Spatial-Temporal Knowledge Distillation for Takeaway Recommendation

Shuyuan Zhao[1,2*], **Wei Chen**[1,2*], **Boyan Shi**[1,2], **Liyong Zhou**[1,2], **Shuohao Lin**[1,2], **Huaiyu Wan**[1,2 †]

[1]School of Computer Science & Technology, Beijing Jiaotong University, Beijing, China
[2]Beijing Key Laboratory of Traffic Data Mining and Embodied Intelligence, Beijing, China
{sy_zhao, w_chen, by_shi, zhouly, shhlin, hywan}@bjtu.edu.cn

Abstract

The takeaway recommendation system aims to recommend users' future takeaway purchases based on their historical purchase behaviors, thereby improving user satisfaction and boosting merchant sales. Existing methods focus on incorporating auxiliary information or leveraging knowledge graphs to alleviate the sparsity issue of user purchase sequences. However, two main challenges limit the performance of these approaches: (1) capturing dynamic user preferences on complex geospatial information and (2) efficiently integrating spatial-temporal knowledge from both graphs and sequence data with low computational costs. In this paper, we propose a novel **s**patial-**t**emporal **k**nowledge **d**istillation model for takeaway **rec**ommendation (STKDRec) based on the two-stage training process. Specifically, during the first pre-training stage, a spatial-temporal knowledge graph (STKG) encoder is trained to extract high-order spatial-temporal dependencies and collaborative associations from the STKG. During the second spatial-temporal knowledge distillation (STKD) stage, a spatial-temporal Transformer (ST-Transformer) is employed to comprehensively model dynamic user preferences on various types of fine-grained geospatial information from a sequential perspective. Furthermore, the STKD strategy is introduced to transfer graph-based spatial-temporal knowledge to the ST-Transformer, facilitating the adaptive fusion of rich knowledge derived from both the STKG and sequence data while reducing computational overhead. Extensive experiments on three real-world datasets show that STKDRec significantly outperforms the state-of-the-art baselines.

Code — https://github.com/Zhaoshuyuan0246/STKDRec

Introduction

Takeaway platforms, such as Yelp, Meituan, and Ele.me, provide convenient online ordering and offline delivery services, playing an increasingly important role in people's daily lives (Zhang et al. 2023). As the core service of these platforms, takeaway recommendation aim to accurately recommend takeaways that align with user preferences based on their historical purchase behaviors. Such recommendation services enhance user satisfaction while increasing visibility and sales opportunities for merchants. In recent years, takeaway recommendations have gained significant attention from the research community and industry (Lin et al. 2022; Du et al. 2023a; Shi et al. 2024).

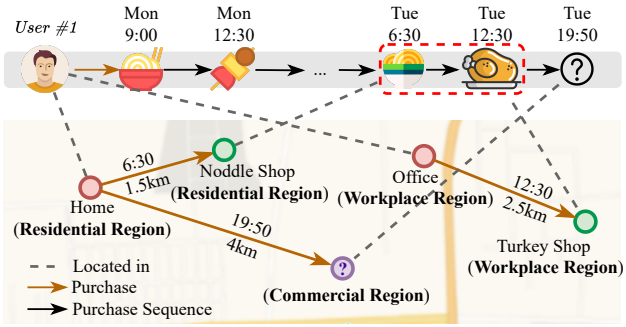

Figure 1: An illustrative example highlighting the importance of capturing dynamic user preferences on complex geospatial information.

Takeaway recommendation is essentially a sequential modeling task that aims to predict the user's future takeaway purchases based on their historical purchase records. Existing methods model these purchase sequences by utilizing deep neural networks, such as recurrent neural networks (RNNs) and self-attention mechanisms. However, the sparsity issue arises since users often purchase only a few takeaways, limiting the performance of recommendations (Kang et al. 2023). Some studies attempt to incorporate various types of auxiliary information, such as brand, category, location, and area of interest (AOI), into purchase sequences. Some methods, such as KGDPL (Liu, Zhu, and Wu 2023), also utilize graph neural networks (GNNs) and Knowledge Graphs (KGs) (Wu et al. 2023; Chen et al. 2024) to explore higher-order user-takeaway relationships or rich semantics of takeaways, alleviating the data sparsity problem. Although these methods all achieve promising performance, two significant challenges still need to be addressed:

(1) Failing to effectively capture dynamic user preferences on complex geospatial information. In takeaway recommendation scenarios, user preferences change dynamically over time and with their current location. For example, users tend to purchase fast food at noon when at the workplace, while they prefer main meals in the evening when at

*These authors contributed equally.
†Corresponding authors.

home. However, complex geospatial information, including the distance between the user and the delivery merchant, the functional region of the merchant, and so on, as an important factor affecting user preferences, has not been adequately explored. As illustrated in Figure 1, *User #1* is primarily located in a workplace region during the day, frequently purchasing food that matches their preferences from nearby shops. However, when the user returns to the residential region in the evening, existing methods prioritize recommending nearby food candidates. Satisfying user preferences for foods from a more distant commercial region is neglected due to insufficient consideration of complex geospatial information. Therefore, capturing dynamic user preferences on complex geospatial information poses a challenge.

(2) How to efficiently integrate spatial-temporal knowledge from both graphs and sequence data with low computational costs. The user's purchase history is sequential data, while KGs are non-Euclidean structure data. Effectively integrating the advantages of these two types of heterogeneous data can alleviate the challenge of data sparsity and improve the accuracy of recommendations. However, due to the typically large scale of KGs, encoding them with GNNs significantly increases computational overhead. Additionally, simple knowledge fusion methods, such as addition or concatenation, are not conducive to integrating these heterogeneous data for subsequent recommendations. Therefore, finding an effective method to fuse spatial-temporal knowledge from both graphs and sequence data while reducing computational costs is crucial.

