_j$  are learnable parameters,  $j$  is the  $j$ -th order in the dimension of embedded sequence ( $j \leq d$ ).  $i$  is the  $i$ -th check-in point in check-in sequence  $\mathcal{T}$ . For any two check-in points in  $\mathcal{T}$ , their relative visited check-in time information can be represented as:

$$\text{TE}(t_i) \text{TE}(t_i + \Delta_t)^\top = \sum_{j=1}^d \cos(\mathbf{w}_j \Delta_t). \quad (3)$$

We obtain the final embedded representations for  $\mathcal{T}$  by  $\mathbf{z}_i = \mathbf{x}_i + \text{TE}(t_i)$  with their corresponding temporal positional encoding  $\text{TE}(\cdot)$ , resulting in  $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_k] \in \mathbb{R}^{k \times d}$ . We denote our temporal transformer encoder as  $f_T(\cdot)$  and feed the embedded representations  $\mathbf{Z}$  into it for encoding. Through this process, we obtain the spatio-temporal context representation vectors  $\mathbf{H} \in \mathbb{R}^{k \times d}$  for the user check-in sequences on a single platform.

### Correlation Attention Block

To capture the similar semantic information of check-in sequences from platforms  $A$  and  $B$ , we introduce the Cor-

relation Attention Block of  $L$  layers to capture the co-occurrence of user check-in sequences. Similar methods have been shown to effectively capture cross-domain relationships by leveraging attention mechanisms (Wang et al. 2024), improving the model’s ability to understand and link data across different domains.

In our bidirectional symmetric structure for platform  $A$ , we obtain the high-level sequence representation  $\mathbf{H}_A$  by encoding the previous step  $\mathbf{Z}_A$  through the temporal transformer  $f_T(\cdot)$  from platform  $A$  as the query vector. *i.e.*,  $\mathbf{Q}_A \in \mathbb{R}^{k \times d}$  after linear transformation and the high-level sequence representation  $\mathbf{H}_B$  by encoding the previous step’s  $\mathbf{Z}_B$  through the temporal transformer  $f_T(\cdot)$  from platform  $B$  as the key and value vectors after linear transformation. *i.e.*,  $\mathbf{K}_B \in \mathbb{R}^{k \times d}$ ,  $\mathbf{V}_B \in \mathbb{R}^{k \times d}$ . We then concatenate the output of the attention layer and the high-level sequence representation from platform  $A$  as the input of the next attention layer, and this process is repeated. For platform  $B$ , the logic is identical to that of  $A$ . This process can be represented as follows:

$$\begin{aligned} \mathbf{Q}_A^{l+1} &= \text{LN} \left( \hat{\mathbf{Q}}_A^l + \text{CrossAttn}(\hat{\mathbf{Q}}_A^l, \mathbf{K}_B, \mathbf{V}_B) \right), \\ \mathbf{Q}_B^{l+1} &= \text{LN} \left( \hat{\mathbf{Q}}_B^l + \text{CrossAttn}(\hat{\mathbf{Q}}_B^l, \mathbf{K}_A, \mathbf{V}_A) \right). \end{aligned} \quad (4)$$

Here  $\text{CrossAttn}(\cdot)$  denotes the multi-head cross attention and  $l \in L$ .  $\hat{\mathbf{Q}}_A^l$  is the  $l$ -th layer’s output and will serve as the input for the next layer.  $\text{LN}(\cdot)$  stands for the Layer Normalization layer.

Finally, we output the attention weight map from the last layer of the block:

$$\begin{aligned} \text{Attn}_{A \rightarrow B}^{|L|} &= \frac{1}{H} \sum_{h=1}^H \text{softmax}(\mathbf{Q}_A^{(|L|,h)} \frac{(\mathbf{K}_B^{(|L|,h)})^T}{\sqrt{d_k}}), \\ \text{Attn}_{B \rightarrow A}^{|L|} &= \frac{1}{H} \sum_{h=1}^H \text{softmax}(\mathbf{Q}_B^{(|L|,h)} \frac{(\mathbf{K}_A^{(|L|,h)})^T}{\sqrt{d_k}}), \end{aligned} \quad (5)$$

where  $h$  represents the  $h$ -th attention head, and we average the attention scores from all  $H$  heads to obtain the final output. Each element in  $\text{Attn}_{A \rightarrow B}^{|L|} \in \mathbb{R}^{N_{Q_A} \times N_{K_B}}$  represents the relevance score between each position in sequence  $\mathcal{T}_A$  and each position in sequence  $\mathcal{T}_B$ . The same logic applies to  $\text{Attn}_{B \rightarrow A}^{|L|} \in \mathbb{R}^{N_{Q_B} \times N_{K_A}}$ .

