MLP-POI: Effective MLP-based Reliable Distillation for POI Recommendation

Anonymous Author(s)*

ABSTRACT

Recent years have witnessed the great success of incorporating spatial-temporal factors into Recurrent Neural Networks (RNNs) or self-attention mechanisms (SAM) in handling Point of Interest (POI) recommendation tasks. However, their computational efficiency and scalability fall short of the practical industry's applications. Although we observe that the simple MLP architecture achieved excellent performance, it failed to capture the spatial-temporal behavior patterns of users, limiting understanding of user preferences. Additionally, when planning to transfer information from the spatial-temporal teacher model to the student MLPs through knowledge distillation, we empirically find that the under-confidence in the prediction distribution by the teacher model resulted in insufficient reliable supervision for the student model. Against this background, we propose an Effective Multi-Layer Perceptrons (MLP) Reliable Distillation Framework (MLP-POI) designed to acquire completed trajectory awareness and a lightweight spatial-temporal-aware MLP through effective knowledge distillation from intricate spatial-temporal models. We initially construct multiple diverse augmented trajectories as well as their confidence levels by leveraging a global transition pattern graph to identify potential missing check-ins. With these as input, a confidence-inspired multi-teacher knowledge distillation approach can adaptively allocate sample-wise reliability for each teacher prediction. Besides, to distill spatial-temporal knowledge, we devised intra-level similarity relationship distribution in the student model and contextual spatial distribution relationship matching in the teacher model. Our method is straightforward yet pragmatic, and comprehensive experiments demonstrate its robustness to sequence length, and high-confidence distributions, consistently achieving state-of-the-art performance.

CCS CONCEPTS

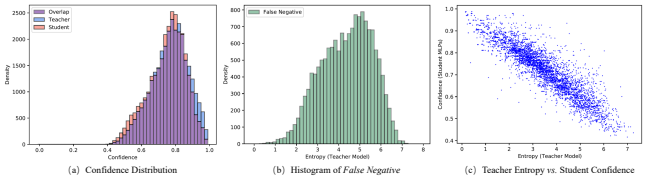
• **Information systems** → **Location based services**; **Data mining**; • **Human-centered computing** → **Ubiquitous and mobile computing design and evaluation methods**.

KEYWORDS

Point-of-Interest recommendation, confidence-aware weighting, multi-teacher distillation, multi-layer perceptrons

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX



(a) Confidence Distribution (b) Histogram of *False Negative* (c) Teacher Entropy *vs.* Student Confidence

Figure 1: (a) Histograms of confidence distributions for correct predictions made by both the spatial-temporal teacher and (distilled) student MLP model. (b) Distribution of "False Negative" sample w.r.t the entropy of predictions by the teacher model, which is predicted correctly by the teacher model but incorrectly by the (distilled) student model. (c) Scatter plot of confidence (student model) and entropy (teacher model) for "True Positive" sample, indicating sample that predicted correctly by both the teacher model and (distilled) student model. Additionally, they are all sampled from the Gowalla dataset.

ACM Reference Format:
Anonymous Author(s). 2018. MLP-POI: Effective MLP-based Reliable Distillation for POI Recommendation. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Point-of-Interest (POI) recommendation system capitalizes on the rapid evolution of mobile technology, emerging as a catalyst for the burgeoning interest in location-based social network (LBSN) services. Facilitated by the substantial profits that POI recommendations yield in real-world business scenarios and their ability to satisfy users' needs, they have garnered significant attention from both the academic and industrial fields. As an important application in smart cities, the POI recommendation system relies on voluntarily updated geographical trajectories (e.g., check-in records) through mobile devices to discern their preferences. It predicts the next location that users are most likely to visit by analyzing their history trajectory and current location, not only assisting users in planning trips to enticing destinations for future movements but also empowering businesses to effectively forecast customer foot traffic.

Traditional next POI recommendation tasks benefit from the sequence dependencies and geographical relationships along user trajectories. It is primarily regarded as a sequential task enhanced by geographical information hidden within the trajectories. In sequence dependencies learning [2, 6, 34], inheriting from natural language processing (NLP), sequence models like Recurrent Neural Networks (RNNs) and self-attention mechanisms (SAM) have demonstrated significant success in POI recommendation tasks by effectively capturing user mobility patterns. Existing efforts primarily focus on employing one of these methods and incorporating

additional components to integrate the inherent spatial-temporal dependencies within trajectory sequences from raw data. STRNN [15] incorporates parameterization of spatial-temporal interval transition matrices fed into RNNs. HST-LSTM [11] and STGN [35] modify the gate mechanism of Long Short-Term Memory (LSTM) by incorporating spatial-temporal distance as an additional input. The recent state-of-the-art works [17, 18, 21, 27] also incorporate relative spatial-temporal interval information between consecutive and non-consecutive POIs into LSTM and attention mechanisms, respectively. Besides, they [14, 16] employ a hierarchical grid with self-attention mechanism to encode geographical information.

Although powerful, these methods are conceptually and technically complicated with their advanced model architectures, rendering their deployment in real-world scenarios intricate. At this juncture, a thought-provoking question arises: Are RNN and SAM always indispensable for POI recommendation prediction? To answer this question, we propose a simple architecture for our student model, entirely based on multi-layer perceptrons (MLPs). However, the straightforward MLP architecture fails to capture the spatial-temporal behavior patterns of users. Therefore, we intend to transfer the knowledge from a pre-trained teacher model that incorporates spatial-temporal information to our student model. Unfortunately, we observed marginal improvements when indiscriminately transferring the teacher model's prediction distribution to the student model using a simple KL-divergence. From three motivational experiments, we discovered that the unsatisfactory performance of the directly distilled MLP model may be attributed to the minimal contribution from the uninformative and unreliable teacher model in the distillation process. In Fig. 1(a), we observe a significant distribution shift between the teacher model and the student model, revealing a problem where the distilled MLP may fail to make predictions as confidently as the spatial-temporal teacher model. In Fig. 1(b) and Fig. 1(c), we realized that a teacher model with high uncertainty (i.e., low reliability) may hinder the student model from making sufficiently confident predictions. Based on the above observations, it is reasonable to hypothesize that the student model's inability to further improve its performance is attributed to the insufficiently reliable supervision provided by the teacher model.

To this end, we propose MLP-POI, a knowledge distillation architecture based on MLPs for POI recommendation, which is designed to enable the lightweight MLP model to acquire spatial-temporal information. Different from other deep learning approaches that focus on designing technically complicated neural architectures to capture user behavior patterns, we choose the lightweight MLP model to summarize trajectory sequences, simplify the neural architecture, and surprisingly achieve outstanding performance. To enable the model to enjoy low computational costs while capturing users' spatial-temporal behavior patterns, we further introduce a knowledge distillation paradigm, using MLP as a lightweight student model and an intricately complex spatial-temporal model as the teacher model. Then, to tackle the challenge of under-confidence in the teacher model caused by supervising each prediction distribution equally in simple knowledge distillation, we design a multi-teacher knowledge distillation scheme. Specifically, it incorporates potential missing check-ins identified from the global transition pattern graph to complete the raw trajectory sequences through weighted sampling, resulting in multiple different augmented trajectories. Therefore, we obtain diverse prediction distributions from multi-teachers, combine them with ground truth labels, and assign weights based on the confidence of each teacher's predictions, we achieve a more reliable knowledge distribution for transfer to the student model. Additionally, we designed a novel scheme to capture the intra-level spatial relationship distribution between check-ins in the historical trajectory sequence and transfer it to the student model's contextual similarity relationship distribution. We summarize the primary contributions of our work as follows:

- We propose a novel MLP-based knowledge distillation framework for POI recommendation, capable of extracting reliable knowledge distributions from the spatial-temporal teacher model and transferring them to the student model. To our best knowledge, this is the first method that transfers spatial-temporal model knowledge to an MLP-based model with a simpler architecture.
- We propose to transfer the intra-level spatial distribution relationships between current location and historical check-ins to the student model, further enhancing the student model's ability to perceive spatial relationships.
- We conducted extensive experiments on two real datasets to demonstrate the effectiveness of our approach. The results also demonstrated that our approach outperforms existing methods across various metrics.