To address these challenges, we propose a novel **S**patial-**T**emporal **K**nowledge **D**istillation model for takeaway **rec**ommendation, termed **STKDRec**. The model distills the offline teacher model's knowledge of the graph structure to better enhance the student model's ability to model users' historical purchase sequences while improving computational efficiency. STKDRec consists of two stages: pre-training and spatial-temporal knowledge distillation (STKD). During the pre-training stage, STKDRec constructs a spatial-temporal knowledge graph (STKG) and extracts high-order spatial-temporal dependencies and collaborative associations between users and takeaways from STKG through training an STKG encoder. During the STKD stage, a spatial-temporal Transformer (ST-Transformer) is presented to capture dynamic user preferences on various types of fine-grained geospatial information from a sequential perspective. Through the STKD strategy, graph-based spatial-temporal knowledge from the STKG encoder is effectively transferred to the ST-Transformer, facilitating heterogeneous knowledge fusion with low computational costs.

The contributions of this work are as follows:

- We propose a novel spatial-temporal knowledge distillation model for takeaway recommendations, utilizing knowledge distillation to fuse spatial-temporal knowledge from both STKG and sequence data, thereby addressing the sparsity issue in sequence data and effectively reducing computational overhead.
- We integrate various types of geospatial information into user sequence data and propose an ST-Transformer to capture dynamic user preferences on complex geospatial information from a sequential perspective.
- Extensive experiments on three real-world datasets show that STKDRec achieves superior recommendation performance over the state-of-the-art baselines.

Related Work

Takeaway Recommendation

Takeaway recommendation methods recommend personalized takeaways to users based on their historical purchase sequences. Early methods used Markov Chains (MC) (Rendle, Freudenthaler, and Schmidt-Thieme 2010) to model user sequences. With the development of deep learning, methods such as GNNs, RNNs, and self-attention mechanisms (Hidasi et al. 2015; Zhang et al. 2022; Chen et al. 2022; Shin et al. 2024) have been used to model users' time-varying interests. Some methods attempt to capture users' spatial-temporal interests by considering spatial-temporal information. StEN (Lin et al. 2022) models the spatial-temporal information of users and foods for click-through rate prediction in location-based service. BASM (Du et al. 2023a) integrates spatial-temporal features to capture user preferences at different times and locations. However, these methods primarily focus on spatial region information (e.g., geohash or AOI) without accounting for fine-grained spatial information of different types, such as spatial distance between the users and the takeaways. They fail to comprehensively and accurately capture dynamic user preferences in takeaway recommendations. Therefore, we propose a spatial-enhanced sequence representation encompassing various types of geospatial information and utilize an ST-Transformer to learn dynamic user preferences from a sequential perspective.

Knowledge Distillation

Knowledge distillation (Hinton, Vinyals, and Dean 2015) achieves lightweight and performance enhancement by transferring knowledge from a teacher model to a student model. Existing methods (Kang et al. 2021; Xia et al. 2022; Zhu et al. 2021) can be categorized into distillation between models of the same type and distillation between models of different types. Distillation between models of the same type aims to reduce parameters, enabling the student model to approach the teacher model's performance with fewer parameters. For example, TinyBERT (Devlin et al. 2018) uses the BERT model for pre-training and fine-tuning, distilling it into the smaller language model. Distillation can also be performed between different types of models. When there are structural or mechanistic differences between the models, the teacher model can provide a unique knowledge background and global understanding that are difficult for the student model to obtain independently. For instance, Graphless Neural Network (Zhang et al. 2021) translates the knowledge from the graph structure of GNNs into the MLP format. Motivated by these approaches, we leverage knowledge distillation to transfer the graph-based spatial-temporal knowledge from the STKG encoder to the ST-Transformer, thereby