The collaborative effect of the multi-head self-attention layer in  $f_T(\cdot)$  and the multi-head cross-attention in the correlation attention block allows for a more comprehensive representation of both intra-sequence and inter-sequence correlations within the check-in sequences  $\mathcal{T}_A$  and  $\mathcal{T}_B$ . The attention weights provide a finer-grained reflection of the relationships between different check-in points within the sequences. Finally, we use the attention weight map from the last layer to guide our masking process.

### Masked Transformer Encoder

Inspired by the random masking strategy used in natural language processing (Devlin et al. 2018; Li et al. 2021), we explored whether a masking strategy could be effectively applied to the User Identity Linkage (UIL) task. However,

a random masking strategy could result in the loss of critical information within the check-in sequences  $\mathcal{T}$ , potentially impacting the model’s overall performance. We propose an attention-guided masking module to preserve the positional correlations better and capture the co-occurrences within user check-in sequences. This module selectively masks non-essential tokens in the sequence, after which the masked sequence is encoded using a masked transformer encoder  $f_M(\cdot)$ .

In the spatio-temporal embedding layer, we obtain the embeddings  $\mathbf{Z}^A$ ,  $\mathbf{Z}^B$  and  $\text{Attn}_{A \rightarrow B}$ ,  $\text{Attn}_{B \rightarrow A}$  from the correlation attention block. Each token in the input embedding sequence is represented by vector sets  $[\mathbf{z}_1, \dots, \mathbf{z}_k] \in \mathbf{Z}$ . The attention weights along the key vector dimensions of  $\text{Attn}_{A \rightarrow B}$  and  $\text{Attn}_{B \rightarrow A}$  are summed as follows:

$$\begin{aligned} N_{mask}^B, \mathbf{I}^B &= \text{IdxValPair}(r, \sum_{j=1}^{|\mathbf{K}|} (\text{Attn}_{B \rightarrow A}^{|L|})_{ij}), \\ N_{mask}^A, \mathbf{I}^A &= \text{IdxValPair}(r, \sum_{j=1}^{|\mathbf{K}|} (\text{Attn}_{A \rightarrow B}^{|L|})_{ij}), \end{aligned} \quad (6)$$

where the  $\text{IdxValPair}(\cdot)$  takes as input two components: a one-dimensional vector, which is the sum of the attention vector along the  $j$ -th dimension of the key vector  $\mathbf{K}$  in the  $i$ -th batch, and the mask ratio  $r$ . It outputs the number of masked tokens  $N_{mask}$  and their corresponding index vector  $\mathbf{I} \in \mathbb{R}^{1 \times |\mathbf{K}|}$ .

Then top- $k$  tokens with the lowest weights are selected as our candidate mask set. *i.e.*,  $\text{set}_B$  and  $\text{set}_A$ . This strategy is represented as follows:

$$(\mathbf{m} \odot \mathbf{z}) = \begin{cases} \mathbf{z}_\alpha, & m_i = 1. \\ \mathbf{z}_i, & m_i = 0. \end{cases} \quad (7)$$

Here,  $(\mathbf{m} \odot \mathbf{z})$  represents the final masked token,  $\mathbf{z}_i$  is the vector at the  $i$ -th token position ( $i \leq k$ ).  $\mathbf{z}_\alpha$  is a learnable mask embedding and  $m_i$  is the  $i$ -th element of the  $\mathbf{m} \in \mathbf{M}_{mask}$ . If  $\text{Attn}_k[i]$  is among the lowest  $r$ ,  $m_i$  is set to 1; otherwise, it is set to 0.  $\text{set}_A$  guides the masking for check-in sequence  $\mathcal{T}_A$ , and  $\text{set}_B$  for check-in sequence  $\mathcal{T}_B$ . The masked embeddings for platforms  $A$  and  $B$ , *i.e.*,  $\mathbf{Z}_A^m$  and  $\mathbf{Z}_B^m$  are obtained through the *Hadamard* product and matrix addition. The final representation  $\mathbf{H}_A^m$  and  $\mathbf{H}_B^m$  is derived by encoding these masked embeddings with  $f_M(\cdot)$ .