2 RELATED WORK

In this section, we will briefly review the related works on multi-teacher knowledge distillation and the next POI recommendation.

2.1 Multi-teacher Knowledge Distillation

Knowledge Distillation (KD) [9] aims to transfer the generalization capability from a cumbersome model (teacher) to a simplified one (student) by leveraging the class probabilities generated by the teacher model as soft targets. In contrast to the typical use of a single-teacher model in KD [26], multi-teacher KD further explores how to better integrate predictions from multiple-teachers. The proposed multi-teacher KD paradigms can be divided into two categories: 1) In the earlier literature [24, 30], a straightforward approach involves assigning average or other fixed weights to the probabilities prediction from each teacher. They enable the student model to benefit from comprehensive guidance provided by multiple teachers. Among them, RLKD [31] introduces reinforcement learning before averaging the prediction logits of each teacher to filter out inappropriate teachers. 2) The adaptive assignment of distinct weights to the outputs of individual teacher models has been recognized as a crucial strategy. EBKD [3] and CA-MKD [32] measure weights using information entropy and cross-entropy with true labels, respectively. MMKD [33] employs a meta-learning network to aggregate logits, guiding the student model at the instance level. AEKD [3] examines ensemble knowledge distillation as a multi-objective optimization problem in the gradient space. KRD [23] dynamically samples reliable knowledge points as multiple teachers and distills their knowledge into student MLPs. However, none of the above methods are suitable for scenarios in POI recommendations where there might be missing check-ins. Moreover, they did not take into account that multiple augmented input data might yield benefits with varying confidence levels.

2.2 Next POI Recommendation

The next POI recommendation in a location-based social model is a key task, which is based on mining user dynamic interest preferences from their historical trajectories. In the early phase, a typical technique involves modeling through Markov-based models, primarily estimating the transition matrix of behavioral probabilities based on historical behavior. Matrix factorization models [12, 19] are proposed to capture the transition of intermittent visits, with subsequent extensions [1, 8] revealing that incorporating explicit spatial and temporal information substantially enhances recommendation performance, particularly in building the transition matrics. Recently, Recurrent Neural Networks (RNNs) and Attention mechanisms have been extended for utilization in POI recommendation tasks, capturing the contextual dependencies within input sequences. DeepMove [4] introduces a historical attention layer to capture long-term periodicity and enhance the short-term sequential patterns captured by gated recurrent units (GRU). Meanwhile, to cope with sparse and incomplete trajectory sequences, the most popular scheme is to seamlessly incorporate additional contextual factors into RNNs or attention mechanisms. STRNN [15] collaboratively incorporates transition matrices parameterized by both spatial-temporal intervals as additional factors modeling successive contextual relationships into the RNN. HST-LSTM [11] introduces a spatial-temporal distance input gating mechanism as an additional input to address the data sparsity issue. STGN [35] adds two pairs of time and distance gates controlled by spatial-temporal distance to LSTMs. Flashback [27] and STAN [17] respectively introduce relative spatial-temporal information between consecutive and non-consecutive POIs into LSTM and attention mechanisms. GeoSAN [14] employs hierarchical gridding with self-attention mechanisms to encode geographical information. Besides, recent approaches [18, 22] strive to leverage graph representation learning to endow POI representations with powerful expressive capabilities and combine them with sequential models. However, such a pattern entails a substantial computational burden, presenting significant challenges for practical deployment.

3 PRELIMINARIES

3.1 Notations and Problem Statement

This section presents the relevant term definitions and problem formulation. Here, the trajectory sequence records the user set, location (POI) set, and timestamp set which are represented as $U = \{u_1, u_2, ..., u_{|U|}\}$, $L = \{l_1, l_2, ..., l_{|L|}\}$, and $T = \{t_1, t_2, ..., t_{|T|}\}$ respectively, where $|U|$, $|L|$ and $|T|$ represent the sizes of their respective sets. And, each location l_i is mapped to a geographical coordinate, denoted as by $g_i = (latitude = \alpha_i, longitude = \beta_i)$.

Definition 1 (Check-in). *Each check-in record consists of a quadruple $r_i^u = (u, l_i, t_i, g_i)$. It represents a check-in behavior where a user u visits a location l_i with an exact geographical position $g_i = (latitude = \alpha_i, longitude = \beta_i)$ at time t_i.*

Definition 2 (User trajectory). *A user trajectory represents a record of the sequence of locations where a user has checked in chronologically. We can formulate the trajectory sequence as $S_u = \{r_1, r_2, \ldots, r_n\}$. Based on geographical coordinates associated with each check-in in the trajectory sequence, we can easily calculate the spatial intervals Δ_{ij}^s between any two locations in the historical trajectory.*

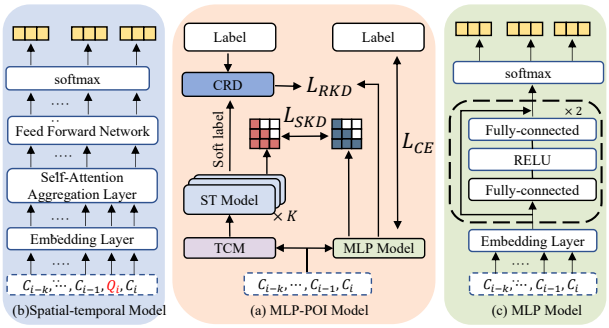


Figure 2: The overall architecture of MLP-POI. (a) consists of the Trajectory Completion Module (TCM), K spatial-temporal teacher models obtaining diverse augmented data inputs, and the MLP student model. (b) is the architecture of the spatial-temporal model. (c) is an MLP student model that includes two layers of MLP architecture.

Definition 3 (Global Transition Pattern Graph). *We denote the global directed transition pattern as graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, , \boldsymbol{w})$. (1) $\mathcal{V} = \{v_1, \cdots, v_N\}$ is a set composed of $N = |\mathcal{V}|$ nodes, where each node represents a POI. (2) $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is a set of edges representing whether nodes are connected. (3) $\boldsymbol{A}$ is the weighted adjacency matrix of global transition pattern $\mathcal{G}$. $a_{ij} = 1$ indicates the oresebce of an edge between node v_i and v_j; otherwise, $a = 0$. (4) $\boldsymbol{w}$ is the weight of the edge, w_{ij} represents the transition possibilities between node w_i and w_j. Noting that the global transition pattern graph we use is modeled by Graph-Flashback, which learns a weighted POI graph to reflect the spatial-temporal transition patterns between POIs better.*

Problem 1 (Next POI recommendation). *Given the user u_i, the trajectory $S_u = \{r_1, r_2, \ldots, r_n\}$ of user u_i and the spatial relation matrix Δ_{ij}^s. The goal of the next POI recommendation task is to output the top-k POIs that the user u_i is most likely to visit next and find the desired output $l_{n+1} \in L$ for the next check-in.*

3.2 Knowledge Distillation

Knowledge distillation was first introduced in [9] to primarily address image data, where knowledge is transferred from a pre-trained teacher model to a simpler student model. It takes the softmax distributions generated by the teacher model as soft labels and imposes a KL-divergence constraint $D_{KL}(\cdot, \cdot)$ on the student model. The optimization objective function is as follows:

$$L_{KD} = \sum_{i=1}^{|L|} KL(\sigma(\frac{\hat{h}_i^t}{\tau}), \sigma(\frac{\hat{h}_i^s}{\tau})), \tag{1}$$

where $\hat{h}_i^t$ is the soft labels generated by the teacher model, and $\hat{h}_i^s$ is the label distribution of the student model. σ represents the softmax function with temperature τ.