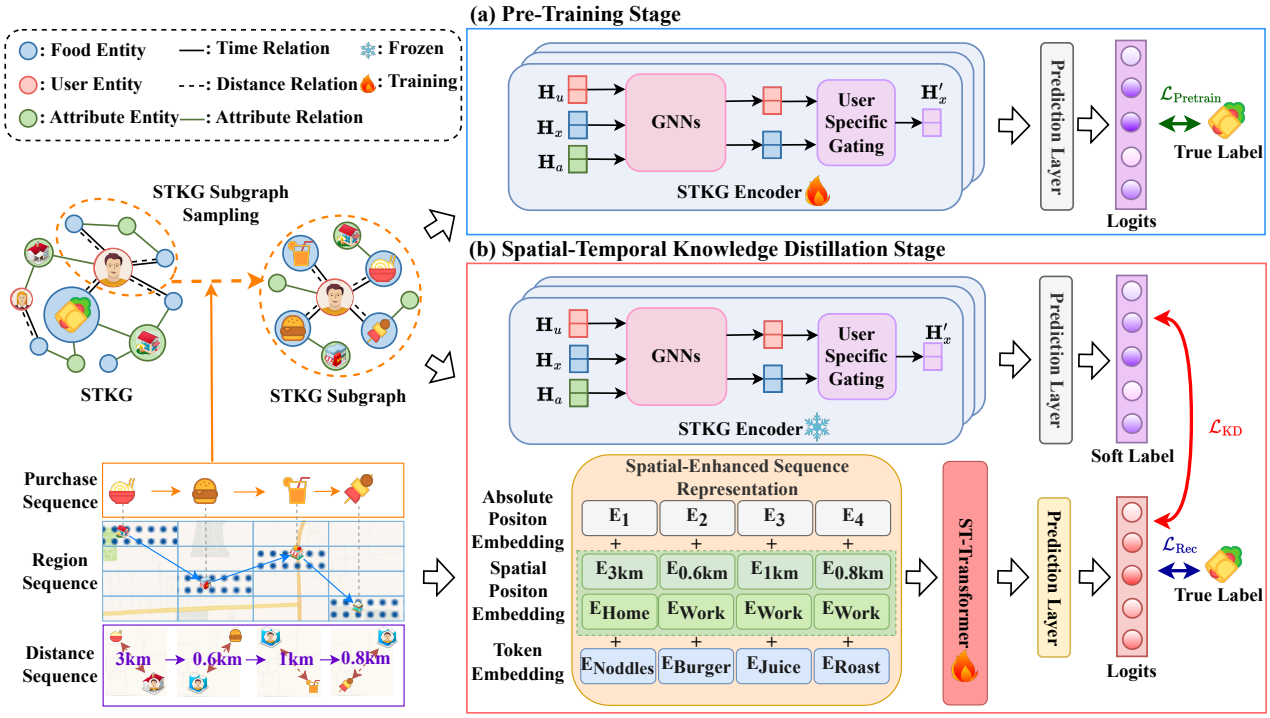

Figure 2: The overall architecture of STKDRec, consisting of two stages: the pre-training stage and the STKD stage.

achieving the purpose of heterogeneous knowledge fusion and reducing computational costs.

Preliminaries

Problem Formulation The goal of takeaway recommendation is to predict the users' next takeaway purchase based on their historical purchase sequences. Given a set of users $\mathcal{U}$ and a set of takeaways $\mathcal{V}$, we can sort the purchased takeaways of each user $u \in \mathcal{U}$ chronologically in a sequence as $x = (v_1, v_2, \ldots, v_{|x|})$, where $v_i \in \mathcal{V}$ denotes the i-th purchased takeaway in the sequence. The task is to recommend a Top-k list of takeaways as potential candidates for the user's next purchase. Formally, we predict $P(v_{|x|+1} \mid x)$.

Spatial-Temporal Knowledge Graph The STKG is represented as $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$, which consists of a set of triples composed of entity-relation-entity. Here, $\mathcal{E}$, $\mathcal{R}$, and $\mathcal{T}$ denote the entity set, the relation set, and the triple set, respectively. In the STKG, entities include users, takeaways, and their associated attributes, while relations encompass *time relation*, *distance relation*, and *attribute relation*. These entities and relations form four different types of triples: time triples of the user purchasing the takeaway $(u, \textit{time}, v)$, distance triples of the user purchasing the takeaway $(u, \textit{distance}, v)$, user-attribute triples $(u, \textit{attribute}, a)$, and takeaway-attribute triples $(v, \textit{attribute}, a)$.

Our Approach

The overall framework of STKDRec is shown in Figure 2, which consists of two stages: the pre-training stage and the STKD stage. During the pre-training stage, an STKG encoder is trained to model the high-order spatial-temporal dependencies and collaborative associations between users and takeaways from the perspective of the graph structure. During the STKD stage, an ST-Transformer is employed to model the dynamic user preferences on complex spatial information from a sequential perspective. The STKD strategy facilitates the fusion of spatial-temporal knowledge from both STKG and sequence data.

Spatial-Temporal Knowledge Graph Encoder

The STKG encoder, as the teacher model, is pre-trained to capture high-order spatial-temporal knowledge from the STKG. Thus, we sample an STKG subgraph from STKG based on the user purchase sequence. The STKG encoder is used to aggregate spatial-temporal knowledge from the subgraph while integrating personalized user features.

STKG Subgraph Sampling In this work, we use the efficient neighborhood sampling method (Hamilton, Ying, and Leskovec 2017) to sample the subgraph from the STKG. Specifically, given a user's purchase sequence x and a maximum sequence length n, the sequence is truncated by removing the earliest takeaways if $|x| > n$ or padded with 0 to get a fixed length sequence $x = (v_1, v_2, \ldots, v_n)$. Each node v_i in x is treated as a center node, and a fixed number s of neighbor nodes are randomly sampled from its adjacent nodes in the STKG. This process retains the relationships between v_i and sampled neighbors. The same process is recursively applied for each neighbor node to sample additional adjacent nodes and relationships, continuing to a specified depth

m. All center nodes, their sampled neighbors, and the retained relations are then integrated into an STKG subgraph, denoted as $\mathcal{G}_x$.

Spatial-Temporal Knowledge Aggregation To model the higher-order spatial-temporal dependencies and collaborative associations between users and takeaways with $\mathcal{G}_x$, we utilize the GNNs to effectively encode the $\mathcal{G}_x$. In the l-th layer of the GNNs, the message passing and aggregation processes are defined as follows:

$$m_{v_i}^l = \text{Aggregate}^l\left(\{\{h_{v_j}^{(l-1)}, h_r\} : \exists\,(v_i, r, v_j) \in \mathcal{G}_x\}\right), \tag{1}$$

$$h_{v_i}^l = \text{Combine}^l\left(m_{v_i}^l, h_{v_i}^{(l-1)}\right), \tag{2}$$

where $h_{v_i}^{(l-1)}$ and $h_{v_j}^{(l-1)}$ represent the embeddings of entity v_i and its neighboring entity v_j at the layer $(l-1)$, respectively. h_r represents the embedding of relationship $r \in \mathcal{R}$ between v_i and v_j. $m_{v_i}^l$ denotes the representation of aggregated neighborhood message for v_i at layer l. Aggregate$(\cdot)$ is a function that aggregates the neighborhood information of the central node v_i, while Combine$(\cdot)$ merges this information to update the entity embeddings. After propagation information through multiple GNN layers on $\mathcal{G}_x$, we obtain the final embeddings of all entities in x, denoted as $\mathrm{H}_x \in \mathbb{R}^{n\times d}$, and the final embedding for user u, denoted as $\mathrm{H}_u \in \mathbb{R}^{1\times d}$, where d is the latent dimension.