### User Identity Linkage Layer

Our user linkage layer draws inspiration from (Chen et al. 2024), using a multi-layer feed-forward neural network as the predictor. We employ a *sigmoid* function as the logistic regression function to generate the final similarity score. The sequence representations  $\mathbf{H}_A^m$  and  $\mathbf{H}_B^m$  are concatenated to form the final feature representation, which is then fed into the predictor to obtain the similarity score. This process can be viewed as a binary classification task, where Eq. (1) determines whether the two check-in sequences belong to the same user.

We optimize the final normalized probability using a binary cross-entropy loss function. The binary cross-entropy

---

**Algorithm 1: The Learning Process of MT-Link**


---

**Input** User spatial-temporal check-in sequences  $\mathcal{T}_A, \mathcal{T}_B$  and timestamp  $t_A$  and  $t_B$   
**Output** Prediction probability  $p$

- 1: Get  $\mathbf{X}_{A/B} = f_{st}(\mathcal{T}_{A/B})$ ;
- 2: Get  $\mathbf{Z}_{A/B} = \mathbf{X}_{A/B} + \text{TE}(t_{A/B})$ ;
- 3: Get  $\mathbf{H}_{A/B} = f_T(\mathbf{Z}_{A/B})$ ;
- 4: **for**  $l : |L|$  **do**
- 5:   Get  $\mathbf{Q}, \mathbf{K}, \mathbf{V}$  from linear( $\mathbf{H}_{A/B}$ );
- 6:    $\text{Attn}_{A \rightarrow B/A \rightarrow B} = \text{CoAttn}(\mathbf{Q}_{A/B}, \mathbf{K}_{B/A}, \mathbf{V}_{B/A})$ ;
- 7: **end for**
- 8: Initialize  $\mathbf{M}_{mask}$  as zero matrix  $\mathbf{0}$ ;
- 9: Get  $N_{mask}^{B/A}$  from  $\text{Attn}_{A \rightarrow B/B \rightarrow A}$ ;
- 10: **for**  $k : N_{mask}^{B/A}$  **do**
- 11:    $\mathbf{M}_{mask}[k] = \text{topk}(\text{sort}(\text{Attn}_{A \rightarrow B/B \rightarrow A}^{[k]}))$ ;
- 12: **end for**
- 13:  $\mathbf{Z}_{B/A}^m = (1 - \mathbf{M}_{mask}) \odot \mathbf{Z}_{B/A} + \mathbf{M}_{mask} \odot \mathbf{z}_\alpha$ ;
- 14:  $\mathbf{H}_{B/A}^m = f_M(\mathbf{Z}_{B/A}^m)$ ;
- 15:  $y = \sigma([\mathbf{H}_A^m; \mathbf{H}_B^m])$ ;
- 16: Calculate  $\mathcal{L}$  with Eq. (8) through  $\hat{y}$  and  $y$ ;
- 17: Back propagation and update parameters in MT-Link;
- 18: **return**  $p$

---

loss for a single sample is given by:

$$\mathcal{L}(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^N \mathbf{w}_i \cdot \ell(y_i, \hat{y}_i), \quad (8)$$

where  $\ell(y_i, \hat{y}_i) = y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)$  and  $y$  represents whether  $u_i^A$  and  $u_j^B$  are the same users, represented by 0 and 1;  $\hat{y}$  represents the classification result of the model.  $N$  represents the total number of samples and  $\mathbf{w}_i$  represents the corresponding weighted item (selected according to the label). The entire process can be referenced in Algorithm 1.

## Experiments

In this section, we evaluate our proposed model using four real-world cross-platform check-in datasets.

### Datasets

We collect two publicly available cross-platform check-in datasets: (1) XSiteTraj (Fu and Li 2023) and (2) ISP-Weibo (Feng et al. 2019). The XSiteTraj dataset includes data from three platforms: Twitter, Foursquare, and Facebook, which allowed us to create three cross-platform datasets: Twitter-Foursquare, Twitter-Facebook, and Foursquare-Facebook. Due to differences in the datasets and significant variations in trajectory lengths, we apply different preprocessing steps to each dataset. To facilitate training, we remove sequences longer than 400, 200, and 200 from the Twitter, Foursquare, and Facebook datasets, respectively, as these represent a small proportion of the data (5.20%, 0.28%, 0.17%). We also remove sequences shorter than 3 from the ISP and Weibo datasets, as they account for a small percentage of the data (3.53%, 0.13%). Table 1 presents the statistical details of the preprocessed datasets.