3.3 Spatial-Temporal Model

POI recommendation, unlike general recommendation tasks, demands the consideration of both temporal and spatial influences on modeling user preferences. Therefore, our goal is for the teacher model to uncover spatial-temporal patterns in the data, serving as a reference for the student model based on these spatial-temporal insights.

Self-attention Aggregation Layer. Inspired by the effectiveness of the self-attention layer in capturing long-term sequential dependencies, we employ it to capture users' temporal preferences. Given the user trajectory embedding vector representation $E(S_u)$, we first construct a mask matrix $M \in R^{n\times n}$ with ones in the upper-left elements and zeros elsewhere. We then project the trajectory embedding vectors to different representation spaces using distinct parameter matrices $W_Q, W_K, W_V \in R^{n\times n}$ to capture diverse and effective contextual meanings, as follows:

$$\begin{aligned} h &= Attention(E(S_u)W_Q, E(S_u)W_K, E(S_u)W_V) \\ Attention(Q,K,V) &= (M \times \sigma(\frac{QK^T}{\sqrt{d}}))V, \end{aligned} \tag{2}$$

By assigning different weights (i.e., $\sigma(\frac{QK^T}{\sqrt{d}})$ in Eq. 2) to historical sequences based on their contextual significance, we assess their impact on the current prediction. In terms of spatial factors, The closer the historical check-ins are to the current location, the more helpful they are in predicting the next one [27]. Therefore, we utilize spatial intervals $\Delta D_{i,j}$ between the current location l_i and past check-in location l_j (i.e., j<i) as a metric to research hidden states with strong predictive capabilities as follows:

$$\begin{aligned} \hat{h}_i &= \frac{\sum_{j=0}^{i} \omega_j * h_j}{\sum_{j=0}^{i} \omega_j} \\ \omega(\Delta d_{i,j}) &= e^{-\eta \Delta D_{i,j}}, \end{aligned} \tag{3}$$

where $\omega(\cdot)$ denotes the similarity function, η represents the spatial decay weight, controlling the rate at which the weights decrease according to $\Delta D_{i,j}$. Finally, we obtain the updated representation of the trajectory sequence through the spatial-temporal model.

Prediction Layer. The output $\hat{h}_t$ of the teacher model or student model at each time step i is concatenated with the user embedding e^u and fed into a fully connected layer to generate the final output as follows:

$$\hat{y}_i^u = W_f[\hat{h}_i; e^u], \tag{4}$$

where $W_f \in R^{|L|\times 2d}$ is a learnable weight matrix to project the output to the probability distribution over labels, $[\cdot;\cdot]$ denotes the concatenation operation. We choose the cross-entropy function to learn the model parameters as follows:

$$L_{CE}(y, \hat{y}) = -\sum_{i\in S_u} \sum_{j=1}^{n} y_{i,j} \log \sigma(\hat{y}_{i,j}), \tag{5}$$

where $y_{i,j}$ is the label at the $i-th$ check-in, $\hat{y}_{i,j}$ is the output logits of the prediction model.

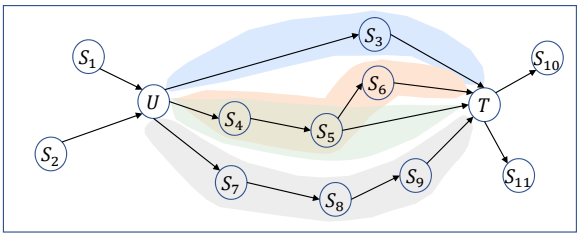


Figure 3: Local Search Strategy and Examples of Probability Simplex. We use backgrounds of varying colors to denote different valid potential missing node search paths. Nodes U and T represent the origin and destination, respectively.

4 METHODS

4.1 MLP Architecture

For efficient inference, MLP is considered the default student model in this paper. During the testing process, only the student model is used for predictions. MLP takes a trajectory sequence as input, where each one obtains a vector x_i of hidden dimension d from the POI look-up embedding table as their representation in the feature space. The POI embedding dictionary here is initialized by Graph-Flashback. We use a 2-layer MLP to summarize sequence relationship information. It is composed of two linear transformation layers, an activation function layer $\sigma = GeLU(\cdot)$, and a dropout function $Dropout(\cdot)$, as follows:

$$\begin{aligned} h_i^{(l)} &= h_i^{(l-1)} + W_2^{(l-1)} Dropout(\sigma(W_1^{(l-1)} h_i^{(l-1)})) \\ h_i^{(0)} &= x_i, \end{aligned} \tag{6}$$

where $W_1^{(l-1)} \in R^{2d\times d}$, $W_2^{(l-1)} \in R^{d\times 2d}$ are weight matrices with the hidden dimension d. In addition to the MLP layers, it also incorporates other standard architectural components: skip-connections [7].Therefore, the computational complexity increases linearly with the length of the sequence input, unlike the attention mechanism model whose complexity is quadratic. While the MLP-based architecture is simple yet efficient, we observed that it can achieve comparable performance to state-of-the-art methods [29] on two datasets for POI recommendation in our ablation experiments. Finally, it use a linear classifier $W_c \in R^{|L|\times 2d}$ as the standard classification head and generate the label distribution $\hat{y}_i = W_c h_i^{(l)}$.

4.2 Trajectory Completion Module (TCM)

As discussed in the Flashback, the raw input S_u is an incomplete trajectory sequence. This presents a formidable challenge for existing sequence-based models aiming to capture user sequence patterns. Therefore, the Trajectory Completion Module is introduced to harness the robust inductive capabilities of trajectory flow graphs, completing any potentially missing check-ins between any two consecutive check-ins in the raw trajectory sequence for all users in the training set. With their support, the spatial-temporal model, capturing richer contextual information, achieves superior performance, and employing them as inputs for the spatial-temporal teacher model results in diverse prediction distributions.

Potential Missing Nodes Search Strategy. We propose a problem of identifying potential missing nodes on the trajectory flow graph. Unlike other methods for identifying missing POIs [25] or context-aware data augmentation methods [13], we do not need additional training or extra modules specifically for recognizing missing POIs. We simply leverage the global transition pattern graph $\mathcal{G}$ to build a prompt memory set of potential missing check-ins before training. We consider any pair of POI nodes on the graph as source U and target nodes T as shown in Fig. (3), respectively. We use a breadth-first sampling (BFS) algorithm on the directed trajectory flow graph to sample a set of transition nodes from the source to the target node, recorded in the prompt memory set $N_s(U, T)$. Additionally, we compute the average weight of the sampled trajectory segments as the weight for each respective trajectory segment. Considering memory limitations, we set the search depth m as a hyperparameter. Additionally, we set a threshold p for trajectory sampling weights, only sampling potential sequences with weights higher than p and lengths less than or equal to m.