To model users' specific spatial-temporal preferences, we introduce a user specific gating mechanism to incorporate personalized user features into H_x. The user specific representation H'_x is defined as:

$$\mathrm{H}'_x = \mathrm{H}_x \otimes \sigma(\mathrm{H}_x\mathrm{W}_1 + \mathrm{W}_2\mathrm{H}_u^\top), \tag{3}$$

where $\mathrm{W}_1 \in \mathbb{R}^{d\times 1}$ and $\mathrm{W}_2 \in \mathbb{R}^{n\times d}$ represent learnable parameters, $\sigma(\cdot)$ denotes the sigmoid activation function, $\otimes$ indicates element-wise multiplication.

Soft Labels To facilitate the subsequent STKD stage, we use the prediction distribution generated by the STKG encoder as soft labels. These labels guide the student model in learning the high-order spatial-temporal knowledge from the STKG. Specifically, after deriving H'_x, an attention mechanism is applied to obtain the representation of x. The soft labels are formally defined as:

$$\mathrm{Y}'_x = \text{Softmax}(\text{AttNet}(\mathrm{H}'_x)\mathrm{E}_\mathcal{V}^\top), \tag{4}$$

where $\mathrm{Y}'_x \in \mathbb{R}^{1\times|\mathcal{V}|}$, the i-th element of Y'_x represents the purchase probability of the i-th takeaway, AttNet$(\cdot)$ denotes the attention network, and $\mathrm{E}_\mathcal{V} \in \mathbb{R}^{|\mathcal{V}|\times d}$ represents the embedding of all takeaways.

Spatial-Temporal Transformer

The ST-Transformer, as the student model, is designed to model dynamic user preferences on complex geospatial information from a sequential perspective. In real-world scenarios, complex geospatial information (e.g., spatial regions and spatial distances) significantly influences user preferences. Spatial region reflects users' general preferences, while spatial distance reveals users' specific preferences across regions. To model these various types of geospatial information, we introduce a spatial-enhanced sequence representation that integrates these diverse geospatial factors, enabling the ST-Transformer to learn dynamic user preferences that vary across these factors.

Spatial-Enhanced Sequence Representation To utilize the sequential order of tokens in the sequence, previous works (Sun et al. 2019; Gao et al. 2023) add an absolute position embedding $\mathrm{E_P}$ to enhance the sequence. Inspired by these, we propose a novel spatial position embedding to enhance user purchase sequences, enabling the ST-Transformer to focus not only on the users' evolving interests over time but also on how these interests shift across various geospatial information. In our approach, we integrate spatial regions and spatial distances to construct the spatial position embeddings. Specifically, the spatial region set $\mathcal{C}$ encompasses predefined geohash6 attributes (Du et al. 2023a) of all takeaways, and the spatial distance set $\mathcal{F}$ encompasses distances between the regions of users and the regions of the takeaways. Given the sequence x of the user u with length n, the embeddings matrix of all takeaways in x is denoted as $\mathrm{E}_x \in \mathbb{R}^{n\times d}$. Similarly, the spatial regions embedding matrix $\mathrm{E}_{x_c} \in \mathbb{R}^{n\times d}$ and the spatial distance embedding matrix $\mathrm{E}_{x_f} \in \mathbb{R}^{n\times d}$ are defined based on the spatial region sequence $x_c = (c_1, c_2, \ldots, c_n)$ and the spatial distance sequence $x_f = (f_1, f_2, \ldots, f_n)$, where $c_i \in \mathcal{C}$ and $f_i \in \mathcal{F}$. By combining E_{x_c} with E_{x_f}, we obtain a learnable spatial position embedding $\mathrm{E_{SP}}$:

$$\mathrm{E_{SP}} = \mathrm{W_{SP}}(\mathrm{E}_{x_c} + \mathrm{E}_{x_f}), \tag{5}$$

where $\mathrm{W_{SP}}$ represents a learnable parameter. Finally, we sum the three embeddings $\mathrm{E}_x, \mathrm{E_{SP}}, \mathrm{E_P}$ to produce the spatial-enhanced sequence representation $\hat{\mathrm{E}}_x \in \mathbb{R}^{n\times d}$,

$$\hat{\mathrm{E}}_x = \mathrm{E}_x + \mathrm{E_{SP}} + \mathrm{E_P}. \tag{6}$$

Spatial-Temporal Context Attention To extract dynamic user preferences that vary across regions and distances from spatial-enhanced sequence representation, we present the spatial-temporal context attention mechanism. Specifically, this mechanism comprises L layers of mask self-attention layers stacked together, transforming the input embedding $\hat{\mathrm{E}}_x$ into the spatial-temporal context representation $\hat{\mathrm{H}}_x \in \mathbb{R}^{n\times d}$. In each layer, we employ three independent linear transformation matrices $\mathrm{W}_i^q, \mathrm{W}_i^k, \mathrm{W}_i^v \in \mathbb{R}^{d\times d'}$ to transform the input embedding $\hat{\mathrm{E}}_x$ into queries, keys, and values for the i-th scaled dot-product attention head, where $d' = \frac{d}{K}$, and $i = 1, 2, \ldots, K$. The function is defined as follows:

$$\hat{\mathrm{h}}_i = \text{Softmax}\left(\frac{(\hat{\mathrm{E}}_x\mathrm{W}_i^q)(\hat{\mathrm{E}}_x\mathrm{W}_i^k)^\top}{\sqrt{d'}}\right)(\hat{\mathrm{E}}_x\mathrm{W}_i^v), \tag{7}$$

where $\hat{\mathrm{h}}_i \in \mathbb{R}^{n\times d'}$ denotes the output representation of the corresponding attention head. We concatenate the outputs from all attention heads to obtain the final spatial-temporal context representation $\hat{\mathrm{H}}_x$,

$$\hat{\mathrm{H}}_x = \text{FFN}(\text{Concatenate}(\hat{\mathrm{h}}_1, \hat{\mathrm{h}}_2, \ldots \hat{\mathrm{h}}_K)), \tag{8}$$

where FFN$(\cdot)$ denotes the feed-forward network.