Datasets	Platforms	#Users	#Records	#Trajs
XSiteTraj	Twitter	11,239	1,187,063	11,239
	Foursquare	8,569	232,932	8,569
	Facebook	7,146	309,664	7,146
ISP-Weibo	ISP	30,405	3,423,865	202,656
	Weibo	20,624	175,895	20,624

Table 1: Statistics of the datasets.

### Baselines

- **DPLink** (Feng et al. 2019): A deep learning model based on RNN that introduces an end-to-end deep learning framework to extract spatio-temporal locality features for user linkage.
- **DPLink-SM** (Feng et al. 2020): An extended version of DPLink, which incorporates a trajectory similarity matcher to aid in the user identity linkage task, improving on DPLink’s capabilities.
- **CPLink** (Ding et al. 2020): A data mining approach that addresses the spatio-temporal UIL task using frequent pattern mining and clustering techniques.
- **CPLink+** (Ma et al. 2022): An enhanced version of CPLink that replaces the binary value function with a real-valued function to calculate the overlap area and uses the original DTW (Dynamic Time Warping) algorithm to achieve optimal distance measurement.

### Evaluation Metrics and Experiment Settings

In our evaluation, we use Macro-Precision, Macro-Recall, Macro-F1, and Area Under the Curve (AUC) to quantify the performance of different methods.

For the baseline models, we use the parameter settings recommended in their respective papers and fine-tune them for optimal performance. The check-in embedding dimension for our model is set to 64, the mask ratio  $r$  to 0.1, and the initial learning rate to 0.001. The dropout rate is set to 0.1, and the correlation attention block consists of 2 layers. We apply early stopping during validation with a patience of 5 to prevent overfitting. Each experiment is repeated five times, and the average results for all methods are reported. The code is available at <https://github.com/DrivenA/MT-Link>. All experiments are conducted on a machine with NVIDIA GeForce RTX 3090 GPU.

### Experimental Results

Table 2 summarizes our experimental results, with the best outcomes highlighted in **bold** and the second-best underlined. Our model aims to address two core challenges in the current spatio-temporal UIL task: 1) *retaining co-occurrence points while filtering out noise*, and 2) *capturing the spatio-temporal co-occurrence between user check-in sequences across different platforms*. Our experimental results validate the effectiveness of MT-Link in tackling these challenges, significantly outperforming existing comparison methods.

	Twitter-Foursquare				Twitter-Facebook				Foursquare-Facebook				ISP-Weibo			
	P(%)	R(%)	F1(%)	AUC(%)	P(%)	R(%)	F1(%)	AUC(%)	P(%)	R(%)	F1(%)	AUC(%)	P(%)	R(%)	F1(%)	AUC(%)
DPLink	51.40	52.38	51.08	51.91	64.52	54.97	55.39	52.29	53.34	51.83	51.88	53.78	24.99	50.00	33.32	50.57
DPLink-SM	55.42	53.02	52.17	52.27	<u>66.69</u>	55.40	55.99	53.21	<u>54.25</u>	51.49	51.02	55.32	50.70	53.09	40.64	52.88
CPLink	53.32	58.57	55.80	74.97	60.34	61.19	60.74	77.21	46.41	<u>62.90</u>	53.40	<u>79.53</u>	<u>63.07</u>	<u>70.72</u>	<u>66.67</u>	<u>80.17</u>
CPLink+	<u>58.76</u>	<u>65.67</u>	<u>62.01</u>	<u>78.96</u>	63.06	<u>68.04</u>	<u>65.44</u>	<u>80.67</u>	47.95	62.28	<u>54.10</u>	79.33	63.01	69.34	66.02	79.57
<b>Ours</b>	<b>66.14*</b>	<b>73.21*</b>	<b>68.06*</b>	<b>82.43*</b>	<b>71.09*</b>	<b>80.46*</b>	<b>73.82*</b>	<b>87.86*</b>	<b>57.34*</b>	<b>72.94*</b>	<b>62.87*</b>	<b>82.80*</b>	<b>80.37*</b>	<b>79.46*</b>	<b>79.85*</b>	<b>86.89*</b>
<i>Impro.</i>	<b>12.55</b>	<b>11.48</b>	<b>9.75</b>	<b>4.39</b>	<b>6.59</b>	<b>18.25</b>	<b>12.80</b>	<b>8.91</b>	<b>5.69</b>	<b>15.96</b>	<b>16.21</b>	<b>4.11</b>	<b>27.42</b>	<b>12.35</b>	<b>19.76</b>	<b>8.38</b>

Table 2: Performance comparison of all models on four real-world datasets. P represents Macro-Precision, R represents Macro-Recall, and F1 represents Macro-F1. Marker \* indicates the results are statistically significant (t-test with p-value < 0.01).