Potential Missing Nodes Completion Strategy. Once we obtain the prompt memory set N_s of potential missing transition nodes between any pair of nodes, we can reconstruct the original trajectory sequence. When given a raw trajectory sequence S_u, we assume the existence of missing trajectory segments between any two consecutive check-ins in the trajectory sequence, where two adjacent check-ins are regarded as the source and target nodes in chronological order. ~~we can consider any pair of consecutive check-ins in the trajectory sequence in temporal order as source and target nodes.~~ For enhanced sampling of potential missing trajectory segments, we utilize the source and target nodes as indices to sample from the prompt set based on their respective weights (i.e., above the weight threshold γ). Finally, we inject the sampled trajectory segments into the positions between the source and target node of the raw trajectory. ~~By consulting the prompt memory set, we obtain potential missing trajectory segments, and based on their weights, we randomly sample and insert them between the consecutive two check-ins in the raw trajectory sequence.~~ The search for potential trajectory nodes is repeated for any two consecutive check-ins in the raw trajectory sequence. Since we complete potential missing trajectory segments by weighted sampling between any two consecutive check-ins in the trajectory, each completion trajectory sequence $S_u^{'}$ generated will be different. Therefore, we input each completed trajectory into every teacher model, enabling each teacher model to capture different perspectives of sequence patterns. ~~Finally, diverse completed trajectory sequences $S_u^{'}$ are generated, providing diverse and augmented trajectory data for modeling various possible rich sequential patterns in teacher models.~~

4.3 Confidence-inspired Reliable Distillation (CRD)

Confidence-based Weight Approach. To effectively mitigate the potential noise introduced by random sampling during trajectory input data completion and simultaneously aggregate reliable prediction distribution from multiple teachers, we cooperatively consider true labels and assign different weights to each teacher model's prediction distribution for the next check-in prediction. Each completed trajectory, modeled by the teacher model, exhibits diverse contextual patterns from multiple views. This provides the student model with nuanced soft labels from various views in the knowledge distillation process. Therefore, in the process of knowledge distillation, the objective is to obtain dependable and multi-views soft labels from the teacher model, which can be effectively transferred to the student model.

The entropy of the predicted distribution by the teacher model reflects the level of confidence in its predictions for the next check-in. Consequently, elevated entropy in soft labels may suggest a higher degree of uncertainty, making them more prone to being incorrect or noisy, and vice versa. Therefore, the student model is more inclined to receive knowledge in the form of soft labels with lower entropy, which is considered more reliable for KD. We can formalize the entropy calculation for soft labels generated by each teacher model using the following:

$$\begin{aligned} H_i &= -\sum_{i=1}^{|L|} \sigma(\frac{\hat{y}_i}{\tau}) \log(\sigma(\frac{\hat{y}_i}{\tau})) \\ C_i &= 1 - \frac{H_i}{\log(|L|)}, \end{aligned} \tag{7}$$

where $\log(|L|)$ denotes the maximum entropy value achieved when all prediction probabilities are equally probable and $C_i \in [0, 1]$ is the confidence score. Subsequently, we integrate confidence-based weighting, leveraging confidence score, to enhance the reliability of aggregating soft labels from multiple teachers as follows:

$$w_{i,k} = \frac{\exp(C_{i,k})}{\sum_{j=1}^{K} C_{i,j}}, \tag{8}$$

where k denotes the $k-th$ teacher. The smaller the entropy H_i, the higher the correlation with the confidence-based weight w_i. So, the knowledge distillation loss function under the supervision of soft labels from multiple teachers is as follows:

$$L_{RKD} = \sum_{i=1}^{|L|} \sum_{k=1}^{K} w_{i,k} KL(\sigma(\frac{\hat{h}_i^t}{\tau}), \sigma(\frac{\hat{h}_i^s}{\tau})), \tag{9}$$

According to Eq. (9), more confident teachers will assign larger weights to guide the student toward making correct judgments.

Label-based Masking Approach. However, entropy-based strategies have notable limitations, as they tend to favor models with low variance, leading to a preference for blindly confident predictions. When the soft label distribution is sharp, large weights will be assigned regardless of whether the most probable distribution is correct or not. This may lead to low-quality predictions from the teacher model, which can misguide the training of the student model and further hurt its distillation performance. Fortunately, with the assistance of true labels, we adeptly navigate this challenge by selectively transferring soft labels from the teacher models when their predictions are accurate.

$$w_{i,k}^{KD} = \begin{cases} w_{i,k} & \text{q=1} \\ 0 & \text{q=0} \end{cases} \tag{10}$$

Dataset	Gowalla	Foursquare
#users	7,768	45,343
#locations	106,994	68,879
#check-ins	1,823,598	9,361,228
#Collection period	02/2009-10/2010	04/2012-01/2014
#Average time between	51.28 hours	58.59 hours
Successive check-ins	2.13 days	2.44 days

Table 1: Basic dataset statistics.

where q represents an indicator function. It equals 1 if the prediction by the $k-th$ teacher model is correct for the $i-th$ attendance prediction; otherwise, q equals 0. The final knowledge distillation loss function L_{RKD}, inspired by confidence weighting, is obtained as follows:

$$L_{RKD} = \sum_{i=1}^{|L|} \sum_{k=1}^{K} w_{i,k}^{KD} KL(\sigma(\frac{\hat{h}_i^t}{\tau}), \sigma(\frac{\hat{h}_i^s}{\tau})), \tag{11}$$

4.4 Spatial Distribution Distillation

In addition, besides transferring knowledge of inter-level prediction distributions within spatial models, we introduce the transfer of intra-level spatial similarity distributions. This involves leveraging the teacher model's computations to enhance contextual relevance for the student model, enabling additional spatial feature transfer. Once again relying on spatial intervals computed by the teacher model, it considers the current location along with the spatial interval distribution from historical check-ins as the inter-class spatial relationship. This inter-class relationship can be transferred to the student model, allowing the contextual relationships of the student model to incorporate spatial correlations. Following the suggestion of previous work, we take the cosine similarity as our metric function. The loss function computation for the distillation process of intra-level spatial distribution is computed as follows:

$$\begin{aligned} L_{SKD} &= \sum_{i=1}^{|L|} KL(\sigma(\frac{\Delta D_{i,:i}}{\tau}), \sigma(\frac{G_{i,:i}}{\tau})) \\ G_{i,j} &= \cos(\hat{y}_i^s, \hat{y}_j^s), \end{aligned} \tag{12}$$

where the element $G_{i,j}$ denotes the contextual relationships within the student model.

4.5 Training Strategy

To perform knowledge distillation, we first train the teacher model using classification loss $L_{teacher} = \frac{1}{|S_u|} \sum_{i \in S_u} CE(y_i, \sigma(\hat{y}_i^t))$, where $CE(\cdot)$ is the cross-entropy loss, y_i is the label, $\hat{y}_i$ corresponds to the predicted values of the teacher model and $\sigma(\cdot)$ is the softmax function. Finally, the overall distillation objective function is defined to distill reliable spatial-temporal knowledge from the experienced teacher model into the MLPs student model is defined as follows:

$$L_{total} = \frac{\lambda}{|S_u|} \sum\nolimits_{i \in S_u} CE(y_i, \hat{y}_i^s) + (1-\lambda)(L_{RKD} + L_{SKD}) \tag{13}$$

where λ is a weight to balance the impact of student model classification loss and knowledge distillation loss.