Prediction Layer To achieve the takeaway recommendation task, the final spatial-temporal context representation $\hat{\mathrm{H}}_x$ is multiplied by the embeddings $\mathrm{E}_{\mathcal{V}}$ of all takeaways to predict the probability of takeaway appearing at $(n+1)$ step:

$$\hat{\mathrm{Y}}_x = \mathrm{Softmax}(\hat{\mathrm{H}}_x \mathrm{E}_{\mathcal{V}}^{\top}), \tag{9}$$

where the j-th element of $\hat{\mathrm{Y}}_x$ denotes the purchase probability of the j-th takeaway.

Spatial-Temporal Knowledge Distillation

Although the ST-Transformer captures dynamic user preferences from a sequential perspective, it fails to capture the high-order spatial-temporal dependencies and collaborative associations between users and takeaways from the STKG. Thus, we propose the STKD strategy, which facilitates the transfer of graph-based spatial-temporal knowledge from the STKG encoder to the more efficient and lightweight ST-Transformer. This approach enables the heterogeneous fusion of spatial-temporal knowledge while significantly reducing the computational costs.

During the pre-training stage, the teacher model STKG encoder supervises the learning process using ground truth labels $\mathrm{Y}_x \in \mathbb{R}^{1\times|\mathcal{V}|}$, generating soft labels Y'_x following pre-training. The pre-training loss is defined as:

$$\mathcal{L}_{\text{Pretrain}} = \mathrm{CrossEntropy}(\mathrm{Y}'_x, \mathrm{Y}_x). \tag{10}$$

During the STKD stage, our goal is to distill valuable spatial-temporal knowledge from the STKG encoder, thereby enhancing the ST-Transformer's ability to capture user preferences from different perspectives and promoting more efficient, streamlined learning. To achieve this, we train the student model, ST-Transformer, to emulate the soft labels of the teacher model, effectively transferring knowledge from the teacher to the student. The distillation loss is defined as:

$$\mathcal{L}_{\text{KD}}(\mathrm{Y}'_x, \hat{\mathrm{Y}}_x) = \mathrm{KL}\left(\mathrm{Y}'_x/\tau \,\middle\|\, \hat{\mathrm{Y}}_x/\tau\right), \tag{11}$$

where KL denotes the Kullback-Leibler divergence, and τ is the temperature coefficient. Furthermore, to ensure that the student model also learns the true labels, we optimize a supervised loss:

$$\mathcal{L}_{\text{Rec}} = \mathrm{CrossEntropy}(\hat{\mathrm{Y}}_x, \mathrm{Y}_x). \tag{12}$$

Ultimately, the student model is trained by jointly optimizing the distillation loss and the supervised loss:

$$\mathcal{L} = \alpha\mathcal{L}_{\text{KD}} + (1-\alpha)\mathcal{L}_{\text{Rec}}. \tag{13}$$

where $\alpha \in [0, 1]$ is the coefficient to balance the distillation loss and the supervised loss.

Experiments

Experimental Setting

Datasets We select three publicly available city takeaway recommendation datasets for evaluation: Wuhan, Sanya, and Taiyuan, which are provided by the well-known takeaway platform Ele.me. Each dataset contains users' purchase sequences and associated attributes. The attributes mainly include geospatial features, such as the geohash6 and AOI, where users and takeaways are located, temporal features, such as the timestamps and weekdays when users purchase takeaways, and additional user and takeaway attributes. To ensure the quality of data, we apply the following processing steps. Initially, we remove the users and takeaways with empty geohash6 attribute values to clean the dataset. Secondly, we retain features from the cleaned dataset and calculate the spherical distance as a spatial distance feature based on the geohash6 values of users and takeaways. Finally, for each user purchase sequence, the last purchased takeaway is designated as test data, the second-to-last as validation data, and the rest as training data. A statistical summary of the processed datasets is shown in Table 1.

City	# Samples	# Users	# Takeaways	# Entities	# Relations	Avg. length
Wuhan	2,125,761	31,047	383,953	447,033	4,906	41.27
Sanya	3,479,272	19,842	168,332	234,307	3,795	42.02
Taiyuan	4,442,984	32,412	306,344	396,153	3,669	39.97

Table 1: Statistics of the processed datasets.

Baselines To evaluate the effectiveness of our model, we compare it with the following nine representative baselines: Caser (Tang and Wang 2018), GRU4Rec (Hidasi et al. 2015), SASRec (Kang and McAuley 2018), BERT4Rec (Sun et al. 2019), DuoRec (Qiu et al. 2022), FEARec (Du et al. 2023b), GCL4SR (Zhang et al. 2022), MAERec (Ye, Xia, and Huang 2023), and BSARec (Shin et al. 2024).

Implementation Details All evaluation methods are implemented in PyTorch. The hyper-parameters for these methods are chosen according to the original papers, and the optimal settings are selected based on model performance on the validation data. For STKDRec, we conduct experiments with the following hyper-parameters. For the STKG encoder, the sampling depth m is set to 2, and the number of neighbor nodes sampled s is selected from {[5,5], [10,10], [15,15], [20,20]}. For the ST-Transformer module, we set the number of self-attention blocks and attention heads to 2 and the embedding dimension to 256. For STKD, the temperature τ is selected from $\{1, 3, 5, 7, 9\}$, and the α is fixed at 0.2. We use Adam as the optimizer, with the learning rate, β_1, and β_2 set to 0.001, 0.9, and 0.98, respectively. The batch size and maximum sequence length n are both set to 128. An early stopping strategy is applied based on the performance of the validation data. The implementation is carried out in PyTorch on a single NVIDIA A40 GPU with 48GB of memory.