First, in the Twitter-Foursquare, Foursquare-Facebook, and Twitter-Facebook datasets, our model excelled in precision and recall, with average improvements of 11.48%, 15.96% and 18.25%, respectively. This demonstrates that MT-Link successfully captures the spatio-temporal co-occurrence information between cross-platform users while effectively preserving key spatio-temporal points. By ignoring noise points that could negatively impact the model, MT-Link focuses on the most relevant features. The spatio-temporal co-linearity information and its relational features effectively guide the masking process, enabling MT-Link to learn robust sequence representations. DPLink and DPLink-SM fail to account for the spatio-temporal co-occurrence of key points between cross-platform users, leading to lower classification performance. On the other hand, CPLink and CPLink+ are limited by larger datasets, which restrict their ability to effectively capture co-occurring key points. This makes it difficult for these models to accurately identify critical spatio-temporal patterns, further impacting their performance in complex cross-platform scenarios.

Second, in ISP-Weibo datasets, our model shows a significant advantage in Precision, with a significant improvement of 27.42%. Our model has demonstrated its effectiveness by maintaining high performance even on large datasets. This indicates that the model successfully captures spatio-temporal co-occurrence between cross-platform users. In contrast, DPLink and DPLink-SM struggle with large datasets and have difficulty maintaining classification accuracy in noisy environments. The lower accuracy of CPLink and CPLink+ on the current datasets suggests that data mining methods are limited by manually constructed features and frequent pattern mining, preventing the models from autonomously learning abstract semantic representations. Furthermore, CPLink and CPLink+ do not leverage real labels to distinguish users, making them more susceptible to false samples and leading to incorrect user identity recognition.

**Ablation Study.** To verify the effectiveness of key components, we conduct ablation experiments (as shown in Figure 3) by removing components of MT-Link. The obtained model variants are as follows:

- **w/o MTE:** We remove the masked transformer encoder from the model and use the output of the correlation at-

Method	Twitter-Foursquare		Twitter-Facebook	
	F1	AUC	F1	AUC
w/o MTE	62.05	64.33	71.63	83.33
w/o CAB	56.17	67.04	67.10	82.20
w/o TTE	54.20	58.38	68.52	84.98
MT-Link	68.06	82.43	73.82	87.86

Table 3: The results of ablation study.

tention block directly for the classification task.

- **w/o CAB:** We remove the correlation attention block from the model and replace it with a random masking strategy to guide the masking process.
- **w/o TTE:** We remove the temporal transformer encoder and replace it with a standard transformer encoder.

As seen in Table 3, MT-Link contributes effectively to its overall performance. *w/o MTE* compares with our MT-Link, it is evident that masked transformer effectively preserves key spatio-temporal co-occurrence points while filtering out noise points. Compared to the *w/o MTE*, MT-Link significantly increases AUC by 23.91% on Twitter-Foursquare. This approach strengthens the model’s ability to effectively filters out noise, leading to improved performance in user identity linkage tasks. *w/o CAB* suggests that the correlation attention block plays a critical role in guiding the retention of key points (those with high attention weights) and in capturing spatio-temporal co-occurrence between user check-in sequences on different platforms. *w/o TTE* shows that temporal position encoding significantly impacts datasets with long sequences, particularly in the Twitter-Foursquare dataset, where AUC significantly decreased by 41.19% when *w/o TTE*. This indicates that the temporal position encoder is crucial for generating spatio-temporal embeddings of check-in sequences.

**Parameter Sensitivity.** We also evaluate the sensitivity of MT-Link to different mask ratios and correlation attention block layers. Figure 3 shows the performance of masking rate and correlation attention block layers on Twitter-Facebook and Twitter-Foursquare.