5 EXPERIMENTS

5.1 Datasets

We evaluate our MLP-POI on two real-world and a general user check-in dataset: Gowalla [1] and Foursquare [2]. Each user check-in record in the dataset includes userID, POIID, latitude, longitude, and timestamp. The number of users, locations, check-ins, collection period, and average time interval are presented in Table 1. To ensure fairness and consistency in our experiments, we follow the preprocessing techniques employed in previous studies. We define users with fewer than one hundred check-in records in the dataset as inactive users and discard them. For the remaining users, their check-in records are sorted in ascending order based on the timestamp. In chronological order, the first 80% of check-in records for each user are used for the training set, while the remaining 20% are designated for the test set. Additionally, to align the lengths of input trajectory sequences, each user's check-in sequence is partitioned into multiple equally sized subsequences (e.g., typically set at 20).

5.2 Baselines

To evaluate the effectiveness of MLP-POI, we compare it with a series of baselines on two datasets:

- **PRME**[5]: An embedding method, which leverages the expressive capabilities of user and POI embeddings to capture personalized sequential transition patterns.
- **STRNN**[15]: An invariant RNN model, which extends spatial-temporal transition matrices into RNN.
- **Deepmove**[4]: It introduces a historical attention layer to capture long-term periodicity and enhance the short-term sequential patterns captured by gated recurrent units (GRU).
- **STGN**[35]: An LSTM-based method, which extends LSTM by introducing spatial-temporal gates for capturing long-term and short-term spatial-temporal factor preferences.
- **SASRec**[10]: An attention-based method that captures users' preferences by identifying the most relevant information from historical sequences.
- **LSTPM**[20]: An LSTM-based method, which uses a nonlocal network and a geo-dilated RNN to respectively capture long-term preference and short-term preference.
- **Flashback**[27]: An RNN-based model, that utilizes spatial-temporal interval features to search for hidden states in historical information with similar contexts to the current one.
- **STAN**[17]: An attention-based model, which further considers the spatial-temporal interval factors between non-adjacent check-ins in historical sequences.
- **GETNext**[28]: It is an encoder-decoder framework and incorporates the global transition patterns, spatial-temporal context, and category embeddings together into the model.
- **Graph-Flashback**[18]: A RNN-based model, which incorporates the weighted POI transition graph to enhance the representation of relationships between POIs.

[1] https://snap.stanford.edu/data/loc-gowalla.html
[2] https://sites.google.com/site/yangdingqi/home

Model	Gowalla				Foursquare			
	Acc@1	Acc@5	Acc@10	MRR	Acc@1	Acc@5	Acc@10	MRR
PRME	0.0740	0.2146	0.2899	0.1503	0.0982	0.3167	0.4064	0.2040
STRNN	0.0900	0.2120	0.2730	0.1508	0.2290	0.4310	0.5050	0.3248
DeepMove	0.0625	0.1304	0.1594	0.0982	0.2400	0.4319	0.4742	0.3270
STGN	0.0624	0.1586	0.2104	0.1125	0.2094	0.4734	0.5470	0.3283
SASRec	0.0787	0.1817	0.2511	0.1434	0.2399	0.4557	0.5479	0.3246
LSTPM	0.0721	0.1843	0.2327	0.1306	0.2484	0.4489	0.5018	0.3365
Flashback	0.1158	0.2754	0.3479	0.1925	0.2496	0.5399	0.6236	0.3805
STAN	0.0891	0.2096	0.2763	0.1523	0.2265	0.4515	0.5310	0.3420
GETNext	0.1343	0.3238	0.4039	0.2237	0.2667	0.5601	0.6382	0.3977
Graph-Flashback	0.1495	0.3399	0.4242	0.2401	0.2786	0.5733	0.6501	0.4109
SNPM	0.1593	0.3514	0.4346	0.2505	0.2899	0.5967	0.6763	0.4278
MLP-POI	**0.1684**	**0.3689**	**0.4561**	**0.2602**	**0.2971**	**0.6167**	**0.6983**	**0.4379**
Improvement (%)	5.7%	4.98%	4.95%	3.87%	2.48%	3.35%	3.25%	2.36%

Table 2: Performance Comparisons with Baselines on the Gowalla and Foursquare Datasets.

- **SNPM**[29]: A RNN-based model, which extracts POI relationships implied in extremely sparse check-in data into the POI similarity graph, and aggregates similar POIs to enhance the model's generalized representations.

Model Settings We implemented our MLP-POI on the PyTorch framework 1.10.1 and conducted all experiments on a Linux server with 128GB RAM, 16-core Intel i9-12900K@5.2GHz CPU, and Nvidia RTX 3090 GPU. In order to ensure a fair comparison with the baseline model, the dimensions of POI embeddings and user embeddings are set to 10 for all methods. For the trajectory completion module, the depth m of BFS is 3 and the weight threshold is set to 0.6. For the training strategy, the balancing weight λ is set to 0.5. We employed the Adam optimizer with a 1e2 learning rate, and the temporal and spatial decay factors are set up the same as in [27].

Evaluation Metrics We apply two standard evaluation metrics following [18] to validate the performance of our method, covering Accuracy@K (*Acc@K*) and mean Reciprocal Rank (*MRR*). *Acc@K* evaluates the rate of POIs that are actually checked in by the user among the top-K recommendations. In our experiments, we empirically reported the recommendation performance for $K = 1, 5, 10$. Unlike *Acc@K*, *MRR* measures the average reciprocal rank of correctly predicted POIs in the ordered recommendation list to emphasize the importance of having correctly predicted POIs ranked higher in the recommendation list. The formula will be shown in Appendix A.

5.3 Performance Comparison

Table. 2 shows the performance of MLP-POI on two datasets and substantiates its effectiveness in comparison to baseline methods. The bold results indicate the best performance, while the underscored results denote the second-best performance. During the comparative analysis of the effectiveness of various methods, we make the following observations: (1) The results on the Gowalla and Foursquare datasets demonstrate a consistent and significant improvement across all evaluation metrics for our proposed method. This is primarily attributed to the excellent performance of the MLP architecture model on the POI recommendation task and the reliable transfer of rich contextual information from the spatial-temporal teacher model, which includes information from the completed trajectories providing rich context as well as spatial-temporal information. (2) The approach of jointly modeling non-continuous spatial-temporal intervals in hidden states (e.g., Graph-Flashback, STAN, Flashback) outperforms the modeling of continuous spatial-temporal gap patterns (e.g., STGN, STRNN). Additionally, LSTPM leveraging geo-dilated POI sequences can also be interpreted as modeling the spatial relationships of non-continuous check-ins in historical trajectories. (3) In comparison, we observe a more pronounced improvement in the effectiveness of our method on the Gowalla dataset. It is noteworthy that the check-in density in the Gowalla dataset (0.002) is lower than that in the Foursquare dataset (0.003). This demonstrates our method's ability to handle sparse data. The substantial improvement is primarily attributed to our trajectory completion module, which provides more complete trajectory data, alleviating the challenge of data sparsity in the POI recommendation task.