Metrics To measure the accuracy of recommendations, we use the widely adopted Top-k metrics HR@k (Hit Rate) and NDCG@k (Normalized Discounted Cumulative Gain), with k set to 5, 10, and 20. HR@k calculates the frequency with which the actual next takeaway appears within the top-k recommendations, while NDCG@k is a position-aware metric assigning higher weights to takeaways ranked higher. To

Dataset	Metric	Caser	GRU4Rec	BERT4Rec	MAERec	DuoRec	FEARec	SASRec	GCL4SR	BSARec	STKDRec
Wuhan	HR@5	0.5797	0.7530	0.6299	0.7441	0.7659	0.7652	0.7420	0.7625	0.7643	**0.7909**
	HR@10	0.6534	0.7854	0.6887	0.7976	0.7913	0.7904	0.7796	0.7938	0.7958	**0.8229**
	HR@20	0.7225	0.8120	0.7485	0.8170	0.8141	0.8133	0.8160	0.8242	0.8252	**0.8659**
	NDCG@5	0.4690	0.6806	0.5631	0.6245	0.7210	0.7198	0.6789	0.7159	0.7216	**0.7463**
	NDCG@10	0.4929	0.6943	0.5821	0.6485	0.7292	0.7279	0.6911	0.7261	0.7318	**0.7586**
	NDCG@20	0.5104	0.7036	0.5972	0.6661	0.7350	0.7338	0.7003	0.7340	0.7393	**0.7694**
Sanya	HR@5	0.8338	0.8580	0.7965	0.7983	0.8561	0.8578	0.8560	0.8566	0.8623	**0.8770**
	HR@10	0.8679	0.8715	0.8294	0.8612	0.8747	0.8769	0.8806	0.8825	0.8838	**0.8940**
	HR@20	0.8931	0.8894	0.8627	0.8862	0.8884	0.8908	0.8987	0.8992	0.9005	**0.9106**
	NDCG@5	0.7326	0.7967	0.7410	0.6478	0.8072	0.8095	0.8004	0.8060	0.8189	**0.8437**
	NDCG@10	0.7446	0.8044	0.7514	0.6750	0.8133	0.8157	0.8084	0.8145	0.8259	**0.8492**
	NDCG@20	0.7512	0.8089	0.7594	0.6840	0.8168	0.8192	0.8130	0.8217	0.8301	**0.8534**
Taiyuan	HR@5	0.7329	0.8259	0.7616	0.8112	0.8343	0.8345	0.8319	0.8541	0.8419	**0.8630**
	HR@10	0.7916	0.8543	0.7985	0.8540	0.8564	0.8533	0.8583	0.8637	0.8661	**0.8789**
	HR@20	0.8396	0.8769	0.8351	0.8781	0.8761	0.8738	0.8813	0.8869	0.8881	**0.8963**
	NDCG@5	0.6175	0.7900	0.7115	0.6965	0.7923	0.7923	0.7774	0.8049	0.8053	**0.8340**
	NDCG@10	0.6366	0.7982	0.7235	0.7138	0.7994	0.7984	0.7860	0.8099	0.8132	**0.8391**
	NDCG@20	0.6488	0.8040	0.7327	0.7225	0.8044	0.8036	0.7918	0.8162	0.8188	**0.8435**

Table 2: Overall performance comparison. The best results are in boldface and the second-best results are underlined.

evaluate STKDRec, we pair the actual takeaway from the test set with 100 randomly sampled negative takeaways that the user has not purchased and subsequently rank them.

Experimental Results

Table 2 presents the overall experimental results of STKDRec and baselines. We have the following vital observations: STKDRec achieves the best performance across all datasets compared to all baselines, highlighting the effectiveness of our STKDRec. The second-best model BSARec considers fine-grained user sequential patterns without accounting for spatial-temporal dependencies and collaborative associations between users and takeaways. These limitations result in BSARec performing worse than STKDRec.

Graph-based methods GCL4Rec and MAERec focus on user-item graphs to capture the collaborative association between users and takeaways but lack the integration of temporal, spatial, and additional auxiliary information, limiting their ability to capture users' preferences over time and geospatial information. In takeaway recommendation scenarios, these factors significantly influence users' purchase behavior. This is why GCL4SR and MAERec perform worse than STKDRec. Feature-enhanced methods DuoRec and FeaRec rely on static features or limited external knowledge, making it difficult to effectively capture the dynamic changes in user preferences. Our STKDRec captures dynamic user preferences on complex geospatial information and integrates the spatial-temporal knowledge within the STKG.

Ablation Studies

To assess the contribution of each component of STKDRec, we perform ablation studies on all datasets. The results of the STKDRec variants are shown in Table 3. *-w/o SP* denotes a variant without spatial position embedding. *-w/o F* and *-w/o C* denote two variants considering only spatial region and spatial distance information separately. *-w/o KD* denotes a variant using only the ST-Transformer module; *-w/o SP + KD* denotes a variant using only the original Transformer to encode sequence. The results in Table 3 show that the variants *-w/o SP*, *-w/o F*, and *-w/o C* consistently perform worse than STKDRec on all datasets, highlighting the importance of geospatial information for modeling dynamic user preferences in takeaway recommendations. The results for *-w/o KD* further indicate that the STKD strategy is crucial for facilitating the fusion of spatial-temporal knowledge from STKG and sequence data. The significant drop in per-

Method	Wuhan		Sanya		Taiyuan	
	H@10	N@10	H@10	N@10	H@10	N@10
-w/o SP+KD	0.7796	0.6911	0.8806	0.8084	0.8583	0.7860
-w/o KD	0.8161	0.7509	0.8838	0.8294	0.8684	0.8277
-w/o SP	0.8097	0.7456	0.8844	0.8348	0.8669	0.8266
-w/o C	0.8181	0.7518	0.8875	0.8379	0.8685	0.8267
-w/o F	0.8178	0.7535	0.8877	0.8384	0.8688	0.8277
STKDRec	**0.8229**	**0.7586**	**0.8940**	**0.8492**	**0.8789**	**0.8391**

Table 3: The results of ablation studies.