We find that the performance on both datasets initially increases with the rise in the mask rate, reaches a critical value,

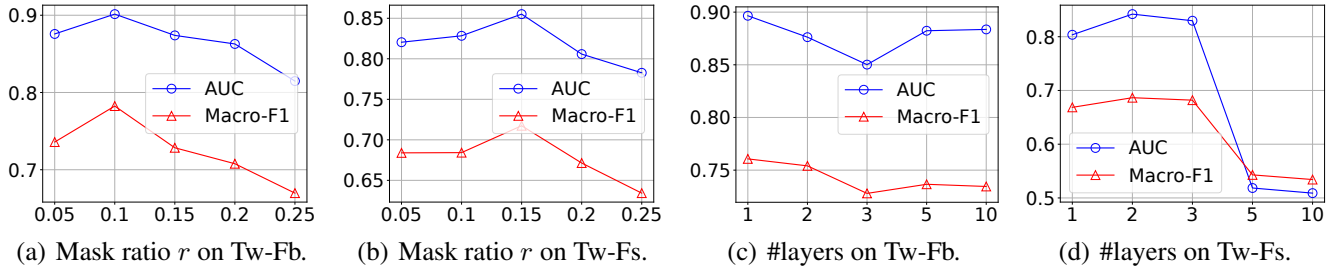


Figure 3: The impact of mask ratio and correlation attention block layers on MT-Link. Tw-Fb represents the Twitter-Facebook dataset, and Tw-Fs represents the Twitter-Foursquare dataset.

and then begins to decline. The optimal mask rates are 0.1 and 0.15, which ensure that most key information is retained while effectively removing noise, meeting the model’s requirements.

In addition, the performance on Twitter-Facebook remains relatively stable across different numbers of correlation attention block layers, with optimal performance at one layer. However, on Twitter-Foursquare, performance drops sharply when the number of layers is increased, suggesting that too many layers can disrupt key features, preventing accurate reflection of cross-platform user check-in co-occurrence.

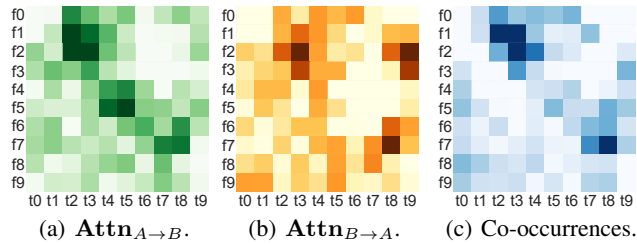


Figure 4: Co-occurrence visualization.

**Co-occurrence Visualization.** We output the attention weight maps learned from the correlation attention block for two trajectories, each with ten check-ins from Twitter and Foursquare, denoted as  $\text{Attn}_{A \rightarrow B}$  and  $\text{Attn}_{B \rightarrow A}$ , as shown in Figure 4(a) and Figure 4(b) (*i.e.*, with the  $x$ -axis representing Twitter check-ins and the  $y$ -axis representing Foursquare check-ins). We also calculate the pairwise distances between check-in locations of two trajectories and normalize them to obtain a co-occurrence matrix, as shown in Figure 4(c). Higher values indicate stronger spatio-temporal co-occurrences between the corresponding check-in points (darker colors), while lower values indicate weaker co-occurrences (lighter colors).

By examining Figure 4(a) and Figure 4(b) alongside the ground truth Figure 4(c), we can conclude that MT-Link correctly learn spatio-temporal co-occurrence patterns at these check-in points, for example,  $\text{Grid}(t8, f7)$  and  $\text{Region}(t2 f1, t3 f2)$  exhibit similar strong spatio-temporal co-occurrence as in Figure 4(c), and  $\text{Grid}(t9, f8)$  and  $\text{Region}(t7 f2, t8 f4)$  also show similar weaker co-occurrence as in Figure 4(c). However,  $\text{Grid}(t5, f5)$  in Figure 4(a) and  $\text{Grid}(t9, f2)$  in Figure 4(b) show stronger co-occurrence

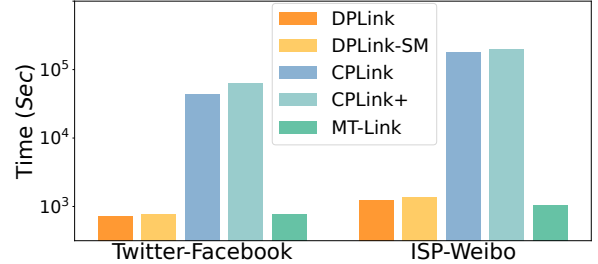


Figure 5: The time costs of MT-Link and other baselines.

than their counterparts in Figure 4(c), suggesting that MT-Link identifies important co-occurring check-in points when modeling correlations across different trajectories. In conclusion, the two learned attention matrices exhibit regional weights similar to those in the co-occurrence matrix, demonstrating that our model effectively captures accurate spatio-temporal co-occurrence information across user check-in sequences.