5.4 Ablation Study

To analyze the effectiveness of different modules in MLP-POI, we conduct ablation experiments in this section and the results on the Gowalla dataset are presented in Table. 3. We denote the based model as MLP-POI, which follows a scheme of transferring knowledge from the spatial-temporal teacher model to the MLP student model. We drop different components to form variants and draw the following conclusions:

Model	Gowalla			
	Acc@1	Acc@5	Acc@10	MRR
MLP (student model)	0.1544	0.3484	0.4328	0.2467
ST model	0.1567	0.3618	0.4505	0.2523
KD	0.1556	0.3517	0.4376	0.2506
KD-single	0.1602	0.3594	0.4487	0.2540
w/o mask	0.1662	0.3637	0.4542	0.2578
w/o ST distill	0.1656	0.3629	0.4536	0.2572
MLP-POI	**0.1684**	**0.3689**	**0.4561**	**0.2602**

Table 3: Effectiveness of different modules in MLP-POI.

- **MLP (student model)** is the performance of only using the MLP architecture model, initialized with the POI embedding from Graph-Flashback, for modeling POI recommendations. We observe with surprise that it achieves results comparable to existing state-of-the-art methods, all while employing a simpler and more efficient architecture. This finding further underscores the effectiveness of MLP in modeling POI sequence data.
- **ST model (teacher model)** is the abbreviation for the spatial-temporal model. It uses the completed trajectories as input data and models them through the spatial-temporal model initialized by Graph-Flashback. We noticeably observe that its performance already surpasses existing state-of-the-art methods, indicating that the trajectory completion method is indeed effective in alleviating the issue of trajectory sparsity.
- **KD** uses a simple KL-divergence objective function to transfer the knowledge from the pre-trained spatial-temporal model to the student model. However, the performance of the student model does not show much improvement compared to using MLP directly. This is primarily attributed to the low confidence in the predicted distribution of the teacher model, making it challenging to transfer useful spatial-temporal information to the student model.
- **KD-single** differs from MLP-POI in that it utilizes only a single teacher model to transfer knowledge to the student model, and uses the Confidence-inspired Reliable Distillation module. This demonstrates that multiple teacher models can offer multi-view spatial-temporal knowledge and contribute to providing highly confident knowledge to the student model.
- **w/o mask** signifies the exclusion of the Label-based Masking Approach in MLP-POI. This indicates that the Label-based Masking Approach can effectively improve the accuracy of knowledge transfer.
- **w/o ST distill** signifies the removal of the Spatial Distribution Distillation module. It indicates that not only transferring the predicted distribution of the teacher model is effective but also transferring the intra-level spatial distribution relationships in the historical trajectory sequence is effective.

5.5 Impact of Sequence Length

In this experiment, we evaluate the impact of the sequence length on the POI recommendation task by varying the length of trajectory sequences that are evenly split as inputs to the model. As shown in Fig. (4), when increasing the length of the split trajectory sequences, we observed a significant performance decline in the MLP

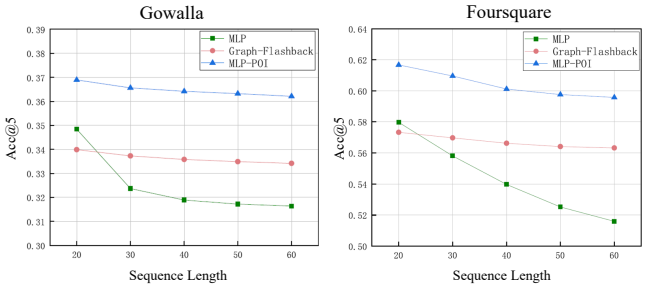


Figure 4: The performance comparison about the sequence length.

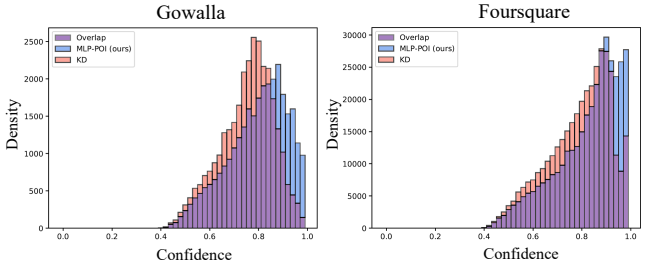


Figure 5: Confidence distribution of the distilled MLPs in general knowledge distillation and MLP-POI on two datasets, where we use the same spatial-temporal teacher model.

model initialized with POI representation from Graph-Flashback. On the other hand, Graph-Flashback maintained the weakest effect, while MLP-POI exhibited a slight downward trend. This indicates that MLP is unable to model long sequence trajectories, which is associated with its simple model architecture and its inability to perceive relative positional relationships within trajectories. However, due to the supervision of spatial-temporal teacher models, MLP-POI empowers the student MLP model to perceive predictive distributions that involve rich contextual relationship modeling, thereby alleviating its apprehension towards long trajectory sequences. This observation verifies the fact that MLP-POI is capable of inheriting the robustness in modeling long sequences from the spatial-temporal teacher model effectively. Other important sensitivity analysis experiments regarding BFS walk length and the number of teacher models are shown in Appendix A.

5.6 Evaluation on Confidence Distribution

To verify whether Confidence-inspired Reliable Distillation provides reliable confidence-based supervision to help the distilled MLP improve prediction confidence, we further visualize the predicted confidence distributions of MLP in MLP-POI on two datasets. We compare it with the optimization using only a simple distillation term defined by Eq. (1). In the Figure. (5), we can observe that the additional reliable supervision provided by the Confidence-inspired Reliable Distillation Module, as defined by Eq. (11), helps to significantly improve the predicted confidence of the student MLPs.

6 CONCLUSION

In this study, we introduce MLP-POI, a novel MLP-based knowledge distillation framework for enhancing the modeling of user behavior patterns in POI recommendation by leveraging completed potential missing check-ins and providing diverse perspectives to multiple teacher models. Specifically, we initially utilize a global transition pattern graph to guide the generation of multiple trajectory sequences with completed potential missing check-ins as diverse inputs for multiple teacher models. Then, we devise a novel confidence-inspired reliable distillation scheme, adaptively aggregating high-confidence teacher knowledge into the student model Additionally, to ensure that the student model captures the geographical relationships between POIs in trajectory sequences, we introduce spatial intra-level knowledge distillation, allowing the student model's contextual distribution to further reflect spatial similarities. We conducted comprehensive ablation experiments, sensitivity analysis of parameters, evaluation of confidence distribution in the student model, and model efficiency analysis in the experimental section. The results of the comparison with baseline models demonstrate the superiority of our model, surpassing state-of-the-art methods.

REFERENCES

[1] Chen Cheng, Haiqin Yang, Michael R. Lyu, and Irwin King. 2013. Where You Like to Go Next: Successive Point-of-Interest Recommendation. In IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013, Francesca Rossi (Ed.). IJCAI/AAAI, 2605–2611. http://www.aaai.org/ocs/index.php/IJCAI/IJCAI13/paper/view/6633

[2] Khalil Damak, Sami Khenissi, and Olfa Nasraoui. 2022. Debiasing the Cloze Task in Sequential Recommendation with Bidirectional Transformers. In KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022, Aidong Zhang and Huzefa Rangwala (Eds.). ACM, 273–282. https://doi.org/10.1145/3534678.3539430

[3] Shangchen Du, Shan You, Xiaojie Li, Jianlong Wu, Fei Wang, Chen Qian, and Changshui Zhang. 2020. Agree to Disagree: Adaptive Ensemble Knowledge Distillation in Gradient Space. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/91c77393975889bd08f301c9e13a44b7-Abstract.html

[4] Jie Feng, Yong Li, Chao Zhang, Funing Sun, Fanchao Meng, Ang Guo, and Depeng Jin. 2018. DeepMove: Predicting Human Mobility with Attentional Recurrent Networks. In Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018, Pierre-Antoine Champin, Fabien Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis (Eds.). ACM, 1459–1468. https://doi.org/10.1145/3178876.3186058