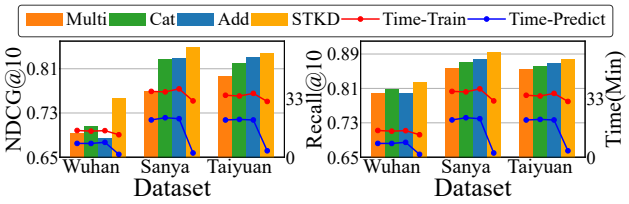

Figure 3: Study on different knowledge fusion methods. Multi refers to *multiplication*, Cat refers to *concatenation*, Add refers to *addition*, STKD denotes our proposed strategy, and Time indicates model training and prediction duration.

formance for *-w/o KD+SP* demonstrates the effectiveness of all the proposed improvements.

To validate the efficacy of the STKD strategy, we replace it with other knowledge fusion strategies, including *concatenation*, *addition*, and *multiplication*. The results are shown in Figure 3. STKDRec achieves superior performance and better training and prediction efficiency than other variants. This is because the knowledge distributions of the STKG encoder and the ST-Transformer differ significantly and exist in distinct vector spaces. Simple combining them fails to effectively integrate the differing distributions and may transform the distribution of the STKG encoder into noise, thereby damaging the distribution of the ST-Transformer. In addition, these variants necessitate additional computations to align the distinct vector spaces, which further compromises the overall efficiency of the model. STKD directly aligns the predicted distributions of the ST-Transformer and STKG encoder, facilitating smoother and more effective knowledge transfer.

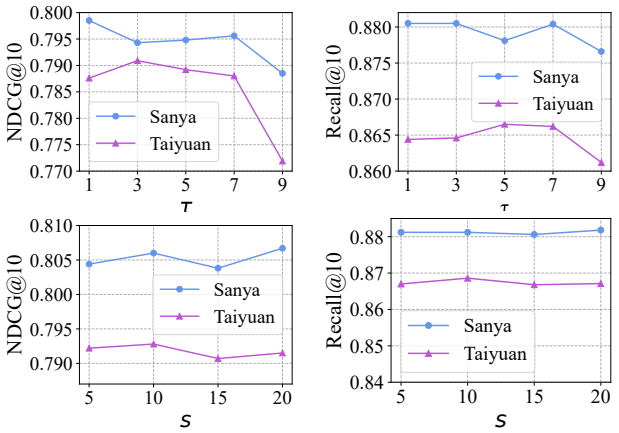

Figure 4: Study on different temperature τ and the number of neighbor nodes sampled s.

Parameter Sensitivity Study

We conduct experiments on the Sanya and Taiyuan datasets to investigate the impact of two hyper-parameters: the number of neighbor nodes sampled s in the STKG and the temperature coefficient τ for STKD. Figure 4 illustrates the performance of STKDRec on the Sanya and Taiyuan datasets under various settings for s and τ. For the neighbor sample size s, changing its value results in minimal changes in model performance, indicating that STKDRec is not sensitive to the sampling size s. Therefore, selecting the appropriate parameter is essential for optimal results. The optimal value for the Sanya dataset is 20, while for the Taiyuan dataset is 10. For the temperature coefficient τ, when τ exceeds 7, the model's performance significantly declines, while when τ is less than 7, the model's performance remains relatively stable. This indicates that an excessively large coefficient weakens the teacher model's information, making it difficult for the student model to learn.

Case Study

To evaluate the impact of the STKD strategy, we conduct a case study comparing the recommendation results of STKDRec and its variant *-w/o KD*. Given user u420's historical purchase sequence, Figure 5 illustrates the Top-5 takeaway recommendation rankings and their corresponding probabilities for u420 at 4:40 PM on Wednesday at their workplace. Coffee and donuts exhibit strong spatial-temporal dependencies and frequent co-purchasing behaviors, driven by their recurrent joint consumption in workplace areas during afternoon tea time. The variant *-w/o KD* only relies on the user's historical purchasing data, failing to account for these collaborative associations, which leads to incorrect recommendations. However, by incorporating STKG, STKDRec can capture the spatial-temporal and collaborative relationships between coffee and donuts, allowing it to recommend the donuts for u420 accurately. This case further demonstrates the effectiveness of utilizing the STKD strategy to integrate the spatial-temporal knowledge from both STKG and sequence data.

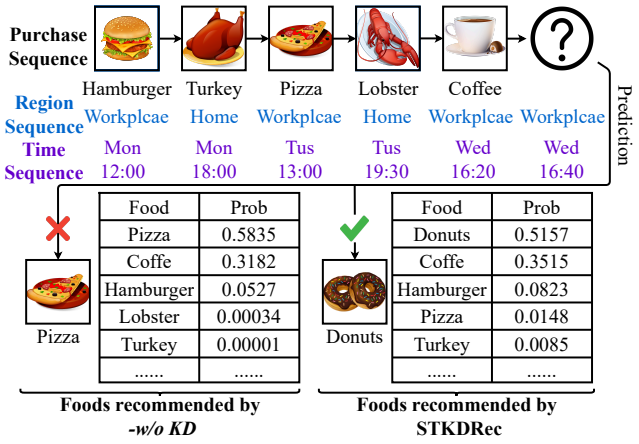

Figure 5: Visualization of case study on recommendation results.