**Computation Time.** To validate the effectiveness of MT-Link in addressing the third challenge, we statistically analyze the runtime of all baselines. As shown in Figure 5, CPLink’s reliance on traditional data mining and clustering tasks results in high time complexity. Specifically, when applied to ISP-Weibo dataset, the running time for traditional data mining methods dramatically increases, about 187 times that of MT-Link, highlighting their limitations on large-scale datasets. Compared with DPLink and its successor DPLink-SM, MT-Link consumes similar running time. This may be because although our MT-Link does not require pre-training, it uses masking strategy and correlation attention block, which requires additional time to mine key information. Despite this, MT-Link has significantly better performance than DPLink and DPLink-SM.

## Conclusion

In this paper, we propose a transformer-based framework, MT-Link, to enhance model performance by learning spatio-temporal co-occurrence patterns of cross-platform users. Our model captures the spatio-temporal co-occurrence between users and utilizes a masked transformer to filter out noise points. In extensive experiments on four real-world cross-platform datasets, our model significantly outperforms state-of-the-art baselines.

## Acknowledgments

This work is partially supported by the National Natural Science Foundation of China under Grant No. 62176243, the Fundamental Research Funds for the Central Universities under Grant No 202442005, and the National Key R&D Program of China under Grant No 2022ZD0117201.

## References

- Basik, F.; Ferhatosmanoğlu, H.; and Gedik, B. 2020. Slim: Scalable linkage of mobility data. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 1181–1196.
- Basik, F.; Gedik, B.; Etemoğlu, Ç.; and Ferhatosmanoğlu, H. 2017. Spatio-temporal linkage over location-enhanced services. *IEEE Transactions on Mobile Computing*, 17(2): 447–460.
- Berisha, V.; Krantsevich, C.; Hahn, P. R.; Hahn, S.; Dasarathy, G.; Turaga, P.; and Liss, J. 2021. Digital medicine and the curse of dimensionality. *npj Digital Medicine*, 4.
- Chen, H.; Yin, H.; Sun, X.; Chen, T.; Gabrys, B.; and Musial, K. 2020. Multi-level graph convolutional networks for cross-platform anchor link prediction. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 1503–1511.
- Chen, W.; Huang, C.; Yu, Y.; Jiang, Y.; and Dong, J. 2024. Trajectory-User Linking via Hierarchical Spatio-Temporal Attention Networks. *ACM Transactions on Knowledge Discovery from Data*, 18(4): 1–22.
- Chen, W.; Li, S.; Huang, C.; Yu, Y.; Jiang, Y.; and Dong, J. 2022. Mutual distillation learning network for trajectory-user linking. *arXiv preprint arXiv:2205.03773*.
- Chen, W.; Wang, W.; Yin, H.; Zhao, L.; and Zhou, X. 2023. HFUL: a hybrid framework for user account linkage across location-aware social networks. *The VLDB Journal*, 32(1): 1–22.
- Chen, W.; Yin, H.; Wang, W.; Zhao, L.; Hua, W.; and Zhou, X. 2017. Exploiting spatio-temporal user behaviors for user linkage. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 517–526.
- Chen, W.; Yin, H.; Wang, W.; Zhao, L.; and Zhou, X. 2018. Effective and efficient user account linkage across location based social networks. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, 1085–1096. IEEE.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Ding, F.; Ma, X.; Yang, Y.; and Wang, C. 2020. User identity linkage across location-based social networks with spatio-temporal check-in patterns. In *2020 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCLOUD/SocialCom/SustainCom)*, 1278–1285. IEEE.
- Feng, J.; Li, Y.; Yang, Z.; Zhang, M.; Wang, H.; Cao, H.; and Jin, D. 2020. User identity linkage via co-attentive neural network from heterogeneous mobility data. *IEEE Transactions on Knowledge and Data Engineering*, 34(2): 954–968.
- Feng, J.; Zhang, M.; Wang, H.; Yang, Z.; Zhang, C.; Li, Y.; and Jin, D. 2019. Dplink: User identity linkage via deep neural network from heterogeneous mobility data. In *The world wide web conference*, 459–469.
- Fu, J.; and Li, Y. 2023. XSiteTraj: A cross-site user trajectory dataset. *Data in Brief*, 51: 109783.
- Goga, O.; Lei, H.; Parthasarathi, S. H. K.; Friedland, G.; Sommer, R.; and Teixeira, R. 2013. Exploiting innocuous activity for correlating users across sites. In *Proceedings of the 22nd international conference on World Wide Web*, 447–458.
- Huang, H.; Ding, F.; Yin, H.; Liu, G.; Wang, C.; and Wu, D. O. 2024. EgoMUIL: Enhancing Spatio-Temporal User Identity Linkage in Location-Based Social Networks With Ego-Mo Hypergraph. *IEEE Transactions on Mobile Computing*, 23(8): 8341–8354.
- Li, B.; Zhu, H.; and Xie, M. 2019. Lisc: location inference attack enhanced by spatial-temporal-social correlations. In *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCOM/IOP/SCI)*, 1083–1092. IEEE.
- Li, S.; Chen, W.; Yan, B.; Li, Z.; Zhu, S.; and Yu, Y. 2023. Self-supervised contrastive representation learning for large-scale trajectories. *Future Generation Computer Systems*, 148: 357–366.
- Li, Z.; Chen, Z.; Yang, F.; Li, W.; Zhu, Y.; Zhao, C.; Deng, R.; Wu, L.; Zhao, R.; Tang, M.; et al. 2021. Mst: Masked self-supervised transformer for visual representation. *Advances in Neural Information Processing Systems*, 34: 13165–13176.
- Long, M.; Chen, S.; Du, X.; and Wang, J. 2023. DegUIL: Degree-Aware Graph Neural Networks for Long-Tailed User Identity Linkage. In De Francisci Morales, G.; Perlich, C.; Ruchansky, N.; Kourtellis, N.; Baralis, E.; and Bonchi, F., eds., *Machine Learning and Knowledge Discovery in Databases: Applied Data Science and Demo Track*, 122–138. Cham: Springer Nature Switzerland.
- Ma, X.; Ding, F.; Peng, K.; Yang, Y.; and Wang, C. 2022. CP-Link: Exploiting continuous spatio-temporal check-in patterns for user identity linkage. *IEEE Transactions on Mobile Computing*, 22(8): 4594–4606.
- Naini, F. M.; Unnikrishnan, J.; Thiran, P.; and Vetterli, M. 2015. Where you are is who you are: User identification by matching statistics. *IEEE Transactions on Information Forensics and Security*, 11(2): 358–372.
- Qi, L.; Zhang, X.; Dou, W.; Hu, C.; Yang, C.; and Chen, J. 2018. A two-stage locality-sensitive hashing based approach for privacy-preserving mobile service recommendation in cross-platform edge environment. *Future Generation Computer Systems*, 88: 636–643.
- Qin, G.; Song, L.; Yu, Y.; Huang, C.; Jia, W.; Cao, Y.; and Dong, J. 2023. Graph structure learning on user mobility