[5] Shanshan Feng, Xutao Li, Yifeng Zeng, Gao Cong, Yeow Meng Chee, and Quan Yuan. 2015. Personalized Ranking Metric Embedding for Next New POI Recommendation. In Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015, Qiang Yang and Michael J. Wooldridge (Eds.). AAAI Press, 2069–2075. http://ijcai.org/Abstract/15/293

[6] Ehsan Gholami, Mohammad Motamedi, and Ashwin Aravindakshan. 2022. PARSRec: Explainable Personalized Attention-fused Recurrent Sequential Recommendation Using Session Partial Actions. In KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022, Aidong Zhang and Huzefa Rangwala (Eds.). ACM, 454–464. https://doi.org/10.1145/3534678.3539432

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. IEEE Computer Society, 770–778. https://doi.org/10.1109/CVPR.2016.90

[8] Ruining He and Julian J. McAuley. 2016. Fusing Similarity Models with Markov Chains for Sparse Sequential Recommendation. In IEEE 16th International Conference on Data Mining, ICDM 2016, December 12-15, 2016, Barcelona, Spain, Francesco Bonchi, Josep Domingo-Ferrer, Ricardo Baeza-Yates, Zhi-Hua Zhou, and Xindong Wu (Eds.). IEEE Computer Society, 191–200. https://doi.org/10.1109/ICDM.2016.0030

[9] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the Knowledge in a Neural Network. CoRR abs/1503.02531 (2015). arXiv:1503.02531 http://arxiv.org/abs/1503.02531

[10] Wang-Cheng Kang and Julian J. McAuley. 2018. Self-Attentive Sequential Recommendation. In IEEE International Conference on Data Mining, ICDM 2018, Singapore, November 17-20, 2018. IEEE Computer Society, 197–206. https://doi.org/10.1109/ICDM.2018.00035

[11] Dejiang Kong and Fei Wu. 2018. HST-LSTM: A Hierarchical Spatial-Temporal Long-Short Term Memory Network for Location Prediction. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden, Jérôme Lang (Ed.). ijcai.org, 2341–2347. https://doi.org/10.24963/IJCAI.2018/324

[12] Yehuda Koren, Robert M. Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. Computer 42, 8 (2009), 30–37. https://doi.org/10.1109/MC.2009.263

[13] Yang Li, Yadan Luo, Zheng Zhang, Shazia W. Sadiq, and Peng Cui. 2021. Context-Aware Attention-Based Data Augmentation for POI Recommendation. CoRR abs/2106.15984 (2021). arXiv:2106.15984 https://arxiv.org/abs/2106.15984

[14] Defu Lian, Yongji Wu, Yong Ge, Xing Xie, and Enhong Chen. 2020. Geography-Aware Sequential Location Recommendation. In KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020, Rajesh Gupta, Yan Liu, Jiliang Tang, and B. Aditya Prakash (Eds.). ACM, 2009–2019. https://doi.org/10.1145/3394486.3403252

[15] Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2016. Predicting the Next Location: A Recurrent Model with Spatial and Temporal Contexts. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA, Dale Schuurmans and Michael P. Wellman (Eds.). AAAI Press, 194–200. https://doi.org/10.1609/AAAI.V30I1.9971

[16] Yan Luo, Haoyi Duan, Ye Liu, and Fu-Lai Chung. 2023. Timestamps as Prompts for Geography-Aware Location Recommendation. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023, Ingo Frommholz, Frank Hopfgartner, Mark Lee, Michael Oakes, Mounia Lalmas, Min Zhang, and Rodrygo L. T. Santos (Eds.). ACM, 1697–1706. https://doi.org/10.1145/3583780.3615083

[17] Yingtao Luo, Qiang Liu, and Zhaocheng Liu. 2021. STAN: Spatio-Temporal Attention Network for Next Location Recommendation. In WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021, Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia (Eds.). ACM / IW3C2, 2177–2185. https://doi.org/10.1145/3442381.3449998

[18] Xuan Rao, Lisi Chen, Yong Liu, Shuo Shang, Bin Yao, and Peng Han. 2022. Graph-Flashback Network for Next Location Recommendation. In KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022, Aidong Zhang and Huzefa Rangwala (Eds.). ACM, 1463–1471. https://doi.org/10.1145/3534678.3539383

[19] Steffen Rendle. 2010. Factorization Machines. In ICDM 2010, The 10th IEEE International Conference on Data Mining, Sydney, Australia, 14-17 December 2010, Geoffrey I. Webb, Bing Liu, Chengqi Zhang, Dimitrios Gunopulos, and Xindong Wu (Eds.). IEEE Computer Society, 995–1000. https://doi.org/10.1109/ICDM.2010.127

[20] Ke Sun, Tieyun Qian, Tong Chen, Yile Liang, Quoc Viet Hung Nguyen, and Hongzhi Yin. 2020. Where to Go Next: Modeling Long- and Short-Term User Preferences for Point-of-Interest Recommendation. In The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020. AAAI Press, 214–221. https://doi.org/10.1609/AAAI.V34I01.5353

[21] En Wang, Yiheng Jiang, Yuanbo Xu, Liang Wang, and Yongjian Yang. 2022. Spatial-Temporal Interval Aware Sequential POI Recommendation. In 38th IEEE International Conference on Data Engineering, ICDE 2022, Kuala Lumpur, Malaysia, May 9-12, 2022. IEEE, 2086–2098. https://doi.org/10.1109/ICDE53745.2022.00202

[22] Zhaobo Wang, Yanmin Zhu, Chunyang Wang, Wenze Ma, Bo Li, and Jiadi Yu. 2023. Adaptive Graph Representation Learning for Next POI Recommendation. In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023, Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete (Eds.). ACM, 393–402. https://doi.org/10.1145/3539618.3591634

[23] Lirong Wu, Haitao Lin, Yufei Huang, and Stan Z. Li. 2023. Quantifying the Knowledge in GNNs for Reliable Distillation into MLPs. In International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA (Proceedings of Machine Learning Research, Vol. 202), Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 37571–37581. https://proceedings.mlr.press/v202/wu23m.html

[24] Meng-Chieh Wu, Ching-Te Chiu, and Kun-Hsuan Wu. 2019. Multi-teacher Knowledge Distillation for Compressed Video Action Recognition on Deep

Neural Networks. In IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019. IEEE, 2202–2206. https://doi.org/10.1109/ICASSP.2019.8682450

[25] Dongbo Xi, Fuzhen Zhuang, Yanchi Liu, Jingjing Gu, Hui Xiong, and Qing He. 2019. Modelling of Bi-Directional Spatio-Temporal Dependence and Users' Dynamic Preferences for Missing POI Check-In Identification. In The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019. AAAI Press, 5458–5465. https://doi.org/10.1609/AAAI.V33I01.33015458

[26] Chen Xu, Quan Li, Junfeng Ge, Jinyang Gao, Xiaoyong Yang, Changhua Pei, Fei Sun, Jian Wu, Hanxiao Sun, and Wenwu Ou. 2020. Privileged Features Distillation at Taobao Recommendations. In KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020, Rajesh Gupta, Yan Liu, Jiliang Tang, and B. Aditya Prakash (Eds.). ACM, 2590–2598. https://doi.org/10.1145/3394486.3403309