Conclusion

In this paper, we propose a novel spatial-temporal knowledge distillation model for takeaway recommendation, termed STKDRec. The model involves two stages: pre-training and STKD. During the pre-training stage, we train an STKG encoder to extract rich spatial-temporal knowledge from STKG. During the STKD stage, we apply an ST-Transformer to capture dynamic user preferences on fine-grained spatial region and spatial distance information from a sequential perspective. The STKD strategy is utilized to integrate heterogeneous spatial-temporal knowledge from both STKG and sequence data while reducing computational overhead. Experimental results on three real datasets show that the performance of STKDRec is significantly better than the state-of-the-art baselines methods.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62272033).

References

Chen, W.; Wan, H.; Guo, S.; Huang, H.; Zheng, S.; Li, J.; Lin, S.; and Lin, Y. 2022. Building and exploiting spatial–temporal knowledge graph for next POI recommendation. *Knowledge-Based Systems*, 258: 109951.

Chen, W.; Wan, H.; Wu, Y.; Zhao, S.; Cheng, J.; Li, Y.; and Lin, Y. 2024. Local-global history-aware contrastive learning for temporal knowledge graph reasoning. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, 733–746. 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*.

Du, B.; Lin, S.; Gao, J.; Ji, X.; Wang, M.; Zhou, T.; He, H.; Jia, J.; and Hu, N. 2023a. BASM: A bottom-up adaptive spatiotemporal model for online food ordering service. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, 3549–3562. IEEE.

Du, X.; Yuan, H.; Zhao, P.; Qu, J.; Zhuang, F.; Liu, G.; Liu, Y.; and Sheng, V. S. 2023b. Frequency enhanced hybrid attention network for sequential recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 78–88.

Gao, Y.; He, Y.; Kan, Z.; Han, Y.; Qiao, L.; and Li, D. 2023. Learning joint structural and temporal contextualized knowledge embeddings for temporal knowledge graph completion. In *Findings of the Association for Computational Linguistics: ACL 2023*, 417–430.

Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Hidasi, B.; Karatzoglou, A.; Baltrunas, L.; and Tikk, D. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Kang, S.; Hwang, J.; Kweon, W.; and Yu, H. 2021. Topology distillation for recommender system. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 829–839.

Kang, S.; Kweon, W.; Lee, D.; Lian, J.; Xie, X.; and Yu, H. 2023. Distillation from heterogeneous models for top-k recommendation. In *Proceedings of the ACM Web Conference 2023*, 801–811.

Kang, W.-C.; and McAuley, J. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*, 197–206. IEEE.

Lin, S.; Yu, Y.; Ji, X.; Zhou, T.; He, H.; Sang, Z.; Jia, J.; Cao, G.; and Hu, N. 2022. Spatiotemporal-enhanced network for click-through rate prediction in location-based services. *arXiv preprint arXiv:2209.09427*.

Liu, H.; Zhu, Y.; and Wu, Z. 2023. Knowledge graph-based behavior denoising and preference learning for sequential recommendation. *IEEE Transactions on Knowledge and Data Engineering*.

Qiu, R.; Huang, Z.; Yin, H.; and Wang, Z. 2022. Contrastive learning for representation degeneration problem in sequential recommendation. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, 813–823.

Rendle, S.; Freudenthaler, C.; and Schmidt-Thieme, L. 2010. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*, 811–820.

Shi, L.; Yang, J.; Lv, P.; Yuan, L.; Kou, F.; Luo, J.; and Xu, M. 2024. Self-derived knowledge graph contrastive learning for recommendation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7571–7580.

Shin, Y.; Choi, J.; Wi, H.; and Park, N. 2024. An attentive inductive bias for sequential recommendation beyond the self-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 8984–8992.

Sun, F.; Liu, J.; Wu, J.; Pei, C.; Lin, X.; Ou, W.; and Jiang, P. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*, 1441–1450.

Tang, J.; and Wang, K. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the eleventh ACM international conference on web search and data mining*, 565–573.

Wu, S.; Wan, H.; Chen, W.; Wu, Y.; Shen, J.; and Lin, Y. 2023. Towards enhancing relational rules for knowledge graph link prediction. *arXiv preprint arXiv:2310.13411*.

Xia, X.; Yin, H.; Yu, J.; Wang, Q.; Xu, G.; and Nguyen, Q. V. H. 2022. On-device next-item recommendation with self-supervised knowledge distillation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 546–555.

Ye, Y.; Xia, L.; and Huang, C. 2023. Graph masked autoencoder for sequential recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 321–330.

Zhang, S.; Liu, Y.; Sun, Y.; and Shah, N. 2021. Graph-less neural networks: Teaching old mlps new tricks via distillation. *arXiv preprint arXiv:2110.08727*.

Zhang, Y.; Liu, Y.; Xu, Y.; Xiong, H.; Lei, C.; He, W.; Cui, L.; and Miao, C. 2022. Enhancing sequential recommendation with graph contrastive learning. *arXiv preprint arXiv:2205.14837*.

Zhang, Y.; Wu, Y.; Le, R.; Zhu, Y.; Zhuang, F.; Han, R.; Li, X.; Lin, W.; An, Z.; and Xu, Y. 2023. Modeling dual period-varying preferences for takeaway recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 5628–5638.

Zhu, Q.; Chen, X.; Wu, P.; Liu, J.; and Zhao, D. 2021. Combining curriculum learning and knowledge distillation for dialogue generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 1284–1295.