data for social relationship inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 4578–4586.

Riederer, C.; Kim, Y.; Chaintreau, A.; Korula, N.; and Lattanzi, S. 2016. Linking users across domains with location data: Theory and validation. In *Proceedings of the 25th international conference on world wide web*, 707–719.

Rodríguez, P.; Bautista, M. A.; Gonzalez, J.; and Escalera, S. 2018. Beyond one-hot encoding: Lower dimensional target embedding. *Image and Vision Computing*, 75: 21–31.

Rossi, L.; and Musolesi, M. 2014. It’s the way you check-in: Identifying users in location-based social networks. In *Proceedings of the second ACM conference on Online social networks*, 215–226.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, H.; Li, Y.; Wang, G.; and Jin, D. 2018. You are how you move: Linking multiple user identities from massive mobility traces. In *Proceedings of the 2018 SIAM International Conference on Data Mining*, 189–197. SIAM.

Wang, Y.; Wu, H.; Dong, J.; Liu, Y.; Qiu, Y.; Zhang, H.; Wang, J.; and Long, M. 2024. Timexer: Empowering transformers for time series forecasting with exogenous variables. *arXiv preprint arXiv:2402.19072*.

Xue, H.; Sun, B.; Si, C.; Zhang, W.; and Fang, J. 2021. KMUL: a user identity linkage method across social networks based on spatiotemporal data. In *2021 IEEE 15th International Conference on Big Data Science and Engineering (BigDataSE)*, 111–117. IEEE.

Zhao, P.; Luo, A.; Liu, Y.; Xu, J.; Li, Z.; Zhuang, F.; Sheng, V. S.; and Zhou, X. 2020. Where to go next: A spatiotemporal gated network for next poi recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 34(5): 2512–2524.