[27] Dingqi Yang, Benjamin Fankhauser, Paolo Rosso, and Philippe Cudré-Mauroux. 2020. Location Prediction over Sparse User Mobility Traces Using RNNs: Flashback in Hidden States!. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020, Christian Bessiere (Ed.). ijcai.org, 2184–2190. https://doi.org/10.24963/IJCAI.2020/302

[28] Song Yang, Jiamou Liu, and Kaiqi Zhao. 2023. GETNext: Trajectory Flow Map Enhanced Transformer for Next POI Recommendation. CoRR abs/2303.04741 (2023). https://doi.org/10.48550/ARXIV.2303.04741 arXiv:2303.04741

[29] Feiyu Yin, Yong Liu, Zhiqi Shen, Lisi Chen, Shuo Shang, and Peng Han. 2023. Next POI Recommendation with Dynamic Graph and Explicit Dependency. In Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023, Brian Williams, Yiling Chen, and Jennifer Neville (Eds.). AAAI Press, 4827–4834. https://doi.org/10.1609/AAAI.V37I4.25608

[30] Shan You, Chang Xu, Chao Xu, and Dacheng Tao. 2017. Learning from Multiple Teacher Networks. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017. ACM, 1285–1294. https://doi.org/10.1145/3097983.3098135

[31] Fei Yuan, Linjun Shou, Jian Pei, Wutao Lin, Ming Gong, Yan Fu, and Daxin Jiang. 2021. Reinforced Multi-Teacher Selection for Knowledge Distillation. In Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021. AAAI Press, 14284–14291. https://doi.org/10.1609/AAAI.V35I16.17680

[32] Hailin Zhang, Defang Chen, and Can Wang. 2022. Confidence-Aware Multi-Teacher Knowledge Distillation. In IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022. IEEE, 4498–4502. https://doi.org/10.1109/ICASSP43922.2022.9747534

[33] Hailin Zhang, Defang Chen, and Can Wang. 2023. Adaptive Multi-Teacher Knowledge Distillation with Meta-Learning. In IEEE International Conference on Multimedia and Expo, ICME 2023, Brisbane, Australia, July 10-14, 2023. IEEE, 1943–1948. https://doi.org/10.1109/ICME55011.2023.00333

[34] Yipeng Zhang, Xin Wang, Hong Chen, and Wenwu Zhu. 2023. Adaptive Disentangled Transformer for Sequential Recommendation. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023, Ambuj K. Singh, Yizhou Sun, Leman Akoglu, Dimitrios Gunopulos, Xifeng Yan, Ravi Kumar, Fatma Ozcan, and Jieping Ye (Eds.). ACM, 3434–3445. https://doi.org/10.1145/3580305.3599253

[35] Pengpeng Zhao, Haifeng Zhu, Yanchi Liu, Jiajie Xu, Zhixu Li, Fuzhen Zhuang, Victor S. Sheng, and Xiaofang Zhou. 2019. Where to Go Next: A Spatio-Temporal Gated Network for Next POI Recommendation. In The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019. AAAI Press, 5877–5884. https://doi.org/10.1609/AAAI.V33I01.33015877

A APPENDIX

A.1 A Implementation Details of MLP-POI

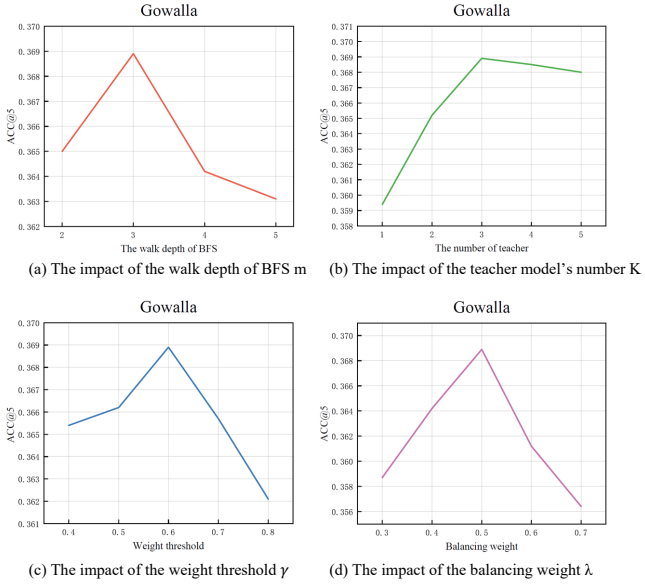
Details of Evaluation Metrics. It is formulated as follows:

$$Acc@K = \frac{\sum_{|Test|} \sigma(trg \in TopK)}{|Test|} \tag{14}$$

where $TopK$ and trg are the set of top-K POIs in the recommendation list and the ground truth, respectively. $|Test|$ denotes the number of test cases. σ is an indicator function that takes the value 1 when the condition is true and 0 otherwise. Unlike $Acc@K$, MRR measures the average reciprocal rank of correctly predicted POIs in the ordered recommendation list to emphasize the importance of having correctly predicted POIs ranked higher in the recommendation list. It is presented as follows:

$$MRR = \frac{1}{|Test|} \sum_{i=1}^{|Test|} \frac{1}{rank_i} \tag{15}$$

where $rank_i$ is the index of the correctly predicted POI in the order recommendation list.



(a) The impact of the walk depth of BFS m (b) The impact of the teacher model's number K

(c) The impact of the weight threshold γ (d) The impact of the balancing weight λ

Figure 6: The performance comparison about the number of the walk depth m of BFS and the number of teacher model K.

Hyperparameter Sensitivity Analysis. We presented parameter sensitivity experiments and analyzed the corresponding experimental results. We analyzed the lengths m of BFS walks in the Trajectory Completion Module and the number K of teachers in the multi-teacher module separately to identify the optimal parameter values on the Gowalla dataset. As shown in the Figure. 6, the results indicate that the optimal length for BFS walks is $m = 3$, the most effective number of teachers in the teacher model is $K = 3$, the most suitable set of weight threshold γ and balancing weight λ is $\gamma = 0.6$ and $\lambda = 0.5$. And, we have the following observations:

- We varied the length of BFS walks from 2 to 5 in Figure. 6(a). It represents the maximum length of subsequences that may be completed between any two consecutive check-ins in the raw trajectory sequence. We observe that using a shorter search length of BFS did not fully exploit the benefits of completing trajectory sequences. The model reached its peak performance at a moderate length of 3, indicating that completing trajectory sequences greatly aids in modeling rich contextual information. Further increasing the search depth impaired the model's performance. This is attributed to the potential noise introduced by excessively long missing sub-trajectories, posing a challenge to effective modeling of trajectory context.
- We conducted a series of experiments with the number K of teachers in the teacher model varying from 2 to 5 in Figure. 6(b). We find that a smaller number of teachers led to poorer results. The optimal performance was achieved when the number of teachers $K = 3$. Subsequently, as the number of teachers increased, the model's performance slightly decreased. Therefore, considering both model effectiveness and experimental efficiency, the best parameter for MLP-POI is $K = 3$.
- Figure. 6(c) shows that $\gamma = 0.6$ is the best weight threshold for potential missing nodes completion strategy on Gowalla. The reason could be that excessively high thresholds filter out many useful potential missing nodes, while overly low thresholds generate more low-reliability potential missing nodes. Therefore, enhancing the potential prompt set with more reliable nodes is essential for improving the overall performance of the model.
- We find in Figure. 6 that the balancing weight achieves the best performance when $\lambda = 0.5$. This also indicates that both the loss of the student model and the distillation loss play equally important roles in optimizing the model.